Nelle Varoquaux^{1,2}

Unfolding the Genome: The Case Study of *P. falciparum*

¹ Statistics, University of California, Berkeley, 367 Evans Hall, Berkeley, California, USA, E-mail: nelle@berkeley.edu.
<http://orcid.org/0000-0002-8748-6546>.

² Berkeley Institute for Data Science, 190, Doe library Berkeley, United States of America, E-mail: nelle@berkeley.edu.
<http://orcid.org/0000-0002-8748-6546>.

Abstract:

The development of new ways to probe samples for the three-dimensional (3D) structure of DNA paves the way for in depth and systematic analyses of the genome architecture. 3C-like methods coupled with high-throughput sequencing can now assess physical interactions between pairs of loci in a genome-wide fashion, thus enabling the creation of genome-by-genome contact maps. The spreading of such protocols creates many new opportunities for methodological development: how can we infer 3D models from these contact maps? Can such models help us gain insights into biological processes?

Several recent studies applied such protocols to *P. falciparum* (the deadliest of the five human malaria parasites), assessing its genome organization at different moments of its life cycle. With its small genomic size, fairly simple (yet changing) genomic organization during its lifecycle and strong correlation between chromatin folding and gene expression, this parasite is the ideal case study for applying and developing methods to infer 3D models and use them for downstream analysis.

Here, I review a set of methods used to build and analyse three-dimensional models from contact maps data with a special highlight on *P. falciparum*'s genome organization.

Keywords: Hi-C, 3D structure, *P. falciparum*, inference

DOI: 10.1515/ijb-2017-0061

Received: August 1, 2017; **Revised:** February 2, 2018; **Accepted:** May 10, 2018

1 Introduction

With more than 216 millions cases and nearly 445 000 deaths in 2016, malaria remains a major disease burden in tropical and subtropical countries and an impediment to economic development. In Africa, where more than 90% of the cases and deaths occur, the effects of malaria extend far beyond direct measures of mortality; the disease is thought to cost more than US\$21 billion a year [1]. The disease is caused by a small unicellular protozoan parasite *Plasmodium* that infects its host through a mosquito bite. Out of the 5 species that can infect humans, *P. falciparum* is by far the deadliest with 99% of all malaria-related deaths associated with it. The parasite has a complex life cycle, with multiple stages both in the human and mosquito hosts (see Figure 1).

Nelle Varoquaux is the corresponding author.

© 2018 Nelle Varoquaux, published by Walter de Gruyter GmbH, Berlin/Boston.

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.

Figure 1: The life cycle of *P. falciparum*.

The human host is infected by *sporozoites* through the bite of an infected female *Anopheles* mosquito. The *sporozoites* quickly migrate to the liver, where they start a two week-long multiplication process. Merozoites are then released in the blood-stream and proceed to infect red blood cells. The parasites then start their “erythrocytic” cycle (through *rings*, *trophozoites* and *schizonts* stages), via another round of replication in red blood cells. This replication occurs via an unusual process of cell division called *schizogony*. In *schizogony* replication, the parasite first undergoes multiple rounds of nuclear replication involving division into 13 to 32 daughter cells, until the red blood cell bursts and provokes the release of *merozoites* and the infective cycle starts anew. This asexual replication cycle is responsible for the symptoms and the complications of the disease: anemia, tertian fever, ...

While effective vaccines still remain a hope and resistance to anti-malaria drugs continues to rise, the focus of many recent genomic-based research studies in malaria have been in the development of novel therapies [2]. However, one of the limiting factors in the development of new drugs is our poor understanding of the mechanisms underlying the parasite’s complex life cycle. While the development of *P. falciparum* through the different stages of its life is thought to be driven by coordinated changes in gene expression, the relative paucity of transcription factors points to unusual gene regulatory mechanisms. Meanwhile, the relative abundance of proteins related to chromatin structures, mRNA decay, and translation rates suggest alternative mechanisms of gene regulation at the epigenetic and post-translational levels [3–7]. Thus an improved understanding of the *P. falciparum* genome architecture, at both local and global scales, will provide clues for developing new therapies.

In recent years, chromosome conformation capture-like methods, broadly referred to as Hi-C, have allowed for the identification of physical interactions between two regions in a genome-wide fashion, yielding information on their relative spatial distance in the nucleus [8]. Hi-C has opened new avenues for more systematic analyses of the three-dimensional folding of the genome, paving the way for a better understanding of the relations between 3D structure and gene regulation, replication timing, epigenetic changes, as well as many other biological processes [9–11].

With the aid of these techniques, researchers have undertaken a wide variety of studies examining the 3D genomic structure of many different organisms, including those on several species of yeast [12–14], bacteria [15], flies [16], plants [17, 18] and numerous human and mouse cell lines [8, 10, 11]. Moreover, there have been two recent studies specifically focusing on the three-dimensional structure of the *P. falciparum* genome. [19] probed several strains of *P. falciparum* to understand the link between 3D structures and the complex regulation of *var* genes, a family of genes involved in the invasion of red blood cells and also responsible for the parasite’s great capacity to evade our immune system. [20] assayed the 3D structure of the genome of the parasite during three key stages of the erythrocytic cycle. The relatively small size of the genome yet relatively complicated genome architecture, the complex life cycle, and the strong link between chromatin folding, gene regulation, and epigenetics makes *P. falciparum* a case of choice for the study of genome folding and its link to gene regulation [19, 20]. [21].

The inference of accurate 3D models plays an essential role in the study of the structure of the genome of *P. falciparum*, as well as that of many other organisms [13, 20, 22]. While these models are interesting as stand-alone entities, they are actually more useful when provided as inputs into various analyses, such as identifying colocalized elements, or distinguishing between open and closed chromatin. One particularly notable use of these 3D models is in their integration with other sources of data such as gene expression or chromatin modi-

fication [13, 20, 22]. In recent years, many methods have been developed for creating such 3D models, either as standalone methods [23, 24] or as context-specific methods whose goal might be to better understand a specific organism [13, 15, 21] or biological process (such as the inactivation of chromosome X) [22, 25].

The methods for creating 3D models broadly fall into two categories: “model-based” and “data-driven.” The former (“model-based”) methods consider the polymer nature of DNA to leverage the theoretical and computational work done in statistical physics of polymers, to build with as few assumptions as possible, many chromosome conformations. Those chromosome conformations are then used to stand against experimental data, such as Hi-C contact count matrices, in order to iteratively improve the models. These models offer mechanistical insights into the folding of DNA. In contrast, the latter (“data-driven”) approaches use the experimental data to infer 3D models, typically by minimizing a cost function ensuring the models are as consistent with the data as possible.

This paper reviews the developments of 3D models using contact maps to better understand genomic and epigenetic processes, with a particular highlight on the *P. falciparum*. It dwells on modeling challenges: why and how to construct 3D models from contact maps, and explains what 3D models can teach us about biological processes. The first and second sections of this paper discuss respectively existing data- and model-driven methods for building 3D structures of DNA using contact maps (see Figure 2). The final section reviews how these models can be used to uncover key roles of genome architecture in biological processes, with a particular focus on uncovered features of *P. falciparum* genome architecture.

Figure 2: Understanding 3D genome structure using contact maps.

Approaches to studying the 3D structure of the genome broadly fall into two categories. The first uses contact count data to infer 3D models, while the second creates models, and validates those using contact count data. The 3D structures can then be used to gain biological insights on the organism of interest.

2 Inferring three-dimensional models of DNA from contact maps

The development of genome-wide and high-throughput protocols to probe samples for their 3D genome architecture naturally paved the way to systematic and in depth studies of the folding mechanisms of DNA. Over the past 10 years, the challenge of building 3D models from contact maps has been tackled by a plethora of methods, some developed for a particular organism, others with the aim of being generalizable to any data sets [13, 15, 20, 22, 24, 26–35]. These methods model chromosomes as a series of beads, and attempt to place the beads in a 3D Euclidean space to accurately represent the contact map. While the methods overall fall into two categories “model-driven” and “data-driven,” each of these categories can itself be broken down further. “Data-driven” methods broadly fall into two groups: (i) *consensus* approaches, that aim at inferring a unique mean structure best representing the contact count data; (ii) *ensemble* methods that yield a population of structures.

Both *consensus* and *ensemble* methods have benefits and drawbacks. *Ensemble* approaches are more biologically accurate: Hi-C data are derived from a population of cells, each of those with a uniquely folded 3D structure. The variability of structures among our cells may thus be better represented by a population of structures. Yet, in addition of being more complex and computationally intensive, *ensemble* approaches raise the question of interpretability. Often, one has to fall back on interpreting the mean structure [36], or a reduced set of struc-

tures [31]. *Consensus* approaches yield a single structure recapitulating the rich information provided by Hi-C data, including hallmarks of genome architecture shared across all cells despite the cell-to-cell variability [37]. More amenable to visualization and analysis, this average structure can be easily integrated with other sources of data, such as RNA-seq, chip-seq etc, which are also population-based. However, interpretation of *consensus* model needs to be handled with care: these structures should not be considered as a true representation of the chromatin folding in the nucleus, but as a summary model of the contact maps.

Table 2 presents a summary of 3D inference methods.

2.1 Notations

In most methods, chromosomes are modeled as a serie of beads in 3D, each bead representing a *loci* or genomic window of given length. I denote by $\mathbf{X} \in \mathbb{R}^{3 \times n}$ the coordinate matrix of the 3D model, where n denotes the number of beads in the genome (for example, for *P. falciparum* at 20 kb resolution, $n = 1173$). I denote by $x_i \in \mathbb{R}^3$ the i -th bead's coordinates, and by d_{ij} the euclidean distance between bead i and j .

The contact map can be summarized by an n -by- n matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$, where each row and each column corresponds to a genomic locus, and each entry c_{ij} to the number of times loci i and j have been seen interacting. Note that the matrix is by construction square and symmetric (See Figure 3).

Figure 3: Contact maps of *P. falciparum*'s chr7.

Both contact maps show enrichment of contact counts in VRSSM clusters (A. Ring stage [19] of IT strains, B. Trophozoite stage of 3D7 [20]). Dark blue corresponds to high interactions, light yellow to low interactions. The strong diagonal reflects the proximity of adjacent regions of the genome. The black line indicates the centromeres.

2.2 Consensus models

Consensus methods aim at inferring a model $\mathbf{X} \in \mathbb{R}^{3 \times n}$ that best represents the contact count matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$, usually through the minimization of an objective function $\mathcal{O}(\mathbf{X}, \mathbf{C})$, sometimes under constraints.

$$\underset{\mathbf{X}}{\text{minimize}} \quad \mathcal{O}(\mathbf{X}, \mathbf{C}),$$

The objective function $\mathcal{O}$, sometimes called the “scoring function,” can be derived from known embedding algorithms (such as multidimensional-scaling methods), from statistical modeling of contact counts, or simply constructed in an *ad hoc* manner to fulfill a set of desirable properties.

2.2.1 Metric MDS-based methods

Early methods cast the 3D inference problem as a multidimensional scaling (MDS) problem: chromosomes are modeled as a chain of beads positioned in 3D such that the distance d_{ij} between bead i and bead j matches as closely as possible a “wish-distance” δ_{ij} derived from contact counts [13, 20, 27, 32]. Such problem can be formulated as follows:

$$\underset{\mathbf{X}}{\text{minimize}} \quad \sum_{(i,j) \in \mathcal{D}} \frac{1}{w_{ij}} (d_{ij} - \delta_{ij})^2, \quad (1)$$

where $\mathcal{D}$ the subset of indices to consider (typically interacting pairs of loci), d_{ij} is the euclidean distance between bead i and j , w_{ij} is a weight, assessing the confidence of the interaction between loci i and j

The first step is thus to find an adequate count-to-distance mapping to convert contact counts into pairwise wish-distances δ_{ij} . Dekker et al. [27] model the 78 pairwise contact counts of *S. cerevisiae*'s chromosome III as a Worm-like chain polymer, Duan et al. [13] use a linear mapping between contact counts and wish distances, Ay et al. [20] rely on the biophysical properties of fractal globule polymers, Tanizawa et al. [32] infer the count-to-distance mapping by fitting the relationship using known pairwise distances obtained through high-resolution FISH measures. In truth, there are as many ways to derive the count-to-distance mapping as there are studies, although many count-to-distance mappings rely on a power-law relationship between contact counts and wish distances: $c \sim \frac{1}{\delta^\alpha}$.

Lesne et al. [29], Hirata et al. [38], propose a different approach to convert counts into distances: they borrow concepts from graph theory to create a distance matrix (either using a shortest-path algorithm or the Dijkstra algorithm). A graph is constructed from the contact map by considering each locus as a node. Loci seen interacting are connected through an edge of weight $\frac{1}{c_{ij}}$. The authors then compute wish-distances δ_{ij} as the shortest path between node i and node j in the graph. Constructing wish-distances in such a way has two advantages: (i) low contact counts do not contribute much to the wish-distances; (ii) the resulting wish-distances form a distance matrix, and thus an optimal solution to the MDS can be found.

In addition to having subtle differences in the derivation of wish distances (which can have important effects on the resulting structures), the different methods also vary by the inclusion of weights to reflect confidence in "wish-distances" [20], or constraints to reflect prior knowledge on the structure [13, 20, 27] (see Table 1 for a summary of the characteristics of each method).

Another approach consists in handcrafting an *ad hoc* optimization function designed to fulfill a number of properties (adding penalization terms on adjacent beads, replacing the least squares with another more complicated form, ...) [33–35].

Table 1: Differences between MDS-based methods.

Paper	Organism	Full or partial	Counts-to-distance mapping	Constraints			Weights w_{ij}
				Adj.	rDNA	C	T
[27]	<i>S. cerevisiae</i>	Chr III	Worm-like chain behavior				
[13]	<i>S. cerevisiae</i>	Whole-genome	Linear relationship	✓	✓	✓	
[32]	<i>S. pombe</i>	Whole-genome	FISH-derived relationship	✓	✓	✓	✓
[20]	<i>P. falciparum</i>	Whole-genome	Fractal globule derived relationship	✓			δ_{ij}^2
[30]	<i>P. falciparum</i>	Whole-genome	Fractal globule derived relationship	✓			δ_{ij}^2
[29]		Whole-genome	<i>ad hoc</i>				

Many methods based on MDS uses slightly different counts-to-distance mapping, constraints and includes weights.

Table 2: A comparison of 3D inference methods.

Publication	Name	Consensus or Ensemble	MDS-based	Statistical model	Available
[27]		C	✓		
[13]		C	✓		✓
[32]		C	✓		
[20]		C	✓		
[26]		C	✓		✓
[23]	Pastis	C		✓	✓
[22]		E			
[15]		E			
[24]	chromSDE	C	✓		✓
[30]	autochrom3D	C	✓		✓
[31]		E		✓	✓
[28]	Bach	E/C		✓	✓
[36]		E			
[29]	ShRec3D	C	✓		✓
[33]		C			
[35]		C			
[34]	MOGEN	C			
[37]		E			

[42]	MBO	C	✓	✓
[38]	RPR	C	✓	

In this table, we summarize properties of published methods to infer the 3D structure of the genome: (1) is it a *consensus* or a *ensemble* based inference? (2) is it an MDS based method? (3) or does it rely on a statistical modeling? (4) is the software available or not (to the best of our knowledge)?

2.2.2 Non-metric MDS-based method

A crucial step of MDS-based methods is the conversion of counts into wish-distances. As described earlier, doing so requires strong assumptions that may not be met in practice. For example, this mapping changes from one organism to another [39], from one resolution to another [24], or even from one time point to another during the cell cycle [20, 40]. To alleviate this problem, after filtering interaction counts based on significance and interpolation of missing values, Ben-Elazar et al. [26] cast the inference as a *non-metric MDS* [41], where the 3D structure is inferred jointly with the wish-distances. Another idea is to parametrize the count-to-distance mapping as a power-law ($d = \beta c^\alpha$) and to infer parameters α and β jointly with the structure [23, 24].

2.2.3 Statistical models for contact counts

Another approach consists in modeling contact counts as random variables, with the 3D model being a latent variable. For example, Varoquaux et al. [23] propose an approach, called *Pastis*, that models contact counts as independent random Poisson variables where the intensity of the process between i and j is a function of the distance: $c_{ij} \sim \text{Poisson}(\beta d_{ij}^\alpha)$. The inference can thus be cast as maximizing the likelihood:

$$\max_{\alpha, \beta, \mathbf{X}} \mathcal{L}(\mathbf{X}, \alpha, \beta) = \sum_{i < j \leq n} c_{ij} \alpha \log d_{ij} + c_{ij} \log \beta - \beta d_{ij}^\alpha \quad (2)$$

Pastis can automatically adjust the parameters $\alpha\beta$ of the counts-to-distance transfer function and infer a genome structure that best explains the observed data. The strength of this method comes from the robustness to low signal-to-noise ratio, thanks to a direct modeling of the noise through the statistical model (see Figure 4).

Figure 4: 3D models of *P. falciparum*'s genome architecture.

Using *pastis*-mds [23] (that implements a weighted MDS model) (A and B) and *pastis*-PO [23] (the Poisson modeling presented section 2.2.3) (C and D) on two Hi-C data sets of *P. falciparum*: a low signal-to-noise ratio data set at a *Ring* stage [19] (A and C) and a high signal-to-noise ratio data set at the *Trophozoite* stage [20] (B and D). *pastis*-MDS recovers a plausible structure for the high quality data set, but the non-uniform distribution of beads on low quality data set suggests that the method is not adequate on such a data set. In contrast, *pastis*-PO recovers plausible structures both on the high and low quality data sets. Each color corresponds to a chromosome. Large white beads mark centromeres, small blue beads telomeres and green beads VRSM clusters. Centromeric cluster is marked with a black line, telomeric cluster is marked with a dashed line.

2.3 Ensemble methods as a means to infer population of structures

Ensemble approaches aim at inferring a population of structures representative of the contact count map. The methods fall into two distinct categories: the first type casts the problem as a restraint-based optimization and samples local minima of the function [15, 22], whereas the second type proposes a statistical modeling of the problem and samples the posterior distribution [28, 31]. In short, the former is the *ensemble* version of MDS-based and *ad hoc* methods, while the latter is the *ensemble* version of statistical based methods.

2.3.1 Sampling local minima

Umbarger et al. [15], Bau et al. [22], Kalhor et al. [36] model chromosomes as a series of beads, linked by restraining oscillators. These oscillators can be thought of as a “force” between beads so that they come into contact or ensure a minimal or maximal distance between those. The model includes two different types of restraints: (i) beads seen interacting are restrained with harmonic oscillators of strengths derived from the contact counts; (ii) adjacent beads are ensured to be neither too close nor too far from one another. This yields an optimization problem with a large number of local minima, which the authors sample from by running 50,000 minimizations starting from random initializations.

2.3.2 Estimating the posterior distribution of a statistical model

Hu et al. [28], Rousseau et al. [31] propose to model contact counts with a formal probabilistic model. Rousseau et al. [31] model observed contact counts c_{ij} as a Gaussian random variable of mean βd_{ij}^α , $\alpha \leq 0$ and variance σ_{ij} estimated directly from the contact count data, whereas Hu et al. [28] model contact counts as random Poisson variables of mean βd_{ij}^α . The authors then sample from the posterior using MCMC. Obtaining a *consensus* structure from such a method can be accomplished by selecting the maximum *a posteriori*.

2.4 Single-cell models

The last category of data-driven methods to infer the 3D architecture of the genome rely on a new protocol to probe single-cells for their 3D structures [37, 43]. Single-cell Hi-C is still in its early days, and, despite potential for assessing the variability of cell-to-cell genome architecture in a genome-wide fashion, only a handful of data sets are today publicly available. The contact maps originating from these data sets are very sparse, and specific methods to infer 3D structures need to be developed specifically for sc-HiC.

Akin to *ensemble*-local minima methods, the first approach is to consider each contact as a constraint and to formulate an under-constrained optimization problem. A population of structures satisfying the constraints can be found by sampling local minima [37]. Akin to *consensus* method, one can attempt to construct a distance matrix, either through manifold-based optimization [42] (by finding a low rank PSD approximation of the sparse contact map) or akin to Lesne et al. [29], by considering the weighted graph of interactions [38]. A classical MDS method applied to such a distance matrix then yields a *consensus* 3D model of the genome.

2.5 Model evaluation and comparison

A substantial difficulty in modeling the 3D structure of the genome is that model evaluation tends to be subjective. What is the relevant measure? “Truth” is generally not fully available, except for a few pairs of loci or in

simulations. Is validating the colocalization of a pair or a few pairs of loci via FISH experiments enough? Are fit to the contact maps or agreements between modeling techniques relevant? Are the conclusions drawn from 3D models in agreement when these are inferred from different methods?

First, methods can be validated and compared against contact maps simulated from a known ground truth [23, 24, 29, 35]. Note that while this is a simple and natural first step for methods inferring a *consensus* structure, comparing and assessing robustness and accuracy of an *ensemble* of 3D models is much more challenging: in fact, it is still an untackled problem. Second, one can assess the stability and robustness of the inference with respect to (1) data bootstrapping; (2) contact map resolution (the models should not change as the resolution of the data varies) [23, 24]; (3) and in between biological replicates [23, 24]. As a cautionary remark, it is important to stress that it is not because a method is very stable that it is “good” with respect to any criterion: I can imagine a number of very stable methods, that would yet provide absolutely no insights on genome organization. Third, models can be compared to other sources of data, such as FISH [13, 20]. Fourth, biological plausibility of the resulting models can be considered: are the beads uniformly distributed in the cell [20]? Are known hallmarks of the genome architecture such as centromeres clustering preserved?

These are a handful of ways to assess plausibility and accuracy of 3D reconstruction, but many avenues in model evaluation and comparison are yet to be explored.

3 The art of modeling genome architecture

“Data-driven” methods, as presented in the previous section, use the experimental contact maps to infer models as consistent as possible with the data. “Model-driven” methods tackle the 3D-modeling challenge exactly in the opposite manner: model in some way a population of structures, and validate this population using the contact map. Consider the following task. You have tens of thousands of randomly placed beads-on-a-string. Can you find the smallest sets of constraints such that these beads interact in overall the same way as a given contact map? This is the daunting task accomplished by “model-driven” approach: chromosomes are modeled as polymers (or random self-excluding fibers) under a small number constraints such that contact maps generated from these models match as closely as possible the observed contact maps. These “model-driven” approaches offer powerful mechanistical insights into the genome architecture, but are difficult to build in practice: each organism, cell type, and time point require hand crafted sets of constraints, built by iteratively improving models.

3.1 Building a yeast nucleus

The budding yeast *S. cerevisiae*’s 3D structure has been extensively studied, both through 3C-type studies [12, 13, 27] and through bio-imaging experiments [44]. The small size of its genome, the well-known hallmarks of its genome architecture and the availability of high resolution contact maps and FISH data set quickly led several teams to investigate the minimal set of constraints needed to reproduce the hallmarks of its genome architecture.

Tjong et al. [45] and Tokuda et al. [46] model *S. cerevisiae*’s chromosomes as flexible random fibers under a small set of constraints. While the exact modeling proposed by the three groups differ, the set of constraints can roughly be summarized as: (i) the chromosomes are constrained into a spherical ball representing the nucleus; (ii) centromeres are constrained into a spherical ball tethered to the nuclear membrane; (iii) telomeres are tethered to the nuclear membrane; (iv) rDNA is constrained into the nucleolus, represented as a spherical ball opposite to the centromeres, (v) a volume-exclusion constraint, preventing the fiber from occupying a space already occupied by the polymer. One can then simulate a large set of random structures fulfilling the constraints, and generate a “volume-exclusion contact map”, or “VE map”, from this population of structures, considering that two beads that are less than 45 nm apart in any of the structures form a contact. The Pearson correlation of the volume-exclusion contact map and the Hi-C one are highly correlated, demonstrating this small set of constraints fully explains the observed counts. In addition, the population of structures also explains FISH experiments previously published.

3.2 Building a *P. falciparum* nucleus?

For at least some stages, *P. falciparum* has a lot of genomic architectural features in common with the budding yeast: the centromeres are strongly co-localized at one end of the nucleus, telomeres are in physical contacts

with one another, ... Could these primary architectural features also arise from a population of constrained but otherwise random of structures? Adapting the set of constraints to match biological knowledge of *P. falciparum*, Ay et al. [20] showed that the resulting simulated VE map not only yielded lower correlations, but also failed to show the same features as the original contact count matrix. In particular, the VRSM genes which display domain-like enrichment in interactions do not appear in the simulated VE map (see Figure 5). Can we add constraints on VRSM gene clusters to improve correlations between true and generated contact maps? Adding a constraint on all beads considered as VRSM clusters is not enough: running 100 experiments yielded structures that did not fulfill all the constraints, demonstrating VRSM genes cannot all cluster together in a cell.

Figure 5: Volume Exclusion Modeling. Observed/expected matrices illustrate either depletion (blue) or enrichment (red) in contact counts for each pair of loci. **A.** Observed/expected map for volume exclusion modeling of *P. falciparum* **B.** Observed/expected map for Hi-C data.

4 Downstream analysis using 3D models: a highlight of the study of *P. falciparum*’s 3D structure

In section 2, I have reviewed data-driven methods to infer either *consensus* or *ensemble* models of the 3D structure. But why go through the effort to obtain such models and not directly study the contact maps? In this section, I will review a number of downstream analyses one can perform on 3D models, highlighting, but not limiting myself to, results on the 3D structure of *P. falciparum*. See Table 3 for a list of available *P. falciparum* Hi-C datasets. Note that while the results presented on *P. falciparum* have been obtained using *consensus* models, the methods presented here can be applied both on models obtained through *consensus* or *ensemble* approaches.

Table 3: Summary of the available *P. falciparum* Hi-C datasets.

Name	Strain	Stage	Resolution	Number of contacts	Perc of cis	Perc of trans	Reference
<i>Ay-rings</i>	3D7	Late Rings	10 kb	16711552	43%	57%	[20]
<i>Ay-trophozoites</i>	3D7	Trophozoites	10 kb	56348498	53%	47%	[20]
<i>Ay-schizonts</i>	3D7	Schizonts	10 kb	11652832	55%	45 %	[20]
Lemieux-A4+	IT/BC6+	Rings	25 kb	18488252	19%	81%	[19]
Lemieux-A4	IT/3G8	Rings	25 kb	19674672	28%	72%	[19]
Lemieux-A44	IT/BC6-	Rings	25 kb	18660594	25%	75%	[19]
Lemieux-DCJ_On	NF54/DCJ on	Rings	25 kb	3098370	26%	74%	[19]
Lemieux-DCJ_Off	NF54/DCF Off	Rings	25 kb	2533470	26%	73%	[19]
Lemieux-B15C2	NF54/B15C2	Rings	25 kb	1022996	12%	88%	[19]

4.1 Structure stability across time points, clustering and other variance analysis

The reader may well ask: “how sensitive are the resulting 3D models to initialization?” Taking the case of the *P. falciparum*, one may wonder whether structures from the same time points but from a different initialization are more alike than structures from different time points. A natural way to answer this question is to perform some dimensionality reduction technique, such as PCA and visualize whether structures from the same time points

cluster with one another. Typically, features would then be the pairwise Euclidean distance of each structure, possibly subsampled to ease computation. Performing such an experiment on 1000 *consensus* structures inferred using the statistical model proposed by Varoquaux et al. [23], and available in the package *pastis* as *pastis-PO*, demonstrates that the results are more stable across initialization than time points (see Figure 6).

Figure 6: Stability of structures across the life cycle.

PCA analysis of the population of structures obtained by running 1000 *Pastis-PO* on the three data sets of Ay et al. [20], corresponding to the *Ring* (Ay-Rin), *Schizont* (Ay-Sch), and *Trophozoite* (Ay-Trop) stages and the B15C2 data set of Lemieux et al. [19], at the ring stage (Lemieux-Rin) demonstrates that structures are more stable across initialization than across time points. Note that the *Ring* stages of Ay et al. [20] and Lemieux et al. [19] do not cluster, reflecting the centromeres strong colocalization in one of the data sets and not the other.

The second question is: “are models locally consistent?” A possible answer to this questions is to divide the structures into overlapping sub-structures ranging from 5 to 20 beads, and to compute the pairwise root mean squared deviation between segments across all structures [22]. Segments overlapping within a certain range for a large number of models can be assessed as locally consistent, while others should be labeled as highly variable.

The last question one can ask is: “how do the structures differ?” Tackling this question is very challenging, but can be reformulated to “are the hallmarks of interest conserved across structures of the same stage?” Ay et al. [20] and Lemieux et al. [19] both identified *P. falciparum* folded in very specific ways, with VRSM genes highly interacting. Ay et al. [20] also observed strong clustering of the centromeres, and enrichment in interaction at the telomeres. These observations can lead to a rigorous approach to identifying and quantifying whether families of loci cluster in the structures.

4.2 Chromatin compaction and chromosome entanglement

3D models can be used to estimate chromatin compaction and chromosome entanglement: distinguishing between open and close chromatin allows to relate the models to gene expression, open region being more accessible to the transcription machinery and thus genes more likely to be expressed. Chromatin compaction can be estimated by looking at the number of base pairs in a region defined either by volume if the scale of the structure is known, or by a percentage of the size of the structure if it is not [22]. Another idea is to sample random beads of a certain diameter, and assess how many loci are seen interacting. Applying this latter method on the three models of *P. falciparum*, Ay et al. [20] show that the trophozoite stage exhibits a more open chromatin than the ring and schizont stages. This finding is consistent with the transcriptionally active state of the parasite during this moment of the life cycle. A similar analysis, but counting the number of inter-chromosomal interactions, can help to assess the chromosome entanglement of the structure.

4.3 3D gene set enrichment

To assess whether groups of genes are colocalized in a 3D model, Ay et al. [20] leverage a statistical method developed by [47], which requires labeling each pair of loci in two groups: “close” or “far.” The authors used varying distance thresholds (10%, 20% and 40% of the nuclear diameter) to deem a locus pair “close” and labeled all remaining pairs in the set as “far.” The authors then compared the enrichment of loci pairs of a group being “close” and “far” by resampling loci among a same chromosome.

This approach dichotomizes loci pairs into two groups, and checks for the enrichment of a label in one of the two groups. Capurso and Segal [48] present an approach, called MPED, that avoids this step, and instead

directly estimates the significance within the 3D model. Briefly, for a group $\mathcal{G}$, MPED computes a test statistic:

$$M = \text{median}_{i,j \in \mathcal{G} | c_i \neq c_j} d_{ij},$$

where d_{ij} is the Euclidean distance between bead i and bead j . The null distribution is estimated empirically by resampling 10^5 times with preservation of the chromosome structure. If the M statistic is smaller than the mean of the null distribution, it is compared to the lower tail of the distribution and indicates co-localization. If the M statistic is larger than the mean of the null distribution, then it is compared to the upper tail of the distribution, and indicates dispersion.

Applying MEPD to the Trophozoite stage, Capurso and Segal [48] confirm that centromeres, telomeres, VRSM genes (both overall, subtelomeric and internal) colocalize.

4.4 Integrative analysis of gene expression and 3D structure using KernelCCA

Last but not least, an exciting contribution of Ay et al. [20] is the integrative analysis of gene expression and 3D structure using an unsupervised learning technique called “kernel Canonical Correlation Analysis” (kCCA) [49]. The goal of this analysis is to explore the relationship between gene expression and 3D structure, by extracting a set of gene expression components that exhibit coherence with respect to the 3D structure. While the components aren’t necessarily an actual gene expression profile, the genes of interest must somehow either be highly positively or negatively correlated with a component, and those genes should exhibit some form of coherence in terms of their 3D structure, like co-location. It can be helpful to think of this procedure as performing a principal component analysis (PCA) gene expression components extracted are correlated with the 3D structure.

Let us take a closer look at how kCCA is formally used in this context. Consider the set of n genes $g \in \mathcal{G}$. Each gene g is represented on the one hand by its gene expression profile $e(g) \in \mathbb{R}^p$ and on the other hand by its 3D position $x(g) \in \mathbb{R}^3$. Assume the set of gene expression profiles is mean-centered and of unit variance.

The goal is to extract a gene expression component $v \in \mathbb{R}^p, \|v\| = 1$ such that it is both representative of set of gene expression profiles but also correlated with the 3D structure.

First, let’s tackle the question of representativeness of the set of gene expression profiles. We can identify a component v with such properties as to compute the percentage of variance explained by this component as follows:

$$V(v) = \sum_{g \in \mathcal{G}} (v^T e(g))^2. \quad (3)$$

Note that maximizing $V(v)$ results in finding the first principal component v of the PCA. We can then define a score $s \in \mathbb{R}^n$ for each gene by computing the projection of each gene expression profile onto the component: $s(g) = v^T e(g)$. Any gene important to component v will be highly negatively or positively correlated with that component and thus either have a strongly negative or positive score $s(g)$.

Now, let’s turn to the question of assessing coherence with respect to the 3D structure. Given a vector of scores $f \in \mathbb{R}^n$, how can we assess the smoothness of these scores along the 3D structure? Ay et al. [20] leverage a standard approach in kernel methods, in which the smoothness of a score f is quantified by the function:

$$S(f) = \frac{f^T K_{3D}^{-1} f}{\|f\|}, \quad (4)$$

where K_{3D} is the Gaussian kernel matrix of the genes’ 3D coordinates. The smaller $S(f)$ is, and the more smoothly f is distributed in 3D.

So far, the two measures of representativeness and smoothness are independent from one another. However, if the scores s and f are required to be as correlated as possible, any genes highly correlated with each gene component v will also be co-localized in space.

To solve this problem, a common trick is to leverage reproducible kernel Hilbert space (RKHS) theory and cast the optimization in dual form. First, it can be shown that any candidate component v can be written as a linear combination of the gene expression profile: $v = \sum_{g \in \mathcal{G}} \alpha(g) e(g)$: α is called the dual coordinate of v . Let $K_g \in \mathbb{R}^{n \times n}$ be the gram matrix of the gene expression profile, obtained by computing the inner product between all expression profiles: $K_g(x, y) = e(x)^T e(y)$. We can thus rewrite equation 3 as:

$$V(s) = \frac{\alpha^T K_g^2 \alpha}{\alpha^T K_g \alpha} \quad (5)$$

As K_{3D} is invertible of dimension $n \times n$, any score f can be written as $f = \sum_{g \in \mathcal{G}} K_{3D} \beta$, and the measure of smoothness as:

$$S(f) = \frac{\beta^T K_{3D} \beta}{\beta^T K_{3D}^2 \beta} \quad (6)$$

The optimization problem can now be cast in dual form as follows:

$$\max_{\alpha, \beta} (\text{corr}(s, f)) = \frac{\alpha^T K_g K_{3D} \beta}{(\alpha^T (K_g^2 + \lambda K_g) \alpha)^{\frac{1}{2}} (\beta^T (K_{3D}^2 + \lambda K_{3D}) \beta)^{\frac{1}{2}}}, \quad (7)$$

where λ is a penalization parameter. Solving this optimization problem identifies s and f such that; (1) the two scores are correlated with one another; (2) s maximizes $V(s)$; and (3) f minimizes $S(f)$. These optimization problems can then be solved efficiently using a generalized eigenvalue decomposition.

Ay et al. [20] apply this method using gene expression profiles and genes' location extracted from the 3D models of the genome architecture and find gene expression profiles highly correlated with the 3D structure. Ranking the genes with their projections onto the gene components, Ay et al. [20] demonstrate that several gene families and Gene Ontology (GO) terms are enriched both close to the telomeres and at the opposite end of the nucleus. This method could easily be extended to other data set such as histone modifications, ATAC-seq, and beyond.

5 Discussion

A plethora of methods for inferring 3D models from contact maps has been developed: some are very specific to an organism or a region of the genome, while others are generalizable easily to many different organisms. As 3C-type methods are democratized, a wider audience of researchers are in need of robust and well-implemented algorithms for building models. Surprisingly, only a handful of methods are publicly available, generalizable, and well-validated. While *ensemble* methods offer a more biologically accurate view of the variety of chromatin folding in a population of cells, their use is more challenging, both as a result of a lack of comparison and validation of the methods, but also as a large body of structures is less amenable to exploration and visualization than *consensus* structures. Yet *consensus* structures need to be interpreted with care, as they best represent a mean of structures and are very unlikely to be a true representation of the chromatin folding.

At the other end of the spectrum, model-driven methods are organism and study specific, and require good understanding of both polymer physics and of the particular organism studied: for each data set, the set of constraints to apply on the structure needs to be revisited. Yet, once built, they provide extensive insights in the average location of genes and may constitute a way to replace high throughput FISH experiments at low cost as a first exploration tool.

While downstream analysis of the *P. falciparum* gave important insights in the relation between gene regulation and genome architecture of the *P. falciparum*, the use of 3D models for downstream analysis remains scarce in the literature and many avenues are left opened to methodological development. For instance, a meaningful comparison of structures from different time points at a genome-scale level is still an open problem.

Code and data availability

All the figures of this paper can be reproduced using the code and instructions at <https://github.com/NelleV/takefive>.

Acknowledgements

I would like to thank R. Barter, C. Holdgraf, D. Morozov and A. Paxton for their feedback on the article.

Funding

This work was supported by the Gordon and Betty Moore Foundation (Grant GBMF3834) and the Alfred P. Sloan Foundation (Grant 2013-10-27) and used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562.

Competing interests: None declared.

Glossary

3C or Chromosome conformation capture Experiment to quantify the number of interactions between a pair of loci. The technology is based on cross-linking the DNA with formaldehyde to “freeze” interactions, digesting the DNA with a restriction enzyme to cut into small fragments, a ligation step favoring ligation of cross-linked DNA, followed by reverse crosslinking. Ligated fragments are then detected using PCR with known primers.

4C or Chromosome conformation capture-on-chip Experiment to quantify the number of interactions between a locus and all the other loci. 4C experiments typically use the same procedure as 3C experiments, with an additional ligation step and inverse PCR. The inverse PCR step allows to amplify the locus of interest as well as the unknown sequences ligated to it.

5C or Chromosome conformation capture carbon copy Experiment to quantify the number of interactions between all loci in a given region, typically less than 1Mb long. The steps to perform a 5C experiments are similar to a 3C experiments, but uses many known primers to ligate to all the fragments in order to identify the loci of interests.

Consensus method Method that aims at inferring a unique mean structure.

Contact count The number of times two genomic windows have been seen interacting in a Hi-C or 3C experiment.

Contact map or contact count matrix A map or a matrix where each row and column corresponds to a genomic loci and each entry to the number of times these two regions have been seen interacting with one another.

Count-to-distance mapping or count-to-distance function A function that takes in input a contact count and returns a wish-distance. The function is often derived from relationships between expected contact counts and euclidean distances, obtained from polymer physics.

Data-driven method Method that uses experimental data to infer 3D models, typically by minimizing a cost function.

Ensemble method Method that aims at inferring a population of structures.

Fluorescence *In Situ* hybridization (FISH) Bio-imaging technique used to localize specific DNA sequences. It uses fluorescent probes that bind to parts of the chromosomes with very high degree of sequence similarity.

Fractal globule polymer A polymer that folds by creating crumpled globules, folded in a hierarchical fashion. This polymer has been proposed as a model for DNA.

Hi-C Experiment to quantify the number of interactions between pairs of loci, in a genome-wide manner. A Hi-C experiment uses the same step as a 3C experiment (crosslinking, digestion, ligation, reverse crosslinking), but identifies the interaction through high-throughput sequencing, hence consider all possible interacting pairs.

Markov chain Monte Carlo (MCMC) Class of algorithms used to sample from a probability distribution.

Model-based method Method that considers the polymer nature of DNA to build, with as few constraints and assumptions as possible, many chromosome conformations.

Multidimensional scaling (MDS) Dimensionality reduction techniques, that aim at placing object in such a way that the distances between each object is preserved as much as possible.

Var genes Family of roughly 60 genes used by the Plasmodium parasite to interact with the human host.

Volume-exclusion (VE) models Models simulated from a constrained flexible random polymer model, with volume-exclusion constraints.

Wish-distance A “wish” distance derived from a contact count, usually using a count-to-distance function estimated from polymer physics.

References

- [1] Onwujekwe O, e. I. F. Malik, S. H. Mustafa, and A. Mnzavaa Do malaria preventive interventions reach the poor? Socioeconomic inequities in expenditure on and use of mosquito control tools in Sudan. *Health Policy Plan.* 2006;21:10–16.
- [2] Kirchner S, Power BJ, Waters AP. Recent advances in malaria genomics and epigenomics. *Genome Med.* 2016;8:92.
- [3] Cui L, Miao J. Chromatin-mediated epigenetic regulation in the malaria parasite *Plasmodium falciparum*. *Eukaryotic Cell.* 2010;9:1138–1149.
- [4] Deutsch K, Duraisingh M, Dzikiowski R, Gunasekera A, Khan S, Le Roch K, Llinas M, Mair G, McGovern V, Roos D, Shock J, Sims J, Wiegand R, Winzeler E. Mechanisms of gene regulation in *Plasmodium*. *Am J Trop Med Hyg.* 2007;77:201–208.
- [5] Duffy MF, Selvarajah SA, Josling GA, Petter M. The role of chromatin in *Plasmodium* gene expression. *Cell Microbiol.* 2012;14:819–828.
- [6] Hoeijmakers WA, Stunnenberg HG, Bartfai R. Placing the *Plasmodium falciparum* epigenome on the map. *Trends Parasitol.* 2012;28:486–495.
- [7] Horrocks P, Wong E, Russell K, Emes RD. Control of gene expression in *Plasmodium falciparum* - ten years on. *Mol Biochem Parasitol.* 2009;164:9–25.
- [8] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.* 2009;326:289–293.
- [9] De S, Michor F. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat Biotechnol.* 2011;29:1103–1108.
- [10] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature.* 2012;485:376–380.
- [11] Rao SS, Huntley MH, Durand N, Neva C, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin v looping. *Cell.* 2014;59:1665–1680.
- [12] Burton JN, Liachko I, Dunham MJ, Shendure J. “Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3 (Bethesda).* 2014;4:1339–1346.
- [13] Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS. A three-dimensional model of the yeast genome. *Nature.* 2010;465:363–367.
- [14] Mizuguchi T, Fudenberg G, Mehta S, Belton J-M, Taneja N, Folco HD, FitzGerald P, Dekker J, Mirny L, Barrowman J, Grewal SI. “Cohesin-dependent globules and heterochromatin shape 3d genome architecture in *S. pombe*. *Nature.* 2014;516:432–435.
- [15] Umbarger MA, Toro E, Wright MA, Porreca GJ, Bau D, Hong S, Fero MJ, Zhu LJ, Marti-Renom MA, McAdams HH, Shapiro L, Dekker J, Church GM. The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Molecular Cell.* 2011;44:252–264.
- [16] Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell.* 2012;148:458–472.
- [17] Feng S, Cokus SJ, Schubert V, Zhai J, Pellegrini M, Jacobsen SE. Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in *Arabidopsis*. *Mol Cell.* 2014;55:694–707.
- [18] Wang C, Liu C, Roqueiro D, Grimm D, Schwab R, Becker C, Lanz C, Weigel D. Genome-wide analysis of local chromatin packing in *Arabidopsis thaliana*. *Genome Research.* 2015;25:246–256.
- [19] Lemieux JE, Kyes SA, Otto TD, Feller AI, Eastman RT, Pinches RA, Berriman M, Su XZ, Newbold CI. Genome-wide profiling of chromosome interactions in *Plasmodium falciparum* characterizes nuclear architecture and reconfigurations associated with antigenic variation. *Mol Microbiol.* 2013;90:519–537.
- [20] Ay F, Bunnik EM, Varoquaux N, Bol SM, Prudhomme J, Vert J-P, Noble WS, Le Roch KG. Three-dimensional modeling of the *P. falciparum* genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Res.* 2014;24:974–988.
- [21] Ay F, Bunnik EM, Varoquaux N, Vert J-P, Noble WS, Le Roch KG. Multiple dimensions of epigenetic gene regulation in the malaria parasite *Plasmodium falciparum*. *Bioessays.* 2015;37:182–194.
- [22] Bau D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom MA. The three-dimensional folding of the -globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol.* 2011;18:107–114.
- [23] Varoquaux N, Ay F, Noble WS, Vert J-P. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics.* 2014;30:i26–i33.
- [24] Zhang Z, Li G, Toh K-C, Sung W-K. Inference of spatial organizations of chromosomes using semi-definite embedding approach and Hi-C data. In: Proceedings of the 17th International Conference on Research in Computational Molecular Biology. *Lecture Notes in Computer Science*, volume 7821, Lecture Notes in Computer Science. Berlin, Heidelberg: Springer-Verlag, 2013:317–332.
- [25] Deng X, Ma W, Ramani V, Hill A, Yang F, Ay F, Berletch JB, Blau CA, Shendure CA, Duan Z, Noble WS, Disteche CM. Bipartite structure of the inactive mouse X chromosome. *Genome Biol.* 2015;16:152.
- [26] Ben-Elazar S, Yakhini Z, Yanai I. Spatial localization of co-regulated genes exceeds genomic gene clustering in the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res.* 2013;41:2191–2201.
- [27] Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science.* 2002;295:1306–1311.
- [28] Hu M, Deng K, Qin Z, Dixon J, Selvaraj S, Fang J, Ren B, Liu JS. Bayesian inference of spatial organizations of chromosomes.” *PLoS Comput Biol.* 2013;9:e1002893.
- [29] Lesne A, Riposo J, Roger P, Cournac A, Mozziconacci J. 3D genome reconstruction from chromosomal contacts. *Nature Methods.* 2014;11:1141–1143.
- [30] Peng C, Fu L-Y, Dong P-F, Deng Z-L, Li J-X, Wang X-T, Zhang H-Y. The sequencing bias relaxed characteristics of Hi-C derived data and implications for chromatin 3D modeling. *Nucleic Acids Res.* 2013;41:e183.

- [31] Rousseau M, Fraser J, Ferraiuolo M, Dostie J, Blanchette M. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinformatics*. 2011;12:414.
- [32] Tanizawa H, Iwasaki O, Tanaka A, Capizzi JR, Wickramasignhe P, Lee M, Fu Z, Noma K. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res*. 2010;38:8164–8177.
- [33] Trieu T, Cheng J. Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. *Nucleic Acids Res*. 2014;42:e52.
- [34] Trieu T, Cheng J. MOGEN: a tool for reconstructing 3D models of genomes from chromosomal conformation capturing data. *Bioinformatics*. 2016;32, 1286–1292.
- [35] Trieu T, Cheng J. 3D genome structure modeling by Lorentzian objective function. *Nucleic Acids Res*. 2017;45:1049–1058.
- [36] Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol*. 2011;30:90–98.
- [37] Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*. 2013;502:59–64.
- [38] Hirata Y, Oda A, Ohta K, Aihara K. Three-dimensional reconstruction of single-cell chromosome structure using recurrence plots. *Sci Rep*. 2016;6:34982.
- [39] Fudenberg G, Mirny LA. Higher-order chromatin structure: bridging physics and biology. *Curr Opin Genet Dev*. 2012;22:115–124.
- [40] Le TB, Imakaev MV, Mirny LA, Laub MT. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science*. 2013;342:731–734.
- [41] Kruskal J. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*. 1964;29:1–27.
- [42] Paulsen J, Gramstad O, Collas P. Manifold based optimization for single-cell 3d genome reconstruction. *PLoS Comput Biol*. 2015;11:e1004396. <http://dx.doi.org/10.1371/journal.pcbi.1004396>.
- [43] Ramani V, Deng X, Qiu R, Gunderson KL, Steemers FJ, Disteche CM, Noble WS, Duan Z, Shendure J. Massively multiplex single-cell Hi-C. *Nat Methods*. 2017;14:263–266.
- [44] Berger AB, Cabal GC, Fabre E, Duong T, Buc H, Nehrbass U, Olivo-Marin J-C, Gadal O, Zimmer C. High-resolution statistical mapping reveals gene territories in live yeast. *Nat Methods*. 2008;5:1031–1037.
- [45] Tjong H, Gong K, Chen L, Alber F. Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome Res*. 2012;22:1295–1305.
- [46] Tokuda N, Terada TP, Sasai M. Dynamical modeling of three-dimensional genome organization in interphase budding yeast. *Biophys J*. 2012;102:296–304.
- [47] Witten DM, Noble WS. On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. 2012;40:3849–3855.
- [48] Capurso D, Segal MR. Distance-based assessment of the localization of functional annotations in 3D genome reconstructions. *BMC Genomics*. 2014;15:992.
- [49] Bach FR, Jordan MI. Kernel independent component analysis. *J Mach Learn Res*. 2002;3:1–48.