From Hi-C Contact Map to Three-dimensional Organization of Interphase Human Chromosomes

Guang Shi[1] and D. Thirumalai[1,*]

[1] *Department of Chemistry, University of Texas at Austin, 78712*

The probabilities that two loci in chromosomes that are separated by a certain genome length can be inferred using chromosome conformation capture method and related Hi-C experiments. How to go from such maps to an ensemble of three-dimensional structures, which is an important step in understanding the way nature has solved the packaging of the hundreds of million base pair chromosomes in tight spaces, is an open problem. We created a theory based on polymer physics and the maximum entropy principle, leading to the HIPPS (Hi-C-Polymer-Physics-Structures) method allows us to go from contact maps to 3D structures. It is difficult to calculate the mean distance ($\langle \bar{r}_{ij} \rangle$) between loci i and j from the contact probability ($\langle \bar{p}_{ij} \rangle$) because the contact exists only in a fraction (unknown) of cell populations. Despite this massive heterogeneity, we first prove that there is a theoretical lower bound connecting $\langle p_{ij} \rangle$ and $\langle \bar{r}_{ij} \rangle$ via a power-law relation. We show, using simulations of a precisely solvable model, that the overall organization is accurately captured by constructing the distance map from the contact map even when the cell population is heterogeneous, thus justifying the use of the lower bound. Building on these results and using the mean distance matrix, whose elements are $\langle \bar{r}_{ij} \rangle$, we use maximum entropy principle to reconstruct the joint distribution of spatial positions of the loci, which creates an ensemble of structures for the 23 chromosomes from lymphoblastoid cells. The HIPPS method shows that the conformations of a given chromosome are highly heterogeneous even in a single cell type. Nevertheless, the differences in the heterogeneity of the same chromosome in different cell types (normal as well as cancerous cells) can be quantitatively discerned using our theory.

INTRODUCTION

The question of how chromosomes are packed in the tight space of the cell nucleus has taken center stage in genome biology, largely due to the spectacular advances in experimental techniques. In particular, the routine generation of a large number of probabilistic contact maps for many species using the remarkable Hi-C technique [1–6] has provided us a glimpse of the genome organization. This in turn has opened several avenues of research with the hope of understanding the many features associated with chromosomes, such as how they are packaged in the nucleus, and how the chromosome organization affects the dynamics, and eventually function. A high contact count between two loci means that they interact with each other more frequently compared to ones with low contact count. Thus, the Hi-C data describes the chromosome structures in statistical terms expressed approximately in terms of a matrix, the elements of which indicate the probability that two loci separated by a specific genomic distance are in contact. The Hi-C data provide only a two-dimensional (2D) representation of the multidimensional organization of the chromosomes. How can we go beyond the genomic contact information to 3D distances between the loci, and eventually the spatial location of each locus is an important unsolved problem. Imaging techniques, such as Fluorescence *In Situ* Hybridization (FISH) and its variations, are the most direct way to measure the spatial distance and coordinates of the genomic loci [7]. But currently, these techniques are limited in scope because currently they provide information on only a small number of loci in a given experimental setup. Is it possible to harness the power of the two methods to construct, at least approximately, 3D structures of chromosomes? Here, we answer this question in the affirmative by building on the precise results for an exactly solvable Generalized Rouse Model for chromosomes [8, 9], and by using certain unusual polymer physics principles governing genome organization.

Several data-driven approaches have been developed in order to go from Hi-C to 3D structure of genomes [10–17] (see the summary in [18] for additional related studies). Although these methods are insightful, they do not predict the physical dimensions of the organized chromosomes nor have the methods been validated, especially when the structures are highly heterogeneous. These are difficult problems to solve using solely data-driven based approaches to infer structures from Hi-C data, without physical considerations, reflected in the polymeric features of the chromosomes. One problem is associated with the difficulty in reconciling Hi-C (contact probability) and the FISH data (spatial distances) [19–22]. For example, in interpreting the Hi-C contact map, one makes the intuitively plausible assumption that loci with high contact probability must also be spatially close. However, it has been demonstrated using Hi-C and FISH data that high contact frequency does not always imply proximity in space [19–22]. Because the cell population is heterogeneous, even though they are synchronized in the Hi-C experiments, a given contact is not present with unit probability in all the cells. Elsewhere [9], we showed that the heterogeneity in the genome or-

* dave.thirumalai@gmail.com

ganization is the reason for the absence of one-to-one relation between contact probability and spatial distance between a pair of loci. The inconsistency between Hi-C and FISH experiments makes it difficult to extract the ensemble of 3D structures of chromosomes using Hi-C data alone without taking into account the physics driving the condensed state of genomes. Even if one were to construct polymer models that produce results that are consistent with the inferred contact map from Hi-C, certain features of the chromosome structures would be discordant with the FISH data, reflecting the heterogeneous genome organization[23].

Despite the difficulties alluded to above, we have created a theory, based on polymer physics concept and the principle of maximum entropy to determine the 3D organization solely from the Hi-C data. The resulting physics-based data-driven method, which translates Hi-C data through polymer physics to average 3D coordinates of each loci, is referred to as HIPPS (Hi-C-Polymer-Physics-Structures). The purposes in the development creating and applications of the HIPPS method are two fold. (1) We first establish that there is a lower theoretical bound connecting the contact probability and the mean 3D distance in the presence of heterogeneity in the genome organization. We prove this concept by using the Generalized Rouse Model for Chromosomes (GRMC) for which accurate simulations can be performed. (2) However, mean spatial distances, $\langle r_{ij}\rangle$s, between the loci do not give the needed 3D structures. In addition, it is important to determine the variability in chromosome structures because massive conformational heterogeneity has been noted both in experiments [23, 30] and computations [9]. In order to solve this non-trivial problem, we use the principle of maximum entropy to obtain the ensemble of individual chromosome structures. The HIPPS method, which allows us to go from the Hi-C contact map to the three-dimensional coordinates, $\boldsymbol{x}_i$ $(i = 1, 2, 3, \cdots, N_c)$, where N_c is the length of the chromosome, may be summarized as follows. First, we construct the mean distances $\langle r_{ij}\rangle$ between all i and j using a power-law relation connecting $\langle p_{ij}\rangle$, the probability that the loci i and j are in contact measured in Hi-C experiments, and $\langle r_{ij}\rangle$. The justification for the power law relation is established using GRMC and polymer physics concepts. Then, using the maximum entropy distribution $P(\{\boldsymbol{x}_i\})$ with $\langle r_{ij}\rangle$ as constraints, we obtained an ensemble of chromosome 3D structures (the 3D coordinates for all the loci).

We tested the HIPPS procedure rigorously using the GRMC, which accounts for the massive heterogeneity noted in recent experiments [23]. The application of our theory to decipher the 3D structure of chromosomes from any species is limited only by the experimental resolution of the Hi-C technique. Comparisons with experimental data for sizes and volumes of chromosomes derived from the calculated 3D structures are made to validate the theory. Our method predicts that the structures of a given chromosome within a single cell and in different cell types is heterogeneous. Remarkably, the HIPPS method can detect the differences in the extent of heterogeneity of a specific chromosome among both normal can cancer cells.

RESULTS

Inferring the mean distance matrix ($\bar{\mathbf{R}}$) from the contact probability matrix ($\mathbf{P}$) for a homogeneous cell population: The elements, $\bar{r}_{ij}$, of the $\bar{\mathbf{R}}$ matrix give the *mean* spatial distance between loci i and j. Note that r_{ij} is the distance value for one realization of the genome conformation in a homogeneous population of cells. In this case a given contact is present with non-zero probability in all the entire cell population. The elements p_{ij} of the $\mathbf{P}$ matrix is the contact probability between loci i and j. We first establish a power law relation between $\bar{r}_{ij}$ and p_{ij} in a precisely solvable model. For the Generalized Rouse Model for chromosomes (GRMC), described in Appendix A, the relation between $\bar{r}_{ij}$ and p_{ij} is given by,

$$\begin{aligned} p_{ij} &= \mathrm{erf}(2r_c/\sqrt{\pi}\bar{r}_{ij}) - (4\pi/r_c\bar{r}_{ij})e^{-4r_c^2/\pi\bar{r}_{ij}^2} \\ &\equiv f_{\mathrm{GRMC}}(\bar{r}_{ij}). \end{aligned} \tag{1}$$

where $\mathrm{erf}(\cdot)$ is the error function, and r_c is the threshold distance for determining if a contact is established. This equation provides a way to calculate the distance matrix ($\bar{\mathbf{R}}$) directly from the contact matrix ($\mathbf{P}$) by inverting $f_{\mathrm{GRMC}}(\bar{r}_{ij})$. Note that $\mathbf{P}$ is inferred only approximately from Hi-C experiments. However, there are uncertainties, in determining both r_c due to systematic errors, and p_{ij} due to inadequate sampling, thus restricting the use of Eq.1 in practice. In light of these considerations, we address the following questions: (a) How accurately can one solve the inverse problem of going from the $\mathbf{P}$ to the $\bar{\mathbf{R}}$? (b) Does the inferred $\bar{\mathbf{R}}$ faithfully reproduce the topology of the spatial organization of chromosomes? We use GRMC to answer these questions.

To answer these two questions, we first constructed the distance map by solving Eq.1 for $\bar{r}_{ij}$ for every pair with contact probability p_{ij}. The $\mathbf{P}$ matrix is calculated using simulations of the GRMC, as described in Appendix B. For such a large polymer, some contacts are almost never formed even in long simulations, resulting in $p_{ij} \approx 0$ for some loci. This would erroneously suggest that $\bar{r}_{ij} \to \infty$, as a solution to Eq.1. Indeed, this situation arises often in the Hi-C experimental contact maps where $p_{ij} \approx 0$ for many i and j. To overcome the practical problem of dealing with $p_{ij} \approx 0$ for several pairs, we apply the block average (a coarse-graining procedure) to $\mathbf{P}$ (described in Appendix C), which decreases the size of the $\mathbf{P}$. The procedure overcomes the problem of having to deal with vanishingly small values of p_{ij} while simultaneously preserving the information needed to solve the inverse problem using Eq.1.

The simulated and constructed distance maps are

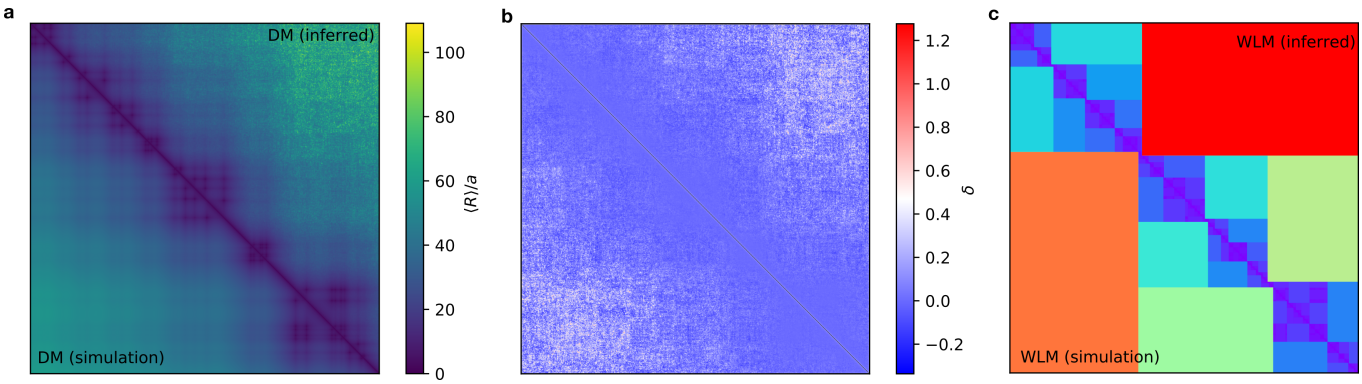

FIG. 1. Comparison of the distance matrices (DMs) for the GRMC. **(a)** The simulated DM (lower triangle) and constructed DM (upper triangle) are compared side by side. The color bar indicates the value of the mean spatial distance, $\langle R_{mn} \rangle$. The constructed DM is obtained by solving Eq.3 using the CM (calculated using Eq.11). The matrix size is 2000×2000 after the block averaging is applied to the raw data (Appendix C). The value of $r_c = 2.0a$. The location of loop anchors are derived from experimental data [6] over the range from 146 Mbps to 158 Mbps for Chromosome 5 in the Human GM12878 cell. **(b)** Relative error δ is represented as a map. The relative error is calculated as, $\delta = (d_{\rm I} - d_{\rm S})/d_{\rm S}$, where $d_{\rm I}$ and $d_{\rm S}$ are the inferred and simulated distances, respectively; δ increases for loci with large genomic distance indicating the tendency to overestimate the distances. **(c)** Ward Linkage Matrices (WLMs) from the simulation and theoretical prediction, shown in the lower and upper triangle, respectively, are in excellent agreement with each other.

shown in the lower and upper triangle, respectively (Fig.1a). We surmise from Fig.1a that the two distance maps are in excellent agreement with each other. There is a degree of uncertainty for the loci pairs with large mean spatial distance (elements far away from the diagonal (Fig.1a,b) due to the unavoidable noise in the contact probability matrix $\mathbf{P}$. The Spearman correlation coefficient between the simulated and theoretically constructed maps is 0.97, which shows that the distance matrix can be accurately constructed. However, a single correlation coefficient is not sufficient to capture the topological structure embedded in the distance map. To further assess the global similarity between the $\bar{\mathbf{R}}$ from theory and simulations, we used the Ward Linkage Matrix [24], which we previously used to determine the spatial organization in interphase chromosomes [25]. Fig.1c shows that the constructed $\bar{\mathbf{R}}$ indeed reproduces the hierarchical structural information accurately. These results together show that the matrix $\bar{\mathbf{R}}$, in which the elements represent the mean distance between the loci, can be calculated accurately, as long as the $\mathbf{P}$ is determined unambiguously. As is well known, this is not possible to do in Hi-C experiments, which renders solving the problem of going from $\mathbf{P}$ to $\bar{\mathbf{R}}$, and eventually the precise three-dimensional structure extremely difficult.

A bound for the spatial distance inferred from contact probability: The results in Fig.1 show that for a homogeneous system (specific contacts are present in all realizations of the polymer), $\bar{\mathbf{R}}$ can be faithfully reconstructed solely from the $\mathbf{P}$. However, the discrepancies between FISH and Hi-C data in several loci pairs [26] suggest that the cell population is heterogeneous, which means that contact between i and j loci is present in only a fraction of the cells. In this case, which one has to contend with in practice [9, 23], the one-to-one mapping between the contact probability and the mean 3D distances (as shown by Eq.1) does not hold, leading to the paradox [19, 20] that high contact probability does not imply small inter loci spatial distance.

Heterogeneity in genome organization implies that given the contact probability, one can no longer determine the mean 3D distance uniquely, which implies that for certain loci the results of Hi-C and FISH must be discordant. Recently, we solved the Hi-C-FISH paradox by calculating the extent of cell population heterogeneity using FISH data and concepts in polymer physics. The distribution of subpopulations could be used to reconstruct the Hi-C data. For a mixed population of cells, the contact probability p_{ij} and the mean spatial distance $\langle \bar{r}_{ij} \rangle$ between two loci m and n, are given by,

$$\langle \bar{r}_{ij} \rangle = \sum_{m}^{S} \eta_{m,ij} \bar{r}_{m,ij} \tag{2}$$

$$\langle p_{ij} \rangle = \sum_{m}^{S} \eta_{m,ij} p_{m,ij} \tag{3}$$

where $\bar{r}_{m,ij}$ and $p_{m,ij}$ are the mean spatial distance and contact probability between i and j in m^{th} subpopulation, respectively. In the above equation, S is total number of distinct subpopulations, and $\eta_{m,ij}$ is the fraction of the subpopulation m, which satisfies the constraint $\sum_m^S \eta_{m,ij} = 1$. Although there exists a one-to-one relation between $p_{m,ij}$ and $\bar{r}_{m,ij}$ in each m^{th} subpopulation, it is not possible to determine $\langle p_{ij} \rangle$ solely from $\langle \bar{r}_{ij} \rangle$ without knowing the values of each $\eta_{m,ij}$ and *vice versa*.

More generally, if we assume that there exists a continuous spectrum of subpopulations, $\langle \bar{r}_{ij} \rangle$ and $\langle p_{ij} \rangle$ can

be expressed as,

$$\langle \bar{r}_{ij} \rangle = \int \mathrm{d}\bar{r}_{ij} K(\bar{r}_{ij}) \bar{r}_{ij} \tag{4}$$

$$\langle P_{ij} \rangle = \int \mathrm{d}p_{ij} Q(p_{ij}) p_{ij} \tag{5}$$

where $\bar{r}_{ij}$ and p_{ij} are the mean spatial distance and the contact probability associated with a single population. $K(\bar{r}_{ij})$ and $Q(p_{ij})$ are the probability density distribution of $\bar{r}_{mn}$ and p_{mn} over subpopulations, respectively.

We have shown [9] that the paradox arises precisely because of the mixing of different subpopulations. The value $\eta_{m,ij}$, $K(\bar{r}_{ij})$ or $Q(p_{ij})$ in Eq. 2-5 in principle could be extracted from distribution of $\langle \bar{r}_{ij} \rangle$, which can be measured using imaging techniques. However, this is usually unavailable or the data are sparse which leads to the question: Despite the lack of knowledge of the composition of cell populations, can we provide an approximate but reasonably accurate relation between $\langle p_{ij} \rangle$ and $\langle \bar{r}_{ij} \rangle$? In other words, rather than answer the question (a) posed in the previous section precisely, as we did for the homogeneous GRMC, we are seeking an approximate solution. The GRMC calculations provide the needed insights to construct the approximate relation to calculate distance matrix from the contact probability matrix.

A key inequality: Let us consider a special case where there are only two distinct discrete subpopulations, and the relation between the $\bar{r}_{ij}(\bar{r})$ and $p_{ij}(p_{ij})$ is given by Eq. 1. According to Eqs. 2-3, we have $\langle \bar{r} \rangle = \eta \bar{r}_1 + (1-\eta)\bar{r}_2 = \eta f^{-1}_{\mathrm{GRMC}}(p_1) + (1-\eta) f^{-1}_{\mathrm{GRMC}}(p_2)$, and $\langle p \rangle = \eta p_1 + (1-\eta) p_2$. Note that f^{-1}_{GRMC} exists since f is a monotonic function. Fig.2a gives a graphical illustration of the inequality $f^{-1}_{\mathrm{GRMC}}(\langle p \rangle) \leq \langle \bar{r} \rangle$. This inequality states that the mean spatial distance of the whole population has a lower bound of $f^{-1}_{\mathrm{GRMC}}(\langle p \rangle)$, which is the mean spatial distance inferred from the measured contact probability $\langle p \rangle$ as if there is only one homogeneous population. This is a powerful result, which is the theoretical basis for constructing the HIPPS method, allowing us to go from Hi-C data to 3D organizations.

The inequality $f^{-1}_{\mathrm{GRMC}}(\langle p \rangle) \leq \langle \bar{r} \rangle$ shows that a theoretical lower bound for $\langle \bar{r}_{ij} \rangle$ exists, given the value of $\langle p_{ij} \rangle$ regardless of the compositions of the whole cell population. In fact, such an inequality can be generalized for arbitrary discrete or continuous distribution of subpopulations. Let us assume that for a homogeneous system, there exists a convex and monotonic decreasing function, ϕ, relating the contact probability p and the mean spatial distance $\bar{r}$, $\bar{r} = \phi(p)$ (we neglect the suffix ij for better readability). Note that ϕ takes the form of Eq. 1 for the GRMC. It can be shown that the following inequality holds (Appendix D),

$$\langle \bar{r} \rangle \geq \phi(\langle p \rangle) \tag{6}$$

The above equation (Eq.6) shows that the lower bound of the mean spatial distance of a heterogeneous population is given by the mean spatial distance computed from the measured contact probability as if the cell population is homogeneous. The equality holds exactly only when the population of cells is precisely homogeneous. This finding is remarkably useful in predicting the approximate spatial organization of chromosomes from Hi-C contact map, as we demonstrate below. For the GRMC, according to Eq. 6, we have $\langle \bar{r}_{ij} \rangle \geq f^{-1}_{\mathrm{GRMC}}(\langle p_{ij} \rangle)$, which is a special case in which only two distinct discrete subpopulations are present. Thus, the precisely solvable model suggests that the approximate power law relating $\langle p_{ij} \rangle$ and $\langle \bar{r}_{ij} \rangle$ could be used as a starting point in constructing the spatial distance matrices using only the Hi-C contact map for chromosomes.

Validation of the lower bound relating $\langle p_{ij} \rangle$ and $\langle \bar{r}_{ij} \rangle$ in heterogeneous cell population: In order to investigate the effect of heterogeneity (contact between i and j for all (i,j) pairs do not exist in all the cells) on the quality of the constructed mean distance matrix $\langle \bar{\mathbf{R}} \rangle$ from the contact probability matrix $\langle \mathbf{P} \rangle$, we simulated a model system with two distinct cell populations. One has all the CTCF mediated loops present (with fraction η), and the other is a polymer chain without any loop constraints (with fraction $1-\eta$). We used the lower bound, $f^{-1}_{\mathrm{GRMC}}(\langle p_{ij} \rangle)$, to infer $\langle \bar{r}_{ij} \rangle$ from $\langle p_{ij} \rangle$. The results, shown in Fig.2b,c,d, provide a numerical verification of the theoretical lower bound linking contact probability and mean spatial distance. Fig.2b shows the scatter plot for $\langle \bar{r}_{ij} \rangle$ versus $\langle p_{ij} \rangle$ from the simulation. The theoretical lower bound, $f^{-1}_{\mathrm{GRMC}}(\langle p_{ij} \rangle)$ is shown in comparison. Fig.2b shows that the lower bound holds with all the points are above it. Using the $f^{-1}_{\mathrm{GRMC}}(\langle p_{ij} \rangle)$, the $\langle \bar{\mathbf{R}} \rangle$ in Fig.2d are calculated from the simulated $\langle \mathbf{P} \rangle$. The comparison between the inferred and the simulated $\langle \bar{\mathbf{R}} \rangle$ (middle ad bottom in Fig.2d) show that the difference between the constructed and simulated DMs is largest near the loops resulting in an underestimate of the spatial distances in the proximity of loops. This occurs because the constructed $\langle \bar{\mathbf{R}} \rangle$ is computed from the simulated $\langle \mathbf{P} \rangle$, which is sensitive to the heterogeneity of the cell population. The difference matrices show that, although the constructed $\langle \bar{\mathbf{R}} \rangle$ underestimated the spatial distances around the loops, most of the pairwise distances are hardly affected. This exercise for the GRMC justifies the use of the lower bound as a practical guide to construct $\langle \bar{\mathbf{R}} \rangle$ from the $\langle \mathbf{P} \rangle$.

To show that the constructed $\langle \bar{\mathbf{R}} \rangle$ using the lower bound gives a good global description of the chromosome organization, we also calculated the often-used quantity $\langle R(s) \rangle$, the mean spatial distance as a function of the genomic distance s, as an indicator of average structure (Fig.2c). The calculated $\langle R(s) \rangle$ differs only negligibly from the simulation results. Notably, the scaling of $\langle R(s) \rangle$ versus s is not significantly changed (inset in

Fig.2c), strongly suggesting that constructing the $\langle\bar{\mathbf{R}}\rangle$ using the lower bound gives a good estimate of the average size of the chromosome segment.

To further assess the quality of the constructed $\langle\bar{\mathbf{R}}\rangle$, we calculated the WLMs for the heterogeneous system with $\eta = (0.1, 0.3, 0.5, 0.7, 0.9, 1.0)$ (see Fig.S1). The results are consistent with the visual comparison of the $\langle\bar{\mathbf{R}}\rangle$; the calculated $\langle\bar{\mathbf{R}}\rangle$ for large η agree significantly better with the simulations compared to small values of η. This is also reflected in the distance correlation [27] between the reconstructed and simulated WLMs (blue curve in Fig.S1b), increasing from ≈ 0.8 to ≈ 1.0 from $\eta < 0.7$ to $\eta > 0.7$. In contrast, the distance correlation coefficients between the reconstructed and simulated $\langle\bar{\mathbf{R}}\rangle$ (red curve in Fig.S1b) stays around 0.95 for all values of η, which would not allow us to distinguish between different models.

It is worth noting that even for small values of η, the distance correlation coefficient is 0.8, which is a high value. This is consistent with the result shown in Fig.2c that the constructed $\langle\bar{\mathbf{R}}\rangle$ gives a rough but reasonable global estimation of the structural organization even though it may deviate from the exact result in details. Taken together these results show that the reconstructed $\langle\bar{\mathbf{R}}\rangle$ provides a fairly accurate description of the conformations in spite of the presence of heterogeneity in the conformations.

The distance correlation gives a global description of the similarity between the simulated and inferred DMs. To further investigate the degree of similarity at different length scales, we computed the Adjusted Mutual Information (AMI) scores between the simulated and constructed clustering result from WLM by varying the number of clusters (Fig.S1). A small number of clusters corresponds to the large scale hierarchical organization whereas a higher number of clusters reveals the structure on the small length scale. For $\eta \leq 0.7$, AMI scores are low (Fig.S1) for the small number of clusters and increases upon increasing the number of clusters up to around 0.8. For $\eta > 0.7$, the AMI scores remain around 0.9 throughout the range of the number of clusters.

Inferring 3D organization of interphase chromosomes from experimental Hi-C contact map: To apply the insights from the results from GRMC to obtain the 3D organization of chromosomes, we conjecture that a power law relation, first suggested using imaging experiments [7] and subsequently established by us [25], relating the contact probability between two genomic loci $\langle p_{ij}\rangle$ and $\langle\bar{r}_{ij}\rangle$ holds generally for chromatins. Thus, we write,

$$\langle\bar{r}_{ij}\rangle = \Lambda\langle p_{ij}\rangle^{-1/\alpha} \tag{7}$$

where α and Λ are unknown coefficients. Again, note that the $\langle\cdot\rangle$ and $\bar{\cdot}$ represent the average over subpopulations and the average over individual conformations in a single subpopulation, respectively. In a homogeneous system, the equalities $\langle\bar{r}\rangle = \bar{r}$ and $\langle p\rangle = p$ hold. For the GRMC, $\Lambda = r_c$ and $\alpha = 3.0$. From the ensemble Hi-C experiments, $\langle p_{ij}\rangle$ can be inferred. For a self-avoiding polymer, $\alpha \approx 3.71$ for two interior loci that are in contact (see Appendix E). Based on experiments [7] and simulations using the Chromosome Copolymer Model [25] a tentative suggestion could be made for a numerical value for $\alpha \approx 4.0$. Given the paucity of data needed to determine α we follow the experimental lead [7] and set it to 4.0, which is an unusually large value not associated with any known polymer model. We show below that the power-law relation given in Eq.7 provides a way to infer the approximate 3D organization of chromosomes from the experimental Hi-C contact map.

Experimental Validation on Eq7 and choice of α: To further show that Eq.7 with $\alpha = 4$ is accurate, we calculated the square of the radius of gyration of all the 23 chromosomes using $R_g^2 = (1/2N_c^2)\sum_{i,j}\langle\bar{r}_{ij}\rangle^2$. The dashed line in Fig.3a is a fit of R_g^2 as a function of chromosome size, which yields $R_g \sim N_c^{0.27}$ where N_c is the length of the chromosome. For a collapsed polymer, $R_g \sim N_c^{1/3}$ and for an ideal polymer to be $R_g \sim N_c^{1/2}$. To ascertain if the unusual value of 0.27 is reasonable, we computed the volume of each chromosome using $(4/3)\pi R_g^3$ and compared the results with experimental data [28]. The scaling of chromosome volumes versus N_c of the predicted 3D chromosome structures using HIPPS is also in excellent agreement with the experimental data (Fig.3b). The exponent $0.27 \lesssim 1/3$ suggests the chromosomes adopt highly compact, space-filling structure, which is also vividly illustrated in Fig.4.

Since the value of Λ (Eq.7) is unknown, we estimate it by minimizing the error between the calculated chromosome volumes and experimental measurements. We find that $\Lambda = 117$ nm, which is the approximate size of a locus of 100 kbps (the resolution of the Hi-C map used in the analysis). It is noteworthy that the genome density computed using the value of $\Lambda = (100 \cdot 10^3/(4/3)\pi\Lambda^3)\text{bps} \cdot \text{nm}^{-3} = 0.015\text{bps} \cdot \text{nm}^{-3}$ is consistent with the typical average genome density of Human cell nucleus $0.012\text{bps} \cdot \text{nm}^{-3}$ [29]. The value of Λ does not change the scaling but only the absolute size of chromosome.

Generating ensembles of 3D structures using the maximum entropy principle: The great variability in the genome organization have been noted before [9, 23, 30]. To investigate the structural heterogeneity of the chromosomes, we ask the question: how to generate an ensemble of structures consistent with the mean pairwise spatial distances between the loci? More precisely, what is the joint distribution of the position of the loci, $P(\{\boldsymbol{x}_i\})$, subject to the constraint that the mean pairwise distance is $\langle||\boldsymbol{x}_i - \boldsymbol{x}_j||\rangle = \langle\bar{r}_{ij}\rangle$? Generally, there

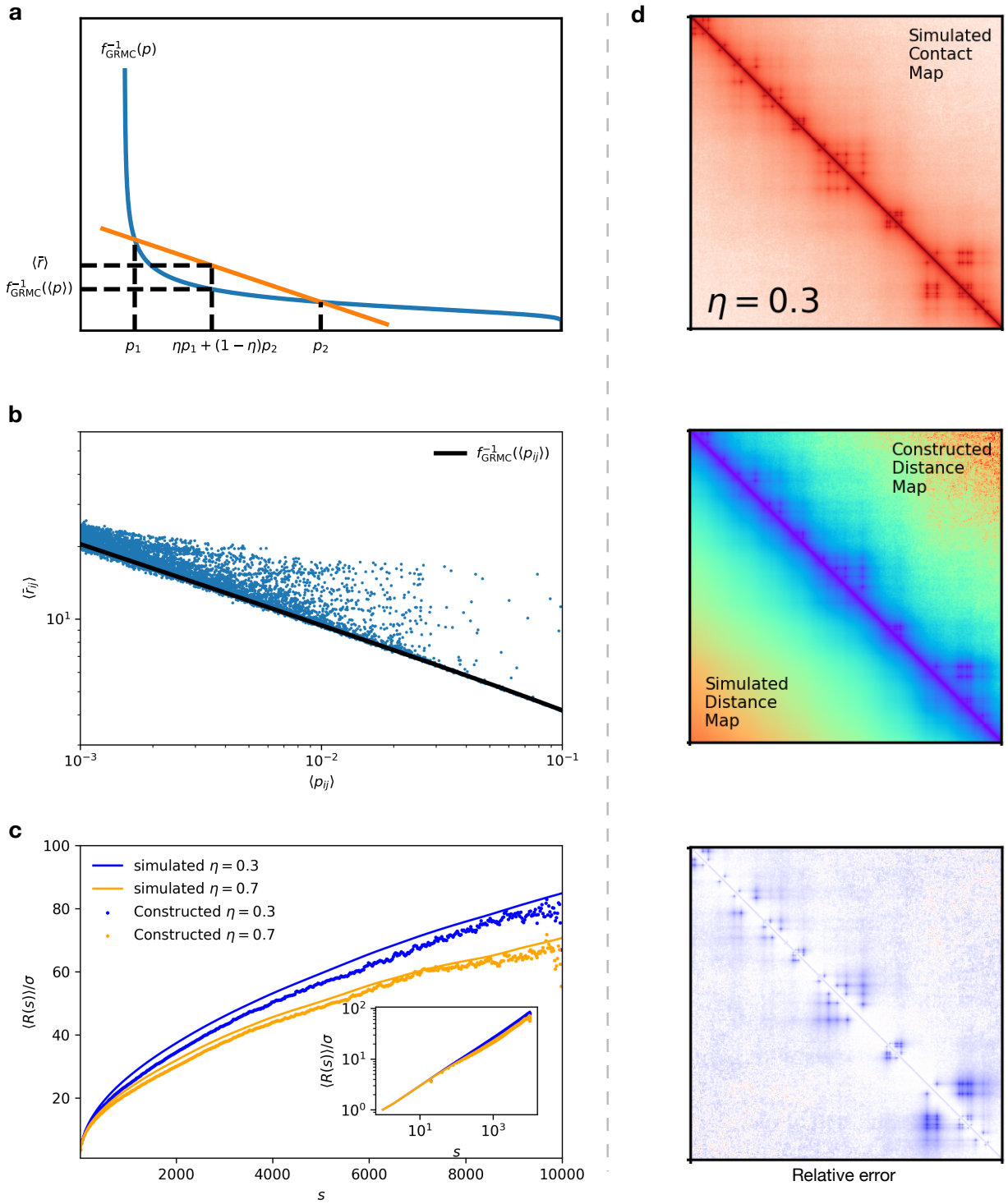

FIG. 2. (Caption next page.)

exists an infinite number of $P(\{\boldsymbol{x}_i\})$, satisfying the constraint of mean pair-wise spatial distances. By adopting the principle of maximum entropy, we seek to find the $P^{\mathrm{MaxEnt}}(\{\boldsymbol{x}_i\})$ with the maximum entropy among all possible $P(\{\boldsymbol{x}_i\})$. The maximum entropy principle has been previously used in the context of genome organization [31, 32] for different purposes. We note parenthetically that the preserving the constraints of mean pairwise distances is equivalent to preserving the constraints of mean squared pairwise distances. In practice,

FIG. 2. **(a)** Lower Bound for mean spatial distance $\langle \bar{r} \rangle$ illustrated graphically. The blue curve is the function $f^{-1}_{\rm GRMC}$ which exists since $f_{\rm GRMC}$ is a monotonic function. The orange line is the secant line between points $(p_1, f^{-1}_{\rm GRMC}(p_1))$ and $(p_2, f^{-1}_{\rm GRMC}(p_2))$. All the points between p_1 and p_2 on x-axis can be expressed as $\eta p_1 + (1-\eta)p_2 \equiv \langle p \rangle$ for some value of $\eta \in [0,1]$. The y-axis value corresponds to $\langle p \rangle$ is $\eta f^{-1}_{\rm GRMC}(p_1) + (1-\eta) f^{-1}_{\rm GRMC}(p_2) \equiv \langle \bar{r} \rangle$ and $f^{-1}_{\rm GRMC}(\langle p \rangle)$ for the orange line and blue curve, respectively. Notice that for any values of p_1, p_2 and η, the orange line is always above the blue curve, which proves the inequality $f^{-1}_{\rm GRMC}(\langle p \rangle) \leq \langle \bar{r} \rangle$. From the graph, it can also be noted the equality holds only when $p_1 = p_2$, which is to say the cell population is homogeneous. **(b)** Scatter plot for mean pair-wise spatial distances versus the contact probability for $\eta = 0.3$. Solid black line is the theoretical lower bound, given by the solution $f^{-1}_{\rm GRMC}(\langle p_{ij} \rangle)$. **(c)** Plots of $\langle R(s) \rangle$ as a function of the genomic distance, s, for $\eta = 0.3$ and 0.7. The inset shows the same data on a log-log scale; $\langle R(s) \rangle$ is calculated using $\langle R(s) \rangle = (1/TM) \sum_{a=1}^{M} \sum_{t=1}^{T} \big(r_{ij}^{(a)}(t) \delta(s - |i-j|)/(N-s) \big)$. The theoretical predictions are in remarkable agreement with simulations. **(d)** Simulated CM (top), simulated DM and inferred DM side by side (middle), and relative error map (bottom) for $\eta = 0.3$ for GRMC. Note that all the maps are block averaged from size 10000 to size 400 as explained in the Appendix C. The inferred DM is obtained using $\langle \bar{r}_{ij} \rangle = f^{-1}_{\rm GRMC}(\langle p_{ij} \rangle)$. Relative error map is shown with blue color indicating larger error. It is clear that the spatial distances are underestimated at the loops

we found that using the constraints of squared distances, $\langle ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2 \rangle = \langle \bar{r}_{ij}^2 \rangle$, yields better convergence. Recall that the $P^{\rm MaxEnt}(\{\boldsymbol{x}_i\})$ with respect to the constraints of the mean squared pairwise spatial distances is,

$$P^{\rm MaxEnt}(\{\boldsymbol{x}_i\}) = \frac{1}{Z} \exp\big(-\sum_{i<j} k_{ij} ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2\big) \qquad (8)$$

where k_{ij} are the Lagrange multipliers that are chosen so that the average values $\langle ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2 \rangle$ matches $\langle r_{ij}^2 \rangle$, which could be either inferred from the Hi-C contact map or directly measured in FISH experiments; Z is the normalization factor. The merit of the maximum entropy distribution (Eq.8) is that it is both data-driven and physically meaningful since the parameters k_{ij} are inferred from experimental data and the term $k_{ij}||\boldsymbol{x}_i - \boldsymbol{x}_j||^2$ can be viewed as pair-wise potential energy between the loci. Indeed, Eq. 8 is exactly the same as the generalized Rouse model [8] where k_{ij} are the spring constants between genomic loci.

The procedure used to generate an ensemble of 3D chromosome structures is the following: First, we compute the mean spatial distance matrix from contact map using Eq. 7 with α set to 4.0. The value of the scaling factor $\Lambda = 117$nm, calculated using additional experimental constraints (see the previous section). Recall that Λ only sets the length scale but has no effect on the conformational ensemble of the chromosome. Using the iterative scaling algorithm, we obtain the values of k_{ij} (Appendix G). Once the values of k_{ij} are obtained, $P^{\rm MaxEnt}$ can be directly sampled as a multivariate normal distribution, thus generating an ensemble of chromosome structures. Fig.5a shows the comparison between the inferred DM and the DM for Chromosome 1 obtained using the maximum entropy principle. It is visually clear that the two DMs are in excellent agreement (see Fig.S2-S7 for the other chromosomes). We should emphasize that the maximum entropy method described here, in principle, can achieve exact match with the inferred DM. The small discrepancies are due to 1) the quality of convergence and 2) the intrinsic error in the Hi-C map and the inferred DM derived from it.

Characteristics of 3D chromosome structures: The 3D conformations are specified by $\boldsymbol{x}_i, i = 1, 2, 3, \cdots, N_c$ where N_c is the number of loci at a given resolution (the centromeres are discarded due to lack to information about them in the Hi-C contact map). The values of N_c for all the 23 chromosomes are given in Table.S1. We generated an ensemble of 1,000 structures for each of the 23 Human interphase chromosomes using the procedure described above. Fig.4a shows the typical conformations of averaged value of radius of gyration for each chromosome. Visually it is clear that there is considerable shape heterogeneity among the chromosomes. To quantify the shape of chromosomes, we obtain the distribution of relative shape anisotropy κ^2 (Appendix H). Fig.4b shows the violin plots of κ^2 for all the 23 chromosomes, ordered by value of $\langle \kappa^2 \rangle$. The chromosomes exhibit considerable variations in κ^2. Chromosome 13 is most spherical and chromosome 19, 9 and 21 have the most elongated shape.

We can draw important conclusions from the calculated 3D structural ensemble with some biological implications that we mention briefly.

Compartments and microphase separation: The probabilistic representations for Chromosome 1 are shown in Fig.5b,c,d where we align all the conformations and superimpose them. First, we note that such a probabilistic representation demonstrates clear hierarchical folding of chromosomes where the loci with small genomic distance (similar color) are also close in space (Fig.5b, see Fig.11 for the other chromosomes). Long-range mixing between the loci is avoided, supporting the notion of crumpled globule [33–35]. Furthermore, the reconstructed structure of the chromosomes shows clear microphase separation (different colors are segregated. These are referred to as A and B compartments (Fig.5c, see Fig.12 for the other chromosomes), representing two epigenetic states (euchromatin and heterochromatin), which we obtained using the spectral clustering [25]. Each compartment predominantly contains loci belonging to either euchromatin or heterochromatin. Contacts within each compartment are enriched between either euchromatin or heterochromatin epigenetic states. In the Hi-C data the compart-

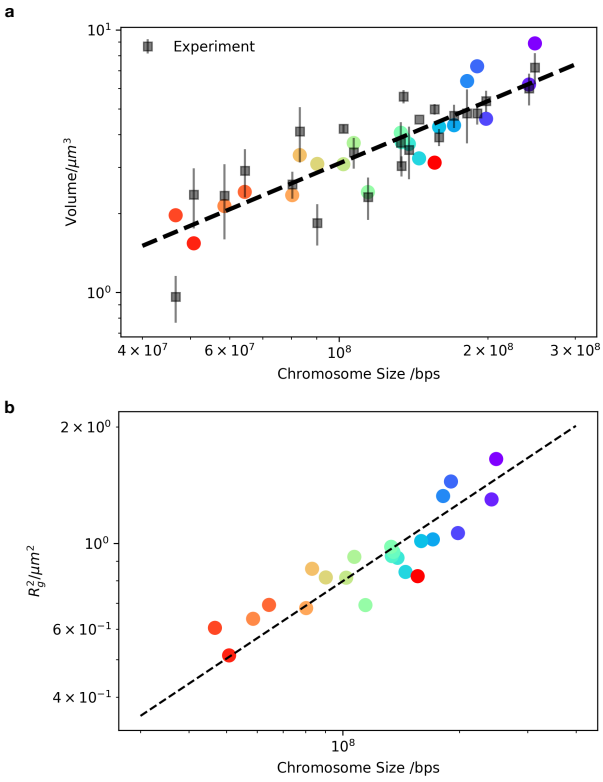

FIG. 3. **(a)** Plot of the square of the radius of gyration R_g^2 as a function of the chromosome size. The dashed line is a fit to the data with the slope 0.54 which implies that $R_g \sim N^{0.27}$. The data are for the 23 chromosomes. **(b)** Volume of each chromosome versus the length in units of base pairs. The experimental values (black squares) are computed using the data in [28]. The dashed line is the fit to the experimental data with slope 0.8. Volume of each chromosome is calculated using $\lambda V_{\rm nuc}$ where λ is the percentage of volume of the nucleus volume $V_{\rm nuc}$. The values of λ are provided in Fig.S5 in [28], and $V_{\rm nuc} = (4/3)\pi r_{\rm nuc}^3$ where $r_{\rm nuc} = 3.5\mu m$ is the radius of Human lymphocyte cell nucleus [28]. Volumes of the reconstructed Chromosome using theory and computation are calculated using $(4/3)\pi R_g^3$ (color circles). The predicted values, without any adjustable parameters, and the experimental values have a Pearson correlation coefficient of 0.79. The good agreement further validates the procedure used to construct the ensemble of 3D genome structures.

ments appear as a prominent checker board pattern in the contact maps. Fig.5c shows that the two compartments are spatially separated and organized in a polarized fashion, which is fully consistent with multiplexed FISH and single-cell Hi-C data[30].

Mapping ATAC-seq to 3D structures: Advances in sequencing technology have been used to infer epigenetic information in chromatin without the benefit of integrating with structures. In particular, the assay for transposase accessible chromatin using sequencing (ATAC-Seq) technique provides chromatin accessibility, which in turn provides insights into gene regulation and other functions. The results obtained using ATAC-seq (see Appendix I for details on processing of ATAC-seq data), also shows microphase separation pattern between high ATAC signal and low ATAC signal region (Fig.5d). With the structures determined by HIPPS in hand, we mapped the ATAC-Seq data onto ensemble of conformations for Chromosome 1 from GM 12878 cell in Fig.5d. It appears that accessibilities in chromosome 1 for various functions (such as nucleosome positioning and transcription factor binding regions) may be spatially segregated. Such segregation between high ATAC signal loci and low ATAC signal loci are also visually clear in other chromosomes (Fig.13). Remarkably, these results, derived from the HIPPS method, follow directly from the Hi-C data *without* creating a polymer model with parameters that are fit to the experimental data.

Structural Heterogeneity: To investigate the heterogeneity in chromosome conformations, we examined the variations among the 1,000 conformations generated for chromosome 5. First, as a global structural characteristic, we computed the radius of gyration of individual structure. R_g. Fig.6a shows the histogram, $P(R_g)$, and conformations with compact, intermediate and expanded conformations as examples. We then wondered what is the degree of variations in the organization of the A/B compartments? Specifically, we want to know whether A/B compartments are spatially separated in a single-cell. To answer this question, we first define a quantitative measure of the degree of mixing between A/B compartments, Q_k,

$$Q_k = \frac{1}{N_c}\sum_i \frac{|n_A(i;k)/\hat{n}_A - n_B(i;k)/\hat{n}_B|}{k} \qquad (9)$$

where k is the number of nearest neighbor of loci i. In Eq. 9 $n_A(i;k)$ and $n_B(i;k)$ are the number of neighbor loci belonging to A compartment and B compartment for loci i out of k nearest neighbor, respectively $(n_A(i;k) + n_B(i;k) = k)$. With $N_c = (N_A + N_B)$, the fraction of loci in the A compartment is $\hat{n}_A = N_A/N_c$ and $\hat{n}_B = N_B/N_c$ is the fraction in the B compartment where N_A, N_B are the number of A and B loci, respectively. The k nearest neighbors of i are computed as follows. First, the distance from i to all loci are calculated. From these distances, the k smallest values are chosen, and this process is repeated for all i. Note that Q_k is length-scale invariant because it is a function of the number of nearest neighbors, which allows us to compare the structures with different values of R_g. Note that $Q_k = 2$ or 0 for perfect demixing and mixing between A and B compartments, respectively. Fig.6b shows Q_k and $P(Q_k)$ histograms for different values of k. The distribution is clearly skewed toward large values, indicating demixing of the A and B compartments on the population level. At the same time, the distribution shows that there exists a small fraction of single cell chromosomes conformations, which have Q_k values close to 0.8, implying that the compartment organization of

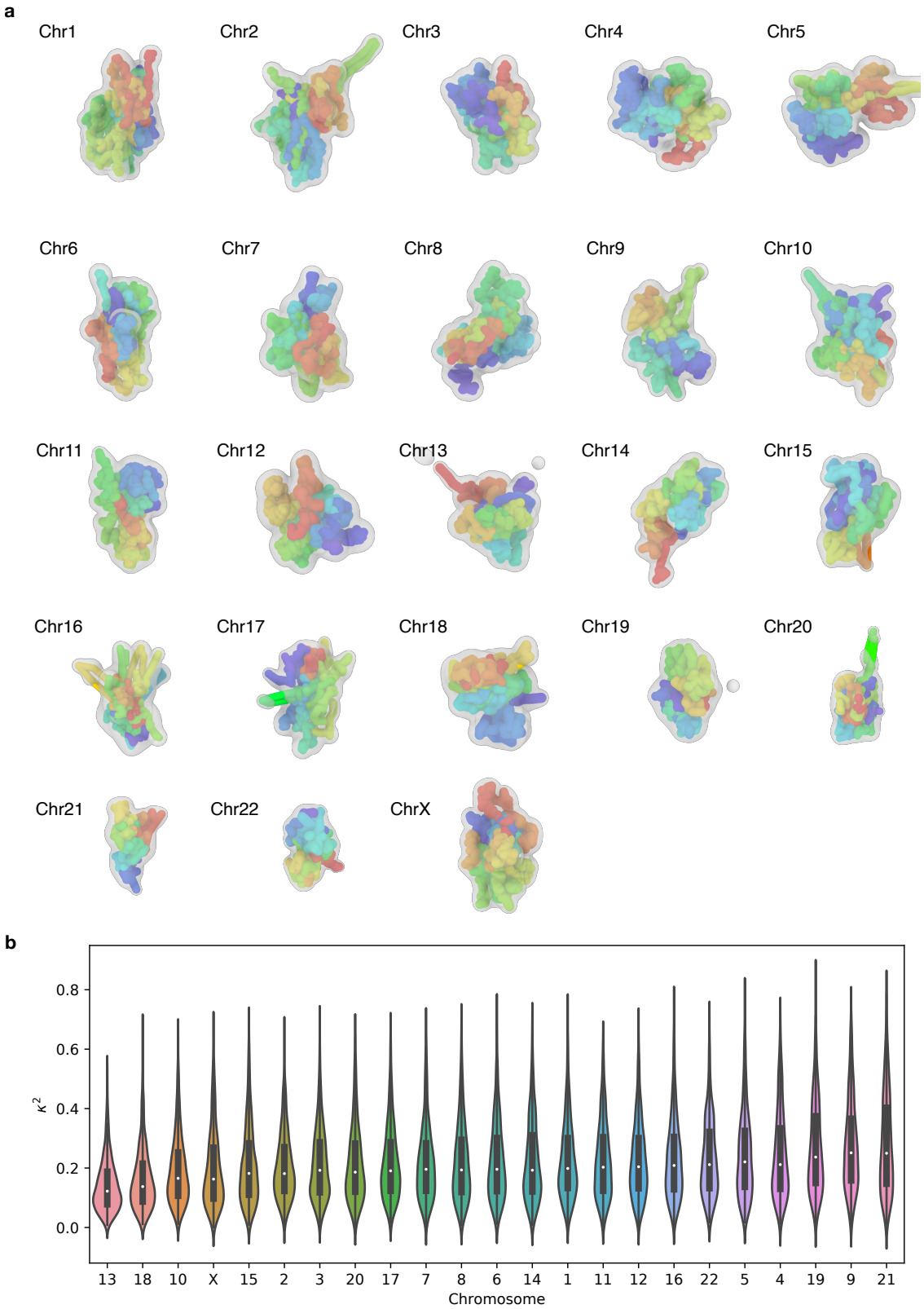

FIG. 4. **(a)** Representative 3D reconstructed structures for all the 23 Human interphase chromosomes using inferred DM, which is calculated using Eq.7 with $\Lambda = 117$ nm and $\alpha = 4.0$. The colors encode the genomic position of the loci. The resolution of loci is 100 kbps. Red and purple represent 5' and 3' ends, respectively. The structures whose radius of gyration closet to the population average value are selected. The structures are rendered using Ovito with bond radius of $\Lambda = 117$nm. **(b)** Violin plot for the relative shape anisotropy κ^2 (Appendix H) for all the 23 chromosomes. The chromosomes are ordered with increasing of $\langle\kappa^2\rangle$. Chromosome 13 and the Chromosome 21 have the most and the least spherical shape, respectively.

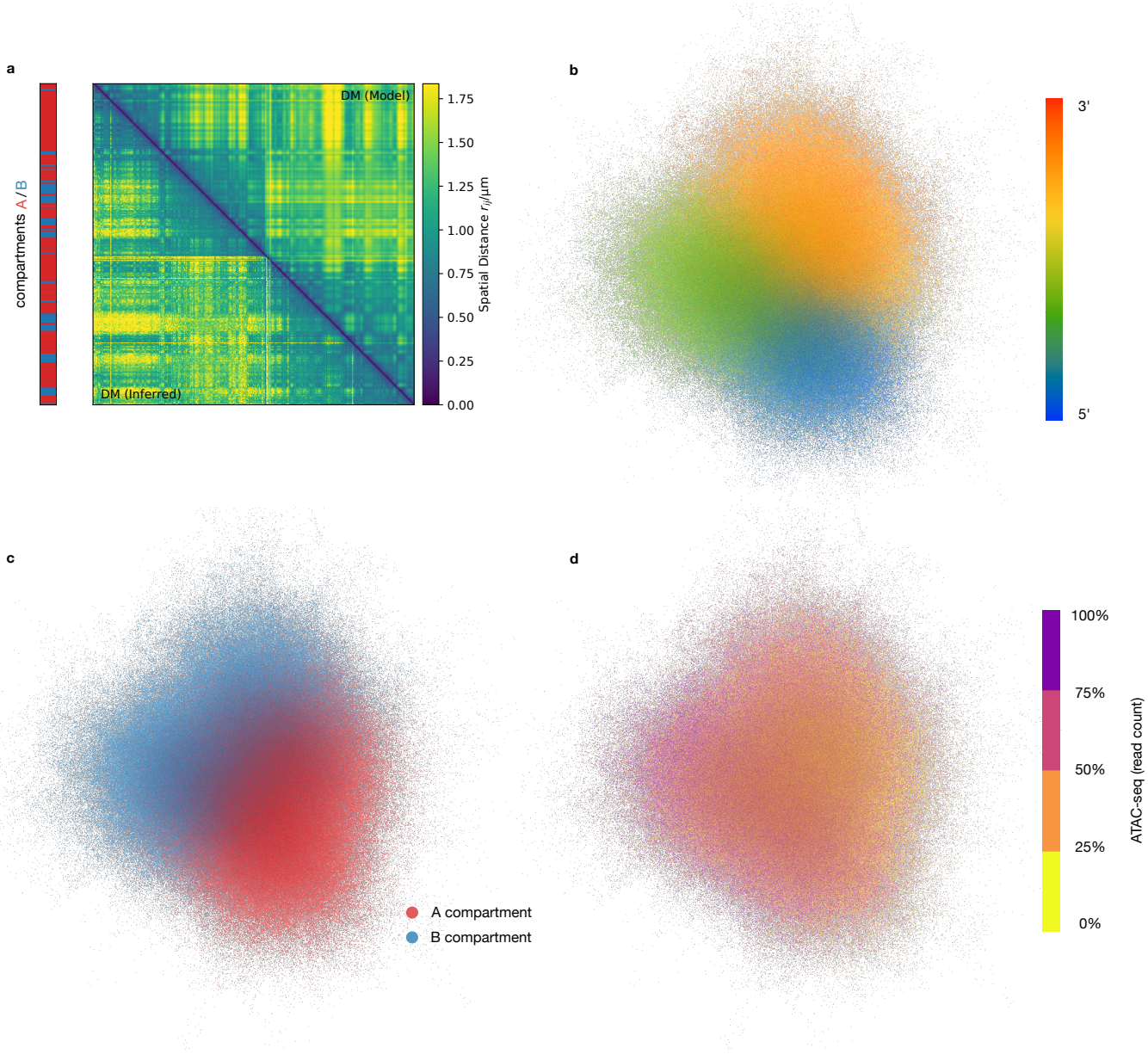

FIG. 5. **(a)** Comparison between the DM inferred from Hi-C data (lower triangle) and the DM calculated from an ensemble of 3D structures for Chr1 using the HIPPS method, $\langle \mathbf{P} \rangle \rightarrow \langle \bar{\mathbf{R}} \rangle \rightarrow$ Ensemble of 3D structures (upper triangle). A/B compartments, determined using spectral biclustering are shown as well [25]. **(b)** Superpositions of an ensemble of 3D structures for Chr1. A total number of 1,000 conformations are aligned and superimposed. Each point represent one loci from one conformation. The cloud representation demonstrates the probabilistic picture of chromosome conformation, with color representing the genomic location of the loci along the genome. It is clear that the loci close along the genome are preferentially located in 3D proximity as well, consistent with the notion of crumpled/fractal globule. The resolution of the loci is 100 kbps. It is worth emphasizing that the structures are entirely determined starting from a contact map, without invoking any energy function. **(c)** Same cloud point representations as (b) with color indicating the A/B compartments. Phase separation between A/B compartments is vividly illustrated. **(d)** Same cloud representation as **(b)** and **(c)** but with ATAC-seq read counts as color coding. The ATAC-seq read counts are obtained and processed (Appendix I) from the data taken from [36] under GEO accession number GSE47753. Then it is binned into four quantiles. It can be observed that the loci with high ATAC signal and low ATAC signal are spatially segregated. For majority of the 23 chromosomes, the spatial pattern of ATAC-seq is consistent with A/B compartments (Fig.13)

chromosome exhibits a degree of heterogeneity.

Chromosome organizations in different cell types: Since single chromosome conformations in a single cell exhibit extensive variations, it is natural to wonder how structurally heterogeneous a given chromosome is in dif-

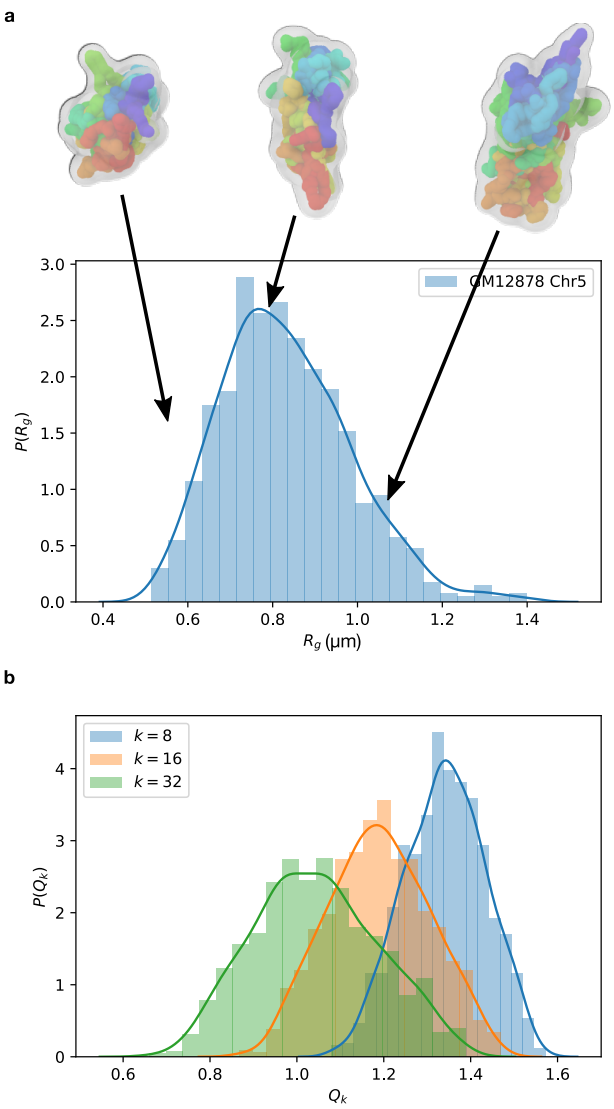

FIG. 6. **(a)** Distribution of the radius of gyration, $P(R_g)$, of Chromosome 5 from GM12878 cell type. Three structures whose value of R_g is in 0.15 quantile, 0.5 quantile and 0.75 quantile, respectively are shown. **(b)** Distribution of the degree of mixing between A/B compartments, $P(Q_k)$ (Eq.9), for Chromosome 5. $Q_k = 2$ for perfect demixing and $Q_k = 0$ for perfect mixing between A/B compartments.

ferent cells types and if the HIPPS method can quantify these differences at the single-cell level? We are searching for differences in the heterogeneity of a specific chromosome in different cell types. From a physical viewpoint this is difficult to answer this question precisely because structural heterogeneity of a chromosome in a given cell type could overwhelm the analysis. Furthermore, one has to contend with high-dimensional data (each conformation has 3N coordinates) in the ensemble of conformations.

In order to delineate the differences in the heterogeneities in the conformations of a specific chromosome in different cell types we used a machine learning method for large data analysis [37]. To compare two single chromosome conformations, we first normalized the distance matrix such that $\sum_{i,j} r_{ij}^2 = 1$. By doing so, we eliminate the effect of overall size of the individual chromosome conformation, thus allowing us to compare them in terms of only their conformations. We generated a total number of 1,000 structures for chromosome 21 from 7 cell types using Hi-C data [6]. Fig.7a shows the tSNE (t-Distributed Stochastic Neighbor Embedding) plot [37] for 7,000 individual chromosome conformations from 7 different cell types (1,000 conformations for each cell type). It is clear that the structural ensembles of chromosome 21 from different cell types have different degrees of overlap with each other. IMR-90 (fibroblast), HUVEC (umbilical vein endothelium), and GM12878 (lymphoblastoid), which are normal human cells, form compact, distinct clusters with negligible overlap with each other. In Fig.7a the conformations of chromosome 21 in the 2D tSNE representation are shown as blue (IMR-90), red (HUVEC), and green (GM12878) dots. In sharp contrast, the conformations of the same chromosome in HMEC (breast epithelial cell), K562 (myeloid leukemia cell in bone marrow), NHEK (epidermal keratinocytes - type of skin cell), and KBM7 (a different leukemia cell) cells display very large variations. They are not as compact and their phase space structure in terms of the low dimensional tSNE coordinates show overlapping regions (Fig.7a).

To further investigate the characteristics of chromosome organization in different cell types, we computed the values of $F(k)$, which quantifies the multi-body long-range interactions of the chromosome structure. We define $F(k)$ as,

$$F(k) = \frac{1}{kN_cF_0(k)} \sum_i \sum_{j \in m_i(k)} |j - i| \qquad (10)$$

where k is the number of nearest neighbors, and $m_i(k)$ is the set of loci that are k nearest neighbors of loci i; $F_0(k) = (1/2)(1 + k/2)$ is the value of $F(k)$ for a straight chain. From Eq.10, it follows that the presence long-range interaction increases the value of $F(k)$. It is worth noting that $F(k)$ can also be viewed as a measure of how well the linear relation along genome is preserved in the 3D structure. Fig.7b show the distributions of $F(k)$ for each cell type. GM12878 cell shows the most enrichment of long-range multi-body clusters whereas NHEK and HMEC cells show the least. However, there is extensive overlap between different cell types for $F(k)$. Remarkably, we find that there are substantial variations in the structural ensembles of chromosome 21, and by implication others as well, not only within a single cell but also among single cells belonging to different tissues. From our perspective, it is most interesting that the HIPPS when combined with machine learning techniques can quantitatively predict the differences.

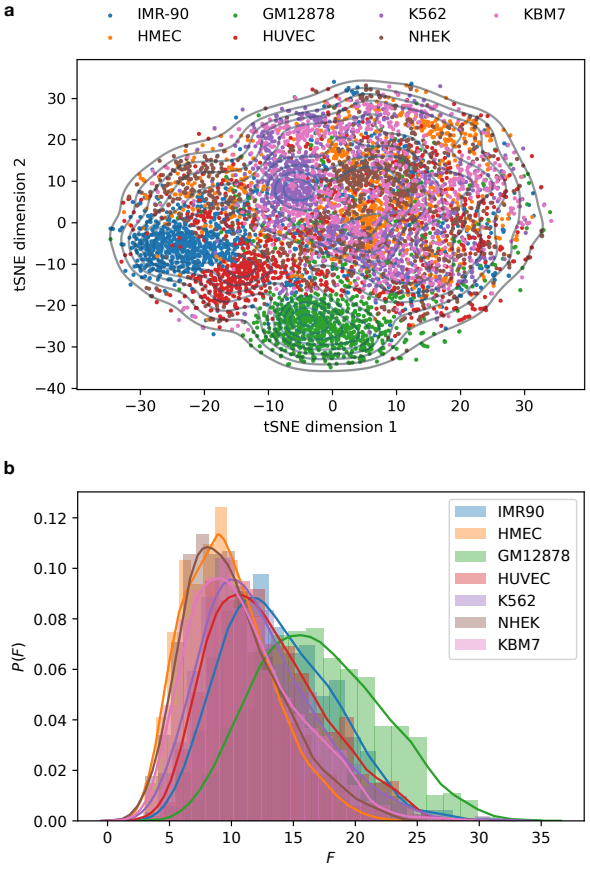

FIG. 7. **(a)** tSNE plot for the ensemble of chromosomes 21 structures for 7 cell types (IMR-90, HMEC, GM12878, HUVEC, K562, NHEK, KBM7). Each cell type has a total number of 1,000 independent conformations. Each conformation is represented by its distance matrix. The metric used to compare two single chromosomes is the squared Euclidean distance between distance matrices. **(b)** The distribution of $F(k)$ (Eq.10) for different cell types. We take $k = 8$, corresponding to 8 nearest neighbors.

DISCUSSION AND CONCLUSION

Using theory, based in polymer physics and the principle of maximum entropy, and precise numerical simulations of a non-trivial model, we have provided an approximate solution to the problem of how to construct an ensemble of the three-dimensional coordinates of each locus from the measured probabilities ($\langle p_{ij} \rangle$) that two loci are in contact. The key finding that makes our theory possible is that $\langle p_{ij} \rangle$ is related to $\langle \bar{r}_{ij} \rangle$ through a power law, which is in accord with experiments [7] as well as accurate polymer models for interphase chromosomes [25]. The inferred mean spatial distances are then used to obtain an ensemble of structures using the maximum entropy principle. Our approach, which is both physically motivated and data-driven procedure, is self-consistently accurate for the precisely solvable GRMC. The physically well-tested theory, leading to the HIPPS method, allowed us to go take the Hi-C contact map and create an ensemble of three-dimensional chromosome structures without any underlying model. Using the HIPPS method we constructed the 3D organization of the twenty-three human chromosomes solely from the Hi-C contact maps. We believe that our theory, with sparse data from Hi-C and FISH experiments, may be combined to produce the 3D structures of chromosomes for any species.

The limitation of many population-based experimental approaches for producing the 3D organization is their inability to extract the single-cell information. Due to the apparent heterogeneity in the cell population [9, 20], Hi-C map, as an ensemble average quantity, does not contain the information about the fluctuations of the organization of genomes. The Hi-C map and the derived $\langle \bar{\mathbf{R}} \rangle$ only characterize the *averaged* structure. In other words, there may not exist a typical single-cell genome that can be described by the Hi-C map, and hence the $\langle \bar{\mathbf{R}} \rangle$ derived from it. Using the maximum entropy principle, we are able to generate an ensemble of structures from $\langle \bar{\mathbf{R}} \rangle$, consistent with observation from imaging data. It is worth noting that our use of the maximum entropy principle with pairwise distances as constraints, leads to a joint distribution of loci coordinates without assuming a predefined energy function.

The HIPPS method may also suffer from the same problem because $\langle \bar{\mathbf{R}} \rangle$ is inferred from ensemble averaged Hi-C map. Thus, we suggest that the actual single-cell experimental measurements are fundamentally crucial to decipher the single-cell genome organization. This can also be reasoned from the following arguments using our simple mixture model system as an example. Every trajectory can be described by either a chain containing all the loops or a chain that is devoid of loops. Therefore, averaging over an ensemble of cells may not be meaningful from an *in vivo* perspective. Using the maximum entropy principle described in this work, a single mode widespread distribution can be obtained instead of a bimodal distribution which characterize two distinct subpopulations. This problem can be overcome by using distribution instead of mean as constraints under the maximum entropy principle. However, such distributions should only be obtained from single-cell measurements. Nevertheless, the theoretical lower bound that we have derived provides a way forward to obtain 3D organization from contact map alone, perhaps even from single-cell Hi-C data.

The HIPPS method could be improved in at least two ways. First, the theory relies on Eq.7, which relates the average contact probability between two loci to the mean distance between them. Even though choosing $\alpha = 4.0$ in Eq.7 provides a reasonable description of the sizes of all the chromosomes it should be treated as a tentative estimate. More precise data accompanied by an analytically solvable polymer model containing consecutive loops, as is prevalent in the chromosomes, could produce more accurate structures. Second, as the resolution of Hi-C map improves the size of the contact matrix will not only increase but the matrix would be increasingly

sparse because of the intrinsic heterogeneity of the chromosome organization. Thus, methods for dealing with sparse matrices will have to be utilized in the HIPPS method for extracting chromosome structures.

We should emphasize that if the chromosome structures are used in conjunction with an underlying model with energy functions that produce the patterns in ensemble averaged Hi-C data then the HIPPS method could be used to predict single cell structures, which would shed light on the heterogeneous organization of chromosomes. Ultimately, this might well be the single most important utility of our theory.

APPENDIX A: SIMULATION DETAILS

The GRMC is a variant of a model introduced previously [8] as a caricature of physical gels. Recently, we used the GRMC [9] as the basis to characterize the massive heterogeneity in chromosome organization. The energy function for the GRMC is [9],

$$U(\boldsymbol{r}_1, ..., \boldsymbol{r}_N) = \sum_{i=1}^{N-1} U_i^S + \sum_{\{p,q\}} U_{\{p,q\}}^L. \quad (11)$$

For the bonded stretch potential, U_i^S, we use,

$$U_i^S = \frac{\kappa}{2}(|\boldsymbol{r}_{i+1} - \boldsymbol{r}_i| - a)^2, \quad (12)$$

where a is the equilibrium bond length. The interaction between the loop anchors is modeled using,

$$U_{\{p,q\}}^L = \frac{\omega}{2}(|\boldsymbol{r}_p - \boldsymbol{r}_q| - a)^2 \quad (13)$$

where the spring constant may be associated with the CTCF facilitated loops. The labels $\{p, q\}$ represent the indices of the loop anchors, which are taken from the Hi-C data [6].

The energy function for the ideal Rouse chain simulated in this work is,

$$U(\boldsymbol{r}_1, ..., \boldsymbol{r}_N) = \sum_{i=1}^{N-1} U_i^S, \quad (14)$$

which is obtained from the energy function for GRMC by eliminating the loop constraints (setting $\omega = 0$ in Eq.13).

In order to accelerate conformational sampling, we performed Langevin Dynamics simulations at low friction [38]. The total number, N, of monomers is $10,000$. We simulated each trajectory for 10^8 time steps, and saved the snapshots every $10,000$ time steps. We generated ten independent trajectories, which are sufficient to obtain reliable statistics, which we illustrate in Fig.S8.

APPENDIX B: DATA ANALYSES OF THE SIMULATION DATA

The contact probability between the m^{th} and n^{th} loci in the simulation is calculated using,

$$P_{mn} = \frac{1}{TM} \sum_{a=1}^{M} \sum_{t=1}^{T} \Theta(r_c - |\boldsymbol{r}_m^{(a)}(t) - \boldsymbol{r}_n^{(a)}(t)|), \quad (15)$$

where $\Theta(\cdot)$ is the Heaviside step function, r_c is the threshold distance for determining the contacts, the summation is over the snapshots along the trajectory, and M is the total number of independent trajectories, and T is the number of snapshots for a single trajectory. The mean spatial distance between the i^{th} and the j^{th} loci in the simulations is calculated using,

$$\langle R_{mn} \rangle = \frac{1}{TM} \sum_{a=1}^{M} \sum_{t=1}^{T} |\boldsymbol{r}_m^{(a)}(t) - \boldsymbol{r}_n^{(a)}(t)|. \quad (16)$$

The objective is to calculate $\langle R_{mn} \rangle$ from P_{mn} , and to determine, if in so doing, we get reasonably accurate results. Because these quantities can be computed precisely for the GRMC, the $[P_{mn}, \langle R_{mn} \rangle]$ relationship can be rigorously tested.

APPENDIX C: BLOCK AVERAGE

Fig.8 shows the procedure used for the block average when dealing with several vanishing (or very small) contact probabilities P_{mn}s. Such a method could be used for (almost) any sparse matrix. Let the original contact matrix (CM) have size $N \times N$. By setting a coarse-grained level n, the original CM is divided into blocks, each with size n. The new coarse-grained CM is constructed in the way the values of elements in the $(N/n) \times (N/n)$ are the arithmetic average of elements in each block. We then demonstrate that this coarse-graining procedure does not alter the structural information embedded in the original CM.

APPENDIX D: DERIVATION OF A LOWER BOUND FOR THE SPATIAL DISTANCE

Let us use $\bar{\cdot}$ and $\langle \cdot \rangle$ as notations for the average over each genome conformations in a single homogeneous population and the average over each individual subpopulations, respectively. Here, $\bar{r}_{ij}$ and p_{ij} are the *mean* spatial distance and the contact probability between loci i and j for a single homogeneous (sub)population. $\langle \bar{r}_{ij} \rangle$ and the $\langle p_{ij} \rangle$ are the *mean* spatial distance and the contact probability between loci i and j measured for the whole

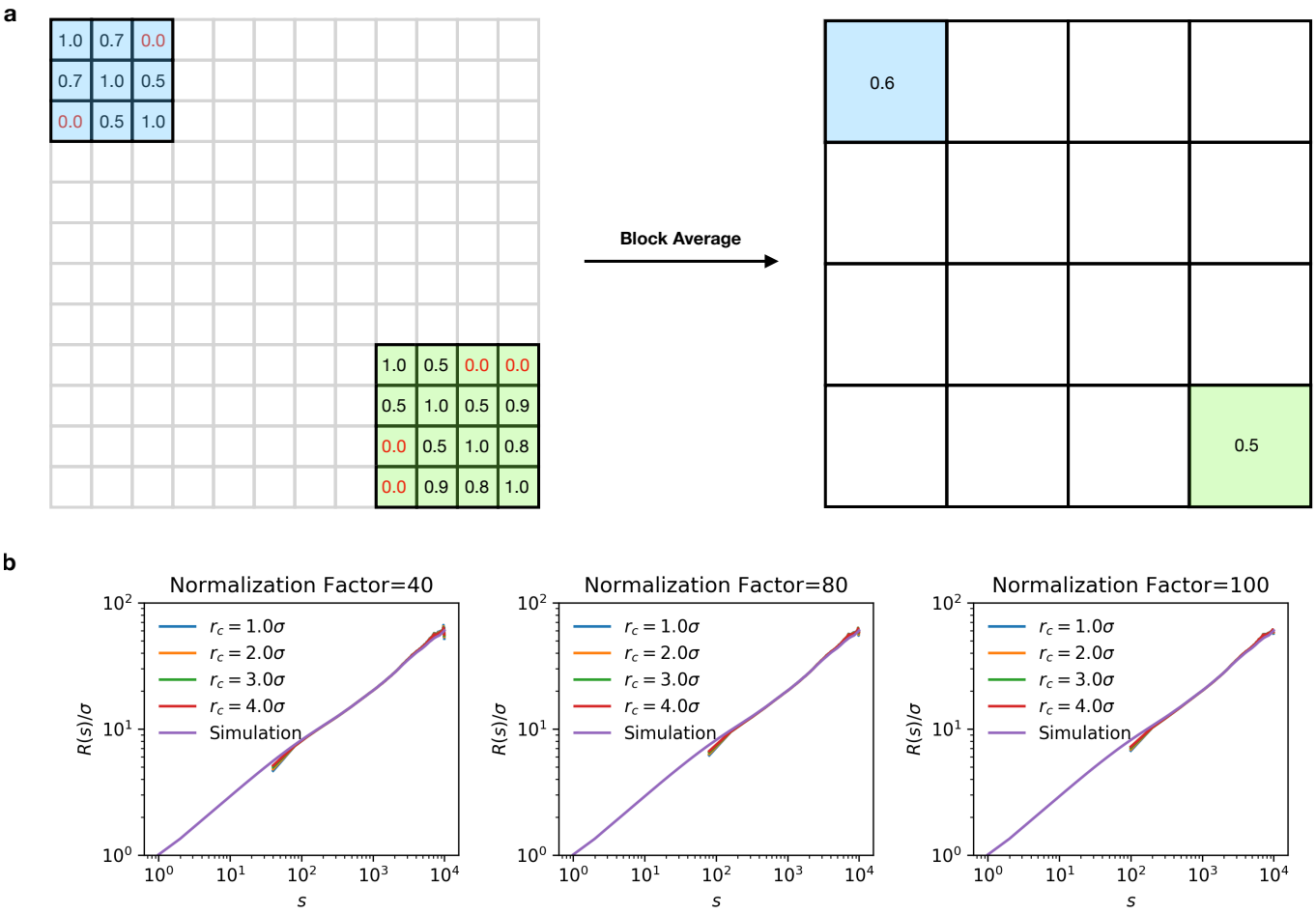

FIG. 8. **(a)** Illustration of block average performed on sparse contact map matrix ($\langle \mathbf{P} \rangle$). There are zero value elements in the original $\langle \mathbf{P} \rangle$ (matrix on the left). When constructing the distance matrix, $\langle \bar{\mathbf{R}} \rangle$, from such $\langle \mathbf{P} \rangle$, the zero value contact probability would naively imply that $\langle \bar{r} \rangle \to \infty$. To overcome this problem, we use block averages. The original $N \times N$ $\langle \mathbf{P} \rangle$ are reconstructed into blocks with size n (red blocks on top left). The value of each block is computed as the mean value of the original elements in each block (matrix on the right). The size of the matrix is reduced from N to N/n where n is the normalization factor. The same procedure could also be applied to $\langle \bar{\mathbf{R}} \rangle$. **(b)** Block average does not alter the information embedded in the original $\langle \mathbf{P} \rangle$ and the calculated $\langle \bar{\mathbf{R}} \rangle$. $R(s)$ is computed for different values of the normalization factor, n. The results in the panel do not depend on the normalization factor. The insensitivity of the results to the block averaging justifies its use in overcoming the problem of missing data points on the $\langle \mathbf{P} \rangle$.

population. It is easy to see that if the population is homogeneous, we have $\langle \bar{r}_{ij} \rangle = \bar{r}_{ij}$ and $\langle p_{ij} \rangle = p_{ij}$.

In this appendix, we prove that there exists a theoretical lower bound for $\langle \bar{r}_{ij} \rangle$ for a given $\langle p_{ij} \rangle$. We assume that for a homogeneous population where only one population is present, there exists a convex and monotonic decreasing function relating the contact probability between two loci and their mean spatial distance, $\bar{r}_{ij} = \phi(p_{ij})$. For better readability, we will neglect the suffix ij from now on. For a heterogeneous population, the contact probability is calculated as,

$$\begin{aligned} \langle p \rangle &= \int_0^{r_c} \int_0^{\infty} \mathrm{d}r \mathrm{d}\bar{r} K(\bar{r}) P(r|\bar{r}) \\ &= \int_0^{\infty} \mathrm{d}\bar{r} K(\bar{r}) \int_0^{r_c} \mathrm{d}r P(r|\bar{r}) \\ &= \int_0^1 pK(\phi(p)) \frac{\mathrm{d}\bar{r}}{\mathrm{d}p} \mathrm{d}p \\ &\equiv \int_0^1 p\psi(p) \mathrm{d}p \end{aligned} \tag{17}$$

where $K(\bar{r})$ is the distribution of $\bar{r}$ for all the subpopulations, and $P(r|\bar{r})$ is the distribution of spatial distance for a single subpopulation given its mean value $\bar{r}$. r_c is the threshold distance for determining the contact. Note that

$p = \int_0^{r_c} \mathrm{d}r P(r|\bar{r})$ by definition. $\psi(p) \equiv K(\phi(p))(\mathrm{d}\bar{r}/\mathrm{d}p)$ is the probability measure of p over individual subpopulation. Since ϕ is a convex function, according to Jensen's inequality, we have,

$$\phi(\langle p \rangle) \leq \langle \phi(p) \rangle = \int \phi(p)\psi(p)\mathrm{d}p \tag{18}$$

Replace the $\psi(p)$ by $K(\phi(p))(\mathrm{d}\bar{r}/\mathrm{d}p)$. We have,

$$\begin{aligned} \phi(\langle p \rangle) &\leq \int \phi(p)K(\phi(p))\frac{\mathrm{d}\bar{r}}{\mathrm{d}p}\mathrm{d}p \\ &= \int \bar{r}K(\bar{r})\mathrm{d}\bar{r} = \langle \bar{r} \rangle \end{aligned} \tag{19}$$

Eq. 19 shows that the lower bound for $\langle \bar{r} \rangle$ is the mean spatial distance inferred from the $\langle p \rangle$ as if the populations of genome conformation is homogeneous, i.e. there is only one single population.

To demonstrate the validity of Eq. 19, we consider the special case where there are two distinct discrete subpopulations. In this case, we $\langle \bar{r} \rangle = \eta\bar{r}_1 + (1-\eta)\bar{r}_2$ and $\langle p \rangle = \eta p_1 + (1-\eta)p_2$. Note that $\bar{r}_1 = \phi(p_1)$ and $\bar{r}_2 = \phi(p_2)$. Let us denote $p_1 = x$ and $p_2 = y$. Given the value of the contact probability $\langle p \rangle$, we show that the lower bound for $\langle \bar{r} \rangle$ is $\phi(\langle p \rangle)$. This is equivalent to the optimization problem,

$$\begin{aligned} &\text{maximize } f(x,y) \\ &\text{subject to } g(x,y) = 0 \end{aligned} \tag{20}$$

where $f(x,y) = -\eta\phi(x) - (1-\eta)\phi(y) \equiv -\langle \bar{r} \rangle$ and $g(x,y) = \eta x + (1-\eta)y - \langle p \rangle$. The Lagrange multiplier is $\mathcal{L}(x,y,\phi) = f(x,y) - \phi g(x,y)$. Using the condition that $\nabla_{x,y,\phi}\mathcal{L}(x,y,\phi) = 0$, it can be shown that $f(x,y)$ is maximized when $x = y$. Thus, we proved that $\langle \bar{r} \rangle$ is minimized when $p_1 = p_2$ and its minimum value is $\phi(\langle p \rangle)$. This is also graphically illustrated in Fig.2a in the main text.

APPENDIX E: CONNECTION BETWEEN THE CONTACT PROBABILITY AND MEAN SPATIAL DISTANCE

For a self-avoiding homopolymer, the distance distribution between two monomers along a polymer chain is [39],

$$P(r|\bar{r}) = A(r/\bar{r})^{2+g}\exp(-B(r/\bar{r})^{\delta}) \tag{21}$$

where r is the distance between two monomers, $\bar{r}$ is the mean distance between them. g is "correlation hole" exponent, and δ is related to the Flory exponent by $\delta = 1/(1-\nu)$. Given the contact threshold, the contact probability p between the two monomers is

$$p = \int_0^{r_c} P(r|\bar{r})\mathrm{d}r \tag{22}$$

When the contact threshold is small compared to the size of the chain $r \ll \bar{r}$, the integral can be approximately evaluated as,

$$\begin{aligned} p &= \lim_{r_c \to 0} \int_0^{r_c} P(r|\bar{r})\mathrm{d}r \\ &= \lim_{r_c \to 0} \int_0^{r_c} A(r/\bar{r})^{2+g}\exp(-B(r/\bar{r})^{\delta})\mathrm{d}r \\ &\sim \bar{r}^{-(3+g)} \end{aligned} \tag{23}$$

Thus, the contact probability between two monomers, p, is connected to their mean distance $\bar{r}$ by a scaling exponent $-(3+g)$. For an ideal chain, $g = 0$, we recover the asymptotically exact relation $p \sim \bar{r}^{-3}$. For a self-avoiding chain, we need to consider three cases [39]: (i) two monomers are at the two ends of the chain. (ii) one monomer is in the chain interior, while the other is at the end. (iii) two monomers are located in the central part of a chain. The correlation hole exponents corresponding to the three cases [39] are $g_1 = 0.273$, $g_2 = 0.46$ and $g_3 = 0.71$. Thus, we have $p = \bar{r}^{-3.273}$ for the contact between two ends of a self-avoiding chain. $p = \bar{r}^{-3.46}$ for contact between two monomers in case (ii), and $p = \bar{r}^{-3.71}$ for the contacts between two monomer located in the chain interior.

For polymers in poor solvents (likely more relevant to the Humam interphase chromosomes), the value of g is not well known. Using simulation, Bohn et al [40] showed that for a equilibrium collapsed homopolymer chain, $g = -0.11$ for two ends of the chain. This leads to the contact probability between two ends of an equilibrium homopolymer globule and the mean distance $p = \bar{r}^{-2.89}$. But the values of g for scenarios (ii) and (iii) are unknown. In addition, copolymer and out of equilibrium states of chromosomes even complicate the theoretical calculations. Hence, the theoretical estimate of the relation between p and $\bar{r}$ for chromosomes is not known rigorously. Nevertheless, we expect based on the arguments given here that a power law connecting p and $\bar{r}$ ought to exist. We determine the precise relation based on experimental data and our previous study [25].

APPENDIX G: ITERATIVE SCALING ALGORITHM FOR MAXIMUM ENTROPY PRINCIPLE

Here, we describe the algorithm for obtaining the k_{ij}s in Eq.8. The algorithm we adopted is iterative scaling.

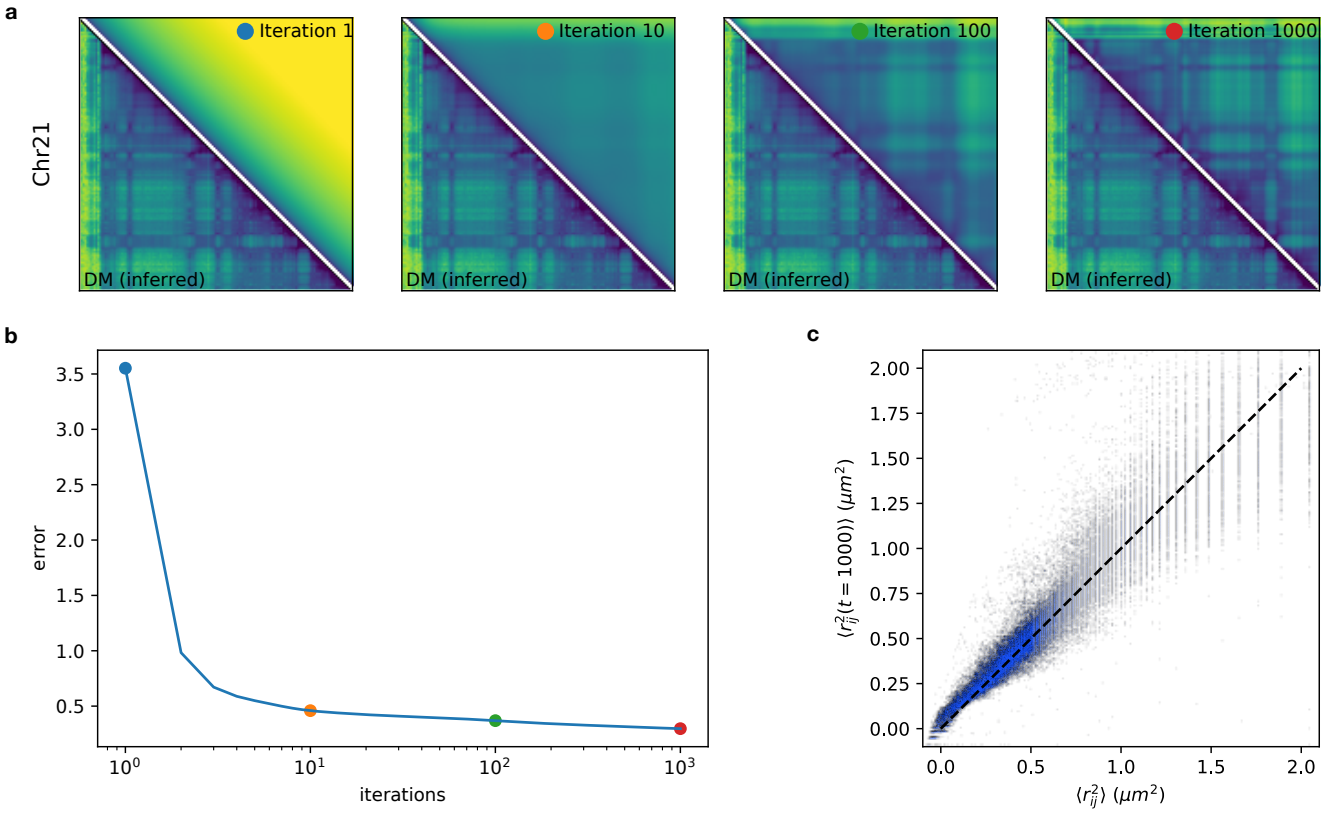

FIG. 9. **(a)** Comparison between the targeted distance map (DM) (lower triangle) and the distance matrix at different iteration steps. At iteration step 1,000, we achieve good agreement with targeted DM. **(b)** The error as a function iteration steps. The error is defined as the L2 norm between targeted DM and simulated DM. **(c)** The scatter plot between targeted $\langle r_{ij}^2\rangle$ and $\langle r_{ij}^2(t)\rangle$ at $t = 1000$. The pearson correlation coefficient between $\langle r_{ij}^2\rangle$ and $\langle r_{ij}^2(t = 1000)\rangle$ is 0.92.

Denote $k_{ij}(t)$ as the value of k_{ij} at t^{th} iteration, it is updated according to,

$$k_{ij}(t+1) = k_{ij}(t) + \frac{r}{\sum_{i<j}\langle r_{ij}^2(t)\rangle} \ln \frac{\langle r_{ij}^2(t)\rangle}{\langle r_{ij}^2\rangle} \qquad (24)$$

where r is the learning rate. $\langle r_{ij}^2(t)\rangle$ is the average squared pairwise distance at t^{th} iteration and $\langle r_{ij}^2\rangle$ is the targeted squared pairwise distance. Generally, the value of $\langle r_{ij}^2(t)\rangle$ can be estimated by simulation under the values of $k_{ij}(t)$. In this particular case, $\langle r_{ij}^2(t)\rangle$ can be numerically computed since P^{MaxEnt} is a multivariate normal distribution.

To demonstrate the effectiveness of the algorithm, Fig.9 shows the comparison between targeted average distance matrix and simulated average distance matrix at different iteration steps. It is clear that after a sufficient number of steps, the simulated distance matrix converges to the targeted one with high accuracy.

APPENDIX H: RELATIVE SHAPE ANISTROPY

To quantify the shape of each chromosome conformation, we calculate the relative shape anistropy (κ^2) as following,

$$\kappa^2 = \frac{3}{2}\frac{\lambda_1^2+\lambda_2^2+\lambda_3^2}{(\lambda_1+\lambda_2+\lambda_3)^2} - \frac{1}{2} \qquad (25)$$

where $\lambda_{1,2,3}$ are the eigenvalues of the gyration tensor. The bounds for κ^2 is $0 \leq \kappa^2 \leq 1$, where 0 is for highly symmetric conformation and 1 corresponds to a rod.

APPENDIX I: PROCESSING ATAC-SEQ DATA

Each monomer/loci in the 3D structures generated is assigned a value representing its ATAC signal. We use ATAC BED file from GEO repository GSE47753. The original data, however, needed to be processed in order to use in our model. The procedure is illustrated in Fig.10. Each line in the BED file corresponds to a ATAC peak, associated with the peak value and the start and end

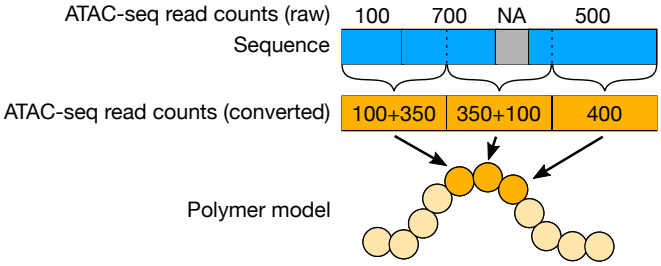

FIG. 10. The procedure for processing ATAC-seq peak data. The raw ATAC-seq read counts data is illustrated at the top track. Each chromatin segment has a read count value. The segments are not distributed uniformly, but have different lengths, and have missing parts. In our model, each monomer represents a fixed length segment. Thus, to estimate the read counts associated with each monomer, we calculate the contribution from the original ATAC-seq segments (blue track) to the segments represented by the monomer (yellow track).

genomic positions of the segment. In our model, each monomer represents a 100kbps genome segment. We count how many basepairs are overlapped between the segment represented by the monomer in our model and the segment in the ATAC-seq data. The contribution to the monomoer's ATAC signal value is computed proportionally from the peak value. For instance, the segment in the ATAC data has a peak value 100, and its length is 50kpbs, and it has overlap of length 30kbps with a given monomer. Then the contribution of ATAC signal from the segment in the ATAC data is $(30/50) * 100 = 60$. If a segment has no corresponding data in the ATAC BED fle, we treat it as it has peak value zero.

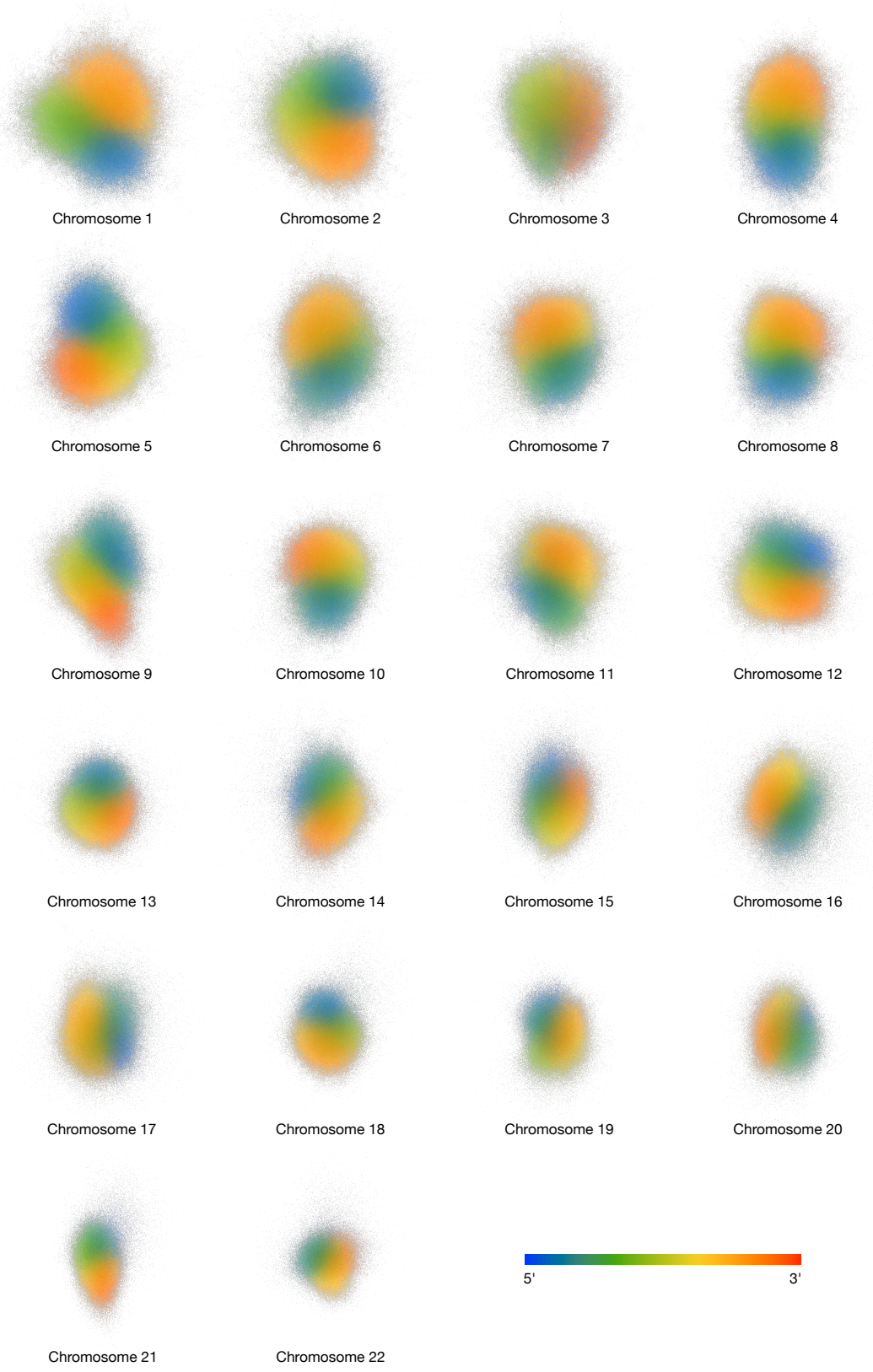

FIG. 11. Superpositions of an ensemble of 3D structures for all 23 chromosomes. A total number of 1,000 conformations are aligned and superimposed for each chromosome. Each point represent one loci from one conformations, with color representing the genomic location of the loci along the genome

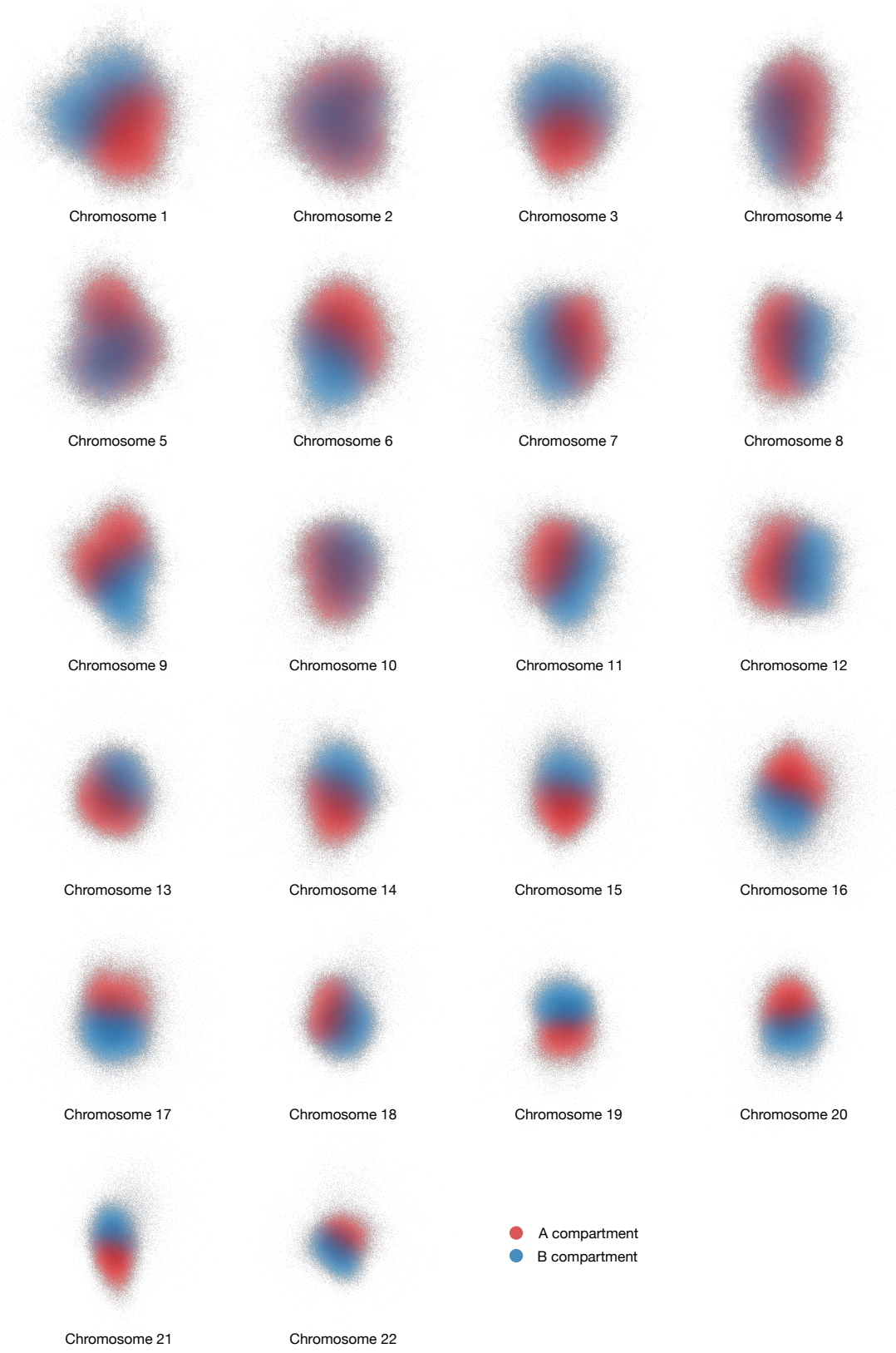

FIG. 12. Superpositions of an ensemble of 3D structures for all 23 chromosomes. A total number of 1,000 conformations are aligned and superimposed for each chromosome. Each point represent one loci from one conformations, with color representing the A/B compartments. Note that the A/B compartments do not necessarily correspond to the same state across different chromosomes since the assignment of label A or label B is arbitrary.

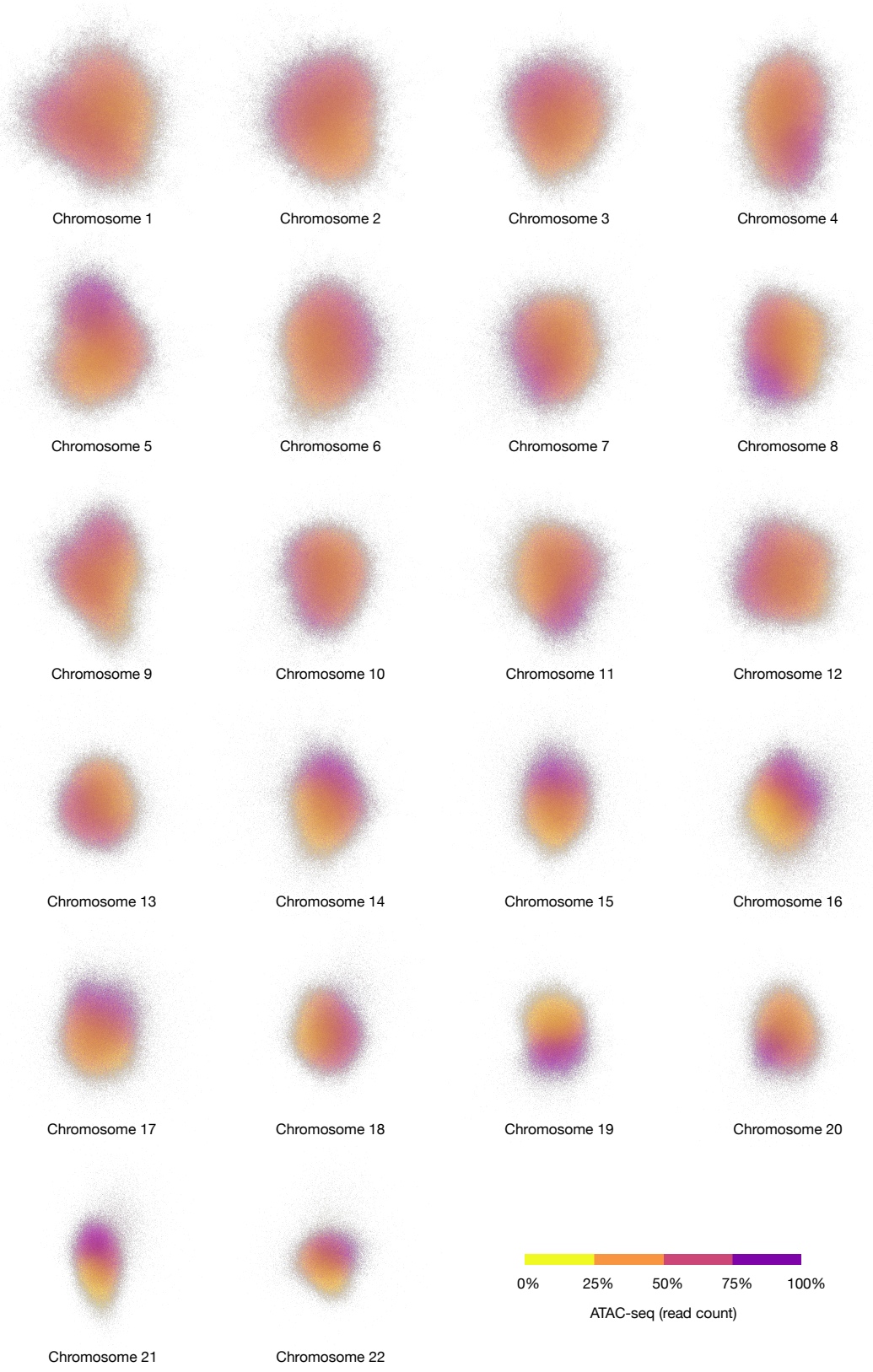

FIG. 13. Superpositions of an ensemble of 3D structures for all 23 chromosomes. A total number of 1,000 conformations are aligned and superimposed for each chromosome. Each point represent one loci from one conformations, with color encoding the ATAC-seq signal values

Acknowledgements: We are grateful to the National Science Foundation (CHE 19-00093) and the Collie-Welch Regents Chair (F-0019) for supporting this work.

[1] Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009). URL https://doi.org/10.1126/science.1181369.

[2] Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012). URL https://doi.org/10.1038/nature11082.

[3] Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the drosophila genome. *Cell* **148**, 458–472 (2012). URL https://doi.org/10.1016/j.cell.2012.01.010.

[4] Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013). URL https://doi.org/10.1038/nature12644.

[5] Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics* **14**, 390–403 (2013). URL https://doi.org/10.1038/nrg3454.

[6] Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

[7] Wang, S. *et al.* Spatial organization of chromatin domains and compartments in single chromosomes. *Science* **353**, 598–602 (2016).

[8] Bryngelson, J. & Thirumalai, D. Internal constraints induce localization in an isolated polymer molecule. *Phys. Rev. Lett.* **76**, 542 (1996).

[9] Shi, G. & Thirumalai, D. Conformational heterogeneity in human interphase chromosome organization reconciles the FISH and hi-c paradox. *Nature Communications* **10** (2019). URL https://doi.org/10.1038/s41467-019-11897-0.

[10] Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465**, 363–367 (2010). URL https://doi.org/10.1038/nature08973.

[11] Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology* **30**, 90–98 (2011). URL https://doi.org/10.1038/nbt.2057.

[12] Rousseau, M., Fraser, J., Ferraiuolo, M. A., Dostie, J. & Blanchette, M. Three-dimensional modeling of chromatin structure from interaction frequency data using markov chain monte carlo sampling. *BMC Bioinformatics* **12**, 414 (2011). URL https://doi.org/10.1186/1471-2105-12-414.

[13] Zhang, Z., Li, G., Toh, K.-C. & Sung, W.-K. 3d chromosome modeling with semi-definite programming and hi-c data. *Journal of Computational Biology* **20**, 831–846 (2013). URL https://doi.org/10.1089/cmb.2013.0076.

[14] Hu, M. *et al.* Bayesian inference of spatial organizations of chromosomes. *PLoS Computational Biology* **9**, e1002893 (2013). URL https://doi.org/10.1371/journal.pcbi.1002893.

[15] Varoquaux, N., Ay, F., Noble, W. S. & Vert, J.-P. A statistical approach for inferring the 3d structure of the genome. *Bioinformatics* **30**, i26–i33 (2014). URL https://doi.org/10.1093/bioinformatics/btu268.

[16] Lesne, A., Riposo, J., Roger, P., Cournac, A. & Mozziconacci, J. 3d genome reconstruction from chromosomal contacts. *Nature Methods* **11**, 1141–1143 (2014). URL https://doi.org/10.1038/nmeth.3104.

[17] Tjong, H. *et al.* Population-based 3d genome structure analysis reveals driving forces in spatial genome organization. *Proceedings of the National Academy of Sciences* **113**, E1663–E1672 (2016). URL https://doi.org/10.1073/pnas.1512577113.

[18] Hua, N. *et al.* Producing genome structure populations with the dynamic and automated PGS software. *Nature Protocols* **13**, 915–926 (2018). URL https://doi.org/10.1038/nprot.2018.008.

[19] Giorgetti, L. & Heard, E. Closing the loop: 3c versus DNA FISH. *Genome Biology* **17** (2016). URL https://doi.org/10.1186/s13059-016-1081-2.

[20] Fudenberg, G. & Imakaev, M. FISH-ing for captured contacts: towards reconciling FISH and 3C. *Nature Methods* (2017).

[21] Bickmore, W. A. & van Steensel, B. Genome Architecture: Domain Organization of Interphase Chromosomes. *Cell* **152**, 1270–1284 (2013). URL https://doi.org/10.1016/j.cell.2013.02.001.

[22] Williamson, I. *et al.* Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Gene. Dev.* **28**, 2778–2791 (2014).

[23] Finn, E. H. *et al.* Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell* **176**, 1502–1515.e10 (2019). URL https://doi.org/10.1016/j.cell.2019.01.020.

[24] Lee, H., Ma, Z., Wang, Y. & Chung, M. K. Topological Distances between Networks and Its Application to Brain Imaging. *arXiv preprint arXiv:1701.04171* (2017).

[25] Shi, G., Liu, L., Hyeon, C. & Thirumalai, D. Interphase human chromosome exhibits out of equilibrium glassy dynamics. *Nature Communications* **9** (2018). URL https://doi.org/10.1038/s41467-018-05606-6.

[26] Fudenberg, G. & Imakaev, M. FISH-ing for captured contacts: towards reconciling FISH and 3c. *Nature Methods* **14**, 673–678 (2017). URL https://doi.org/10.1038/nmeth.4329.

[27] Székely, G. J., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* 2769–2794 (2007).

[28] Branco, M. R. & Pombo, A. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol.* **4**, e138 (2006).

[29] Rosa, A. & Everaers, R. Structure and dynamics of interphase chromosomes. *PLoS Computational Biology* **4**, e1000153 (2008). URL https://doi.org/10.1371/

journal.pcbi.1000153.

[30] Stevens, T. J. *et al.* 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59–64 (2017). URL https://doi.org/10.1038/nature21429.

[31] Pierro, M. D., Zhang, B., Aiden, E. L., Wolynes, P. G. & Onuchic, J. N. Transferable model for chromosome architecture. *Proceedings of the National Academy of Sciences* **113**, 12168–12173 (2016). URL https://doi.org/10.1073/pnas.1613607113.

[32] Farré, P. & Emberly, E. A maximum-entropy model for predicting chromatin contacts. *PLOS Computational Biology* **14**, e1005956 (2018). URL https://doi.org/10.1371/journal.pcbi.1005956.

[33] Grosberg, A. Y., Nechaev, S. K. & Shakhnovich, E. I. The role of topological constraints in the kinetics of collapse of macromolecules. *J. Phys-paris.* **49**, 2095–2100 (1988).

[34] Grosberg, A., Rabin, Y., Havlin, S. & Neer, A. Crumpled globule model of the three-dimensional structure of DNA. *Europhys. Lett.* **23**, 373 (1993).

[35] Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).

[36] Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10**, 1213–1218 (2013). URL https://doi.org/10.1038/nmeth.2688.

[37] Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

[38] Honeycutt, J. & Thirumalai, D. The nature of folded states of globular proteins. *Biopolymers* **32**, 695–709 (1992).

[39] Des Cloizeaux, J. Short range correlation between elements of a long polymer in a good solvent. *Journal de Physique* **41**, 223–238 (1980).

[40] Bohn, M. & Heermann, D. W. Conformational properties of compact polymers. *The Journal of chemical physics* **130**, 174901 (2009).