Title: Exploring Chromosomal Structural Heterogeneity Across Multiple Cell Lines

Authors: Ryan R. Cheng[1, *], Vinicius Contessoto[1,2], Erez Lieberman Aiden[1, 3], Peter G. Wolynes[1, 4, 5], Michele Di Pierro[1, 6, *], José N. Onuchic[1, 5, *]

Affiliations:

[1] Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005.

[2] Brazilian Biorenewables National Laboratory - LNBR, Brazilian Center for Research in Energy and Materials – CNPEM, Campinas, SP, Brazil.

[3] Center for Genome Architecture, Baylor College of Medicine, Houston, Texas 77030.

[4] Department of Chemistry, Rice University, Houston, Texas 77005.

[5] Department of Physics & Astronomy, Rice University, Houston, Texas 77005.

[6] Department of Physics, Northeastern University, Boston, Massachusetts 02115.

*To whom the correspondence should be addressed: ryan.r.cheng@gmail.com, m.dipierro@northeastern.edu, jonuchic@rice.edu

Abstract:

We study the structural ensembles of human chromosomes across different cell types. Using computer simulations, we generate cell-specific 3D chromosomal structures and compare them to recently published chromatin structures obtained through microscopy.

We demonstrate using a combination of machine learning and polymer physics simulations that epigenetic information can be used to predict the structural ensembles of multiple human cell lines. The chromosomal structures obtained *in silico* are quantitatively consistent with those obtained through microscopy as well as DNA-DNA proximity ligation assays. Theory predicts that chromosome structures are fluid and can only be described by an ensemble, which is consistent with the observation that chromosomes exhibit no unique fold. Nevertheless, our analysis of both structures from simulation and microscopy reveals that short segments of chromatin make transitions between a closed conformation and an open dumbbell conformation. This conformational transition appears to be consistent with a two-state process with an effective free energy cost of about four times the effective information theoretic temperature. Finally, we study the conformational changes associated with the switching of genomic compartments observed in human cell lines. Genetically identical but epigenetically distinct cell types appear to rearrange their respective structural ensembles to expose segments of transcriptionally active chromatin, belonging to the A genomic compartment, towards the surface of the chromosome, while inactive segments, belonging to the B compartment, move to the interior. The formation of genomic compartments resembles hydrophobic collapse in protein folding, with the aggregation of denser and predominantly inactive chromatin driving the positioning of active chromatin toward the surface of individual chromosomal territories.

Introduction:

The 3D spatial organization of the chromosomes in the nucleus of eukaryotic cells appears to be cell type specific[1-7]. What determines this cell type specific organization and how that organization relates to patterns of gene expression remain crucial questions in structural genomics.

DNA-DNA ligation experiments have revealed spatial compartmentalization, generally termed A/B compartmentalization[8], and CTCF-mediated loop domains. It was observed that the A compartment chromatin contains a larger amount of the expressed genes while the B compartment chromatin is less transcriptionally active. Similar A/B compartmentalization has been observed across human cell lines[1-3] as well as in other species [2, 4, 9-12], suggesting that compartmentalization is a conserved feature of genome organization across evolution. While single-cell structures can be interrogated using proximity ligation assays[13, 14], high resolution has so far only been achieved through ligation methods when the experiments are performed over a large population of cells, thus averaging over the respective individual 3D structures.

Recent microscopy approaches have begun to reveal the 3D structures of segments of chromatin longer than a megabase at a spatial resolution on the nanometer scale[15-18]. These approaches not only allow for the quantification of pairwise and higher order interactions between loci, but also allow some quantification of the structural variability in a population of cells. One consistent observation from the imaging approaches, as well as from single cell Hi-C [13, 14, 19], has been the high degree of structural variability seen within an apparently homogeneous population of synchronized cells of a single cell type.

Despite this variability, well-defined cell type-specific DNA-DNA ligation maps for the ensemble emerge after population averaging the single cell results.

Polymer models [20-29] that describe the process of chromosome organization have been proposed. In particular, physics-based modeling established using energy landscape theory[26, 27, 30-32] directly simulates the microscopic details of structural variability that have been shown to recapitulate the population averaged DNA-DNA ligation maps. The Minimal Chromatin Model (MiChroM)[27] is a coarse-grained polymeric representation of chromosomes confined in a nucleus. MiChroM is an effective energy landscape model describing chromosomes as polymers subject to interactions which depend on chromatin biochemical composition and on the genomic distance separating any two loci[26, 27]. In MiChroM, transient binding interactions among the chromosomal loci result in the emergence of compartmentalization through a process of phase separation, in which chromatin of the same type preferentially co-localizes. It was subsequently shown that the propensity toward phase separation can be reliably predicted using epigenetic marking data; thus, the information contained within the 1D epigenetic marking patterns decorating the chromatin polymer appears to be sufficient to predict the ensemble of 3D chromosome structures for a given cell type. A neural network called MEGABASE[30] was trained to quantify the statistical relationship between the experimental sub-compartment annotations and the histone methylation and acetylation markings tracks, as assayed using chromatin immunoprecipitation data. Once trained, MEGABASE can be used to predict the compartmentalization patterns of a chromosome using a set of epigenetic ChIP-Seq tracks

as the sole input. Combining MEGABASE and MiChroM, we are able to simulate the structural dynamics of chromosomes using Langevin dynamics on the energy landscape.

In this manuscript we provide an analysis of the chromosomal structural ensembles of different cell types, using computer simulations based on MEGABASE+MiChroM and compare them with the structural ensembles obtained through microscopy[15]. This comparison further validates the present theory and computational models, showing that not only the population averages but also the structural heterogeneity that is observed in human chromosomes in the interphase are consistent with the energy landscape model. We first use the MEGABASE+MiChroM computational pipeline[30] to predict the 3D ensemble of chromosomal structures for several well-studied cell types: HMEC, HUVEC, IMR90, K562, HeLa-S3, and H1-hESC. To test these simulated 3D ensembles, we then generate ensemble averaged simulated ligation maps that are compared directly to population averaged DNA-DNA ligation maps[1, 2]. For the cell lines IMR90 and K562, we also use energy landscape tools to analyze the structures obtained through diffraction-limited microscopy by Bintu et al[15] for short ~2 Mb segments of chromatin and compare the experimental structural ensembles directly with the corresponding regions of the simulated chromosome 21 for IMR90 and K562. Chromosomes do not adopt a single structure in the interphase, but rather, exhibit a high structural variability characteristic of a phase-separated liquid. Using order parameters commonly used to quantify structural similarity in protein folding theory, we uncover two dominant clusters of structures in both the experimental and simulated structural ensembles. The transition from a closed structure to an open set of structures appears to be governed by a two-state process with an apparent

free energy cost of about four times the effective information theoretic temperature. Additionally, we further examine the structural differences between whole chromosomes belonging to different cell types. The simulations show that inactive segments of chromatin move to the interior of the chromosome, while gene active chromatin moves to the chromosome surface. This effect appears to be driven by the favorable effective interactions between loci belonging to the B compartment, which forms a stable interior core; a phenomenon reminiscent of the hydrophobic collapse much studied in protein folding.

Results & Discussion

A polymer model of chromatin based on epigenetics features captures chromosome organization across different cell types.

We previously developed a computational pipeline that can predict the 3D ensemble of chromosome structures by using chromatin immunoprecipitation tracks for histone modifications as input [30]. This approach was successfully used to predict the 3D chromosome structures for human lymphoblastoid cells (GM12878) using the experimental ChIP-Seq tracks for 11 histone modifications[30], i.e., H2AFZ, H3K27ac, H3K27me3, H3K4me1, H3K4me2, H3K4me3, H3k79me2, H3K9ac, H3K9me3, and H4K20me1. Predicted chromosome structures for human lymphoblastoid cells (GM12878) were found to be consistent with both DNA-DNA ligation and fluorescence in situ

hybridization (FISH) experiments[1]. Here we test the robustness of the model's predictions beyond GM12878, to other well studied cell lines for which we have found sufficient epigenetic marking data.

Using the MEGABASE neural network, which was previously trained using data from GM12878, and sourcing from the Encyclopedia of DNA Elements (ENCODE) database the ChIP-Seq tracks for the same 11 histone modifications previously used, sub-compartment annotations for all the autosomes of cell lines were generated that had never been used in the training phase of the neural network. These sequences of sub-compartment annotations, or chromatin types, then serve as input for molecular dynamics simulations using the Minimal Chromatin Model (MiChroM)[27]. Using this combined approach, the chromosomal structural ensembles for 6 additional cell lines were generated: human fetal lung cells (IMR-90), human umbilical vein endothelial cells (HUVEC), immortalized myelogenous leukemia cells (K562), human mammary epithelial cells (HMEC), human embryonic stem cells (H1-hESC), and HeLa-S3 cells.

For each cell type, averaging the simulated ensemble generates *in silico* DNA-DNA ligation maps, which are in excellent agreement with those determined experimentally. Figure 1 shows the comparison between simulated and experimental maps for IMR90 (Figure 1A), HUVEC (Figure 1B) and K562 (Figure 1C), demonstrating quantitative agreement. Corresponding comparisons of the compartmentalization patterns are also provided in Figure S1 for additional cell types HMEC, HeLa-S3, and H1-hESC, as well as for GM12878 in Ref:[30]. The annotations generated by MEGABASE agree with the

compartment annotations obtained from the ligation maps (Figures 1A-C; Figure S1). In particular, the Pearson's *R* between the simulated and experimental maps of matching cell type as a function of genomic distance shows that the long-range patterns of compartmentalization are captured over tens of mega-bases. To establish a term of comparison we calculated the Pearson's *R* between the experimental DNA-DNA ligation maps of mismatching cell types. While the experimental observations on different cell lines do correlate with each other, computational modeling delineates the difference between cell type and appears to best match the experimental map when the cell types of simulation and experiment are matched up. This last result demonstrates that the present theoretical model discriminates well between different cell lines.

While we have focused so far on the spatial organization of entire chromosomes on the micrometer length scale, for a better comparison with the structures of chromosome 21 of IMR90 and K562 obtained from microscopy[15], we have also incorporated in the polymer physics simulation the loops mediated by the activity of the protein CTCF.

Figure 1D shows that the inclusion of CTCF-loops, which are easily be incorporated into the model, improves the quality of the results for the shorter range features of chromosome organization within 10Mb in genomic distance; at larger length scales the model appears to be completely insensitive of CTCF-mediated loops.

Figure 1E shows the Pearson's *R* between the AB annotation vectors derived directly from the DNA-DNA ligation maps and those obtained from MiChroM simulations for different

cell types. The diagonal of Figure 1E corresponds to the Pearson's *R* between AB annotations derived from experiment and simulation of matching cell types. The simulated and experimental annotations for the same cell types agree well with each other. Figure S1E shows the Pearson's *R* between AB annotations derived from experiment alone for the different cell types. Notably, the high degree of correlation between the myelogenous leukemia cell line K562 and human lymphoblastoids (GM12878) maps observed in Figure 1E is apparent from DNA-DNA ligation maps alone (Figure S1E). The agreement between the simulated and experimental A/B annotations is the highest quality (Pearson's $r \sim 0.9$) for the DNA-DNA ligation maps of GM12878, which is not surprising since the GM12878 has an order of magnitude more reads than any other map and consequently has the highest resolution.

Taken together, these results demonstrate that long range compartmentalization observed in the DNA-DNA ligation maps is well captured by the simulated structural ensembles for these well-studied cell lines using only information about the epigenetic marking patterns as input.

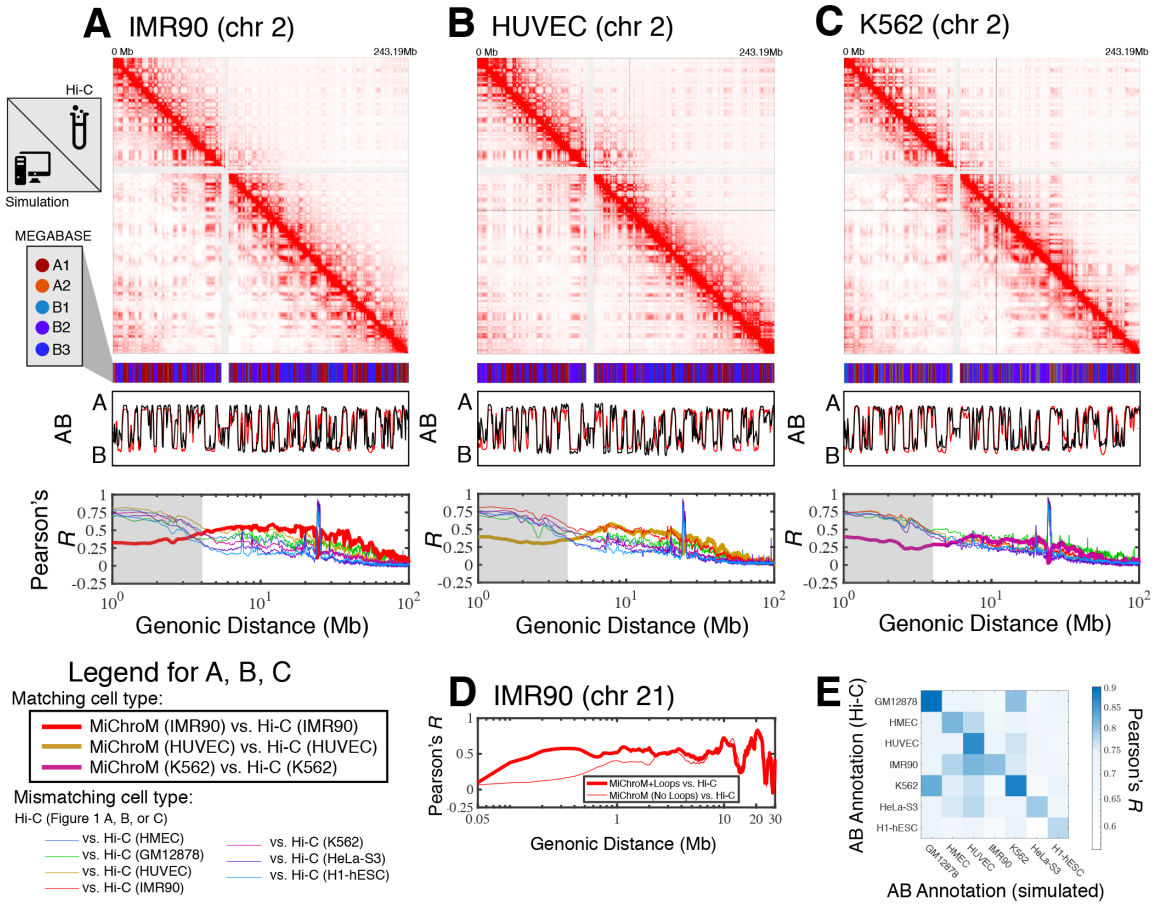

Figure 1. Prediction of chromosome structures for differentiated cell lines and for immortalized leukemia cells. The 3D ensemble of chromosome structures was predicted for the cell types (A) IMR90, (B) HUVEC, and (C) K562 using the ChIP-Seq histone modification tracks for the respective cell lines found on ENCODE—shown are the structural predictions for chromosome 2. As validation, the chromosome structures were compared with the DNA-DNA ligation experiments of Rao et al[1], where the simulated map is shown on the top right triangle and the experimental map is shown on the bottom left triangle. The datasets are visualized using Juicebox[33]. The MEGABASE chromatin type annotation is shown as a color vector under the contact probability map, followed by the A/B compartment

annotation[1] for the simulated map (red) and the experimental map (black), respectively. The Pearson's *R* between the simulated and experimental contact maps for fixed genomic distances are plotted for the cell types IMR90, HUVEC, and K562, respectively, in thick lines. The Pearson's *R* between the experimental maps of mismatching cell types are also shown with thin lines—See Legend. The shaded region highlights that at relatively short genomic distances (<10 Mb), excluding CTCF-mediated loops from the simulation results in disagreement between the simulated and experimental maps. When loops are included in the simulations, the agreement between the simulation and experiment is recovered at the short genomic distances. (D) Pearson's *R* as a function of genomic distance is plotted between the experimental map for chromosome 21 (IMR90) and MiChroM simulation with loops (thick red line) and without loops (thin red line). (E) A matrix of Pearson's *R* between the AB annotation of the experimental ligation map and the simulated contact maps for different cell types, respectively. The high Pearson's *R* signifies the consistency between the simulated maps and the experimental DNA-DNA ligation maps.

Chromatin structural ensembles from DNA tracing reveal coexistence of open and closed structures.

Recent developments in DNA-tracing have allowed the direct experimental determination of three-dimensional structures using diffraction-limited and super-resolution microscopy[15-18]. DNA tracing is a technique that labels consecutive stretches of DNA

with optical probes, which can be used to spatially resolve the positions of those probes using microscopy. It has become increasingly clear that unlike the situation for folded globular proteins, which typically can be reasonably well approximated for many purposes by a single native structure corresponding to the average conformation, chromatin appears to be highly dynamical and cannot be characterized by any single conformation. The heterogeneity of the chromosomal structural ensembles was first suggested by the analysis of the free energy landscape of chromosomes[26, 27] and has been indirectly observed through single cell Hi-C experiments[13, 14, 19]. The heterogeneity has now been confirmed by direct imaging of individual chromosomal structures[15-17]. As a consequence of this conformational plasticity, statistical ensembles [26, 27, 30, 31, 34-39] must be used in order to describe chromosomal structures *in vivo*.

In order to improve our understanding of the genomic structural ensembles, we directly compare chromatin structures imaged using microscopy with those obtained from the simulated structural ensembles.

We first focus on the traced structures of Bintu et al[15], who obtained hundreds of images structures for short ~2Mb segments of chromatin belonging to chromosome 21. These regions are 29.37-31.32Mb (referred to here as Segment 1) of IMR90 and K562 cell types and 20.0-21.9Mb (referred to as Segment 2) of IMR90. Only structures where the positions of over 90% of the loci were resolved are used in the present analysis. There are then 692 usable structures for IMR90 Segment 1, 752 usable structures of IMR90 Segment 2, and 244 usable structures of K562 Segment 1. While MEGABASE+MiChroM, provides us

with structures of entire chromosomes, we focus specifically on Segment 1 and Segment 2 of chromosome 21 for our direct comparison.

As previously reported[15-17], the traced structures can be used to generate a population averaged contact maps, which turn out to be consistent with DNA-DNA ligation maps. Shown in Figure S2A-C are the averaged contact maps for the chromatin Segments 1 (IMR90 & K562) and Segment 2 (IMR90), respectively. Nevertheless, information is lost when converting from a 3D structural ensemble to a 2D contact map.

Focusing on the structural details that cannot be found in a contact map, we make a close examination of the types of structures observed in the tracing dataset using an order parameter commonly used in studying protein folding landscapes, Q, which quantifies the structural similarity between two structures α and β:

$$Q^{\alpha\beta} = \frac{1}{N}\sum_{i<j}\exp\left(-\left(\frac{\left(r_{ij}^{(\alpha)} - r_{ij}^{(\beta)}\right)^2}{2\delta^2}\right)\right) \qquad \text{(Eq. 1)}$$

where $r_{ij}^{(\alpha)}$ and $r_{ij}^{(\beta)}$ are the distances between chromatin loci i and j in structures α and β:, respectively, N is the number of pairs of loci included in the summation, and $\delta = 0.165\mu m$ is the resolution length scale for which deviations in the distances between structures α and β are treated as being similar. The Q between any two structures ranges from 0 (dissimilar) to 1 (identical) over the entire set of pairwise distances between loci. The order parameter Q is not solely based on contacts; a pair of chromatin loci can

contribute to Q even if they are not spatially proximate if they are separated in both structures by a similar distance as set by δ. In this way Q measures structure more stringently than a simple contact map does.

Using $1-Q$ to define the distance between any two structures, hierarchical clustering of the traced structures for Segment 1 was applied to identify clusters having distinct structural features. These cluster sub-ensembles can be considered distinct conformational states. To see whether the segment 1 structures for IMR90 and K562 exhibit a high degree of structural similarity, we combined their datasets before clustering.

When applied to the 936 combined experimental structures for Segment 1, the clustering algorithm yields three distinct clusters. These correspond to a closed dumbbell (Cluster 1), an open dumbbell (Cluster 2), and a highly dense chromatin state (Cluster 3) shown in Figure 2. The closed dumbbell, where the head and tail globular domains are in contact with one another, is the dominant state observed for Segment 1 in both IMR90 and K562, accounting for 97.4% of the imaged structures ($N_{closed} = 912$). Cluster 1 can further be sub-divided into subgroups 1a, 1b, 1c, and 1d (Figure 2), which account for 75.5% of the structures in Cluster 1. The subgroups appear to capture various stages of the process of opening. The structures in subgroup 1a are fully collapsed, while structures in 1b, 1c, and 1d capture the progressive opening of the closed dumbbell. The distribution of the Radius of Gyration for structures belonging to sub-clusters 1a-1d is shown in Figure S3. The open dumbbell structures where the head and tail domains have dissociated from one another, account for approximately 1.8% of the imaged data ($N_{open} = 17$). Additionally, 7 dense,

highly compact structures were identified from clustering. Representative structures from the three clustered structural groups are shown in Figure 2 and the corresponding population averaged contact maps are shown in Figure 2B and 2C for the closed and open structures, respectively.

The high density chromatin, cluster 3, which was found when imaging both Segment 1 and Segment 2[15], is characterized by an extraordinarily high density of DNA $5.8\times10^4 kb/\mu m^3 \sim 4\times10^7 mg/ml$, as estimated for naked dsDNA. For comparison, the density of heterochromatin that is estimated using microscopy data is $\sim 200 mg/ml$ [40](5 orders of magnitude smaller); for this reason, we believe that these chromatin conformations are likely artifacts of the experimental protocol. We therefore have excluded Cluster 3 from further analysis.

Assuming that the opening of the chromatin Segment 1 is in an effective thermodynamic equilibrium would imply a relative stability of $\log\left(N_{closed}/N_{open}\right) = E_{open} - E_{closed} \sim 4k_B T$, where $E_{open} - E_{closed}$ is an apparent free energy difference between the closed and open states and T is an information theoretic temperature characterizing the ensemble[41]. Interestingly, the relative number of open and closed structures found in the simulations (discussed in the proceeding section) is in remarkable agreement with this experimental finding.

We then used the radius of gyration, R_g, as an additional order parameter for the structural ensembles of segment 1 belonging to IMR90 and K562 (Figure 2D). A corresponding potential of mean force can be extracted from the distribution of R_g as $PMF = -k_B T \log P(R_g)$, which also shows the free energy difference of $\sim 4k_B T$ between the closed (Cluster 1) and open (Cluster 2) structural sub-ensembles. The distributions of R_g are also shown for Clusters 1 and 2 in Figure 2D. The open conformations (Cluster 2) possibly belong to a free energy minima in the PMF located between between $R_g \sim 0.6 - 0.8 \mu m$, although additional statistics would be necessary to firmly establish the presence of this additional conformational mode. Interestingly, the vast majority of genes appear to be positioned along the linker region connecting the two globular domains (Figure 2E).

Unlike Segment 1, Segment 2 of IMR90 completely lacks loop domains and, consequently, the averaged contact maps for Segment 2 exhibit no additional features beyond the decay in contact probability as a function of genomic distance (Figure S2). Structural analysis reveals that, without the presence of loop domains, Segment 2 is highly disordered; while clustering reveals open and closed structures, the lack of loop domains and domain boundaries results in the loss of dumbbell-like structures (Figure S4). It should be noted that unlike Segment 1, Segment 2 has an absence of genes (Figure S7).

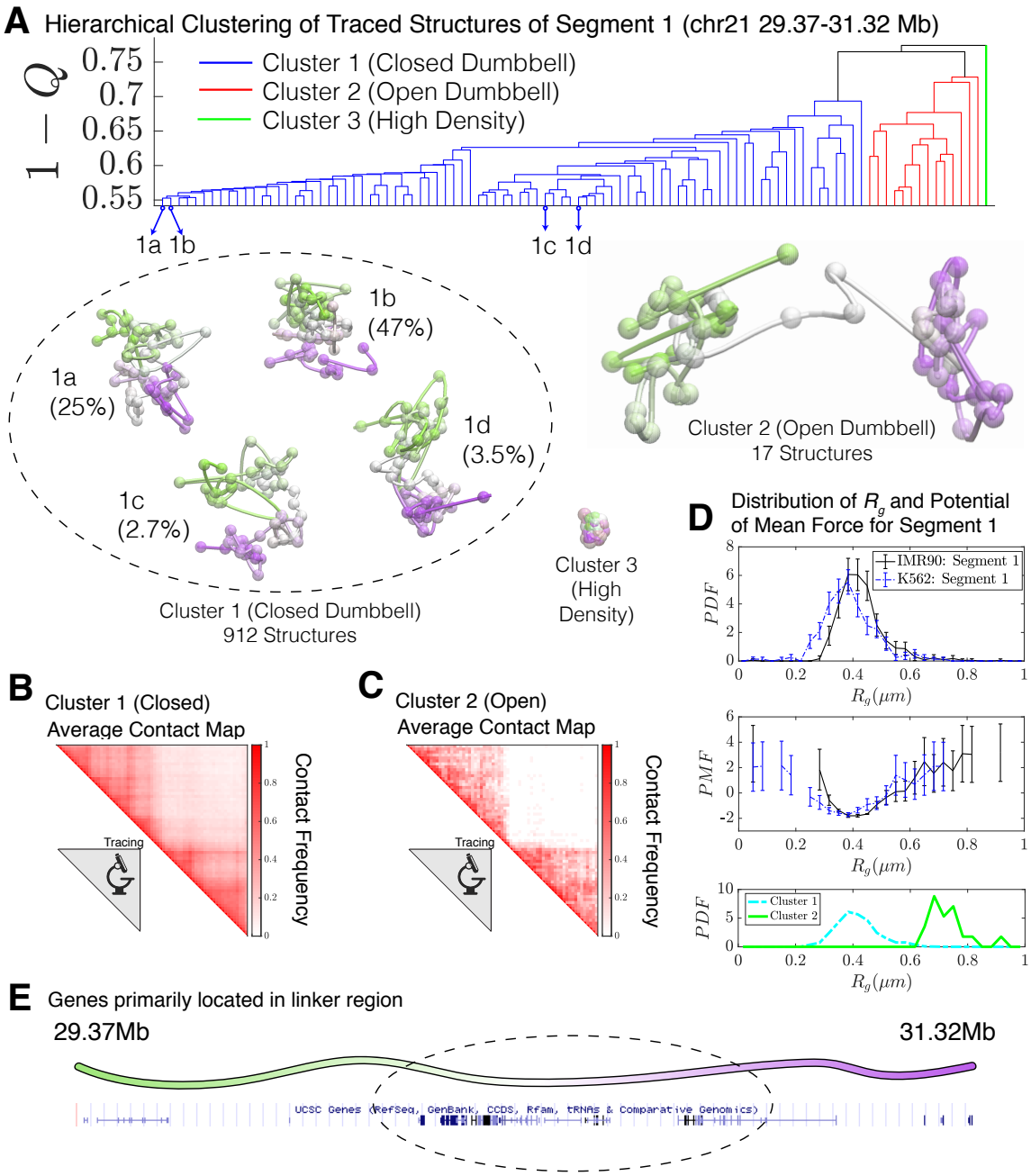

Figure 2. Hierarchical Clustering and the detailed structural analysis of traced Segment 1. (A) The dendrogram representation of the hierarchical clustering of Segment 1 (chr21 29.37-31.32Mb for IMR90 and K562 of [15]), where $1-Q$ is used as the distance between two structures. The clustering reveals three main clusters: closed dumbbell, open

dumbbell, and highly dense structures. Further analysis of Cluster 1 reveals the presence of sub-clusters labeled 1a-1d that represent the gradual opening of the closed dumbbell. Representative traced structures are shown for each of the clusters and sub-clusters. The population averaged contact maps for the closed and open structure clusters are shown respectively in (B) and (C), where 165nm is used to define a contact between two 30kb loci. (D) The distribution of the Radius of Gyration (top), the corresponding potential of mean force (center), and the distributions of Radius of Gyration for Cluster 1 and Cluster 2 (bottom) are shown for the traced structures of Segment 1 of IMR90 and K562. The distribution exhibits a heavy tail to the right of the average value, indicating the existence of open, elongated structures. (E) The UCSC Genes track is plotted along the genomic positions of Segment 1 using the Genome Browser [42].

The chromosomal structures obtained from physical modeling are consistent with those observed with microscopy.

We compare the chromosome structures sampled in the simulations to the diffraction-limited microscopy structures of Bintu et al[15], finding that the conformational states observed using microscopy are also found in the simulated structural ensemble without any calibration or fine tuning of parameters.

It is important to note that the simulated model, and the structural variability that it captures, was derived from the energy landscape learned from population-averaged DNA-DNA ligation data using the principle of maximum entropy[27]. MiChroM has been shown to be consistent with experimental ligation maps (Figure 1 and Refs:[27, 30, 32]), as well

as the distribution of distances between Fluorescence in Situ Hybridization (FISH) probes[30] and several observations regarding chromatin dynamics[31].

Using the $1-Q$ as the distance between all simulated structures for Segment 1, we now performed hierarchical clustering of the simulated structures. The dendrogram of this clustering is shown in Figure 3A, which uncovers two main clusters in the structural ensemble: a closed dumbbell (Cluster 1) and an open dumbbell (Cluster 2). The closed and open structures are consistent with those observed in the Bintu et al[15] datasets. The representative structures of the closed and open conformations are shown in Figure 3, alongside the averaged contact maps for each of the clusters (Figure 3B-C), which are consistent with those determined experimentally (Shown in Figure 2B-C). The simulated Cluster 1 can again further be sub-divided into subgroups; 1α, 1β, 1γ, and 1δ represent the 4 most populated sub-groups (Figure 3), which comprise 66% of the simulated structures of Cluster 1. The subgroups appear to capture various stages of the process of opening. The structures in subgroup 1α are fully collapsed, while structures in 1β, 1γ, and 1δ capture the progressive opening of the closed dumbbell. The Radius of Gyration of sub-clusters 1α-1δ are shown in Figure S5.

No highly dense structures exist in the simulations. Such structures would collapse the entire chromatin segment to the volume of a single monomer, an occurrence that is prohibited by the energy function used to model the system. This is in harmony with our view that Cluster 3 seen in the experiments are artifacts of some sort.

For Segment 1, we performed our analysis on a set of 6400 structures, a representative subset of the simulated trajectories. Both closed ($N_{closed} = 6275$) and open structures $N_{open} = 125$) were identified by the clustering algorithm. Since MiChroM assumes an effective equilibrium thermodynamics representation of chromosome structures and dynamics, we can quickly calculate the relative stability between closed and open structures in the simulated ensemble as $\log\left(N_{closed} / N_{open}\right) = E_{open} - E_{closed} \sim 4k_B T$, where $E_{open} - E_{closed}$ is the effective free energy difference between the closed and open states. This free energy difference is remarkably consistent with the value estimated using only the experimentally traced structures in the preceeding section.

Finally, we calculated the distribution of the radius of gyration, R_g , for the experimetal traced structures of Bintu et al [15] and for the simulated MiChroM structures for Segment 1 belonging to IMR90 and K562 (shown in Figure 3D and Figure 3E respectively). Using a length scale calibrated previously[30] from a single FISH experiment of 0.165μm yields excellent quantitative agreement between the experimentally observed structures and those predicted *de novo* from simulation. It is particularly remarkable that any discrepancies between the experimental and simulated datasets can in fact be captured within 5% error of our original length estimate (Figure S6). Similarly, Figure 3F shows the direct comparison between the distribution of R_g for Segment 1 as well as the corresponding potential of mean force. We see then that MiChroM appears to reproduce the relative apparent free energy differences between open and closed structures found using the experimentally traced structures.

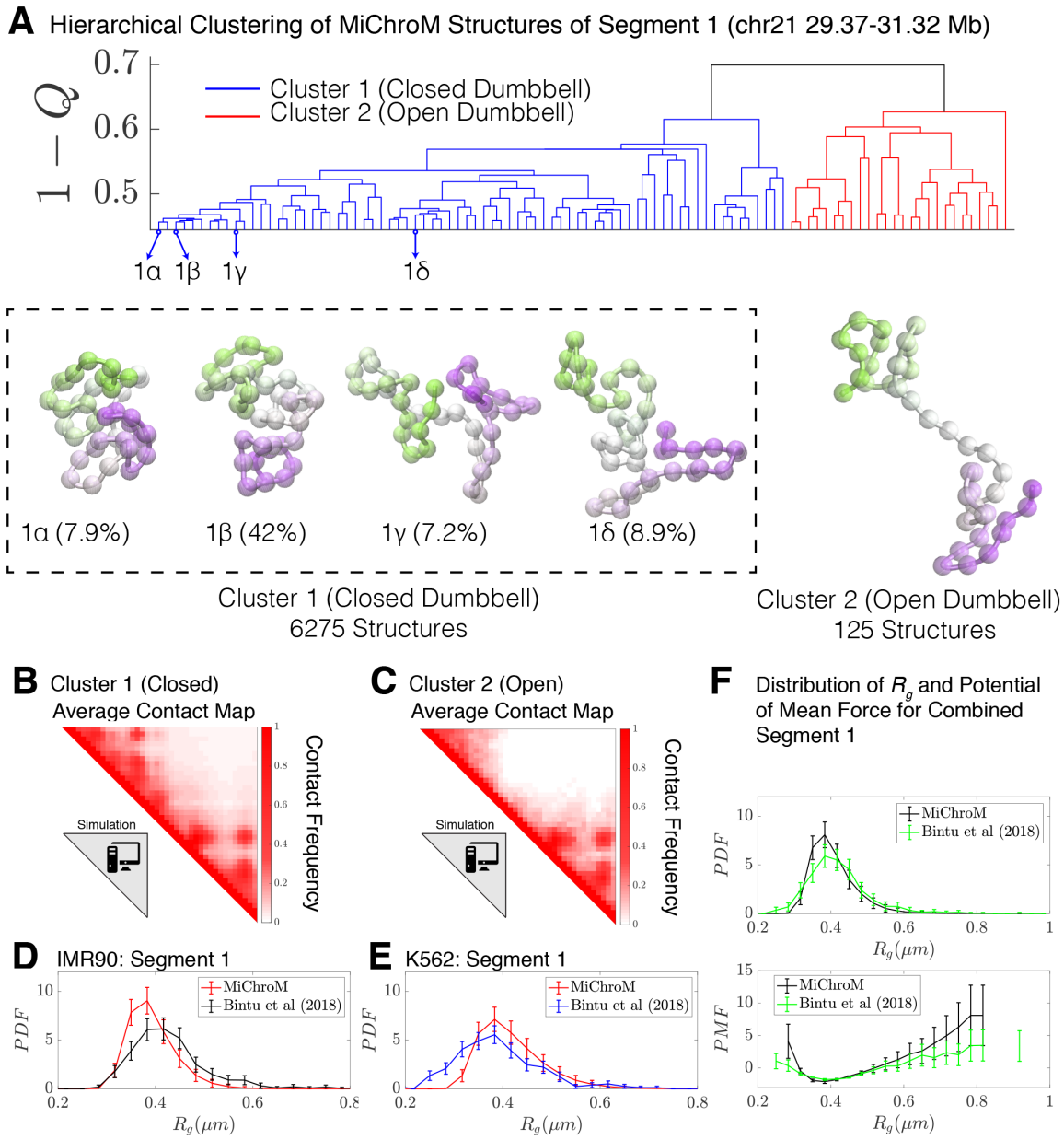

Figure 3. Hierarchical Clustering and the detailed structural analysis of simulated chromatin segment. (A) The dendrogram representation of the hierarchical clustering of simulated Segment 1 (chr21 29.37-31.32Mb for IMR90 and K562) where $1-Q$ is used as the distance between two structures. The clustering reveals two main clusters: closed dumbbell (6275 out of 6400 structures) and open dumbbell (125 out of 6400 structures).

The closed dumbbell can be subdivided into sub-clusters labeled 1α-1δ that represent the opening transition of the closed dumbbell. Representative structures are shown for each of the clusters and sub-clusters. The population averaged contact maps for the clusters are shown respectively in (B) and (C), where 165nm is used to define a contact between two 50kb loci of the MiChroM model. The distribution of the Radius of Gyration is shown for Segment 1 IMR90 (D) and K562 (E) traced structures in comparison with the experimental structures. (F) Distribution of the Radius of Gyration and the corresponding potential of mean force is shown for both experiment and simulation for all of the structures of Segment 1.

Comparative analysis of the chromosomal structural ensembles of different cell lines: connecting the epigenetic markings of loci with their radial positioning within territories

The frequency of chromatin type annotations predicted by MEGABASE over different cell types is shown in Figure 4A as a stacked bar chart that represents the distribution of chromatin type annotations predicted for each locus of chromosome 2 over all of the cell types. It is evident that certain loci have similar epigenetic markings patterns in all the cell types that we examined, either by being generally transcriptionally active loci, thus likely belonging to the A compartment, or by being transcriptionally inactive B compartment

loci. On the other hand, several segments of chromatin switch compartments between different cell types.

Using the structural ensemble from the simulations based on the predicted compartments we then quantified the conformational differences between different cell types. On the chromosomal scale, structural differences emerge primarily from changes in the phase separation of epigenetically marked segments of chromatin. An example is illustrated in Figure 4B, which focuses on the region 39.9-40.6Mb of chromosome 2 for HMEC, HUVEC, and IMR90. The MEGABASE classification (Figure 4B) identifies the segment in HMEC and IMR90 as belonging to the A compartment, whereas the segment for HUVEC should belong to the B compartment. Representative 3D structures for this segment for each of the respective cell types are shown in Figure 4C.

A plot of the radial density of A compartment loci and B compartment loci is shown in Figure 4D. These radial densities are consistent with previously reported simulations[27]. Taking a look at the radial distance of the center of mass of the segment of chromatin in each of the cell types, one finds that the A compartment loci tend to localize towards the surface of the chromosome, while the B compartment loci of the HUVEC cell type tend to localize in the interior (Figure 4E). A similar positioning of transcriptionally active chromatin toward the periphery of chromosomal territories was also observed by Nagano et al[13] in mouse cells using Hi-C experiments.

We additionally use simulations to predict and examine the spatial positioning of the segments of chromatin examined by Bintu et al in the context of the entire chromosome 21. The experimental traced structures could not be used to ascertain the spatial positioning of those chromatin segments within chromosome 21 since only short segments were imaged rather than the entire chromosome. Figure 4F shows a stacked bar chart that represents the distribution of chromatin types predicted by MEGABASE for each genomic position of chromosome 21. Figure 4G shows the MEGABASE predictions for the traced segments, showing that IMR90 Segment 2 (20.0-21.9Mb) is composed of A-type chromatin while IMR90 Segment 1 (29.37-31.32Mb) primarily is composed of B compartment chromatin types. K562 Segment 1 (29.37-31.32Mb) appears to contain both A and B chromatin types. Figure 4H shows the radial distance distribution of the center of mass of these segments of chromatin, showing that IMR90 Segment 2 tends to be in the interior, IMR90 Segment 1 tends to lie near the chromosome surface, and K562 Segment 1 occupies an intermediate region.

The finding that there exists a radial ordering associated with the spatial compartmentalization is consistent with the fact that according to the MiChroM potential[27], contact interactions between B loci exhibit the most favorable energetic stabilization of all chromatin interactions. On the other hand, A to B or A to A type interactions are significantly weaker than the B to B interaction, but are comparably strong to each other (See Table S1). In other words, according to the MiChroM energetic parameters (which were originally learned from Hi-C maps), B loci drive the phase separation of the chromosomes. Much like a hydrophobic-polar model from protein

folding, the B compartmentalization forms the stable core of the simulated chromosome and the weaker interactions between A compartment loci with A or B loci tends towards the surface, to minimize the free energy of the molecular assembly. Our theoretical model thus corroborates recent experiments[43, 44] that suggests heterochromatin phase separation is a major driving force behind genome organization, further highlighting the important role of phase separation in biological organization [27, 45-48].

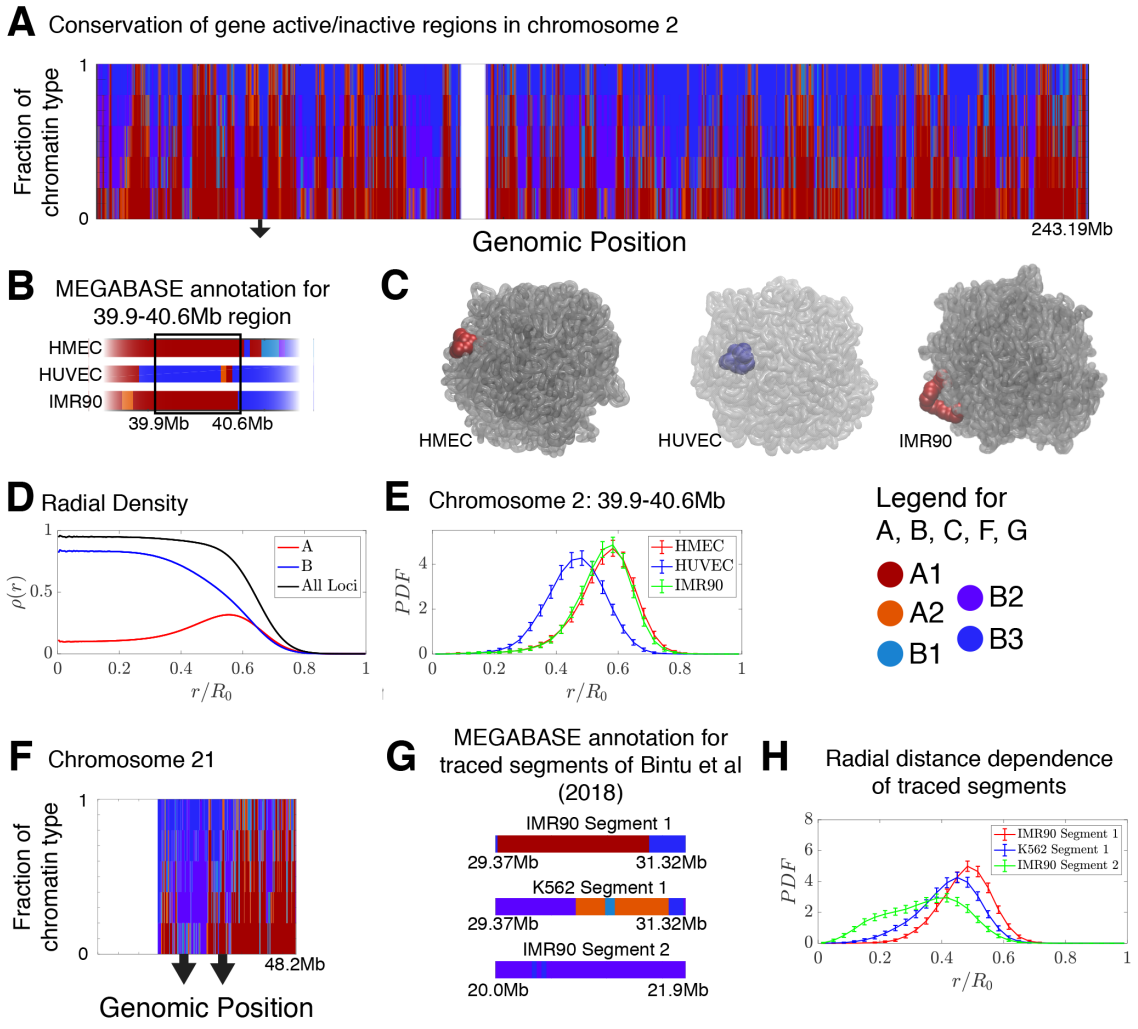

Figure 4. Conservation of compartmentalization across cell types and the radial dependence of marked chromatin. (A) A stacked bar chart is used to represent the distribution of chromatin type annotations predicted by MEGABASE

as a function of the genomic position along chromosome 2 (hg19). The colors correspond the chromatin types given in the Figure Legend. For a given genomic position, the relative height of a particular color indicates the fraction of that particular chromatin type predicted at that locus. (B) The MEGABASE prediction of the chromatin type is shown for the chromatin segment 39.9-40.6 Mb of chromosome 2 for HMEC, HUVEC, and IMR90. A black arrow in (A) highlights the location of this segment. (C) The chromatin segment 39.9-40.6 Mb of chromosome 2 is shown in a representative structure for each of the cell types, where the color of the segment denotes its MEGABASE annotation. For HMEC and IMR90, the segment of chromatin tends towards the chromosome surface, whereas the segment tends towards the interior for HUVEC. (D) The radial density as a function of the normalized radial distance is plotted for A compartment loci, B compartment loci, and all loci. (E) The probability density functions of the radial distance are shown for the center of mass of the segment 39.9-40.6 Mb of chromosome 2 for HMEC, HUVEC, and IMR90, respectively. (F) A stacked bar chart is used to represent the distribution of chromatin type annotations predicted by MEGABASE as a function of the genomic position along chromosome 21 (hg19). The arrows indicate the locations of the traced segments of Bintu et al[15]: Segment 1 (29.37-31.32Mb) and Segment 2 (20.0-21.9Mb). (G) The MEGABASE annotation of the traced chromatin segments are given for IMR90 and K562. (H) The distribution of radial distances of the center of mass of each traced segment is shown.

Discussion

DNA-tracing combined with diffraction-limited or super-resolution microscopy is beginning to shed light on the high degree of variability that is characteristic of chromosomal structures in the interphase[15-18]. These studies add to the growing body of evidence that a unique chromosomal fold simply does not exist in the interphase. Chromosome structures in the nucleus appear to be highly dynamical, owing to the many non-equilibrium processes in the cell, such as the activity of motor proteins.

The advances in genome imaging and the molecular simulation of chromosomes allows the development of order parameters able to quantify the structural similarities between different chromosome structures, and the degree of heterogeneity in the ensemble of structures. Our results demonstrate that the order parameter Q, commonly used in protein folding studies and structural biophysics, is a good order parameter for characterizing the structural ensemble of a segment of chromatin. Despite the high degree of conformational plasticity, it appears that for segments of chromatin as short as the ones imaged by Bintu et al [15] (~2Mb in length), there do exist distinct clusters of chromatin structures that can be distinguished using the Q order parameter. The dominant structures found for chromatin Segment 1 (chr21 29.37-31.32Mb) examined using data from microscopy as well as from simulation can be described as being a closed dumbbell and an open dumbbell, where the ends of the dumbbell are the globular domains at the head and tail of the chromatin segment.

It is known that CTCF proteins bound along the genome acts as gene insulators, probably through their suppressing activity toward loop extrusion[49-51]. Interestingly, a survey of the positioning of genes along Segment 1 reveals that the vast majority of genes appear clustered in the linker region of the chromatin segment (Figure 2E; Figure S7A), sandwiched between the head and tail loop domains. On the other hand, there is an absence of genes located on Segment 2 (chr21 20.0-21.9Mb) (Figure S7B), which contains no loop domains. Classification of the experimentally imaged structures of Segment 2 reveal structures that transition between open and closed conformations similar to Segment 1, but unlike Segment 1 lack the loop domains that segregate the head and the tail of the chromatin segment into globular domains. These findings suggest a possible role in transcriptional regulation for the opening and closing of organized dumbbell structures. How open and closed structures would achieve such regulation of the transcriptional activity remains to be investigated. It is however clear that understanding the structure-function relationship in the genome is a crucial question that can only be answered using an accurate statistical characterization of the conformational ensembles.

Finally, our work refines the classical view of the spatial compartmentalization of chromatin. We find a striking dependence between radial positioning of chromatin and epigenetic marking patterns. Our theoretical model, MiChroM, predicts that transcriptionally active loci, typically belonging to the A compartment, move towards the surface of the chromosomal territory, while B compartment loci, typically inactive, move to the interior[27]. Since interactions among B-B loci result in the greatest energetic stabilization, aggregation of these loci seems to be driving force behind both the phase

separation of epigenetically similar chromatin into compartments and the expulsion of the active chromatin toward the periphery of chromosomal territories. In other words, according to the present energy landscape model, when the epigenetic marking patterns of a locus are rewritten from A to B, the locus moves towards the interior of the chromosome, perhaps affecting the transcriptional activity of the associated genes.

Notes

Unless explicitly stated otherwise, all genomic positions are reported using the positions of the hg19 assembly. All of the simulated chromosome structures discussed in this manuscript were deposited in the Nucleome Data Bank (NDB)[32] found at https://ndb.rice.edu.

Acknowledgements

The authors would like to thank Ting Wu for helpful discussions. This work was supported by the Center for Theoretical Biological Physics sponsored by the National Science Foundation NSF Grant PHY-1427654. J.N.O. was also supported by the NSF-CHE-1614101 and by the Welch Foundation (Grant C-1792). J.N.O. is a CPRIT Scholar in Cancer Research sponsored by the Cancer Prevention and Research Institute of Texas. V.G.C. is a Robert A. Welch Postdoctoral Fellow and was also funded by FAPESP (São Paulo Research Foundation and Higher Education Personnel: Grant 2016/13998-8), and

CAPES (Higher Education Personnel Improvement Coordination: Grant 2017/09662-7). Additional support to P.G.W. was provided by the D. R. Bullard-Welch Chair at Rice University (Grant C-0016). E. L. A. was also supported by an NIH New Innovator Award (1DP2OD008540-01), the NHGRI Center for Excellence for Genomic Sciences (HG006193), the Welch Foundation (Q-1866), an NVIDIA Research Center Award, an IBM University Challenge Award, a Google Research Award, a Cancer Prevention Research Institute of Texas Scholar Award (R1304), a McNair Medical Institute Scholar Award, an NIH 4D Nucleome Grant (U01HL130010), an NIH Encyclopedia of DNA Elements Mapping Center Award (UM1HG009375), and the President's Early Career Award in Science and Engineering.

1. Rao, S.S.P., et al., *A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.* Cell, 2014. **159**(7): p. 1665-1680.
2. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-380.
3. Dixon, J.R., et al., *Chromatin architecture reorganization during stem cell differentiation.* Nature, 2015. **518**(7539): p. 331-336.
4. Rowley, M.J., et al., *Evolutionarily Conserved Principles Predict 3D Chromatin Organization.* Molecular Cell, 2017. **67**(5): p. 837-852.e7.
5. Dekker, J. and E. Heard, *Structural and functional diversity of Topologically Associating Domains.* FEBS Letters, 2015. **589**(20PartA): p. 2877-2884.
6. Yu, M. and B. Ren, *The Three-Dimensional Organization of Mammalian Genomes.* Annual Review of Cell and Developmental Biology, 2017. **33**(1): p. 265-289.
7. Tang, Z., et al., *CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription.* Cell, 2015. **163**(7): p. 1611-1627.
8. Lieberman-Aiden, E., et al., *Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.* Science, 2009. **326**(5950): p. 289-293.
9. Dudchenko, O., et al., *The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000.* bioRxiv, 2018: p. 254797.
10. Sexton, T., et al., *Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome.* Cell, 2012. **148**(3): p. 458-472.

11. Eagen, K.P., T.A. Hartl, and R.D. Kornberg, *Stable Chromosome Condensation Revealed by Chromosome Conformation Capture.* Cell, 2015. **163**(4): p. 934-946.
12. Zhang, Y., et al., *Spatial Organization of the Mouse Genome and Its Role in Recurrent Chromosomal Translocations.* Cell, 2012. **148**(5): p. 908-921.
13. Nagano, T., et al., *Single-cell Hi-C reveals cell-to-cell variability in chromosome structure.* Nature, 2013. **502**(7469): p. 59-64.
14. Stevens, T.J., et al., *3D structures of individual mammalian genomes studied by single-cell Hi-C.* Nature, 2017. **544**(7648): p. 59-64.
15. Bintu, B., et al., *Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells.* Science, 2018. **362**(6413): p. eaau1783.
16. Boettiger, A.N., et al., *Super-resolution imaging reveals distinct chromatin folding for different epigenetic states.* Nature, 2016. **529**(7586): p. 418-+.
17. Nir, G., et al., *Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling.* PLOS Genetics, 2018. **14**(12): p. e1007872.
18. Beliveau, B.J., et al., *Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH probes.* Nature Communications, 2015. **6**(1): p. 7147.
19. Finn, E.H., et al., *Extensive Heterogeneity and Intrinsic Variation in Spatial Genome Organization.* Cell, 2019. **176**(6): p. 1502-1515.e10.
20. Barbieri, M., et al., *Complexity of chromatin folding is captured by the strings and binders switch model.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(40): p. 16173-16178.
21. Jost, D., et al., *Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains.* Nucleic Acids Research, 2014. **42**(15): p. 9553-9561.
22. Gursoy, G., Y. Xu, and J. Liang, *Spatial organization of the budding yeast genome in the cell nucleus and identification of specific chromatin interactions from multi-chromosome constrained chromatin model.* Plos Computational Biology, 2017. **13**(7).
23. Brackley, C.A., et al., *Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domains.* Nucleic Acids Research, 2016. **44**(8): p. 3503-3512.
24. Tjong, H., et al., *Physical tethering and volume exclusion determine higher-order genome organization in budding yeast.* Genome Research, 2012. **22**(7): p. 1295-1305.
25. Nuebler, J., et al., *Chromatin organization by an interplay of loop extrusion and compartmental segregation.* Proceedings of the National Academy of Sciences, 2018. **115**(29): p. E6697.
26. Zhang, B. and P.G. Wolynes, *Topology, structures, and energy landscapes of human chromosomes.* Proceedings of the National Academy of Sciences of the United States of America, 2015. **112**(19): p. 6062-6067.
27. Di Pierro, M., et al., *Transferable model for chromosome architecture.* Proceedings of the National Academy of Sciences of the United States of America, 2016. **113**(43): p. 12168-12173.
28. Wong, H., et al., *A Predictive Computational Model of the Dynamic 3D Interphase Yeast Nucleus.* Current Biology, 2012. **22**(20): p. 1881-1890.

29. MacPherson, Q., B. Beltran, and A.J. Spakowitz, *Bottom–up modeling of chromatin segregation due to epigenetic modifications.* Proceedings of the National Academy of Sciences, 2018. **115**(50): p. 12739.
30. Di Pierro, M., et al., *De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture.* Proceedings of the National Academy of Sciences of the United States of America, 2017. **114**(46): p. 12126-12131.
31. Di Pierro, M., et al., *Anomalous diffusion, spatial coherence, and viscoelasticity from the energy landscape of human chromosomes.* Proceedings of the National Academy of Sciences, 2018. **115**(30): p. 7753.
32. Contessoto, V.G., et al., *The Nucleome Data Bank: Web-based Resources to Simulate and Analyze the Three-Dimensional Genome.* bioRxiv, 2019: p. 2019.12.20.885145.
33. Durand, N.C., et al., *Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom.* Cell Systems, 2016. **3**(1): p. 99-101.
34. Zhang, B. and P.G. Wolynes, *Shape Transitions and Chiral Symmetry Breaking in the Energy Landscape of the Mitotic Chromosome.* Phys Rev Lett, 2016. **116**(24): p. 248101.
35. Di Pierro, M., *Inner workings of gene folding.* Proceedings of the National Academy of Sciences, 2019. **116**(11): p. 4774.
36. Goundaroulis, D., E. Lieberman Aiden, and A. Stasiak, *Chromatin Is Frequently Unknotted at the Megabase Scale.* Biophysical Journal, 2019.
37. Bascom, G.D., C.G. Myers, and T. Schlick, *Mesoscale modeling reveals formation of an epigenetically driven HOXC gene hub.* Proceedings of the National Academy of Sciences, 2019. **116**(11): p. 4955.
38. Dekker, J., M.A. Marti-Renom, and L.A. Mirny, *Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data.* Nature Reviews Genetics, 2013. **14**(6): p. 390-403.
39. Kalhor, R., et al., *Genome architectures revealed by tethered chromosome conformation capture and population-based modeling.* Nature Biotechnology, 2012. **30**(1): p. 90-98.
40. Imai, R., et al., *Density imaging of heterochromatin in live cells using orientation-independent-DIC microscopy.* Molecular Biology of the Cell, 2017. **28**(23): p. 3349-3359.
41. Zhang, B. and P.G. Wolynes, *Genomic Energy Landscapes.* Biophysical Journal, 2017. **112**(3): p. 427-433.
42. Kent, W.J., et al., *The Human Genome Browser at UCSC.* Genome Research, 2002. **12**(6): p. 996-1006.
43. Falk, M., et al., *Heterochromatin drives compartmentalization of inverted and conventional nuclei.* Nature, 2019. **570**(7761): p. 395-399.
44. Strom, A.R., et al., *Phase separation drives heterochromatin domain formation.* Nature, 2017. **547**(7662): p. 241-245.
45. Hnisz, D., et al., *A Phase Separation Model for Transcriptional Control.* Cell, 2017. **169**(1): p. 13-23.
46. Brangwynne, Clifford P., P. Tompa, and Rohit V. Pappu, *Polymer physics of intracellular phase transitions.* Nature Physics, 2015. **11**(11): p. 899-904.

47. Banani, S.F., et al., *Compositional Control of Phase-Separated Cellular Bodies.* Cell, 2016. **166**(3): p. 651-663.
48. Shin, Y. and C.P. Brangwynne, *Liquid phase condensation in cell physiology and disease.* Science, 2017. **357**(6357): p. eaaf4382.
49. Sanborn, A.L., et al., *Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes.* Proceedings of the National Academy of Sciences of the United States of America, 2015. **112**(47): p. E6456-E6465.
50. Fudenberg, G., et al., *Formation of Chromosomal Domains by Loop Extrusion.* Cell Reports, 2016. **15**(9): p. 2038-2049.
51. Vian, L., et al., *The Energetics and Physiological Impact of Cohesin Extrusion.* Cell, 2018. **173**(5): p. 1165-1178.e20.