Biophysical Journal
Article

Biophysical Society

Heterogeneous Loop Model to Infer 3D Chromosome Structures from Hi-C

Lei Liu,[1] Min Hyeok Kim,[1] and Changbong Hyeon[1,*]
[1]School of Computational Sciences, Korea Institute for Advanced Study, Seoul, Republic of Korea

ABSTRACT Adapting a well-established formalism in polymer physics, we develop a minimalist approach to infer three-dimensional folding of chromatin from Hi-C data. The three-dimensional chromosome structures generated from our heterogeneous loop model (HLM) are used to visualize chromosome organizations that can substantiate the measurements from fluorescence in situ hybridization, chromatin interaction analysis by paired-end tag sequencing, and RNA-seq signals. We demonstrate the utility of the HLM with several case studies. Specifically, the HLM-generated chromosome structures, which reproduce the spatial distribution of topologically associated domains from fluorescence in situ hybridization measurement, show the phase segregation between two types of topologically associated domains explicitly. We discuss the origin of cell-type-dependent gene-expression level by modeling the chromatin globules of α-globin and SOX2 gene loci for two different cell lines. We also use the HLM to discuss how the chromatin folding and gene-expression level of Pax6 loci, associated with mouse neural development, are modulated by interactions with two enhancers. Finally, HLM-generated structures of chromosome 19 of mouse embryonic stem cells, based on single-cell Hi-C data collected over each cell-cycle phase, visualize changes in chromosome conformation along the cell-cycle. Given a contact frequency map between chromatic loci supplied from Hi-C, HLM is a computationally efficient and versatile modeling tool to generate chromosome structures that can complement interpreting other experimental data.

SIGNIFICANCE The packaging of chromosomes, giant macromolecules made of hundreds-of-megabase-long DNA, into a micrometer-sized cell nucleus is truly remarkable. Recent advances in Hi-C techniques have ushered in a new era of research on genome organization. We developed a computationally efficient and versatile approach, called the heterogeneous loop model, to generate chromosome structural ensemble from Hi-C. The heterogenous-loop-model-generated three-dimensional chromosome structures not only substantiate the chromosome organizations implicated by diverse experimental data but also allow us to decipher the structural origin of genome function and variation of gene-expression level along the cell cycle and across different cell types.

INTRODUCTION

Recent advances in chromosome conformation capture techniques combined with parallel sequencing (1–5) and fluorescence imaging microscopies have ushered in a new era of chromosome research over the past decade. Along with post-translational histone modifications, which have led to the conceptualization of epigenomes (6), the critical findings from fluorescence imaging and Hi-C data that the spatial organization of chromatin varies with the tissue or cell types (7,8), cell cycle (4), and pathological states (9–11) have brought a new dimension to our understanding of genome functions.

Among others, maps of genome-wide contact frequencies, quantified by Hi-C data, offer unprecedented opportunities to infer 3D chromosome structures in cell nuclei (12–22). In a nutshell, Hi-C provides the contact frequencies of genomic loci pairs based on the statistics of PCR-amplified DNA fragments digested from formaldehyde cross-linked cells (1,2). One could interpret that Hi-C measures the population-sampled contact probability between pair of genomic loci, say i and j, p_{ij}. A proper mathematical mapping of p_{ij} to the spatial distance r_{ij} is of critical importance for interpreting fluorescence imaging data (23,24) in comparison with Hi-C data.

The advent of fluorescence in situ hybridization (FISH) followed by C-based techniques have engendered much

Submitted March 5, 2019, and accepted for publication June 25, 2019.
*Correspondence: hyeoncb@kias.re.kr
Editor: Wilma Olson.
https://doi.org/10.1016/j.bpj.2019.06.032

devotion to capturing the principle underlying the three-dimensional (3D) folding of chromosomes. This has led to development of a series of polymer-based models over the decades, which include the "multiloop subcompartment model" (25,26), the "random loop model" (RLM) (27–29), the "strings and binders switch" model (12,15,30) and its derivative (17,31,32), the "loop extrusion model" (13–15,33), the "minimal chromatin model" (34), and, more recently, the "chromosome copolymer model" (22). Among them, although applicability is limited to the associated spatiotemporal scale of the model being considered, some were developed by keeping a specific molecular mechanism in mind or by incorporating "one-dimensional" information of epigenetic modification and/or DNA accessibility along genomic loci as an input to a heteropolymer model (22,32,35). On the other hand, partly sacrificing model simplicity, others were developed solely for the purpose of reconstructing more precise 3D chromatin structures from Hi-C (20,36–38) and other experiments (39).

As the cell imaging data over different cell types are rapidly growing, comparative study of chromosome conformations has become imperative. In the abovementioned models, however, a physically sound mapping of p_{ij} from Hi-C to the spatial distance r_{ij} (see review (40)) is still lacking, and computational costs are still high. To this end, here we develop a minimalist model that allows us to generate chromatin conformations from Hi-C data in a most efficient way and to study the structural characteristics of chromosome at a length scale of interest corresponding to the resolution of the given data. To achieve such a goal in the most simplifying manner, one could learn much from the literature of generic polymer problems, such as the collapse transition of an isolated polymer chain or macromolecular networks with increasing numbers of internal bonds (41–44) and polymer conformation and dynamics inside confinement (45–47).

Pushing the polymer physics idea to its extreme, we propose a minimalist approach, termed the heterogeneous loop model (HLM), that allows us to build 3D structures of chromosomes from Hi-C data. The HLM adapts the RLM, which was originally developed based on a randomly cross-linked polymer chain (27,28,48). In the RLM, which represents chromosome conformation in terms of the sum of harmonic potentials, pairwise contact probabilities are expressed analytically in terms of a few model parameters. Here, without sacrificing the mathematical tractability and simplicity of the RLM, we extend the RLM to the HLM by allowing the loop interactions to be nonuniform and heterogeneous such that the resulting loop interactions can best represent a given Hi-C data set.

In this study, we apply the HLM to various regions of human and mouse genomes that span 1–100 Mb at 5–500 kb resolution and generate the corresponding conformational ensemble of chromosomes. We demonstrate the utilities of the HLM by comparing the structural information extracted from an HLM-generated chromosome ensemble with those implicated by the measurements from FISH (23,24,28), chromatin interaction analysis by paired-end tag sequencing (49,50), and previous modeling studies (28,32,37,51,52). Through multiple examples, this study will demonstrate that the HLM is an excellent approach to infer 3D structures from Hi-C data.

METHODS

Description of the HLM

The full energy potential of the HLM consists of two parts.

$$U_{\mathrm{HLM}}(\mathbf{r}) = U_{\mathcal{K}}(\mathbf{r}) + U_{\mathrm{nb}}(\mathbf{r}) \tag{1}$$

In what follows, we delineate the first and second terms of Eq. 1 (see Supporting Materials and Methods for technical details).

First, decomposed into two parts, $U_{\mathcal{K}}(\mathbf{r})$ describes the harmonic constraints on a chain of N monomers (27),

$$\begin{aligned} U_{\mathcal{K}}(\mathbf{r}) &= \sum_{i=1}^{N-1}\frac{k}{2}\left(\vec{r}_i-\vec{r}_{i-1}\right)^2+\sum_{i=0}^{N-3}\sum_{j=i+2}^{N-1}\frac{k_{ij}}{2}\left(\vec{r}_i-\vec{r}_j\right)^2, \\ &= \frac{3}{2}\mathbf{r}^T\mathbf{K}\mathbf{r}, \end{aligned} \tag{2}$$

where successive monomers along the backbone and nonsuccessive monomers forming loops are both harmonically restrained. In the second line, $U_{\mathcal{K}}(\mathbf{r})$ is written in a compact form with $\mathbf{r}=\left(\vec{r}_1,\vec{r}_2,\cdots,\vec{r}_{N-1}\right)^T$ and $\mathbf{K}$ representing the Kirchhoff matrix. $\mathbf{K}$ can be built from the interaction strength matrix $\mathcal{K}$, which takes $k_{ij}=(\mathcal{K})_{ij}$ as its matrix element. The interaction strengths ought to be non-negative ($k_{ij}\geq 0$) for all i and j-th monomer pairs. In the HLM, if $k_{ij}\neq 0$, then the i and j-th monomer has a potential to form a (chromatin) loop. After removing the translational degrees of freedom by setting $\vec{r}_0=(0,0,0)$ in Eq. 2, we obtain the probability density of pairwise distance as (27)

$$P\left(r_{ij}\right) = 4\pi\left(\gamma_{ij}/\pi\right)^{3/2}r_{ij}^2e^{-\gamma_{ij}r_{ij}^2}, \tag{3}$$

where

$$\gamma_{ij} = \begin{cases} \dfrac{1}{2\left(\sigma_{ii}+\sigma_{jj}-2\sigma_{ij}\right)}, & i>0 \\ \dfrac{1}{2\sigma_{jj}}, & i=0 \end{cases} \tag{4}$$

and $\sigma_{ij}[=\langle\delta\vec{r}_i\cdot\delta\vec{r}_j\rangle]$ is the covariance between the positions of i and j-th monomers, which can be obtained from an inverse of $\mathbf{K}$-matrix as

$$\sigma_{ij} = \left(\mathbf{K}^{-1}\right)_{ij} \tag{5}$$

One can obtain the contact probability p_{ij} by integrating the pairwise distance $P(r_{ij})$ (Eq. 3) up to a certain capture radius (r_c) (53,54), $p_{ij}=\int_0^{r_c}P(r_{ij})dr_{ij}$, which gives

$$p_{ij} = \mathrm{erf}\left(\sqrt{\gamma_{ij}r_c^2}\right)-2\sqrt{\frac{\gamma_{ij}r_c^2}{\pi}}e^{-\gamma_{ij}r_c^2}, \tag{6}$$

where $\mathrm{erf}(x)=(2/\sqrt{\pi})\int_0^x dte^{-t^2}$. Therefore, a one-to-one analytical mapping between p_{ij} and k_{ij} follows from the precise mappings between p_{ij} and σ_{ij} from Eqs. 4 to 6 and between σ_{ij} and k_{ij} from Eq. 5.

Although it is tempting to directly use the mathematical relation between p_{ij} and k_{ij} to obtain $\mathcal{K}$ from Hi-C data, there is an unavoidable numerical issue (see Supporting Materials and Methods and Figs. S3–S5 for details). In practice, we calculate the $\tilde{\mathcal{K}}$-matrix that approximates $\mathcal{K}$ by selecting only the significant contacts in $\mathcal{P}$. More specifically, we evaluate the significance of contact probability p_{ij} by calculating z_{ij}, which is defined as (see the *matrix elements* in the *upper diagonal part* of Fig. 1 *B*)

$$z_{ij} = \frac{p_{ij}}{P(s)}, \tag{7}$$

where $P(s) = (1/N-s)\sum_{i=0}^{N-s-1} p_{i,i+s}$ is the mean contact probability for monomer pairs separated by the arc length s along the contour. The greater the value of z_{ij}, the more significant the contacts are deemed. We then select the top $2N$ (i, j) pairs ranked in terms of the values of z_{ij} (>1) (the *matrix elements* in the *lower diagonal part* of Fig. 1 *B*). For these $2N$ pairs whose contact probability p_{ij} is given in $\mathcal{P}$, the precise value of γ_{ij}^* (or equivalently $\langle r_{ij}^{2*}\rangle = \int_0^\infty r_{ij}^2 P(r_{ij})dr_{ij} = (3/2\gamma)_{ij}^*$) can be determined using Eq. 6. Then, starting from a Rouse chain configuration as an initial input, we add nonsuccessive bonds with varying interaction strengths ($0 \le k_{ij} \le 10\, k_BT/a^2$) until we minimize the objective function $\mathcal{F}(\mathcal{K})$

$$\mathcal{F}(\mathcal{K}) = \sum_{(i,j)}^{2N} \omega_{ij}\left(\frac{\left\langle r_{ij}^2(\{k_{\alpha\beta}\})\right\rangle}{\left\langle r_{ij}^{2*}\right\rangle} - 1\right)^2 \tag{8}$$

so as to determine the optimal values of $\tilde{\mathcal{K}} = \{\tilde{k}_{\alpha\beta}\} = \min_{\{k_{\alpha\beta}\}}\mathcal{F}(\mathcal{K})$. Here, the weight factor ω_{ij}, which is used to normalize the statistical bias from chromatin loops of different sizes, is defined as

$$\omega_{ij} = \omega(|i-j|) = \omega(s) = \frac{n^{-1}(s)}{\sum_s n^{-1}(s)}, \tag{9}$$

where $n(s) = \sum_{(i,j)}\delta(|i-j|-s)$ is the number of loops of size s. The gradient-descent algorithm (L-BFGS-B method in SciPy package) was used to determine the optimal parameters $\{\tilde{k}_{\alpha\beta}\}$. A fully convergent solution of $\tilde{\mathcal{K}}$-matrix (Fig. 1 *C*) could be obtained within a few minutes when N was not too large ($\le$200). This $\tilde{\mathcal{K}}$-matrix determining process, termed "constrained optimization," faithfully reproduces the original $\mathcal{K}$ matrix with a relative error smaller than 5% (see also Figs. S3–S5).

In fact, the number of selected top contact pairs (n_c) could have been $3N$, N, or even $N/2$. But we found that when $n_c \ge 2N$, the quality of the resulting interaction strength matrix $\tilde{\mathcal{K}}$ is already good enough that the Pearson correlation (PC) between the original Hi-C and the contact probability matrix obtained from $\tilde{\mathcal{K}}$ saturates for $n_c > 2N$ (Fig. S6). Thus, to build the interaction strength matrix by simultaneously minimizing the computational cost, we chose $n_c = 2N$.

After obtaining $\tilde{\mathcal{K}}$ (Fig. 1 *C*) and hence $U_{\mathcal{K}}(\mathbf{r})$, we added a nonbonded interaction term $U_{\rm nb}(\mathbf{r})$, defined for all i and j pairs to the full energy potential $U_{\rm HLM}(r)$ (Eq. 1):

$$U_{\rm nb}(\mathbf{r}) = \sum_{ij}\chi_{t_i,t_j}u_{\rm LJ}(r_{ij}), \tag{10}$$

where $u_{\rm LJ}(r)$ is the Lennard-Jones potential truncated for $r \ge r_c$ where $r_c = 5a/2$ with $\epsilon = 0.45\ k_BT$,

$$u_{\rm LJ}(r) = \epsilon\left[\left(\frac{a}{r}\right)^{12} - 2\left(\frac{a}{r}\right)^{6}\right]\Theta(r_c - r) \tag{11}$$

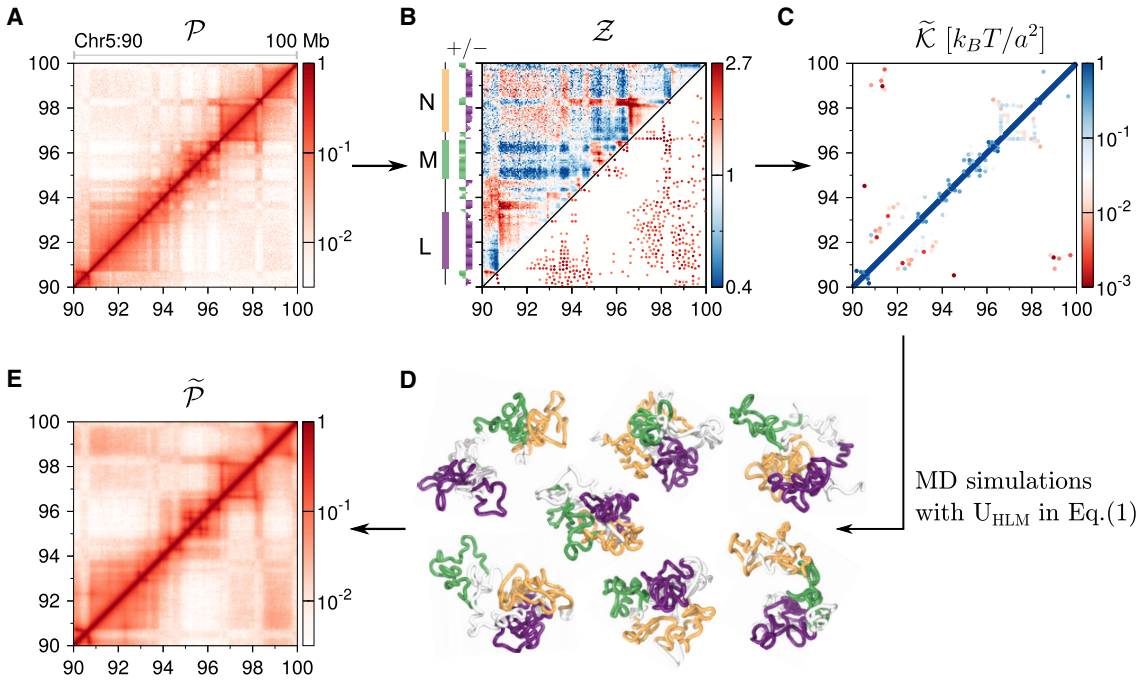

FIGURE 1 The pipeline of the HLM. (*A*) Contact probability matrix $\mathcal{P}$ of a 10-Mb genomic region of chr5 in GM12878 cells is shown. (*B*) $\mathcal{Z}$ matrix calculated from Eq. 7 is shown above the diagonal. The significant contacts selected from $\mathcal{Z}$ are shown below the diagonal. The sign of the first principal component of $\mathcal{Z}$ is provided on the left-hand side of the panel, and we divide the whole chromosome domain into "L" (*purple*), "M" (*green*), and "N" (*orange*) accordingly. (*C*) The interaction strength matrix $\tilde{\mathcal{K}}$ calculated by the constrained optimization is shown. (*D*) The conformational ensemble of chromosomes generated from HLM potential defined by $\tilde{\mathcal{K}}$ and $\mathcal{Z}$ (Eq. 1) is illustrated with the L, M, and N domains colored in purple, green, and orange, respectively, following the domain labels assigned in (*B*). (*E*) shows $\tilde{\mathcal{P}}$-matrix calculated using a conformational ensemble produced from molecular dynamics simulations. The PC between $\mathcal{P}$ (*A*) and $\tilde{\mathcal{P}}$ (*E*) is 0.96. To see this figure in color, go online.

If $\epsilon = \epsilon_\theta$ ($= 0.34\ k_BT$) with $\chi_{t_i,t_j} = 1$, then $U_{\rm nb}(\mathbf{r})$ leads to θ-solvent condition for an infinitely long chain, putting the second virial coefficient to zero, i.e., $\nu_2 = (1/2)\int_0^{r_c}(1 - e^{-\beta u_{\rm NB}(r)})d^3r = 0$. We chose ε ($= 0.45\ k_BT$) slightly greater than ε_θ and assigned a loci-pair-type-dependent prefactor χ_{t_i,t_j}. Each monomer i is assigned with a type t, either "$-$" or "$+$," based on the sign of the first principal component of $\mathcal{Z}$ (see the *track* on *top* of Fig. 1 *B*). The value of prefactor χ_{t_i,t_j} (>0)—depending on the types of two loci i and j, which are $t_it_j = ++$, $--$, or $-+$ —is evaluated by averaging over all the monomer pairs of the corresponding types, such that $\chi_{p,q} = \langle z_{ij}\rangle_{t_i=p,t_j=q}$. The values of χ_{t_i,t_j} are determined based on a given Hi-C data set. For the case shown in Fig. 1, we obtain $\chi_{-,-} = 1.18$, $\chi_{-,+} = 0.79$, and $\chi_{+,+} = 1.19$. According to the Flory-Huggins theory (55), the condition $\chi^{\rm eff}_{-,+} = (1/2)(\chi_{-,-} + \chi_{+,+}) - \chi_{-,+} \approx 0.4 > 0$ leads to spatial separation between $+$ and $-$ type loci, which indeed is realized and reflected in the characteristic checkerboard pattern of Hi-C data. It should be noted, however, that the classification of type $-/+$ is not necessarily identical to the A/B compartment of chromatin. Whereas A/B compartments are genome-wide characteristics usually defined based on Hi-C data at low (Mb) resolutions (2,3), the monomers in the HLM can be always classified into types $-/+$ regardless of the resolution of the model.

Finally, we sampled 3D chromosome structures using molecular dynamics simulation, implementing the full energy potential $U_{\rm HLM}(r)$, and calculated the contact probability matrix based on an HLM-generated conformational ensemble. In the specific example demonstrated for the Hi-C data of the 10-Mb genomic region of chr5 in the GM12878 cell line (Fig. 1), $\tilde{\mathcal{P}}$ (Fig. 1 *E*) obtained from HLM-generated chromosome conformations (Fig. 1 *D*; see also the clustering analysis that highlights the conformational variability of chromosomes in Supporting Materials and Methods; Fig. S7) displays a notable resemblance to the input $\mathcal{P}$ (Fig. 1 *A*) (PC of 0.96; Spearman correlation of 0.92). Despite the simplicity of the HLM potential (Eq. 1), the similarity between $\mathcal{P}$ and $\tilde{\mathcal{P}}$, as well as the chromosome conformations ensemble generated during the procedure, is remarkable.

Structure characterization

We quantified the structural feature of HLM-generated chromosome ensemble by means of several quantities:

1) The compactness of a (sub)chain of length N is quantified in terms of r_g^3/N, where r_g is the gyration radius of the (sub)chain.
2) The asphericity (A) is calculated by $A = \sum_{i=1}^{3}(\lambda_i - \bar{\lambda})^2/6\bar{\lambda}^2$, where λ_i ($i = 1, 2, 3$) are the three eigenvalues of the moment of inertia tensor and $\bar{\lambda}$ is their mean (56,57). $A = 0$ for a sphere, and $A > 0$ for a nonspherical shape.
3) The roughness of the surface of a (sub)chain was evaluated using the Voronoi diagram (58), which tessellates the 3D space occupied by the chain. An upper bound for the volume of each monomer was set using a dodecahedron with a diameter of $2a$. The Voronoi diagram provides a well-defined volume V and surface area S of the (sub)chain. Because the surface area of a perfect sphere with the volume V is $S_0 = (36\pi V^2)^{1/3}$, we quantified the surface roughness using $S/S_0 \geq 1$.
4) To visualize an ensemble of structures with considerable variability, we first divided the chain into a few segments (domains). Next, the distribution of the distances between the geometric centers of these domains were computed based on the ensemble of structures. Several configurations of chromosomes were then randomly selected from the most populated state (in terms of interdomain distances), aligned, and rendered.

RESULTS

The HLM is effectively a multiblock copolymer model in which monomer-monomer interactions (loops) are harmonically restrained with varying interaction strengths (k_{ij}) (Methods; Supporting Materials and Methods). Mapping the pairwise contact probabilities p_{ij} from Hi-C to the model parameters k_{ij} is the essence of the HLM. By incorporating a standard Lennard-Jones nonbonded potential slightly below the θ-condition, which takes into account the short-range excluded-volume interaction between monomers as well as the global thermodynamic driving force that induces spatial separation between different monomer types, the HLM allows us to generate a conformational ensemble of chromosome structures that reproduces a contact probability matrix that displays close resemblance to an original inputted Hi-C data set. We used the HLM to model various genomic regions (see Table 1). HLM-generated chromosome conformations were used to interpret the currently available experimental results.

TABLE 1 Genomic Regions Simulated in This Work

Species	Cell line	Hi-C experiment	Chromosome	Start (bp)	End (bp)	Resolution (kb)	N	PC[a]	Time (min)[b]	Figure
Human	GM12878	(3)	chr5	90,000,000	100,000,000	50	200	0.96	4.8	Fig. 1
	IMR90	(3)	chr21	14,000,000	48,000,000	250	137	0.97	0.8	Fig. 2
	IMR90	(3)	chr11	59,000,000	94,000,000	250	140	0.98	1.7	Fig. S8
	IMR90	(3)	chr1	150,000,000	180,000,000	250	120	0.98	0.8	Fig. S9
	K562	(3)	chr16	60,000	560,000	5	100	0.94	0.2	Fig. 3
	GM12878	(3)	chr16	60,000	560,000	5	100	0.92	0.4	Fig. 3
	hESC	(88)	chr3	179,000	184,000	40	125	0.94	1.4	Fig. 4
	HUVEC	(3)	chr3	179,000	184,000	40	125	0.95	1.8	Fig. 4
Mouse	mESC	(75)	chr2	105,000,000	106,000,000	8	125	0.94	1.4	Fig. 5
	NPC	(75)	chr2	105,000,000	106,000,000	8	125	0.96	1.2	Fig. 5
	CN	(75)	chr2	105,000,000	106,000,000	8	125	0.97	1.3	Fig. 5
	ncx_NPC	(75)	chr2	105,000,000	106,000,000	8	125	0.97	0.8	Fig. 5
	ncx_CN	(75)	chr2	105,000,000	106,000,000	8	125	0.97	1.1	Fig. 5
	mESC	(4)	chr19	1	61,342,430	500	117	0.92[c]	1.4	Fig. 6

[a]The similarity between contact probabilities (p_{ij}) from Hi-C and those from modeling is quantified by the PC (see also discussions in Supporting Materials and Methods).
[b]It takes a few minutes to determine the interaction strength parameters by the constrained optimization, namely obtaining $\tilde{\mathcal{K}}$ from $\mathcal{P}$.
[c]From the post-M to pre-M phase, PC of mESCs is 0.77, 0.96, 0.96, 0.96, 0.97, and 0.91, respectively.

Spatial distribution of TADs inferred from HLM in comparison with FISH measurement

Intrachromosomal distances between topologically associated domains (TADs) in human IMR90 cells, measured by Wang et al. through a multiplexed FISH method (23), have been used as a benchmark for different models (38). To show the utility of the HLM, we model a 34-Mb genomic region on chr21 of IMR90 cells, which contains 33 labeled TADs (Table S1 provides the genomic positions of these TADs).

First, the contact probability matrix $\tilde{\mathcal{P}}$ constructed from HLM-generated structures captures the characteristic checkerboard pattern of the heatmap of Hi-C data, $\mathcal{P}$; the mean contact probability $P_{\mathrm{HLM}}(s)$ of the HLM is consistent with $P_{\text{Hi-C}}(s)$ calculated from Hi-C over all length scales, including the wiggly pattern at large s (Fig. 2, *A* and *B*).

The heatmap calculated for inter-TAD distances using the HLM-generated conformational ensemble (*lower diagonal part* of Fig. 2 *C*) can directly be compared with the FISH measurement (*upper diagonal part*). The square block pattern along the diagonal axis of the heatmap indicates that four to five adjacent TADs constitute an aggregate, reminiscent of meta-TAD (30), and the patterns in the off-diagonal part (highlighted by the *magenta boxes*) suggest long-range clustering of TADs. The error of the inter-TAD distance heatmap relative to FISH is 0.184, which is comparable to the value of the GEM model (38) and better than others (see Fig. 4 *D* in (38)). A principal component analysis of this matrix (*top left part of the matrix* in Fig. 2 *C*) divides TADs into A/B types (23). Fig. 2 *D* demonstrates the polarized organization of A- and B-type TADs by aligning the geometric centers of HLM-generated A- and B-type TADs along an axis that best separates the two types of TADs (23).

The intrachain end-to-end distance $r(s) = \sum_{i=0}^{N-s-1} r_{i,i+s}/(N-s)$ displays a scale-dependent scaling relationship with the genomic distance s, $r(s) \sim s^{\nu}$ (Fig. 2 *E*). In qualitative agreement with the FISH measurement (23), there is a crossover around $s = 7$ Mb, such that $\nu \approx 1/3$ for $s < 7$ Mb and $\nu \approx 0.21$ for $s > 7$ Mb.

We explore the relationship between contact probability p_{ij} and the corresponding distance r_{ij} of two loci. It is expected that the looping probability of polymer is inversely proportional to the volume of space (V) explored by the two loci as $P_{\mathrm{loop}} \sim 1/V$. Because the volume V scales with the spatial separation (R) between the two loci in d-dimension as $V \sim R^d$, it follows that (59–61)

$$P_{\mathrm{loop}} \sim \frac{1}{R^d} f\left(\frac{r_c}{R}\right) \sim \frac{1}{R^d}\left(\frac{r_c}{R}\right)^g \tag{12}$$

The correlation hole exponent g is $g = 0$ for a Gaussian chain (55). According to the Flory theorem (62–65), the

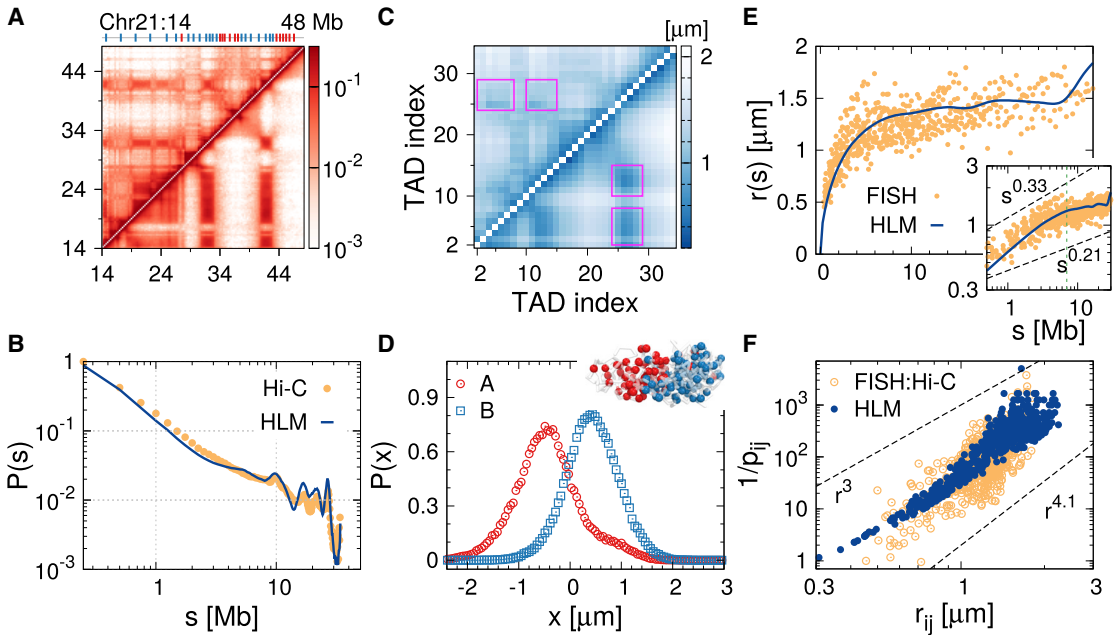

FIGURE 2 A 34-Mb-genomic region of chr21 in IMR90 cells modeled by the HLM. (*A*) A heatmap of contact probabilities from Hi-C ($\mathcal{P}$, *upper diagonal part*) and HLM ($\tilde{\mathcal{P}}$, *lower diagonal part*) is given. The PC between $\mathcal{P}$ and $\tilde{\mathcal{P}}$ is 0.97. The positions of TADs, labeled by sticks, are displayed above the heatmap. The type of each domain, A or B, is depicted in red or blue, respectively. (*B*) Plotted are the mean contact probabilities $P(s)$'s calculated from Hi-C (*orange data*) and HLM (*blue line*). (*C*) The heatmap of inter-TAD distances measured by FISH (*upper diagonal part*) is compared with that calculated from the HLM (*lower diagonal part*). (*D*) Distributions of A- and B-type TADs projected on the x axis, along which the geometric centers of different types of TADs are aligned, are indicative of the spatial separation between the two TAD types. An ensemble of structures is also shown. (*E*) Intrachain end-to-end distances $r(s)$'s as a function of arc length s from FISH (*orange data*) and the HLM (*blue line*) are shown. The inset shows $r(s)$'s in log-log scale. (*F*) Inverse of pairwise contact probability between TADs, p_{ij}^{-1} versus inter-TAD distance r_{ij} is shown in log-log scale. To see this figure in color, go online.

ideal chain statistics is a good approximation for a chain in polymer melts or for a subchain in a fully equilibrated globule. Because $d = 3$ for 3D, we expect $P_{\text{loop}} \sim R^{-3}$, or equivalently, $p_{ij} \sim r_{ij}^{-3}$ (see also Fig. S1 *B*). In fact, this scaling relation is observed for the data point generated by HLM for $r_{ij} < 1$ μm (Fig. 2 *F*). Although Wang et al., who combined Hi-C and FISH data, reported a scaling relation of $p_{ij} \sim r_{ij}^{-4.1}$ for the entire range, it is not clear whether the relation can straightforwardly be extended to the range of $r_{ij} < 1$ μm in which the data point from their measurement might be less accurate. According to the HLM-generated data, a more proper scaling should be $p_{ij} \sim r_{ij}^{-3}$ for $r_{ij} < 1$ μm and $p_{ij} \sim r_{ij}^{-4.1}$ for $r_{ij} > 1$ μm.

Next, to demonstrate another analysis on FISH measurement, we applied the HLM to the q-arm of chr11 in IMR90 cells, whose intrachain pairwise distances between genomic loci had been measured with FISH (28,66) (see Table S2 for the position of FISH probes in the genome and in the model). The model produces the contact probability matrix $\tilde{\mathcal{P}}$ with a PC of 0.98 relative to Hi-C data ($\mathcal{P}$) (see Fig. S8, *A* and *B*). The HLM enables us to calculate the spatial distances between specific pairs of loci (Fig. S8 *C*), with a mean relative error of 0.189 (with respect to FISH data). The HLM-generated structural ensemble also indicates that compared to the gene-poor and transcriptionally inactive antiridge domain, the transcriptionally active ridge domain is less compact, less spherical, and has a rougher domain surface (Fig. S8, *D–F*), all of which are in agreement with the FISH experiment (66). Modeling another 30-Mb region on chr1 of IMR90 cells leads to similar results (Fig. S9; Table S3).

Visualization of chromatin globules

α-Globin gene

Cis-regulatory elements generally mediate the transcription of neighboring genes within a range smaller than 1 Mb (67). The α-globin gene domain, a 500-kb genomic region known as ENm008 located at the left telomere of human chr16, has previously been studied to decipher the relationship between chromatin structure and transcription activity (37,51,52). RNA-seq data (68–70) indicate that the α-globin genes (including ζ-, μ-, α2-, α1-, and θ-globin genes) are expressed in K562 cell lines but silenced in GM12878 (*tracks* on the *left side of the Hi-C heatmaps* in Fig. 3 *A*). According to 3C/5C measurements (51,71), the α-globin gene forms long-range looping interactions with multiple regulatory elements upon gene activation. Among them, of particular interest is one of the DNase I-hypersensitive sites (DHSs), HS40, located at ~70 kb upstream of the α1 gene.

The HLM-generated structural ensemble at 5-kb resolution for ENm008 of two cell lines (K562 and GM12878) suggests that the contact probability $P(s)$ decreases slightly faster in K562 than in GM12878 cells at large s (Fig. 3 *B*). The α-globin domains of K562 and GM12878 cell lines visualized with FISH (51) indicate that K562 is less compact than GM12878, which is confirmed straightforwardly by the compactness calculated using the HLM-generated structures (Fig. 3 *C*). Compared with GM12878 cells, the α-globin domain in K562 cells adopts a less spherical shape (Fig. 3 *D*; (51,52)) (see also Fig. S10, where individual loci are classified into different groups based on their 3D coordinates, clarifying the spatially separated domains of K562 cells).

Next, we examined the changes in the distances between the α1-globin gene and other loci upon activation of the gene. Even though the whole domain in K562 cells is relatively more expanded, HS40 is closer to the α1 gene in K562 than in GM12878 cells (Fig. 3 *E*), which is consistent with the expectation based on the higher contact enrichment between HS40 and the α1 gene observed in K562 by 3C/5C measurements (e.g., Fig. 2 in (51)). Through inter-cell-line comparison between K562 and GM12878 for the rest of the region using distance distribution to the α-globin gene locus, we identified a group of loci other than HS40 that are significantly closer to α-globin genes in K562 cells (Mann-Whitney U test, $p < 1 \times 10^{-5}$). Their genomic positions are marked using red sticks in Fig. 3 *E*. According to the independent chromatin interaction analysis by paired-end tag sequencing experiments (49,50) designed to capture the chromatin loop interactions mediated by specific protein factors, the structural variation associated with α-globin genes is mainly orchestrated by Pol II (see Table S4). HLM captures 83% of Pol-II-mediated chromatin loops specific to K562 cells (Fig. 3 *F*).

These results indicate that the HLM captures both the tissue-specific variation in the global packing of the α-globin gene domain and variation in the structure of the gene locus. The multiple K562-specific interactions, substantiated by the HLM, suggest that a cooperative action of multiple regulatory elements, including HS40, is responsible for the activation of α-globin genes (37). HLM-generated conformations indeed confirm the notion of chromatin globule proposed in (51).

SOX2 gene

As another example of transcription-dependent chromatin folding, we studied the human SOX2 gene locus, which encodes a transcription factor involving the regulation of embryonic development. The SOX2 gene is transcribed in human embryonic stem cells (hESCs) but not in umbilical vein epithelial cells (HUVECs) (Fig. 4 *A*). To compare the results from the HLM with a recent modeling study (32), we measured the distances between the SOX2 gene and two possible regulatory elements located at regions ~800 kb upstream and ~650 kb downstream. Whereas both elements are closer to the SOX2 locus in transcriptionally active hESCs than in inactive HUVECs, the chromatin fiber is less compact in hESCs (Fig. 4 *D*; see also the *snapshots* in

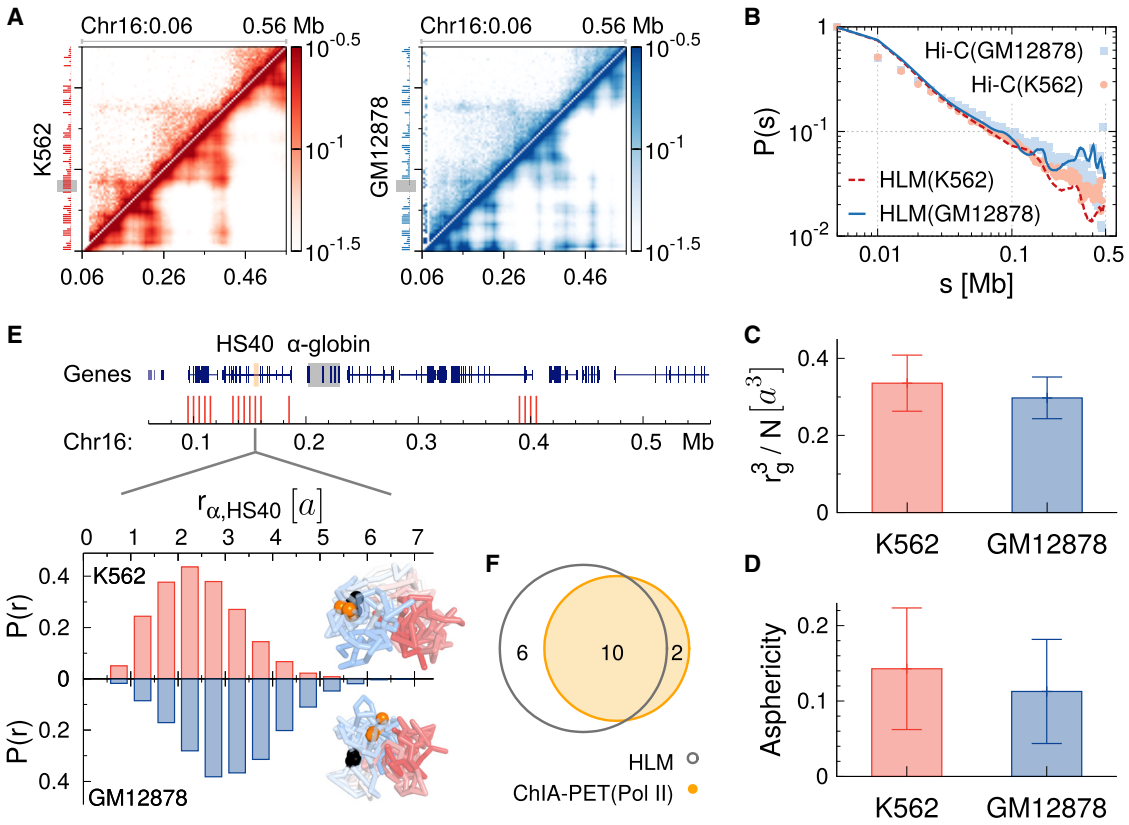

FIGURE 3 α-Globin gene domain modeled by the HLM for two different cell lines. (*A*) Shown are the heatmap ($\mathcal{P}$) of contact probabilities measured by Hi-C (*upper diagonal part*) and the corresponding map ($\tilde{\mathcal{P}}$) obtained from the HLM (*lower diagonal part*) in K562 (*left*) and GM12878 (*right*) cells. RNA-seq signals (68) are displayed on the left side of the heatmaps, and the location of the α-globin gene is depicted in gray shading. (*B*) Mean contact probability $P(s)$ is plotted. (*C*) Compactness and (*D*) asphericity of the domain are shown. (*E*) Genomic positions of the loci, closer to the α1 gene in K562 than in GM12878 cells, are labeled using red sticks. Contrasted below are the distance distributions between the α1 gene and HS40, $P(r_{\alpha,\mathrm{HS40}})$, for two cell lines. For each cell line, an ensemble of structures is shown for comparison with chains colored by the genomic position from the telomere (*blue*) to centromere (*red*). The α-globin gene and HS40 are rendered using a black and an orange sphere, respectively. (*F*) Pol-II-mediated chromatin interactions (49), involving α-globin genes and specific to K562 cells, are compared with the model. To see this figure in color, go online.

Fig. 4, *E* and *F*). HLM-generated structures demonstrate the dependence of chromatin folding on the transcription level at SOX2 gene loci, and this trend comports well with the prediction made in (32), which also employed polymer model simulation.

Chromatin interaction at Pax6 gene loci

The efficacy of the HLM was further tested for the genomic loci of the Pax6 gene, which involve the development of mouse neural tissues. Flanked by two neighboring genes (Pax6os1 and Elp4), the expression level of the Pax6 gene is considered to be regulated by multiple long-range elements, including two regulatory regions located at ~50 kb upstream (URR) and ~95 kb downstream (DRR) (Fig. 5 *A*). The DRR contains several DHSs and the SIMO enhancer, which was identified in transgenic reporter gene studies of developing mouse embryos (72,73). Another *cis*-regulatory element within the URR, PE3, has recently been identified from mouse pancreatic β cells (β-TC3) (74).

A study combining Capture-C, FISH, and simulations (32) has reported a nontrivial correlation between the expression level of the Pax6 gene and the spatial separation from the Pax6 gene to the URR and DRR. Among the three types of mouse cells (β-TC3, MV+, and RAG cells) studied in (32), the Pax6 gene maintained the largest separation from the DRR in the β-TC3 cells that displayed the highest expression level of Pax6. Therefore, it was suggested (32) that the enhancer at the DRR is not involved in upregulation of Pax6 in β-TC3 cells or that some unclear upregulation mechanisms that do not require the spatial proximity to enhancers are responsible for the activity of the Pax6 gene.

To study the origin of the possible complex interplay between the Pax6 gene and neighboring genetic elements, we applied the HLM to the same genomic region of five different mouse cell types whose Hi-C data are currently available: 1) embryonic stem cells (mESCs), 2) neural progenitors (NPCs), 3) cortical neurons (CNs), 4) ncx_NPCs, and 5) ncx_CNs, with the prefix "ncx_ " indicating that the cells are directly purified from the developing mouse

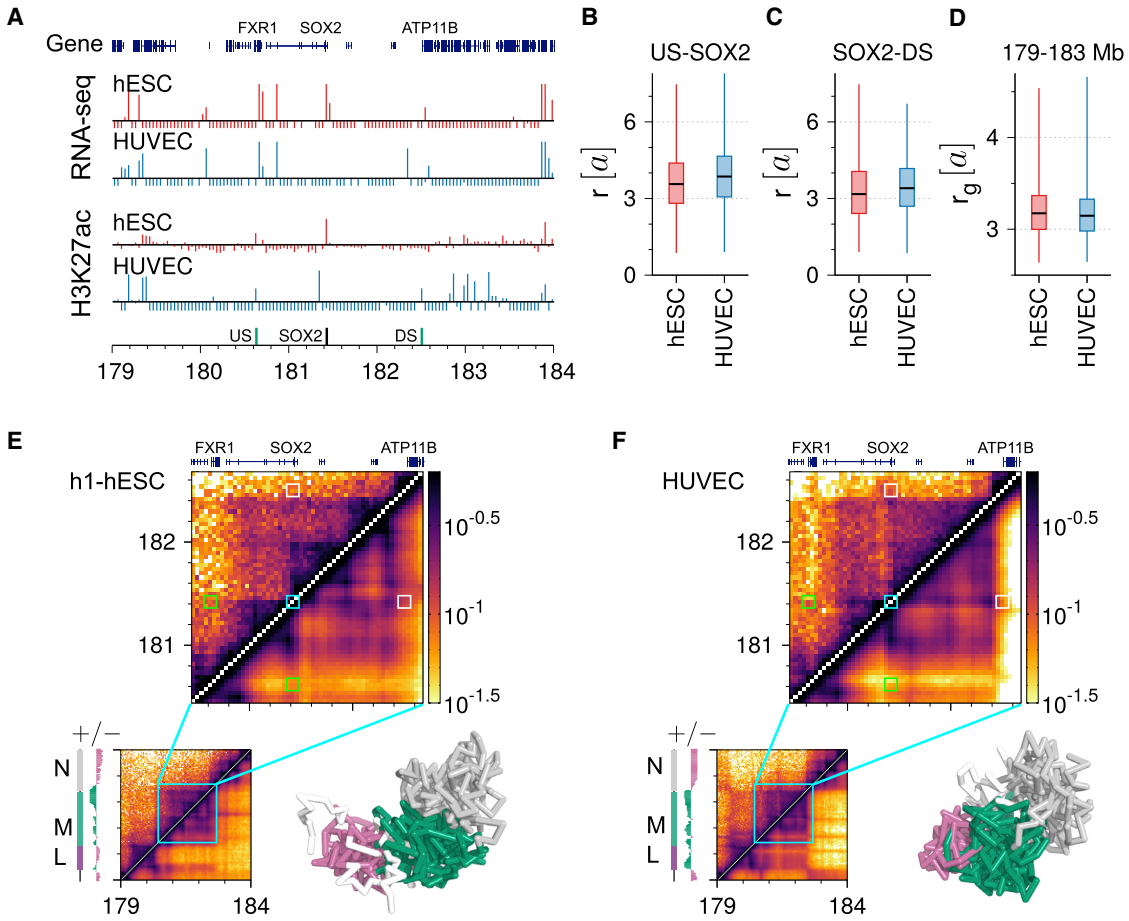

FIGURE 4 Comparison of a 5-Mb genomic region on chr3 modeled by the HLM between hESCs and HUVECs, which includes the SOX2 gene. (*A*) Genes annotated in this region are aligned with RNA-seq (68) and H3K27ac signals (89) of two cell lines. The genomic positions of three "simulated" FISH probes (32) are labeled in the bottom track. (*B*) The distance between upstream and SOX2, (*C*) the distance between SOX2 and downstream, and (*D*) the gyration radius calculated from our model are given. (*E*) A heatmap of contact probabilities for hESCs measured by Hi-C (88) (*upper diagonal part*) and calculated from the HLM (*lower diagonal part*) is displayed. Based on the first principal component of the significance matrix (*track* on the *left side* of heatmap), we divided the region into three domains and colored the chromatin chain accordingly in the snapshot of a typical structural ensemble. (*F*) Analysis was carried out for HUVECs with Hi-C data from (3). To see this figure in color, go online.

embryonic neocortex in vivo. Each cell type displays distinct transcriptional activity patterns of Pax6 and its neighboring genes (75) (Fig. 5 *A*). According to the fragments per kilobase of transcript per million scores from RNA-seq analysis, which is higher when the gene transcription is more active, the five cell types display Pax6 activity in the following order: ncx_NPC > NPC > CN > ES > ncx_CN (Fig. 5 *B*).

The contact probabilities calculated from our HLM-generated conformations reasonably reproduce the Hi-C data at 8-kb resolution (75) (see Fig. S11; Table 1). The Hi-C contact profiles of three genomic loci (URR, Pax6, and DRR) with other genomic regions (*histograms* in Fig. 5 *C*) are well-captured by HLM-generated conformations (*lines* in Fig. 5 *C*). Compared with the distance of Pax gene promoter (P) to the upstream enhancer (UE), Pax6 gene activity is better correlated with the distance to the downstream enhancer (DE) (see Fig. 5 *D*); the closer to the DE, the higher the Pax gene activity is. The highest Pax gene activity is seen in ncx_NPCs. Notice that the most enriched Hi-C contacts between Pax6 and DRR are indeed found in ncx_NPCs, which is marked with a red star in Fig. 5 *C*. We note that our finding on contacts between Pax6 and the DRR using a different set of cell lines differs from the result based on β-TC3 cells (see Fig. 2 *A* in (32)). This, however, underscores that the mechanism or the chromatin conformations responsible for the Pax6 gene activity depend strongly on the cell type. It is clear that the mechanism of Pax6 gene regulation in ncx_NPCs differs from that in β-TC3 cells.

Next, given that Hi-C data are obtained from a collection of millions of cells, heterogeneity of chromatin conformations is inevitable in analyses, which has indeed been highlighted in (32). To characterize the heterogeneity in the HLM-generated conformational ensembles, we classified each chromatin structure into five groups based on the separations between

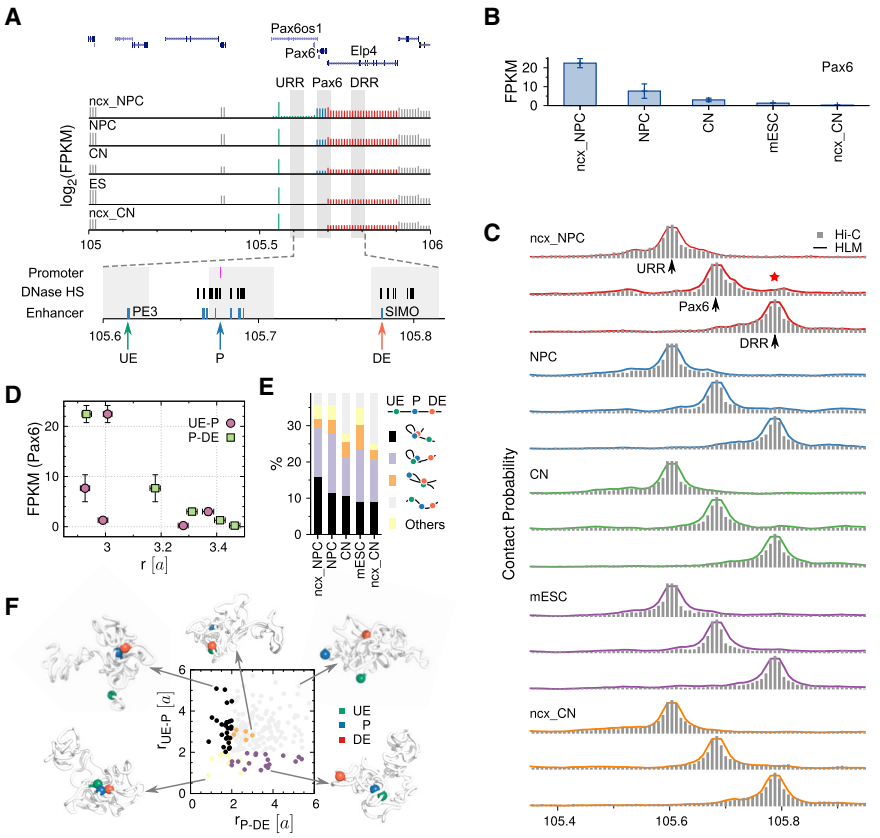

FIGURE 5 Pax6 gene locus modeled by the HLM for five different mouse cell types. (*A*) Genes in a 1-Mb region on chr2, where the genomic positions of the URR, Pax6, and the DRR are labeled with gray shading, are shown in alignment with the fragments per kilobase of transcript per million score measured from RNA-seq analysis (75). Pax6 gene promoters and enhancers and nearby DHSs are zoomed in at the bottom (74,90), where the positions of the upstream enhancer (UE), promoter (P), and downstream enhancer (DE) are marked with arrows. (*B*) Expression levels of the Pax6 gene are provided for different cell types. (*C*) Contact profiles for five different cell types are shown. The profiles were constructed using Hi-C data (*gray bars*) for the URR, Pax6, and the DRR (from *top* to *bottom*) relative to other genomic regions and calculated using the HLM (*solid lines*). (*D*) Pax6 expression level is shown as a function of the average distance between two enhancers (UE and DE) and the promoter (P). (*E*) Percentage of chromatin conformations belonging to each group classified based on the distances between UE, P, and DE is shown. (*F*) Shown are the scatter plot of the distances r_{P-DE} and R_{UE-P} of 200 structures, which were randomly selected from the conformational ensemble of ncx_NPCs. Typical structures are presented for each group in which the three sites are labeled using different colors. To see this figure in color, go online.

the Pax6 gene promoter (P) and two enhancers (UE and DE) (Fig. 5 *E*). To visualize the conformational diversity, we randomly selected 200 structures and characterized them by the promoter-enhancer distances (Fig. 5 *F*). Except for the "gray" group, in which all three separations are large, the population of conformational ensemble consists mainly of the "black" group (P is close to DE but not to UE) and the "purple" group (P is close to UE but not to DE), which are suspected to be responsible for the high expression level of the Pax6 gene. Consistent with our analysis of the ensemble-averaged distance to enhancers for different cells (Fig. 5 *D*), the proportion of the "black" group shows a decreasing trend as Pax6 becomes less active (Fig. 5 *E*), suggesting a more important role of DE than UE in regulating the Pax6 gene for the five cell lines.

Although an indirect upregulation of Pax6 gene by DRR as seen in β-TC3 cells (32) cannot entirely be ruled out, the correlation of gene activity level with the spatial proximity of the Pax6 gene to the DRR is clearly demonstrated, at least across the five cell lines that we studied using the HLM. The mechanism of indirect upregulation and the mechanism of cell-type-dependent choices deserve further study.

Chromosome in different phases of the cell cycle

Most Hi-C data are obtained over a population of "unphased" cells. Here, we employ the HLM to model the global architecture of chromosome at different phases of the cell cycle during the interphase based on single-cell Hi-C (4). Accumulating the data from tens to hundreds of binary contact matrices of single cells into an input matrix $\mathcal{P}$, we built a 500-kb-resolution model of chromosome for the post-M, early-S, mid-S, late-S/G2, and pre-M phases of chr19 in mESCs (*above the diagonal* in Fig. 6 *A*). $\tilde{\mathcal{P}}$ matrices computed using the HLM (*below the diagonal* in Fig. 6 *A*) display reasonable correlation with the original Hi-C data (PC $>$ 0.9), except for the post-M phase (PC $=$ 0.77); unlike other phases, the lower PC value with the $\tilde{\mathcal{P}}$-matrix at the post-M phase, characterized with a uniform and featureless pattern, is due to the smaller number of sampling cells (N_c).

The local compactness of the chromosome conformation was quantified in terms of the average volume occupied by a single monomer ($\overline{v} = V/N$) based on the Voronoi tessellation (Fig. 6 *B*). After mitosis, the chromosome continues to expand until the late-S/G2 phase. The gyration radius also captures this trend (Fig. 6 *B*), except that the model has the largest value of r_g in the post-M phase. A partial condensation of the chain (decreases in r_g^3 and $\overline{v}$) is observed before entering the pre-M phase. This decondensation-condensation cycle is also captured with the asphericity of structures generated from the HLM (Fig. 6 *C*), which decreases dramatically from the post-M to G1 phases and then increases gradually after the G1 phase. The same

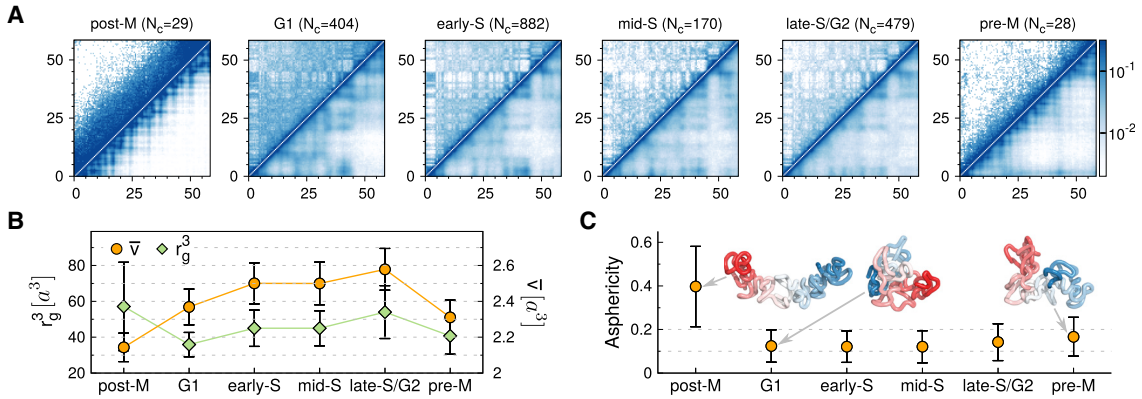

FIGURE 6 Chr19 of mESCs modeled by the HLM at 500-kb resolution. (*A*) A heatmap of contact probabilities from Hi-C (*upper diagonal part*) and HLM (*lower diagonal part*) is given. From post-M phase to pre-M phase, Pearson correlations (PCs) are 0.77, 0.96, 0.96, 0.96, 0.97, and 0.91, respectively. The Hi-C matrices are the outcomes from a sum of N_c binary contact matrices of single cells in the same phases of the cell cycle. (*B*) Plotted are r_g^3 and the average volume ($\bar{v}$) occupied by a single monomer. (*C*) Asphericity of the chromosome in different phases along the cell cycle is calculated. Depicted at post-M, G1, and pre-M phases are the snapshots of the HLM-generated structure, which are colored from the centromere (*blue*) to telomore (*red*). To see this figure in color, go online.

conclusion can be drawn from the probability density of pairwise distance between monomers (see Fig. S12).

DISCUSSION

The HLM is similar to previous polymer models of chromatin that also convert information on spatial proximity into effective harmonic restraints between monomers (25,76,77). In fact, our use of harmonic potential is based on our observation that the pairwise loci distance distributions measured in many FISH experiments (23,36,78,79) are reasonably represented by the variations under harmonic restraints. For example, the distance distribution between seven pairs of FISH probes in mESCs (36) can be reasonably represented by the probability density of the pairwise distance of the HLM (Fig. S1 *C*). Of course, we cannot rule out the possibility that the cell population is too heterogeneous to capture by using single harmonic restraint.

The HLM adopts a "mean-field" approach of using a population-sampled Hi-C map as the sole input data. Fundamental concerns as to the use of single-input data in solving the inverse problem can still be raised to many modeling approaches employing information such as epigenetic marks and DNA accessibility, which are also population-averaged, not single-cell based. Nevertheless, the nature of contact pairs is still probabilistic, giving rise to variations in pairwise distances (Fig. S1 *A*). More importantly, topological and energetic frustrations that arise from the competition between the chain connectivity and long-range interaction defined in Hi-C data are inherent in the polymeric system (80). It is generally not possible to obtain a single chromatin structure that satisfies all the probabilistic constraints given in the Hi-C map. As a result of computationally solving the inverse problem of inferring 3D structures from population-sampled Hi-C data, we always observe structural heterogeneity in the chromosome conformation ensemble (e.g., see Fig. S7). Of course, it is in principle questionable whether or not such a heterogeneous structural ensemble acquired from a Hi-C-map-based HLM represents the true heterogeneous population of chromosome; however, as demonstrated in this study, using the HLM, we can still extract an amount of meaningful information that can complement diverse experimental measurements.

To demonstrate that the choice of energy potential in HLM is optimal over similar alternatives, we examined HLM and its three variants on a 10-Mb genomic region on chr5 of GM12878 cells (Fig. S13). Unlike the HLM, which faithfully reproduced the domain edges of enriched contacts observed by Hi-C (highlighted by *cyan boxes* in Fig. S13 *A*) that were regarded as a distinct feature of loop extrusion (14), two alternative copolymer models, which retain uniform strength of loop interaction, could not properly reproduce the diagonal-block patterns of Hi-C data (Fig. S13, *B* and *C*). In a homopolymer model, in which $\chi_{-,-}$, $\chi_{-,+}$, and $\chi_{+,+}$ are all set to 1 (see Methods), the long-range checkboard pattern was not reproduced (Fig. S13 *D*). The PC of contact probabilities contrasted between Hi-C and other models at different genomic separation, PC(s), shows that HLM outperforms others (Fig. S13 *E*).

Di Stefano et al. have performed steered molecular dynamics simulations of a polymer model of the whole genome of hESC and IMR90 cells, based on physical restraints derived from Hi-C (21). Their model features the nuclear positioning of different chromosomes and functional genomic regions observed in vivo. To compare two models, we computed the Kendall rank correlation between the Hi-C contact matrix and the HLM distance matrix in the simulated genomic regions of both cell types (Fig. S13 *F*). The Kendall rank correlation value gets closer to -1 as the correlation between the model and Hi-C increases. In

comparison with steered molecular dynamics, structures generated with the HLM are better correlated with Hi-C, especially at short genomic distance.

The minimal chromosome model (MiChroM) is another interesting Hi-C-based polymer model (34), which we have used in our previous study (35). We simulated chr5 of GM12878 cells at 50-kb resolution with MiChroM (Fig. S13 *G*) and compared PC(s) of a 10-Mb region based on both models (Fig. S13 *H*). Even though MiChroM considers six types of monomers to describe nonbonded interactions, the HLM outperforms MiChroM in terms of short-range correlations. It also shows comparable long-range correlations with Hi-C.

In addition to the overall PCs listed in Table 1, we calculated PC(s) for all the genomic regions discussed in this study (Fig. S14). Compared with the modeling based on ensemble Hi-C data (Fig. S14, *A* and *B*), the model of chr19 of mESC based on single-cell Hi-C shows lower correlation (Fig. S14 *C*). We found that the PC in general decreases at large genomic separation, but there are two groups of exceptional cases in which the models maintain high correlation with Hi-C at large value of s. The first group is the model of IMR90 cells, which has the lowest model resolution (i.e., the largest genomic size of each monomer) among the human genome models in Fig. S14 *A*. The second group is the model of mouse neuron cells (Fig. S14 *B*), in which the input Hi-C library has higher depth of coverage than the Hi-C data used in others (75). These results suggest that the quality of the resulting structures depends on the accuracy of Hi-C contact probabilities, which can be improved by lowering the model resolution and choosing ultradeep Hi-C. For a specific genomic region of interest, one may improve the model quality further by fine-tuning the value of χ.

As shown for different chromosomes, cell types, and species with a flexible choice of model resolution, one of the greatest advantages of the HLM is its versatile application. Although all of the output conformations exhibit great variability (see discussions in Figs. S7 and 5 *F*), the population-sampled contact map faithfully reproduces the input Hi-C data. For a given Hi-C data set, the two sets of model parameters $\tilde{\mathcal{K}}$ and $\{\chi_{t_i,t_j}\}$ can be determined in a few minutes using a personal computer without any manual intervention (Table 1).

In summary, we demonstrated that the HLM is a computationally efficient approach with which to investigate the genome function. The conformational ensemble generated by the HLM shows that depending on the chromatin states, different types of chromatin domains have different compactness and shapes, and spatial phase separation between domains takes places in human genome. The inter-cell-line comparison of human α-globin and SOX2 loci shows that although the submegabase gene domain becomes less compact upon gene activation, the most critical regulatory element comes closer to the gene, and that its expression is likely affected by many other elements. The activity of the Pax6 gene during mouse neuron development is mostly modulated by the proximity between Pax6 promoter and the DE, whereas the distance to the upstream regulatory element shows nonmonotonic variations with its activity for the cell types we studied. The HLM was also used to visualize the cell-cycle dynamics of chromosome organization based on single-cell Hi-C. Although the HLM is not designed based on assumptions of molecular mechanisms of genome organization, the principle of transcription regulation can be inferred from the changes of chromatin conformations. With Hi-C data being accumulated, the HLM would be of great use to provide complementary structural information that is not easily accessible to current experiments.

SUPPORTING MATERIAL

Supporting Material can be found online at https://doi.org/10.1016/j.bpj.2019.06.032.

AUTHOR CONTRIBUTIONS

L.L., M.H.K., and C.H. designed and performed the research, analyzed the data, and wrote the manuscript.

ACKNOWLEDGMENTS

We thank the Center for Advanced Computation in Korea Institute for Advanced Study for providing computing resources.

C.H. acknowledges a partial support from the National Research Foundation of Korea (NRF-2018R1A2B3001690).

SUPPORTING CITATIONS

References (81–87) appear in the Supporting Material.

REFERENCES

1. Dekker, J., K. Rippe, …, N. Kleckner. 2002. Capturing chromosome conformation. *Science.* 295:1306–1311.
2. Lieberman-Aiden, E., N. L. van Berkum, …, J. Dekker. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.* 326:289–293.
3. Rao, S. S., M. H. Huntley, …, E. L. Aiden. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell.* 159:1665–1680.
4. Nagano, T., Y. Lubling, …, A. Tanay. 2017. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature.* 547:61–67.
5. Du, Z., H. Zheng, …, W. Xie. 2017. Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature.* 547:232–235.
6. Rivera, C. M., and B. Ren. 2013. Mapping human epigenomes. *Cell.* 155:39–55.
7. Jin, F., Y. Li, …, B. Ren. 2013. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature.* 503:290–294.

8. Leung, D., I. Jung, …, B. Ren. 2015. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature.* 518:350–354.
9. Dryden, N. H., L. R. Broome, …, O. Fletcher. 2014. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res.* 24:1854–1868.
10. Jäger, R., G. Migliorini, …, R. S. Houlston. 2015. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat. Commun.* 6:6178.
11. Baca, S. C., D. Prandi, …, L. A. Garraway. 2013. Punctuated evolution of prostate cancer genomes. *Cell.* 153:666–677.
12. Barbieri, M., M. Chotalia, …, M. Nicodemi. 2012. Complexity of chromatin folding is captured by the strings and binders switch model. *Proc. Natl. Acad. Sci. USA.* 109:16173–16178.
13. Sanborn, A. L., S. S. Rao, …, E. L. Aiden. 2015. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. USA.* 112:E6456–E6465.
14. Fudenberg, G., M. Imakaev, …, L. A. Mirny. 2016. Formation of chromosomal domains by loop extrusion. *Cell Rep.* 15:2038–2049.
15. Bianco, S., D. G. Lupiáñez, …, M. Nicodemi. 2018. Polymer physics predicts the effects of structural variants on chromatin architecture. *Nat. Genet.* 50:662–667.
16. Jost, D., P. Carrivain, …, C. Vaillant. 2014. Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. *Nucleic Acids Res.* 42:9553–9561.
17. Brackley, C. A., J. M. Brown, …, D. Marenduzzo. 2016. Predicting the three-dimensional folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models. *Genome Biol.* 17:59.
18. Wang, S., J. Xu, and J. Zeng. 2015. Inferential modeling of 3D chromatin structure. *Nucleic Acids Res.* 43:e54.
19. Szalaj, P., P. J. Michalski, …, D. Plewczynski. 2016. 3D-GNOME: an integrated web service for structural modeling of the 3D genome. *Nucleic Acids Res.* 44:W288–W293.
20. Tjong, H., W. Li, …, F. Alber. 2016. Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proc. Natl. Acad. Sci. USA.* 113:E1663–E1672.
21. Di Stefano, M., J. Paulsen, …, C. Micheletti. 2016. Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization. *Sci. Rep.* 6:35985.
22. Shi, G., L. Liu, …, D. Thirumalai. 2018. Interphase human chromosome exhibits out of equilibrium glassy dynamics. *Nat. Commun.* 9:3161.
23. Wang, S., J. H. Su, …, X. Zhuang. 2016. Spatial organization of chromatin domains and compartments in single chromosomes. *Science.* 353:598–602.
24. Boettiger, A. N., B. Bintu, …, X. Zhuang. 2016. Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature.* 529:418–422.
25. Münkel, C., and J. Langowski. 1998. Chromosome structure predicted by a polymer model. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics.* 57:5888–5896.
26. Münkel, C., R. Eils, …, J. Langowski. 1999. Compartmentalization of interphase chromosomes observed in simulation and experiment. *J. Mol. Biol.* 285:1053–1065.
27. Bohn, M., D. W. Heermann, and R. van Driel. 2007. Random loop model for long polymers. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 76:051805.
28. Mateos-Langerak, J., M. Bohn, …, S. Goetze. 2009. Spatially confined folding of chromatin in the interphase nucleus. *Proc. Natl. Acad. Sci. USA.* 106:3812–3817.
29. Hofmann, A., and D. W. Heermann. 2015. The role of loops on the order of eukaryotes and prokaryotes. *FEBS Lett.* 589:2958–2965.
30. Fraser, J., C. Ferrai, …, M. Nicodemi; FANTOM Consortium. 2015. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Syst. Biol.* 11:852.
31. Brackley, C. A., J. Johnson, …, D. Marenduzzo. 2016. Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domains. *Nucleic Acids Res.* 44:3503–3512.
32. Buckle, A., C. A. Brackley, …, N. Gilbert. 2018. Polymer simulations of heteromorphic chromatin predict the 3D folding of complex genomic loci. *Mol. Cell.* 72:786–797.e11.
33. Chiariello, A. M., C. Annunziatella, …, M. Nicodemi. 2016. Polymer physics of chromosome large-scale 3D organisation. *Sci. Rep.* 6:29775.
34. Di Pierro, M., B. Zhang, …, J. N. Onuchic. 2016. Transferable model for chromosome architecture. *Proc. Natl. Acad. Sci. USA.* 113:12168–12173.
35. Liu, L., G. Shi, …, C. Hyeon. 2018. Chain organization of human interphase chromosome determines the spatiotemporal dynamics of chromatin loci. *PLoS Comput. Biol.* 14:e1006617.
36. Giorgetti, L., R. Galupa, …, E. Heard. 2014. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell.* 157:950–963.
37. Gürsoy, G., Y. Xu, …, J. Liang. 2017. Computational construction of 3D chromatin ensembles and prediction of functional interactions of alpha-globin locus from 5C data. *Nucleic Acids Res.* 45:11547–11558.
38. Zhu, G., W. Deng, …, J. Zeng. 2018. Reconstructing spatial organizations of chromosomes through manifold learning. *Nucleic Acids Res.* 46:e50.
39. Li, Q., H. Tjong, …, F. Alber. 2017. The three-dimensional genome organization of *Drosophila melanogaster* through data integration. *Genome Biol.* 18:145.
40. Serra, F., M. Di Stefano, …, M. A. Marti-Renom. 2015. Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS Lett.* 589:2987–2995.
41. Goldbart, P. M., and A. Zippelius. 1993. Amorphous solid state of vulcanized macromolecules: a variational approach. *Phys. Rev. Lett.* 71:2256–2259.
42. Solf, M. P., and T. A. Vilgis. 1995. Statistical mechanics of macromolecular networks without replicas. *J. Phys. Math. Gen.* 28:6655.
43. Bryngelson, J. D., and D. Thirumalai. 1996. Internal constraints induce localization in an isolated polymer molecule. *Phys. Rev. Lett.* 76: 542–545.
44. Zwanzig, R. 1997. Effect of close contacts on the radius of gyration of a polymer. *J. Chem. Phys.* 106:2824–2827.
45. Cacciuto, A., and E. Luijten. 2006. Self-avoiding flexible polymers under spherical confinement. *Nano Lett.* 6:901–905.
46. Kang, H., Y. G. Yoon, …, C. Hyeon. 2015. Confinement-induced glassy dynamics in a model for chromosome organization. *Phys. Rev. Lett.* 115:198102.
47. Gürsoy, G., Y. Xu, …, J. Liang. 2014. Spatial confinement is a major determinant of the folding landscape of human chromosomes. *Nucleic Acids Res.* 42:8223–8230.
48. Zhang, Y., and D. W. Heermann. 2011. Loops determine the mechanical properties of mitotic chromosomes. *PLoS One.* 6:e29225.
49. Li, G., X. Ruan, …, Y. Ruan. 2012. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell.* 148:84–98.
50. Tang, Z., O. J. Luo, …, Y. Ruan. 2015. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell.* 163:1611–1627.
51. Baù, D., A. Sanyal, …, M. A. Marti-Renom. 2011. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.* 18:107–114.
52. Peng, C., L. Y. Fu, …, H. Y. Zhang. 2013. The sequencing bias relaxed characteristics of Hi-C derived data and implications for chromatin 3D modeling. *Nucleic Acids Res.* 41:e183.

53. Meluzzi, D., and G. Arya. 2015. Efficient estimation of contact probabilities from inter-bead distance distributions in simulated polymer chains. *J. Phys. Condens. Matter.* 27:064120.

54. Fudenberg, G., and M. Imakaev. 2017. FISH-ing for captured contacts: towards reconciling FISH and 3C. *Nat. Methods.* 14:673–678.

55. de Gennes, P. G. 1979. Scaling Concepts in Polymer Physics. Cornell University Press, Ithaca and London.

56. Aronovitz, J. A., and D. R. Nelson. 1986. Universal features of polymer shapes. *J. Phys. (Paris).* 47:1445–1456.

57. Hyeon, C., R. I. Dima, and D. Thirumalai. 2006. Size, shape, and flexibility of RNA structures. *J. Chem. Phys.* 125:194905.

58. Rycroft, C. H. 2009. VORO++: a three-dimensional voronoi cell library in C++. *Chaos.* 19:041111.

59. Friedman, B., and B. O'Shaughnessy. 1991. Short time behavior and universal relations in polymer cyclization. *J. Phys. II.* 1:471–486.

60. Hyeon, C., and D. Thirumalai. 2006. Kinetics of interior loop formation in semiflexible chains. *J. Chem. Phys.* 124:104905.

61. Toan, N. M., G. Morrison, …, D. Thirumalai. 2008. Kinetics of loop formation in polymer chains. *J. Phys. Chem. B.* 112:6094–6106.

62. Flory, P. J. 1949. The configuration of real polymer chains. *J. Chem. Phys.* 17:303–310.

63. Grosberg, A. Y., and A. R. Khokhlov. 1994. Statistical Physics of Macromolecules. AIP Press, New York.

64. Halverson, J. D., J. Smrek, …, A. Y. Grosberg. 2014. From a melt of rings to chromosome territories: the role of topological constraints in genome folding. *Rep. Prog. Phys.* 77:022601.

65. Liu, L., and C. Hyeon. 2016. Contact statistics highlight distinct organizing principles of proteins and RNA. *Biophys. J.* 110:2320–2327.

66. Goetze, S., J. Mateos-Langerak, …, R. van Driel. 2007. The three-dimensional structure of human interphase chromosomes is related to the transcriptome map. *Mol. Cell. Biol.* 27:4475–4487.

67. Levine, M., C. Cattoglio, and R. Tjian. 2014. Looping back to leap forward: transcription enters a new era. *Cell.* 157:13–25.

68. Mortazavi, A., B. A. Williams, …, B. Wold. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods.* 5:621–628.

69. Trapnell, C., B. A. Williams, …, L. Pachter. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28:511–515.

70. ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 489:57–74.

71. Vernimmen, D., F. Marques-Kranc, …, D. R. Higgs. 2009. Chromosome looping at the human α-globin locus is mediated via the major upstream regulatory element (HS -40). *Blood.* 114:4253–4260.

72. Kleinjan, D. A., A. Seawright, …, V. van Heyningen. 2001. Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6. *Hum. Mol. Genet.* 10:2049–2059.

73. Bhatia, S., H. Bengani, …, D. A. Kleinjan. 2013. Disruption of autoregulatory feedback by a mutation in a remote, ultraconserved PAX6 enhancer causes aniridia. *Am. J. Hum. Genet.* 93:1126–1134.

74. Buckle, A., R. S. Nozawa, …, N. Gilbert. 2018. Functional characteristics of novel pancreatic Pax6 regulatory elements. *Hum. Mol. Genet.* 27:3434–3448.

75. Bonev, B., N. Mendelson Cohen, …, G. Cavalli. 2017. Multiscale 3D genome rewiring during mouse neural development. *Cell.* 171:557–572.e24.

76. Fritsche, M., S. Li, …, P. A. Wiggins. 2012. A model for *Escherichia coli* chromosome packaging supports transcription factor-induced DNA domain formation. *Nucleic Acids Res.* 40:972–980.

77. Di Stefano, M., A. Rosa, …, C. Micheletti. 2013. Colocalization of coregulated genes: a steered molecular dynamics study of human chromosome 19. *PLoS Comput. Biol.* 9:e1003019.

78. Finn, E. H., G. Pegoraro, …, T. Misteli. 2017. Comparative analysis of 2D and 3D distance measurements to study spatial genome organization. *Methods.* 123:47–55.

79. Szabo, Q., D. Jost, …, G. Cavalli. 2018. TADs are 3D structural units of higher-order chromosome organization in Drosophila. *Sci. Adv.* 4:eaar8082.

80. Thirumalai, D., and C. Hyeon. 2005. RNA and protein folding: common themes and variations. *Biochemistry.* 44:4957–4970.

81. Knight, P. A., and D. Ruiz. 2013. A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* 33:1029.

82. Veitshans, T., D. Klimov, and D. Thirumalai. 1997. Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Fold. Des.* 2:1–22.

83. Limbach, H. J., A. Arnold, …, C. Holm. 2006. ESPResSo – an extensible simulation package for research on soft matter systems. *Comput. Phys. Commun.* 174:704–727.

84. Yang, T., F. Zhang, …, Q. Li. 2017. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.* 27:1939–1949.

85. Ester, M., H.-P. Kriegel, …, X. Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *In* Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. E. Simoudis, J. Han, and U. Fayyad, eds. The AAAI Press, pp. 226–231.

86. Pedregosa, F., G. Varoquaux, …, E. Duchesnay. 2011. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12:2825–2830.

87. Kuhn, R. M., D. Haussler, and W. J. Kent. 2013. The UCSC genome browser and associated tools. *Brief. Bioinform.* 14:144–161.

88. Dixon, J. R., S. Selvaraj, …, B. Ren. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature.* 485:376–380.

89. Ernst, J., P. Kheradpour, …, B. E. Bernstein. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature.* 473:43–49.

90. Pruitt, K. D., G. R. Brown, …, J. M. Ostell. 2014. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* 42:D756–D763.

Biophysical Journal, Volume *117*

Supplemental Information

Heterogeneous Loop Model to Infer 3D Chromosome Structures from Hi-C

Lei Liu, Min Hyeok Kim, and Changbong Hyeon

SUPPLEMENTARY INFORMATION

Hi-C data preparation

The contact frequency matrix obtained from a population of cells, the raw data of which are listed in Table 1, was first normalized by using the Knight-Ruiz (KR) method [1], so that the sum of each row and column of the matrix is unity. We then rescaled the KR-normalized matrix so that it satisfies $P(s) = 1$ at $s = 1$, where s is the genomic separation in the unit of the model resolution, and used it as the input contact probability matrix $\mathcal{P}$. Since genomic loci have different coordinates in sequence database for different genome assemblies, if necessary, we converted the genomic coordinates of human and mouse by choosing Hg19 and mm10, respectively, as their references.

Pairwise contact probability in heterogeneous loop model

Similar to RLM [2], the harmonic-restraining energy potential of HLM for a chain of N monomers can be written

$$U_{\mathcal{K}}(\mathbf{r}) = \frac{3}{2}\mathbf{r}^T\mathbf{K}\mathbf{r}, \tag{S1}$$

where $\mathbf{r} = (\vec{r}_1, \vec{r}_2, \cdots, \vec{r}_{N-1})^T$ and the translational degrees of freedom was removed by setting $\vec{r}_0 = (0, 0, 0)$. $\mathbf{K}$ is the Kirchhoff matrix:

$$\begin{pmatrix} \sum_{j=0,j\neq 1}^{N-1} k_{1j} & -k_{12} & \cdots & -k_{1,N-1} \\ -k_{21} & \sum_{j=0,j\neq 2}^{N-1} k_{2j} & \cdots & -k_{2,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ -k_{N-1,1} & -k_{N-1,2} & \cdots & \sum_{j=0,j\neq N-1}^{N-1} k_{N-1,j} \end{pmatrix}. \tag{S2}$$

Then, the probability density of the distance between the i and j-th monomer ($i < j$) projected on one dimension is

$$\begin{aligned} P(x_{ij};\gamma_{ij}) &= \langle \delta[x_{ij} - (x_i - x_j)] \rangle \\ &= \int dx_1 \cdots dx_{N-1} \delta\left[x_{ij} - (x_i - x_j)\right] P(\mathbf{x}) \\ &\propto \int_0^\infty dq e^{iqx_{ij}} \int dx_1 \cdots dx_{N-1} e^{-iq(x_i - x_j)} e^{-\frac{1}{2}\mathbf{x}^T\mathbf{K}\mathbf{x}} \\ &\propto \int_0^\infty dq e^{iqx_{ij}} e^{-\frac{q^2}{4\gamma_{ij}}} \\ &\propto e^{-\gamma_{ij}x_{ij}^2}, \end{aligned} \tag{S3}$$

where we have used $\langle \exp\left(\sum_n \xi_n x_n\right)\rangle = \exp\left(\frac{1}{2}\sum_{nm}(K^{-1})_{nm}\xi_n\xi_m\right)$. The value of γ_{ij} depends on the topology of 'vulcanized' polymer chain, dictated by $\mathbf{K}$ matrix, and is related with the covariance between the positions of i and j-th monomer $\sigma_{ij} = \langle \delta\vec{r}_i \cdot \delta\vec{r}_j \rangle$ as follows

$$\gamma_{ij} = \begin{cases} \frac{1}{2(\sigma_{ii}+\sigma_{jj}-2\sigma_{ij})}, & i > 0 \\ \frac{1}{2\sigma_{jj}}, & i = 0 \end{cases} \tag{S4}$$

where $\sigma_{ij} (= (\Sigma)_{ij})$ is the elements of inverse matrix $\Sigma = \mathbf{K}^{-1}$. Finally, the probability density of pairwise distance in 3D is [2]

$$P(r_{ij};\gamma_{ij}) = 4\gamma_{ij}^{3/2}/\sqrt{\pi} r_{ij}^2 e^{-\gamma_{ij}r_{ij}^2}. \tag{S5}$$

Typical profiles of $P(r_{ij};\gamma_{ij})$ for varying γ_{ij} are shown in Fig. S1A with the relation between p_{ij} and γ_{ij} given in the inset.

Eq. S5 enables us to evaluate a few quantities of interest directly. For a pair of monomers in contact with the condition of $r_{ij} < r_c$, the pairwise contact probability, p_{ij}, is given by

$$\begin{aligned} p_{ij} &= \int_0^{r_c} P(r_{ij}) dr_{ij} \\ &= \mathrm{erf}(\gamma_{ij}^{1/2} r_c) - 2r_c\sqrt{\frac{\gamma_{ij}}{\pi}} e^{-\gamma_{ij}r_c^2}, \end{aligned} \tag{S6}$$

with $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x dt e^{-t^2}$. In addition, the mean pairwise distance is

$$\langle r_{ij} \rangle = \int_0^\infty r_{ij} P(r_{ij}) dr_{ij} = \frac{2}{\sqrt{\pi}\gamma_{ij}^{1/2}}, \tag{S7}$$

and the mean square distance equals to

$$\langle r_{ij}^2 \rangle = \int_0^\infty r_{ij}^2 P(r_{ij}) dr_{ij} = \frac{3}{2\gamma_{ij}}. \tag{S8}$$

For $\gamma_{ij}r_c^2 (\equiv g_{ij}) \ll 1$ (or $\langle r_{ij} \rangle \gg r_c$), p_{ij} is approximated as

$$\begin{aligned} p_{ij} &= \frac{2}{\sqrt{\pi}}\Big\{ \int_0^{g_{ij}^{1/2}} [1 - t^2 + \mathcal{O}(t^4)] dt - g_{ij}^{1/2}[1 - g_{ij} + \mathcal{O}(g_{ij}^2)] \Big\} \\ &= \frac{1}{3\sqrt{\pi}} g_{ij}^{3/2} + \mathcal{O}(g_{ij}^{5/2}). \end{aligned} \tag{S9}$$

Replacing g_{ij} with $\langle r_{ij} \rangle$ using Eq. S7, one obtains a scaling relation between p_{ij} and $\langle r_{ij} \rangle$ as

$$p_{ij} \sim \Big[\frac{4r_c^2}{\pi\langle r_{ij}\rangle^2}\Big]^{3/2} \sim \langle r_{ij} \rangle^{-3}. \tag{S10}$$

As shown in Fig. S1B, the scaling $p_{ij}^{-1} \sim \langle r_{ij} \rangle^3$ holds for large $\langle r_{ij} \rangle$.

RLM was developed to understand the scaling of the spatial distance between two genomic loci with respect to their genomic distance [2]. The original RLM assumes that all loops have the same interaction strength (i.e., $k_{ij} = 3$ or $0\ k_BT/a^2$), and the overall compactness of the chain was adjusted by the total number of loops

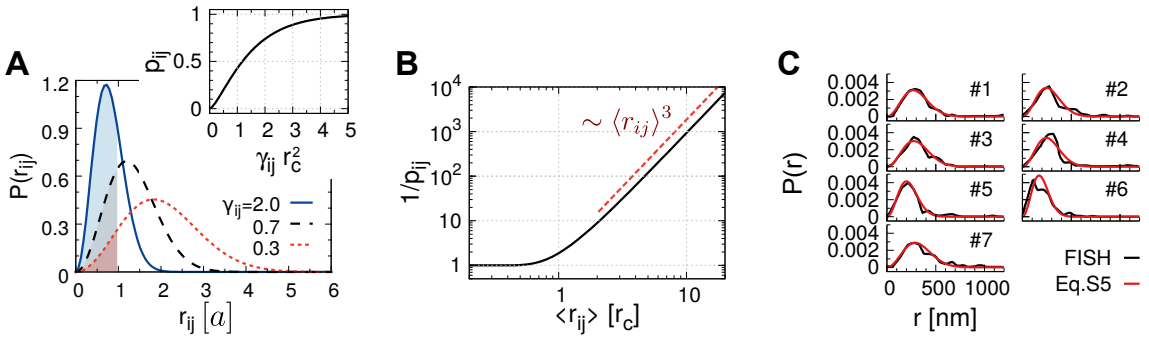

FIG. S1. (A) Probability density of pairwise distance $P(r_{ij})$ at various values of γ_{ij} (Eq. S5), where the inset shows contact probability p_{ij} as a function of $\gamma_{ij}r_c^2$ (Eq. S6). (B) Inverse of p_{ij} is plotted as a function of the mean pairwise distance $\langle r_{ij}\rangle$ in a log-log scale, where it shows $p_{ij}^{-1} \sim \langle r_{ij}\rangle^3$ at $\langle r_{ij}\rangle \gg r_c$ (Eq. S10). (C) $P(r_{ij})$ between seven FISH probe pairs in the *Tsix/Xist* region on chrX of mESC. The experimental data (black lines) were digitized from Fig. 2F in Ref. [3], and their corresponding fits to Eq. S5 are shown in red.

($N_l = \sum_{i>j} \delta(k_{ij} - 3)$). Therefore, any quantity of interest, e.g., $P(r_{ij})$, needs to be averaged over different instances of $\mathcal{K}$ with the same value of N_l. In the worst case, this requires $2^{(N-1)(N-2)/2}$ implementations of $\mathcal{K}$, which renders a precise evaluation of $P(r_{ij})$ impractical. In this study, instead of varying N_l, we relax the constraint on k_{ij} so that it can take any non-negative value.

Inferring interaction strengths by direct inversion

In HLM, p_{ij} increases *monotonically* with γ_{ij} at given r_c (Fig. S1A), which allows one to determine the value of γ_{ij} from p_{ij}. Given a contact probability matrix $\mathcal{P}$ of elements p_{ij}, we can further derive the interaction strength matrix $\mathcal{K}$ of elements k_{ij} in three steps

$$\mathcal{P} \to \Sigma \to \mathbf{K} \to \mathcal{K}. \tag{S11}$$

More specifically, the interaction strength matrix $\mathcal{K}$ for HLM can be obtained from $\mathcal{P}$ via the following steps, which we call the *direct inversion*:

1. Construct the matrix Σ using the relations of diagonal elements $\sigma_{ii} = \frac{1}{2\gamma_{0i}}$ for $i > 0$; and the off-diagonal elements $\sigma_{ij} = \frac{1}{2}\left(\sigma_{ii} + \sigma_{jj} - \frac{1}{2\gamma_{ij}}\right)$ for $i, j > 0, i \neq j$.

2. Invert the matrix ($\Sigma \to \Sigma^{-1}$) to get the Kirchhoff matrix $\mathbf{K}$.

3. Determine the interaction strengths $k_{ij}(= (\mathcal{K})_{ij})$ from (i) $k_{ij} = -(\mathbf{K})_{ij}$ for $i, j > 0, i \neq j$; (ii) $k_{0i} = \sum_j (\mathbf{K})_{ij}$ for $i, j > 0$.

However, even a small sampling error, if any, in the contact probability matrix $\mathcal{P}$ may result in unphysical values of $\mathcal{K}$ with this protocol. We demonstrate this issue clearly using molecular dynamics (MD) simulations of three toy models ($N = 20$) characterized with different intra-chain loops (i.e., different $\mathcal{K}$-matrix): (i) a chain with a single loop (Fig. S3); (ii) a chain with two nested loops (Fig. S4); (iii) a chain composed of two blocks of monomers without any inter-block attraction (Fig. S5). For the case of the single-loop polymer (model (i)), $\mathcal{P}$ estimated based on the conformational ensemble from MD simulation resembles $\mathcal{P}^*$, where the superscript * denotes the *true* value, with a small relative error of 0.018; however, $\mathcal{K}$ obtained from $\mathcal{P}$ using the aforementioned direct inversion gives rise to unphysical matrix elements $k_{ij} < 0$. The same issue ($k_{ij} < 0$) was encountered for the two other cases (models (ii) and (iii)). For the three toy models, we circumvented the issue of $k_{ij} < 0$ resulting from the direct inversion through the constrained optimization, which is explained in the METHOD section in the main text and illustrated along the magenta arrows in Figs. S3-S5.

We note that Hi-C data is still an outcome of sampling over finite number of cell population, and hence a small but finite amount of error is inevitably included in Hi-C data. The same issue arises when the direct inversion is applied to Hi-C. Therefore, in the framework of HLM, we use the constrained optimization to determine $\widetilde{\mathcal{K}}$-matrix that can serve as a proxy of $\mathcal{K}$, which are followed by MD simulations using HLM potential.

Molecular dynamics simulations

To produce a conformational ensemble of chromosome using HLM via the enhanced sampling, we performed the low-friction Langevin simulations by numerically integrating the following equation [4],

$$m\frac{d^2\vec{r}_i}{dt^2} = -\zeta_{\mathrm{MD}}\frac{d\vec{r}_i}{dt} - \vec{\nabla}_{\vec{r}_i}U(\vec{r}_1, \vec{r}_2, \ldots) + \vec{\xi}(t), \tag{S12}$$

We chose a friction coefficient $\zeta_{\rm MD} = 1.0m/\tau_{\rm MD}$ and a time step $\delta t = 0.01\tau_{\rm MD}$ with the characteristic time scale $\tau_{\rm MD} = (ma^2/\epsilon)^{1/2}$. The whole simulation was carried out with three steps. (i) Random Gaussian chains were first equilibrated for 500 $\tau_{\rm MD}$ under the energy potential $U_{\mathcal{K}}(\mathbf{r})$ without nonbonded interaction term. At this stage, excluded volume interaction is absent. (ii) The simulations were performed under the full HLM potential $U_{\rm HLM}(\mathbf{r})$ for 100 $\tau_{\rm MD}$ but with extra care. Excessive overlaps between monomers generated from the foregoing stage, were eliminated by gradually increasing the contribution from the short-range repulsive potential, which was achieved by using the LJ potential term $u_{\rm LJ}(r_{ij}) = \min\{u_c, u_{\rm LJ}(r_{ij})\}$ with gradually increasing u_c. (iii) The production runs were generated for $5 \times 10^5 \tau_{\rm MD}$, during which chain configurations were collected every 50 $\tau_{\rm MD}$. Simulations were all carried out by using ESPResSo 3.3.1 package [5]. For any chromosome ensemble discussed in this work, it took less than 5 hours on a single CPU to generate them.

Quantifying the similarity between contact probabilities from Hi-C and modeling

It has been recently proposed [6] that compared with a global Pearson or Spearman correlation coefficient, the reproducibility of Hi-C data can be better assessed by measuring the Pearson correlation PC(s), at each value of genomic separation s, which minimizes the dependence of contact frequency on the genomic distance (e.g., see Fig. S13E). A underlying assumption for this quantity is that the mean contact probability as a function of genomic distance, $P(s)$, does not change too much. Whereas this is probably true for different replicates in Hi-C experiment, it is not guaranteed for modeling.

Two examples are shown in Fig. S2A and B. Although the contact probability matrix I and II (III) are perfectly correlated at each genomic scale (Fig. S2D), their overall similarity is low due to the different profiles of $P(s)$ (Fig. S2C). With these possible artifacts in mind, we calculated both properties to quantify the similarity of contact probabilities between Hi-C and our model (Table 1 and Fig. S14).

Clustering analysis on chromosomes with conformational variability

To characterize the variability in the conformational ensemble in Fig. 1D quantitatively, we cluster the HLM-generated structures using hierarchical clustering algorithm. We first defined three domains labeled as "L", "M" and "N" based on the sign of the first principle component of $\mathcal{Z}$ matrix (see the left tracks of Fig. 1B). Structures were hierarchically clustered according to structural similarity assessed by the distance-based root-mean-square deviation (dRMSD). For any two structures, say α and β, their similarity was measured by

$$\mathrm{dRMSD}_{\alpha,\beta} = \sqrt{\sum_{\{\mathrm{X\text{-}Y}\}} \frac{1}{3}(r^{\alpha}_{\mathrm{X\text{-}Y}} - r^{\beta}_{\mathrm{X\text{-}Y}})^2}, \qquad \text{(S13)}$$

where $r_{\mathrm{X\text{-}Y}}$ is the distance between the geometric centers of two different domains X and Y (X,Y $\in$ {L,M,N}). The dendrogram depicted in Fig. S7A identifies *at least* 4 main classes of conformations. Chromatins fold into compact globules in the class-1, but adopt elongated conformations in the class-4. The class-2 and -3 can be identified separately from the class-1 and -4 in the 2D phase plane drawn as a function of $r_{\mathrm{L\text{-}M}}$ and $r_{\mathrm{L\text{-}N}}$ (Fig. S7B). Since the structural interconversion among different chromosome conformations is an unusually time-consuming, glass-like process [7], the contact probability matrix $\mathcal{P}$ (Fig. 1E) is in effect an outcome of *quenched*-average [2] over distinct conformations.

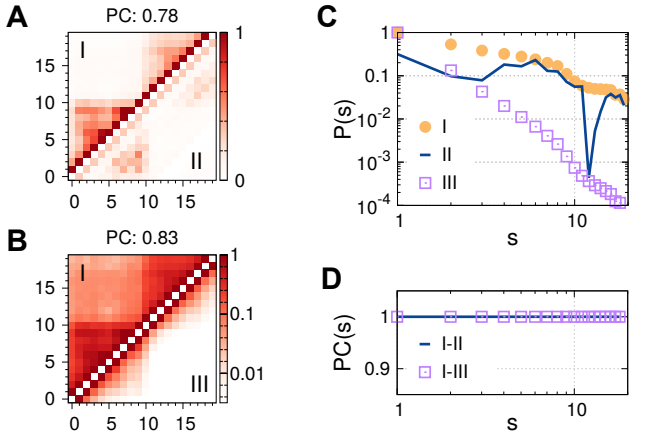

FIG. S2. Similarity between contact probablity matrices. Comparison of contact probability matrices I and II (A), I and III (B), with the overall Pearson correlations labeled on top. (C) Mean contact probability, (D) Pearson correlation as a function of genomic distance.

[1] Knight, P. A., and D. Ruiz, 2013. A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* 33:1029.

[2] Bohn, M., D. W. Heermann, and R. van Driel, 2007. Random loop model for long polymers. *Phys. Rev. E.* 76:051805.

[3] Giorgetti, L., R. Galupa, E. Nora, T. Piolot, F. Lam, J. Dekker, G. Tiana, and E. Heard, 2014. Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription. *Cell* 157:950 – 963.

[4] Veitshans, T., D. Klimov, and D. Thirumalai, 1997. Protein folding kinetics: timescales, pathways and energy

landscapes in terms of sequence-dependent properties. *Folding Des.* 2:1–22.

[5] Limbach, H. J., A. Arnold, B. A. Mann, and C. Holm, 2006. ESPResSo – An Extensible Simulation Package for Research on Soft Matter Systems. *Comput. Phys. Commun.* 174:704–727.

[6] Yang, T., F. Zhang, G. G. Yardimci, F. Song, R. C. Hardison, W. S. Noble, F. Yue, and Q. Li, 2017. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.* 27:1939–1949.

[7] Shi, G., L. Liu, C. Hyeon, and D. Thirumalai, 2018. Interphase Human Chromosome Exhibits Out of Equilibrium Glassy Dynamics. *Nat. Commun.* 9:3161.

[8] Mateos-Langerak, J., M. Bohn, W. de Leeuw, O. Giromus, E. M. M. Manders, P. J. Verschure, M. H. G. Indemans, H. J. Gierman, D. W. Heermann, R. van Driel, and S. Goetze, 2009. Spatially confined folding of chromatin in the interphase nucleus. *Proc. Natl. Acad. Sci. USA* 106:3812–3817.

[9] Ester, M., H.-P. Kriegel, J. Sander, and X. Xu, 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *In* Proc. of 2nd International Conference on Knowledge Discovery and Data Mining. 226–231.

[10] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12:2825–2830.

[11] Baù, D., A. Sanyal, B. R. Lajoie, E. Capriotti, M. Byron, J. B. Lawrence, J. Dekker, and M. A. Marti-Renom, 2011. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.* 18:107–114.

[12] Di Stefano, M., J. Paulsen, T. G. Lien, E. Hovig, and C. Micheletti, 2016. Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization. *Sci. Rep.* 6:35985.

[13] Di Pierro, M., B. Zhang, E. L. Aiden, P. G. Wolynes, and J. N. Onuchic, 2016. Transferable model for chromosome architecture. *Proc. Natl. Acad. Sci. USA* 113:12168–12173.

[14] Wang, S., J.-H. Su, B. J. Beliveau, B. Bintu, J. R. Moffitt, C.-t. Wu, and X. Zhuang, 2016. Spatial organization of chromatin domains and compartments in single chromosomes. *Science* 353:598–602.

[15] Kuhn, R. M., D. Haussler, and W. J. Kent, 2013. The UCSC genome browser and associated tools. *Brief Bioinform* 14:144–161.

[16] Li, G., X. Ruan, R. Auerbach, K. Sandhu, M. Zheng, P. Wang, H. Poh, Y. Goh, J. Lim, J. Zhang, H. Sim, S. Peh, F. Mulawadi, C. Ong, Y. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C.-L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. Fullwood, E. Cheung, E. Liu, W.-K. Sung, M. Snyder, and Y. Ruan, 2012. Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. *Cell* 148:84–98.

[17] Tang, Z., O. Luo, X. Li, M. Zheng, J. Zhu, P. Szalaj, P. Trzaskoma, A. Magalska, J. Wlodarczyk, B. Ruszczycki, P. Michalski, E. Piecuch, P. Wang, D. Wang, S. Tian, M. Penrad-Mobayed, L. Sachs, X. Ruan, C.-L. Wei, E. Liu, G. Wilczynski, D. Plewczynski, G. Li, and Y. Ruan, 2015. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* 163:1611–1627.

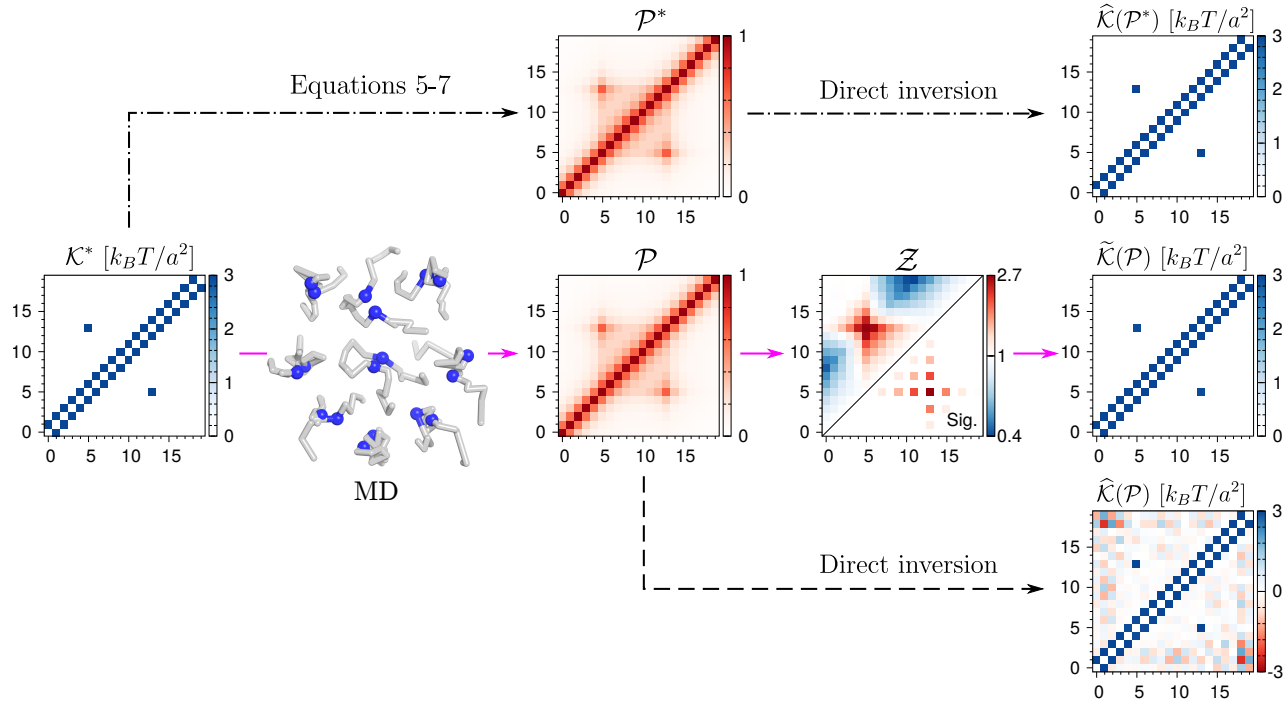

FIG. S3. HLM illustrated by a model of single-loop polymer chain. Starting from a true spring strength matrix $\mathcal{K}^*$, one can obtain either the contact probability matrix $\mathcal{P}^*$ analytically, or $\mathcal{P}$ (as an estimate of $\mathcal{P}^*$) numerically. Through the direct inversion (Eq. S11), $\mathcal{P}^*$ reproduces $\widehat{\mathcal{K}}(\mathcal{P}^*) = \mathcal{K}^*$, but $\mathcal{P}$, which is similar to $\mathcal{P}^*$ except for small deviation, generates $\widehat{\mathcal{K}}(\mathcal{P})$ that contains unphysical negative elements due to numeric errors. By contrast, $\mathcal{K}^*$ can be still inferred from $\mathcal{P}$, $\widetilde{\mathcal{K}}(\mathcal{P}) \approx \mathcal{K}^*$ with a relative error of 8×10^{-4}, using the constrained optimization (Eq. 9) on the significant contacts selected from $\mathcal{Z}$ (lower diagonal region).

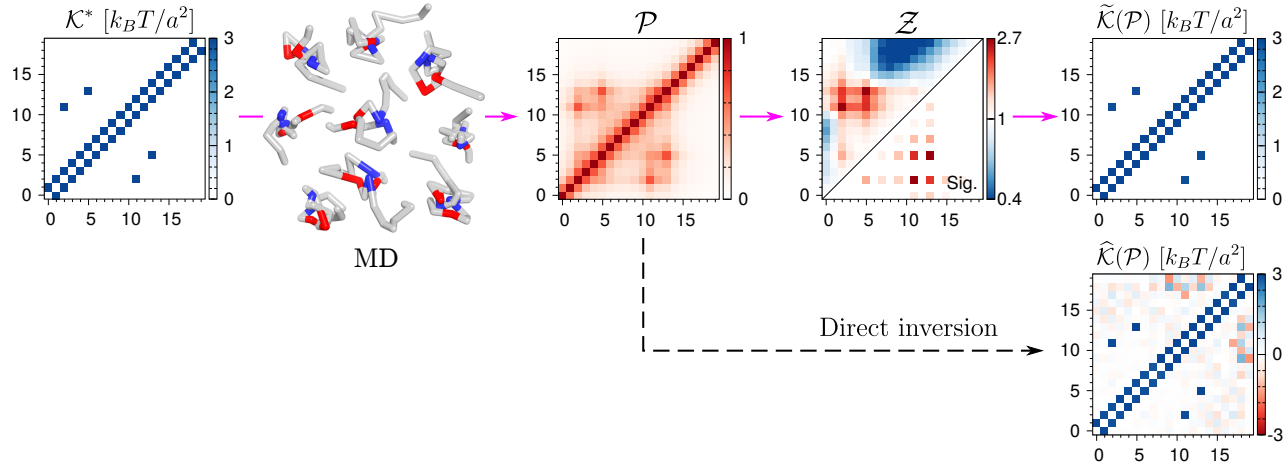

FIG. S4. HLM illustrated by a model of two-nested-loop polymer chain. The chain is composed of 20 monomers. One loops is anchored between the second and 11-th monomers, and the other is between the 5-th and 13-th monomers. In contrast to the direct inversion which generates unphysical negative elements in $\widehat{\mathcal{K}}(\mathcal{P})$, constrained optimization leads to $\widetilde{\mathcal{K}}(\mathcal{P})$, that is similar to $\mathcal{K}^*$ with a relative error of 0.002.

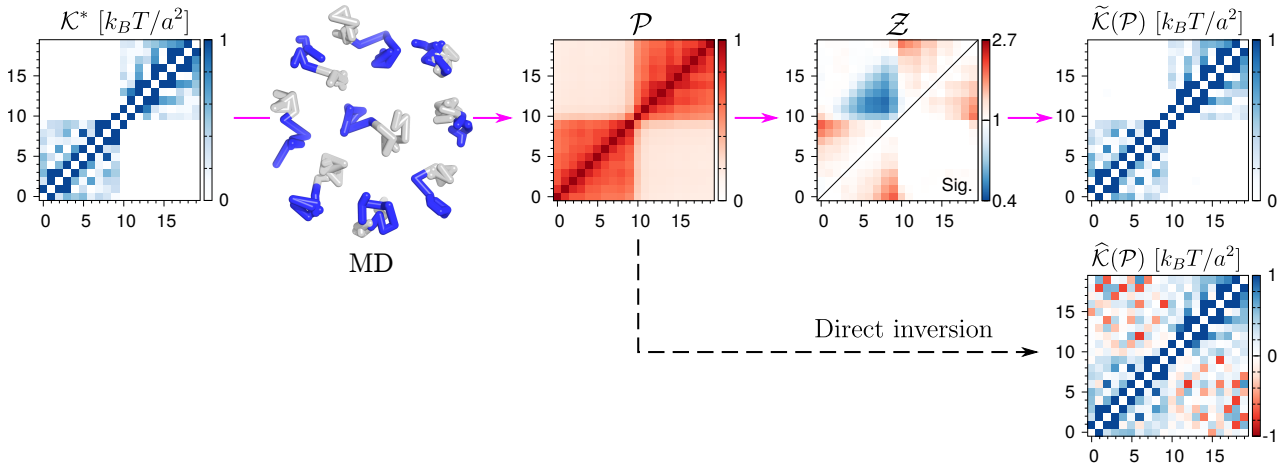

FIG. S5. HLM illustrated by a model of diblock copolymer chain. The chain is composed of 20 monomers. $\mathcal{K}^*$-matrix indicates the presence of heterogeneous intra-block attractions within the first $0 \le i,j < 10$ and the second block $10 \le i,j < 20$, but there is no such interaction between the two blocks. Compared to $\widetilde{\mathcal{K}}(\mathcal{P})$ calculated from direct inversion of $\mathcal{P}$-matrix which is fraught with many negative the inter-block interaction strength demonstrated in the off-diagonal block, $\widetilde{\mathcal{K}}(\mathcal{P})$ inferred by the constrained optimization display a better resemblance to $\mathcal{K}^*$ with a relative error of 0.044.

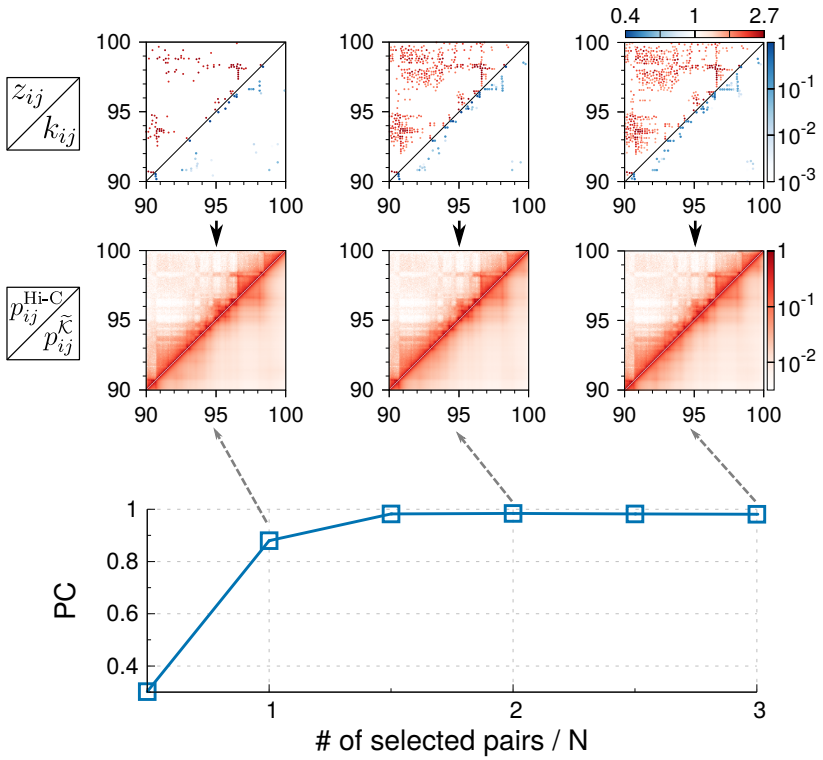

FIG. S6. Dependences of the interaction strength matrix and contact probability matrix on the number of selected significant pairs, are demonstrated using the 10 Mb region on chr5 of GM12878 cells. When top N, $2N$, and $3N$ contact pairs are selected, the heatmaps of the contact significance (z_{ij}) and interaction strength (k_{ij}) are shown on the top; the contact probability from Hi-C ($p_{ij}^{\text{Hi-C}}$) and from k_{ij} without considering non-bonded interactions ($p_{ij}^{\widetilde{\mathcal{K}}}$) are shown in the middle row. The Pearson correlation between $p_{ij}^{\text{Hi-C}}$ and $p_{ij}^{\widetilde{\mathcal{K}}}$ as a function of the number of selected pairs is plotted at the bottom.

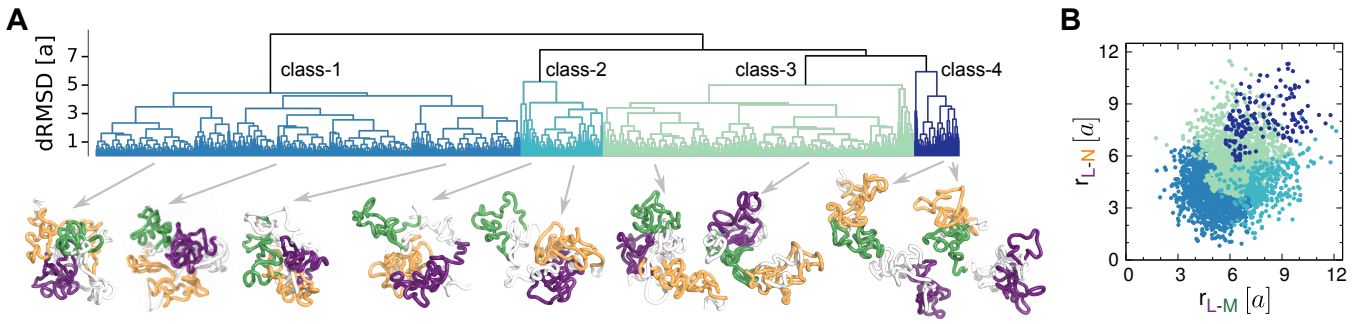

FIG. S7. Variability in the HLM-generated conformational ensemble for a 10 Mb-genomic region of chr5 in GM12878 cells (Fig. 1D). (A) Dendrogram of chromosome conformations from hierarchical clustering. Illustrated are the chromosome conformations from the four classes with L, M, N domains colored in purple, green, orange, respectively, following the domain labels assigned in Fig. 1B. (B) Scatter plot of inter-domain distances $r_{\text{L-M}}$ versus $r_{\text{L-N}}$ of structures in different classes.

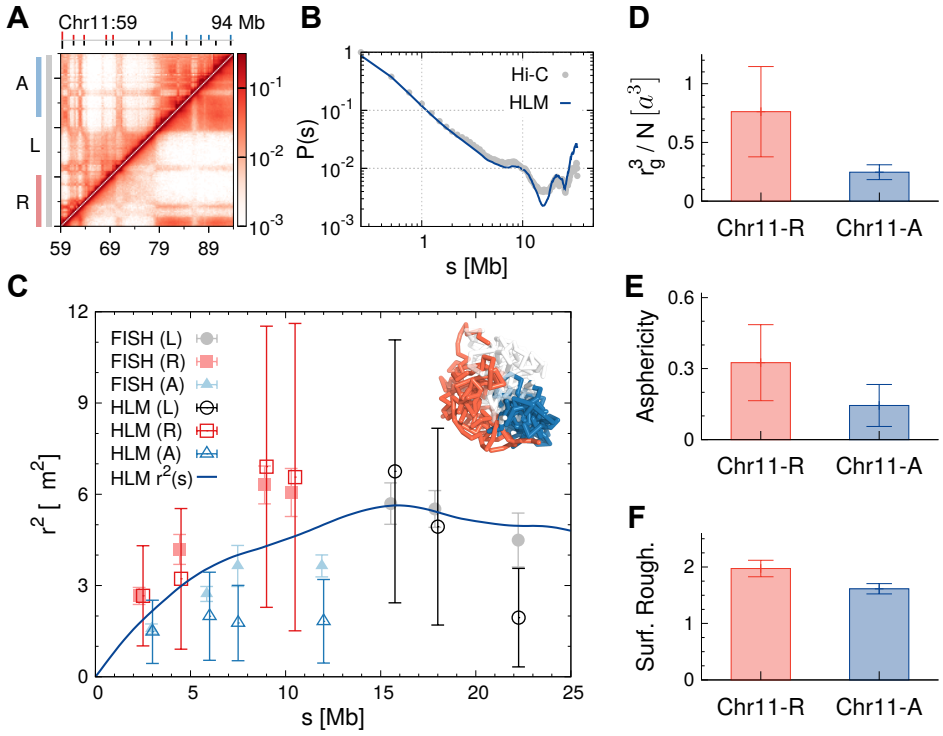

FIG. S8. A 35 Mb genomic region on chr11 in IMR90 cells modeled by HLM. Mateos-Langerak *et al.* have performed FISH experiment [8] in this region, with FISH probes distributed within a transcriptionally active ridge domain, inactive anti-ridge domain, and a longer region including both. They are labeled as "R","A" and "L", respectively. The genomic positions of the probes are labeled by sticks at the top of (A), below which is the heatmap of contact probabilities from Hi-C (upper diagonal region) and HLM (lower diagonal region). (B) Mean contact probability $P(s)$. (C) Pairwise square distance r_{ij}^2 between the FISH probes as a function of the genomic distance s, as well as the mean square distance $r^2(s) = \sum_{i=0}^{N-s-1} r_{i,i+s}^2/(N-s)$ (the solid line). Structural ensemble is illustrated with the ridge and anti-ridge domains colored red and blue, respectively. (D) Compactness, (E) asphericity, and (F) roughness of the surface of the domains.

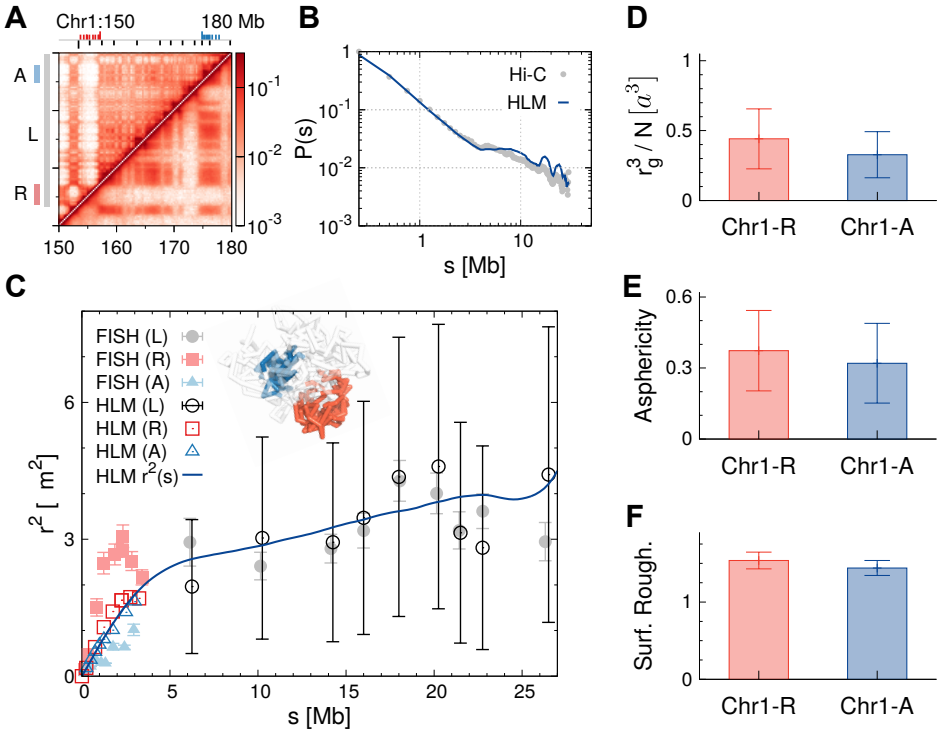

FIG. S9. A 30 Mb genomic region on chr1 in IMR90 cells modeled by HLM (see also the caption of Fig. S8).

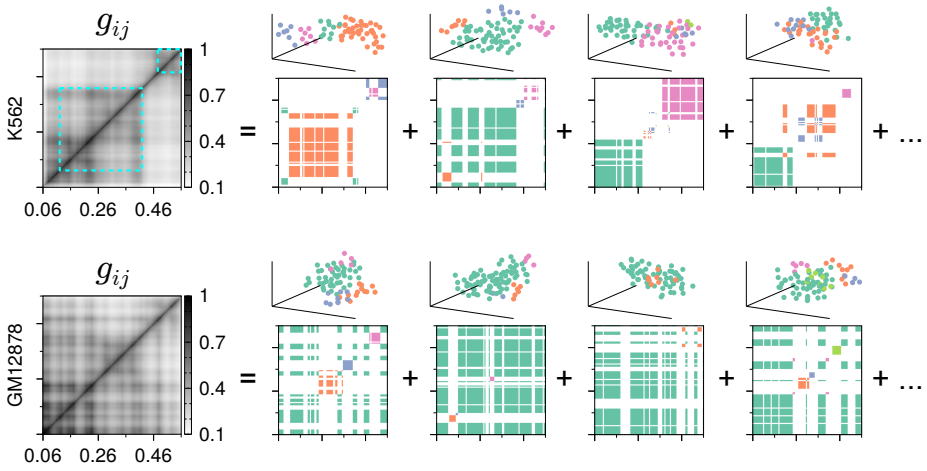

FIG. S10. Structural domains at α-globin gene loci. For each chain conformation, monomers are classified into different groups by applying DBSCAN (Density Based Spatial Clustering of Applications with Noise [9]) algorithm. Four conformations and clustering results are shown as examples for each cell type, where different colors label different groups. The leftmost heatmaps are ensemble-averaged grouping probabilities, determined as $g_{ij} = \sum_{c=1}^{M} g_{ij}^c/M$ where M is the total number of conformations. For the c-th conformation, $g_{ij}^c = 1$ if the i and j-th monomers belong to the same group; otherwise, $g_{ij}^c = 0$. DBSCAN clustering was carried out by using the Scikit-learn package [10], with a threshold distance of 7 a. The cyan dashed boxes mark the previously reported domain boundaries in Ref. [11].

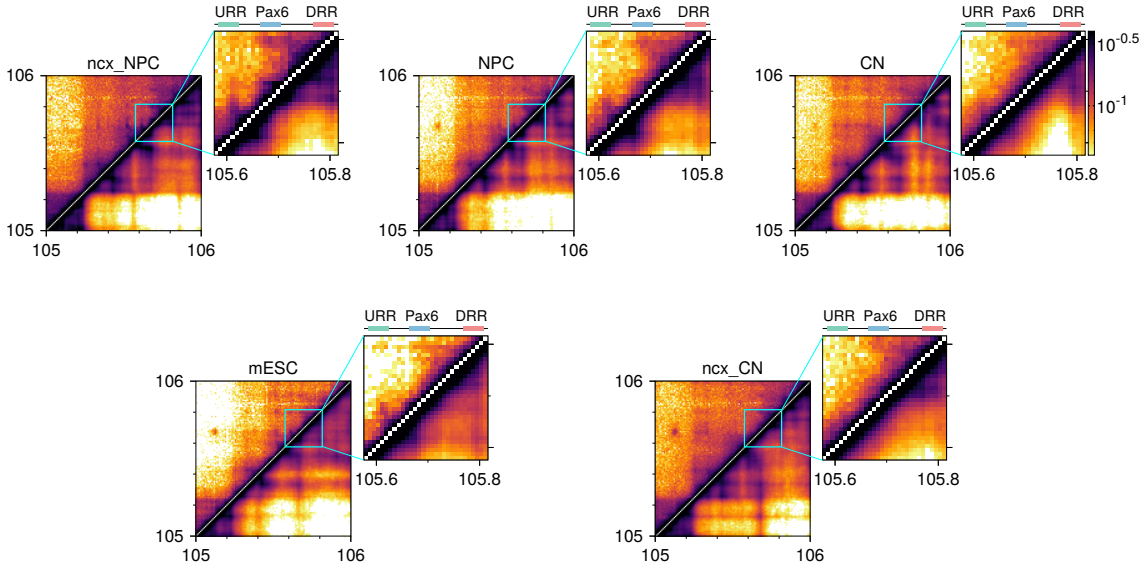

FIG. S11. Contact probabilities calculated from Hi-C (upper diagonal region) and HLM (lower diagonal region) around the mouse Pax6 gene for five different cell types. A 200 kb region was zoomed in, which highlights the contacts between the three "simulated" FISH probes (URR, Pax6, and DRR).

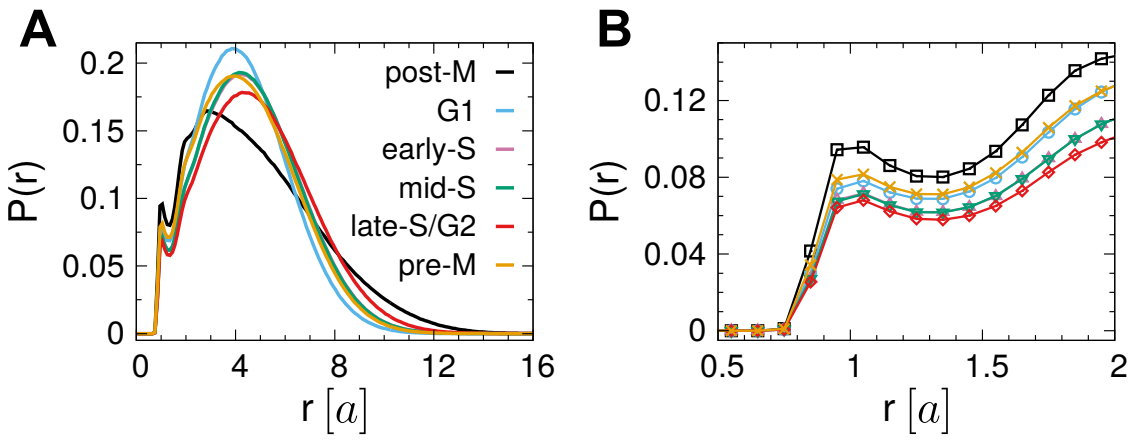

FIG. S12. (A) Probability densities of pairwise distance $P(r) = \frac{2}{N(N-1)} \sum_{i>j} \delta(r_{ij} - r)$ in chr19 of mESC at different phases of the cell cycle. (B) A zoom-in view of $P(r)$ at small values of r. Because of the elongated shape, the post-M phase has $P(r)$ with a fatter tail than other phases. In terms of local compactness, $P(r)$ at $r < 2a$ (Fig. S12B) suggests that post-M phase is the most compact, followed by pre-M phase and so forth (post-M > pre-M > G1 > early-S ≥ mid-S > late-S/G2), the order of which is identical to that assessed by the monomer volume ($\overline{v}$) (Fig. 6B).

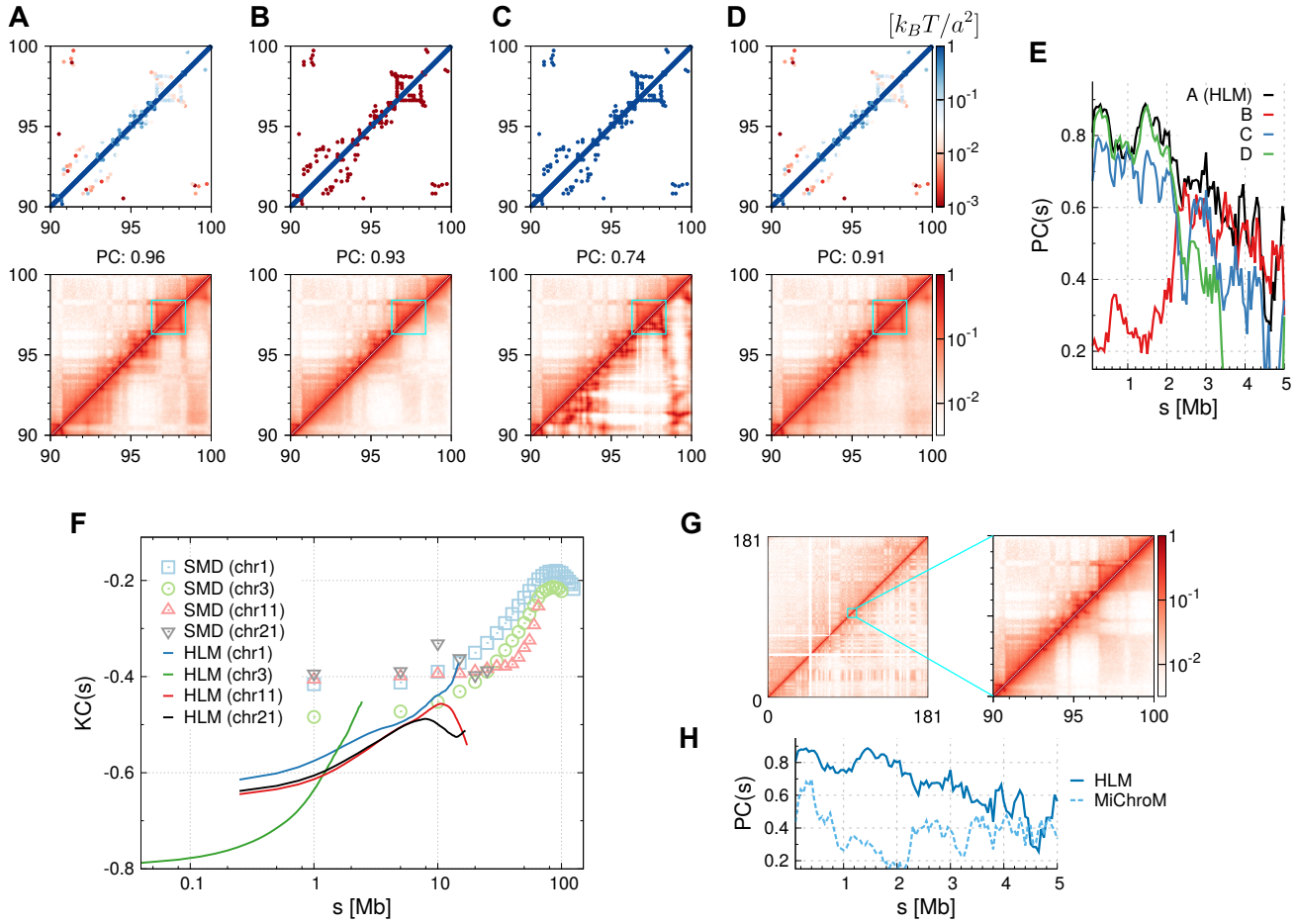

FIG. S13. Comparison of HLM with other models by modeling a 10 Mb region of GM12878 cells. Four models (HLM and its three variants) are considered: multi-block copolymer (A,B,C) and homopolymer (D) model with harmonic springs of various (A,D) or uniform strengths (B,C). For each model, the top panel shows the corresponding spring strength matrix, and the resulting contact probabilities (lower diagonal) are contrasted with those measured by Hi-C (upper diagonal) at the bottom panel. A domain with edges showing enriched contacts is highlighted by a cyan box. (E) Pearson correlation of contact probabilities between Hi-C and modeling as a function of the genomic separation. (F) Comparing HLM with the polymer model based on steered molecular dynamics simulations (SMD) [12]. Kendall rank correlation coefficients (KC) are computed between Hi-C contact probabilities and mean pairwise distances based on HLM, considering monomer pairs with genomic distance greater than s. The results of SMD (markers) are adpated from Fig. S4 in Ref. [12]. The correlation is higher if the value of KC is closer to -1. (G,H) Comparing HLM with MiChroM [13]. (G) Heatmaps of contact probabilites from Hi-C (upper diagonal part) and from MiChroM (lower diagonal part) of chr5 in GM12878 cells. A zoom-in view of a 10 Mb region of interest is on the right. (H) Pearson correlations between Hi-C and two models at different genomic separation.

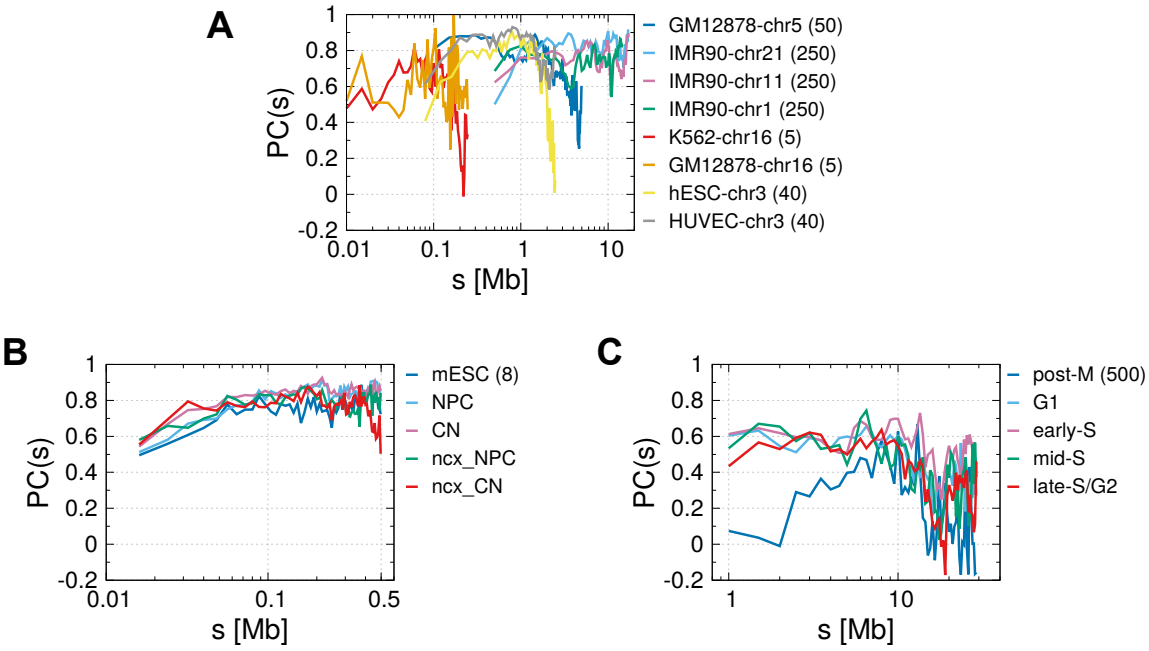

FIG. S14. Pearson correlations of contact probabilities from Hi-C and HLM at different genomic separation of (A) human, (B) mouse genomic regions based on ensemble Hi-C data, and (C) chr19 of mESC at different phases of cell cycle based on single-cell Hi-C data. The numbers in the parentheses are the resolutions of HLM models (i.e., the genomic size of each monomer in the unit of kb. See also Table 1.)

TABLE S1. TADs on chr21 of human IMR90 cells whose pairwise distances were measured in Ref.[14], and computed in Fig. 2.

TAD index	Start(bp)	End(bp)	Center(bp)	i	A/B
2	13,280,000	16,160,000	14,720,000	2	B
3	16,160,000	18,280,000	17,220,000	12	B
4	18,320,000	21,080,000	19,700,000	22	B
5	21,240,000	23,160,000	22,200,000	32	B
6	24,320,000	25,680,000	25,000,000	44	B
7	26,160,000	27,000,000	26,580,000	50	B
8	27,040,000	28,000,000	27,520,000	54	A
9	28,000,000	29,320,000	28,660,000	58	B
10	29,360,000	29,680,000	29,520,000	62	B
11	29,760,000	31,360,000	30,560,000	66	B
12	31,400,000	31,960,000	31,680,000	70	B
13	31,960,000	32,600,000	32,280,000	73	B
14	32,600,000	32,920,000	32,760,000	75	B
15	33,080,000	33,800,000	33,440,000	77	B
16	33,840,000	34,200,000	34,020,000	80	A
17	34,200,000	34,680,000	34,440,000	81	A
18	34,680,000	34,920,000	34,800,000	83	A
19	34,920,000	36,440,000	35,680,000	86	A
20	36,440,000	36,640,000	36,540,000	90	A
21	36,680,000	37,440,000	37,060,000	92	A
22	37,440,000	37,880,000	37,660,000	94	B
23	37,960,000	38,720,000	38,340,000	97	B
24	38,720,000	39,760,000	39,240,000	100	B
25	39,760,000	41,440,000	40,600,000	106	B
26	41,440,000	42,120,000	41,780,000	111	B
27	42,120,000	42,840,000	42,480,000	113	B
28	42,840,000	43,120,000	42,980,000	115	B
29	43,160,000	44,040,000	43,600,000	118	A
30	44,040,000	44,360,000	44,200,000	120	A
31	44,360,000	45,040,000	44,700,000	122	A
32	45,040,000	45,400,000	45,220,000	124	A
33	45,440,000	46,160,000	45,800,000	127	A
34	46,160,000	46,944,323	46,552,161	130	A

i is the index of the corresponding monomer in HLM. A/B type is determined by a principal component analysis of the inter-TAD distance matrix measured in the experiment.

TABLE S2. FISH probes in chr11 of human IMR90 cells whose pairwise distance were measured in Ref.[8], and computed in Fig. S8.

Domain	q	Start(bp)	End(bp)	Center(bp)	i
R	1	59,145,349	59,328,495	59,236,922	0
	0	61,446,617	61,635,131	61,540,874	10
	0	63,606,476	63,728,827	63,667,651	18
	0	68,035,816	68,234,792	68,135,304	36
	0	69,453,280	69,614,785	69,534,032	42
A	1	81,410,783	81,515,783	81,463,283	89
	0	84,330,302	84,491,603	84,410,952	101
	0	87,242,814	87,391,556	87,317,185	113
	0	88,840,806	89,052,495	88,946,650	119
	0	93,270,300	93,463,527	93,366,913	137
L	1	59,145,349	59,328,495	59,236,922	0
	0	74,678,334	74,845,650	74,761,992	63
	0	77,016,132	77,155,090	77,085,611	72
	0	81,410,783	81,515,783	81,463,283	89
	0	84,330,302	84,491,603	84,410,952	101
	0	87,242,814	87,391,556	87,317,185	113
	0	90,287,090	90,448,063	90,367,576	125
	0	93,270,300	93,463,527	93,366,913	137

In the experiment [8], distances were measured between the reference probe ($q = 1$) and other probes ($q = 0$) within a transcriptionally active ridge domain (R), a transcriptionally inactive anti-ridge domain (A), and a longer genomic region including both (L). The genomic positions of probes were lifted from hg15 to hg19 [15]. i is the index of the corresponding monomer in HLM.

TABLE S3. FISH probes in chr1 of human IMR cells. The pairwise distance were measured in Ref.[8], and computed in Fig. S9.

Domain	q	Start(bp)	End(bp)	Center(bp)	i
R	0	153,688,049	153,838,214	153,763,131	15
	0	154,258,113	154,423,159	154,340,636	17
	0	154,756,480	154,933,673	154,845,076	19
	0	154,813,142	154,963,617	154,888,379	19
	0	155,236,093	155,386,538	155,311,315	21
	0	155,869,571	156,011,182	155,940,376	23
	0	156,245,828	156,422,950	156,334,389	25
	0	156,763,312	156,949,996	156,856,654	27
	0	156,918,444	157,130,858	157,024,651	28
	1	157,089,739	157,266,762	157,178,250	28
A	1	174,780,621	174,961,968	174,871,294	99
	0	174,960,409	175,130,220	175,045,314	100
	0	175,283,924	175,434,463	175,359,193	101
	0	175,600,401	175,773,331	175,686,866	102
	0	175,886,407	176,036,727	175,961,567	103
	0	176,108,104	176,295,598	176,201,851	105
	0	176,558,298	176,714,279	176,636,288	106
	0	177,180,236	177,391,475	177,285,855	109
	0	177,747,748	177,891,719	177,819,733	111
L	1	153,367,866	153,518,504	153,443,185	13
	0	155,275,054	155,425,545	155,350,299	21
	0	157,394,838	157,556,421	157,475,629	29
	0	159,499,529	159,658,201	159,578,865	38
	0	163,510,707	163,671,832	163,591,269	54
	0	167,491,540	167,680,298	167,585,919	70
	0	169,360,041	169,519,360	169,439,700	77
	0	171,415,729	171,565,827	171,490,778	85
	0	173,507,237	173,672,089	173,589,663	94
	0	176,530,621	176,711,968	176,621,294	99
	0	177,858,104	178,045,598	177,951,851	104
	0	179,679,177	179,836,567	179,757,872	119

Some column names are explained in the footnote of Table S2.

TABLE S4. Chromatin interactions captured by ChIA-PET between the α-globin gene of α-globin domain of human chr16 in two distinct cell lines (K562 and GM12878) and the rest of the domain.

Cell line/	*i*		*j*	
Protein	Start(bp)	End(bp)	Start(bp)	End(bp)
K562/	228,606	232,911	101,957	105,248
Pol II[16]	223,848	234,650	112,365	122,815
	225,207	234,180	123,062	131,022
	228,466	231,601	140,986	144,385
	230,572	233,311	144,684	148,565
	223,675	234,904	148,901	176,075
	212,504	217,730	187,975	191,457
	228,181	231,216	193,260	195,050
	227,544	232,194	281,991	287,514
	228,369	231,250	337,052	340,267
	231,278	233,446	400,120	404,207
	227,874	230,293	415,011	417,686
K562/	231,119	231,985	115,418	116,419
CTCF[16]	230,063	230,936	115,816	116,534
	231,349	231,926	118,457	119,027
	229,989	230,841	146,609	147,247
	231,126	232,019	146,696	147,500
	229,955	230,849	156,728	157,605
	231,083	232,073	156,717	157,821
	230,241	231,130	157,831	158,483
	229,997	230,876	167,564	168,163
GM12878/	230,169	231,829	115,412	117,393
CTCF[17]	231,254	231,916	146,941	147,379
	233,653	235,304	157,106	157,280
	230,334	232,616	154,799	158,610
	230,238	232,162	166,032	168,889
	231,353	231,816	412,084	412,434

Pol II-mediated interaction, involving α-globin genes and the rest of the domain, is absent in GM12878 cells [17]. Since CTCF-mediated interactions are mostly overlapped between K562 and GM12878 cells, the K562-specific interactions are mainly mediated by Pol II.