A method for extracting effective interactions from Hi-C data with applications to interphase chromosomes and inverted nuclei

Sucheol Shin,[1] Guang Shi,[1,2] and D. Thirumalai[1,3,*]

[1]*Department of Chemistry, The University of Texas at Austin, Austin, TX 78712, USA.*

[2]*Department of Chemistry, University of Illinois at Urbana-Champaign, Illinois, 61801, USA.*

[3]*Department of Physics, The University of Texas at Austin, Austin, TX 78712, USA.*

(Dated: March 29, 2023)

Abstract

Contact probabilities between loci that as a function of the genomic distance for a number of cell types have been reported using genome-wide conformation capture (Hi-C) experiments. How to directly use the experimental data, without an underlying model, to extract the effective interaction energies between active euchromatin (A) and inactive heterochromatin (B) is an open problem. Here we first calculate the pairwise effective interaction energies (A-A, A-B, or B-B) for interphase chromosomes based on Hi-C data by using the concept of Statistical Potential (SP). The assumption in the SP evaluation is that the interaction energy between two loci is proportional to the logarithm of the frequency with which they interact. Polymer simulations, using the extracted interaction energy values without any parameter, naturally results in the segregation between A and B type loci (compartments), and the emergence of topologically associating domains (TADs), features that are prominent in the Hi-C data for interphase chromosomes. Remarkably, the values of the SP automatically satisfy the Flory-Huggins phase separation criterion for all the chromosomes, which explains the compartment formation in interphase chromosomes. Strikingly, simulations using the SP that accounts for pericentromeric constitutive heterochromatin (C), show hierarchical structuring with the density of C in the nuclear center, followed by localization of the B type loci, with euchromatin being confined to the nuclear periphery, which is in accord with imaging data. Such an unusual organization of chromosomes is found in inverted nuclei of photoreceptor rods in nocturnal mammals. The proposed parameter free method and applications show that compartment formation in conventional and inverted nuclei is best explained by the inequality between the effective interaction energies, with heterochromatin attraction being the dominant driving force.

* Corresponding author: dave.thirumalai@gmail.com

INTRODUCTION

Knowledge-based potentials, often referred to as statistical potentials (SPs), have been used to extract effective pairwise interactions between amino acid residues from the data base of non-redundant folded structures. The essence of the idea was first introduced by Tanaka and Scheraga [1], and subsequently developed by Miyazawa and Jernigan [2, 3] and others [4–8]. The frequency of contact between specific amino acid (aa) residues is used to estimate the free energy of the interaction. The set of free energies, which is proportional to the aa contact frequencies in the set of the PDB structures, constitute the approximate strengths of tertiary interactions between aa side chains or between the backbone and aa side chains. The resulting SPs have been successful, especially when combined with coarse-grained simulations, to predict protein folding thermodynamics and kinetics [9], peptide binding to MHC complexes [10], protein-ligand binding [11], and protein-protein interactions [12]. Most recently, an AI-based approach using the SPs demonstrated the prediction of native protein structures with remarkably high accuracy [13].

The concept of SP has also been used to extract stacking interactions between nucleotides in RNA [14], which is important in determining the stability of RNA folds [15], by exploiting the PDB structures that were available in 2005. The calculated values of the stacking interactions are in excellent agreement with experimental measurements, which were obtained by studying the melting profiles of oligonucleotides [16, 17]. By using gapless threading and the SPs for RNA, we identified in excess of 70% of native base pairs in the secondary structure for RNA molecules with less than 700 nucleotides. Our study on RNA and the ones on proteins established that knowledge-based methods are successful in extracting the values of the interaction parameters, which could then be profitably used in simulations for a variety of purposes.

Here, we explore if the large number of genomic contact maps (CMs), available from the chromosome conformation capture experiments (referred to as Hi-C [18] from now on), for a number of species and under a variety of growth conditions, could be used as a guide for calculating the effective free energies of interactions between the distinct loci in interphase chromosomes. For this purpose, we classify the loci as euchromatin (A-type locus) and heterochromatin (B-type locus). The free energy scales for three different locus pair interactions (A-A, B-B, and A-B interactions) and their distributions are calculated from

the CMs using a generalization of the formulation used for proteins and RNA. By using the mean values of the free energies as the interaction parameters in polymer simulations based on Chromosome Copolymer Model (CCM) [19], we show that the relevant organization features found in the Hi-C CMs are accurately captured. The mean free energy values result in the effective interaction parameter, $\chi_{\rm FH}$, from the Flory-Huggins theory [20, 21] to be greater than zero, which naturally explains the microphase separation between euchromatin and heterochromatin that is routinely observed across virtually all eukaryotic chromosomes [22, 23]. The ensemble of three-dimensional structures predicted by the SP-based CCM (SP-CCM) polymer simulations is in very good agreement with that observed in the imaging experiment using DNA fluorescence *in situ* hybridization (FISH) [24]. The SP-CCM simulations also resolve TAD structures in the CM, when the CTCF-mediated loop anchors were included in the model. Strikingly, the SPs extracted from the interchromosome Hi-C data for the inverted nuclei [25] when used in the SP-CCM simulations result in the atypical spatial pattern of nuclear compartmentalization in which the euchromatin are localized in the nuclear periphery whereas heterochromatin structures are in the interior. Our method, which provides the effective pairwise interaction energies calculated directly from the Hi-C data *without any parameters*, may be used in polymer simulations to predict the structural and dynamical properties of chromosomes over a broad range of length scales.

RESULTS

Effective interaction energies between loci: We first calculated the values of the SP between individual locus pairs, ΔG_{ij} [Eq. (2)], for chromosome 2 (Chr2) from the IMR90 cell line using the Hi-C CM at 100-kb resolution [26]. The distributions of the SPs, $P(\Delta G_{ij}/k_BT)$, where k_B is the Boltzmann constant and T is the temperature, are plotted for each locus pair type in Fig. 1**a**. The locus types, A (active or euchromatin) or B (inactive or heterochromatin), are inferred from the experimental CM using the principal component analysis (see Methods for details). Although the distribution of effective interactions is broad, Fig. 1**a** show that the mean A-A interaction is modestly more favorable than between B-B, which is different from the prevailing view in the literature. The interaction between A-B is less favorable than between A-A or B-B.

The mean values, $\Delta G_{\rm AA}$, $\Delta G_{\rm BB}$, and $\Delta G_{\rm AB}$, set the effective energy scales in the CCM

simulations (see Methods and Eq. (8) for details). We used $\epsilon_{\alpha\beta} = -\Delta G_{\alpha\beta}$, where $\alpha, \beta =$ A or B (Fig. 1**a**). The calculation yields $\epsilon_{\rm AA} = 1.62 k_B T$, $\epsilon_{\rm BB} = 1.41 k_B T$, and $\epsilon_{\rm AB} = 0.95 k_B T$. Note that $\chi_{\rm FH} = [(\epsilon_{\rm AA} + \epsilon_{\rm BB})/2 - \epsilon_{\rm AB}]/k_B T > 0$, which implies the use of these effective energy scales in the polymer simulations should result in microphase separation between A and B type loci. It is worth emphasizing that the calculated interaction energies emerge naturally from Eq. (2) with the CM being the only input. In other words, these values were not adjusted to fit with any experimental values unlike in a majority of previous studies [19, 25, 27–32]. The validity of the calculated SP values can only be assessed by polymer simulations.

Compartment formation from SP-CCM simulations: Next we assessed the accuracy of the SP-based energetic parameters by performing polymer simulations, as described in Methods. The SP-CCM simulations accounts for phase separation between A and B loci, which accords well with experiment (Fig. 1**b**). For a quantitative comparison, we computed the Pearson correlation matrix from the CM [Eq. (10)]. The correlation matrix, $\rho(i, j)$, highlights the checker-board pattern of compartments vividly. In Fig. 1**c**, the Pearson correlation matrices for the Hi-C and simulated CMs show good agreement. Visually the segregation between A and B loci appears to be stronger in the calculated CM using the SP-CCM simulations compared to Hi-C data. However, the distributions of the Pearson correlation coefficients demonstrate that the compartmentalization predicted by the SP-CCM is in quantitative agreement with the Hi-C result (Fig. 1**d**). The Jensen-Shannon divergence (JSD) [Eq. (23)] between the distributions for the p-arm (q-arm) is 0.086 (0.13), which shows that the simulated and experimentally inferred distributions are in excellent agreement.

We then compared the Chr2 structures generated by the SP-CCM simulations to those observed in the super-resolution imaging experiment based on DNA FISH [24]. The simulated and experimental distance maps (DMs), showing the mean pairwise distance, $\langle r_{ij} \rangle$, where the angular brackets, $\langle \cdots \rangle$, denote an ensemble average over multiple cells (in experiments) or trajectories (in simulations) are in good agreement (Fig. 1**e** and Figs. S1a-S1c). Although both the DMs feature the checker board patterns corresponding to A/B compartments, the simulated DM shows modestly larger difference between A-A (or B-B) and A-B interaction blocks than observed in the experiment. To characterize the differences quantitatively, we compare the simulated and experimental probability distributions of the pairwise distance r_{ij} for two locus pairs, $(i, j) =$ (24 Mb, 32.1 Mb) and (32.4 Mb, 42.6 Mb), the A-A and

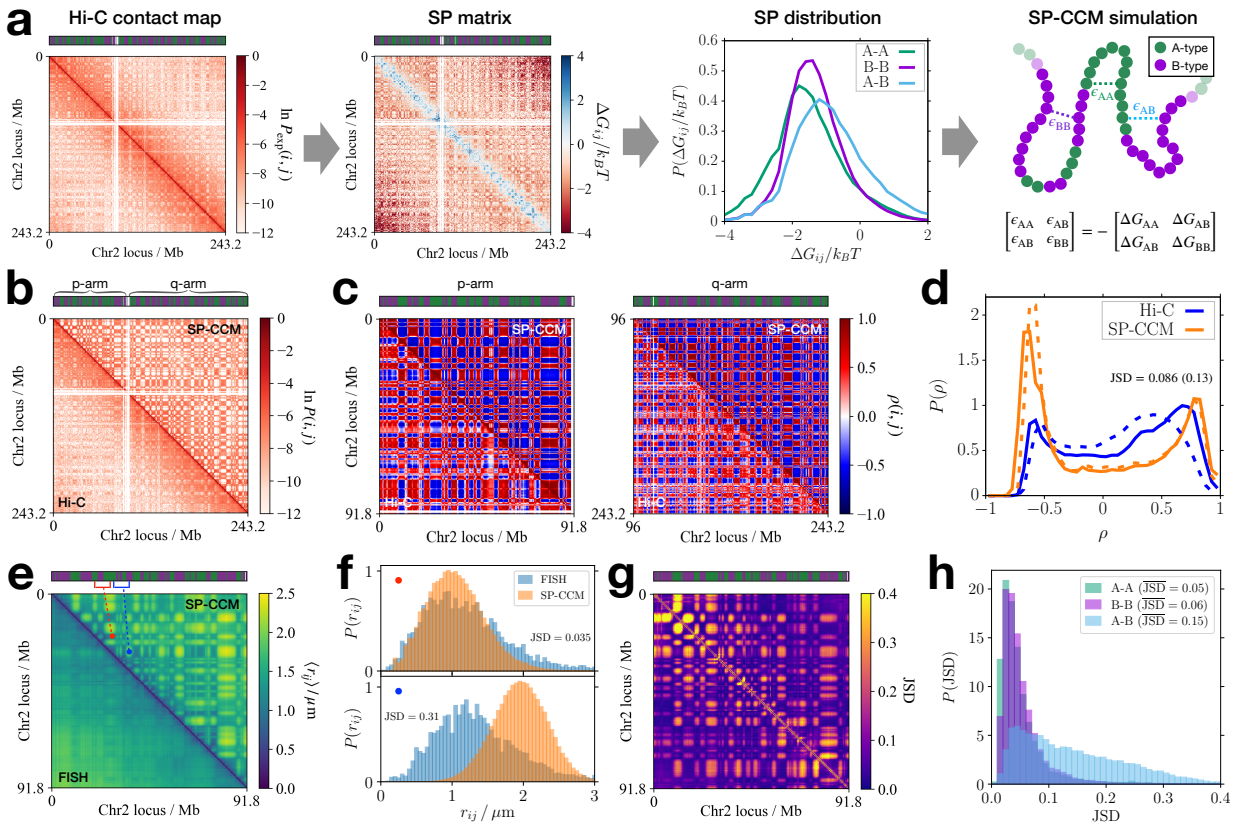

FIG. 1. **SP-CCM simulations for Chr2 predict the chromosome structures with A/B compartmentalization.** (**a**) The Hi-C contact map is converted to the SP matrix [Eq. (2)], where the color bar above each heat map indicates the A/B type (green/purple) of the individual loci. The calculated SP values are sorted into the distributions based on the locus type, whose mean values set the relative strength of A-A, B-B, and A-B interactions that are used in the polymer simulations. (**b**) Comparison between the contact matrices obtained from the Hi-C experiment (lower triangle) and the SP-CCM simulation (upper triangle), where the red-color shading shows the logarithm of the contact frequency, given in the color bar on the right. (**c**) Pearson correlation matrices, corresponding to the contact matrices in panel **b**, computed separately for the p- (left) and q-arms (right). (**d**) Probability distributions of the Pearson correlation coefficients in panel **c**, computed from the Hi-C and SP-CCM, shown in blue and orange solid (dashed) lines for the p-arm (q-arm), respectively. (**e**) Comparison between the mean pairwise distance matrices, $\langle r_{ij} \rangle$, obtained from the imaging experiment (lower triangle) and the SP-CCM simulations (upper triangle). The A-A and A-B pairs, (i, j) = (24 Mb, 32.1 Mb) and (32.4 Mb, 42.6 Mb), are highlighted in red and blue colors, respectively. (**f**) Probability distributions of the distance between the locus pairs specified by the red (top) and blue (bottom) circles in panel **e**. (**g**) Heatmap of the JSD matrix, where each element indicates the value of the JSD between the probability distributions, $P(r_{ij})$, for a given locus pair as shown in panel **f**. (**h**) Histograms of JSD shown in panel **g** for a given type of locus pair.

A-B pairs, which have the similar genomic distance ($|i-j| \sim$ 8-10 Mb; see the red and blue markers in Fig. 1**e**). For the (24 Mb, 32.1 Mb) pair, the distance distributions, $P(r_{ij})$, obtained from the FISH and the SP-CCM are in excellent agreement with JSD = 0.035 (Fig. 1**f**, top). On the other hand, for the (32.4 Mb, 42.6 Mb) pair, the agreement is not as good (JSD = 0.31; Fig. 1**f**, bottom). The difference between the distance distributions for the (32.4 Mb, 42.6 Mb) pair arises because the structural ensemble from the FISH experiment is not identical to the one implied by the Hi-C data (see Fig. S3). The simulations are based on the SPs extracted from the Hi-C data, which quantitatively differ from the SPs based on the FISH data (see Discussion).

To assess the overall similarity between the structural ensembles determined from experiments and the simulations, we quantified the JSD values between the experimental and the simulated distributions, $P(r_{ij})$, for all the pairs (Fig. 1**g**). The distribution of JSD for a given pair type shows that the spatial arrangements for A-A or B-B locus pairs predicted by the SP-CCM are quantitatively close to those extracted from the imaging experiment (Fig. 1**h**). There is a small difference between the A-B pair distance distributions, with $\overline{\text{JSD}} = 0.15$ where the bar denotes a sample mean (*e.g.*, average over the pairs). The results suggest that the simulated structures exhibit modestly stronger A/B segregation while the spatial arrangements between the same type of loci are accurate. Taken together, we surmise that the SP-CCM simulations predict the structural ensemble that is similar to the FISH data, which we find to be most interesting because our theory and simulations contain no free parameters. We also obtained similar results for the q-arm of Chr2 (Figs. S1d-S1e).

The results for the chromosome 21 (Chr21) show similar trends as well. We computed ΔG_{ij} from the Hi-C CM at 50-kb resolution for IMR90 Chr21 (14Mb-46.7Mb). The distribution of ΔG_{ij} for each pair type is plotted in Fig. S4a. From the mean values of the SPs, we obtain $\epsilon_{\text{AA}} = 0.46 k_B T$, $\epsilon_{\text{BB}} = 1.04 k_B T$, and $\epsilon_{\text{AB}} = 0.49 k_B T$ ($\chi_{\text{FH}} > 0$) for use in the CCM simulations. Note that $\epsilon_{\text{BB}} > \epsilon_{\text{AA}} \approx \epsilon_{\text{AB}}$ for this chromosome unlike the values of interaction parameters for Chr2. With this choice of the energetic parameters, the SP-CCM simulations reproduce the compartments in the CM (Figs. S4b-S4c). The simulated mean distance map also captures the pattern of compartments shown in the distance map from the FISH experiment [24] (Fig. S4d). The matrix and distribution of $\text{JSD}(P_{\text{exp}}(r_{ij})||P_{\text{sim}}(r_{ij}))$ demonstrate that the structural ensemble from the SP-CCM simulations is near quantitative agreement with the FISH data (Figs. S4e-S4f).

TAD structures from SP-CCM simulations: TADs are the average structures that appear largely, but not always, in conjunction with the formation of CTCF-cohesin loops [26, 33, 34] on the scale starting from $\sim$ 500 kbps. To account for TAD formation in the simulations, we not only increased the resolution of the SP-CCM from 100kb to 50kb but also included the structural elements for loops in a CCM polymer chain. The loops are represented by bonding interactions between specific pairs of loci that are identified as the loop anchors with CTCF motifs from the Hi-C experiment [26]. In Chr2 from the IMR90 cell line, there are 126 loops uniquely detected with CTCF motifs (see Method for details). The calculated value of the energetic parameters for Chr2, at 50kb resolution, are $\epsilon_{\rm AA} = 2.30 k_B T$, $\epsilon_{\rm BB} = 2.10 k_B T$, and $\epsilon_{\rm AB} = 1.74 k_B T$, which also leads to $\chi_{\rm FH} > 0$. Note that the interaction scales are smaller at the lower (100 relative to 50kb) resolution, which follows from general arguments given in the SI.

Figure 2 shows the results for TAD formation in the SP-CCM simulations for the p-arm of Chr2. In a Hi-C contact map, TADs appear as enrichment of contacts along the diagonal [33, 34]. This feature is also prominent in the CM calculated using the SP-CCM simulations (Fig. 2**a**). To compare the TADs between the simulation and Hi-C results, we determined the TAD boundaries using the insulation score (IS), defined in Eq. (11), which quantifies the average number of interactions between the downstream and upstream regions from a given locus. A small value of IS indicates enhanced insulation around a given locus, so the associated minima would indicate the TAD boundaries [35]. The profile of IS and the corresponding TAD boundaries for the SP-CCM simulations agree well with the Hi-C results (Fig. S5). Although the simulations predict less number of TADs than expected from Hi-C (78 versus 137), $\sim$ 94% of the predicted boundaries are within 200 kb from those expected in the experimental CM. The probability distribution of the length of the TAD, $L_{\rm TAD}$, defined as the genomic distance between the inferred boundaries, also gives quantitative agreement between the simulations and the experiment, as indicated by the small JSD value ($\sim$ 0.08; Fig. 2**b**).

We also compare the simulation results with the FISH experimental data for the three-dimensional characteristics of the TADs. In Fig. 2**c**, the mean DM exhibits the regions with enhanced pair proximity, which are qualitatively similar to the contact enrichment shown in the corresponding CM (righthand, Fig. 2**a**). While the resolution of the experimental data (250-kb resolution) is too low to perform an analysis on individual TADs, we considered the

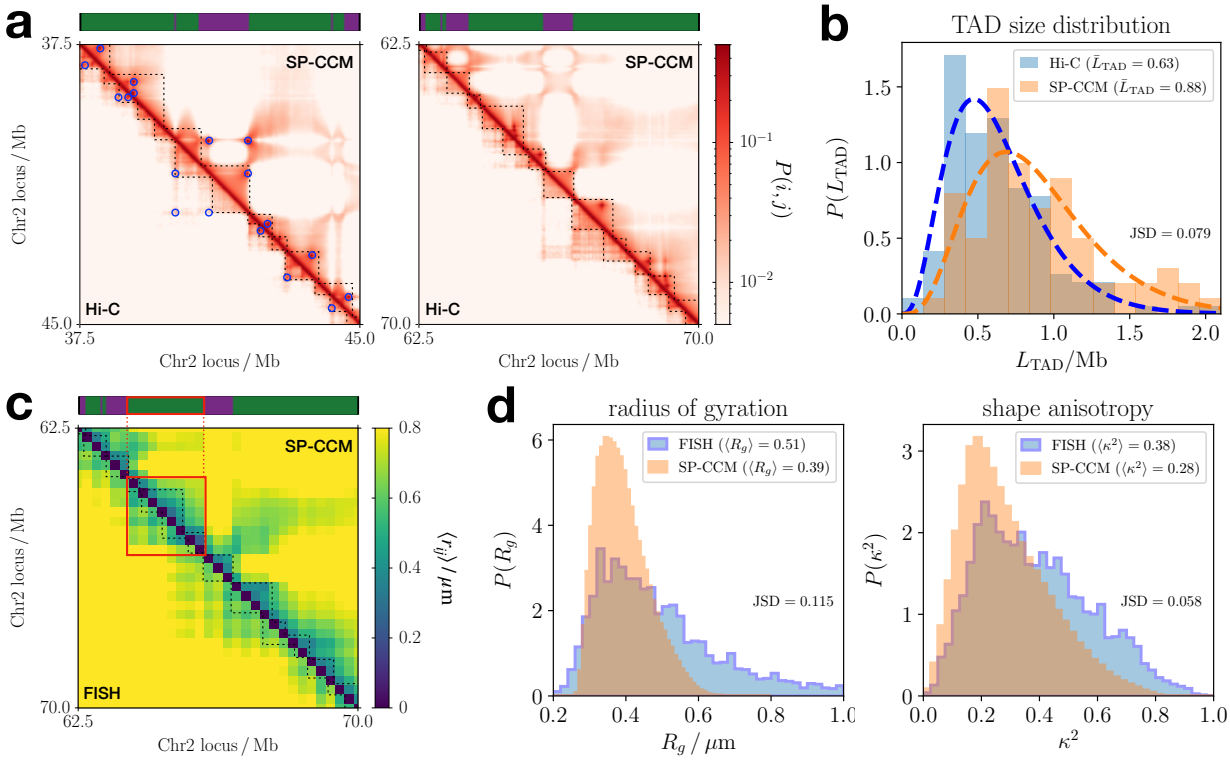

FIG. 2. **Resolving TADs using SP-CCM simulations.** (**a**) Comparison between the CMs from the Hi-C experiment and the SP-CCM simulations for a region of Chr2 spanning 37.5Mb-45Mb and 62.5Mb-70Mb on the left and right panels, respectively. The TAD boundaries inferred from the insulation score of each CM are delineated by the dashed lines. On the left panel, the CTCF loop anchors are marked by blue circles. (**b**) Comparison of the probability distribution for the genomic length of TADs in the p-arm of Chr2, between the Hi-C (blue) and the SP-CCM (orange). The dashed lines are the Gamma-distribution fits, $f(x;a,b) = (x/b)^{a-1}e^{-x/b}/\int_0^\infty t^{a-1}e^{-t}dt$, with $(a,b) = (3.99, 1.58)$ and $(4.57, 1.93)$ for the Hi-C and the SP-CCM, respectively. The mean TAD length, $\bar{L}_{\rm TAD}$, in Mb unit, is shown at the top. (**c**) Comparison between the mean distance matrices from the FISH experiment and simulation for Chr2: 62.5Mb-70Mb, overlaid with the TAD boundaries shown in panel **a**. The red box illustrates a 2Mb-region that is enriched in A-type, containing 3 to 4 TADs. (**d**) Probability distributions of the radius of gyration (left) and relative shape anisotropy (right) of the red-boxed domain in panel **c**, compared between the FISH (blue) and SP-CCM (orange) results.

structural features in a 2-Mb region (Chr2: 63.85Mb-65.85Mb), which includes 3-4 TADs of A-type loci (red box, Fig. 2**c**). We computed the distributions of the radius of the gyration, R_g, and the relative shape anisotropy, κ^2 [Eqs. (13)-(14)], which are displayed in Fig. 2**d**. Despite the minor difference in the widths of the R_g distributions between the experiment

and the SP-CCM polymer simulations (standard deviation = 0.23μm vs. 0.07μm), the median values are similar (0.45μm vs. 0.38μm), and the JSD value is small (≈ 0.12). The calculated and measured distributions of κ^2 are in excellent agreement with each other, with a JSD value ≈ 0.06. The non-zero value of κ^2 implies that these structures are anisotropic even though they adopt globular structures on the whole. These comparisons show that the TAD structures predicted by the SP-CCM simulations are in quantitative agreement with both Hi-C and FISH experiments. It is worth emphasizing that the extent of agreement is obtained *without any adjustable parameter* in the polymer simulations, which only used the experimental CM as input in the theory.

Comparison of single-chromosome SP values: From the Hi-C data at 50-kb resolution [26], we calculated the values of the single-chromosome SPs for all other chromosomes from the IMR90 cell line (Table S2). The SP values, $\epsilon_{\alpha\beta}^{(n)}$, depend on n, the chromosome number (Fig. 3**a**), to a minor extent. The differences between $\epsilon_{\rm AA}$ and $\epsilon_{\rm BB}$ in all chromosomes are modest. In about 9 chromosomes $\epsilon_{\rm BB} > \epsilon_{\rm AA}$ whereas in other the reverse holds. In addition to Chr2 and Chr21, we also performed polymer simulations for a few other chromosomes using the calculated single-chromosome SPs (Table S2, Fig. 3**a**). The results in Fig. S6 ensure that the CMs for $n = 5$, 15, and X are accurately predicted from the SP-CCM simulations of a single chromosome.

Using $\epsilon_{\alpha\beta}^{(n)}$ for all n, we calculated the effective Flory-Huggins interaction parameter, $\chi_{\rm FH}^{(n)}$, defined as,

$$\chi_{\rm FH}^{(n)} = \frac{(\epsilon_{\rm AA}^{(n)} + \epsilon_{\rm BB}^{(n)})/2 - \epsilon_{\rm AB}^{(n)}}{k_B T} \ , \tag{1}$$

which predicts the propensity of A and B type loci to phase-separate. Fig. 3**b** shows that the values of $\chi_{\rm FH}^{(n)}$ are narrowly distributed around a mean value, 0.45. The X chromosome is a notable exception with $\chi_{\rm FH}^{(\rm X)} \approx 0.06$, which is mainly due to the loss of A/B compartments in the inactive X chromosome [26, 36, 37]. As pointed out above, $\chi_{\rm FH}^{(n)} > 0$ ensures that A and B loci would microphase separate, subject only to the constraint of chain connectivity. Figure 3**c** shows $\epsilon_{\rm max}^{(n)} = \max[\epsilon_{\rm AA}^{(n)}, \epsilon_{\rm BB}^{(n)}, \epsilon_{\rm AB}^{(n)}]$ as a function of the chromosome length, $L_{\rm chr}^{(n)}$, which is the difference between the start and end positions of the Hi-C reads for a given chromosome. There is a strong correlation between the overall energy scale in the SP for each individual chromosome and the logarithm of the length (Pearson correlation coefficient = 0.96). The length dependence of the SP suggests that there is a larger free energetic cost

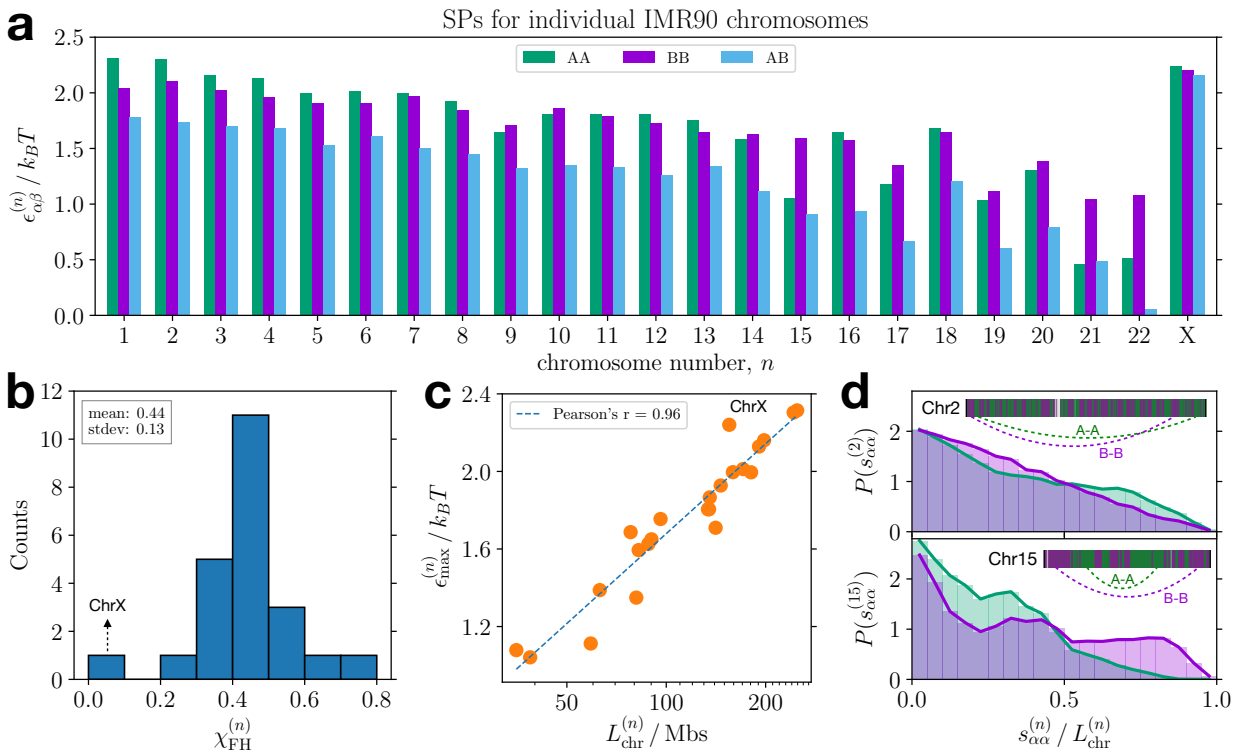

FIG. 3. **SP values reflect intrinsic energies of a chromosome.** (**a**) Bar graph showing the SP energetic parameters calculated using 50-kb resolution for individual IMR90 chromosomes. (**b**) Histogram of the effective Flory-Huggins χ parameter computed from the extracted energetic parameters, where the arrow indicates the data point for ChrX. (**c**) Scatter plot showing the correlation between the maximum value of the energetic parameter for a given chromosome as a function of chromosome length along with a linear fit given by the dashed line. The x-axis is shown in a log scale. (**d**) Probability distributions of the genomic distance for A-A (green) or B-B (purple) pairs in chromosomes 2 and 15, plotted in the top and bottom panels respectively, where A/B sequence information of each chromosome is shown along with the most probable but farthest interactions for A-A and B-B pairs.

for a longer chromosome to collapse into a compact configuration. If we consider the contact between the start and end loci for collapse, then $\Delta G_{1,N} = -k_BT \ln[P(1,N)/P_{\rm ref}(1,N)] = k_BT \ln N^{-3/2} + \text{constant}$, so $|\Delta G^{(n)}_{1,N}| \sim \ln L^{(n)}_{\rm chr}$. The proportionality between the SP and the logarithm of the contour length can be also demonstrated using homopolymer simulations with different number of monomers (Fig. S7a). The X chromosome is again an outlier from the trend, as it has large energetic parameters compared to its length. The relatively large value of $\epsilon^{(\rm X)}_{\alpha\beta}$ should be attributed to the formation of "super-loops" and "super-domains," which enhance the long-range contacts over 10 to 100 Mbs in the inactive X chromosome

[26, 37, 38].

The difference between $\epsilon_{\rm AA}^{(n)}$ and $\epsilon_{\rm BB}^{(n)}$, which depends on the chromosome number, can be understood by using the similar argument to $\epsilon_{\rm max}^{(n)} \propto \ln L_{\rm chr}^{(n)}$ for the A-A and B-B locus pairs along a given chromosome. That is, the overall scales of $\epsilon_{\rm AA}^{(n)}$ and $\epsilon_{\rm BB}^{(n)}$ is determined by the contact free energies for the A-A and B-B pairs separated by large genomic distances. In Fig. 3**d**, the distributions of the genomic distances for A-A and B-B pairs, $s_{\rm AA}$ and $s_{\rm BB}$, are compared between Chr2 and Chr15. For Chr2, the distributions of $s_{\rm AA}^{(2)}$ and $s_{\rm BB}^{(2)}$ are close to each other, which implies A and B loci are similarly positioned along the chromosome (Fig. 3**d**, top). Accordingly, $\epsilon_{\rm AA}^{(2)}$ and $\epsilon_{\rm BB}^{(2)}$ have similar effective energies, while the presence of A-A pairs at large distances leads to $\epsilon_{\rm AA}^{(2)} \gtrsim \epsilon_{\rm BB}^{(2)}$. On the other hand, these distributions for Chr15 are significantly different (Fig. 3**d**, bottom). The B loci are located predominantly near the ends of the chromosome whereas A loci are inside, which yields a larger value of $\epsilon_{\rm BB}^{(15)}$ relative to $\epsilon_{\rm AA}^{(15)}$. Hence, $\epsilon_{\rm AA}^{(n)}/\epsilon_{\rm BB}^{(n)}$ shows a better correlation with the mean genomic distance ratio, $\bar{s}_{\rm AA}^{(n)}/\bar{s}_{\rm BB}^{(n)}$, than with the locus number fraction, $N_{\rm A}^{(n)}/N_{\rm B}^{(n)}$ (Fig. S8). We also demonstrated the relationship between $\epsilon_{\rm AA}/\epsilon_{\rm BB}$ and $\bar{s}_{\rm AA}/\bar{s}_{\rm BB}$, where the SP values were inferred from a homopolymer contact map with different A/B sequences assumed (Figs. S7b-S7d).

For additional comparison, we also calculated the single-chromosome SPs using the Hi-C CMs from GM12878 cell line [26]. The mean SP values for individual chromosomes, $\epsilon_{\alpha\beta}^{(n)}$, have a range of magnitudes similar to those for IMR90 chromosomes (Fig. S9a). Unlike the SPs for IMR90, $\epsilon_{\rm AA}^{(n)}$ is larger than $\epsilon_{\rm BB}^{(n)}$ ($\bar{\epsilon}_{\rm AA}/\bar{\epsilon}_{\rm BB} \approx 1.2$) (Fig. S9b), whereas $\chi_{\rm FH}^{(n)}$ shows a similar trend ($\bar{\chi}_{\rm FH} = 0.46 \pm 0.16$) (Fig. S9c). We confirm that SP values for GM12878 are also related to the chromosome length and A/B sequence in the same way as for IMR90 (Figs. S9d-S9e). As such, the SP can differentiate between the interaction scales for individual chromosomes from distinct cell lines.

SP-CCM simulations for inverted nuclei: In typical interphase nuclei, gene-poor inactive heterochromatin (B-type) is localized on the nuclear periphery whereas gene-rich active euchromatin (A-type) tends to be located in the interior of a nucleus [23]. The conventional picture for the spatial arrangement of euchromatin and heterochromatin is inverted in the mouse rod photoreceptor cells [39]—heterochromatin loci are found at the center and euchromatin loci are localized on the periphery (Fig. 4**a**). The SP concept used to extract interaction energies is general that it is applicable to any chromosome for which the Hi-C

data is available. We investigate if the SPs extracted for the inverted nuclei yield results that characterize the experiments well [25]. In particular, we calculated the *inter-chromosome* SPs for use in polymer simulations of multiple chromosomes in a nucleus.

Falk and coworkers surmised that the interactions involving pericentromeric heterochromatin (referred to as C-type; see Fig. 4**b** and Methods) are predominant in driving the spatial pattern in the inverted nuclei. Using a search of all the parameter permutations (720 sets) in the interaction parameter space, ϵ_{AA}, ϵ_{BB}, ϵ_{CC}, ϵ_{AB}, ϵ_{AC}, and ϵ_{BC} [25], they obtained the set of 6 parameters that produces the best agreement between their simulations and experimental measurements [25]. Our calculations of the SPs for intra-chromosome interactions from the Hi-C data for the inverted nuclei, show that $\epsilon_{\mathrm{AA}}^{(n)} > \epsilon_{\mathrm{BB}}^{(n)} > \epsilon_{\mathrm{CC}}^{(n)}$ for each individual chromosome (Figs. S10a-S10b). On the other hand, when we calculated $\epsilon_{\alpha\beta}^{(n,m)}$ from the inter-chromosome CMs, which is the mean value of the SP for the loci of types α and β belonging to distinct chromosomes n and m, respectively [Eqs. (16)-(18)], we found that $\epsilon_{\mathrm{CC}}^{(n,m)} > \epsilon_{\mathrm{BB}}^{(n,m)} > \epsilon_{\mathrm{AA}}^{(n,m)}$ for most pairs of distinct chromosomes (Fig. S10c). The trend in $\bar{\epsilon}_{\alpha\beta}$ [Eq. (20)], the SP values averaged over all distinct chromosome pairs (Fig. 4**c**), is similar to that in the optimized energy parameter set from Falk *et al.* (*cf.* Fig. 2d in Ref. 25). The values of $\bar{\epsilon}_{\alpha\beta}$ are roughly equal to the ones estimated by Falk *et al.* with an offset (see Table S3), which is surprising because we used an entirely different method to extract the effective energies. The comparison shows that it is the relative effective interaction values that determine the organization of chromosomes. Note that the difference between the energy scales for distinct pair types is small as $\sim 0.05 k_B T$. Using $\bar{\epsilon}_{\alpha\beta}$, we obtain $\chi_{\mathrm{FH}}^{(\mathrm{AB})} = [(\bar{\epsilon}_{\mathrm{AA}} + \bar{\epsilon}_{\mathrm{BB}})/2 - \bar{\epsilon}_{\mathrm{AB}}]/k_B T = 0.12$ and $\chi_{\mathrm{FH}}^{(\mathrm{AC})} = [(\bar{\epsilon}_{\mathrm{AA}} + \bar{\epsilon}_{\mathrm{CC}})/2 - \bar{\epsilon}_{\mathrm{AC}}]/k_B T = 0.12$, which predicts the microphase separation between euchromatin (A) and heterochromatin (B and C). On the other hand, $\chi_{\mathrm{FH}}^{(\mathrm{BC})} = [(\bar{\epsilon}_{\mathrm{BB}} + \bar{\epsilon}_{\mathrm{CC}})/2 - \bar{\epsilon}_{\mathrm{BC}}]/k_B T = 0.006$, which is close to a typical value of Flory-Huggins χ for polymer blends [40], which suggests a less prominent segregation between the B and C loci.

Using $\bar{\epsilon}_{\alpha\beta}$ for the interaction parameters, we performed SP-CCM simulations by confining 20 different chromosomes to a sphere to mimic the rod cell nucleus (see Methods for simulation details). Remarkably, without any parameter adjustments, our simulations accurately capture the trends in the distributions of relative radial position, r/R_0, obtained from the FISH experiment [39], where r is the distance of a given locus from the nuclear center and R_0 is the nuclear radius (Fig. 4**d**). Our polymer simulations show that the euchromatin

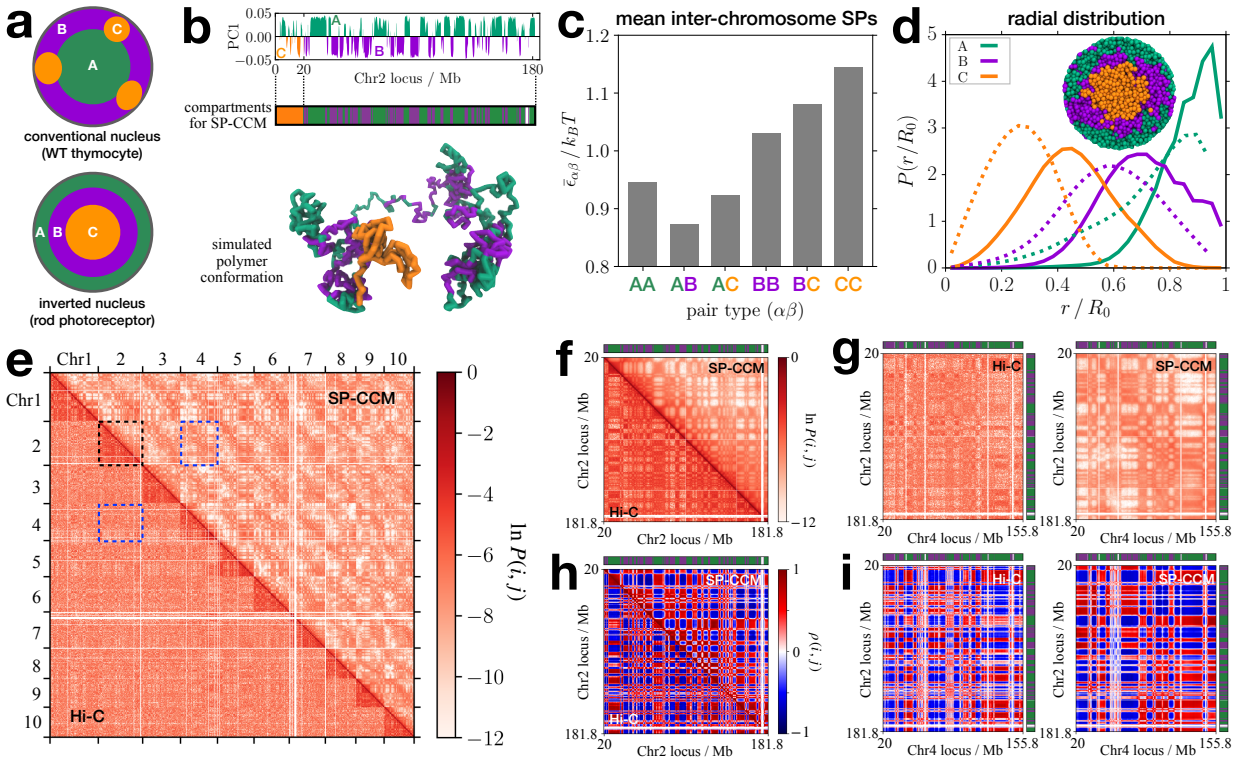

FIG. 4. **Applications of SPs for inverted nuclei.** (**a**) Schematic depictions of the conventional and inverted nuclei investigated by Falk *et al.* [25], showing the different spatial patterns of subnuclear segregation among euchromatin (A), heterochromatin (B), and pericentromeric heterochromatin (C). (**b**) Compartment types for each individual loci determined using the first principal component (PC1) from the Hi-C CM (top). In polymer simulations, the first 20Mb of each chromosome is assigned the C-type (center). The bottom panel shows a simulated conformation for Chr2. (**c**) Bar graph showing the average of the inter-chromosome SP for each pair type. (**d**) Distributions of the relative radial position for different locus types in the inverted nuclei, comparing the SP-CCM simulation results (solid lines) with the experimental data (dotted lines). A cross section of the simulated inverted nucleus is shown at the top. (**e**) Comparison of the CM (Chr1 to Chr10) between the Hi-C experiment (lower triangle) and the SP-CCM simulations (upper triangle). Gaussian smoothing with the standard deviation of the half bin size (100kb) was applied to the Hi-C CM to reduce noises. (**f**) Enlarged view of the black dotted square in panel **e**, showing the Hi-C and the simulated CMs for Chr2. (**g**) Enlarged view of the blue dotted squares in panel **e**, showing the Hi-C (left) and the simulated (right) CMs between Chr2 and Chr4. (**h**, **i**) Pearson correlation matrices, corresponding to the CMs in panels **f** and **g**.

is on the periphery and heterochromatin is predominantly in the interior, which is a key characteristic of inverted nuclei that differentiates them from conventional nuclei. We also

verified that individual chromosomes in the simulated nucleus exhibit the structures consistent with the Hi-C data (Fig. 4**e**). Figure 4**f** shows that the CM for Chr2 calculated from polymer simulations is in good agreement with the Hi-C inferred CM. In Fig. 4**g**, the similar extent of visual agreement holds for the inter-chromosome CM between Chr2 and Chr4. The corresponding Pearson correlation matrices (Figs. 4**h**-4**i**) show that the compartment formation is in excellent agreement between the simulation results and the Hi-C data (Fig. S10d). We compared the distributions for the Pearson coefficients from the simulations and the Hi-C, using the JSD between the distributions (Figs. S10e-S10g). The small JSD values ($\overline{\text{JSD}} = 0.061$) confirm the near quantitative agreement for the microphase separated chromosome organizations between the simulations and the experiments.

DISCUSSION

We demonstrated that the effective interaction energies between locus pairs are heterogeneous and differ depending on the pair type (A-A, B-B, or A-B). Without tweaking any parameters, the calculated SP values averaged over each pair type, ϵ_{AA}, ϵ_{BB}, and ϵ_{AB}, automatically satisfy $\chi_{\text{FH}} = [(\epsilon_{\text{AA}} + \epsilon_{\text{BB}})/2 - \epsilon_{\text{AB}}]/k_B T > 0$, which readily accounts for the microphase separation between A and B loci in interphase chromosomes. The use of the mean SP values in the polymer simulations generates the structural ensemble with A/B compartmentalization, which is in very good agreement with both Hi-C and FISH results. The SP-CCM simulations at a sufficiently high resolution (*e.g.*, higher than 50 kb) also resolve the TAD structures. Surprisingly, the SP theory faithfully captures the intrinsic differences in the inter-chromosome interactions in the inverted nuclei. Polymer simulations for multiple chromosomes, with the inter-chromosome SPs, predict the chromosome organizations in the inverted nuclei accurately. We find it remarkable that using only the measured CM and with no parameters in the simulations, we can nearly quantitatively describe the 3D structures of both the interphase chromosomes and those found in inverted nuclei.

Because the SP concept is general, it can be expanded and applied to other experimental data, such as super-resolution DNA FISH [24, 41], Micro-C [42, 43], GAM [44], and SPRITE [45]. Although the values of the calculated energy scales may differ depending on the experiment and resolution, the qualitative features should be conserved for a given system. For instance, the SPs based on the FISH data for the IMR90 Chr2 (Fig. S11) show

qualitatively the same trend as those inferred from the corresponding Hi-C data (Fig. 1**a**), that is, $\Delta G_{\rm AA} \lesssim \Delta G_{\rm BB} < \Delta G_{\rm AB}$. Our method based on the SP should serve as a guiding basis for characterizing the effective interaction scales encoded in experimental results on chromosome structures. In conjunction with polymer simulations, they provide a method for calculating chromosome dynamics as well.

METHODS

Statistical potential for a single interphase chromosome

We calculated the statistical potential (SP) between distinct, non-adjacent loci i and j (*i.e.*, $|i-j| \geq 2$) on a single chromosome by using,

$$\Delta G_{ij} = -k_B T \ln \left[\frac{P_{\rm exp}(i,j)}{P_{\rm ref}(i,j)} \right] , \tag{2}$$

where $P_{\rm exp}(i,j)$ and $P_{\rm ref}(i,j)$ are the measured contact probabilities for the pair (i,j) in a Hi-C experiment and in a reference system, respectively. We consider an ideal homopolymer as the reference system for which $P_{\rm ref}(i,j) = |i-j|^{-3/2}$. With this choice, ΔG_{ij} specifies the effective interaction for the contact pair, which excludes the free energy cost associated with the entropic contribution from the polymer backbone. For $P_{\rm exp}(i,j)$, the intrachromosome contact matrix, C, from a Hi-C experiment is converted to a probability matrix using $P_{\rm exp}(i,j) = \max[1, C(i,j)/\mathcal{N}]$, where $\mathcal{N}$ is a normalization factor. We choose $\mathcal{N}$ to be the mean value of the diagonal entries of the contact matrix, because a diagonal entry is the contact frequency for the genomic points that are almost always in contact at a given resolution and thus expected to have nearly unit contact probability.

In the calculation of the SPs for globular proteins, it is known that the choice of the reference system is important [6, 7]. The choice of the reference system for proteins and RNA may be assessed only by comparing predictions based on folding simulations using the knowledge-based potentials to experimental data. We follow a similar procedure here, and demonstrate that the ideal polymer is a reasonable choice.

Chromosome Copolymer Model (CCM)

In order to demonstrate the efficacy of the knowledge-based potential, calculated using Eq. (2), we performed polymer simulations using the CCM [19]. A chromosome is modeled as a flexible self-avoiding copolymer chain with two locus types, A and B, representing euchromatin and heterochromatin, respectively. Each monomer corresponds to a region of the genomic DNA binned at a given resolution (e.g. 10kbps, 50 kbps, 100 kbps, etc.). The locus type is determined using the procedure described below. The interactions involving the bonded and non-bonded pairs are described by the finite extensible nonlinear elastic (FENE) [46, 47] and the Lennard-Jones (LJ) potentials, respectively, which are given by,

$$u_{\mathrm{b}}(r) = -\frac{1}{2}K_{\mathrm{S}}b_{\mathrm{max}}^2 \ln\left(1 - \frac{r^2}{b_{\mathrm{max}}^2}\right) + u_{\mathrm{WCA}}(r) \ , \tag{3}$$

and

$$u_{\mathrm{nb}}(r|\epsilon,\sigma) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \ , \tag{4}$$

where K_S is the FENE spring constant, b_{max} is the maximum bond length, ϵ is the depth of the non-bonding potential well, and σ is the diameter of a locus. In Eq. (3), the excluded volume interaction between bonded loci is represented by the Weeks-Chandler-Anderson (WCA) potential [48], given by,

$$u_{\mathrm{WCA}}(r) = \left[u_{\mathrm{nb}}(r|\epsilon_{\mathrm{b}},\sigma) + \epsilon_{\mathrm{b}}\right]\Theta(r_{\mathrm{LJ}}^* - r) \ , \tag{5}$$

where $\Theta(x)$ is the Heaviside step function such that $\Theta(x) = 1$ if $x > 0$ and $\Theta(x) = 0$ if $x \leq 0$. The WCA potential $u_{\mathrm{WCA}}(r)$ is the repulsive tail of the LJ potential, which is shifted and truncated at $r_{\mathrm{LJ}}^* = 2^{1/6}\sigma$, which makes it decay to 0 smoothly ($u_{\mathrm{nb}}(r|\epsilon_{\mathrm{b}},\sigma)$ at the minimum at r_{LJ}^*). For the CTCF-mediated loops, we used a harmonic potential between loop-anchoring loci identified in the Hi-C experiment [26], defined as,

$$u_{\mathrm{loop}}(r) = K_{\mathrm{L}}(r-a)^2, \tag{6}$$

where K_{L} is the harmonic spring constant and a is the equilibrium bond length between the pair of loop anchors. Thus, the total potential energy of the CCM polymer chain with N loci is given by,

$$U(r^N) = \sum_{i=1}^{N-1} u_{\mathrm{b}}(r_{i,i+1}) + \sum_{i=1}^{N-2}\sum_{j=i+2}^{N} u_{\mathrm{nb}}(r_{i,j}|\epsilon_{\nu(i)\nu(j)},\sigma) + \sum_{\{p,q\}} u_{\mathrm{loop}}(r_{p,q}) \ , \tag{7}$$

where $r^N = \{\mathbf{r}_1, \cdots, \mathbf{r}_N\}$ is the set of positions of all the loci, $r_{i,j} = |\mathbf{r}_i - \mathbf{r}_j|$ is the distance between the ith and j^{th} loci, and $\{p, q\}$ is the set of indices for the loop anchors. In the second summation term in Eq. (7), $\nu(i) = \text{A or B}$, so there are three interaction parameters, ϵ_{AA}, ϵ_{BB}, and ϵ_{AB}, which set the depths of the attractive pairwise potential wells for given pair types. We take the diameter for A and B type loci to be identical.

In the previous study [19], we assumed that $\epsilon_{\text{AA}} = \epsilon_{\text{BB}} = \epsilon$, and performed a single-parameter search with the constraint that $\chi_{\text{FH}} = \epsilon - \epsilon_{\text{AB}} > 0$, which ensures that A and B loci undergo microphase separation, a procedure that was adopted recently [25]. In the present study, the interaction parameters are calculated using Eq. (2) with an experimentally measured contact map as the sole input. More precisely, we performed an average of the calculated SPs over a given pair type,

$$\epsilon_{\alpha\beta} = -\Delta G_{\alpha\beta} = -\frac{\sum_{i=1}^{N-2}\sum_{j=i+2}^{N} \Delta G_{ij}\delta_{\nu(i)\alpha}\delta_{\nu(j)\beta}}{\sum_{i=1}^{N-2}\sum_{j=i+2}^{N} \delta_{\nu(i)\alpha}\delta_{\nu(j)\beta}} \ , \tag{8}$$

where $\alpha, \beta = \text{A or B}$ and δ_{ij} is the Kronecker delta (*i.e.*, $\delta_{ij} = 1$ if $i = j$, or 0 else). The use of $\delta_{\nu(i)\alpha}$ and $\delta_{\nu(j)\beta}$ in Eq. (8) ensures that the summation includes the locus pairs of given type only. Because we use the Hi-C data at face value, there are no adjustable parameters in the energy function. Therefore, the SP theory and the polymer simulations are parameter free.

Determination of compartment types

From the experimental CM, the compartment (A or B) type of a given locus is determined by the standard procedure [18]. We first define a normalized contact matrix, C^*, whose elements are the observed contact frequency divided by the expected value at given genomic distance such that,

$$C^*(i,j) = \frac{(N - |i-j|)C(i,j)}{\displaystyle\sum_{k=1}^{N}\sum_{l\geq k+1}^{N} C(k,l)\delta_{|k-l|,|i-j|}} \ , \tag{9}$$

where the sum is taken over all the locus pairs. The normalized contact matrix is then converted to the Pearson correlation matrix, $\rho(i,j)$, which is the normalized covariance between two row vectors,

$$\rho(i,j) = \frac{\text{cov}(C_i^*, C_j^*)}{[\text{cov}(C_i^*, C_i^*)\,\text{cov}(C_j^*, C_j^*)]^{1/2}} \ . \tag{10}$$

Then we performed the principal component analysis (PCA) on $\rho(i,j)$. The sign (positive or negative) of each element in the first eigenvector indicates the compartment type of the corresponding locus (see Fig. 4**b**). By comparing the PCA vector with specific histone marker tracks (*e.g.*, H3K4me3 or H3K4me1 for active, H3K27me3 for inactive) for the reference human genomes [49, 50], we determine the sign that corresponds to A or B type (active or inactive). In the present study, the resulting compartment types of individual loci are used as the locus type, $\nu(i)$, for the CCM simulations.

TAD analysis

Given a contact matrix C, the insulation score for the nth locus is given by,

$$\mathrm{IS}(n) = \frac{1}{w^2}\sum_{i=1}^{w}\sum_{j=1}^{w} C(n-i, n+j) \ , \tag{11}$$

where w is the number of loci across which the contact frequencies are averaged. We used $w = 10$ at 50-kb resolution. In other words, the upstream and downstream regions of 500 kbs are considered. The computation of the insulation profile can be visualized by sliding a square of width w along the diagonal of the CM over which the contact frequencies are averaged [35]. The first and last w bins are not assigned any IS, as the insulation square would go beyond the given chromosome region. The local minima of the calculated IS profile correspond to the boundary positions of TADs, as a small value of IS indicates a region insulated from the contacts with neighboring regions.

For an identified TAD region, we calculated the radius of gyration, R_g, and the relative shape anisotropy, κ^2 (Fig. 2**d**). These quantities are determined from the gyration tensor, defined by,

$$S_{\alpha\beta} = \frac{1}{N(\theta)}\sum_{i\in\theta}^{N(\theta)}(r_i^\alpha - r_{\mathrm{cm}}^\alpha)(r_i^\beta - r_{\mathrm{cm}}^\beta) \ , \tag{12}$$

where θ is a given TAD region having $N(\theta)$ loci, r_{cm} is the position of the center of mass, and $\alpha, \beta = x, y$, or z so the superscripts on positions specify the three-dimensional coordinates. R_g and κ^2 are defined in terms of the gyration tensor, λ_α, as

$$R_g = (\lambda_x + \lambda_y + \lambda_z)^{1/2} \ , \tag{13}$$

and,

$$\kappa^2 = \frac{3(\lambda_x^2 + \lambda_y^2 + \lambda_z^2)}{2R_g^4} - \frac{1}{2} \ , \tag{14}$$

respectively.

Simulation details for the single-chromosome SP-CCM

To sample the conformations of the CCM for a specific chromosome, using the effective pair interaction energies extracted from the CM, we performed Langevin dynamics simulations by integrating the equation of motion,

$$m\ddot{\mathbf{r}}_i + \zeta\dot{\mathbf{r}}_i = -\frac{\partial}{\partial \mathbf{r}_i} U(\mathbf{r}_1, \cdots, \mathbf{r}_N) + \mathbf{R}_i(t) \ , \tag{15}$$

where m is the mass of a locus and ζ is the friction coefficient. The Gaussian random force, $\mathbf{R}_i(t)$, mimicking thermal fluctuations, has the mean, $\langle \mathbf{R}_i(t) \rangle = 0$, and the variance, $\langle \mathbf{R}_i(t) \cdot \mathbf{R}_j(t') \rangle = 6k_BT\zeta\delta_{ij}\delta(t-t')$. For the structural parameters in the energy function, U [Eq. (7)], we set $K_{\mathrm{S}} = 30k_BT/\sigma^2$, $b_{\max} = 1.5\sigma$, $\epsilon_{\mathrm{b}} = 1.0k_BT$, $K_{\mathrm{L}} = 300k_BT/\sigma^2$, and $a = 1.13\sigma$, which are similar to the values used in previous studies [19, 47, 51]. The interaction parameters, ϵ_{AA}, ϵ_{BB}, and ϵ_{AB}, were determined from the mean value of ΔG_{ij} for each locus pair type [see Eq. (8)]. The loop anchors were determined using the locations of CTCF loops identified by Rao *et al.* [26]. We only took the loops with CTCF motifs "uniquely" called at both the anchors. For the Chr2 simulations, whose results are shown in Fig. 1, we did not include CTCF loop anchors in order to focus on the A/B compartmentalization.

All the simulations were carried out in the reduced units ($m = \sigma = k_B = T = 1$), using the LAMMPS molecular dynamics program [52]. We chose the integration time step as $\Delta t_L = 0.01$ in units of $\tau_L = \sqrt{m\sigma^2/k_BT}$. The simulation temperature, $T = 1$, which corresponds to the room temperature in reduced units, was maintained using the Langevin thermostat. Each trajectory starts from a random configuration corresponding to a self-avoiding walk polymer, which is relaxed for 10^7–10^8 steps (varied depending on N) after which we find that the total energy and the radius of gyration fluctuate around plateau values. Subsequently, we propagate the system for additional 10^7–10^8 steps from which we chose 10^4 conformations that are equally spaced along the production run. To obtain an ensemble-averaged CM, a minimum of 20 independent trajectories were generated.

SPs for inter-chromosome interactions

The effective energy scales for inter-chromosome interactions may also be calculated using the SP theory. By generalizing Eq. (2), we define the SP between the ith and j^{th} loci belonging to chromosomes n and m, respectively, as,

$$\Delta G_{ij}^{(n,m)} = -k_B T \ln \left[\frac{P_{\text{exp}}(i,j|n,m)}{P_{\text{ref}}(i,j|n,m)} \right] , \tag{16}$$

where $P_{\text{exp}}(i,j|n,m)$ is the inter-chromosome contact probability for the pair i and j, inferred from a Hi-C experiment. For the reference contact probability, $P_{\text{ref}}(i,j|n,m)$, we take the average contact probability over all the inter-chromosome locus pairs, by assuming that chromosomes can intermingle with one another in the reference system. We define $P_{\text{ref}}(i,j|n,m)$ using,

$$P_{\text{ref}}(i,j|n,m) \equiv P_{\text{ref}}^{(\text{inter})} = \frac{2}{M(M-1)} \sum_{n=1}^{M-1} \sum_{m=n+1}^{M} \bar{P}_{\text{exp}}^{(n,m)} , \tag{17}$$

where M is the total number of chromosomes (*e.g.*, $M = 23$ for human and $M = 20$ for mouse) and $\bar{P}_{\text{exp}}^{(n,m)} = \frac{1}{N^{(n)}N^{(m)}} \sum_{i=1}^{N^{(n)}} \sum_{j=1}^{N^{(m)}} P_{\text{exp}}(i,j|n,m)$ is the average contact probability for the locus pairs between chromosomes n and m, which have $N^{(n)}$ and $N^{(m)}$ loci, respectively. The mean value of the SP for the locus types α and β, in a given chromosome pair, is computed using,

$$\epsilon_{\alpha\beta}^{(n,m)} = -\Delta G_{\alpha\beta}^{(n,m)} = -\frac{\sum_{i=1}^{N^{(n)}} \sum_{j=1}^{N^{(m)}} \Delta G_{ij}^{(n,m)} \left(\delta_{\nu(i)\alpha}\delta_{\nu(j)\beta} + \delta_{\nu(i)\beta}\delta_{\nu(j)\alpha} \right)}{N_\alpha^{(n)} N_\beta^{(m)} + N_\beta^{(n)} N_\alpha^{(m)}} , \tag{18}$$

where $N_\alpha^{(n)}$ and $N_\beta^{(m)}$ are the numbers of loci with types α and β in chromosomes n and m, respectively.

Following Falk *et al.* [25], we assigned the compartment types, A, B, or C to individual chromosome loci binned at 200-kb resolution for the mouse rod cell. In each chromosome, the A and B types for euchromatin and heterochromatin were determined using the same procedure as described above. The B-type loci in the first 20 Mb were classified as the C type for pericentromeric heterochromatin (see Fig. 4**b**). Based on the locus types, we obtained $\epsilon_{\alpha\beta}^{(n,m)}$ for $\alpha, \beta =$ A, B, or C, whose distribution is shown in Fig. S10c.

Polymer simulations for an inverted nucleus

We prepared a system with 20 polymer chains, which model chromosomes 1 to X from a mouse rod cell at 200-kb resolution. There are 13,203 loci in total, whose types are determined as described above. In each chromosome, the first 20Mb is collectively redefined as the C-type loci which forms the chromocenter involving centromeres and pericentromeric heterochromatin (Fig. 4**b**), so there are 2,000 C-type loci $\sim 15\%$ of the entire system (*cf.* 16% was assigned the C loci in [25]). The diameters (σ's) for the A, B, and C type loci are identical. The potential energy for the multi chain system is,

$$\begin{aligned} U_{\rm multi}(r^{N^{(1)}}, \cdots, r^{N^{(M)}}) &= \sum_{n=1}^{M} U_{\rm intra}^{(n)}(r^{N^{(n)}}) + \sum_{n=1}^{M-1} \sum_{m=n+1}^{M} U_{\rm inter}^{(n,m)}(r^{N^{(n)}}, r^{N^{(m)}}) \\ &= \sum_{n=1}^{M} \left[\sum_{i=1}^{N^{(n)}-1} u_{\rm b}(r_{i,i+1}^{(n,n)}) + \sum_{i=1}^{N^{(n)}-2} \sum_{j=i+2}^{N^{(n)}} u_{\rm nb}(r_{i,j}^{(n,n)} | \epsilon_{\nu(i)\nu(j)}^{(n)}, \sigma) \right] \\ &\quad + \sum_{n=1}^{M-1} \sum_{m=n+1}^{M} \sum_{i=1}^{N^{(n)}} \sum_{j=1}^{N^{(m)}} u_{\rm nb}(r_{i,j}^{(n,m)} | \epsilon_{\nu(i)\nu(j)}^{(n,m)}, \sigma) \ , \end{aligned} \tag{19}$$

where $r^{N^{(n)}} = \{\mathbf{r}_1, \cdots, \mathbf{r}_{N^{(n)}}\}$ are the positions of all the loci in the $n^{\rm th}$ chromosome, and $r_{i,j}^{(n,m)} = |\mathbf{r}_i^{(n)} - \mathbf{r}_j^{(m)}|$ is the distance between the $i^{\rm th}$ and $j^{\rm th}$ loci in chromosomes n and m, respectively. Following Falk *et al.* [25], we assumed that the interaction parameters for both intra- and inter-chromosome pairs are given by,

$$\bar{\epsilon}_{\alpha\beta} = \frac{2}{M(M-1)} \sum_{n=1}^{M-1} \sum_{m=n+1}^{M} \epsilon_{\alpha\beta}^{(n,m)} \ . \tag{20}$$

In other words, $\epsilon_{\alpha\beta}^{(n)} = \epsilon_{\alpha\beta}^{(n,m)} = \bar{\epsilon}_{\alpha\beta}$ for all n and m. In some instances, it is important to distinguish between intra- and inter-chromosome interactions in order to capture the relative positions of chromosomes and the chromosome territories [53]. Nevertheless, the non-discrimination between intra- and inter-chromosome interactions is a reasonable assumption for simulating the inverted nuclei because the inter-chromosome contacts are higher in mouse rod cells than in other cell types [25]. In recent studies [54, 55], whole-genome polymer simulations yielded the inter-chromosome CMs that are in good agreement with the Hi-C data, even without differentiating between inter- and intra-chromosome ones.

The simulations were performed using the same reduced units as in the single-chromosome simulations. The CCM chains were initially placed on a square lattice such that they were

equally spaced from one another (by 2σ) in the linearly extended configurations. The chains were relaxed to a collapsed state using the Langevin thermostat. The collapsed polymer chains were then equilibrated for $5 \times 10^7 \Delta t_L$ under spherical confinement (the radius of the sphere is $R_0 = 15\sigma$), which ensures that the volume density is similar to that for the mouse rod cell nucleus with $\sigma \approx 0.125\,\mu$m (the nuclear diameter of a mouse rod cell is $\approx 4.8\,\mu$m so the rod cell chromosomes are more compact than the human IMR90 chromosomes). The equilibrated system was propagated for an additional $5 \times 10^7 \Delta t_L$ from which 10^4 conformations were sampled along the production run. We generated 50 independent trajectories to obtain the statistics for the radial distributions and the contact map that are shown in Figs. 4**d** and 4**e** in the main text.

Analysis for interchromosomal CMs

For quantitative comparison of the compartment patterns between the Hi-C and the simulated interchromosome CMs, we calculated the correlation matrix for all the genomic region of our interest (Chr1 to Chr10, as shown in Figs. 4**e** and S10d). We first rescaled the normalized contact matrix, $C^{(n)*}$, defined for chromosome n (Eq. (9), by the mean interchromosome contacts,

$$\tilde{C}^{(n)*}(i,j) = \frac{C^{(n)*}(i,j)}{\overline{C^{(n)*}}(M-1)} \sum_{k \neq n}^{M} \bar{C}^{(n,k)} \ , \tag{21}$$

where $\overline{C^{(n)*}}$ is the mean value of all the elements in $C^{(n)*}$, and $\bar{C}^{(n,k)}$ is the mean contact frequency in the interchromosomal CM, $C^{(n,k)}$ (from either Hi-C or simulations), between chromosomes n and k. The rescaling removes statistical bias in the intrachromosomal CM for a given chromosome relative to the contacts with other chromosomes. Hence, the normalized and rescaled CM for the entire genomic region can be written as the following block matrix form,

$$\tilde{C}^* = \begin{bmatrix} \tilde{C}^{(1)*} & C^{(1,2)} & \dots & C^{(1,M)} \\ C^{(2,1)} & \tilde{C}^{(2)*} & \dots & C^{(2,M)} \\ \vdots & \vdots & \ddots & \vdots \\ C^{(M,1)} & C^{(M,2)} & \dots & \tilde{C}^{(M)*} \end{bmatrix} \ . \tag{22}$$

This matrix is then converted to the Pearson correlation matrix, $\rho(i,j)$, as defined in Eq. (10). For the Hi-C data, gaussian smoothing with the standard deviation of the bin

size (200kb) was applied to $\tilde{C}^*$ before computing the correlation matrix.

Jensen-Shannon divergence

The Jensen-Shannon divergence (JSD) between two probability distributions, $P_1(x)$ and $P_2(x)$, is defined as,

$$\mathrm{JSD}(P_1||P_2) = \frac{1}{2}\int dx \sum_{i=1}^{2} P_i(x)\log_2\left[\frac{P_i(x)}{P_{\mathrm{M}}(x)}\right] , \tag{23}$$

where $P_{\mathrm{M}}(x) = (P_1(x) + P_2(x))/2$. The value of JSD is zero for $P_1(x) = P_2(x)$, and is unity if the distributions do not have any overlap, *i.e.*, $\int P_1(x)P_2(x)dx = 0$. For instance, if $P_1(x)$ and $P_2(x)$ are Gaussian distributions with identical variance σ_{s}^2 but different means, μ_1 and μ_2, then $\mathrm{JSD}(P_1||P_2) \approx 0.16$ for $|\mu_1 - \mu_2| = \sigma_{\mathrm{s}}$ and $\mathrm{JSD}(P_1||P_2) \approx 0.40$ for $|\mu_1 - \mu_2| = 2\sigma_{\mathrm{s}}$. If $P_1(x)$ and $P_2(x)$ have an identical mean but their standard deviations differ by a factor of two ($\sigma_{\mathrm{s},1} = 2\sigma_{\mathrm{s},2}$), then $\mathrm{JSD}(P_1||P_2) \approx 0.13$.

ACKNOWLEDGEMENT

We acknowledge Hung Nguyen, Mauro Mugnai, Debayan Chakraborty, and Davin Jeong for useful discussions. This work was supported by grants from the National Science Foundation (CHE 19-00033) and the Welch Foundation (F-0019) administered through the Collie-Welch Regents Chair.

---

[1] S. Tanaka and H. A. Scheraga. Medium-and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins. *Macromolecules*, 9(6): 945–950, 1976.

[2] S. Miyazawa and R. L. Jernigan. Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation. *Macromolecules*, 18(3):534–552, 1985.

[3] S. Miyazawa and R. L. Jernigan. Residue – Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. *Journal of Molecular Biology*, 256(3):623–644, 1996.

[4] M. J. Sippl. Knowledge-based potentials for proteins. *Current Opinion in Structural Biology*, 5(2):229–235, 1995.

[5] P. D. Thomas and K. A. Dill. An iterative method for extracting energy-like quantities from protein structures. *Proceedings of the National Academy of Sciences of the United States of America*, 93(21):11628–11633, 1996.

[6] J. Skolnick, L. Jaroszewski, A. Kolinski, and A. Godzik. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? *Protein Science*, 6(3):676–688, 1997.

[7] M. R. Betancourt and D. Thirumalai. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Science*, 8(2):361–369, 1999.

[8] R. I. Dima, G. Settanni, C. Micheletti, J. R. Banavar, and A. Maritan. Extraction of interaction potentials between amino acids from native protein structures. *Journal of Chemical Physics*, 112(20):9151–9166, 2000.

[9] E. P. O'Brien, G. Ziv, G. Haran, B. R. Brooks, and D. Thirumalai. Effects of denaturants and osmolytes on proteins are accurately predicted by the molecular transfer model. *Proceedings of the National Academy of Sciences of the United States of America*, 105(36):13403–13408, 2008.

[10] O. Schueler-Furman, Y. Altuvia, A. Sette, and H. Margalit. Structure-based prediction of binding peptides to MHC class I molecules: Application to a broad range of MHC alleles. *Protein Science*, 9(9):1838–1846, 2000.

[11] R. S. DeWitte, A. V. Ishchenko, and E. I. Shakhnovich. SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 2. Case studies in molecular design. *Journal of the American Chemical Society*, 119(20):4608–4617, 1997.

[12] H. Lu, L. Lu, and J. Skolnick. Development of unified statistical potentials describing protein-protein interactions. *Biophysical Journal*, 84(3):1895–1901, 2003.

[13] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, and D. Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[14] R. I. Dima, C. Hyeon, and D. Thirumalai. Extracting stacking interaction parameters for RNA from the data set of native structures. *Journal of Molecular Biology*, 347(1):53–69, 2005.

[15] Q. Vicens and J. S. Kieft. Thoughts on how to think (and talk) about RNA structure. *Proceedings of the National Academy of Sciences of the United States of America*, 119(17): 1–9, 2022.

[16] A. E. Walter, D. H. Turner, J. Kim, M. H. Lyttle, P. Müller, D. H. Mathews, and M. Zuker. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proceedings of the National Academy of Sciences of the United States of America*, 91(20):9218–9222, 1994.

[17] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5):911–940, 1999.

[18] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950):289–293, 2009.

[19] G. Shi, L. Liu, C. Hyeon, and D. Thirumalai. Interphase human chromosome exhibits out of equilibrium glassy dynamics. *Nature Communications*, 9(1):3161, 2018.

[20] M. L. Huggins. Solutions of long chain compounds. *The Journal of Chemical Physics*, 9(5): 440, 1941.

[21] P. J. Flory. Thermodynamics of High Polymer Solutions. *The Journal of Chemical Physics*, 9(8):660, 1941.

[22] D. W. Fawcett. On the occurrence of a fibrous lamina on the inner aspect of the nuclear envelope in certain cells of vertebrates. *American Journal of Anatomy*, 119(1):129–145, 1966.

[23] T. Misteli. The Self-Organizing Genome: Principles of Genome Architecture and Function. *Cell*, 183(1):28–45, 2020.

[24] J. H. Su, P. Zheng, S. S. Kinrot, B. Bintu, and X. Zhuang. Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. *Cell*, 182(6):1641–1659.e26, 2020.

[25] M. Falk, Y. Feodorova, N. Naumova, M. Imakaev, B. R. Lajoie, H. Leonhardt, B. Joffe, J. Dekker, G. Fudenberg, I. Solovei, and L. A. Mirny. Heterochromatin drives compartmen-

talization of inverted and conventional nuclei. *Nature*, 570(7761):395–399, 2019.

[26] S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7): 1665–1680, 2014.

[27] L. Giorgetti, R. Galupa, E. P. Nora, T. Piolot, F. Lam, J. Dekker, G. Tiana, and E. Heard. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell*, 157(4):950–963, 2014.

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

[29] M. Di Pierro, B. Zhang, E. L. Aiden, P. G. Wolynes, and J. N. Onuchic. Transferable model for chromosome architecture. *Proceedings of the National Academy of Sciences*, 113(43):12168–12173, 2016.

[30] Q. Szabo, D. Jost, J. M. Chang, D. I. Cattoni, G. L. Papadopoulos, B. Bonev, T. Sexton, J. Gurgo, C. Jacquier, M. Nollmann, F. Bantignies, and G. Cavalli. TADs are 3D structural units of higher-order chromosome organization in Drosophila. *Science Advances*, 4(2):1–14, 2018.

[31] S. Bianco, D. G. Lupiáñez, A. M. Chiariello, C. Annunziatella, K. Kraft, R. Schöpflin, L. Wittler, G. Andrey, M. Vingron, A. Pombo, S. Mundlos, and M. Nicodemi. Polymer physics predicts the effects of structural variants on chromatin architecture. *Nature Genetics*, 50(5): 662–667, 2018.

[32] A. Perez-Rathke, Q. Sun, B. Wang, V. Boeva, Z. Shao, and J. Liang. CHROMATIX: Computing the functional landscape of many-body chromatin interactions in transcriptionally active loci from deconvolved single cells. *Genome Biology*, 21(1):1–17, 2020.

[33] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

[34] T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, and G. Cavalli. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, 148(3):458–472, 2012.

[35] E. Crane, Q. Bian, R. P. McCord, B. R. Lajoie, B. S. Wheeler, E. J. Ralston, S. Uzawa, J. Dekker, and B. J. Meyer. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, 523(7559):240–244, 2015.

[36] L. Giorgetti, B. R. Lajoie, A. C. Carter, M. Attia, Y. Zhan, J. Xu, C. J. Chen, N. Kaplan, H. Y. Chang, E. Heard, and J. Dekker. Structural organization of the inactive X chromosome in the mouse. *Nature*, 535(7613):575–579, 2016.

[37] E. M. Darrow, M. H. Huntley, O. Dudchenko, E. K. Stamenova, N. C. Durand, Z. Sun, S. C. Huang, A. L. Sanborn, I. Machol, M. Shamim, A. P. Seberg, E. S. Lander, B. P. Chadwick, and E. L. Aiden. Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proceedings of the National Academy of Sciences of the United States of America*, 113(31):E4504–E4512, 2016.

[38] C. Y. Wang, T. Jégu, H. P. Chu, H. J. Oh, and J. T. Lee. SMCHD1 Merges Chromosome Compartments and Assists Formation of Super-Structures on the Inactive X. *Cell*, 174(2): 406–421.e25, 2018.

[39] I. Solovei, M. Kreysing, C. Lanctôt, S. Kösem, L. Peichl, T. Cremer, J. Guck, and B. Joffe. Nuclear Architecture of Rod Photoreceptor Cells Adapts to Vision in Mammalian Evolution. *Cell*, 137(2):356–368, 2009.

[40] M. Rubinstein and R. H. Colby. *Polymer physics*. New York: Oxford university press, 2003.

[41] B. Bintu, L. J. Mateo, J.-H. Su, N. A. Sinnott-Armstrong, M. Parker, S. Kinrot, K. Yamaya, A. N. Boettiger, and X. Zhuang. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, 362(6413):419, 2018.

[42] T. H. S. Hsieh, A. Weiner, B. Lajoie, J. Dekker, N. Friedman, and O. J. Rando. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell*, 162(1):108–119, 2015.

[43] T. H. S. Hsieh, C. Cattoglio, E. Slobodyanyuk, A. S. Hansen, O. J. Rando, R. Tjian, and X. Darzacq. Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. *Molecular Cell*, 78(3):539–553.e8, 2020.

[44] R. A. Beagrie, A. Scialdone, M. Schueler, D. C. Kraemer, M. Chotalia, S. Q. Xie, M. Barbieri, I. De Santiago, L. M. Lavitas, M. R. Branco, J. Fraser, J. Dostie, L. Game, N. Dillon, P. A. Edwards, M. Nicodemi, and A. Pombo. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, 543(7646):519–524, 2017.

[45] S. A. Quinodoz, N. Ollikainen, B. Tabak, A. Palla, J. M. Schmidt, E. Detmar, M. M. Lai, A. A. Shishkin, P. Bhat, Y. Takei, V. Trinh, E. Aznauryan, P. Russell, C. Cheng, M. Jovanovic, A. Chow, L. Cai, P. McDonel, M. Garber, and M. Guttman. Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell*, 174(3):744–757.e24, 2018.

[46] H. R. Warner. Kinetic Theory and Rheology of Dilute Suspensions of Finitely Extendible Dumbbells. *Industrial and Engineering Chemistry Fundamentals*, 11(3):379–387, 1972.

[47] K. Kremer and G. S. Grest. Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. *The Journal of Chemical Physics*, 92(8):5057–5086, 1990.

[48] J. D. Weeks, D. Chandler, and H. C. Andersen. Role of repulsive forces in determining the equilibrium structure of simple liquids. *The Journal of Chemical Physics*, 54(12):5237–5247, 1971.

[49] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[50] Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–329, 2015.

[51] H. Kang, Y.-G. Yoon, D. Thirumalai, and C. Hyeon. Confinement-Induced Glassy Dynamics in a Model for Chromosome Organization. *Physical Review Letters*, 115(19):198102, 2015.

[52] S. Plimpton. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics*, 117:1–19, 1995.

[53] Y. Qi, A. Reyes, S. E. Johnstone, M. J. Aryee, B. E. Bernstein, and B. Zhang. Data-Driven Polymer Model for Mechanistic Exploration of Diploid Genome Organization. *Biophysical Journal*, 119(9):1905–1916, 2020.

[54] S. Fujishiro and M. Sasai. Generation of dynamic three-dimensional genome structure through phase separation of chromatin. *Proceedings of the National Academy of Sciences of the United States of America*, 119(22), 2022.

[55] V. G. Contessoto, O. Dudchenko, E. L. Aiden, P. G. Wolynes, J. N. Onuchic, and M. Di Pierro. Interphase chromosomes of the Aedes aegypti mosquito are liquid crystalline and can sense mechanical cues. *Nature Communications*, 14(1):1–12, 2023.