Nucleic Acids Research, 2023, 1–17
https://doi.org/10.1093/nar/gkad1121
Gene regulation, Chromatin and Epigenetics

OXFORD

Hi-BDiSCO: folding 3D mesoscale genome structures from Hi-C data using brownian dynamics

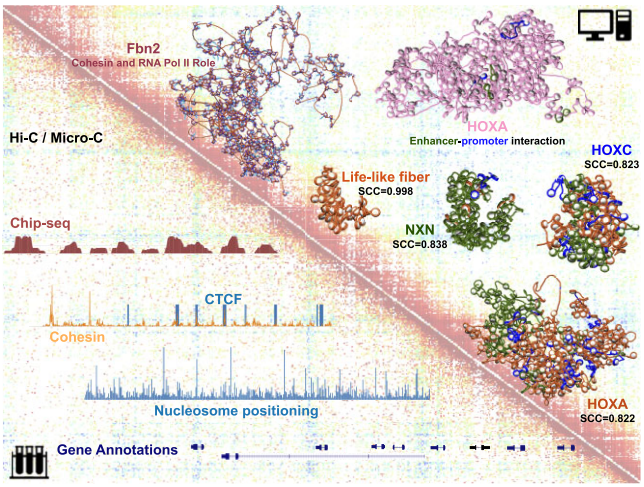
Zilong Li [1,4] **and Tamar Schlick** [1,2,3,4,*]

[1]Department of Chemistry, 100 Washington Square East, Silver Building, New York University, New York, NY 10003, USA
[2]Courant Institute of Mathematical Sciences, New York University, 251 Mercer St., New York, NY 10012, USA
[3]New York University-East China Normal University Center for Computational Chemistry, New York University Shanghai, Shanghai 200122, China
[4]Simons Center for Computational Physical Chemistry, 24 Waverly Place, Silver Building, New York University, New York, NY 10003, USA
[*]To whom correspondence should be addressed. Tel: +1 212 998 3116; Email: schlick@nyu.edu

Abstract

The structure and dynamics of the eukaryotic genome are intimately linked to gene regulation and transcriptional activity. Many chromosome conformation capture experiments like Hi-C have been developed to detect genome-wide contact frequencies and quantify loop/compartment structures for different cellular contexts and time-dependent processes. However, a full understanding of these events requires explicit descriptions of representative chromatin and chromosome configurations. With the exponentially growing amount of data from Hi-C experiments, many methods for deriving 3D structures from contact frequency data have been developed. Yet, most reconstruction methods use polymer models with low resolution to predict overall genome structure. Here we present a Brownian Dynamics (BD) approach termed Hi-BDiSCO for producing 3D genome structures from Hi-C and Micro-C data using our mesoscale-resolution chromatin model based on the Discrete Surface Charge Optimization (DiSCO) model. Our approach integrates reconstruction with chromatin simulations at nucleosome resolution with appropriate biophysical parameters. Following a description of our protocol, we present applications to the NXN, HOXC, HOXA and Fbn2 mouse genes ranging in size from 50 to 100 kb. Such nucleosome-resolution genome structures pave the way for pursuing many biomedical applications related to the epigenomic regulation of chromatin and control of human disease.

Graphical abstract



Introduction

The 3D architecture of the genome is critical for gene expression (1,2), regulation (3,4), transcription (5) and other fundamental biological functions. To understand the 3D organization of the eukaryotic genome, much progress has been realized on both experimental and computational fronts over the past decade.

As described in a recent review (6), chromosome conformation capture (3C) methods have risen exponentially since 2002 for genome interrogation. The 3C method developed by Dekker et al. (7) was followed by higher resolution techniques such as 4C (8), 5C (9), Hi-C (10), Micro-C (11) and RCMC (12). Today, Hi-C-like experiments with resolution as high as 50 bp (base pairs) are feasible (12), with interactions across the entire genome detectable. In addition, single-

Received: July 6, 2023. **Revised:** October 12, 2023. **Editorial Decision:** November 6, 2023. **Accepted:** November 22, 2023

cell Hi-C techniques (13,14) measure 3D structures of individual genomes directly and provide data for the analysis of chromatin folding in rare cell types like stem (15) and totipotent (16) cells, thereby contributing to a better understanding of basic mechanisms in development, differentiation and human disease. For example, single-cell Hi-C has been applied to examine changes in genome organization during the cell cycle (17,18), cell-type-specific chromosomal architecture (19), and structure-expression interactions (20).

Methods for predicting and reconstructing 3D genome structures from such Hi-C data have also naturally emerged (21–23). Physical polymer models, with beads representing bp to kbp, are commonly used to reconstruct the whole genome or sub elements, like genes, TADs. The general workflow for Hi-C reconstruction of the polymer model is described in our recent review (23). In brief, the **input data** include Hi-C or single-cell Hi-C maps with possibly additional information, such as FISH and Chip-seq data. **Correction** of these data may be performed by data normalization (24) or probability-based methods (25) that treat the problem as Bayesian inference or maximum likelihood problem, so that uncertainties in the experimental Hi-C data can be interpreted in probabilistic terms. The **conversion** of Hi-C data can be made into distances (26,27) or contacts (28–30). Distance-based methods usually convert the frequencies between two loci i and j (f_{ij}) to spatial 3D distances between those loci (d_{ij}) by an inverse relationship $d_{ij} \propto 1/f_{ij}^{\alpha}$, where α is a parameter related to the genomic distance and resolution of the Hi-C map. Contact-based methods use Hi-C contacts directly as restraints for modeling, such as by adding fictitious 'bonds' between loci. A reconstruction **model** can be based on an analytical framework or simulations. Finally, the **output structures** are categorized by consensus, resampling and population-based methods (31). Consensus methods generate a single structure from ensemble Hi-C data; resampling methods generate an ensemble of structures that satisfy distance data derived from the Hi-C map; and population-based methods generate a population of individual structures from ensemble Hi-C data.

Apart from these physical models, as outlined in the process above, statistical models together with machine learning algorithms (32–34) also provide insights into various genome patterns such as relationships between histone modifications, transcription factor binding, and chromatin interactions.

When reconstructing 3D genome structures, several challenging problems arise, including the level of resolution and usage of artificial constraints. Many polymer reconstruction models predict coarse polymer chain trajectories. Some approaches add fictitious constraints (e.g., specific binders) to match Hi-C data without biophysical meaning.

To address some of these limitations, our Brownian Dynamics (BD) based approach incorporates our mesoscale model for structure refinement to capture the spatial relationship between nucleosomes and DNA; the resulting structure from BD reconstruction can be used directly to simulate and further probe the genome system's dynamics independently of any constraints/restraints.

We illustrate performance on three gene systems, showing a two-stage approach. The BD stage reproduces the patterns on the Hi-C maps in minutes for fibers in the size range of 100 kb, and the MC stage resolves clashes and accounts for histone tails and linker histones without compromising the contact patterns derived from the original Hi-C map. Reconstruction from known structures further demonstrates that Hi-BDiSCO generates biophysically sound fibers. Then we use Hi-BDiSCO to reconstruct the HOXA gene and examine the role of Pol II pausing in enhancer–promoter interactions. Similarly, we reconstruct the Fbn2 gene to study the effect of cohesin and transcription inhibition on chromatin architecture. The resulting Hi-BDiSCO folded gene structures at mesoscale resolution provide a wealth of mechanistic insights into 3D spatial gene structure to probe many biological features and processes.

Materials and methods

Chromatin mesoscale model

Our mesoscale chromatin model (Figure 1 and see recent reviews in (6,35–37)) has coarse grained elements at different levels of resolution. The nucleosome cores are treated by a Discrete Surface Charge Optimization (DiSCO) model (38) as charged disks according to the atomistic core particle. Linker DNA, histone tails and linker histones (LH) are treated as beads. Histone tails and LH can be turned on or off to study details at different resolutions.

We currently consider flexible histone tails (39) and their acetylation (40,41), two LH variants (H1E and H1C) with several binding modes (on- versus off-dyad) (42,43), and non-uniform linker DNA lengths (44), among other features (6).

Parameters for mesoscale model

Nucleosome positioning

Nucleosome positions are obtained from MNase-seq data. We first use DANPOS (45) to retrieve the peaks of the MNase-seq data, which represent the nucleosome locations. To eliminate overlapping nucleosomes, we perform a 'greedy' algorithm, that is, from the first nucleosome, we select the next nucleosome when the coordinates between the two adjacent nucleosome starting positions are at least 166 bp away (147 bp of nucleosomal DNA plus 19 bp of linker DNA). The numbers and relative distances (in bp) between nucleosomes are then used to build our mesoscale model.

Tail acetylation

Tail acetylation regions are obtained from H3K27ac Chip-seq data. The peaks of H3K27ac Chip-seq data are calculated using MACS (46), and the overlapping with their peaks genomic coordinates are used to identify nucleosomes in these regions.

Linker Histone (LH) density

The linker histone (LH) density ρ measured as LH per nucleosome is approximated from experimental Chip-seq measurements (GEO accession: GSE46134 (47)). Besides Chip-seq, in some cases, we also use the average LH density observed from experiments (48). We first calculate the number of nucleosomes that have LH attached ($N_{C_{LH}}$) according to ρ, $N_{C_{LH}} = \rho \times N_C$ (N_C is the total number of nucleosome cores). We then select $N_{C_{LH}}$ nucleosomes associated with the peaks (highest values) in the Chip-seq data and assign LHs to them.

Examples of annotated mesoscale gene models for HOXA, HOXC and NXN are shown in Figure 2.

Note that although we use a specific setting for nucleosome positions, tail acetylation marks, and LHs, an ensemble of fibers with different distributions, as considered in (49), could also be used. The experimental data sources from mESC used for these distributions are summarized in Supplementary Table S1.

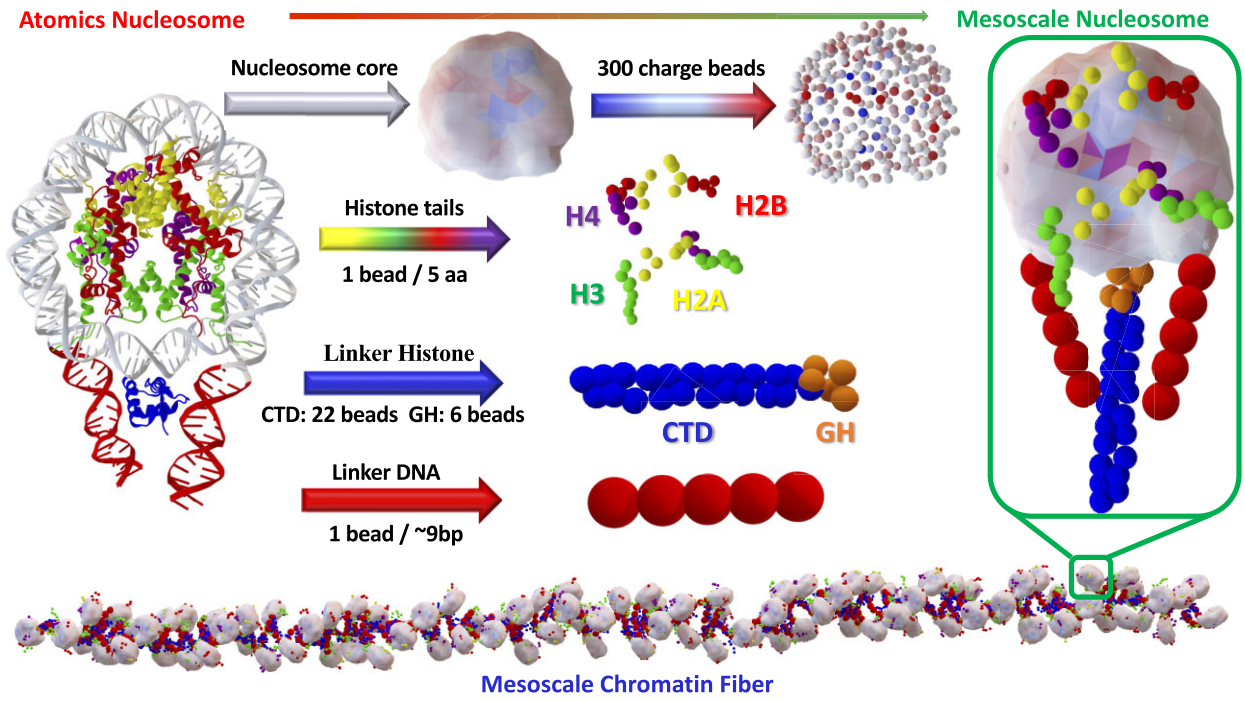


Figure 1. Our chromatin nucleosome-resolution mesoscale model (6,35–37). A linear 100-nucleosome chromatin fiber at bottom with the enlarged basic unit (chromatosome) at top, from a side view. The Nucleosome core is represented by the Discrete Surface Charge Optimization (DiSCO) model shown at right as an irregular surface and 300 distributed charge beads. Histone tails are shown as green (H3), yellow (H2A), red (H2B) and blue (H4) beads. Linker histones are shown as orange (globular head) and cyan (C-terminal domain) beads. Linker DNA units are shown as red beads.

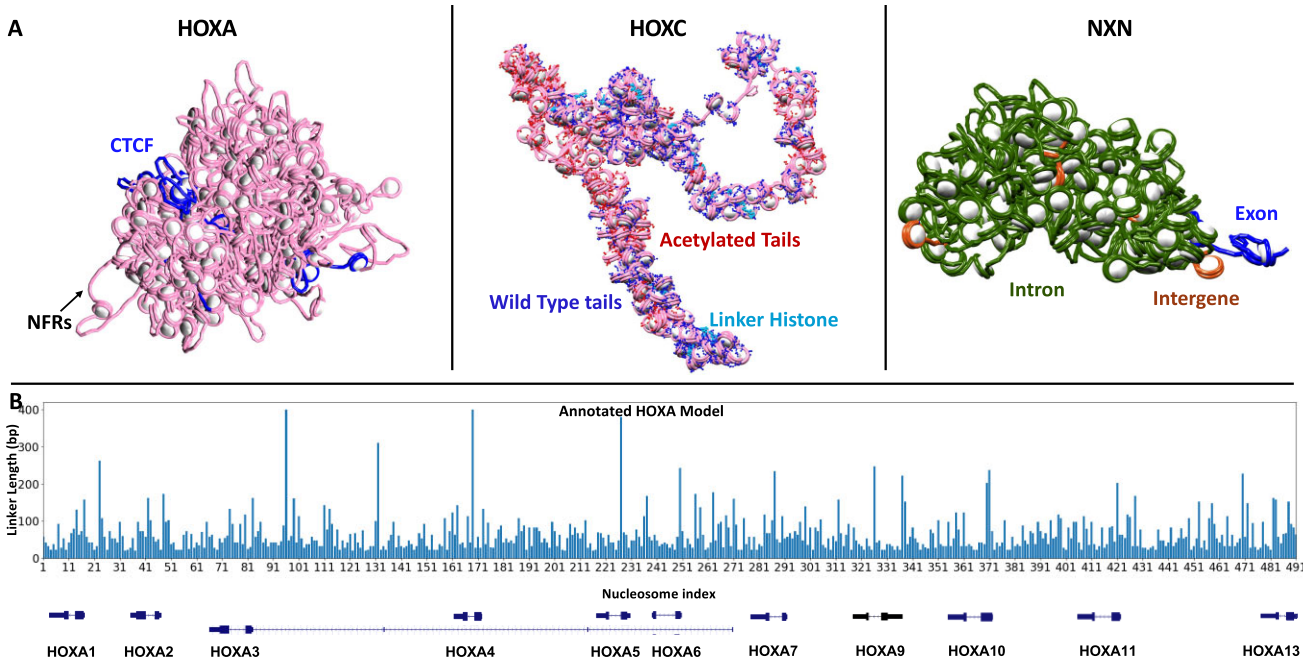


Figure 2. Examples of annotated gene systems. (**A**) Mesoscale models of HOXA, HOXC and NXN gene systems with different annotated information and reconstruction stages. Left: HOXA gene with blue CTCF binding sites after the MC subsequent simulation; middle: HOXC gene with wildtype tails as blue beads, acetylated tails as red beads and LH as teal beads, initial random structure; right: NXN gene system with intron DNA in dark green, exon DNA in blue and intergenic DNA in red after BD reconstruction. (**B**) DNA linker lengths of HOXA model, plotted against the nucleosome index, 492 nucleosomes in total obtained from mESC MNase-seq data (GEO accession: GSM2083105), along with annotated gene locations.

Hi-BDiSCO Brownian dynamics simulation and Monte Carlo sampling protocol

Our mesoscale Brownian dynamics (BD) simulation protocol was recently developed for CUDA implementation (37), demonstrating physically accurate behavior and rapid performance that makes feasible the study of chromatin fibers in the range of kb, or hundreds of nucleosomes. The bottleneck is the storage requirement for diffusion tensors and force matrices for large systems and the associated Cholesky factorization. For example, ∼100 nucleosomes or 20 kb chromatin fiber can be simulated with full tail and LH beads, but a 50 kb fiber would occupy 64 GB memory, too large for the typical 16 GB or 32 GB GPU memory.

Our mesoscale Monte Carlo (MC) sampling protocol has proven useful for simulating gene-level (∼500 nucleosomes or 100 kb chromatin fiber) systems with full details of histone tails and LH (50,51).

Both BD and MC have the total potential energy function of the model as:

$$E(\mathbf{r}) = E_S + E_B + E_T + E_{tS} + E_{tB} + E_{lhS} + E_{lhB} + E_V + E_C, \quad (1)$$

where $\mathbf{r}$ is the collective position vector of the system; E_S, E_B, E_T are the stretching, bending and twisting terms for the linker DNA and nucleosome core; E_{tS} and E_{tB} are the stretching and bending terms for histone tails; E_{lhS} and E_{lhB} are the stretching and bending terms for LH; E_V and E_C are the excluded volume (repulsive term, applied to all particle–particle interactions to avoid close contact) and electrostatic terms for all beads, respectively.

The details of our BD and MC protocols can be found in (37) and (39,51,52).

Brownian dynamics reconstruction implementation

The idea of the reconstruction is as follows (see Figure 3):

Hi-C map and initial structures

We first map intervals between the 2D Hi-C/Micro-C and the 3D mesoscale model in the chromatin system corresponding to the chromosome coordinates. The contact interaction map ($M \times M$ matrix) from the Hi-C/Micro-C experiment provides a frequency f_{ij} for each $\{i, j\}$ genomic contact. From this experimental map (a Micro-C map in this case), we select the region of interest (i.e., HOXA shown in Figure 3Ai)). In our mesoscale model, each DNA bead represents ∼9 bp region. Each nucleosome core represents ∼147 bp region (shown as 16 grey nucleosomal DNA beads in Figure 3Aii)). The mesoscale fiber is then evenly partitioned into n regions according to the resolution of Micro-C data, as shown in Figure 3A. There we select a 1000 bp region for illustration, so that the Micro-C resolution of 100 bp corresponds to $n = 10$, or ten 100-bp regions. Based on the chromosome coordinates, we map the regions in the 3D mesoscale chromatin fiber to the 2D Micro-C data. For example, the 11 beads in the red circle of the zoomed mesoscale model correspond to chr6:52197500–52197600, and this region maps to the region marked with a blue triangle in the Micro-C map. We then generate a number (N_{rep}) of replicas (Figure 3A) with the same configurations (nucleosome positioning, Micro-C mapping regions, etc.) and different random initial spatial structures for the following steps.

Restraints definition and distribution

All the Hi-C/Micro-C $\{f_{ij}\}$ values are then distributed into the replicas to reproduce the original contact frequency distribution. That is, based on the frequency f_{ij} and the number of replicas N_{rep}, we calculate the number of copies onto which we assign specific contacts $N_{R_{ij}}$, a subset of the N_{rep} replicas, $N_{R_{ij}} = f_{ij} \times N_{rep}$. As shown in Figure 3B, the simplified illustration has $f_{31} = 0.6$ and $N_{rep} = 8$, so we calculate $N_{R_{31}} = f_{31} \times N_{rep} = 0.6 \times 8 \approx 5$. Thus, in this case, we distribute R_{31} contacts into 5 out of the 8 replicas. We determine the specific subset of replicas using the random sampling method in Python (*random.sample*(*population*, *k*)), where the *population* is the number of replicas (N_{rep}) and k is the target number of restraints ($N_{R_{ij}}$). This will return k elements chosen from the *population*, and the output will differ every time. In practice, we start by randomly choosing one element from the *population*, then moving the selected element into a vacant replica and selecting another element from the other elements in *population* until k elements are selected. As shown in Figure 3B, four restraints R_{ij} are assigned: R_{31} is assigned to five replicas (1, 2, 4, 7, 8); R_{64} is assigned to all eight replicas; R_{83} is assigned to two replicas (3,7); and R_{97} is assigned to replica 3 only. In practice, all the restraints R_{ij} (for non-adjacent i, j) are assigned to replicas. Clearly, there are infinite ways of distributing restraints to reproduce the Hi-C/Micro-C maps. This random distribution, one of many possible ways, has proven reasonable for reproducing 3D structures and related experimental contact maps.

The restraints are modeled as distance based with linear force $R_{ij} = h_m(d_{ij} - d_0)$, where h_m is the stretching constant, d_{ij} is inter-bead distance corresponding to the regions i and j, and d_0 is the target d_{ij} (see parameter selection in the following subsections.) Since each partitioned region has multiple DNA beads (linker or nucleosomal) in our mesoscale model (∼11 beads for 100 bp resolution Micro-C data), the restraints are applied only to the center bead of the region; if the center bead is within the nucleosome core, the restraint is applied to the core.

Simulation

We then perform a simplified mesoscale BD simulation (with tail/LH interactions and tail/LH beads omitted) for 100 ns using $\Delta t = 1$ps for 100 000 steps with these restraints in each replica to fold the gene fibers. This simplification makes it possible to simulate ∼500 nucleosome gene systems by BD and quickly fold the system (minutes of computational time). As shown in Figure 3Ci), after the BD simulation, the structures fold into condensed configurations. The restraint regions i, j in replicas are close to each other (e.g., regions 1 and 3, and 4 and 6 are close in replica 1 after BD).

To fully utilize parallel computing with CUDA, we feed the restraints as a hash table, and assign N threads, where N is the number of restraints. Thus, each thread i only performs one calculation for the ith restraint. With the CUDA implementation of our BD simulation (37), which executes the compute-intensive computations (diffusion tensor, force, etc.) in parallel, the chromatin fiber can be quickly folded into a condensed structure compatible with the Hi-C/Micro-C data.

Finally, we perform full-scale nucleosome-resolution mesoscale Monte Carlo (MC) simulations (39,50) with restraints released for 10 million steps (Figure 3C). The MC simulation takes into account the effect of histone tails and

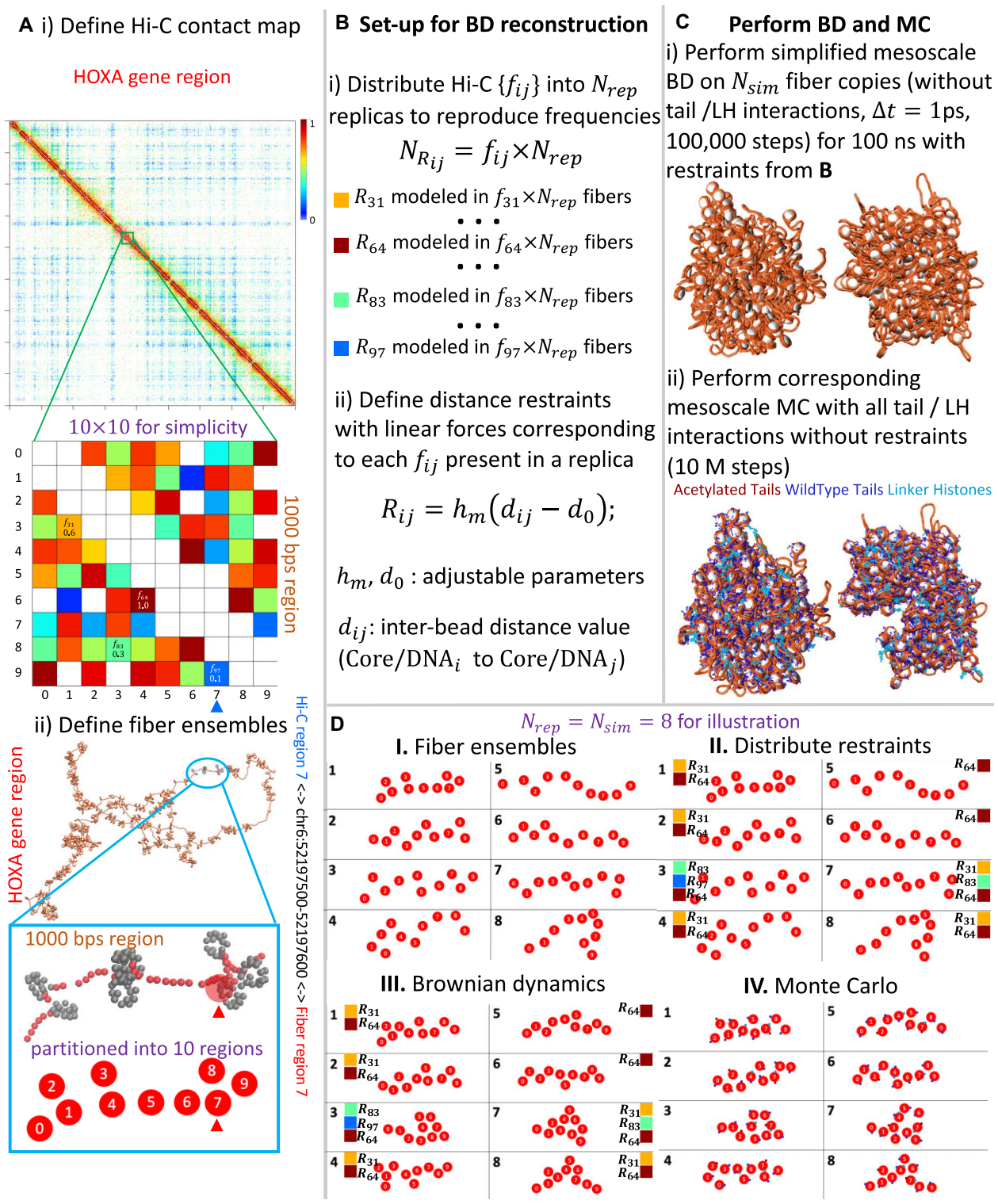


Figure 3. Hi-BDiSCO reconstruction strategy from Micro-C maps using BD (with artificial values for illustration and $N_{rep} = N_{sim} = 8$). (**A**) i) A Micro-C contact map (100 bp resolution) with the HOXA gene region, enlarged in a 1000 bp region (10 × 10 for simplicity). Illustrative values (e.g., $f_{31} = 0.6$) are displayed to show the frequency of contacts between those regions. ii) Ensemble of fibers: illustrative mesoscale model with the HOXA gene region, with enlarged 1000 bp region. In this configuration, each bead represents a ∼9 bp region, with linker DNA beads (red) and nucleosomal DNA beads (grey). We evenly partition this fiber into 10 regions (100 bp each) and map the 2D Micro-C data region into the 3D mesoscale chromatin fiber based on the chromosome coordinates. (**B**) Set-up of BD simulation from Micro-C data. i) Distribute randomly Micro-C (f_{ij}) into replicas to reproduce frequencies. Based on the frequency f_{ij} and the number of replicas (N_{rep}), we know which restraints are distributed into which replicas. ii) Apply distance restraints with linear forces. (**C**) Simulation. i) Perform simplified mesoscale BD with restraints in (**B**) for N_{sim} fiber copies. ii) Perform corresponding mesoscale MC (with tail/LH interaction) for 10 million steps without restraints. The two replicas of simulated structures are also shown with slightly different overall configurations from those in (**C**) i), and with fully histone tail and linker histone details. (**D**) Simple illustration. I. We first generate N_{rep} replicas with random initial structures ($N_{rep} = 8$). II. Then we distribute restraints to replicas based on the frequencies. III. The $N_{sim} = 8$ structures are folded by simplified BD according to the distributed restraints and IV. updated and with tail and LH details by MC.

linker histones, as well as uses biophysical parameters to resolve non-physical contacts resulting from the artificial BD restraints.

Distance/force parameter choices

The stretching force between the beads in our BD simulation (37) is given by $F_{S_i} = h(l_i - l_0)$, where h is the stretching constant, l_i is the distance between two beads, and l_0 is the equilibrium length. For the restraint forces, $R_{ij} = h_m(d_{ij} - d_0)$, we run simulations with different stretching constants h_m and d_0 to experiment with the effect of different force magnitudes and target equilibrium length. The stretching constant for the connecting DNA beads and nucleosome cores in our mesoscale model is $h = 100k_BT/l_0^2$, where $l_0 = 3$ nm, k_B is the Boltzmann constant, and T is the temperature. For the restraints, we have tested both constant and weighted parameters.

For the constant parameter, we run simulations with $h_m = \frac{h}{50}, \frac{h}{20}, \frac{h}{5}$, and h. When the stretching force is strong ($h_m = h$), the contacted beads will be pulled together strongly and the overall result is a very condensed structure (Supplementary Figure S1); when the stretching force is weak ($h_m = \frac{h}{50}$), the chromatin fiber folds more smoothly and becomes more open (Supplementary Figure S1). Both strong and weak force choices reproduce Micro-C patterns, but strong forces have more long-range contacts than soft forces. Supplementary Figure S2 shows the sedimentation coefficient, packing ratio, volume and radius of gyration of resulting chromatin configurations obtained with different force choices. As the force increases, sedimentation and packing ratios are also higher, and the volume and radius of gyration are lower. For the target equilibrium length, we run simulations with $d_0 = l_0$, $5 \times l_0$, $10 \times l_0$ or $d_0 = 3$, 15, 30 nm respectively. Very similar structures result after BD simulation with different d_0 using the same initial structure, and the same restraints. Runs with smaller d_0 yield slightly more condensed systems compared to larger d_0 (Supplementary Figure S3). In addition, the MC simulation has a greater effect on systems with larger d_0, because open structures can move more freely.

Following this testing, we chose $h_m = \frac{h}{20}$ and $d_0 = 3$ nm as default values. Then we run simulations with weighted parameters based on the distance of the regions (beads). The distance of the beads is denoted by the difference in the bead indices, i.e., bead_i and bead_j have the distance of $|j - i|$. For both $h_{m_{ij}}$ and $d_{0_{ij}}$, we applied the weighted function $h_{m_{ij}} = h_m \times |j - i| \times \alpha$ and $d_{0_{ij}} = d_0 + |j - i| \times \beta$, where α and β are adjustable variables. We analyzed the nucleosome interactions for an artificial 50-nucleosome fiber with different choice of α and β, and found that $\alpha = 0.01$ and $\beta = 0.1$ are good choices for maintaining overall zigzag fibers.

As a result, the final choice for the stretching parameter is $h_{m_{ij}} = 0.05k_BT \times |j - i|/l_0^2$ and $d_{0_{ij}} = 3 + 0.1 \times |j - i|$ nm.

Choice of the number of replicas (N_{rep}) and the number of simulated copies (N_{sim})

The experimental Micro-C data are obtained by accumulating the contacts among millions of cells. To reproduce structures with biological meaning, the choice of N_{rep} should ideally be the same as the number of cells used in the experiments. However, it is computationally prohibitive to run millions of simulations. Hence, we use two strategies to assign N_{rep}.

The first strategy is using a small number (e.g., $N_{rep} = 100$, 200, etc.) as the number of total replicas, scaling the contact frequency (f_{ij}) from the experimental Micro-C data to be between 0 and 1, and generating $N_{rep} = N_{sim}$ copies to assign restraints and run BD simulations. The resulting N_{sim} 3D structures are used to reproduce contact maps to compare to the scaled Micro-C maps.

The second strategy is to use the original experimental Micro-C data and set the number of replicas to be the same as the number of cells used in the experiment (e.g., $N_{rep} = 1M$ for the Micro-C data we used (53)) for the distribution of the restraints. Then we simulate a smaller number of copies (e.g., $N_{sim} = 100 < N_{rep}$) to obtain an ensemble of 3D structures. The resulting N_{sim} 3D structures are used to study the realistic biophysical properties. In this case, reproduction of the Micro-C map is not possible, because we are omitting many contacts by simulating only a portion of the representatives of the whole population. The choice of N_{rep} and N_{sim} is dictated computational resources. An optimal choice would be N_{sim} as close to N_{rep} as possible and N_{rep} around cell size number.

Choice of the chromosome conformation capture data

From a reconstruction point of view, most methods generate polymer models with the number of beads the same as the number of bins on the Hi-C map. Our mesoscale chromatin model has high-resolution particles (9-bp), so it is compatible with Micro-C, which can have resolution as high as 100 bp. Three sets of data are obtained for five gene systems as follows: (i–iii) the NXN (54), HOXC (55,56) and HOXA (57,58) genes in mouse embryonic stem cells (mESC). We focused on these three regions for validation purposes using Micro-C data from (53) (GEO accession: GSE130275) with 100 bp resolution. (iv) We study the HOXA gene region for the enhancer–promoter contact to explore the role of Pol II pausing and use Micro-C from (59) (GEO accession: GSE206131) at 200 bp resolution. (v) We study the Fbn2 gene region to understand cohesin and transcription inhibition roles on chromatin architecture using Region Capture Micro-C (RCMC) data from (12) (GEO accession: GSE207225) at 50 bp resolution.

Model assessment

Although some efforts have been made to correlate nucleosome resolution modeling with super-resolution 'immuno-OligoSTORM' images (60), to date, there are seldom gene-level experimental structures to compare with reconstructed 3D structures. To validate and evaluate the accuracy of our method, we calculate the Spearman correlation coefficient (SCC) (61) between the Micro-C map and our reconstructed contact map. We calculate the root mean square deviation (RMSD) between our reconstructed 3D structures at different stages (before and after MC) or with 3D structures predicted from other methods (e.g., HOXC (51), Supplementary Figure S3). Finally we calculate chromatin volume concentrations (CVC = $V_{\text{chrom}}/120$ nm^3) and compared them with the experimental values (62) in SI.

SCC

SCC is a common measure for evaluating the correlation between two matrices (scaled from –1 to 1), and it is given by:

$$SCC = \frac{\sum_{i=1}^{n}(rx_i - \overline{rx})(ry_i - \overline{ry})}{\sqrt{\sum_{i=1}^{n}(rx_i - \overline{rx})^2 \sum_{i=1}^{n}(ry_i - \overline{ry})^2}} \quad (2)$$

where rx_i and ry_i are the rank variables of the two sets of data, and $\overline{rx}$ and $\overline{ry}$ are the means of rank.

SCC applies to ranked data, which is useful for non-normally distributed variables, and is robust to outliers. 'Moderate', 'Strong' and 'Very strong' correlations correspond to SCC beyond 0.4 (to 0.7), 0.7 (to 0.9) and 0.9 (to 1), respectively.

RMSD

RMSD is given by:

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}||v_i - w_i||^2} \quad (3)$$

where $\boldsymbol{v}$ and $\boldsymbol{w}$ are two sets of points. We compare structures based on the positions of the nucleosome cores.

Results

Hi-BDiSCO method

Full details of our reconstruction method are described in 'MATHERIALS AND METHODS' and Figure 3. The essential ideas are: (i) build the model for the system of interest by using experimental nucleosome positions, acetylation islands, and LH positions and densities; (ii) select a number of replicas N_{rep} (preferably in the order of cells used to produce the Micro-C data), and scale the Micro-C map based on the choice of N_{rep}, so that the contact frequencies (f_{ij}) are between 0 and 1; (iii) distribute distance restraints derived from the (scaled/original) Micro-C map into these replicas to reproduce the experimental information; (iv) perform BD simulations with restraints as added force, without LH and histone tails, for N_{sim} replicas, where $N_{sim} \leq N_{rep}$; (v) perform MC simulations with restraints released and histone tails and LH included to resolve clashes/outliers due to the omitted contacts in the simulated sample.

The resulting ensemble of MC structures is our representation of the genome structures compatible with the experimental data. To assess the structures where reasonable (i.e., $N_{sim} \approx N_{rep}$), we compute the Spearman correlation coefficient (SCC) between an average matrix corresponding to our N_{sim} systems and the scaled experimental map. SCC is a statistical covariance measurement of the correlation between two matrices (see section 'Model assessment').

BD reconstructed 3D structures reproduce the patterns of experimental Micro-C Maps

To validate the performance of our Hi-BDiSCO reconstruction method, we use the scaled Micro-C map to generate and simulate $N_{rep} = N_{sim} = 100$ for each gene system (NXN, HOXC and HOXA), with different random initial structures and restraints as described in 'MATERIALS AND METHODS'. We choose $N_{sim} = N_{rep}$ to capture all contacts derived from the Micro-C map, so that the calculated contact map can be assessed by SCC calculations. We run the BD simulations until the systems converge (the energy difference between two consecutive steps is less than 10^{-21} J). We then use the reconstructed structures to compute the contact map (the choice of cut-offs is described in the subsection 'Distance/Force Parameter Choices' under 'Materials and methods'), and the frequencies are averaged among all the replicas.

Figure 4 displays the resulting configuration replicas and comparisons to the scaled Micro-C data. The reconstructed contact map (upper right triangle) reproduces qualitatively the patterns (blocks and stripes and frequencies) of the original experimental Micro-C map (bottom left triangle). The computed Spearman correlation coefficients (SCCs) (61) between these two matrices for each system are >0.74, indicating a strong correlation (reconstructions with SCC >0.7 are considered to be accurate (63)). For validation purposes, we use $N_{rep} = N_{sim} = 100$ replicas to reproduce scaled frequencies. We additionally tested $N_{rep} = N_{sim} = 1000$ and obtained an SCC = 0.838 (see next subsection). The SCC value also shows that the performance of Hi-BDiSCO is consistent for the three systems.

We see from the matrices in Figure 4 that the simulation results are darker (higher frequency) than the experimental values for the near-diagonal region. This is because scaling the Micro-C map to correspond to 100 replicas instead of 1 million cells increases the contact frequencies f_{ij}, making the reconstructed structures compact. By simulating more replicas (see next subsection), results would be more reasonable.

From the reconstructed 3D structures, we see that representative structures vary, reflecting the chromatin heterogeneity across cells and fluidity within the cell.

The time performance of BD reconstruction (averaging among 100 replicas) in Table 1 show that although the wall time increases with system size due to the data transfer and some functions with linear time (e.g., Cholesky decomposition) (37), the longest simulation only takes 37.5 minutes.

Increasing N_{rep} and N_{sim} improves the accuracy of the reconstruction and subsequent MC fixes structural anomalies

In the prior subsection, our Hi-BDiSCO runs with $N_{rep} = 100$ from the scaled Micro-C map yielded good performance on assessing 2D contact maps. Here we show how reconstruction changes with N_{rep} due to scaling.

When N_{rep} is small, each replica receives more restraints during distribution, which results in more compact structures. Simulating 1 million copies is not feasible, but here we run $N_{rep} = N_{sim} = 200$ and $N_{rep} = N_{sim} = 1000$ for three gene systems to assess how performance changes with increasing N_{rep}. That means that the same restraints are distributed into more copies. As shown in Figure 5A and Table 1, SCC increases from 0.784 (Figure 4, where $N_{rep} = N_{sim} = 100$) to 0.800 and 0.838 for NXN, from 0.740 to 0.786 and 0.823 for HOXC, and from 0.745 to 0.775 and 0.822 for HOXA, when N_{rep} is 200 and 1000, respectively. The concomitant decrease in short-range contact frequencies is also evident.

The representative structures for each system are also shown in Figure 5A. With $N_{rep} = 200$ (left), fibers are more compact than those with $N_{rep} = 1000$ (right). To quantitatively compare these structures, we calculated their volume and radius of gyration, prior to MC simulations. As shown

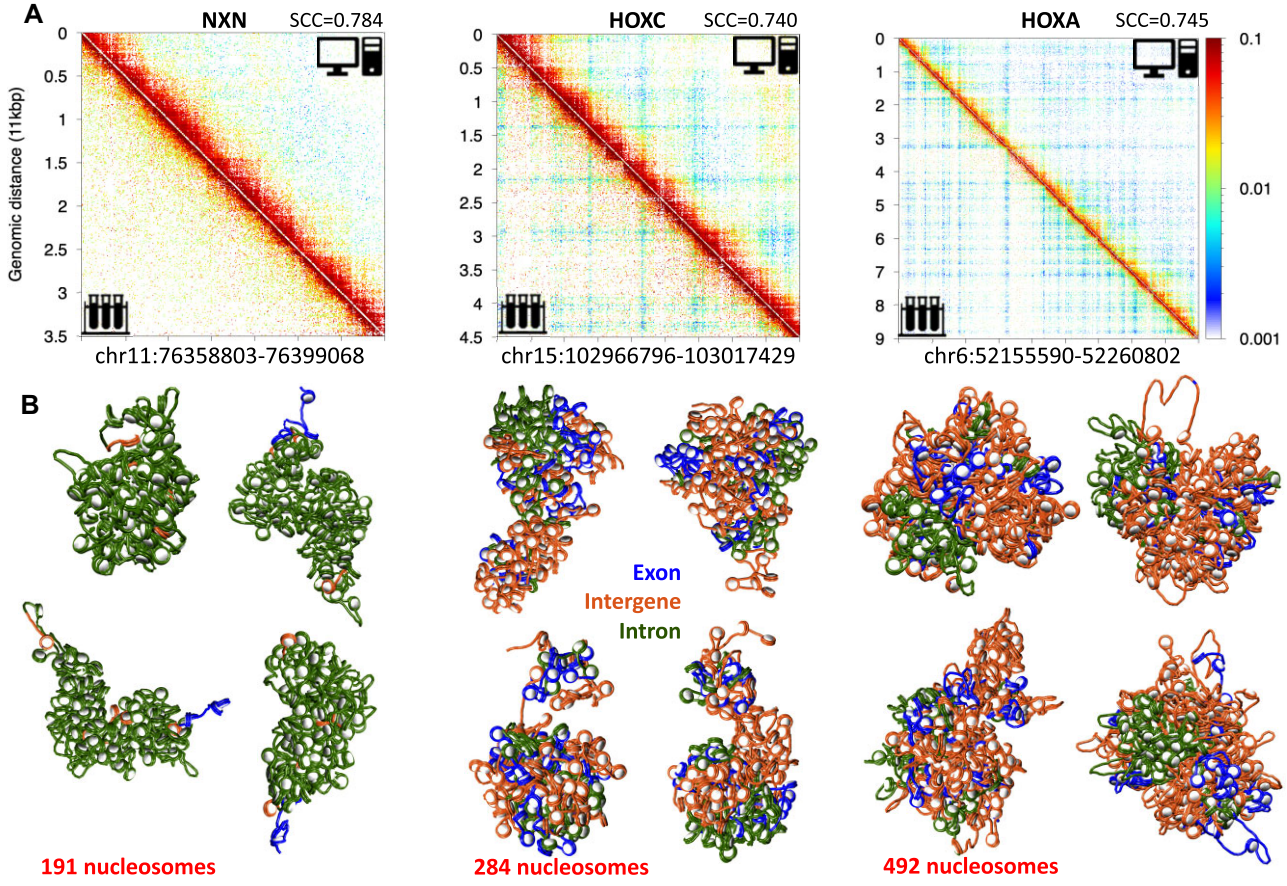


Figure 4. Evaluation of Hi-BDiSCO reconstructed genome systems, with $N_{rep} = N_{sim} = 100$ replicas, prior to MC simulations. (**A**) For each gene system (NXN, HOXC and HOXA), the lower triangle of the contact map is the scaled experimental Micro-C data, and the upper triangle of the contact map is calculated from BD reconstructed structures, along with the Spearman correlation coefficients (SCC) (see Materials and methods). All gene systems show good reproduction of the frequency (colors) and patterns. (**B**) For each gene, we also show representative reconstructed 3D structures colored with exon (blue), intron (dark green) and intergene (red).

Table 1. Performance of BD reconstruction for different gene systems

Gene	System size (kbp)	$N_{rep} = N_{sim}$	Steps for convergence	Time (min)	SCC
HOXA	105.2	100	11963	37.52	0.745
		200			0.775
		1000			0.822
HOXC	50.6	100	11826	12.18	0.740
		200			0.786
		1000			0.823
NXN	40.3	100	13151	10.08	0.784
		200			0.800
		1000			0.838

in Figure 5B, both properties increase with N_{rep}. Thus, the reconstruction is more accurate as N_{rep} and N_{sim} increase.

Although we studied scaled Micro-C maps with $N_{sim} = N_{rep}$ to validate the reconstruction, more realistic structures could be obtained if unscaled Micro-C data are used with N_{rep} close to the number of cells used in the experiment (usually millions). Simulating such large number of replicas is not feasible, but as discussed in the next subsection, we can simulate a subset $N_{sim} < N_{rep}$ to obtain representative structures.

Choosing $N_{sim} < N_{rep}$ and N_{rep} near cell population size yields good reconstruction

Although we have shown that BD simulations with Micro-C restraints can artificially fold 3D chromatin fibers to yield agreement with the 2D Micro-C map, realistic structures without artificial restraints and satisfying biophysical conditions are needed. Since the choice of N_{rep} and the scale of the Micro-C maps affects the compaction level of the resulting structures, here we choose different N_{rep} depending on the original (non-scaled) Micro-C map to study the performance of subsequent MC simulations applied to each system with different compaction levels.

The experimental Micro-C map (53) corresponds to 1 million cells. For the three gene systems, during the restraints distribution (as described in 'MATERIALS AND METHODS'), we choose: $N_{rep} = 2M$ (larger population than experimental) for NXN; $N_{rep} = 1M$ (experimental population) for HOXC; and $N_{rep} = 100$ (smaller population than experimental) for the HOXA gene. However, for each system's simulations (BD followed by MC), we select only $N_{sim} = 100$ out of these total copies, effectively neglecting some contacts when $N_{sim} < N_{rep}$.

As shown in Figure 6A, the subsequent MC simulations maintain the Micro-C contact patterns for all three gene systems, despite the lower number of simulated fibers. In addition, structural features, such as hierarchical loops (50,64), remain after the MC simulations.

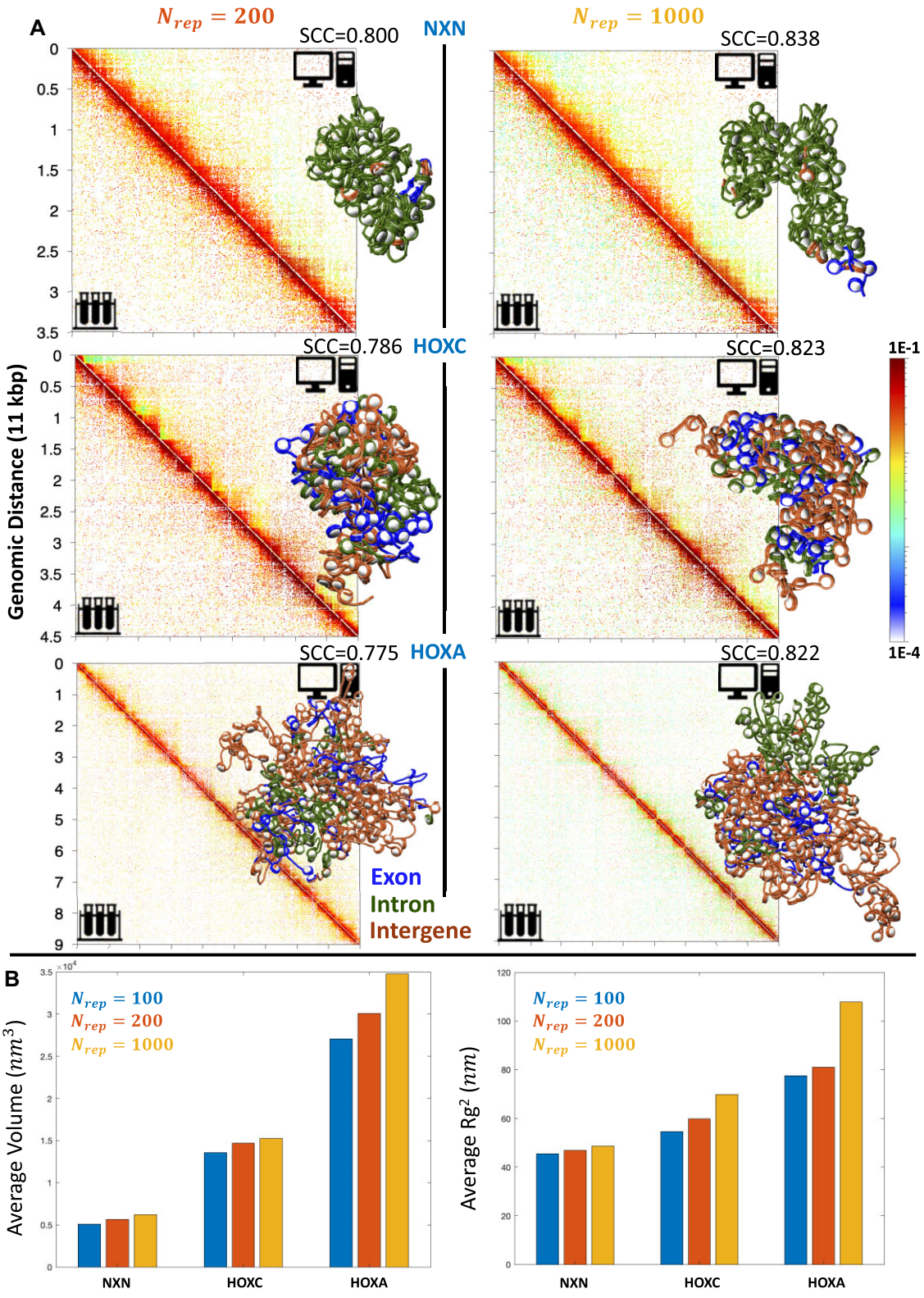
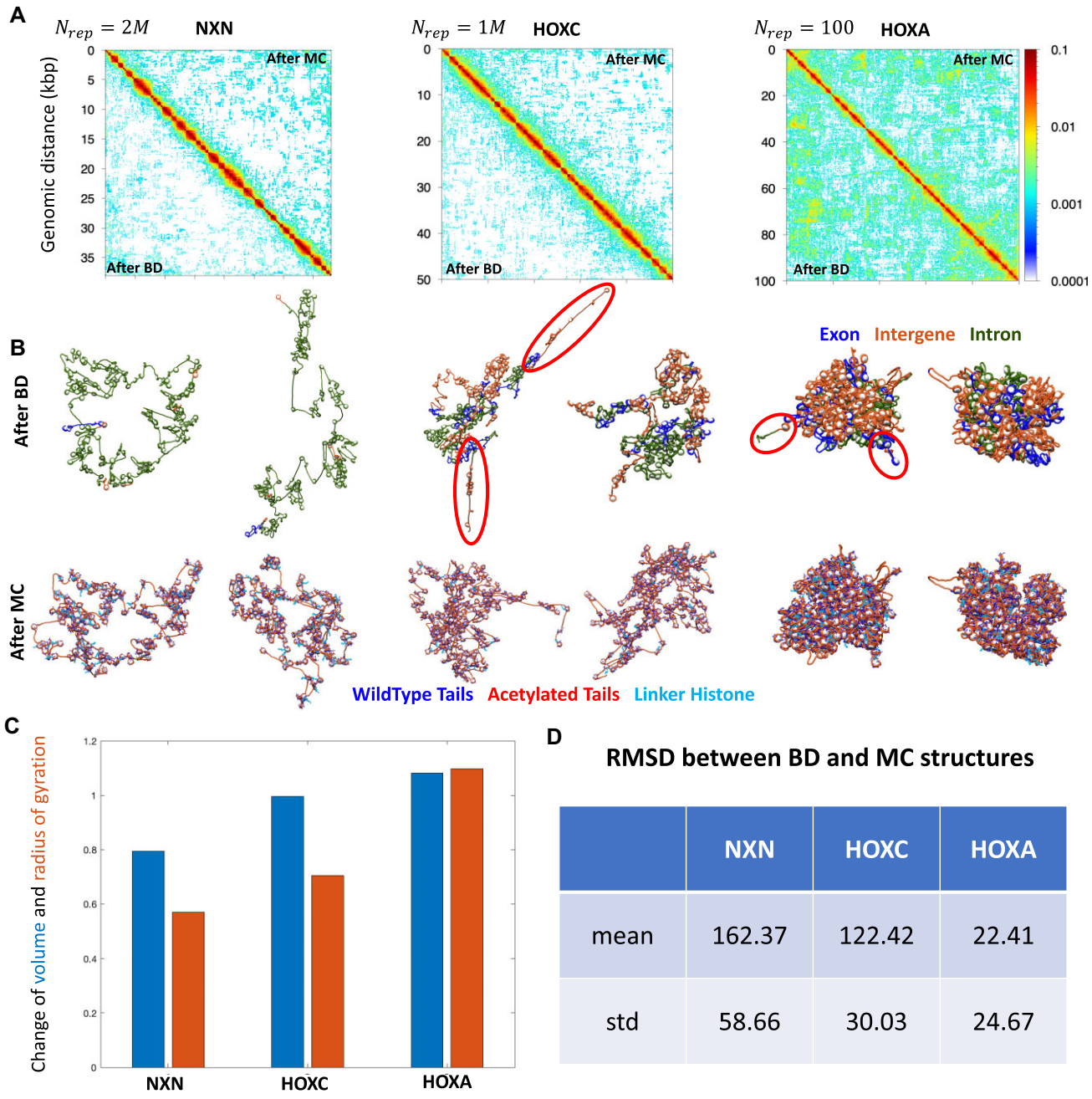


Figure 5. Hi-BDiSCO reconstruction of the three genes of Figure 4 (where $N_{rep} = N_{sim} = 100$) with larger replica values $N_{rep} = N_{sim} = 200$ and 1000. (**A**) Comparison between scaled experimental Micro-C data and BD reconstructed contact maps, prior to MC simulations. The lower triangle is scaled experimental Micro-C data, and the upper triangle of the contact map is calculated from BD reconstructed structures. Spearman correlation coefficient (SCC) values are shown for each case. Representative 3D structures from $N_{rep} = 200$ and $N_{rep} = 1000$ can be compared to Figure 4, when $N_{rep} = 100$. (**B**) The volume and radius of gyration for three different N_{rep} values (100, 200 and 1000) as averaged across the first 100 replicas, prior to MC simulations.



	NXN	HOXC	HOXA
mean	162.37	122.42	22.41
std	58.66	30.03	24.67

Figure 6. Comparison between reconstructed chromatin conformations before and after MC simulations for three gene systems (NXN, HOXC and HOXA) with choices of N_{rep} based on experimental context (see text). (**A**) contact maps calculated after BD (lower triangle) and after MC (upper triangle). (**B**) Representative 3D structures from different replicas (exon (blue), intron (dark green) and intergene (red)) after BD and after MC (wildtype histone tails (blue beads), acetylated histone tails (red beads) and LH (teal beads)). The red circles for HOXC and HOXA represent the dangling fiber ends as unphysical spatial problems. (**C**) Change of volume (blue) and radius of gyration (red) for three gene systems. (**D**) RMSD that represents the structural changes in the three gene systems.

Figure 6B shows that after the BD reconstructions, the NXN configurations are open structures, whereas the HOXA configurations are compact. This is because $N_{rep} = 2M$ for NXN and thus few restraints are included in the 100 simulated copies, but for HOXA, all restraints are included in the simulated copies. After MC, however, NXN structures become more compact, and HOXA structures become more open.

To quantitatively analyze the effects of MC refinement on the systems, we calculate the change in volume and radius of gyration (Figure 6C) and RMSD (Figure 6D) after MC. For the NXN gene system, MC reduces the volume because N_{rep} is larger than the experimental number of cells. For HOXA, the condensed fibers, due to small N_{rep}, are decondensed by MC. HOXC with N_{rep} the same as the experimental number of cells shows unchanged volume after MC. The radii of gyration in Figure 6C and RMSD values further demonstrate that MC following BD makes subtle modifications for the over-compact structures (HOXA). Note that we cannot assess the structures by SCC because $N_{sim} \ll N_{rep}$.

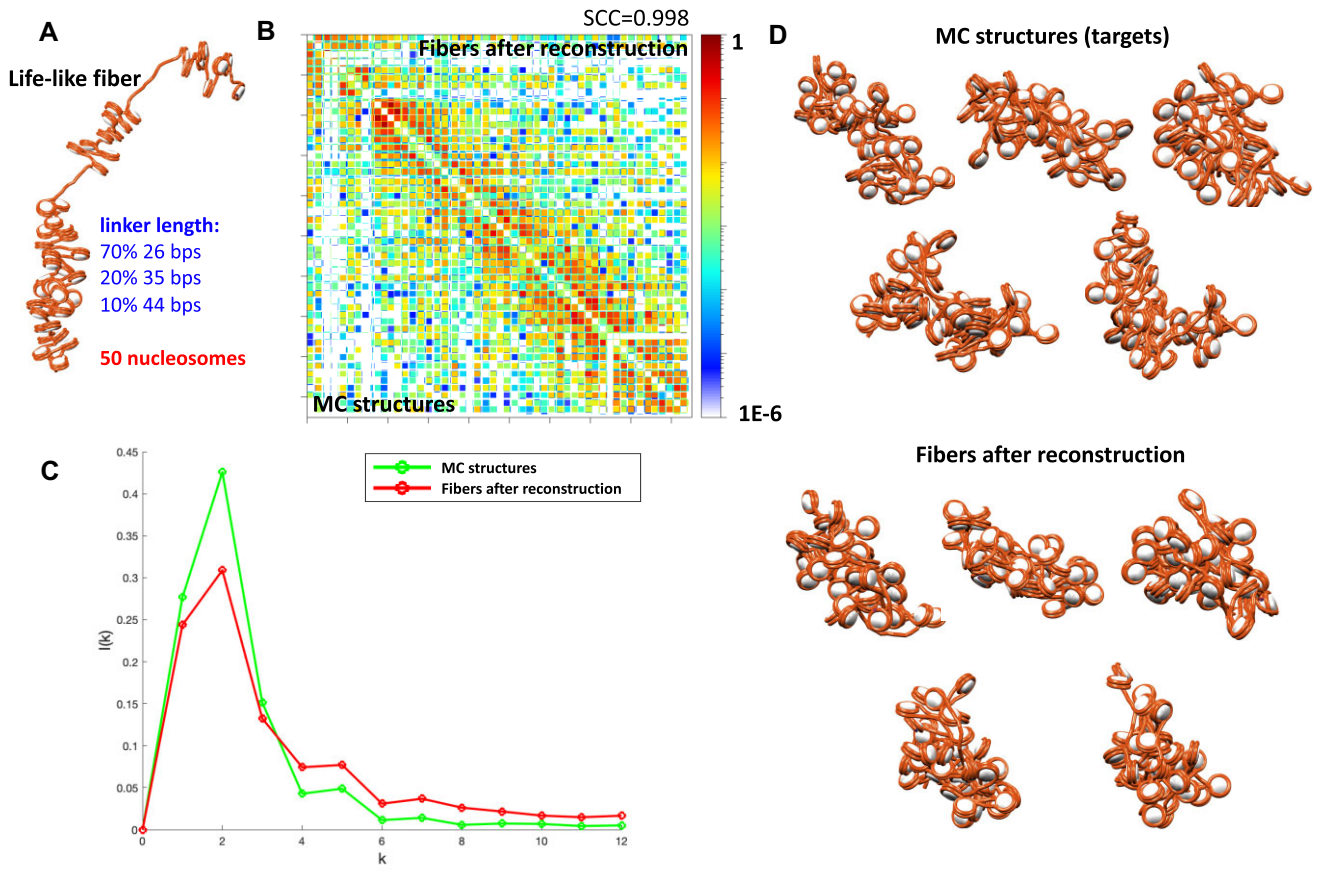


Figure 7. Evaluation of Hi-BDiSCO reconstruction for a designed life-like 50-nucleosome fiber. (**A**) Target life-like fiber with a distribution of linker lengths motivated by experimental patterns (structures shown in D). (**B**) Comparison between contact maps of 100 initial random structures (bottom triangle) and reconstructed structures (top triangle) along with Spearman correlation coefficient (SCC). (**C**) Nucleosome interactions of 100 initial random structures (green) versus reconstructed structures (red). (**D**) Five representative configurations of initial random structures and reconstructed structures.

Thus, MC simulations after BD are essential for our Hi-BDiSCO reconstruction to incorporate biophysical features and resolve unphysical problems. The optimal choice of N_{rep} is the number of cells used in the experiment, but quantitative measurements of reconstruction are still good with scaled input data and $N_{sim} = 100$. To further validate that N_{rep} near cell population size provides an optimal reconstruction, we compare calculated to experimental CVC values for the three gene systems in Supplementary Table S2.

Because we simulated the folding of the HOXC gene de novo without Micro-C data (previously in (51)), we also compare this de novo prediction to Hi-BDiSCO reconstructed structures in SI; there, we show that Hi-BDiSCO and MC predicted genome structures are overall quite similar.

Reconstruction of known structures demonstrates biophysically sound fibers

So far we have assessed Hi-BDiSCO primarily by comparing the contact maps, but many 3D structures can map onto the same 2D contact map. Because there are no high-resolution experimental comparisons for our 3D nucleosome resolution gene-level structures, simulated life-like 3D structures provide a reasonable choice for Hi-BDiSCO validation. We generate a 100-configuration ensemble of life-like fibers (70% of linkers are 26 bp, 20% are 35 bp, and 10% are 44 bp) (65) by MC sampling with 50 nucleosomes using random initial zigzag structures (with wildtype tails and without LH). We then compute the contact map for these 100 random replicas and use this as a target 'Micro-C data'. From Figure 7B, which compares the contact maps of 100 initial random structures (bottom triangle) to that of reconstructed structures (top triangle), we note excellent agreement, with SCC=0.998. This indicates that Hi-BDiSCO can reproduce the original 3D structures very well. To analyze our nucleosome resolution structure further, we compare the nucleosome-nucleosome interaction frequencies before and after reconstruction in Figure 7C. We see zigzag peaks at $k \pm 2$, and slightly more long-range interactions for the reconstructed fibers. The peaks at $k \pm 2 \approx 0.3$ agree with the dominant zigzag structure in discrete life-like fibers (64). Thus, Hi-BDiSCO reconstructed fibers are biophysically meaningful. The five representative structures in Figure 7D appear slightly more compact after reconstruction than the initial random structures.

See SI for more information on life-like fibers and nucleosome-nucleosome interactions.

HOXA gene cluster regulation by RNA Pol II

Beyond applying our nucleosome resolution mesoscale model to reconstruct 3D structures with biophysically reasonable zigzag patterns, we also apply Hi-BDiSCO to help interpret biological mechanisms such as enhancer–promoter interactions.

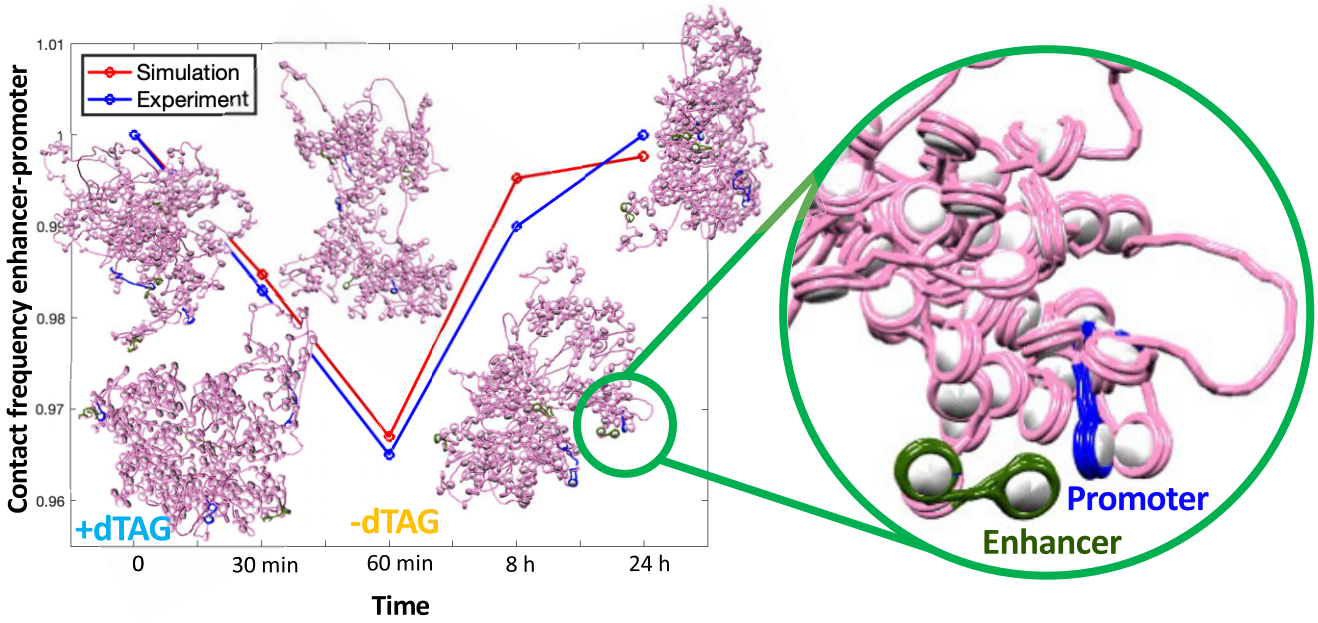


Figure 8. Regulation of enhancer–promoter interactions by RNA Pol II assessed by Hi-BDiSCO reconstruction. Left. Contact frequency plot for enhancer–promoter interactions at increasing time of RNA Pol II inhibition were determined experimentally and by Hi-BDiSCO. The plots compare total contact values for experimental and simulation results for 11 enhancer–promoter interactions in the system (relative to the untreated $T = 0$ control) at different time snapshots (0, 30 min, 60 min, 8 h and 24 h). Representative 3D structures are also shown colored with enhancers (green) and promoters (blue) of the HOXA genes, as enlarged at right.

How enhancers regulate target gene expression across large genomic distances remains unclear. Gilad *et al.* (59) explore the role of Pol II pausing by introducing dTAG, a Pol II inhibitor, and removing it after an hour. They record the contacts at five time points (0, 30 min, 60 min, 8 h and 24 h) with Micro-C, as shown in Supplementary Figure S4. From the experiment, Gilad et al. observed a decreased number of enhancer–promoter interactions after adding dTAG, but increased after removing dTAG.

We applied Hi-BDiSCO to study how chromatin folding regulates the transcriptional activity of the HOXA gene cluster as follows. We reconstructed 3D structures, with $N_{rep} = 1M$ and $N_{sim} = 100$, for each of the 5 Micro-C maps (different times). For the structures corresponding to each time assessed, we measure the enhancer–promoter contacts relative to the untreated T=0 control for all enhancer–promoter contacts as follows. Enhancer-promoters are considered in contact if they are <50 nm from one another. Each gene in the HOXA cluster has one promoter and one enhancer. Thus, in total, there are 11 distances for 11 enhancer–promoter pairs. As shown in Figure 8, we reproduce the same trend as captured experimentally, supporting the conclusion that paused Pol II decreases enhancer–promoter contact levels.

Reconstruction of Fbn2 gene reveals cohesin and RNA Pol II roles on chromatin architecture

Goel et al. recently applied Region Capture Micro-C (RCMC) (12), a technique that improves Micro-C data resolution to 50 bp, to study genome organization. In particular, they study several gene regions, such as Sox2, Klf1 and Fbn2 in mESC, and obtained highly nested 3D interactions showing microcompartments. In their first experiment, they compared a system without cohesin (−cohesin) to a control system (+cohesin control). In their second experiment, they inhibited transcription by removing RNA Pol II and determining genome architecture at increasing times of RNA Pol II elimination treatment. Thus, three RCMC datasets were produced: wildtype data; RNA Pol II elimination for 45 min (−Pol II 45 min); and RNA Pol II elimination for 4 h (−Pol II 4 h).

We used Hi-BDiSCO (with $N_{rep} = 1M$ and $N_{sim} = 100$) to reconstruct 3D structures of the Fbn2 gene region (100 kb, chr18:58110000..58210000) under the five conditions (Figure 9A, B). By studying the Fbn2 structure under different conditions, we can assess the role of cohesin and RNA Pol II on chromatin architecture. To quantify Fbn2 architecture, we calculate the average volume and radius of gyration among the 100 simulated replicas (Figure 9C), and the interaction frequencies against genomic distance (Figure 9D).

As shown in Figure 9A, experimental contact maps for '−cohesin' and '+cohesin control' appear similar, but the SCC between these two experimental maps is around 0.82, indicating subtle differences. Hi-BDiSCO reconstruction suggests what these differences are in fiber condensation. As shown in Figure 9C and D, after cohesin depletion, there is a volume increase and a decrease of long range interactions (internucleosome interaction frequencies plotted as interaction frequency versus genomic distance). Cohesin depletion induces a loss of loops and thus results in more open fibers.

For the study of transcription inhibition, Figure 9B shows a reduction of contacts in the contact maps when we compare the wildtype system at 45 min and 4 hr treatment conditions. The reconstructed 3D structures (Figure 9B) and the volume change (Figure 9C) show the decreased fiber compaction upon RNA Pol II inhibition. As discussed in the prior subsection for the HOXA region, a reduction of enhancer–promoter interactions is due to the pausing/removing of Pol II, and thus the fiber is less compact. When comparing chromatin global

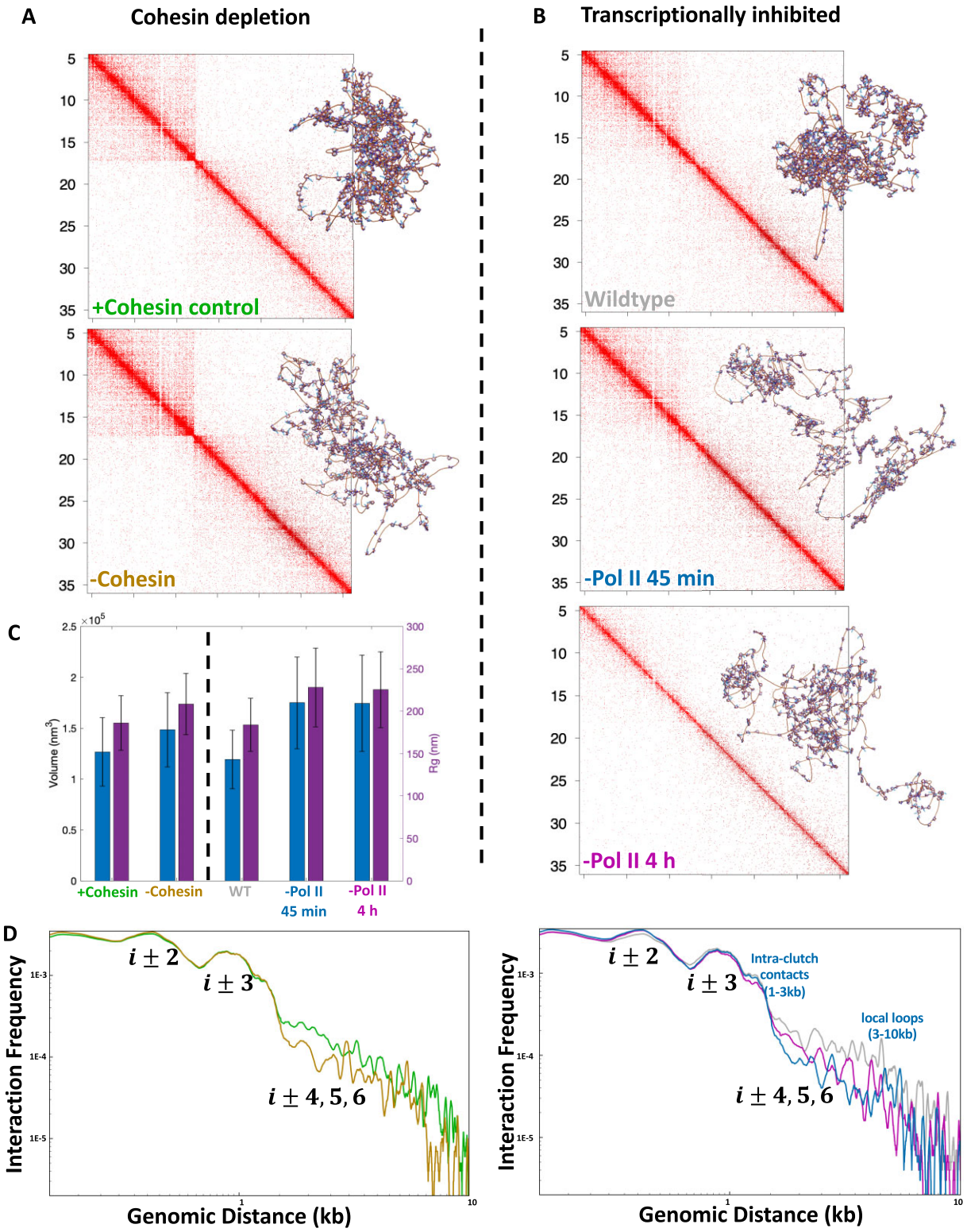


Figure 9. Role of cohesin and RNA pol II in genome architecture evaluated by Hi-BDiSCO reconstruction. A and B. Experimental Region Capture Micro-C (RCMC) maps (12) and representative Hi-BDiSCO reconstructed 3D structures for part of the Fbn2 gene region (chr18:58110000..58210000). In (**A**), +cohesin control map is shown at top, and cohesin depletion treatment (−cohesin) map at bottom. In (**B**), wildtype control map is shown at top, the inhibition of transcription by removing RNA Pol II for 45 min in the middle, and inhibition of transcription map by removing RNA Pol II for 4 h at the bottom. (**C**) Associated volume and radius of gyration (Rg) for each system, where blue represents the volume and purple represents the Rg for each system. (**D**) Analysis of internucleosome interaction contacts for each system. The left shows the +cohesin (green) and −cohesin (brown); the right shows the wildtype control (grey), −Pol II 45 min (blue) and −Pol II 4 h (purple).

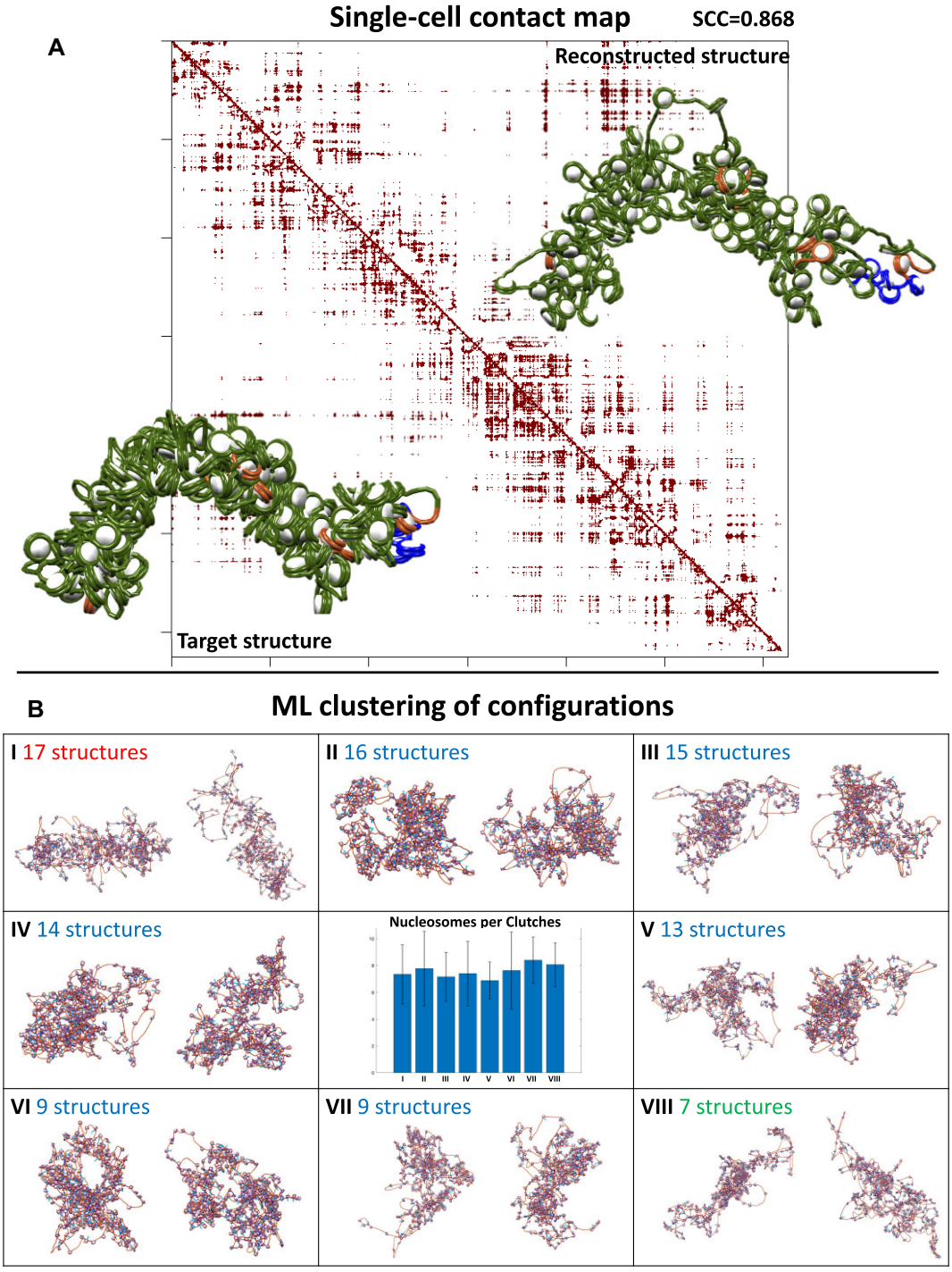


Figure 10. Further applications of Hi-BDiSCO to single cell data and ensemble analysis. (**A**) Contact map derived for a single structure is used to demonstrate scHi-C reconstruction. Bottom left: target structure and related contact map. Top right: reconstructed structure and contact map. (**B**) Unsupervised clustering of wildtype Fbn2 gene reconstructed structures using a *K*-means clustering algorithm in the Python scikit-learn package. Two representative structures for each of the 8 distinct clusters resulting of the 100 structures are shown. Nucleosome clutch analysis (middle) shows the average number of nucleosomes per clutch in each cluster as one feature that can be analyzed (see SI for more).

structure at 45 min and 4 h treatment, not much difference is seen; the volume and radius of gyration do not change. However, intra-clutch contacts increase, and local loop contacts decrease (Figure 9D).

These results demonstrate that the structural information obtained from the reconstructed 3D structures can complement and expand the information obtained from 2D maps, underscoring the utility of 3D reconstruction in studying genome organization and function.

Discussion

Many reconstruction methods are now available (21,23) aiming to bridge the 2D chromosome conformation capture data with 3D genome organization, including gene folds, TADs, compartments and whole genomes. We have developed an efficient nucleosome-resolution strategy for folding gene-level chromatin fibers based on chromosome conformation capture information. Namely, we use a contact-based reconstruction method with a population-based BD sampling algorithm with our nucleosome-level resolution mesoscale model to fold gene-level structures. We have demonstrated the reproduction of Micro-C patterns for several systems, including HOXA, HOXC and NXN, with different choices of replicas (N_{rep}) to distribute the restraints into, and different choices of simulated copies ($N_{sim} \leq N_{rep}$). We showed that the BD simulation without considering tails/LHs is sufficient to reproduce the contact frequencies of the scaled Micro-C map when scaled data are used with $N_{sim} = N_{rep}$; alternatively, $N_{sim} \approx N_{rep}$ with N_{rep} similar to the number of cells used in the experiments yields good SCC values for the original Hi-C map. The subsequent MC simulation is essential to resolve any spatial problems and provide biological context by adding histone tails and linker histones so that the final structures provide realistic insights into the gene folding motifs and mechanisms. The strategy itself is not limited to our mesoscale model and can be applied to any 3D chromatin/chromosome model.

By studying the enhancer–promoter interaction of the HOXA gene region with paused RNA Pol II, and exploring the roles of cohesin and transcription inhibition of chromatin architecture of the Fbn2 gene region, we showed that our reconstructed structures can reproduce the information from 2D maps and provide insights into structural features and mechanisms.

Besides Hi-BDiSCO, other reconstruction approaches with nucleosome or near-nucleosome resolution include MiOS (60) and POSSUMM (66). As in Hi-BDiSCO, both methods consider bioinformatics data (in different ways), such as Chip-seq data, and use Hi-C/Micro-C data as a reference to guide the chromatin folding. Although the Hi-C/Micro-C maps used have different resolutions, the resulting structures have similar physical properties. Clearly, the problem remains that the reconstructed structural ensemble is only a possible subset to many possible real-life structures that correspond to the Hi-C/Micro-C data. In our recent review (23), we discussed the validation of the reconstructed structures by statistical covariances (SCC or PCC close to 1), or low-to-high resolution microscopy (e.g., FISH (28,67), cryo soft X-Ray (68) and OligoSTORM (60)). However, many reconstructed genome structures can reproduce similar Hi-C/Micro-C maps, and the resolution of the experimental structures is much lower than the resolution of the reconstructed 3D structures. Statistically, all such models may provide valuable structural properties and associated mechanisms. See (63) for extended discussion on examples of structural assessment.

From a modeling point of view, challenges for 3D reconstructed models from 2D maps remain, such as: the lack of a true structure to compare with; obtaining high resolution 3D models of large genomes from Hi-C data; simulating millions of replicas to reproduce cell populations; and handling biases and noises of the Hi-C maps. Our approach is most suitable for Micro-C data and fibers at kb lengths, and currently uses the raw Micro-C data without correction. From these Micro-C data, we generate an ensemble of structures that aims to represent structures in single cells. Although single-cell Hi-C (scHi-C) data are available, the resolution is not sufficiently high for Hi-BDiSCO. Yet it is possible to create structures for scHi-C data, as done, for example, in Figure 10A, where one Hi-C map corresponds to one structure (see Supplementary Figure S5 for more details). We are also limited by the size that MC and BD can handle (about 500 nucleosomes). To study larger systems like TADs, compartments, or even whole genomes, a further coarse-grained model (or polymer model) can be used with the same strategy to distribute Hi-C data as restraints. Data interpolation is an alternative strategy for incorporating lower-resolution data. With such adjustments, larger systems could be studied to explore compartment mechanisms (66). In addition, clustering the resulting configurations by machine learning algorithms can also help analyze cluster features. Our clustering of structures for the Fbn2 gene in Figure 10B (see Supplementary Figure S6 for details) reveals sub-population of structures that might correlate with different cell subtypes. Overall, such representative ensembles of 3D structures at the mesoscale level can help shed insight into important chromosomal activity and structure/function relationships.

Data availability

The working Hi-BDiSCO code has been deposited in GitHub under the Schlick lab group: https://github.com/Schlicklab/Hi-BDiSCO, and in Zenodo: https://doi.org/10.5281/zenodo.8400541. There, we provide Python scripts for converting experimental data (e.g., MNase, Chip-seq, Hi-C/Micro-C) into the input structures of our mesoscale structures and restraints, and the binary executables for BD reconstruction and MC subsequent simulations. The executable code runs on the popular Linux CentOS 7, RedHat/Roky 7 and 8, Ubuntu 20 and 22 platforms and works with both Intel and AMD CPUs.

Users may incorporate the provided example to test the code or replace the experimental data with their data of choice in the 'data' folder and run the provided shell scripts to perform simulations and obtain reconstructed 3D structures. If no MNase or Chip-seq data are provided, life-like fibers (with user defined density of tail acetylation and LH) will be generated for Hi-BDiSCO to perform the simulation.

More details are provided in the README on the GitHub repository. Interested users may email us for further help.

Supplementary Data

Supplementary Data are available at NAR Online.

Acknowledgements

We gratefully acknowledge support from the Simons Foundation through the NYU Simons Center for Computational Physical Chemistry. This work was supported in part through the NYU IT High Performance Computing (HPC) group, which provided resources, services and expert advice. We thank Dr Stephanie Portillo-Ledesma for critical discussions and careful reading of the manuscript, and Shenglong Wang from the NYU HPC group for helping prepare a distributable code.

Funding

National Institutes of Health, National Institutes of General Medical Sciences [R35-GM122562]; National Science Foundation RAPID Award [2030377] from the Division of Mathematical Sciences and Chemistry, and Award 2151777 from the Division of Mathematical Sciences, and Philip-Morris International (to T.S.).

Conflict of interest statement

None declared.

References

1. De Laat,W. and Grosveld,F. (2003) Spatial organization of gene expression: The active chromatin hub. *Chromosome Res.*, **11**, 447–459.
2. Spencer,V.A., Xu,R. and Bissell,M.J. (2010) Gene expression in the third dimension: The ECM-nucleus connection. *J. Mammary Gland Biol. Neoplasia*, **15**, 65–71.
3. Dekker,J. (2008) Gene regulation in the third dimension. *Science*, **319**, 1793–1794.
4. Dekker,J., Marti-Renom,M.A. and Mirny,L.A. (2013) Exploring the three-dimensional organization of genomes: Interpreting chromatin interaction data. *Nat. Rev. Genet.*, **14**, 390–403.
5. Gorkin,D.U., Leung,D. and Ren,B. (2014) The 3D genome in transcriptional regulation and pluripotency. *Cell Stem Cell*, **14**, 762–775.
6. Portillo-Ledesma,S. and Schlick,T. (2020) Bridging chromatin structure and function over a range of experimental spatial and temporal scales by molecular modeling. *WIREs Comput. Mol. Sci.*, **10**, e1434.
7. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.
8. Simonis,M., Klous,P., Splinter,E., Moshkin,Y., Willemsen,R., De Wit,E., Van Steensel,B. and De Laat,W. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.*, **38**, 1348–1354.
9. Dostie,J., Richmond,T.A., Arnaout,R.A., Selzer,R.R., Lee,W.L., Honan,T.A., Rubio,E.D., Krumm,A., Lamb,J., Nusbaum,C., *et al.* (2006) Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, **16**, 1299–1309.
10. Lieberman-Aiden,E., Van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O., *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
11. Hsieh,T.S., Fudenberg,G., Goloborodko,A. and Rando,O.J. (2016) Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome. *Nat. Methods*, **13**, 1009–1011.
12. Goel,V.Y., Huseyin,M.K. and Hansen,A.S. (2023) Region capture Micro-C reveals coalescence of enhancers and promoters into nested microcompartments. *Nat. Genet*, **55**, 1048–1056.
13. Ulianov,S.V., Tachibana-Konwalski,K. and Razin,S.V. (2017) Single-cell Hi-C bridges microscopy and genome-wide sequencing approaches to study 3D chromatin organization. *BioEssays*, **39**, 1700104.
14. Chi,Y., Shi,J., Xing,D. and Tan,L. (2022) Every gene everywhere all at once: high-precision measurement of 3D chromosome architecture with single-cell Hi-C. *Front. Mol. Biosci.*, **9**, 959688.
15. Bonora,G., Ramani,V., Singh,R., Fang,H., Jackson,D.L., Srivatsan,S., Qiu,R., Lee,C., Trapnell,C., Shendure,J., *et al.* (2021) Single-cell landscape of nuclear configuration and gene expression during stem cell differentiation and X inactivation. *Genome Biol.*, **22**, 279.
16. Flyamer,I.M., Gassler,J., Imakaev,M., Brandão,H.B., Ulianov,S.V., Abdennur,N., Razin,S.V., Mirny,L.A. and Tachibana-Konwalski,K. (2017) Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, **544**, 110–114.
17. Nagano,T., Lubling,Y., Várnai,C., Dudley,C., Leung,W., Baran,Y., Mendelson Cohen,N., Wingett,S., Fraser,P. and Tanay,A. (2017) Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, **547**, 61–67.
18. Zhen,C., Wang,Y., Geng,J., Han,L., Li,J., Peng,J., Wang,T., Hao,J., Shang,X., Wei,Z., *et al.* (2022) A review and performance evaluation of clustering frameworks for single-cell Hi-C data. *Brief. Bioinform.*, **23**, bbac385.
19. Tan,L., Xing,D., Chang,C., Li,H. and Xie,X.S. (2018) Three-dimensional genome structures of single diploid human cells. *Science*, **361**, 924–928.
20. Tan,L., Ma,W., Wu,H., Zheng,Y., Xing,D., Chen,R., Li,X., Daley,N., Deisseroth,K. and Xie,X.S. (2021) Changes in genome architecture and transcriptional dynamics progress independently of sensory experience during post-natal brain development. *Cell*, **184**, 741–758.
21. Oluwadare,O., Highsmith,M. and Cheng,J. (2019) An overview of methods for reconstructing 3-D chromosome and genome structures from Hi-C data. *Biol. Proced. Online*, **21**, 7.
22. Belokopytova,P. and Fishman,V. (2021) Predicting genome architecture: challenges and solutions. *Front. Genet.*, **11**, 617202.
23. Li,Z., Portillo,S. and Schlick-Ledesma,T. (2023) Techniques for and challenges in reconstructing 3D genome structures from 2D chromosome conformation capture data. *Curr. Opin. Cell Biol.*, **83**, 102209.
24. Lyu,H., Liu,E. and Wu,Z. (2020) Comparison of normalization methods for Hi-C data. *BioTechniques*, **68**, 56–64.
25. Rosenthal,M., Bryner,D., Huffer,F., Evans,S., Srivastava,A. and Neretti,N. (2019) Bayesian estimation of three-dimensional chromosomal structure from single-cell Hi-C Data. *Comput. Biol.*, **26**, 1191–1202.
26. Hua,K.-J. and Ma,B.-G. (2019) EVR: reconstruction of bacterial chromosome 3D structure models using error-vector resultant algorithm. *BMC Genomics*, **20**, 738.
27. Trieu,T., Oluwadare,O. and Cheng,J. (2019) Hierarchical reconstruction of high-resolution 3D models of large chromosomes. *Sci. Rep.*, **9**, 4971.
28. Abbas,A., He,X., Niu,J., Zhou,B., Zhu,G., Ma,T., Song,J., Gao,J., Zhang,M.Q. and Zeng,J. (2019) Integrating Hi-C and FISH data for modeling of the 3D organization of chromosomes. *Nat. Commun.*, **10**, 2049.
29. Zhu,G., Deng,W., Hu,H., Ma,R., Zhang,S., Yang,J., Peng,J., Kaplan,T. and Zeng,J. (2018) Reconstructing spatial organizations of chromosomes through manifold learning. *Nucleic Acids Res.*, **46**, E50.
30. Lappala,A., Wang,C., Kriz,A., Michalk,H., Tan,K., Lee,J.T. and Sanbonmatsu,K.Y. (2021) Four-dimensional chromosome reconstruction elucidates the spatiotemporal reorganization of the mammalian X chromosome. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2107092118.

31. Hua,N., Tjong,H., Shin,H., Gong,K., Zhou,X.J. and Alber,F. (2018) Producing genome structure populations with the dynamic and automated PGS software. *Nat. Protoc.*, **13**, 915–926.
32. Schwessinger,R., Gosden,M., Downes,D., Brown,R.C., Oudelaar,A.M., Telenius,J., Teh,Y.W., Lunter,G. and Hughes,J.R. (2020) DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat. Methods.*, **17**, 1118–1124.
33. Fudenberg,G., Kelley,D.R. and Pollard,K.S. (2020) Predicting 3D genome folding from DNA sequence with Akita. *Nat. Methods*, **17**, 1111–1117.
34. Belokopytova,P.S., Nuriddinov,M.A., Mozheiko,E.A., Fishman,D. and Fishman,V. (2020) Quantitative prediction of enhancer–promoter interactions. *Genome Res.*, **30**, 72–84.
35. Bascom,G.D. and Schlick,T. (2017) Mesoscale modeling of chromatin fibers. *Nucl. Archit. Dyn.*, **2**, 123–147.
36. Portillo-Ledesma,S., Li,Z. and Schlick,T. (2022) Genome modeling: from chromatin fibers to genes. *Curr. Opin. Struct. Biol.*, **78**, 102506.
37. Li,Z., Portillo-Ledesma,S. and Schlick,T. (2022) Brownian dynamics simulations of mesoscale chromatin fibers. *Biophys. J.*, **122**, 2884–2897.
38. Zhang,Q., Beard,D.A. and Schlick,T. (2003) Constructing irregular surfaces to enclose macromolecular complexes for mesoscale modeling using the discrete surface charge optimization (DiSCO) algorithm. *J. Comput. Chem.*, **24**, 2063–2074.
39. Arya,G. and Schlick,T. (2009) A tale of tails: how histone tails mediate chromatin compaction in different salt and linker histone environments. *J. Phys. Chem.*, **113**, 4045–4059.
40. Collepardo-Guevara,R., Portella,G., Vendruscolo,M., Frenkel,D., Schlick,T. and Orozco,M. (2015) Chromatin unfolding by epigenetic modifications explained by dramatic impairment of internucleosome interactions: a multiscale computational study. *J. Am. Chem. Soc.*, **137**, 10205–10215.
41. Bascom,G.D. and Schlick,T. (2018) Chromatin fiber folding directed by cooperative histone tail acetylation and linker histone binding. *Biophys. J.*, **114**, 2376–2385.
42. Luque,A., Collepardo-Guevara,R., Grigoryev,S. and Schlick,T. (2014) Dynamic condensation of linker histone C-terminal domain regulates chromatin structure. *Nucleic Acids Res.*, **42**, 7553–7560.
43. Perisic,O., Portillo-Ledesma,S. and Schlick,T. (2019) Sensitive effect of linker histone binding mode and subtype on chromatin condensation. *Nucleic Acids Res.*, **47**, 4948–4957.
44. Collepardo-Guevara,R. and Schlick,T. (2014) Chromatin fiber polymorphism triggered by variations of DNA linker lengths. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 8061–8066.
45. Chen,K., Xi,Y., Pan,X., Li,Z., Kaestner,K., Tyler,J., Dent,S., He,X. and Li,W. (2013) DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res.*, **23**, 341–351.
46. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W., *et al.* (2008) Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
47. Cao,K., Lailler,N., Zhang,Y., Kumar,A., Uppal,K., Liu,Z., Lee,E.K., Wu,H., Medrzycki,M., Pan,C., *et al.* (2013) High-resolution mapping of h1 linker histone variants in embryonic stem cells. *PLoS Genet.*, **9**, e1003417.
48. Woodcock,C.L., Skoultchi,A.I. and Fan,Y. (2006) Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length. *Chromosom. Res.*, **14**, 17–25.
49. Zhou,X., Blocker,A.W., Airoldi,E.M. and O'Shea,E. K. (2016) A computational approach to map nucleosome positions and alternative chromatin states with base pair resolution. *Elife*, **5**, e16970.
50. Bascom,G.D., Sanbonmatsu,K.Y. and Schlick,T. (2016) Mesoscale modeling reveals hierarchical looping of chromatin fibers near gene regulatory elements. *J. Phys. Chem. B*, **120**, 8642–8653.
51. Bascom,G.D., Myers,C.G. and Schlick,T. (2019) Mesoscale modeling reveals formation of an epigenetically driven HOXC gene hub. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 4955–4962.
52. Perišić,O., Collepardo-Guevara,R. and Schlick,T. (2010) Modeling studies of chromatin fiber structure as a function of DNA linker length. *J. Mol. Biol.*, **403**, 777–802.
53. Hsieh,T.S., Cattoglio,C., Slobodyanyuk,E., Hansen,A.S., Rando,O.J., Tjian,R. and Darzacq,X. (2020) Resolving the 3D landscape of transcription-linked mammalian chromatin folding. *Mol. Cell*, **78**, 539–553.
54. Idelfonso-García,O.G., Alarcón-Sánchez,B.R., Vásquez-Garzón,V.R., Baltiérrez-Hoyos,R., Villa-Treviño,S., Muriel,P., Serrano,H., Pérez-Carreón,J.I. and Arellanes-Robledo,J. (2022) Is nucleoredoxin a master regulator of cellular redox homeostasis? its implication in different pathologies. *Antioxidants*, **11**, 670.
55. Kuraku,S. (2011) Hox gene clusters of early vertebrates: do they serve as reliable markers for genome evolution?. *Genom. Proteom. Bioinform.*, **9**, 97–103.
56. Bhatlekar,S., Fields,J.Z. and Boman,B.M. (2014) HOX genes and their role in the development of human cancers. *J. Mol. Med.*, **92**, 811–823.
57. Yao,Q., Wang,C., Wang,Y., Zhang,X., Jiang,H. and Chen,D. (2022) The integrated comprehension of lncRNA HOXA-AS3 implication on human diseases. *Clin. Transl. Oncol.*, **24**, 2342–2350.
58. Ekanayake,D.L., Małopolska,M.M., Schwarz,T., Tuz,R. and Bartlewski,P.M. (2022) The roles and expression of HOXA/Hoxa10 gene: a prospective marker of mammalian female fertility?. *Reprod. Biol.*, **22**, 100647.
59. Barshad,G., Lewis,J.J., Chivu,A.G., Abuhashem,A., Krietenstein,N., Rice,E.J., Rando,O.J., Hadjantonakis,A.-K. and Danko,C.G. (2023) RNA polymerase II and PARP1 shape enhancer–promoter contacts. *Nat. Genet.*, **55**, 1370–1380.
60. Neguembor,M.V., Arcon,J.P., Buitrago,D., Lema,R., Walther,J., Garate,X., Martin,L., Romero,P., AlHaj Abed,J. and Gut,M., (2022) MiOS, an integrated imaging and computational strategy to model gene folding with nucleosome resolution. *Nat. Struct. Mol. Biol.*, **29**, 1011–1023.
61. Spearman,C. (1904) The proof and measurement of association between two things. *Am. J. Psychol.*, **15**, 72.
62. Ou,H.D., Phan,S., Deerinck,T.J., Thor,A., Ellisman,M.H. and O'Shea,C.C. (2017) ChromEMT: visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science*, **357**, eaag0025.
63. Trussart,M., Serra,F., Baù,D., Junier,I., Serrano,L. and Marti-Renom,M.A. (2015) Assessing the limits of restraint-based 3D modeling of genomes and genomic domains. *Nucleic Acids Res.*, **43**, 3465–3477.
64. Grigoryev,S.A., Bascom,G., Buckwalter,J.M., Schubert,M.B., Woodcock,C.L. and Schlick,T. (2016) Hierarchical looping of zigzag nucleosome chains in metaphase chromosomes. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 1238–1243.
65. Bascom,G.D., Kim,T. and Schlick,T. (2017) Kilobase pair chromatin fiber contacts promoted by living-system-like DNA linker length distributions and nucleosome depletion. *J. Phys. Chem. B*, **121**, 3882–3894.
66. Harris,H.L., Gu,H., Olshansky,M., Wang,A., Farabella,I., Eliaz,Y., Kalluchi,A., Krishna,A., Jacobs,M., Cauer,G., *et al.* (2023) Chromatin alternates between A and B compartments at kilobase scale for subgenic organization. *Nat. Commun.*, **14**, 3303.
67. Wang,H., Yang,J., Zhang,Y., Qian,J. and Wang,J. (2022) Reconstruct high-resolution 3D genome structures for diverse cell-types using FLAMINGO. *Nat. Commun.*, **13**, 2645.
68. Tjong,H., Li,W., Kalhor,R., Dai,C., Hao,S., Gong,K., Zhou,Y., Li,H., Zhou,X.J., Le Gros,M.A., *et al.* (2016) Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E1663–E1672.

Received: July 6, 2023. **Revised:** October 12, 2023. **Editorial Decision:** November 6, 2023. **Accepted:** November 22, 2023