bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PRINCIPAL CURVE APPROACHES FOR INFERRING 3D CHROMATIN ARCHITECTURE

A PREPRINT

Elena Tuzhilina
Department of Statistics
Stanford University
Stanford, CA 94305
elenatuz@stanford.edu

Trevor J. Hastie
Department of Statistics
Stanford University
Stanford, CA 94305
hastie@stanford.edu

Mark R. Segal [*]
Department of Epidemiology
and Biostatistics
University of California
San Francisco, CA 94143
mark.segal@ucsf.edu

June 15, 2020

ABSTRACT

Three dimensional (3D) genome spatial organization is critical for numerous cellular processes, including transcription, while certain conformation-driven structural alterations are frequently oncogenic. Genome architecture had been notoriously difficult to elucidate, but the advent of the suite of chromatin conformation capture assays, notably Hi-C, has transformed understanding of chromatin structure and provided downstream biological insights. Although many findings have flowed from direct analysis of the pairwise proximity data produced by these assays, there is added value in generating corresponding 3D reconstructions deriving from superposing genomic features on the reconstruction. Accordingly, many methods for inferring 3D architecture from proximity d hyperrefata have been advanced. However, none of these approaches exploit the fact that single chromosome solutions constitute a one dimensional (1D) curve in 3D. Rather, this aspect has either been addressed by imposition of constraints, which is both computationally burdensome and cell type specific, or ignored with contiguity imposed after the fact. Here we target finding a 1D curve by extending principal curve methodology to the metric scaling problem. We illustrate how this approach yields a sequence of candidate solutions, indexed by an underlying smoothness or degrees-of-freedom parameter, and propose methods for selection from this sequence. We apply the methodology to Hi-C data obtained on IMR90 cells and so are positioned to evaluate reconstruction accuracy by referencing orthogonal imaging data. The results indicate the utility and reproducibility of our principal curve approach in the face of underlying structural variation.

***K*eywords** 3D structure · Genome reconstruction · Hi-C assay · Metric scaling · Multiplex FISH.

[*] To whom correspondence should be addressed.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 Introduction

The three-dimensional (3D) configuration of chromosomes within the eukaryote nucleus is important for several cellular functions, including gene expression regulation, and has also been linked to translocation events and cancer driving gene fusions (Mitelman et al., 2007). While direct visualization of 3D architecture has improved (see Section 2.10), imaging challenges pertaining to chromatin compaction and dynamics persist. However, the ability to *infer* chromatin architectures at increasing resolution has been enabled by chromosome conformation capture (3C) assays (Dekker et al., 2002). In particular, when coupled with next generation sequencing, such Hi-C methods (Lieberman-Aiden et al., 2009, Duan et al., 2010) yield an inventory of pairwise, genome-wide chromatin interactions, or contacts. In turn, the contact data form the basis for *reconstructing* 3D configurations (Zhang et al., 2013, Varoquaux et al., 2014, Ay et al., 2014, Zou et al., 2016, Rieber and Mahony, 2017). While many novel conformational-related findings have flowed from direct analysis of contact level data, added value of performing downstream analysis based on attendant 3D reconstructions has been demonstrated. These benefits derive from the ability to superpose genomic features on the reconstruction. Examples include co-localization of genomic landmarks such as early replication origins in yeast (Witten and Noble, 2012, Capurso and Segal, 2014), gene expression gradients in relation to telomeric distance and co-localization of virulence genes in the malaria parasite (Ay et al., 2014), the impact of spatial organization on double strand break repair (Lee et al., 2016), and elucidation of '3D hotspots' corresponding to (say) overlaid ChIP-Seq transcription factor extremes which can reveal novel regulatory interactions (Capurso et al., 2016).

The contact or interaction matrices resulting from Hi-C assays, which are typically performed on bulk cell populations, are depicted as heatmaps, which record the frequency with which pairs of binned genomic loci are cross-linked, reflecting spatial proximity of the respective loci bins within the nucleus. A common first step toward 3D reconstruction is the conversion of contact frequencies into *distances*, typically assuming inverse power-law relationships (Varoquaux et al., 2014, Ay et al., 2014, Shavit et al., 2014, Rieber and Mahony, 2017), from which 3D chromatin architecture can be obtained via versions of the multi-dimensional scaling (MDS) paradigm. In response to (i) the bulk cell population underpinnings of contact data, (ii) computational challenges posed by the dimensionality of the MDS reconstruction problem as governed by bin extent, and (iii) accommodating biological considerations, several competing reconstruction algorithms have been advanced. However, none of these take advantage of the fact that the 3D solution for individual chromosomes corresponds to a one-dimensional (1D) curve in 3-space. Rather, this aspect has been addressed by imposition of constraints (Duan et al., 2010, Ay et al., 2014, Stevens et al., 2017), which are cell type specific and require prescription of constraint parameters. These parameters can be difficult to specify and their inclusion substantially increases the computational burden. Other approaches (Zhang et al., 2013, Park and Lin, 2017, Rieber and Mahony, 2017) do not formally incorporate contiguity but impose it post hoc, creating chromatin reconstructions by "connecting the dots" of the 3D solution according to the ordering of corresponding genomic bins.

Here we directly target chromosome reconstruction by finding a 1D curve approximation to the contact matrix via extending principal curve methodology (Hastie and Stuetzle, 1989) to the metric scaling problem. After reviewing problem formulation and current reconstruction techniques in Section 2.1 we develop three approaches (i) Principal Curve Metric Scaling (PCMS; Section 2.2–2.3), (ii) a weighted generalization of PCMS (WPCMS; Section 2.4–2.5), and (iii) Poisson Metric Scaling (PoisMS; Section 2.7–2.8). Strategies for selecting a specific reconstruction from a degrees-of-freedom indexed series of solutions are described in Section 2.9. Methods for appraising the accuracy of candidate reconstructions by referencing to orthogonal imaging data are outlined in Section 2.10. Results from applying the methodology to Hi-C data from IMR90 cells are presented in Section 3, while the Discussion indicates directions for future work.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2 Methods

2.1 Existing approaches to 3D chromatin reconstruction from Hi-C assays

Our focus is on reconstruction of *individual* chromosomes; whole genome architecture can follow by appropriately positioning these solutions (Segal and Bengtsson, 2015, Rieber and Mahony, 2017). As is standard, we disregard complexities deriving from chromosome pairing arising in diploid cells (which can be disentangled at high resolutions (Rao et al., 2014)) and defer issues surrounding inter-cell variation to the Discussion.

The result of a Hi-C experiment, following important preprocessing and normalization steps (Imakaev et al., 2012), is the *contact map*, a symmetric matrix $C = [C_{ij}] \in \mathbb{Z}_{+}^{n\times n}$ of contact counts between n (binned) genomic loci i, j on a genome-wide basis (see Figure 7 for an example of a contact matrix). This matrix can be exceedingly sparse even after binning. The 3D chromatin reconstruction problem is to use the contact matrix C to obtain a 3D point configuration $x_1, \ldots, x_n \in \mathbb{R}^3$ corresponding to the spatial coordinates of loci $1, \ldots, n$ respectively.

Many approaches have been proposed to tackle this problem with broad distinction between optimization and model-based methods (Varoquaux et al., 2014, Rieber and Mahony, 2017). A common first step is conversion of the contact matrix into a distance matrix $D = [D_{ij}]$ (Duan et al., 2010, Varoquaux et al., 2014, Ay et al., 2014, Shavit et al., 2014), followed by solving the *multi-dimensional scaling* (MDS; Hastie et al., 2009) problem: position points (corresponding to genomic loci) in 3D so that the resultant interpoint distances best conform to the distance matrix.

A variety of methods have also been used for transforming frequencies into distances. At one extreme, in terms of imposing biological assumptions, are methods that relate observed intra-chromosomal contacts to genomic distances and then ascribe *physical* distances based on organism specific findings on chromatin packing (Duan et al., 2010) or relationships between genomic and physical distances for crumpled polymers (Ay et al., 2014). Such distances inform the subsequent optimization step as they permit incorporation of known biological constraints that can be expressed in terms of physical separation. Importantly, these constraints include prescriptions on the 3D separation between contiguous genomic bins. It is by this means that obtaining a 1D curve is indirectly facilitated. However, obtaining physical distances requires both strong assumptions and organism specific data (Fudenberg and Mirny, 2012). More broadly, a number of approaches (Zhang et al., 2013, Varoquaux et al., 2014, Zou et al., 2016, Rieber and Mahony, 2017) utilize power law transfer functions to map contacts to (non-physical) distances $D_{ij} = \begin{cases} (C_{ij})^{-\alpha} \text{ if } C_{ij} > 0, \\ \infty \text{ if } C_{ij} = 0. \end{cases}$ Adoption of the power law derives from empirical and theoretical work but again constitutes a strong assumption (Fudenberg and Mirny, 2012).

Once we have a distance matrix, D, optimization approaches seek a 3D configuration $x_1, \ldots, x_n$ that best fits D according to an MDS criterion. If $\|\cdot\|$ designates the Euclidean norm, then an example of MDS loss incorporating weights and penalty (Zhang et al., 2013) is

$$\ell(x_1, \ldots, x_n) = \sum_{\{i,j|D_{ij}<\infty\}} W_{ij}(\|x_i - x_j\| - D_{ij})^2 - \lambda \sum_{\{i,j|D_{ij}=\infty\}} \|x_i - x_j\|^2 \tag{1}$$

with the corresponding optimization problem

$$\text{minimize } \ell(x_1, \ldots, x_n) \text{ w.r.t. } x_1, \ldots, x_n \in \mathbb{R}^3. \tag{2}$$

Here common choices for the weights W_{ij} include D_{ij}^{-1} (Zhang et al., 2013) and D_{ij}^{-2} (Varoquaux et al., 2014), these being analogous to precision weighting since large C_{ij} (small D_{ij}) are more accurately measured. Similarly, the

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

penalty (second) term maximizes the pairwise distances for loci bins with $C_{ij} = 0$ under the presumption that such loci should not be too close.

It is worth noting that (1), and related criteria, correspond to a nonconvex, nonlinear optimization problem that is NP hard and while various devices have been employed to mitigate the computational burden (e.g., Zhang et al., 2013), computational concerns, particularly for high resolution (many loci bins) problems, remain forefront.

Probabilistic methods postulate models for the coontact counts C_{ij}. The optimization goal is to maximize the corresponding log-likelihood

$$\ell(x_1, \dots, x_n) = \sum_{i=1}^{n-1} \sum_{j=i}^{n} \log \mathbf{P}(C_{ij} | x_1, \dots, x_n). \tag{3}$$

Poisson models $C_{ij} \sim Pois(\lambda_{ij})$ are widely adopted (Varoquaux et al., 2014, Zou et al., 2016, Park and Lin, 2017), where $\lambda_{ij} = \lambda_{ij}(x_1, \dots, x_n)$ is some function depending on genomic loci spatial coordinates $x_1, \dots, x_n$. For example, Rosenthal et al. [2019] prescribe exponential dependence between the Poisson rate parameter and interpoint distances: $\lambda_{ij} = \beta \|x_i - x_j\|^{\alpha}$. Although probabilistic methods are natural for modeling biological phenomena, here they incur computational costs and impose additional assumptions.

All existing approaches implicitly represent chromatin as a polygonal chain. Constraints on the geometrical structure of the polygonal chain can be imposed via penalties on edge lengths and angles between successive edges, with even quaternion-based formulations employed (Caudai et al., 2015). Rosenthal et al. [2019] utilize penalties to control smoothness of the resulting conformations. However, despite imparting targeted properties to the resulting reconstruction, such penalty-based approaches increase the complexity of the objective, its gradient and Hessian, both slowing and limiting, especially with respect to resolution, associated algorithms.

Here we develop a suite of novel approaches that directly model chromatin configuration as a 1D curve in 3D. As a baseline method we introduce *Principal Curve Metric Scaling* (PCMS). The PCMS optimization problem, inspired by MDS, has a simple solution that can be found via the singular value decomposition. Subsequently, we develop *Weighted Principal Curve Metric Scaling* (WPCMS), a weighted generalization of PCMS that permits control over the influence of particular elements of the contact matrix on the resulting reconstruction. Finally, we develop *Poisson Metric Scaling* (PoisMS) that uses WPCMS as a building block for fitting a Poisson model for contact counts. PoisMS provides an efficient way to find a 1D chromatin reconstruction that combines advantages of both MDS and probabilistic models.

2.2 Principal curve metric scaling formulation

PCMS is based on (classical) MDS with the contact matrix C treated as a similarity matrix and approximated by an inner product matrix (Buja et al., 2008). Let $X = \begin{pmatrix} -x_1^T- \\ -x_2^T- \\ \dots \\ -x_n^T- \end{pmatrix} \in \mathbb{R}^{n \times 3}$ be the matrix of genomic loci coordinates and $\| \cdot \|_F$ the Frobenius norm. The goal is to minimize the *Strain* objective:

$$\ell(x_1, \dots, x_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} (C_{ij} - \langle x_i,\ x_j \rangle)^2 \iff \ell(X) = \|C - XX^T\|_F^2. \tag{4}$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Note that inner product based PCMS bypasses the need to convert contacts to distances, as is required for metric MDS.

Instead of adding a smoothness penalty to the objective, we impose an additional constraint:

$$x_1, \dots, x_n \in \gamma, \text{ where } \gamma \text{ is a smooth one-dimensional curve in } \mathbb{R}^3. \tag{5}$$

As emphasized, this constraint captures the inherent contiguity of chromatin. We model the curve γ by a cubic spline with k degrees-of-freedom as follows (Hastie et al., 2009). Suppose $h_1(t), \dots, h_k(t)$ are cubic spline basis functions in $\mathbb{R}^1$ then

$$\gamma(t) = (\gamma_1(t), \gamma_2(t), \gamma_3(t))^T, \text{ where } \gamma_j(t) = \textstyle\sum_{\ell=1}^{k} \Theta_{\ell j}\; h_\ell(t) \text{ for } j = 1, 2, 3.$$

Suppose t_i index genomic loci x_i in the parametrization space of t, i.e. $x_i = \gamma(t_i)$, and let $H \in \mathbb{R}^{n \times k}$ be the matrix of spline basis evaluations at t_i, i.e. $H_{i\ell} = h_\ell(t_i)$. Since binning typically results in evenly spaced genomic loci, it is convenient to set $t_1 = 1,\ t_2 = 2, \dots, t_n = n$, although irregular spacing is readily handled. So, the constraint (5) can be written as $X_{ij} = \sum_{\ell=1}^{k} \Theta_{\ell j}\; h_\ell(t_i)$, or equivalently, in matrix form as $X = H\Theta$ leading to the optimization problem

$$\text{minimize } \ell_{PCMS}(\Theta) = \|C - H\Theta\Theta^T H^T\|_F^2 \text{ w.r.t. } \Theta \in \mathbb{R}^{k \times 3}. \tag{6}$$

Hereafter we denote the corresponding solution $\hat{\Theta} = \text{PCMS}(C, H)$, the contact matrix approximation $\hat{C} = H\hat{\Theta}\hat{\Theta}^T H^T$, and the resulting chromatin reconstruction $\hat{X} = H\hat{\Theta}$.

2.3 PCMS solution

Note that the parameter Θ in the PCMS problem (6) is unconstrained. Since Θ is defined up to a multiplication by a full-rank matrix, one can always assume H to be a matrix with orthogonal columns. To find the PCMS solution the following lemma is useful.

LEMMA 2.1 If $H \in \mathbb{R}^{n \times k}$ is a matrix with orthogonal columns, i.e. $H^T H = I$, then problem (6) is equivalent to

$$\text{minimize } \tilde{\ell}_{PCMS}(\Theta) = \|H^T C H - \Theta\Theta^T\|_F^2 \text{ w.r.t. } \Theta \in \mathbb{R}^{k \times 3}. \tag{7}$$

Proof. Suppose $\tilde{H} = \begin{pmatrix} H & H_\perp \end{pmatrix}$ is a column-wise combination of H and its orthogonal complement $H_\perp$. Therefore, $\tilde{H}$ is a square orthogonal matrix and for any $B \in \mathbb{R}^{n \times n}$ the following relation holds

$$\|B\|_F^2 = \|\tilde{H}^T B\|_F^2 = \|\tilde{H}^T B \tilde{H}\|_F^2 = \|H^T B H\|_F^2 + \|H_\perp^T B H_\perp\|_F^2.$$

Substituting $B = C - H\Theta\Theta^T H^T$ we conclude that

$$\|C - H\Theta\Theta^T H^T\|_F^2 = \|H^T C H - \Theta\Theta^T\|_F^2 + \|H_\perp^T C H_\perp\|_F^2 = \tilde{\ell}_{PCMS}(\Theta) + const.$$

In the last equation the constant term does not depend on the parameter Θ implying the equivalence of the optimization problems (6) and (7). □

Further, note that minimizing the objective $\tilde{\ell}_{PCMS}(\Theta)$ can be interpreted as a low-rank approximation of the matrix $H^T C H$ by a positive semi-definite rank 3 matrix $\Theta\Theta^T$. Assuming that the symmetric matrix $H^T C H$ has at least three positive eigenvalues the solution can be found via eigen-decomposition of $H^T C H$: let $H^T C H = Q\Lambda Q^T$ for orthogonal Q and diagonal $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$, then

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$$\Theta = Q\sqrt{\Lambda_3}, \text{ where } \sqrt{\Lambda_3} = \mathrm{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \sqrt{\lambda_3}, 0, \ldots, 0).$$

The computational efficiency of PCMS derives from the fact that it relies on eigen-decomposition of a small $k \times k$ matrix, requiring only $O(k^3)$ additional operations.

2.4 WPCMS problem statement

Our weighted generalization of PCMS was motivated by reducing the influence of elements of the contact matrix on the resulting reconstruction, in particular to counteract diagonal dominance (Yang et al., 2017). We introduce a matrix of weights $W \in [0,1]^{n\times n}$ and consider the *weighted* Strain objective

$$\ell(x_1,\ldots,x_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(C_{ij} - \langle x_i,\ x_j\rangle)^2 \iff \ell(X) = \|\sqrt{W} * (C - XX^T)\|_F^2 \tag{8}$$

where $*$ refers to the Hadamard product. The *Weighted Principal Curve Metric Scaling* (WPCMS) problem can be stated as follows:

$$\text{minimize } \ell_{WPCMS}(X) = \|\sqrt{W} * (C - H\Theta\Theta^T H^T)\|_F^2 \text{ w.r.t. } \Theta \in \mathbb{R}^{k\times 3}. \tag{9}$$

The corresponding solution is denoted $\hat{\Theta} = \mathrm{PCMS}_W(C, H)$.

2.5 WPCMS iterative algorithm

Optimization problem (9) can be elegantly solved via an iterative algorithm that repeatedly applies PCMS to a convex combination of the current contact matrix reconstruction $\hat{C}$ and the original contact matrix C:

1. **[Initialize]** Generate random $\Theta \in \mathbb{R}^{k\times 3}$.
2. *Repeat until convergence*:
   2.1 **[Reconstruct]** Calculate the current guess for the contact matrix approximation
   $$\hat{C} = H\Theta\Theta^T H^T.$$
   2.2 **[Mix]** Combine the current approximation and the original contact matrices
   $$C^* = W * C + (1 - W) * \hat{C}.$$
   2.3 **[PCMS]** Update Θ via computing the corresponding PCMS solution
   $$\Theta := \mathrm{PCMS}(C^*, H).$$

Convergence is assessed via the stopping criterion $\left|\frac{\ell_{WPCMS}(\Theta_{old}) - \ell_{WPCMS}(\Theta_{new})}{\ell_{WPCMS}(\Theta_{old})}\right| < \epsilon_1$, where ϵ_1 is some pre-chosen accuracy rate, Θ_{old} is the value of Θ calculated at the previous iteration and Θ_{new} is the updated value of Θ. Further details and extensions of the WPCMS algorithm are provided in the supplementary materials.

2.6 WPCMS algorithm connections

The core of the WPCMS approach is an alternating algorithm for solving *weighted low-rank approximation* problems introduced by Srebro and Jaakkola [2003]. It can also be viewed as a generalization of the *soft-impute* technique, devised to find a low-rank matrix approximation in the presence of missing values (Mazumder et al., 2010). We present WPCMS here as a *projected gradient descent* (PGD) procedure (Hastie et al., 2015), broadly used to solve

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

constrained optimization problems. Specifically, note that the WPCMS problem (9) can be restated as an optimization problem on a matrix manifold

$$\text{minimize } \ell_{WPCMS}(M) = \|\sqrt{W} * (C - HMH^T)\|_F^2 \text{ w.r.t. } M \in S_+^k(3) \tag{10}$$

where $S_+^k(3) = \{M \in \mathbb{R}^{k \times k} : M \succeq 0, \ \mathrm{rk}(M) = 3\}$ is the manifold of positive semidefinite matrices of rank 3. Hence, this problem can be solved via PGD iteration:

$$\textbf{[Gradient]} \ \ M := M - \nabla f(M) \quad \text{and} \quad \textbf{[Projection]} \ \ M := \mathrm{proj}_{S_+^k(3)}(M).$$

These steps are equivalent to the **[Mix]** and **[PCMS]** steps of the WPCMS algorithm.

2.7 PoisMS problem statement

Our Poisson Metric Scaling model is based on Poisson distributed contact counts C_{ij}. However, unlike previous formulations (*cf* Rosenthal et al. [2019] above), our prescribed dependence of the Poisson parameters on genomic loci coordinates confers the considerable advantage of convexity of the resulting optimization problem.

Consider the model (3) with the contact counts each having a Poisson distribution:

$$C_{ij} \sim Pois(\lambda_{ij}), \quad \log(\lambda_{ij}) = \alpha\langle x_i, x_j\rangle + \beta, \tag{11}$$

with $\alpha > 0$ and $\beta \in \mathbb{R}$ hyperparameters. The negative log-likelihood objective is

$$\ell_{PoisMS}(X) = \sum_{1 \le i,j \le n} \left[e^{\alpha\langle x_i, x_j\rangle + \beta} - C_{ij}\left(\alpha\langle x_i, x_j\rangle + \beta\right)\right] \tag{12}$$

and the MLE optimization problem under the smooth curve constraint 5 is

$$\text{minimize } \ell_{PoisMS}(X) \text{ w.r.t. } X \text{ subject to } X = H\Theta. \tag{13}$$

We denote the corresponding solution as $\hat{\Theta} = \ \mathrm{PoisMS}(C, H)$.

2.8 PoisMS iterative algorithm

A virtue of the Poisson model is that the second order Taylor approximation of the negative log-likelihood (12) is simply the weighted Frobenius norm. Moreover, it is well known that the optimal value of this second order approximation (SOA) amounts to one step of the Newton method for optimizing the original loss function. We use these facts to develop an iterative algorithm based on the WPCMS technique, which is equivalent to a projected Newton Method.

First, we review the SOA of the negative Poisson log-likelihood in the univariate case. Suppose $c \sim Pois(\lambda)$. The negative log likelihood $\ell(\lambda) = \lambda - c\log\lambda$ can be reparametrized in terms of the natural parameter $\eta = \log(\lambda)$ leading to $\ell(\eta) = e^\eta - c\eta$. Then the SOA of the reparametrized negative log-likelihood at some point $\eta_0 = \log\lambda_0$, up to scaling and shifting by a constant, is:

$$\ell(\eta) \approx \ell_{SOA}(\eta) = w(z - \eta)^2 \ \text{ where } \ w = e^{\eta_0} = \lambda_0 \ \text{ and } \ z = \eta_0 + \frac{c - \lambda_0}{\lambda_0}.$$

The multivariate version can be stated as follows.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

LEMMA 2.2 Suppose $C \in \mathbb{Z}_+^{n\times n}$ where $C_{ij} \sim Pois(\lambda_{ij})$ and $\eta_{ij} = \log(\lambda_{ij})$. Let the respective matrices of Poisson and natural parameters be $\Lambda = [\lambda_{ij}] \in \mathbb{R}_+^{n\times n}$ and $\mathcal{H} = [\eta_{ij}] \in \mathbb{R}^{n\times n}$. Then the SOA of the negative log-likelihood at some point $\mathcal{H}_0$, up to scale and shift constants, is

$$\ell(\mathcal{H}) \approx \ell_{SOA}(\mathcal{H}) = \|\sqrt{W} * (Z - \mathcal{H})\|_F^2 \text{ where } W = e^{\mathcal{H}_0} = \Lambda_0 \text{ and } Z = \mathcal{H}_0 + \frac{C - \Lambda_0}{\Lambda_0}.$$

Here $*$ is the Hadamard (element-wise) product, with matrix exponentiation and division also being interpreted as element-wise operations.

Recall that in the Poisson model (11) the natural parameter depends linearly on the matrix of spatial coordinate inner products: $\mathcal{H} = \log \Lambda = \alpha XX^T + \beta$. So, the SOA can be rewritten as

$$\ell_{SOA}(\mathcal{H}) = \|\sqrt{W} * (Z - \alpha XX^T - \beta)\|_F^2 = \alpha^2 \|\sqrt{W} * (\tilde{Z} - XX^T)\|_F^2 \text{ for } \tilde{Z} = \frac{Z - \beta}{\alpha}.$$

Suppose that the current guess for the chromatin reconstruction is X_0 with corresponding natural parameter value $\mathcal{H}_0 = \alpha X_0 X_0^T + \beta$. Then we have the following approximation of the Poisson loss (12) at point $\mathcal{H}_0$ again up to scaling and shifting by a constant:

$$\ell_{PoisMS}(X) \approx \ell_{SOA}(X) = \|\sqrt{W} * (Z - XX^T)\|_F^2$$
$$\text{where } W = e^{\alpha X_0 X_0^T + \beta} \text{ and } Z = X_0 X_0^T + \tfrac{1}{\alpha}(\tfrac{C - W}{W}).$$

It is not difficult to check that under the smooth curve constraint $X = H\Theta$ the loss function $\ell_{SOA}(X)$ coincides with the WPCMS loss (9). Therefore, we obtain a nice application of the WPCMS algorithm, with the solution to the second order approximation of problem (13):

$$\text{minimize } \ell_{SOA}(\Theta) = \|\sqrt{W} * (Z - H\Theta\Theta^T H^T)\|_F^2 \text{ w.r.t. } \Theta \tag{14}$$

being exactly $\Theta = \text{PCMS}_W(Z, H)$. This observation can be applied to simplify computations for the Poisson model and underlies our PoisMS algorithm. The algorithm repeatedly approximates the Poisson objective at current guess Θ by a quadratic function and shifts Θ towards the global minimum of this quadratic approximation:

1. **[Initialize]** Generate random $\Theta \in \mathbb{R}^{k\times 3}$, calculate the current guess for the contact matrix approximation $\hat{C} = H\Theta\Theta^T H^T$.

2. *Repeat until convergence*:

   2.1 **[SOA]** Calculate the SOA for the Poisson loss at the current guess

   $$W = e^{\alpha\cdot\hat{C} + \beta} \quad \text{and} \quad Z = \hat{C} + \tfrac{1}{\alpha}(\tfrac{C - W}{W}).$$

   2.2 **[WPCMS]** Update current guesses for Θ and $\hat{C}$ via the WPCMS approach

   $$\Theta := \text{PCMS}_W(Z, H) \quad \text{and} \quad \hat{C} := H\Theta\Theta^T H^T.$$

The stopping rule for the PoisMS algorithm is similar to WPCMS: for some fixed accuracy rate ϵ_2 we check if the updated Θ_{new} meets the criteria $\left|\frac{\ell_{PoisMS}(\Theta_{old}) - \ell_{PoisMS}(\Theta_{new})}{\ell_{PoisMS}(\Theta_{old})}\right| < \epsilon_2$ after each iteration of steps 2.1–2.2. The Supplement contains further details, including computational complexities of WPCMS and PoisMS, as well as algorithmic extensions.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2.9 Determination of principal curve degrees-of-freedom

The main hyperparameter for the PCMS and PoisMS approaches described above is the spline degrees-of-freedom df (spline basis size), which controls the smoothness of the resulting reconstruction. To determine the optimal value, for each df we create the spline basis matrix H_{df}, find the corresponding solution $\hat{\Theta}_{df}$ and the contact matrix approximation $\hat{C}_{df} = H_{df}\hat{\Theta}_{df}\hat{\Theta}_{df}^T H_{df}^T$. We measure the error rate by the normalized loss function, i.e

$$\text{PCMS:}\;\; err(\hat{C}_{df}) = \frac{1}{n^2}\|C - \hat{C}_{df}\|_F^2 \tag{15}$$

$$\text{PoisMS:}\;\; err(\hat{C}_{df}) = \frac{1}{n^2}\sum_{1\le i,j\le n} \lambda_{ij} - C_{ij}\log\lambda_{ij}, \text{ where } \log\Lambda = \alpha\hat{C}_{df} + \beta. \tag{16}$$

Initially, we tried cross-validation to find the optimal value of df, as is common for smoothing (penalty) parameter determination. However, the complex and structural dependencies that characterize contact matrices made this approach problematic. As an alternative we adopted an approach based on identifying the “elbow" that is prototypic in graphs of resubstitution error, here $err(\hat{C}_{df})$, versus model complexity, here df. The logic as to why this change point constitutes a basis for model complexity determination is described in Breiman et al. [1984] in terms of bias-8variance tradeoff. Elbow identification is also used for determining appropriate numbers of principal components (Jolliffe, 2002) and clusters (Hastie et al., 2009), as well as dimension in MDS (Kruskal and Wish, 1978) and non-negative matrix factorization (NMF; see Hutchins et al., 2008) problems.

2.10 Accuracy assessment via multiplex FISH

While the prescription in Section 2.9 provides a means for selecting a particular PCMS or PoisMS model it does not address the accuracy of the chosen model. The absence of gold standards makes such assessment challenging. In comparing competing 3D genome reconstructions several authors have appealed to simulation (Zhang et al., 2013, Varoquaux et al., 2014, Zou et al., 2016, Park and Lin, 2017), however, real data referents are preferable. To that end, many of the same reconstruction algorithm developers have made recourse to fluorescence in situ hybridization (FISH) imaging as a basis for gauging accuracy. This proceeds by comparing distances between imaged probes with corresponding reconstruction-based distances. But such methods are necessarily limited by the sparse number of probes ($\sim 2-6$; see Lieberman-Aiden et al., 2009, Shavit et al., 2014, Park and Lin, 2017) and the modest resolution thereof, many straddling over 1 megabase (Mb). The recent advent of *multiplex* FISH (Wang et al., 2016) transforms 3D genome reconstruction accuracy evaluation by providing an order of magnitude more probes and hence two orders of magnitude more inter-probe distances than conventional FISH. Moreover, the probes are at higher resolution (more precisely localized) and centered at previously defined topologically associated domains (TADs; see Dixon et al., 2012). We use this imaging data, along with companion accuracy assessment approaches (Segal and Bengtsson, 2018) to evaluate our PCMS and PoisMS reconstructions, as outlined next.

The image-based 3D genomic coordinates furnished from multiplex FISH serve to define the gold standard by which we assess reconstructions. The existence of numerous multiplex FISH replicates is crucial for this task and three steps are necessary to effect such evaluation.

Obtaining the gold standard. Given N multiplex FISH replicates denote the matrix of the spatial coordinates for replicate $i \in \{1,\ldots,N\}$ by $X_i^0 \in \mathbb{R}^{n_0\times 3}$ where n_0 denotes the number of distinct multiplex FISH loci (probes) over all replicates. We start by defining the *medoid replicate*. For a pair of 3D conformations $X_1, X_2 \in \mathbf{R}^{n_0\times 3}$ denote the number of observed loci by $n(X_1, X_2)$ and suppose $d_{proc}(X_1, X_2)$ is the Procrustes distance from X_2 to X_1

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

following alignment allowing translation, rotation and scaling (Hastie et al., 2009). Then the dissimilarity between X_1 and X_2 is defined by

$$d(X_1, X_2) = \frac{1}{n(X_1, X_2)} d_{proc}(X_1, X_2). \tag{17}$$

Using this dissimilaity one can calculate the medoid replicate as the replicate whose average dissimilarity to the other replicates is minimal:

$$i^* = \text{argmin}_{i=1,\dots,N} \sum_{j=1}^{N} d(X_i^0, X_j^0). \tag{18}$$

Next, let X_j^* be the Procrustes alignment of X_j^0 to the medoid $X_{i^*}^0$. The *average Procrustes conformation* $\bar{X}$, defined as the locus-wise average of the X_j^*, then serves as a gold standard. Our application of Procrustes alignment prior to this (noise reducing) averaging accommodates translation, rotation, and scaling and differences between replicate conformations.

Computing the reference distribution. Treating the average Procrustes conformation $\bar{X}$ as our gold standard we can obtain a reference distribution by measuring the dissimilarity between it and the multiplex FISH replicates, i.e. $d(\bar{X}, X_i^0)$. The resulting empirical distribution captures experimental variation around the gold standard. A fine point is that, by construction, this distribution will exhibit reduced dispersion compared to its target population quantity owing to data re-use since X_i^0 contributes to $\bar{X}$. While this concern could be mitigated by employing leave-one-out techniques the large number of available replicates (> 110) renders this approach unnecessary (Segal and Bengtsson, 2018).

Evaluating chromatin reconstructions. To evaluate reconstructions resulting from the PCMS and PoisMS approaches we first need to align the reconstruction with the gold standard. This may involve preliminary coarsening of one or other coordinate sets to yield comparable resolution. Here, the genomic coordinate ranges for each multiplex FISH probe are coarser than the Hi-C bins used in our reconstructions. So we calculate the average of the reconstruction coordinates falling in the corresponding multiplex FISH bins to obtain a lower resolution reconstruction $\hat{X}$ of the same dimension as $\bar{X}$. To quantify how close this reconstruction is to the gold standard $\bar{X}$ we again measure dissimilarity following alignment $d(\bar{X}, \hat{X})$ Interpretations of this quantity in the context of the reference distribution are presented in the Results section.

2.11 A contrasting reconstruction algorithm: HSA

To compare our PCMS and PoisMS solutions with an alternate reconstruction algorithm we make recourse to HSA (Zou et al., 2016). This technique provides an interesting contrast in that it employs a similar Poisson formulation to (12) but instead of contiguity being captured via principal curves per (5), it is indirectly imparted by constraints that induce dependencies on a hidden Gaussian Markov chain over the solution coordinates. Obtaining these spatial coordinates is achieved via simulated annealing with further smoothness effected via distance-based penalization. This procedure is embedded in an alternating algorithm with parameters analogous to α, β in (11) estimated by Poisson regression.

HSA has performed well in some benchmarking studies and features several compelling attributes: (i) it can simultaneously handle multiple data tracks allowing for integration of replicate contact maps; (ii) it can adaptively estimate the power-law index whereby contacts are transformed to distances, the importance of which has been previously emphasized (Zhang et al., 2013); and (iii) by using Hamiltonian dynamics based simulated annealing it can purportedly efficiently optimize over the space of genomic loci 3D coordinates. Nonetheless, in contrast to PCMS

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and PoisMS, HSA incurs a substantial compute and memory burden, and questions surrounding robustness have been raised (Rieber and Mahony, 2017).

To compare PCMS and PoisMS performance with HSA we use the approach described in Section 2.10: we measure the dissimilarity between the HSA reconstructions and the gold standard and interpret the obtained quantities in the context of the attendant reference distribution and the corresponding PCMS and PoisMS values (see Section 3.3).

3 Results

3.1 Chromosome reconstructions

We present PCMS and PoisMS reconstructions for IMR90 cell chromosome 20 at 100kb resolution for which multiplex FISH and Hi-C data acquisition and processing has been previously described (Segal and Bengtsson, 2018). Results for chromosome 21 are presented in the Supplement.

For the Poisson model (11) the inner product matrix XX^T depends on the Poisson parameters Λ: $XX^T = \frac{\log \Lambda - \beta}{\alpha}$. So, to make PCMS and PoisMS results comparable, we apply the algorithms to $C^{\log} = \frac{\log(C+\epsilon)-\beta}{\alpha}$ and C, respectively. Here ϵ is a small constant introduced to avoid taking the logarithm of zero and β can be interpreted as a centering constant. In our experiments we take $\alpha = 1$, $\beta = \log\left(\frac{\sum_{i,j=1}^{n} C_{ij}}{n}\right)$ and $\epsilon = 10^{-3}$ (the recommended range for this parameter is $10^{-1}, \ldots, 10^{-4}$). The corresponding transformed contact matrix is shown in Figure 7; the resulting contact matrix approximations $\hat{C}_{df}$ and the chromatin reconstructions $\hat{X}_{df}$ computed via the PCMS and PoisMS techniques are presented in Figures 8 and 9, respectively.

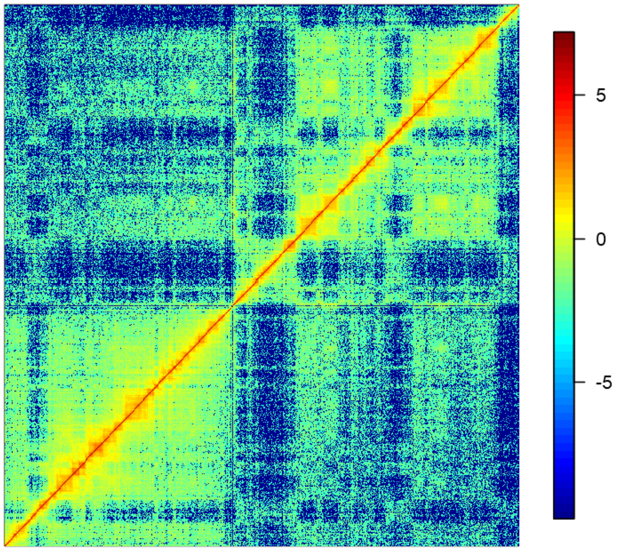
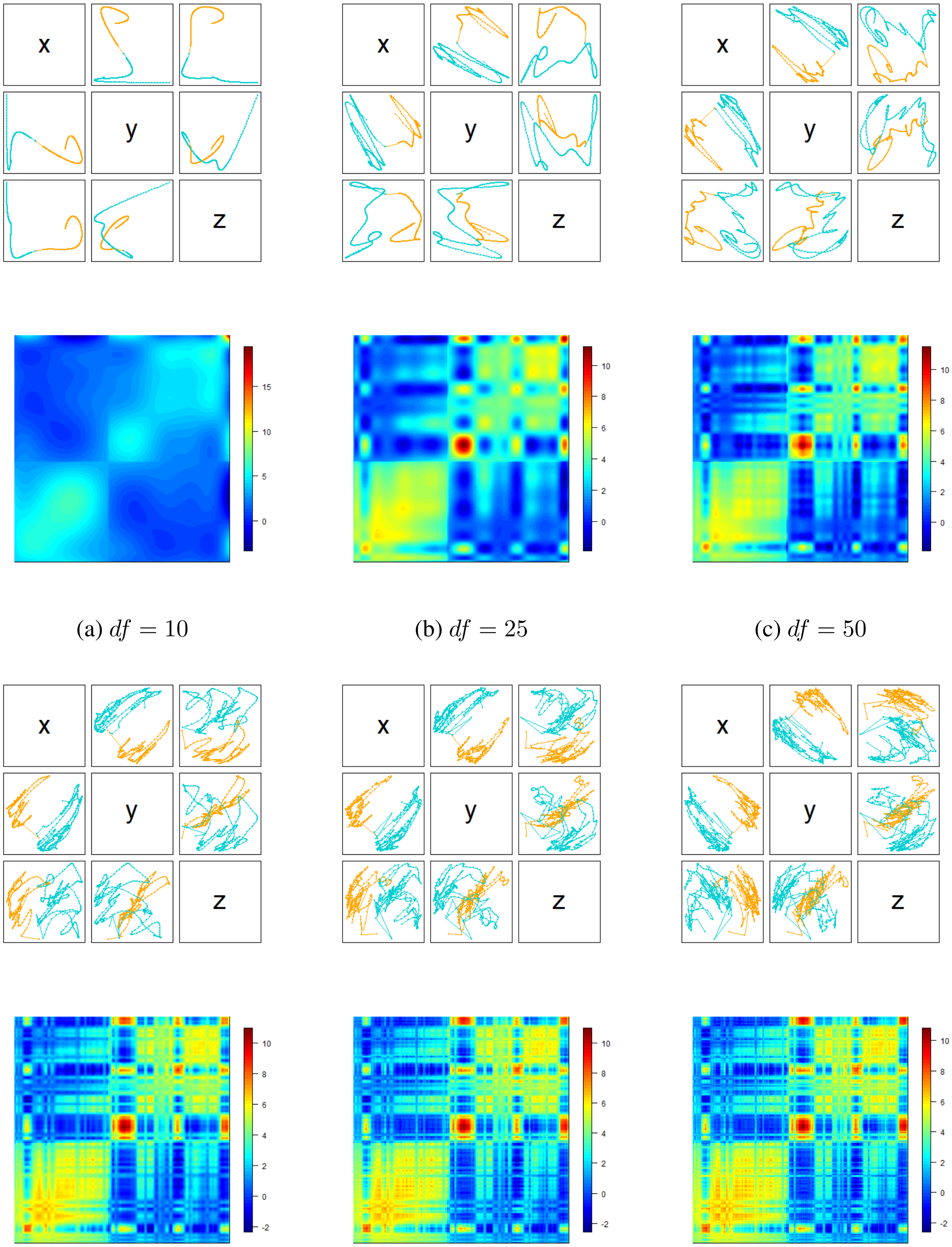


Figure 1: Log-transformed contact matrix $C^{\log} = \frac{\log(C+\epsilon)-\beta}{\alpha}$ for $\epsilon = 0.001$, $\alpha = 1$ and $\beta = \log\left(\frac{\sum_{i,j=1}^{n} C_{ij}}{n}\right)$.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) $df = 10$

(b) $df = 25$

(c) $df = 50$

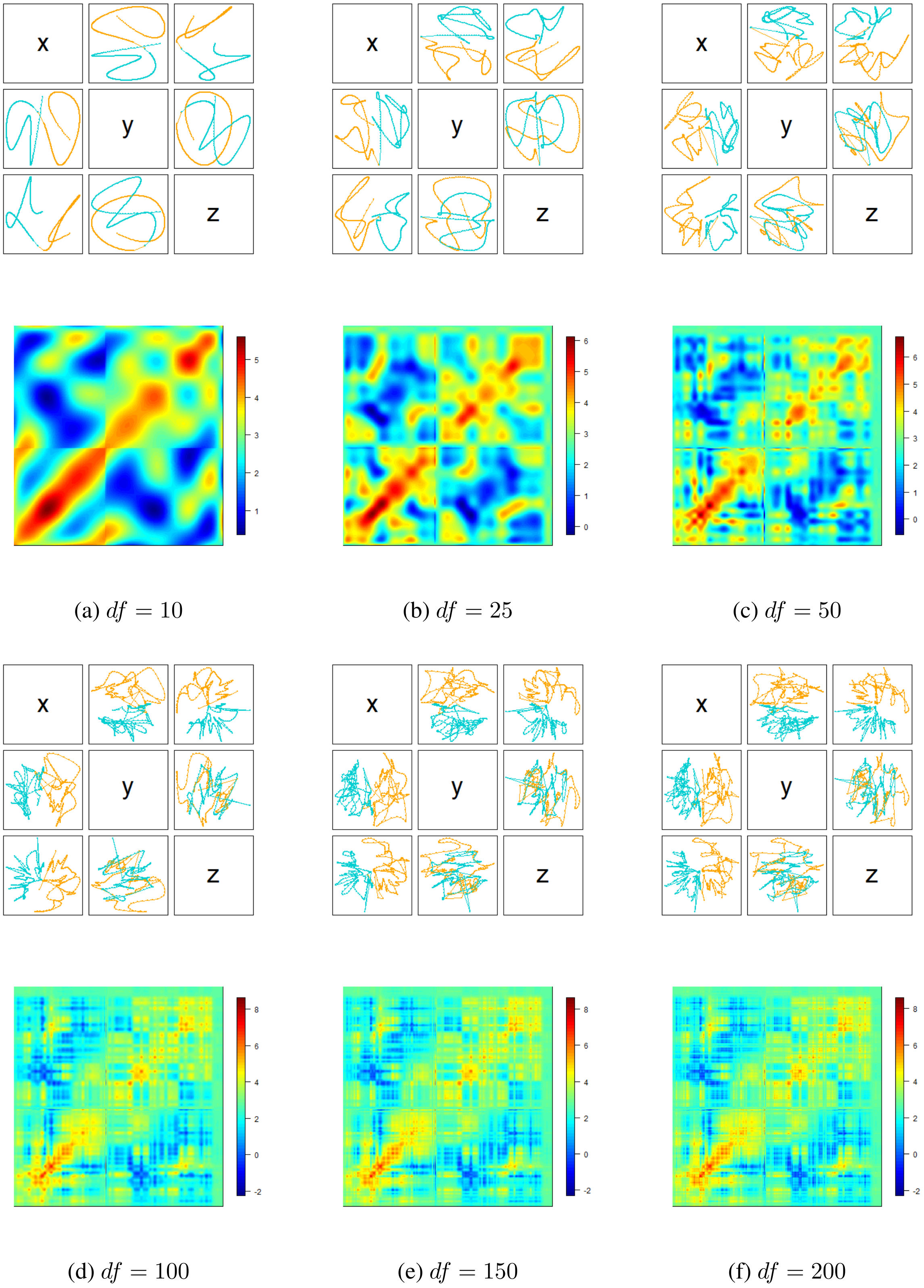
(d) $df = 100$

(e) $df = 150$

(f) $df = 200$

Figure 2: $\hat{X}_{df}$, the projections of the resulting reconstruction, and $\alpha\hat{C}_{df} + \beta$, the approximation of $\log(C)$, obtained via PCMS for different degrees-of-freedom values df.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) $df = 10$

(b) $df = 25$

(c) $df = 50$

(d) $df = 100$

(e) $df = 150$

(f) $df = 200$

Figure 3: $\hat{X}_{df}$, the projections of the resulting reconstruction, and $\alpha\hat{C}_{df} + \beta$, the approximation of $\log(C)$, obtained via PoisMS for different degrees-of-freedom values df.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3.2 Determining degrees-of-freedom

To estimate degrees-of-freedom for both PCMS and PoisMS we plot error rate $err(\hat{C}_{df})$ vs df. For PCMS, error rate decreases rapidly up to $df = 25$ with subsequent decline being gradual. A similar pattern pertains for PoisMS with the gradual error decline evident after $df = 20$ (Figure 10). Estimated df according to the elbow (slope change point) heuristic are obtained using the R package `segmented` (Muggeo, 2008) and depicted in the respective Figures.

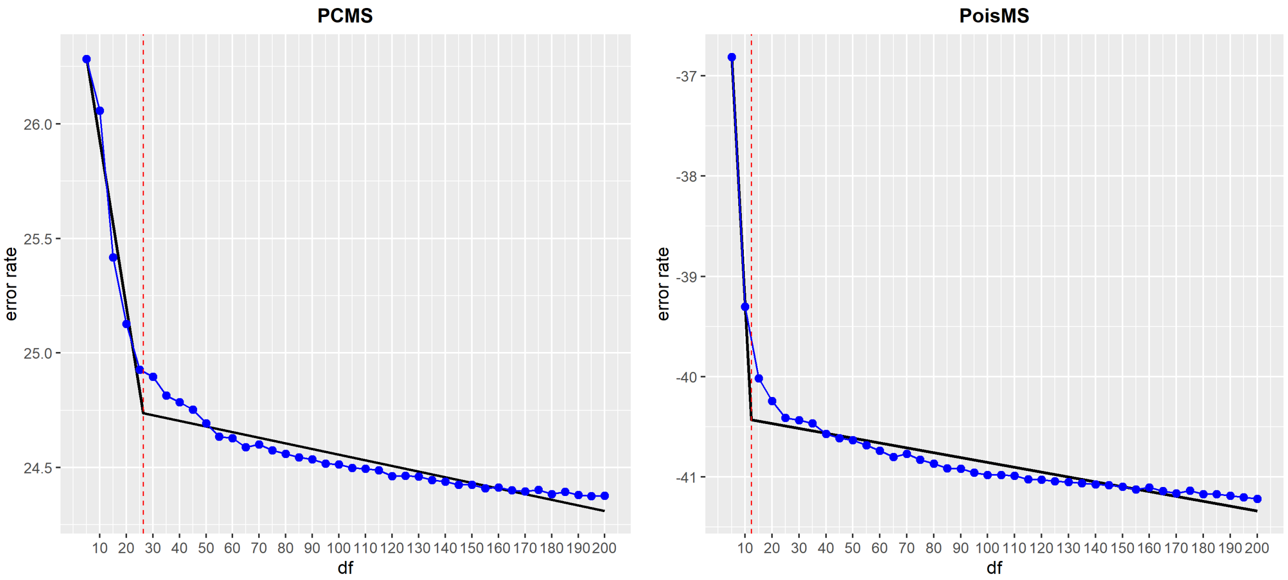


Figure 4: Error rate, here $err(\hat{C}_{df}) = \frac{1}{n_{obs}^2}\|C^{\log} - \hat{C}_{df}\|_F^2$ for PCMS and $err(\hat{C}_{df}) = \frac{1}{n_{obs}^2}\sum_{1\leq i,j\leq n} \lambda_{ij} - C_{ij}\log\lambda_{ij}$ with $\log\Lambda = \alpha\hat{C}_{df} + \beta$ for PoisMS, vs. degrees-of-freedom, here df. The segmented regression is given by the piecewise linear fit (black) with the degrees-of-freedom selected via kink estimation indicated by the red vertical line. Segmentation change point indicated for PCMS is $df = 26.32$ and for PoisMS is $df = 12.27$. Note that the optimal df value obtained via this method tends to be conservative.

3.3 Evaluating reconstructions via the multiplex FISH referent

Procrustes alignment of 3D conformations, and calculation of the corresponding Procrustes distances $d_{proc}(\cdot,\cdot)$, was performed using the R package `vegan` (Oksanen et al., 2016). We obtain the multiplex FISH medoid conformation based on the smallest row sum (18) of the dissimilarity matrix of normalized Procrustes distances (17) as described (Section 2.10). The 111 multiplex FISH replicate conformations are then aligned to the medoid as a prelude to calculating the average Procrustes conformation — our gold standard — presented in Figure 11. Next, we plot the multiplex FISH dissimilarity reference distribution of dissimilarities between replicates and the gold standard. We position the PCMS and PoisMS reconstruction dissimilarities with the gold standard in the resulting histograms. Figures 12 presents results for a series of reconstructions based on select degrees-of-freedom: $df = 5, 10, 25, 50, 100, 150, 200$. Finally, HSA reconstruction dissimilarity values are included in the plot.

The following conclusions can be drawn from Figure 12. For chromosome 20 (see Supplement for chromosome 21), all fits for both PCMS and PoisMS lie within the range of the multiplex FISH dissimilarity distribution that reflects experimental variation. The fact that the PCMS and PoisMS dissimilarity values are in the left tail of this distribution

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

indicates the accuracy of the proposed reconstructions, highlighting the utility of the proposed methodologies. Further, the larger HSA highlights the potential of our methods.

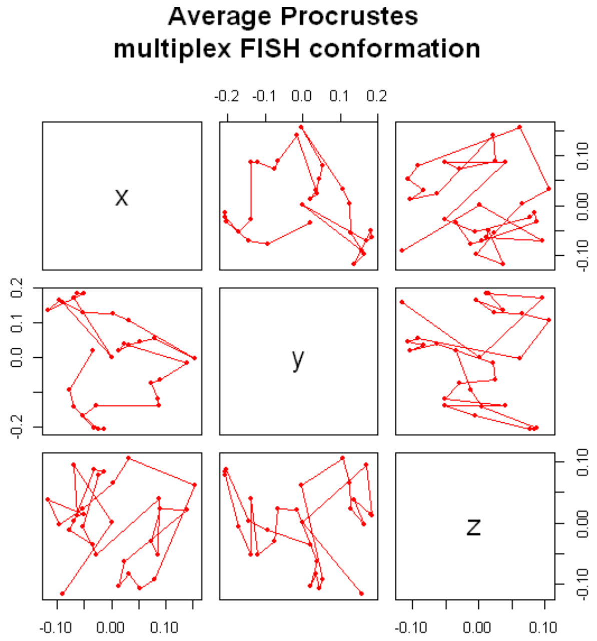


Figure 5: The average Procrustes conformation computed for 111 multiplex FISH replicate conformations for chromosome 20.

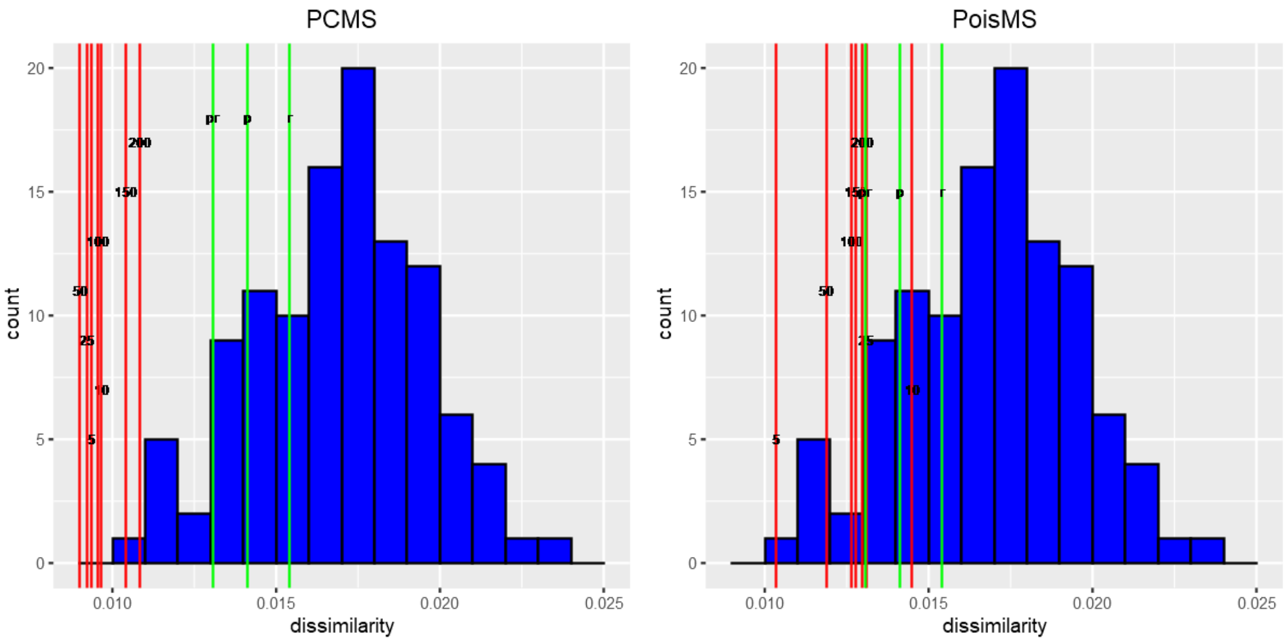


Figure 6: Reference distribution measuring the dissimilarity between the gold standard $\bar{X}$ and 111 multiplex FISH replicate conformations X_i^0 for chromosome 20. The vertical red lines correspond to the dissimilarity between $\bar{X}$ and the low-resolution reconstruction $\hat{X}_{df}$ calculated for different df values; the green lines represent three types of the HSA reconstruction(see Sections 2.10 and 2.11).

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

4 Discussion

Central to our principal curve based approaches to 3D chromatin reconstruction is that the configuration of an individual chromosome within the nucleus can be treated as a contiguous 1D curve since the diameter of the chromatin fiber is negligible compared to the nuclear volume. The extent to which the curve is “smooth” is determined by an adaptively selected degrees-of-freedom parameter. As mentioned in the introduction, previous reconstruction methods either impart contiguity indirectly by prescribing constraints, which are difficult to specify, or impose it post hoc. In comparison, our methods based on principal curves are computationally efficient, readily scale to high resolution contact data and are parsimonious with regard tuning parameters.

Our implementation of PCMS and PoisMS utilizes cubic spline basis functions, which contribute to this computational efficiency. However, the nature of chromatin folding and attendant Hi-C data is such that these bases will be less effective in capturing fine 3D structure, as opposed to global backbone architecture. This derives from the hierarchical, domain-based organization of chromatin, aspects that have been tackled by some reconstruction algorithms using strategies that synthesize solutions obtained at differing scales (Rieber and Mahony, 2017, Trieu et al., 2019). We will investigate whether principal curve solutions can similarly serve as building blocks in addition to exploring the use of alternate basis functions, notably wavelets.

Our analyses of Hi-C data from IMR90 cells was motivated by the availability of corresponding multiplex FISH data enabling accuracy assessment. However, the extent and resolution of multiplex FISH imaging is limited, narrowing the applicability of this means of evaluation. An even more fundamental issue pertains to attempting chromatin reconstruction using *bulk* Hi-C data from large cell populations. As has been emphasized (Lando et al., 2018), the presence of numerous conflicting contacts suggests that the notion of a consensus underlying 3D conformation is questionable and that there is substantial cell-to-cell structural variation. This places a premium on pursuing single cell reconstructions as enabled by the recent emergence of single cell Hi-C protocols (Ramani et al., 2017). That one of these advances (Stevens et al., 2017) also provides parallel imaging data, putatively enabling reconstruction accuracy determination, underscores the importance of applying reconstruction methods in single cell settings, despite contact map sparsity, and is the subject of future work.

5 Software

Proposed methods are implemented in the R package `PoisMS`; the software is available from Github (`https://github.com/ElenaTuzhilina/PoisMS`).

Funding

This work was supported by the National Institutes of Health [GM109457 to M.S.]

Acknowledgments

Conflict of Interest: None declared.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

F. Mitelman, B. Johansson, and F. Mertens. The impact of translocations and gene fusions on cancer causation. *Nature Reviews Cancer*, 7:233–245, 2007.

J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295:1306–1311, 2002.

E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range contacts reveals folding principles of the human genome. *Science*, 326:289–293, 2009.

Z. Duan, M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble. A three-dimensional model of the yeast genome. *Nature*, 465:363–367, 2010.

Z. Zhang, G. Li, Toh K.-C., and W.-K. Sung. 3D chromosome modeling with semi-definite programming and Hi-C data. *Journal of Computational Biology*, 20:831–846, 2013.

N. Varoquaux, F. Ay, W. S. Noble, and J. P. Vert. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics*, 30:26–33, 2014.

F. Ay, E. M. Bunnik, N. Varoquaux, S. M. Bol, J. Prudhomme, J. P. Vert, W. S. Noble, and K. G. Le Roch. Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Research*, 24:974–88, 2014.

C. Zou, Y. Zhang, and Z. Ouyang. HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure. *Genome Biology*, 17:40, 2016.

L. Rieber and S. Mahony. miniMDS: 3D structural inference from high-resolution hi-c data. *Bioinformatics*, 33: 261–266, 2017.

D. M. Witten and W. S. Noble. On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic Acids Research*, 40:3849–3855, 2012.

D. Capurso and M. R. Segal. Distance-based assessment of the localization of functional annotations in 3D genome reconstructions. *BMC Genomics*, 15:992, 2014.

C. S. Lee, R. W. Wang, H. H. Chang, D. Capurso, M. R. Segal, and J. E. Haber. Chromosome position determines the success of double-strand break repair. *Proceedings of the National Academy of Science*, 113:146–154, 2016.

D. Capurso, H. Bengtsson, and M. R. Segal. Discovering hotspots in functional genomic data superposed on 3D chromatin configuration reconstructions. *Nucleic Acids Research*, 44:2028–2035, 2016.

Y. Shavit, F. K. Hamey, and P. Lio. FisHiCal: an R package for iterative FISH-based calibration of Hi-C data. *Bioinformatics*, 30:3120–3122, 2014.

T. J. Stevens, D. Lando, S. Basu, L. P. Atkinson, Y. Cao, S. F. Lee, M. Leeb, K. J. Wohlfahrt, W. Boucher, A. O'Shaughnessy-Kirwan, J. Cramard, A. J. Faure, M. Ralser, E. Blanco, L. Morey, M. Sanso, M. G. S. Palayret, B. Lehner, L. Di Croce, A. Wutz, B. Hendrich, D. Klenerman, and E. D. Laue. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 544:59–64, 2017.

J. Park and S. Lin. A random effect model for reconstruction of spatial chromatin structure. *Biometrics*, 73:52–62, 2017.

T. J. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 406:502–516, 1989.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

M. R. Segal and H. L. Bengtsson. Reconstruction of 3D genome architecture via a two-stage algorithm. *BMC Bioinformatics*, 16:373, 2015.

S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159:1665–1680, 2014.

M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, 9:999–1003, 2012.

T. J. Hastie, R. J. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2009.

G. Fudenberg and L. A. Mirny. Higher-order chromatin structure: bridging physics and biology. *Current Opinions in Genetics & Development*, 22:115–124, 2012.

M. Rosenthal, D. Bryner, F. Huffer, S. Evans, A. Srivastava, and N. Neretti. Bayesian Estimation of 3D Chromosomal Structure from Single Cell Hi-C Data. *Journal of Computational Biology*, 26:1191–1202, 2019.

C. Caudai, E. Salerno, M. Zopp, and A. Tonazzini. Inferring 3d chromatin structure using a multiscale approach based on quaternions. *BMC Bioinformatics*, 16:234, 2015.

A. Buja, D. F. Swayne, M. L. Littman, N. Dean, H. Hofman, and L. Chen. Data Visualization With Multidimensional Scaling. *Journal of Computational and Graphical Statistics*, 17:444–472, 2008.

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

N. Srebro and T. Jaakkola. Weighted Low-Rank Approximations. *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, 20:720–727, 2003.

R. Mazumder, T. J. Hastie, and R. J. Tibshirani. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

T. J. Hastie, R. J. Tibshirani, and M. J. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall, New York, 2015.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman and Hall, New York, 1984.

I. Jolliffe. *Principal Component Analysis*. Springer, New York, 2002.

J. B. Kruskal and M. Wish. *Multidimensional Scaling*. Sage, Newbury Park, 1978.

L. N. Hutchins, S. M. Murphy, P. Singh, and J. H. Graber. Position-dependent motif characterization using non-negative matrix factorization. *Bioinformatics*, 24:2684–2690, 2008.

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

J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin contacts. *Nature*, 485:376–380, 2012.

M. R. Segal and H. L. Bengtsson. Improved accuracy assessment for 3D genome reconstructions. *BMC Bioinformatics*, 19:196, 2018.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

V. M. Muggeo. segmented: an R package to fit regression models with broken-line relationships. *Rnews*, 8:20–25, 2008.

J. Oksanen, F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, H. Stevens, E. Szoecs, and H. Wagner. vegan: Community Ecology Package. *R package version*, 2:4–1, 2016. URL `http://cran.r-project.org/package-vegan`.

T. Trieu, O. Oluwadare, and J. Cheng. Hierarchical reconstruction of high-resolution 3D models of large chromosomes. *Scientific Reports*, 9:4971, 2019.

D. Lando, T. J. Stevens, S. Basu, and E. D. Laue. Calculation of 3D genome structures for comparison of chromosome conformation capture experiments with microscopy: An evaluation of single-cell Hi-C protocols. *Nucleus*, 9: 190–201, 2018.

V. Ramani, X. Deng, K. L. Gunderson, F. J. Steemers, C. M. Disteche, W. S. Noble, Z. Duan, and J. Shendure. Massively multiplex single-cell Hi-C. *Nature Methods*, 14:263–266, 2017.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Material

6 Algorithms details and extensions

6.1 WPCMS algorithm details

1. Since the WPCMS algorithm is equivalent to Projected Gradient Descent (PGD), many extensions can be naturally derived. For example, a learning rate δ can be introduced by replacing W by δW at the **[Mix]** step; furthermore, it is not difficult to add line search to the algorithm:

$$\textbf{[Mix]}\ \ C^* = W * C + (1 - W) * \hat{C} \quad \Longrightarrow \quad \begin{aligned} \textbf{[Search]}\ C^*_\delta &= \delta W * C + (1 - \delta W) * \hat{C} \\ \delta_1 &= \operatorname{argmin}_\delta \ell_{WPCMS}(\mathrm{PCMS}(C^*_\delta, H)) \\ \textbf{[Mix]}\ C^* &= \delta_1 W * C + (1 - \delta_1 W) * \hat{C} \end{aligned}$$

2. One more improvement of the WPCMS algorithm could be done for the **[Initialize]** step.

$$\textbf{[Initialize]}\ \ \text{Generate random } \Theta \in \mathbb{R}^{k\times 3} \quad \Longrightarrow \quad \textbf{[Initialize]}\ \ \Theta := \mathrm{PCMS}(C, H)$$

It has been experimentally shown that using $\mathrm{PCMS}(C, H)$ as a warm start could significantly decrease the number of iterations required for the algorithm to converge.

6.2 PoisMS algorithm details

1. It is not necessary to wait until the complete convergence of the objective $\ell_{SOA}(\Theta)$ at the **[WPCMS]** step. According to our experiments, the PoisMS algorithm can be significantly accelerated if one does just a few steps of WPCMS between each update of W and Z. This can be easily done varying the accuracy rate ϵ_1 for the WPCMS stopping rule $\left|\frac{\ell_{SOA}(\Theta_{old}) - \ell_{SOA}(\Theta_{new})}{\ell_{SOA}(\Theta_{old})}\right| < \epsilon_1$.
2. First, note that $\ell_{PoisMS}(X)$ depends only on the scalar products $\langle x_i, x_j \rangle$ and, therefore, can be efficiently reparametrized in terms of XX^T. Moreover, under the smooth curve constraint $X = H\Theta$ the Poisson loss function could be further reparametrized in terms of the similarity matrix $S = H\Theta\Theta^T H^T$, i.e $\ell_{PoisMS}(S) = \sum_{1\le i,j\le n} e^{\alpha S_{ij}+\beta} - c_{ij}(\alpha S_{ij} + \beta)$. Finally, the optimal value of the Poisson loss quadratic approximation at current guess $\mathcal{H}_0$ coincides with the result of one step of the Newton Method initialized at $\mathcal{H}_0$. Combining these together, one can add a learning rate as well as line search to the Poisson Metric Scaling algorithm.

$$\begin{aligned} \textbf{[WPCMS]}\ \Theta &:= \mathrm{PCMS}_W(Z, H) \\ \hat{C} &:= H\Theta\Theta^T H^T \end{aligned} \quad \Longrightarrow \quad \begin{aligned} \textbf{[WPCMS]}\ \Theta &:= \mathrm{PCMS}_W(Z, H) \\ C^* &= H\Theta\Theta^T H^T \\ \textbf{[Search]}\ C^*_\delta &= (1-\delta)\hat{C} + \delta C^* \\ \delta_2 &= \operatorname{argmin}_\delta \ell_{PoisMS}(C^*_\delta) \\ \textbf{[Mix]}\ \hat{C} &:= (1-\delta_2)\hat{C} + \delta_2 C^* \end{aligned}$$

3. Similarly to the WPCMS algoritm, the PCMS technique can be used to find a warm start at the **[Initialize]** step. According to the Poisson model under consideration, the Poisson parameters matrix Λ depend on the

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

genomic loci spatial coordinates X via the following relation $\Lambda = e^{\alpha XX^T+\beta}$. Therefore, for the initialization step it is reasonable to assume the following approximation $C \approx e^{\alpha XX^T+\beta}$, which is essentially the same as if the log transformed contact matrix $\frac{\log C-\beta}{\alpha}$ was approximated by the inner product matrix XX^T. To avoid taking logarithm of zero one can replace $\log C$ by $\log(C+\epsilon)$ for a small ϵ and consider the new initialization as follows:

$$\textbf{[Initialize]}\ \text{Generate random } \Theta \in \mathbb{R}^{k\times 3} \qquad \Longrightarrow \qquad \textbf{[Initialize]}\ \Theta := \text{PCMS}\left(\frac{\log(C+\epsilon)-\beta}{\alpha}\right)$$

$$\hat{C} := H\Theta\Theta^T H^T \qquad\qquad \hat{C} := H\Theta\Theta^T H^T$$

4. The hyperparameter α does not play any important role in the reconstruction as it is responsible for the resulting conformation scaling only. On the other hand, for the fixed Θ one can optimize the negative log-likelihood w.r.t. the hyperparameter β and hence add an additional **[Update β]** step to the algorithm:

$$4.\ \textbf{[Update } \beta\textbf{]}\ \ \beta := \log\left(\frac{\sum_{1\le i,j\le n} C_{ij}}{\sum_{1\le i,j\le n} e^{\alpha\cdot S_{ij}}}\right)$$

6.3 Computational complexity

The computational complexity of PCMS is $O(n^2k)$ operations for calculating the product H^TCH and extra $O(k^3)$ for eigen-decomposition (negligible for small k). Each iteration of WPCMS has the same complexity as PCMS, plus $O(n^2)$ operations for computing Hadamard products with matrix W. Finally, the complexity of a PoisMS iteration is equivalent to WPCMS one and demands $O(n^2)$ extra operations for calculating the second order approximation.

7 Reconstruction results calculated for chromosome 21

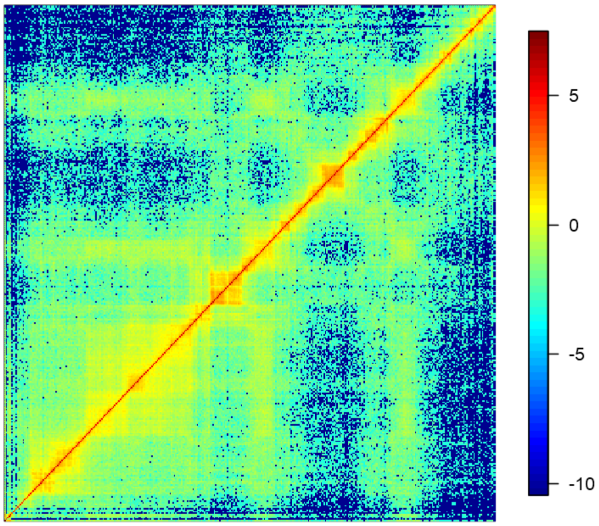
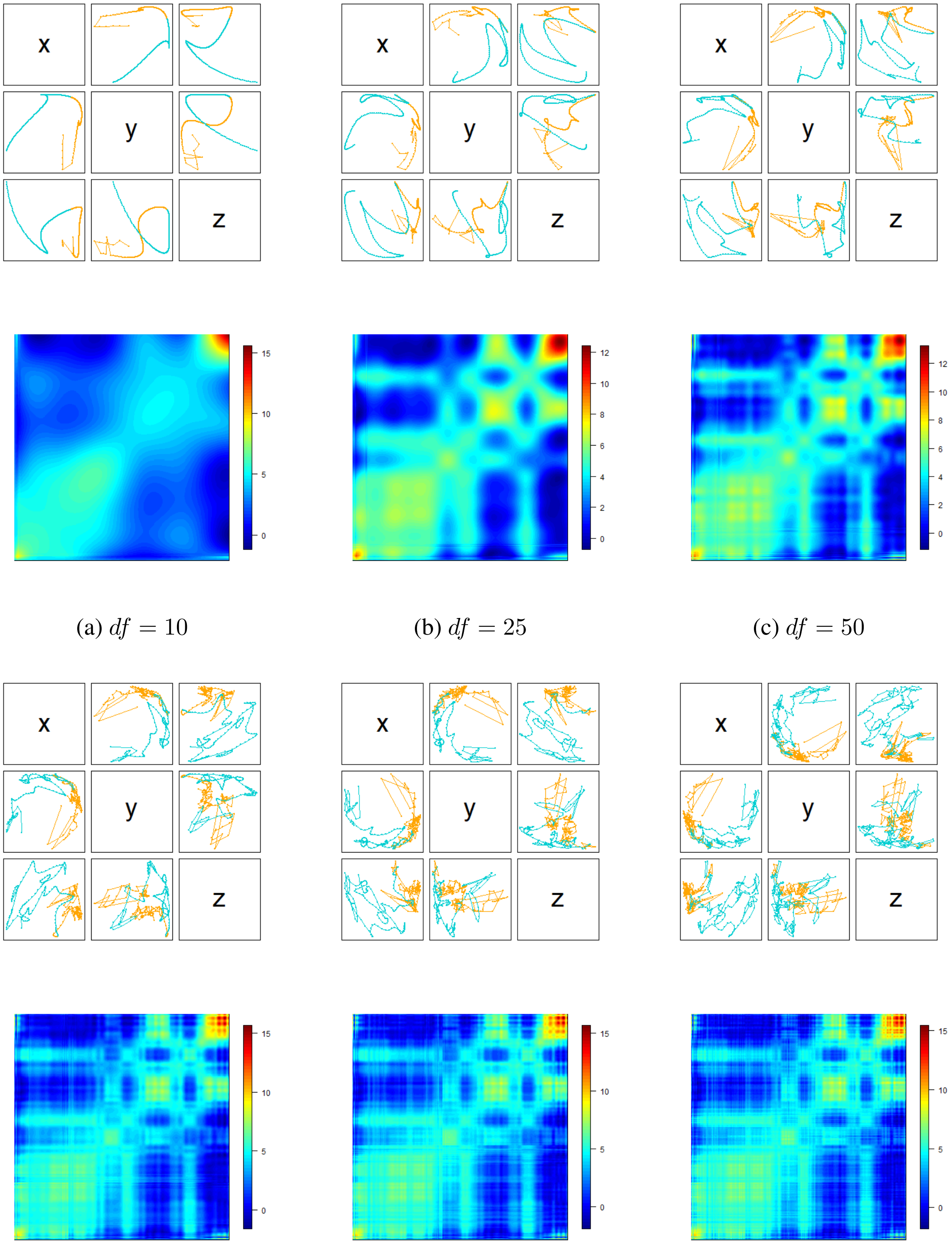


Figure 7: Log-transformed contact matrix $C^{\log} = \frac{\log(C+\epsilon)-\beta}{\alpha}$ for $\epsilon = 0.001$, $\alpha = 1$ and $\beta = \log\left(\frac{\sum_{i,j=1}^n C_{ij}}{n}\right)$.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) $df = 10$

(b) $df = 25$

(c) $df = 50$

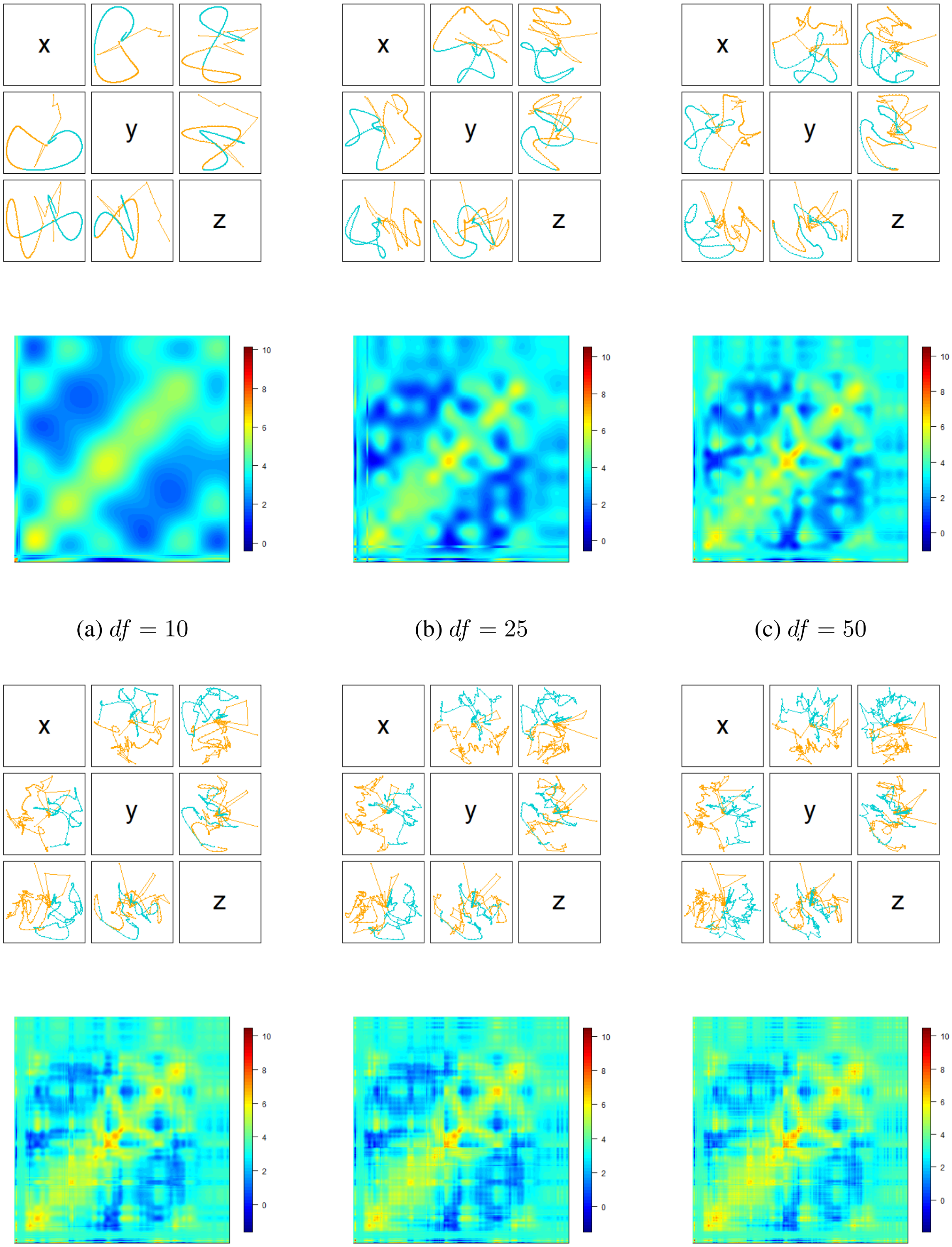
(d) $df = 100$

(e) $df = 150$

(f) $df = 200$

Figure 8: $\hat{X}_{df}$, the projections of the resulting reconstruction, and $\alpha\hat{C}_{df} + \beta$, the approximation of $\log(C)$, obtained via PCMS for different degrees of freedom values df.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) $df = 10$

(b) $df = 25$

(c) $df = 50$

(d) $df = 100$

(e) $df = 150$

(f) $df = 200$

Figure 9: $\hat{X}_{df}$, the projections of the resulting reconstruction, and $\alpha\hat{C}_{df} + \beta$, the approximation of $\log(C)$, obtained via PoisMS for different degrees of freedom values df.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

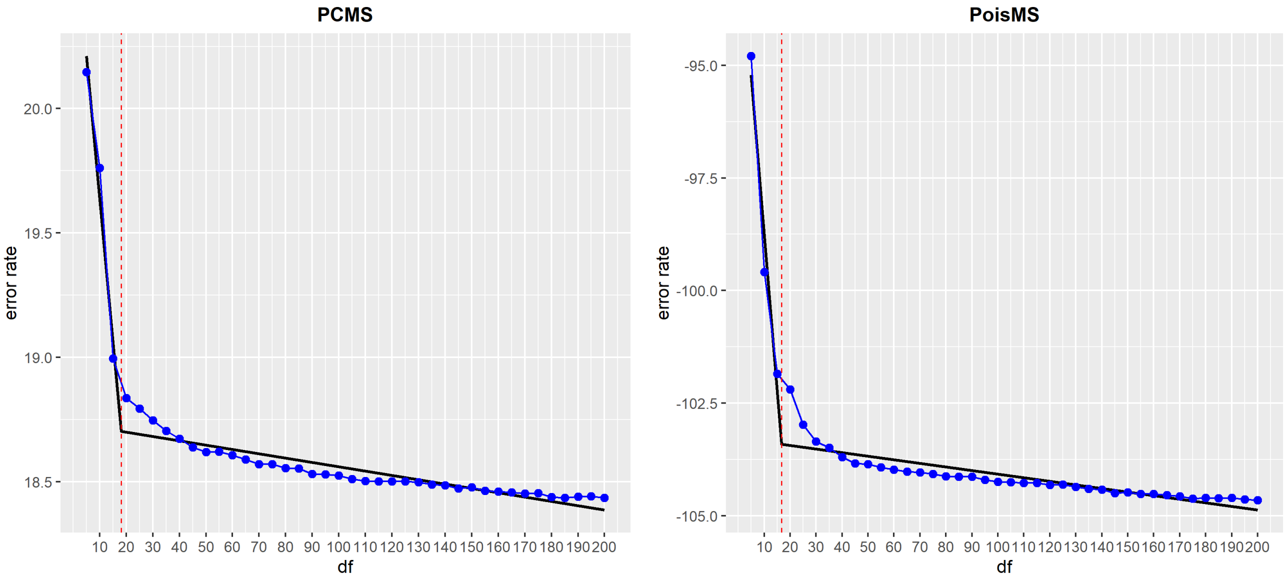


Figure 10: Error rate, here $err(\hat{C}_{df}) = \frac{1}{n_{obs}^2}\|C^{\log} - \hat{C}_{df}\|_F^2$ for PCMS and $err(\hat{C}_{df}) = \frac{1}{n_{obs}^2}\sum_{1\leq i,j\leq n}\lambda_{ij} - C_{ij}\log\lambda_{ij}$ with $\log\Lambda = \alpha\hat{C}_{df} + \beta$ for PoisMS, vs. degrees-of-freedom, here df. The segmented regression is given by the piecewise linear fit (black) with the degrees-of-freedom selected via kink estimation indicated by the red vertical line. Segmentation change point indicated for PCMS is $df = 18.1$ and for PoisMS is $df = 16.62$. Note that the optimal df value obtained via this method tends to be conservative.

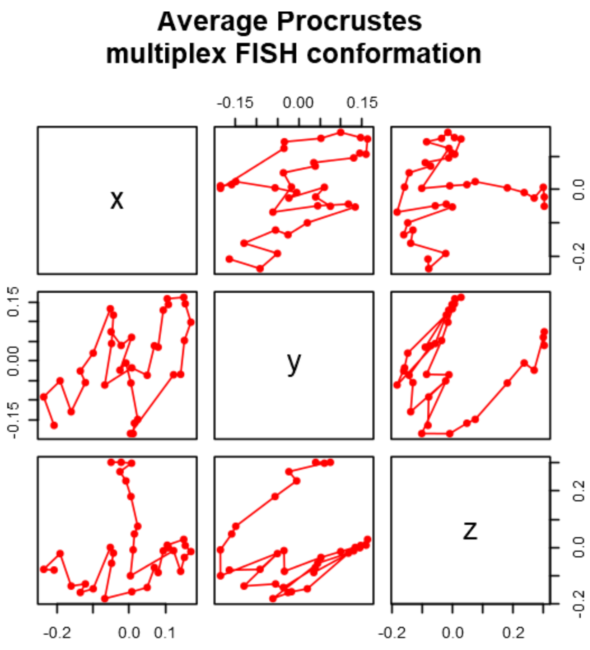


Figure 11: The average Procrustes conformation computed for 111 multiplex FISH replicate conformations for chromosome 21.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.15.151688; this version posted June 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

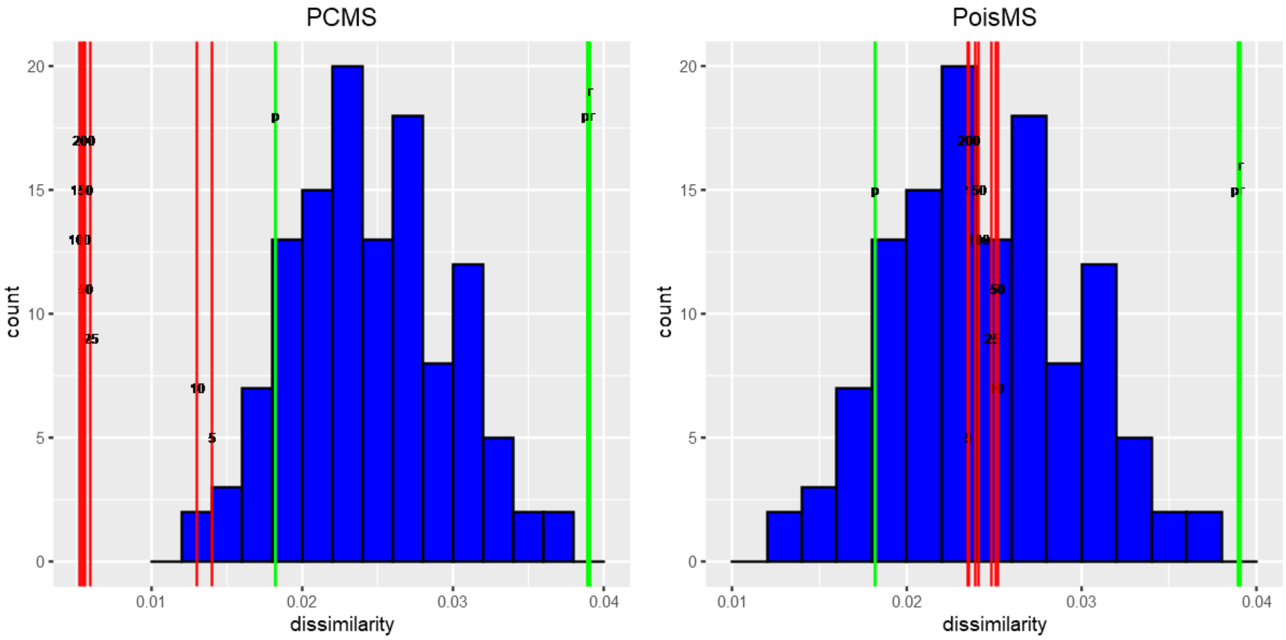


Figure 12: Reference distribution measuring the dissimilarity between the gold standard $\bar{X}$ and 111 multiplex FISH replicate conformations X_i^0 for chromosome 21. The vertical red lines correspond to the dissimilarity between $\bar{X}$ and the low-resolution reconstruction $\hat{X}_{df}$ calculated for different df values; the green lines represent three types of the HSA reconstruction.