JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 26, Number 11, 2019
Mary Ann Liebert, Inc.
Pp. 1191–1202
DOI: 10.1089/cmb.2019.0100

Research Articles

# Bayesian Estimation of Three-Dimensional Chromosomal Structure from Single-Cell Hi-C Data

MICHAEL ROSENTHAL,[1,*] DARSHAN BRYNER,[1,*] FRED HUFFER,[2] SHANE EVANS,[3] ANUJ SRIVASTAVA,[2] and NICOLA NERETTI[3,4]

## ABSTRACT

**The problem of three-dimensional (3D) chromosome structure inference from Hi-C data sets is important and challenging. While bulk Hi-C data sets contain contact information derived from millions of cells and can capture major structural features shared by the majority of cells in the sample, they do not provide information about local variability between cells. Single-cell Hi-C can overcome this problem, but contact matrices are generally very sparse, making structural inference more problematic. We have developed a Bayesian multiscale approach, named Structural Inference via Multiscale Bayesian Approach, to infer 3D structures of chromosomes from single-cell Hi-C while including the bulk Hi-C data and some regularization terms as a prior. We study the landscape of solutions for each single-cell Hi-C data set as a function of prior strength and demonstrate clustering of solutions using data from the same cell.**

**Keywords:** Bayesian estimation, chromosome 3D structure, Hi-C.

## 1. INTRODUCTION

THE USE OF WHOLE-GENOME CONFORMATION CAPTURE TECHNIQUES (3C) such as Hi-C (Lieberman-Aiden et al., 2009) has revealed that the three-dimensional (3D) organization of the genome plays a key role in regulating fundamental cellular processes such as transcriptional regulation, cell cycle progression, and cellular differentiation (Lieberman-Aiden et al., 2009; Naumova et al., 2013; Dixon et al., 2015). These studies generate contact maps describing the probability of observing interactions between any two regions of the genome, which can be associated with distance matrices between pairs of genomic loci. Methods developed to infer the 3D structure of chromosomes from these contact maps typically rely either on optimization-based strategies to minimize the difference between the inferred structure and the distance matrix (Rousseau et al., 2011; Hu et al., 2013; Varoquaux et al., 2014; Zou et al., 2016; Park and Lin, 2016), or on probabilistic modeling to find the most likely structure(s) given the observed contact probabilities (Baù and Marti-Renom, 2012; Zhang et al., 2013; Lesne et al., 2014; Szałaj et al., 2016; Adhikari et al., 2016; Rieber and Mahony, 2017; Trieu and Cheng, 2017).

[1]Science and Technology Department, Naval Surface Warfare Center, Panama City Division, Panama City, Florida.
[2]Department of Statistics, Florida State University, Tallahassee, Florida.
[3]Center for Computational Molecular Biology, Brown University, Providence, Rhode Island.
[4]Department of Molecular Biology, Cell Biology, and Biochemistry, Brown University, Providence, Rhode Island.
*These authors contributed equally to this work.

While Hi-C is typically collected on bulk samples containing millions of cells, it is not clear how much the organizational features present in these population data sets reflect the 3D organization of chromosomes in individual cells. For example, it is not guaranteed that all observed long-range contacts appear simultaneously in each cell (Tjong et al., 2016). Thanks to recent advances in Hi-C technology, we can now study long-range interactions at the single-cell level (Nagano et al., 2013; Flyamer et al., 2017; Ramani et al., 2017; Stevens et al., 2017; Nagano et al., 2017). Single-cell Hi-C has confirmed many organizational principles described in bulk experiments, but their interpretation is not straightforward. For example, it is not yet clear whether topologically associated domains are 3D structural units in individual cells or a population feature that emerges when many cells are aggregated in bulk Hi-C experiments (Nagano et al., 2013; Flyamer et al., 2017), although recent work in *Drosophila* points to the former (Szabo et al., 2018).

The primary difficulty in inferring 3D chromosome structures from single-cell Hi-C data is the sparseness of the contact maps. Currently available methods rely on inference of missing data (Paulsen et al., 2015) or on polymer models with the optimization based on Markov chain Monte Carlo (Carstens et al., 2016) or simulated annealing techniques (Nagano et al., 2013; Stevens et al., 2017; Nagano et al., 2017). However, recovery of potentially missing long-range interactions in the contact matrix relies exclusively on the information contained within individual single-cell matrices.

Here we present a solution using Structural Inference via Multiscale Bayesian Approach (SIMBA3D), which utilizes bulk Hi-C to aid in recovering the contribution of interactions potentially missed in single-cell Hi-C contact maps. Our strategy is similar in principle to the one used in Tjong et al. (2016), where bulk Hi-C is decomposed into an ensemble of single-cell 3D structures. We build a generalized Bayesian framework that utilizes penalties associated with folding constraints and a prior derived from bulk Hi-C samples to infer 3D chromosome structure in single cells. The SIMBA3D software is available at (https://github.com/nerettilab/SIMBA3D).

## 2. METHODS

### 2.1. Proposed framework

The primary goal of the inference is to efficiently explore a vast space of potential chromosomal structures and seek optimal solutions using contact matrices and other contextual data. This requires constructing objective functions with desirable properties and developing scalable algorithms to reach interpretable conformations in times that are practical for large-scale computations. As stated above, the problem of estimating chromosomal structure from single-cell data is challenging because these data are very sparse and noisy. To reach more realistic solutions, we implement a Bayesian approach that supplements the single-cell data with the bulk data. This technique helps fill the missing parts with structures corresponding to the population of cells and additionally imposes certain penalties to improve the quality of estimated structures. The penalties are designed in particular to favor uniform placement of points on the estimated curve and to force the curve itself to be smoother.

Suppose that the genome is partitioned into $n$ equally sized, disjoint segments, or bins. Let $C$ be the $n \times n$ data matrix obtained from an Hi-C experiment. The $ij$'th entry of $C$, call it $c_{ij}$, represents the number of observed interactions between segments $i$ and $j$, and thus, $C$ is naturally a symmetric matrix. Suppose further that another data matrix $C'$ is available to us and represents the collective results of prior Hi-C experiments. For example, in the case of a single-cell Hi-C matrix $C$, we could also have available to us bulk Hi-C data from the same type of cell, or, alternatively, $C'$ could be equal to the sum of several other single-cell Hi-C matrices. Let $x_i \in \mathbb{R}^3$ be the center of mass of the $i$th segment, and let $X \in \mathbb{R}^{n\times 3}$ be the collection of all such $x_i$'s. The problem at hand is twofold: (1) to estimate the structure $X$ from the Hi-C data matrix, and (2) to quantify differences in structures obtained either from the same or different data matrices. In both the estimation and analysis stages, we consider a structure $X$ to be equivalent modulo scale, translation, and rigid rotation/reflection.

We define a posterior energy function $E$ on the space of potential curves, $\mathcal{X} = \mathbb{R}^{n\times 3}$—each curve containing $n$ points (or nodes) in $\mathbb{R}^3$—and use a gradient-based approach to solve for an optimal solution

$$\hat{X} = \arg\inf_{X\in\mathcal{X}} E(X|C, C', a, b, \lambda), \tag{1}$$

where $C=(c_{ij})$ and $C'=(c'_{ij})$ are the single-cell and bulk contact matrices, respectively, $\lambda$ is a vector of weights, and $a$ and $b$ are predetermined model parameters (described later). Let $M=\sum_{j>i} c_{ij}$ and $M'=\sum_{j>i} c'_{ij}$, respectively. The energy function $E$ has several terms, each contributing to a certain aspect of the estimated curve:

$$E(X|C, C', a, b, \lambda)=\frac{1}{M}g(X|C, a, b)+\frac{\lambda_3}{M'}g(X|C', a, b)+\lambda_1 h_1(X)+\lambda_2 h_2(X). \tag{2}$$

We discuss one by one these quantities that comprise $E$.

*2.1.1. Negative log-likelihood term.* The first term $g(X \mid C, a, b)$ in Equation (2) is the negative log-likelihood of the contact matrix $C$ given a curve $X$. This term follows a Poisson model (Varoquaux et al., 2014) with $a$, $b$ being predetermined model parameters, as follows. Varoquaux et al. (2014) link the $ij$'th interaction count with the $ij$'th pairwise distance via the following probability model:

$$C_{ij}\sim \text{Poisson}(b||x_i-x_j||^a) \tag{3}$$

for $j > i$, with $\|\cdot\|$ being the standard Euclidean norm, and for scalars $a<0$ and $b>0$. Since $a<0$, the expected number of interactions between segments $i$ and $j$ is larger when the segments are located closer together in space, and this expected number behaves according to a power law with power $a$. Varoquaux et al. (2014) derive the theoretically optimal value of $a=-3$ from principles of polymer physics. Furthermore, the parameter $b$ acts as a scaling parameter.

The probability mass function for the Poisson random variable given in Equation (3) is given by the following:

$$P(C_{ij}=c_{ij}|X, a, b)=\frac{(b||x_i-x_j||^a)^{c_{ij}}e^{-b||x_i-x_j||^a}}{c_{ij}!}\ .$$

Thus, given an Hi-C matrix with independent entries, the log-likelihood function is written as follows:

$$\ell(X, a, b|C)=\log\left(\prod_{i=1}^{n-1}\prod_{j=i+1}^{n} P(C_{ij}=c_{ij}|X, a, b)\right)$$

$$=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(c_{ij}\log(b)+ac_{ij}\log(||x_i-x_j||)-b||x_i-x_j||^a-\log(c_{ij}!)\right).$$

From the log-likelihood function above, one can see that for a given value of $a$, the parameter $b$ is nonidentifiable because $\ell(X, a, \gamma^a b|C)=\ell(\gamma X, a, b|C)$ for any scalar $\gamma > 0$. That is, changing $b$ is equivalent to changing the scale of $X$, and since $X$ is considered equivalent modulo scale, the choice of $b$ is arbitrary. Define the function $g$ as the negative log-likelihood function, dropping terms that are constant with respect to $X$, that is,

$$g(X|C, a, b)=-\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(ac_{ij}\log(||x_i-x_j||)-b||x_i-x_j||^a)\ . \tag{5}$$

In structure estimation, we consider the parameters $a$ and $b$ to be fixed and known values; thus, we include them with the data $C$ as given when writing the function $g$. The maximum likelihood estimate (MLE) of $X$ is computed as the minimizer of $g$; that is,

$$\hat{X}=\underset{X\in\mathbb{R}^{n\times 3}}{\arg\min}\ g(X|C, a, b). \tag{6}$$

*2.1.2. Penalty terms.* In situations when the data matrix $C$ is sparse or noisy, the standard MLE in Equation (6) can be biologically unrealistic or even fail to converge. Thus, we design several additive penalty terms to regulate the maximum likelihood solution. The remaining terms in $E$ as defined in Equation (2) can be viewed as imposing a prior distribution on the curve. These terms represent the prior belief that the curve displays some regularity. The term involving $h_1$ penalizes variation in the distances

between adjacent points on the estimated curves, while the term with $h_2$ penalizes deviations from straightness. Since the weights $\lambda_1$ and $\lambda_2$, are typically small, these terms essentially discourage excessive variation in the distances between points and excessive bending of the curve. The second term in Equation (2) involves the negative log-likelihood of the bulk contact matrix $C'$ and represents the prior belief that the curve $X$ will bear some resemblance to those found in the population. Together, these three terms drive the solution toward a smoother, more interpretable curve that conforms to both single-cell and bulk data.

These are constructed as follows.

1. **First Penalty**: Define the first penalty as

$$h_1(X) = (n-1)\frac{\sum_{i=1}^{n-1} \|x_{i+1} - x_i\|^2}{\left(\sum_{i=1}^{n-1} \|x_{i+1} - x_i\|\right)^2} - 1. \tag{7}$$

The interpretation of $h_1$ is the following. Define $L(X) = \sum_{i=1}^{n-1} ||x_{i+1} - x_i||$ as the length of $X$, and let $u_i = \frac{n-1}{L(X)} ||x_{i+1} - x_i||$ be the distance between the $i$th pair of adjacent points in $X$, when $X$ has been rescaled to have length $n-1$. Then, one can show that $h_1(X) = \sigma_u^2$, where $\sigma_u^2$ is the variance of the $u_i$'s, and therefore, the effect of the penalty $h_1$ is to reduce the variability of the distances between adjacent points of $X$. The configuration that minimizes $h_1(X)$ is such that all the $u_i$'s are equal to 1, that is, adjacent points in $X$ are all the same distance apart. The minimum value of $h_1$ is 0 regardless of the value of $n$. Furthermore, notice that since $h_1(\gamma X) = h_1(X)$ for any $\gamma > 0$, $h_1(X+y) = h_1(X)$ for any $y \in \mathbb{R}^p$, and $h_1(XR) = h_1(X)$ for any $3 \times 3$ orthogonal matrix $R$, the penalty $h_1$ is invariant to scale, translation, and rotation/reflection.

2. **Second Penalty**: Define the second penalty as

$$h_2(X) = \frac{1}{n-2} \sum_{i=2}^{n-1} \frac{(x_{i-1} - x_i) \cdot (x_{i+1} - x_i)}{||x_{i-1} - x_i|| ||x_{i+1} - x_i||}. \tag{8}$$

The interpretation of $h_2$ is the following. If $\theta_i$ is the angle created by the triplet of points $(x_{i-1}, x_i, x_{i+1})$, and $y_i = \cos(\theta_i)$, then $h_2(X) = \bar{y}$, the sample mean of all the $y_i$'s. Therefore, the minimizer of $h_2$ is such that $\cos(\theta_i) = -1$ for all $i = 2, \ldots, n-1$. This occurs when $X$ is a straight line with all $\theta_i = \pi$; hence, the effect of the penalty $h_2$ is to enforce a level of smoothness to $X$. The penalty $h_2$ has a minimum value of $-1$ regardless of the value of $n$ and is invariant to scale, translation, and rotation/reflection.

3. **Bulk Prior**: The second term in Equation (2; the bulk prior) may be written as $\lambda_3 h_3(X)$ where we define

$$h_3(X) = \frac{1}{M'} g(X|C', a, b). \tag{9}$$

Notice that if we let $\tilde{C} = C + \frac{\lambda_3 M}{M'} C'$ and $\tilde{b} = (1 + \frac{\lambda_3 M}{M'})b$, then the sum $g(X \mid C, a, b) / M + \lambda_3 h_3(X)$ is equal to $g(X \mid \tilde{C}, a, \tilde{b}) / M$. Therefore, the effect of the penalty $h_3$ is essentially to add a scalar multiple of $C'$ to the data $C$ and perturb the parameter $b$. If $\lambda_3$ is chosen to be small enough, then this penalty term will only slightly alter $C$ and not overwhelm the original data with the bulk data. If $C$ is sparse and $C'$ is not, then the addition of this penalty term with a small enough $\lambda_3$ eliminates the sparseness of $C$ by replacing many of the 0 entries with small numbers that are biologically more meaningful than random noise.

There are many other possibilities for penalty terms. However, for practical implementation of the optimization problem in Equation (1), we use only the penalties $h_1$, $h_2$, and $h_3$ for three reasons. First, each penalty term has a straightforward and biologically meaningful interpretation. Second, the formula for each term is relatively simple and inexpensive to compute—in particular, since $h_3$ becomes absorbed into the function $g$, the addition of this penalty requires an essentially zero increase in overall computation time. Third, along with the function $g$, we can write an analytical expression for the gradient of each penalty term, and therefore, we can write a gradient expression for $E$. By inputting the expression of $\nabla E$ to a numerical solver, we can maintain computational tractability on a personal computer for the large values of $n$ typically seen in real data sets. Using the penalty functions, $E$ can be written as

$$E(X|C, C', a, b, \lambda) = \frac{1}{M} g(X|\tilde{C}, a, \tilde{b}) + \lambda_1 h_1(X) + \lambda_2 h_2(X)\ , \tag{10}$$

where $\tilde{C}$ and $\tilde{b}$ are defined in the text following Equation (9), and we consider this objective function for the remainder of this work. The choice of the penalty weights $\lambda$ is left to the user and can influence the solution greatly.

### 2.2. *Multiscale gradient optimization for improved inference*

The biggest challenge in solving the optimization problem given in Equation (1) comes from a non-convex energy function and an extremely high-dimensional search space, which brings about multiple local optima and a tremendous computational cost. While the presence of the bulk and penalty terms helps to mitigate these issues by steering the search toward more realistic solutions, the computational complexity still remains a major hurdle.

*2.2.1. Gradient-based optimization.* We take a multiscale, gradient-based approach, where the optimization at each iteration is performed using a gradient-based technique called Broyden–Fletcher–Goldfarb–Shanno (BFGS; Gill et al., 1981; Nocedal and Wright, 2006). Thus, it is helpful to derive an analytical expression for $\nabla E$ to a numerical optimizer. The gradient of $E$ at $X$ is written as

$$\nabla E(X|C, C', a, b, \lambda) = \frac{1}{M}\nabla g(X|\tilde{C}, a, \tilde{b}) + \lambda_1 \nabla h_1(X) + \lambda_2 \nabla h_2(X) \; ; \tag{11}$$

therefore, to build the expression for $\nabla E$, we need to compute the expressions for $\nabla g$, $\nabla h_1$, and $\nabla h_2$, where $g$, $h_1$, and $h_2$ are defined in Equations (5), (7), and (8), respectively. The expressions for these gradients are presented in Supplementary Data.

*2.2.2. Multiscale optimization.* To reduce the computation time to allow for a practical full-genome reconstruction, we implement a multiscale optimization technique. Compared with a standard approach that computes the full resolution optimization using a random initialization, the multiscale approach reduces computation time and limits the local solutions obtained. As shown in Figure 1B, this leads to solutions with smaller energies.

The multiscale optimization technique used in SIMBA3D is as follows. First, from a given full resolution contact matrix, we generate a series of new matrices decreasing in resolution, that is, decreasing in size, by recursively combining adjacent pairwise interaction counts to reflect a merging of adjacent genomic bins. One iteration of this process cuts the dimension of the contact matrix roughly in half. For each matrix generated in the series, we ignore the diagonal elements as we would in the original full contact matrix. Once we generate the multiscale series of matrices, we execute the series of optimizations in the reverse order, beginning with the smallest matrix and ending with the full matrix. We initialize the smallest optimization randomly from a standard multivariate normal distribution, obtain a solution, and then upsample this solution (i.e., interpolate between the solution nodes) to initialize the next larger optimization problem in the series. We continue this iterative process of solving successively larger optimizations, using an upsampled version of the current solution as an initialization to the next higher resolution problem, until we finish with the full solution.

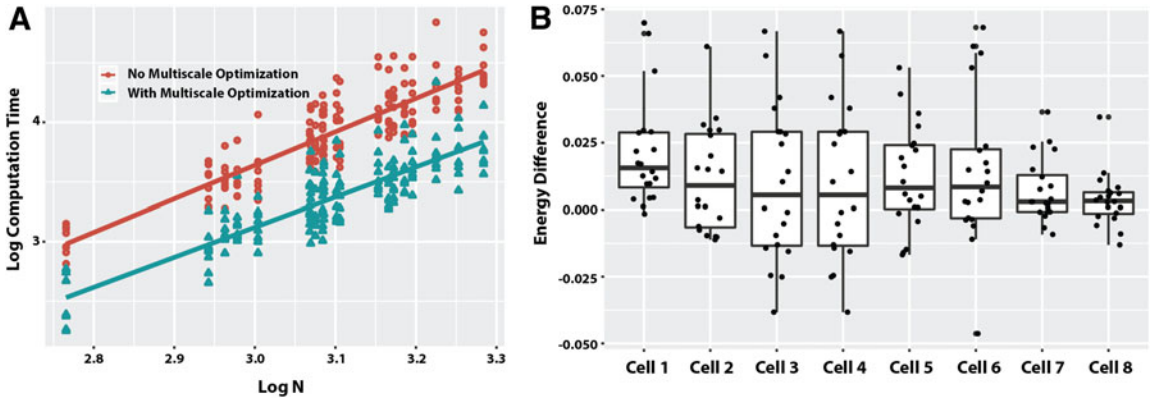


**FIG. 1.** Improvement in computation time and final energy with multiscale optimization. Comparison of results, with and without the multiscale approach executed on all 20 chromosomes in each of the 8 cells, available in the mESC data set. **(A)** The log scale computation time is plotted over log scale number of nodes with regression lines fitted. **(B)** Boxplot of the difference in final energies obtained for each of the eight cells. The difference is computed by subtracting the energy obtained via the multiscale approach from the energy obtained without the multiscale approach. That is, a positive number here indicates that the multiscale approach yielded a lower energy solution.

SIMBA3D implements this multiscale technique in Python, through which at each scale the optimization is solved using the BFGS method (Gill et al., 1981; Nocedal and Wright, 2006) with analytical gradient.

Although we solve several optimization problems in the above multiscale approach compared with just one in a standard approach, the computation time is significantly reduced. Since the smaller optimizations are relatively fast compared with the full resolution version, the multiscale approach is essentially a systematic way of cheaply providing a good initialization to the full problem. The combined cost of producing this initialization and executing the full resolution optimization is less than the cost of executing the full resolution optimization with a full resolution random initialization. Moreover, our experimental results show that we achieve on average a lower energy, that is, better quality solution using the multiscale approach compared with that of the standard approach of random full resolution initialization. An additional consequence of using the multiscale approach is that by design, the space of obtainable full resolution local solutions is limited by the initial smallest resolution. In extreme cases, the very small resolution problems may only have one solution; therefore, if one wishes to explore the local solution space of the full resolution by using different random initializations while still enjoying the benefits of reduced computation time, one must strike a reasonable compromise in initial scale size.

Shown in Figure 2 is an example of an estimated structure obtained using SIMBA3D on chromosome 19 in the mouse embryonic stem cell (mESC) data set (Stevens et al., 2017). The left panel shows the single-cell contact matrix $C$ (from cell 1) and the middle panel shows the ensemble matrix $C'$. The result of the estimation $\hat{X}$ is shown in the rightmost panel. The algorithm was applied for the parameter values $\lambda_1 = 0.5$, $\lambda_2 = 1.0$, and $\lambda_3 = 0.1$.

## 3. RESULTS

### 3.1. Simulated data

A complete estimation solution for a simulated configuration, intended as an illustration of the estimation process, is presented in Supplementary Figure S1. To verify the capabilities of SIMBA3D on more realistic data sets, we designed and executed a series of experiments on single-cell Hi-C data matrices that have been simulated from a known set of ground truth structures. The ground truth structure set $\{X_k \in \mathbb{R}^{100\times 3}, k = 1, \ldots, K\}$ was designed to exhibit cell-to-cell chromatin shape variability from a mean structure that was obtained from downsampling and smoothing an SIMBA3D solution from real Hi-C data, more extensively described in Section 3.2. For each $X_k$, one can easily simulate a corresponding $100 \times 100$ single-cell Hi-C matrix $C_k$ using the Poisson model described in Equation (3). Any solution from SIMBA3D using $C_k$ can then be optimally scaled and aligned to the ground truth structure and then compared via the root mean square distance (RMSD) metric. With carefully designed experiments, one can make inferences about the solution quality that SIMBA3D produces under various circumstances using this RMSD to ground truth metric.

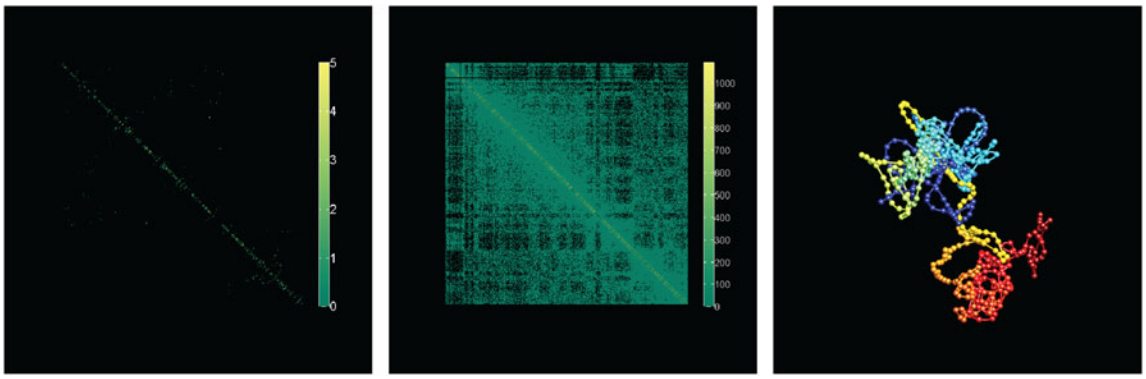


**FIG. 2.** An illustration of chromosome structure estimation using SIMBA3D: the single-cell sparse contact matrix (left), the ensemble contact matrix (middle), and the final estimated structure (right).

The first experiment is designed to verify that SIMBA3D can recover the ground truth structure exactly, up to a small error tolerance, when there are a sufficient amount of contact data, that is, when the data matrix is dense. From one selected ground truth structure, we simulate a dense Hi-C matrix from the Poisson model using $a=-3$ and a large value of $b=10^6$. Since in this situation there is ample contact data available to accurately reconstruct the curve, we use small penalty weights $\lambda_1=\lambda_2=0.001$, and we set $\lambda_3=0$ to forgo the unnecessary use of the population prior. Figure 3 analyzes the solution quality of 200 structure estimations obtained from SIMBA3D using 200 random initializations and without using the multiscale approach. Figure 1 shows that the 50 solutions with the lowest RMSD are essentially the same and recover the ground truth structure correctly up to a small error tolerance. However, due to the high dimensionality of the optimization problem, there are many other local solutions, for example, Solution 131, that are nearly globally optimal but exhibit a flipped or reflected portion of the structure. For this reason, RMSD and energy are only loosely correlated. The darker squares along the diagonal of the pairwise distance matrix in the upper right panel show evidence for the clustering of local solutions with respect to the RMSD metric.

The second experiment is designed to verify that incorporating the population prior in SIMBA3D when the single-cell Hi-C matrix is sparse improves the quality of the estimated solution. The experimental setup is the following. First, for each $X_k$, $k=1, \ldots, K$, we simulated a sparse Hi-C matrix $C_k$ from the Poisson model using a value of $b=10$. We then selected the first 8 matrices to perform the structure estimation with

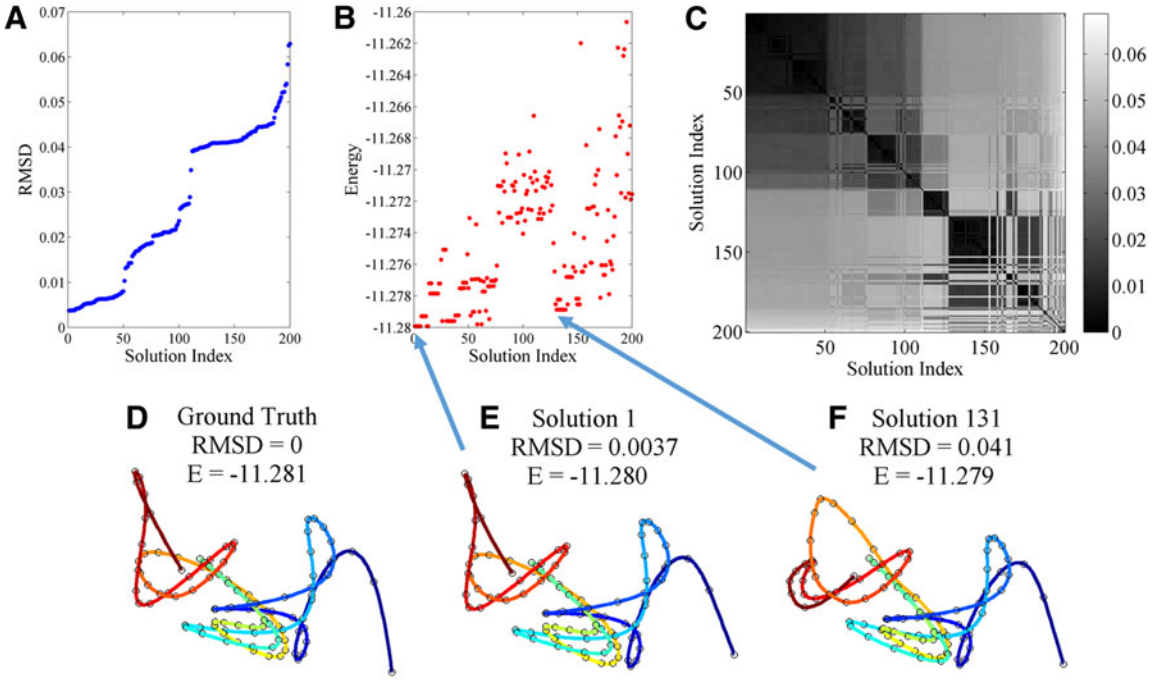


**FIG. 3.** From one selected ground truth structure containing 100 nodes, we simulate a dense Hi-C matrix (not shown) from the Poisson model using $a=-3$ and $b=10^6$. We obtain 200 solutions from SIMBA3D using 200 random initializations and with penalty weights $\lambda_1=\lambda_2=0.001$, and $\lambda_3=0$. **(A)** A plot of RMSD to ground truth versus the solution index, where the solution index is ordered in ascending RMSD. Using the same solution set and index ordering, **(B)** a plot of the energy versus solution index. **(C)** The $200\times200$ pairwise RMSD matrix for each solution pair, also using the same index ordering as in **(A)** and **(B)**. **(D, E, F)** A 100-node structure with nodes connected via a colorized spline interpolated curve. The beginning node with index 1 is located at the blue end of the curve, and the ending node with index 100 is located at the red end of the curve. Also shown above each curve is its label, its RMSD to ground truth, and its energy value **(E)**. **(D)** The ground truth curve. **(E)** Solution 1, the solution with the lowest RMSD of the 200; and **(F)** Solution 131, the solution with the 131st lowest RMSD. Solution 1 is a near-perfect reconstruction with respect to ground truth, and Solution 131 is a similarly optimal solution with respect to energy, but exhibits a local reflection at the red portion of the structure, which leads to a much higher RMSD value. RMSD, root mean square distance; SIMBA3D, Structural Inference via Multiscale Bayesian Approach.

SIMBA3D, using several values of $\lambda_3$, including $\lambda_3 = 0$, and fixed values $\lambda_1 = \lambda_2 = 0.1$. When estimating the structure from matrix $C_k$, the population prior makes use of the bulk matrix given by the sum of all matrices in the data set of size $K$ excluding $C_k$. For each value of $\lambda_3$ and for each of the 8 chosen matrices, we obtain 20 solutions from SIMBA3D using 20 random initializations, again forgoing the use of the multiscale optimization feature in the software. Figure 4 plots the RMSD to ground truth value averaged over all $8 \times 20 = 160$ solutions for each value of $\lambda_3$ tested. We repeat this experiment three times using $K = 10$, $K = 100$, and then the full $K = 1000$ structures in the data set to show how the amount of available bulk data can affect the results. In all three cases, the average RMSD drops immediately as $\lambda_3$ increases from 0 and reaches a minimum value for some $\lambda_3 > 0$. This result shows that including the proposed population prior improves the accuracy of chromatin reconstruction in simulated sparse single-cell Hi-C data, and the improvement is greater when more bulk data are available.

### 3.2. Real data

We applied SIMBA3D to the reconstruction of chromosome structures from single-cell Hi-C data from mESC (Stevens et al., 2017). To highlight the influence of parameter selection on the results, Figure 5

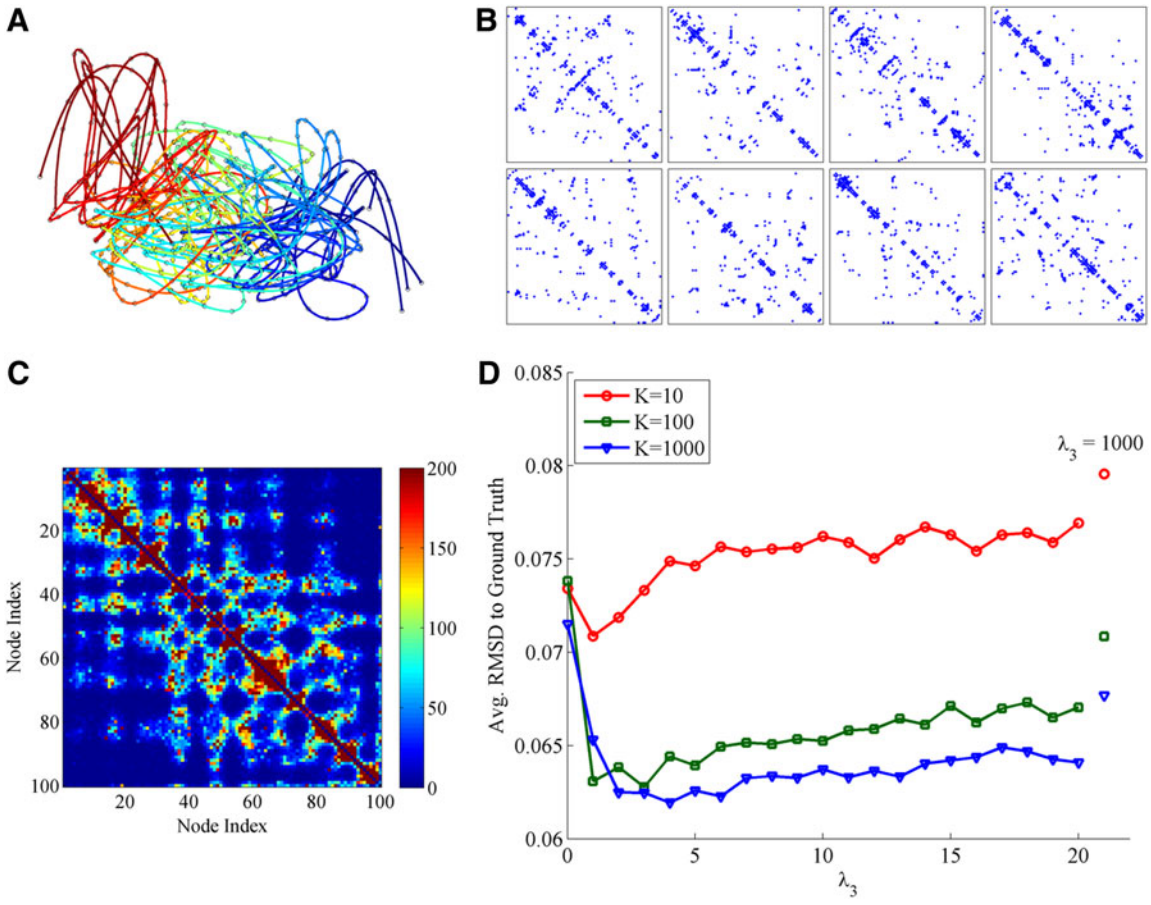


**FIG. 4.** **(A)** The first eight ground truth structures in the simulated data set, all with the same scale, centroid, and mutual alignment over rigid body transformations. As with the structures shown in Figure 1, each structure has 100 nodes connected via a colorized spline interpolated curve. The beginning node with index 1 is located at the blue end of the curve, and the ending node with index 100 is located at the red end of the curve. **(B)** The nonzero elements of the simulated sparse Hi-C matrices $C_k$ for each respective ground truth curve $X_k$ in **(A)**. **(C)** The bulk data matrix that results from summing all $C_k$'s together, excluding $C_1$, for a data set of size $K = 1000$. **(D)** Plots RMSD versus $\lambda_3$ averaged over 20 SIMBA3D solutions obtained for each of the first 8 simulated matrices (i.e., each data point represents the average RMSD to ground truth for $20 \times 8 = 160$ structures). The plot shows three curves corresponding to the scenario of using a data set of size $K = 10$, 100, and 1000 single-cell matrices.

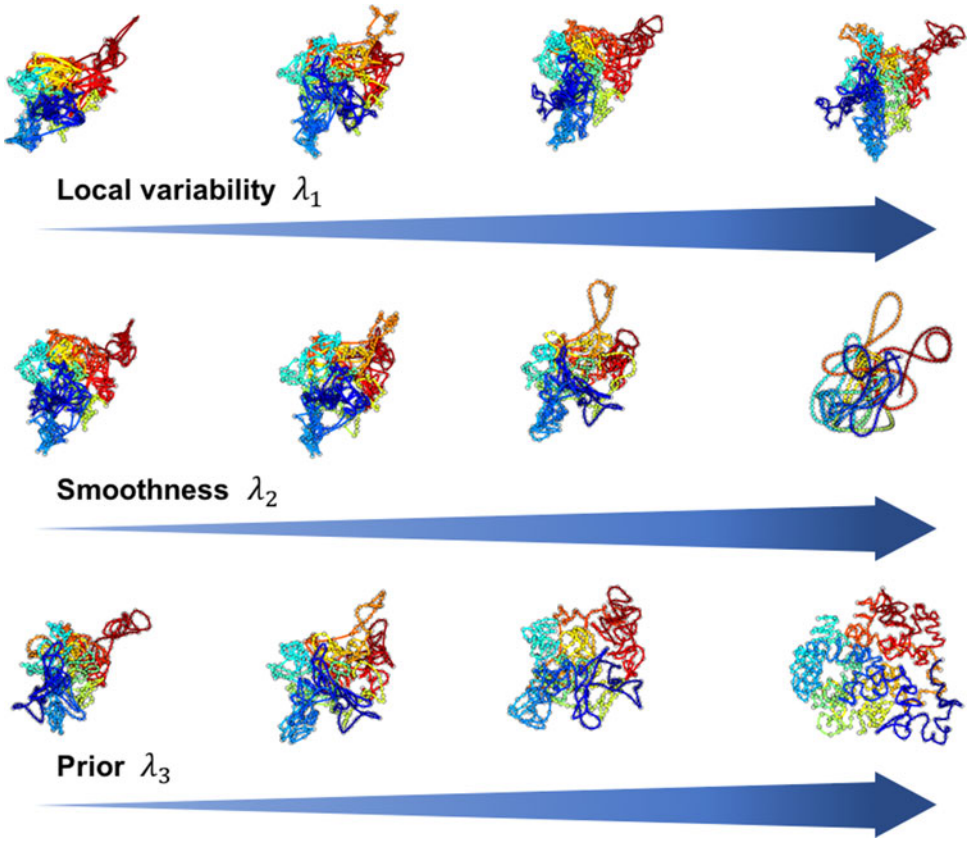


**FIG. 5.** Effect of model parameters on 3D reconstruction quality. We compute the 3D reconstruction as a function of parameter values using the Hi-C data matrix associated with chromosome 19 in cell 1 of the mESC data set. The top row of structures from left to right shows the effect of an increased weight $\lambda_1$ on the parameterization penalty $h_1$. We vary $\lambda_1 = 0.01, 0.1, 1, 10$ and fix $\lambda_2 = \lambda_3 = 0$ to obtain four solution curves with exponentially increasing penalty weight. The center row of structures from left to right shows the effect of an increasing weight $\lambda_2$ on the smoothing penalty $h_2$. Here we vary $\lambda_2 = 0.01, 0.1, 1, 10$ and fix $\lambda_1 = 0.5$ and $\lambda_3 = 0$ to obtain these four solution curves. Finally, to show the effect of incorporating the bulk data—the mESC chromosome 19 population matrix—in the analysis, we vary $\lambda_3 = 0.01, 0.1, 1, 10$ and fix $\lambda_1 = 0.5$, $\lambda_2 = 1$ to obtain the four structures on the bottom row. We computed all structures using the multiscale approach with $n = 73, 146, 292, 584$. 3D, three-dimensional; mESC, mouse embryonic stem cell.

illustrates the effect of the relative values of the three weights—$\lambda_1$, $\lambda_2$, and $\lambda_3$—on the resulting estimated structures. As expected, higher values of these weights lead to increases in the respective properties they emphasize. For instance, an increase in $\lambda_3$ leads to the chromosome structure bearing more resemblance to the structure estimated from the bulk data alone.

Figure 6 studies the nature of solutions resulting from different initializations on the same data. Due to the vast search space in which the structure estimation is performed as well as the nonconvexity of the objective function, the optimization procedure in SIMBA3D cannot ensure convergence to a *unique* global solution. Instead, the output structure represents one of many different local optima that can be reached depending on the initialization. Despite the existence of several local optima, multiple configurations resulting from the same cells do in fact cluster together in the shape space, as illustrated using a pairwise RMSD matrix and dendrogram in this figure. The clustering observed here lends further validity to the inferred structures.

The use of the multiscale optimization technique is beneficial for several reasons. It first estimates broader, coarser structures and then adds smaller details, thereby avoiding the abundance of local traps present at the highest resolution. In addition to reaching a lower energy solution on average, it also speeds the algorithm significantly due to low-dimensional searches in the early stages. Figure 6 quantifies gains in computational cost and final energies due to this multiscale approach. An illustration of this method is shown in Supplementary Movie S1.

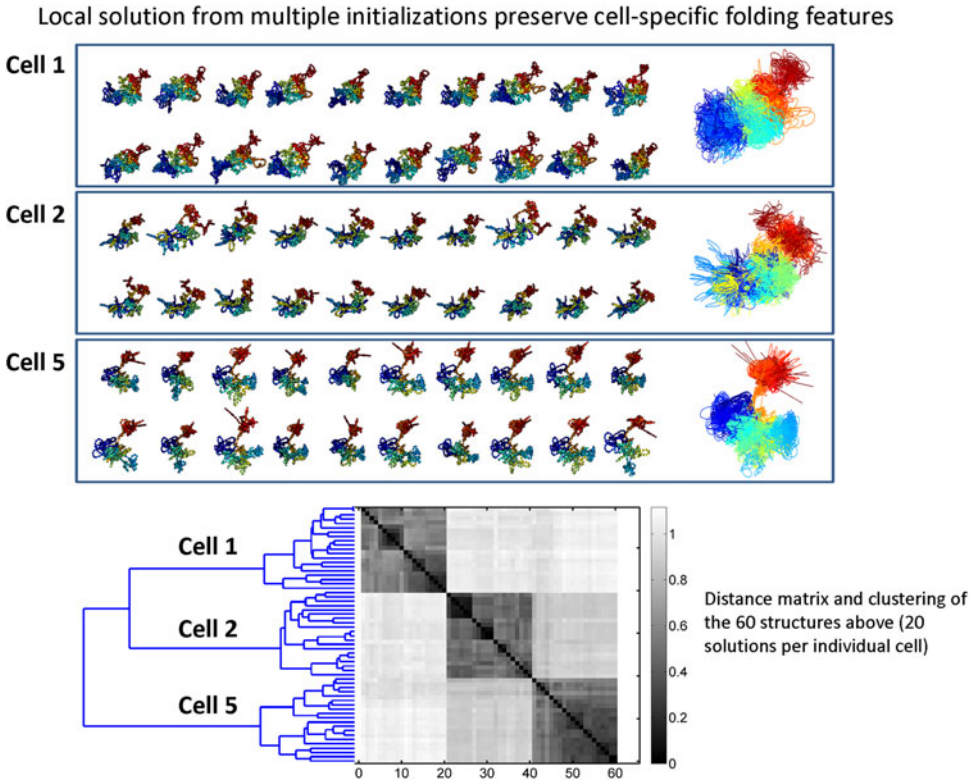


**FIG. 6.** Similarity between ensembles of solutions across cells. Here we show twenty local solutions obtained for chromosome 19 in each of three cells—cell 1, cell 2, and cell 5—in the mESC data set using fixed $\lambda = (0.5, 1, 0.1)$. We computed all structures using the multiscale approach with $n = 73, 146, 292, 584$, and for each cell, we used the same twenty random initializations at the smallest scale. All displayed solutions are rotationally aligned. For each cell, we show the twenty obtained solutions separately, and in addition, to help visualize the variability inherent to the local solutions within cells, we plot the three groups of solutions on top of each other in three respective windows. We then show the clustering of all 60 solutions in shape space via a $60 \times 60$ pairwise RMSD matrix and associated dendrogram plot.

## 4. CONCLUSIONS

In conclusion, SIMBA3D is a Bayesian framework for estimating 3D chromosome structures from single-cell Hi-C data, using penalties for regularization of the estimated structures and using additional information from the bulk Hi-C data. Using multiscale optimization tools and a BFGS routine, it generates computationally efficient inferences and compares these across different initializations and different data (cells). Clustering of solutions in the shape space from the same cell data supports the validity of these solutions.

## ACKNOWLEDGMENTS

This work was supported by the NIH Common Fund Program, grant U01CA200147, as a Transformative Collaborative Project Award (TCPA) to TCPA-2017-NERETTI to N.N. and A.S. Also, we acknowledge the support of Dr. Frank Crosby at the Naval Surface Warfare Center Panama City Division for funding M.R. and D.B. through the In-House Laboratory Independent Research (ILIR) program.

## AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

## SUPPLEMENTARY MATERIAL

Supplementary Data
Supplementary Figure S1
Supplementary Table S1
Supplementary Table S2
Supplementary Movie S1

## REFERENCES


Adhikari, B., Trieu, T., and Cheng, J. 2016. Chromosome3D: Reconstructing three-dimensional chromosomal structures from Hi-C interaction frequency data using distance geometry simulated annealing. *BMC Genomics*. 17, 886.

Baù, D., and Marti-Renom, M.A. 2012. Genome structure determination via 3C-based data integration by the Integrative Modeling Platform. *Methods*. 58, 300–306.

Carstens, S., Nilges, M., and Habeck, M. 2016. Inferential structure determination of chromosomes from single-cell Hi-C data. *PLoS Comput. Biol*. 12, e1005292.

Dixon, J.R., Jung, I., Selvaraj, S., et al. 2015. Chromatin architecture reorganization during stem cell differentiation. *Nature*. 518, 331–336.

Flyamer, I.M., Gassler, J., Imakaev, M., et al. 2017. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature.* 544, 110–114.

Gill, P., Murray, W., and Wright, M. 1981. *Practical Optimization*. Academic Press, London.

Hu, M., Deng, K., Qin, Z., et al. 2013. Bayesian inference of spatial organizations of chromosomes. *PLoS Comput. Biol*. 9, e1002893.

Lesne, A., Riposo, J., Roger, P., et al. 2014. 3D genome reconstruction from chromosomal contacts. *Nat. Methods.* 11, 1141–1143.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*. 326, 289–293.

Nagano, T., Lubling, Y., Stevens, T.J., et al. 2013. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*. 502, 59–64.

Nagano, T., Lubling, Y., Várnai, C., et al. 2017. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*. 547, 61–67.

Naumova, N., Imakaev, M., and Fudenberg, G., 2013. Organization of the mitotic chromosome. *Science (New York, N.Y.)*. 342, 948–953.

Nocedal, J., and Wright, S.J. 2006. *Numerical Optimization*, 2nd edn. Springer Series in Operations Research and Financial Engineering. Springer, New York.

Park, J., and Lin, S. 2016. Impact of data resolution on three-dimensional structure inference methods. *BMC Bioinformatics*. 17, 1–13.

Paulsen, J., Gramstad, O., and Collas, P. 2015. Manifold based optimization for single-cell 3D genome reconstruction. *PLoS Comput. Biol*. 11, 1–19.

Ramani, V., Deng, X., Qiu, R., et al. 2017. Massively multiplex single-cell Hi-C. *Nat. Methods*. 14, 263–266.

Rieber, L., and Mahony, S. 2017. MiniMDS: 3D structural inference from high-resolution Hi-C data. *Bioinformatics*. 33, i261–i266.

Rousseau, M., Fraser, J., Ferraiuolo, M.A., et al. 2011. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinformatics*. 12, 414.

Stevens, T.J., Lando, D., Basu, S., et al. 2017. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*. 544, 59–64.

Szabo, Q., Jost, D., Chang, J.M., et al. 2018. TADs are 3D structural units of higher-order chromosome organization in Drosophila. *Sci. Adv*. 4, eaar8082.

Szałaj, P., Tang, Z., Michalski, P., et al. 2016. An integrated 3-dimensional genome modeling engine for data-driven simulation of spatial genome organization. *Genome Res.* 26, 1697–1709.

Tjong, H., Li, W., Kalhor, R., 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.

Trieu, T., and Cheng, J. 2017. 3D genome structure modeling by Lorentzian objective function. *Nucleic Acids Res*. 45, 1049–1058.

Varoquaux, N., Ay, F., Noble, W.S., et al. 2014. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics*. 30, i26–i33.

Zhang, Z., Li, G., Toh, K.C., et al. 2013. 3D chromosome modeling with semi-definite programming and Hi-C data. *J. Comput. Biol*. 20, 831–846.

Zou, C., Zhang, Y., and Ouyang, Z. 2016. HSA: Integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure. *Genome Biol*. 17, 1–14.

Address correspondence to:
*Dr. Nicola Neretti*
*Department of Molecular Biology, Cell Biology, and Biochemistry*
*Brown University*
*70 Ship Street*
*Providence, RI 02903*

*E-mail:* nicola_neretti@brown.edu
*Dr. Anuj Srivastava*
*Department of Statistics*
*Florida State University*
*Tallahassee, FL 32304*

*E-mail:* anuj@stat.fsu.edu