Statistical Inference on Three-Dimensional Structure of Genome by Truncated Poisson Architecture Model

Jincheol Park and Shili Lin

Abstract In recent years, next generation sequencing technology, coupled with an assay that is capable of detecting genome-wide chromatin interactions, has produced a massive amount of data and led to a greater understanding of long-range, or spatial, gene regulation mechanisms. Hence, the traditional one-dimensional linear view of a genome, which is especially prevalent in statistical and mathematical modeling, is inadequate in many genomic studies. Instead, it is essential, in studying genomic functions, to estimate the three-dimensional (3D) structure of a genome. The availability of genome-wide interaction data necessitates the development of analytical methods to recover the underlying 3D spatial chromatin structure, but challenges abound. One particular issue is the excess of zeros, especially with higher resolution, or inter-chromosomal, data. This leads to questions concerning the appropriateness of using the Poisson distribution to model such data. In this article, we introduce a truncated Poisson Architecture Model (tPAM) to directly model sequencing counts with many zeros. We carried out an extensive simulation study to evaluate tPAM and to compare its performance with an existing method that uses the Poisson distribution to model the counts. We applied tPAM to reconstruct the underlying 3D structures of two data sets, one of human and one of mouse, to demonstrate its utility. The analysis of the human data set considered chromosomes 14 and 22 jointly, thereby illustrating tPAM's capability of analyzing inter-chromosomal data. On the other hand, the mouse analysis was focused on a region on chromosome 2 to evaluate tPAM's performance for recovering structure with loci in different topologically associated domains.

Keywords Spatial interactions · Hi-C · Excess of zeros · Chromatin looping · Data resolution

J. Park
Department of Statistics, Keimyung University, 1095 Dalgubeol-daero,
Daegu 704-701, South Korea
e-mail: park.jincheol@gw.kmu.ac.kr

S. Lin (✉)
Department of Statistics, Ohio State University, 1958 Neil Avenue,
Columbus 43210, USA
e-mail: shili@stat.osu.edu

© Springer International Publishing Switzerland 2015

P.K. Choudhary et al. (eds.), *Ordered Data Analysis, Modeling and Health Research Methods*, Springer Proceedings in Mathematics & Statistics 149,
DOI 10.1007/978-3-319-25433-3_15

1 Introduction

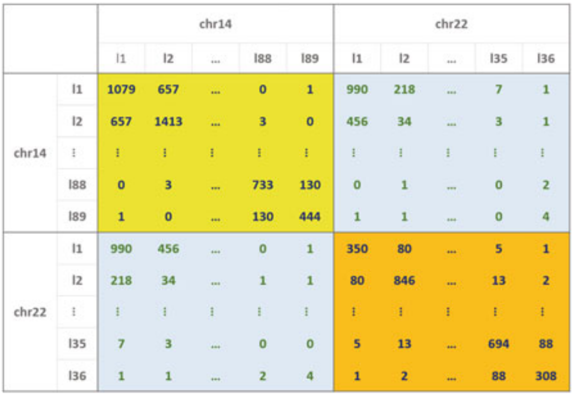
The spatial (three-dimensional, or 3D) organization of a genome is closely linked to its biological function, and thus, full understanding of the genomic structure is essential. In recent years, the ability to identify long-range chromatin interactions genome-wide, known as looping, aided by next generation sequencing technology, has been truly revolutionary in genomic and epigenetic research. The most well-known assay for detecting chromatin interaction, Hi-C [14], produces a library of products that are pairs of fragments in close proximity to each other in the cell nucleus but may be far apart in terms of their chromosomal locations (and may even be on different chromosomes). The library is then analyzed through massively parallel DNA sequencing, producing a catalog of interacting fragments that can be organized into a two-dimensional matrix (known as a contact matrix) of contact counts. Figure 1 provides an example of a contact matrix for chromosomes 14 and 22 based on data from [14], showing only some of the contact counts for illustration purposes. In addition to Hi-C, other assays for detecting genome-wide long-range interactions have also been developed, such as ChIA-PET [6] and TCC [12].

Despite spectacular advances in molecular technologies that allow for unprecedented identifications of genome-wide chromatin interactions, our understanding of 3D organization of genomes is still coarse and incomplete, especially for complex organisms such as humans and mice. This is partly due to the massive amount of data that prove to be extremely difficult to analyze. In addition to its size, the features of the data also pose challenges, rendering conventional statistical methods ineffective. To tackle these issues, analytical approaches have been proposed to understand the spatial organization of the genome based on Hi-C long-range looping data. The approaches can be classified into optimization-based and modeling-based.

For optimization-based approaches, the idea is to first translate each pairwise contact count into a distance using a biophysical property. One then obtains a consensus 3D structure by minimizing some objective function, such as the total "differences" between the translated distances and those inferred from the hypothesized 3D architecture [1, 4, 5, 13, 17, 21]. Many of the optimization methods are based on metric or non-metric multi-dimensional scaling [2, 4, 17]. For this type of approach, normalization of the data is key [11].

Modeling-based approaches, on the other hand, are all based on probability models that describe the relationship between the contact counts with the 3D physical distance. The contact counts are modeled either by a normal distribution to account for variability in the estimation [16] or by a Poisson distribution [10, 18] with its intensity parameter assumed to be related to the physical distance by an inverse relationship. Statistical inferences on the 3D structure (together with other model parameters) are made either by maximum likelihood [18] or through casting the problem into a Bayesian framework [10, 16].

As discussed earlier, a Hi-C experiment produces contact counts that are organized as a 2D matrix for a given resolution. For example, the data matrix shown in Fig. 1 is based on a 1 Mb (megabases) resolution. If there is sufficient sequencing depth,



		chr14					chr22				
		l1	l2	...	l88	l89	l1	l2	...	l35	l36
chr14	l1	1079	657	...	0	1	990	218	...	7	1
	l2	657	1413	...	3	0	456	34	...	3	1
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	l88	0	3	...	733	130	0	1	...	0	2
	l89	1	0	...	130	444	1	1	...	0	4
chr22	l1	990	456	...	0	1	350	80	...	5	1
	l2	218	34	...	1	1	80	846	...	13	2
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	l35	7	3	...	0	0	5	13	...	694	88
	l36	1	1	...	2	4	1	2	...	88	308

Fig. 1 Contact matrix of Hi-C data. The two *diagonal blocks* correspond to intra-chromosomal contacts among loci in chromosome 14 and 22, respectively, while the two *off-diagonal blocks* depict inter-chromosomal contacts between loci in chromosomes 14 and 22. Note that the matrix is symmetric

a higher resolution matrix can lead to a finer and more useful 3D structure, but there tends to be more zero entries in the contact matrix, rendering the Poisson distribution inadequate for modeling the data. To remedy the problem, in this paper, we propose a truncated Poisson Architecture Model (tPAM) by using a truncated Poisson distribution without the zero counts. We carried out an extensive simulation study to evaluate tPAM and to compare its performance with an existing method [10] that uses the Poisson distribution to model the counts. We applied tPAM to reconstruct the underlying 3D structures of two data sets, one of human and one of mouse, to demonstrate its utility. The analysis of the human data set considered chromosomes 14 and 22 jointly, thereby illustrating its capability of analyzing inter-chromosomal data. On the other hand, the mouse analysis was focused on a region on chromosome 2 to evaluate tPAM's performance for recovering a structure with loci in different topologically associated domains (TADs).

2 Methods

2.1 *The tPAM Model*

Consider a set of n fragments (also referred to as loci), each being represented by a point in the 3D space. Collectively, they are denoted by $\Omega \equiv \{\mathbf{p}_i = (p_i^x, p_i^y, p_i^z);\ i = 1, \ldots, n\}$. Let d_{ij} denote the Euclidean distance between loci i and j, that is,

$$d_{ij} = \sqrt{(p_i^x - p_j^x)^2 + (p_i^y - p_j^y)^2 + (p_i^z - p_j^z)^2}. \tag{1}$$

The contact counts of these n loci are organized into a 2D matrix, with y_{ij} denoting the contact count (the (i, j) entry of the matrix), which represents the interaction intensity between loci i and j. Based on these data ($\mathbf{y} = \{y_{ij}, 1 \leq i < j \leq n\}$; note that the matrix is symmetric), the goal is to make inference about the coordinates, Ω, of the 3D structure.

We assume that the contact counts follow a truncated Poisson distribution, with its intensity parameter linked to the 3D distance and other covariates through a log-linear model. More specifically, the Poisson model was built under the assumption that two loci in close proximity in 3D space are likely to interact more, which leads to the following model for the Poisson intensity parameter λ_{ij}:

$$\log \lambda_{ij} = \alpha_0 + \alpha_1 \log d_{ij} + \mathbf{x}_{ij}^T \beta, \tag{2}$$

where $\mathbf{x}_{ij}^T = (x_{ij}^1, \ldots, x_{ij}^K)$ and $\beta = (\beta_1, \ldots, \beta_K)^T$ denote the vector of K covariates and its associated vector of coefficients, respectively. Typical covariates include GC content, fragment length, mappability score, and potentially also restriction enzyme to take care of systematic bias and to normalize data [10, 20]. Under the assumption that the physical 3D distance between two loci is inversely related to the contact counts [14], the restriction of $\alpha_1 < 0$ is imposed in the model.

Letting θ denote the collection of all model parameters, we have the following log-likelihood function:

$$\log p(\mathbf{y}|\theta, \Omega) \propto \sum\sum_{(i,j)\in\mathscr{I}} \left\{ y_{ij} \log \lambda_{ij} - \log(e^{\lambda_{ij}} - 1) \right\}, \tag{3}$$

where $\mathscr{I}$ denotes the index set of non-zero contact counts, that is, $\mathscr{I} = \{(i, j); y_{ij} \neq 0, 1 \leq i < j \leq n\}$. This model, which excludes the zero contact counts, is referred to as the truncated Poisson Architecture Model (tPAM).

We remark that model (2) suffers from non-identifiability because the estimated structure, $\hat{\Omega}$, is not invariant to scale, rotation, reflection, and translation. To resolve this issue, without loss of generality, we can fix α_0 to be an arbitrarily predefined quantity. Note that α_0 controls the scale of the 3D structure, thus fixing α_0 will effectively lead to the structure being estimated only up to a scale. However, this is not an issue since the relative distance does not affect the predicted structure and its correlation with genomic functions [21]. Following [10], we further place the following restrictions on Ω to make it estimable, as four conditions on the structure are sufficient to uniquely determine the 3D structure: $\mathbf{p}_1 = (0, 0, 0)$, $\mathbf{p}_2 = (p_2^x, 0, p_2^z)$ with $p_2^z > 0$, $\mathbf{p}_3 = (p_3^x, p_3^y, p_3^z)$ with $p_3^y > 0$, and $\mathbf{p}_n = (p_n^x, 0, 0)$ with $p_n^x > 0$.

2.2 MCMC Procedure for Parameter Estimation

To make inferences about the 3D coordinates, we devise a Markov chain Monte Carlo (MCMC) sampling procedure as follows. We write the posterior distribution of Ω (main parameters of interest), together with nuisance parameters θ, as

$$p(\Omega, \theta|\mathbf{y}) \propto p(\mathbf{y}|\Omega, \theta) p(\Omega) p(\theta). \tag{4}$$

The first component of Eq. (4) corresponds to the likelihood as given in (3), that is,

$$p(\mathbf{y}|\Omega, \theta) = \prod\prod_{(i,j)\in\mathscr{I}} \mathscr{L}_P\{\lambda_{ij}(\Omega, \theta)\}, \tag{5}$$

where $\mathscr{L}_P(.)$ denotes the zero-truncated Poisson distribution and

$$\lambda_{ij}(\Omega, \theta) = \exp\left(\alpha_0 + \alpha_1 \log d_{ij} + \mathbf{x}_{ij}^T \beta\right). \tag{6}$$

The remaining parts of (4) describe the distributions for $\mathbf{p}$ and θ, which are assigned non-informative priors: $p(\Omega) \propto 1$, $p(\alpha_1) \propto I(\alpha_1 < 0)$, and $p(\beta) \propto 1$.

To accommodate the estimable conditions imposed on Ω, we consider an isometric transformation, with details provided in Appendix A. To sample from the posterior distributions of θ, we use Metropolis-Hastings algorithms, and in particular the Gibbs sampler whenever the conditional distribution of a parameter is of a commonly known one. In sampling the posterior of Ω, we employ Hamiltonian MCMC to more effectively handle the high correlations among the samples [7]. In the following, we briefly describe the updating schemes. Let ϑ denote the current estimates of (Ω, θ) at iteration t, and ϑ_{-a} denote ϑ without the element a.

- Updating of α_1.
  We base on the current α_1^t to sample a candidate α_1^* from proposal distribution $J_\alpha(\alpha_1^*|\alpha_1^t)$, a normal distribution with mean α_1^t and predefined proposal $\sigma_{\alpha_1}^2$, and calculate the ratio of the densities

  $$r = \frac{p(\alpha_1^*|\mathbf{y}, \vartheta_{-\alpha_1})}{p(\alpha_1^t|\mathbf{y}, \vartheta_{-\alpha_1})}, \tag{7}$$

  where $p(\alpha_1^*|\mathbf{y}, \vartheta_{-\alpha_1}) \propto p(\mathbf{y}|\vartheta_{-\alpha_1}, \alpha_1^*)$. Accept α_1^* as α_1^{t+1} with probability equal to $\min(r, 1)$; otherwise $\alpha_1^{t+1} = \alpha_1^t$.
- Updating of β_k, $k = 1, \ldots, K$.
  We base on the current β_k^t to sample a candidate β_k^* from proposal distribution $J_\beta(\beta_k^*|\beta_k^t)$, a normal distribution with mean β_k^t and predefined proposal σ_β^2, and calculate the ratio of the densities

  $$r = \frac{p(\beta_k^*|\mathbf{y}, \vartheta_{-\beta_k})}{p(\beta_k^t|\mathbf{y}, \vartheta_{-\beta_k})}, \tag{8}$$

  where $p(\beta_k^*|\mathbf{y}, \vartheta_{-\beta_k}) \propto p(\mathbf{y}|\vartheta_{-\beta_k}, \beta_k^*)$. Accept β_k^* as β_k^{t+1} with probability equal to $\min(r, 1)$; otherwise $\beta_k^{t+1} = \beta_k^t$.
- Updating of Ω.
  Based on an analogy with physical systems, Hamiltonian Monte Carlo introduces an additional parameter vector $\mathbf{v}_i = (v_i^x, v_i^y, v_i^z)^T$ corresponding to parameter $\mathbf{p}_i$ and updates both of them together in a new Metropolis-Hastings algorithm. Specifically, we use Hamiltonian functions defined by $H(\mathbf{p}_i, \mathbf{v}_i) = U(\mathbf{p}_i) +$

$K(\mathbf{v}_i)$, where $U(\mathbf{p}_i)$, a potential energy, is assigned $-\log\{p(\mathbf{p}_i|\mathbf{y}, \vartheta_{-\mathbf{p}_i})\}$, while $K(\mathbf{v}_i)$, a kinetic energy, is defined as $\mathbf{v_i}^T\mathbf{v_i}/2$. Then we consider the following joint density of $(\mathbf{p}_i, \mathbf{v}_i|\mathbf{y}, \vartheta_{-\mathbf{p}_i})$ using the Hamiltonian function $H(\mathbf{p}_i, \mathbf{v}_i)$:

$$p(\mathbf{p}_i, \mathbf{v}_i|\mathbf{y}, \vartheta_{-\mathbf{p}_i}) \propto \exp\{-H(\mathbf{p}_i, \mathbf{v}_i)\} = \exp\{-U(\mathbf{p}_i)\}\exp\{-K(\mathbf{v}_i)\}. \tag{9}$$

Hamiltonian MCMC then proceeds in three stages. First, we sample random auxiliary variables v_i^x, v_i^y, and v_i^z from $N(0, 1)$. Then we simultaneously update $(\mathbf{p}_i, \mathbf{v}_i)$ to obtain a proposal vector $(\mathbf{p}_i^*, \mathbf{v}_i^*)$ using a leapfrog method (see Appendix B). In the last stage, we accept the proposed vector $(\mathbf{p}_i^*, \mathbf{v}_i^*)$ using the Metropolis-Hastings method where the ratio is given by

$$r = \exp\{-H(\mathbf{p}_i^*, \mathbf{v}_i^*) + H(\mathbf{p}_i, \mathbf{v}_i)\}. \tag{10}$$

Accept $\mathbf{p}_i^*$ as $\mathbf{p}_i^{t+1}$ with probability $\min(r, 1)$; otherwise $\mathbf{p}_i^{t+1} = \mathbf{p}_i^t$.

3 Application to Two Hi-C Datasets

We demonstrate the utility of tPAM by applying it to two Hi-C datasets. The application to the first dataset illustrates tPAM's ability of analyzing inter-chromosomal data with many zero contact counts. Its performance is also evaluated by comparing the structure inferred to distances obtained from limited experimental validation data. The second application aims to explore how tPAM performs with modularized structures, the TADs, also known as topological domains [3].

3.1 Human Lymphoblastoid Cell Line Hi-C Data

We applied tPAM to the Hi-C data produced by [14]. In fact, there are two Hi-C experiments performed on the same karyotypical normal human lymphoblastoid cell line, which are combined into a single data set in our analysis given their high reproducibility [14]. We focused on chromosome 14 and 22, as experimental validation data based on Fluorescence In Situ Hybridization (FISH) are available for several loci on these two chromosomes and are publicly available [14]. Specifically, [14] discussed interesting features of spatial interactions, based on the FISH measures, among 4 loci on chromosome 14 (L_1, L_2, L_3, and L_4, located in that linear order) and 4 loci on chromosome 22 (L_5, L_6, L_7, and L_8, in that linear order) using the FISH experiment. In particular, the spatial 3D distance between L_2 and L_4 was observed by FISH experiments to be smaller than that between L_2 and L_3, despite the fact that L_2 is farther apart from L_4 than from L_3 in terms of their linear 1D distances. A similar observation was made for (L_6, L_7, L_8), in that the spatial 3D distance between L_6

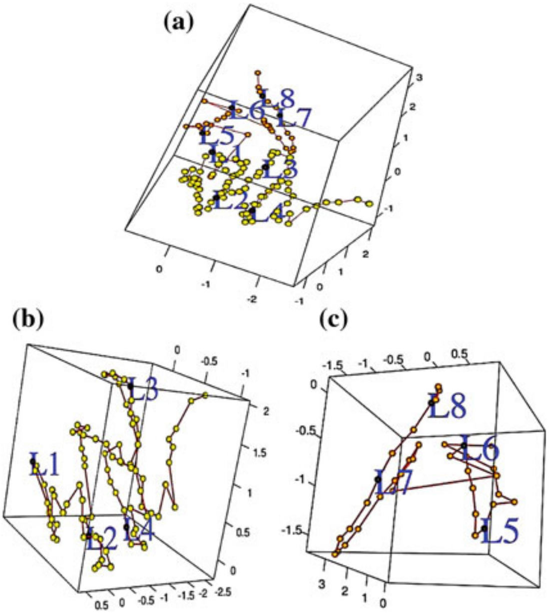


Fig. 2 Reconstructed 3D structure of chromosomes 14 and 22. **a** Joint 3D structure of chromosomes 14 and 22, with each loci marked by a ball, among them positions of L_1 through L_8 are labeled and marked by *black balls*; **b** 3D structure of chromosome 14, with a different orientation than that of the joint structure for better visualization; **c** 3D structure of chromosome 22, with a different orientation than that of the joint structure for better visualization. These figures were drawn using the R package 'rgl'

and L_8 is significantly smaller than that between L_6 and L_7. The resolution used is 1 Mb, which leads to 89 loci in chromosome 14 and 36 loci in chromosome 22.

We ran the MCMC procedure for 1.1×10^6 iterations, with the first 10^5 iterations for burn-in and the remaining 10^6 iterations for obtaining 10,000 posterior samples after thinning. The convergence of the posterior samples was confirmed by several diagnostic statistics, including those developed by [8, 9, 15]. The 3D structure identified by tPAM is given in Fig. 2a. For a better visualization of the structure in each of the chromosomes, we also provide Fig. 2b, c with different orientations. We can see from these figures that, indeed, L_2 and L_4 are much closer in terms of their spatial distance compared to L_2 and L_3, and L_6 and L_8 are closer compared to L_6 and L_7. These observations are consistent with the results of [14] that the pairs of (L_2, L_4) and (L_6, L_8) are brought to close proximity through chromatin looping.

To further evaluate the performance of tPAM, we compare its estimates of pairwise distances to those of FISH, the gold standard measurements. To make it possible to compare due to scale differences (recall we set α_0 arbitrarily), we first calculated a unitless distance $\tilde{d}(L_i, L_j)$ by dividing each distance $d(L_i, L_j)$ by the median distance between L_3 and L_4 (the largest distance among all pairs). Note that the median is taken over 100 measurements for FISH and 10,000 estimates for tPAM. The results, given in Fig. 3, show that the tPAM estimates agree well with the FISH measurements. In fact, the FISH measurements (100 measures for each pair) are much more variable compared to the tPAM estimates, as evident from the larger

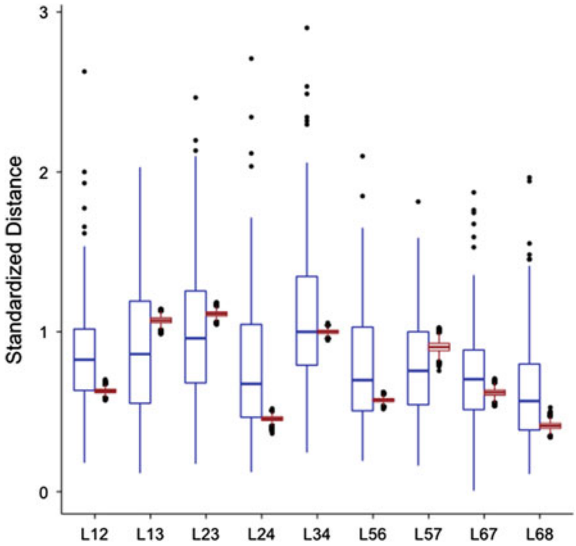


Fig. 3 Assessment of performance of tPAM in comparison with FISH measurements. For each pair of loci for which FISH measurements are available, boxplots are used to summarize the results for the 100 FISH measurements (*left box*) and 10,000 tPAM estimates (*right box*)

boxes, longer whiskers, and existence of outliers in the boxplots. The results also confirm that the distance between L_2 and L_4 is indeed smaller than that between L_2 and L_3 or L_3 and L_4, and L_6 is located closer to L_8 than to L_7.

3.2 Mouse Embryonic Stem Cell Hi-C Data

We applied tPAM to a mouse embroyonic stem cell line [3] generated at 40 Kb resolution (i.e. interaction frequencies are available for regions of 40 Kb in length). We used the bias-corrected Hi-C count data directly, as libraries of factors that are known to cause systematic biases are not available to us. In particular, we focused on the segment of chromosome 2 from base pair (bp) 73720001 to bp 75440000, as this segment is believed to contain two TADs [3]. Loci within the same domain interact with each other much more than across domains, and thus the two domains should be well separated in 3D space. The data based on a 40 Kb resolution lead to a contact matrix of dimensions 43 by 43. Application of tPAM yielded the estimated 3D structure depicted in Fig. 4. We can see, from the figure, that the 19 loci within the segment from bp 73720001 to bp 74480000 are located close to one another in 3D space (red balls), whereas the remaining 24 loci within the segment from bp 74480001 to bp 75440000 make up the other cluster (green balls) in 3D space. As it turns out, these two clusters of loci do correspond to the two TADs discussed in [3]. In MCMC sampling, 3×10^5 and 7×10^5 iterations were executed respectively for burn-in and statistical inference. Thinning resulted in 10,000 posterior samples for structure estimation. Convergence of the sample was confirmed by the diagnostic measures described in Sect. 2.

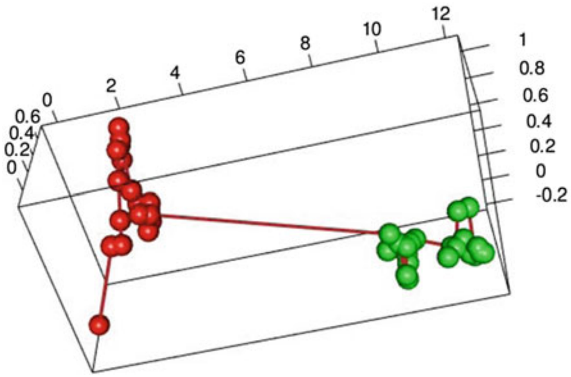


Fig. 4 Reconstructed 3D structure of mouse data. Loci within the two topological domains are denoted by two different colors

4 Simulation Study

As we can see from the analysis results of the human Hi-C data, the inferred 3D structure from tPAM leads to consistent results with FISH experimental data. Nevertheless, the aptness of the 3D structure as a whole was not adequately assessed due to the limited number of loci involved in the FISH experiment. Similarly, although the analysis of the Hi-C mouse data yielded results that support the concept of compartmentalization of a chromosome [3, 14], the within compartment (domain) organization was not assessable. Therefore, to more fully evaluate the performance of tPAM, we conducted a simulation study in this section using two underlying 3D structures, which will serve as the "gold standard". We further compared the performance of tPAM with BACH, a Bayesian inference method proposed by [10] based on the Poisson model. The simulation settings and results are presented in two subsections below, but we first describe several assessment criteria for comparing the performances between tPAM and BACH.

4.1 Performance Assessment

We consider three criteria to assess the performance of the methods. The first is the overall goodness of fit of a model by comparing the observed with their predicted values from the model. More specifically, our measure is the Pearson χ^2 goodness of fit statistic, which is given by

$$\chi^2 = \sum\sum_{(i,j)\in\mathscr{I}} \frac{(y_{ij} - \hat{\lambda}_{ij})^2}{\hat{\lambda}_{ij}} / n(\mathscr{I}), \tag{11}$$

where $\mathscr{I}$ is the index set denoting all non-zero contact counts as defined in Sect. 2 and $n(\mathscr{I})$ denotes a size of the set $\mathscr{I}$.

Given that, in our simulation, the underlying structure is known, we can also devise two other criteria that make use of the true underlying distance between a pair of loci. Recall that the structure estimated is accurate up to a scaling factor, γ, which is estimated by the least squares model as follows:

$$\hat{\gamma} = \arg\min_{\gamma} \left\{ \sum\sum_{1 \le i < j \le n} (d_{ij} - \gamma \hat{d}_{ij})^2 \right\}. \tag{12}$$

Note that, as mentioned above, the fact that tPAM or BACH can only estimate the structure up to a scale is not an issue, because the relative distance does not affect the predicted structure nor its correlation with genomic functions [21]. After scaling the estimated structure $\hat{\Omega}$ by the factor estimate $\hat{\gamma}$, we can compare the true structure with the estimated structure after appropriate isometric transformation. This leads to the proposal of the following two measures:

$$\mathscr{D}_{mean} = \frac{1}{n} \sum_{i=1}^{n} \frac{||\mathbf{p}_i - \hat{\gamma}\hat{\mathbf{p}}_i||}{\bar{d}_{\mathbf{p}}} \times 100 \tag{13}$$

$$\mathscr{D}_{max} = \max_{1 \le i \le n} \frac{||\mathbf{p}_i - \hat{\gamma}\hat{\mathbf{p}}_i||}{\bar{d}_{\mathbf{p}}} \times 100, \tag{14}$$

where $\bar{d}_{\mathbf{p}}$ is the average pairwise distance derived from the true underlying structure Ω. Thus, these two measures compute respectively the average- and the maximum-coordinate departure of loci (based on the estimated architecture) from the corresponding true ones (based on the true architecture). As we will see below, the true structures are being specified completely either based on the helix model or the estimated mouse model for the purpose of the simulation study.

4.2 Helix Structure

We consider a helix model with 50 loci. We chose this model for our first simulation as a helix structure has been used as a means of modeling chromatin in the statistical literature [19]. We denote the helix structure by $\Omega^h = \{\mathbf{p}_i, i = 1, \dots, 50\}$. The 3D location of each locus, $\mathbf{p}_i = (p_i^x, p_i^y, p_i^z)$, is constructed as

$$p_i^x = \cos(\theta_i),\ p_i^y = \sin(\theta_i),\ p_i^z = L\theta_i/(2\pi), \tag{15}$$

where $L = 0.2$ and $\theta_i = \pi i/4$. To mimic real data, we also include three covariates, $\{x_{l,i}, x_{g,i}, x_{m,i}, i = 1, \dots, 50\}$, to capture systematic bias, leading to the following simulation model:

$$\log \lambda_{ij} = \alpha_0 + \alpha_1 \log d_{ij} + \beta_l \log(x_{l,i}x_{l,j}) + \beta_g \log(x_{g,i}x_{g,j}) + \log(x_{m,i}x_{m,j}). \tag{16}$$

We set $\alpha_0 = 3.5$, and $\alpha_1 = -1.5$, $\beta_l = \beta_g = 0.3$ and simulated $x_{l,i} \sim \text{Unif}(0.2, 0.3)$, $x_{g,i} \sim \text{Unif}(0.4, 0.5)$ and $x_{m,i} \sim \text{Unif}(0.9, 1)$, where Unif(.) denotes a uniform distribution. To simulate the excess of zero situation in real data, we considered the following zero-inflated Poisson model:

$$P(Y_{ij} = 0) = \pi + (1-\pi)e^{-\lambda_{ij}},$$
$$P(Y_{ij} = y_{ij}) = (1-\pi)\frac{\lambda_{ij}^{y_{ij}} e^{-\lambda_{ij}}}{y_{ij}!}, \quad y_{ij} = 1, 2, \ldots. \tag{17}$$

In other words, the above represents a mixture of a point mass at 0 and a Poisson distribution with intensity parameter λ_{ij}, with the mixing proportion being π. In our simulation, we considered four mixing proportions: $\pi = 0.0, 0.1, 0.2$, and 0.3. Note that the setting with $\pi = 0.0$ corresponds to the BACH model of [10] and as such, BACH is expected to perform well.

The results are presented in Table 1. In MCMC sampling, $10^5 \sim 10^6$ iterations were run for burn-in and an additional $10^6 \sim 2 \times 10^6$ iterations were executed for posterior sampling to obtain 10^4 realizations for inference after thinning. The convergences of the posteriors were confirmed by the diagnostics described in Sect. 2. As we can see from the table, across all three criteria, tPAM performs significantly better than BACH for the settings when $\pi \neq 0$. More specifically, tPAM yielded significantly smaller average and maximum relative departure from the true Ω^h (all p-values $<10^{-3}$ based on paired-t tests). This is to be expected as BACH, based on Poisson, cannot adequately accommodate the excess of zeros. We are also reassured to see that, even when $\pi = 0$, the underlying setting of BACH, tPAM still performs as well as BACH or may even be viewed as slightly better based on all three criteria. We can further observe that the results of tPAM are fairly consistent for different zero inflation proportions (i.e. similar values under the same criterion), demonstrating the robustness of tPAM to excess of zeros in the observed data, and hence data with different resolutions. In contrast, BACH's performance gets worse (with larger criterion value) as the inflation proportion becomes larger.

Table 1 Performance evaluation of tPAM and BACH with the Ω^h 3D structure

π	Model	$\mathscr{D}_{mean}$ (%)	$\mathscr{D}_{max}$ (%)	χ^2
0.0	BACH	26.37 (17.70)	63.99 (39.26)	1.04 (0.13)
	tPAM	23.70 (11.15)	60.11 (29.99)	0.98 (0.11)
0.1	BACH	39.14 (17.79)	96.66 (35.83)	2.03 (0.24)
	tPAM	23.65 (12.94)	57.12 (32.38)	0.98 (0.13)
0.2	BACH	61.07 (25.41)	140.84 (51.43)	3.94 (0.44)
	tPAM	25.79 (11.74)	59.96 (28.19)	0.95 (0.19)
0.3	BACH	62.65 (20.06)	142.05 (40.83)	7.16 (0.70)
	tPAM	26.49 (16.67)	65.56 (46.17)	0.88 (0.07)

Table 2 Performance evaluation of tPAM and BACH with the Ω^m 3D structure

π	Model	$\mathscr{D}_{mean}$ (%)	$\mathscr{D}_{max}$ (%)	χ^2
0.0	BACH	49.85 (5.14)	93.60 (7.43)	1.23 (0.04)
	tPAM	39.57 (7.55)	74.16 (15.08)	1.80 (0.76)
0.1	BACH	65.26 (11.20)	109.40 (15.63)	1.65 (0.17)
	tPAM	42.51 (9.45)	77.26 (14.93)	1.42 (0.56)
0.2	BACH	77.65 (13.52)	124.67 (19.56)	3.43 (0.36)
	tPAM	43.00 (8.63)	79.56 (15.25)	1.62 (0.71)
0.3	BACH	84.41 (15.36)	139.94 (23.14)	6.90 (0.88)
	tPAM	46.67 (20.76)	89.78 (45.03)	1.36 (0.52)

4.3 *Mouse Model*

Using the mouse structure $\hat{\Omega}^m$ and the $\hat{\alpha}_1$ value estimated by tPAM in Sect. 3.2, we let $\log \lambda_{ij} = 3 + \hat{\alpha}_1 \log d_{ij}$, where d_{ij} is the pairwise distance inferred from the estimated structure $\hat{\Omega}^m$. We simulated datasets of $\{Y_{ij}\}$ from the zero-inflated Poisson model (17) with $\pi = 0.0, 0.1, 0.2$, and 0.3. In MCMC sampling, $7 \times 10^5 \sim 10^6$ iterations were run for burn-in, and afterward $5 \times 10^5 \sim 10^6$ iterations were run to obtain 10^4 realizations for inference after thinning. As with the helix simulation, the convergences of the posteriors were confirmed by the diagnostics described in Sect. 2. The results are given in Table 2, from which, one can see that tPAM clearly outperforms BACH for $\pi \neq 0$ (all p-values $\leq 10^{-4}$ based on paired-t tests), consistent with the results for the helix model. Similarly, when $\pi = 0.0$, the underlying model for BACH, tPAM is seen to perform just as well. The robustness of tPAM to the proportion of zero-inflation component, and the lack of such for BACH, is once again observed.

5 Conclusion and Discussion

The spatial organization of a genome has gained a great deal of continuing attention in recent years, as the structure is intimately linked to the biological functions of the genome, especially on long-range gene regulation. To turn experimental data into accurate estimates of spatial chromatin structures, a number of analytical methods have been proposed, including those that make use of the Poisson distribution to model the contact counts. Recognizing the sparsity of the contact matrix for inter-chromosomal interactions and with higher resolutions, in this paper, we propose a truncated Poisson model as a solution to accommodate this feature of data so that it is robust to resolution specification. Applications of tPAM to two existing data sets, one human and one mouse, illustrate its utility, as the results are consistent with those obtained from the limited FISH validation data. For the mouse data, with a 40

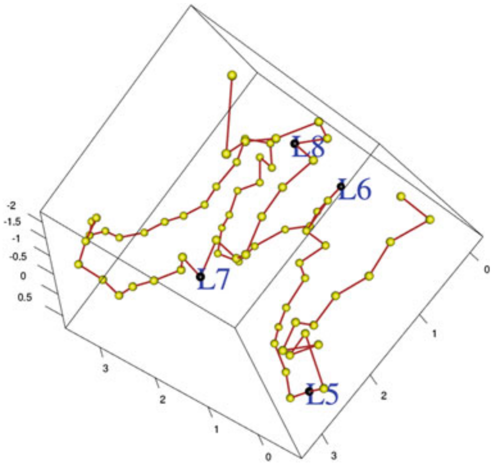


Fig. 5 Reconstructed 3D structure of chromosome 22 with 500 Kb resolution

Kb resolution, we see two clear TADs, reflecting chromatin long-range interaction in a “domain scale”. Within each domain, with such an intermediate resolution, we can see looping within each domain, perhaps representing spatial interaction within a gene structure. For the human data, the analysis was performed at a 1 Mb resolution following the original analysis [14], which appears to capture the broad looping feature of chromatin organization, but fine scale looping within gene structures are largely unobserved. Inspired by the mouse data results with intermediate resolution, we carried out an additional analysis for constructing the 3D structure of chromosome 22 at a 500 Kb resolution. We observe that the result (Fig. 5) preserves the “domain level” looping, with locus L_6 still closer to L_8 than to L_7. Furthermore, the finer structure now also depicts more “local level” looping. Nevertheless, a more comprehensive study with even higher resolution is needed to study spatial interactions within gene structures, especially between promoters and enhancers.

Our simulation study, with two underlying structures, further substantiates the appropriateness of tPAM for analyzing Hi-C data, and more clearly showcases its ability to handle the sparsity of the contact matrix. The different mixing proportions in the zero-inflated model can be viewed as representing different resolutions, thus clearly demonstrating the robustness of tPAM to varying resolution level. This is in contrast to an existing method based on the Poisson model, in which one can see that the results are quite sensitive to the level of resolution: as the resolution gets finer and finer, the deviation from the “true” gets larger and larger for each of the evaluation criteria, compared to the stable feature of the tPAM values.

Computational feasibility is a major concern for genomic data, but the concern is even greater for chromatin interaction data as the size of the data is $O(n^2)$ when there are n genomic loci, an order of magnitude increase compared to analysis of linear chromosomal data. In this regard, tPAM has the added advantage as its computational cost is greatly reduced by excluding the zero counts. As such, higher resolution data, which lead to a much larger contact matrix (i.e. larger n), does not necessarily result in more computational cost due to the sparsity nature of the matrix. In contrast, for methods based on the Poisson distribution, the computational cost increases with higher resolution data.

Acknowledgments This work was supported in part by the National Science Foundation grants DMS-1042946 and DMS-1220772, the National Institute of Health Grant IROIGM114142-01, and by the Bisa Research Grant of Keimyung University in 2014. This material was also based upon work partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Appendices

A. Isometric Transformation

To make Ω uniquely estimable, instead of incorporating the restrictions on Ω into prior, we employed a group of isometric (distance preserving) mappings. Suppose we sample Ω^t at iteration t. For simplicity, we let Ω denote the transformed one throughout the rest of this appendix.

Step 1. $\mathbf{p}_1 \rightarrow (0, 0, 0)$.

To place $\mathbf{p}_1^t$ at the origin (0, 0, 0), we apply a translation operation $\mathscr{R}_\tau$ such that

$$\mathscr{R}_\tau : \mathbf{p}_i^t \rightarrow \mathbf{p}_i^t - \mathbf{p}_1^t. \tag{18}$$

Let $\Omega = \{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$ be the translated architecture.

Step 2. $\mathbf{p}_n \rightarrow (p_n^x, 0, 0)$ with $p_n^x > 0$.

a. $\mathbf{p}_n \rightarrow (p_n^x, 0, p_n^z)$.

To place $\mathbf{p}_n$ on the xz-plane, we apply a rotation operation $\mathscr{R}_{\overset{\circ}{z}}$ with associated matrix $R_{\overset{\circ}{z}}$, clockwise-rotation matrix on $\mathbf{p}_n$ about the z-axis, sending it to the xz-plane:

$$R_{\overset{\circ}{z}} = \begin{bmatrix} \cos\phi_1 & \sin\phi_1 & 0 \\ -\sin\phi_1 & \cos\phi_1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

where,

$$\cos\phi_1 = p_n^x / \sqrt{(p_n^x)^2 + (p_n^y)^2},$$
$$\sin\phi_1 = p_n^y / \sqrt{(p_n^x)^2 + (p_n^y)^2}.$$

Let $\varOmega = \{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$ be the rotated architecture.

b. $\mathbf{p}_n \to (p_n^x, 0, 0)$.

To place $\mathbf{p}_n$ on the x-axis, we apply a rotation operation $\mathscr{R}_{\mathring{y}}$ with associated matrix $R_{\mathring{y}}$, a clockwise-rotation matrix around the y-axis:

$$R_{\mathring{y}} = \begin{bmatrix} \cos\phi_2 & 0 & \sin\phi_2 \\ 0 & 1 & 0 \\ -\sin\phi_2 & 0 & \cos\phi_2 \end{bmatrix},$$

where

$$\cos\phi_2 = p_n^x / \sqrt{(p_n^x)^2 + (p_n^z)^2},$$
$$\sin\phi_2 = p_n^z / \sqrt{(p_n^x)^2 + (p_n^z)^2}.$$

Let $\varOmega = \{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$ be the rotated architecture.

Step 3. $\mathbf{p}_2 \to (p_2^x, 0, p_2^z)$ with $p_2^z > 0$.

To place $\mathbf{p}_2$ on the xz-plane, we apply a counter-clockwise rotation about the x-axis $\mathscr{R}_{\mathring{x}}$ with associated matrix $R_{\mathring{x}}$:

$$R_{\mathring{x}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi_3 & -\sin\phi_3 \\ 0 & \sin\phi_3 & \cos\phi_3 \end{bmatrix},$$

where

$$\cos\phi_3 = p_2^z / \sqrt{(p_2^y)^2 + (p_2^z)^2},$$
$$\sin\phi_3 = p_2^y / \sqrt{(p_2^y)^2 + (p_2^z)^2}.$$

Let $\varOmega = \{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$ be the rotated architecture.

Step 4. $\mathbf{p}_3 \to (p_3^x, p_3^y, p_3^z)$ such that $p_3^y > 0$.

To satisfy $p_3^y > 0$, if $p_3^y < 0$, reflect $\mathbf{p}$ as

$$\mathscr{R}_{rfl} : p_i^y \to -p_i^y. \tag{19}$$

Let transformation $\mathscr{I}$ be the composite of the five isometric transformations, $\mathscr{R}_\tau$, $\mathscr{R}_{\mathring{z}}$, $\mathscr{R}_{\mathring{y}}$, $\mathscr{R}_{\mathring{x}}$, and $\mathscr{R}_{rfl}$ in the following way: $\mathscr{I} \equiv R_{rfl}\mathscr{R}_{\mathring{x}}\mathscr{R}_{\mathring{y}}\mathscr{R}_{\mathring{z}}\mathscr{R}_\tau$. Then $\mathscr{I}$ is an isometric (distance-preserving) transformation and the transformed coordinates satisfy the following estimability conditions on $\mathbf{p} : \mathbf{p}_1 = (0, 0, 0)$, $\mathbf{p}_2 = (p_2^x, 0, p_2^z)$ with $p_2^z > 0$, $\mathbf{p}_3 = (p_3^x, p_3^y, p_3^z)$ with $p_3^y > 0$, and $\mathbf{p}_n = (p_n^x, 0, 0)$ with $p_n^x > 0$.

B. Leapfrog Method for Hamiltonian MCMC

In the second stage of Hamiltonian MCMC, we simultaneously update $(\mathbf{p}_i, \mathbf{v}_i)$ to obtain a proposal vector $(\mathbf{p}_i^*, \mathbf{v}_i^*)$ using a leapfrog method which involves a leap scale ε and a repetition number L:

(1) For each of x, y, z, update v_i^x, v_i^y, v_i^z as

$$v_i^{(.)} \leftarrow v_i^{(.)} + \frac{1}{2}\varepsilon \frac{d \log p(p_i^{(.)}|\mathbf{y}, \vartheta_{-\mathbf{p}_i})}{dp_i^{(.)}}. \tag{20}$$

(2) Repeat the following updates $L - 1$ times:

$$v_i^{(.)} \leftarrow v_i^{(.)} + \frac{1}{2}\varepsilon \frac{d \log p(p_i^{(.)}|\mathbf{y}, \vartheta_{-\mathbf{p}_i})}{dp_i^{(.)}}, \qquad p_i^{(.)} \leftarrow p_i^{(.)} + \varepsilon v_i^{(.)}. \tag{21}$$

(3) Update v_i^x, v_i^y, v_i^z as

$$v_i^{(.)} \leftarrow v_i^{(.)} + \frac{1}{2}\varepsilon \frac{d \log p(p_i^{(.)}|\mathbf{y}, \vartheta_{-\mathbf{p}_i})}{dp_i^{(.)}}. \tag{22}$$

(4) The updated $\mathbf{p}_i$ and $\mathbf{v}_i$ constitute a proposal vector $(\mathbf{p}_i^*, \mathbf{v}_i^*)$.
In the leapfrog method, the essential quantities to evaluate are

$$\frac{d \log p(p_i^x|\mathbf{y}, \vartheta_{-\mathbf{p}_i})}{dp_i^x} = \sum_{j \neq i} \left(y_{ij} - \lambda_{ij} \frac{e^{\lambda_{ij}}}{e^{\lambda_{ij}} - 1} \right) \alpha_1 \frac{p_i^x - p_j^x}{\delta_{ij}^2}, \tag{23}$$

$$\frac{d \log p(p_i^y|\mathbf{y}, \vartheta_{-\mathbf{p}_i})}{dp_i^y} = \sum_{j \neq i} \left(y_{ij} - \lambda_{ij} \frac{e^{\lambda_{ij}}}{e^{\lambda_{ij}} - 1} \right) \alpha_1 \frac{p_i^y - p_j^y}{\delta_{ij}^2}, \tag{24}$$

$$\frac{d \log p(p_i^z|\mathbf{y}, \vartheta_{-\mathbf{p}_i})}{dp_i^z} = \sum_{j \neq i} \left(y_{ij} - \lambda_{ij} \frac{e^{\lambda_{ij}}}{e^{\lambda_{ij}} - 1} \right) \alpha_1 \frac{p_i^z - p_j^z}{\delta_{ij}^2}. \tag{25}$$

References

1. Baù, D., A. Sanyal, B.R. Lajoie, E. Capriotti, M. Byron, et al. 2011. The three-dimensional folding of the a-globin gene domain reveals formation of chromatin globules. *Nature Structural and Molecular Biology* 18: 107–114.
2. Ben-Elazar, S., et al. 2013. Spatial localization of co-regulated genes exceeds genomic gene clustering in the saccharomyces cerevisiae genome. *Nucleic Acids Research* 41: 2191–2201.
3. Dixon, J.R., S. Selvaraj, F. Yue, et al. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485: 376–380.

4. Duan, Z., M. Andronescu, K. Schutz, S. McIlwain, et al. 2010. A three-dimensional model of the yeast genome. *Nature* 465: 363–367.
5. Fraser, J., M. Rousseau, S. Shenker, M.A. Ferraiuolo, et al. 2009. Chromatin conformation signatures of cellular differentiation. *Genome biology* 10: R37+.
6. Fullwood, M.J., M.H. Liu, Y.F. Pan, J. Liu, et al. 2011. TAn oestrogen-receptor-[agr]-bound human chromatin interactome. *Nature* 462: 58–64.
7. Gelman, A., J.B. Carlin, H.S. Stern, D.B. Dunson, et al. 2013. *Bayesian Data Analysis, Third Edition (Chapman and Hall/CRC Texts in Statistical Science).* Chapman and Hall/CRC
8. Geweke, J. 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Bayesian Statistics (Vol. 4, pp. 169–193). Oxford: Oxford University Press.
9. Heidelberger, P., and P.D. Welch. 1983. Simulation Run Length Control in the Presence of an Initial Transient. *Operations Research* 31: 1109–1145.
10. Hu, M., K. Deng, Z. Qin, et al. (2013). Bayesian inference of spatial organizations of chromosomes. *PLOS Computational Biology* 9: e1002893+.
11. Imakaev, M., G. Fudenberg, R. McCord, et al. 2012. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature Methods* 9: 999–1003.
12. Kalhor, R., H. Tjong, N. Jayathilaka, et al. 2012. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology* 30: 90–98.
13. Lesne, A., J. Riposo, P. Roger, et al. (2014). 3D genome reconstruction from chromosomal contacts. *Nature Biotechnology*, advance online publication.
14. Lieberman-Aiden, E., N.L. van Berkum, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326: 289–293.
15. Raftery, A.E., and S.M. Lewis. (1995). The number of iterations, convergence diagnostics and generic Metropolis algorithms, *In Practical Markov Chain Monte Carlo*, (pp. 115–130).
16. Rousseau, M., J. Fraser, M. Ferraiuolo, J. Dostie, and M. Blanchette. (2011). Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling, *BMC Bioinformatics* 12: 414+.
17. Tanizawa, H., O. Iwasaki, A. Tanaka, et al. 2010. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Research* 38: 8164–8177.
18. Varoquaux, N., F. Ay, W.S. Noble, and J. Vert. 2014. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics* 30: 26–33.
19. Xiao, G., X. Wang, and A.B. Khodursky. 2011. Modeling three-dimensional chromosome structures using gene expression data. *Journal of the American Statistical Association* 106: 61–72.
20. Yaffe, E., and A. Tanay. 2011. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics* 43: 1059–1065.
21. Zhang, Z., Li, G., K. Toh, and W. Sung. 2013. Inference of spatial organizations of chromosomes using semi-definite embedding approach and Hi-c data. *Proceedings of the 17th International Conference on Research in Computational Molecular Biology* 16: 317–332.