

SCUDDO: Single-cell Clustering Using Diagonal Diffusion Operators

Luka Maisuradze^{1,*}, Mark D. Shattuck², and Corey S. O’Hern^{3,4}

¹Department of Molecular Biophysics and Biochemistry, Yale University

²Benjamin LeVich Institute and Physics Department, The City College of New York

³Department of Mechanical Engineering, Yale University

⁴Graduate Program in Computational Biology and Biomedical Informatics, Yale University

*Corresponding author: lgmaisura@gmail.com

Abstract

Motivation: Advances in high-throughput chromatin conformation capture have provided insight into the three-dimensional structure and organization of chromatin. While bulk Hi-C experiments capture spatio-temporally averaged chromatin interactions across millions of cells, single-cell Hi-C experiments report on the chromatin interactions of individual cells. Supervised and unsupervised algorithms have been developed to embed single-cell Hi-C maps and identify different cell types. However, single-cell Hi-C maps are often difficult to cluster due to their high sparsity, with state-of-the-art algorithms achieving a maximum Adjusted Rand Index (ARI) of only $\lesssim 0.4$ on several datasets while requiring labels for training.

Results: We introduce a novel unsupervised algorithm, Single-cell Clustering Using Diagonal Diffusion Operators (SCUDDO), to embed and cluster single-cell Hi-C maps. We evaluate SCUDDO on three previously difficult-to-cluster single-cell Hi-C datasets, and show that it can outperform other current algorithms in ARI by $\gtrsim 0.2$. Further, SCUDDO outperforms all other tested algorithms even when we restrict the number of intrachromosomal maps for each cell type and when we use only a small fraction of contacts in each Hi-C map. Thus, SCUDDO can capture the underlying latent features of single-cell Hi-C maps and provide accurate labeling of cell types even when cell types are not known *a priori*.

Availability: SCUDDO is freely available at www.github.com/lmaisuradze/scuddo. The tested datasets are publicly available and can be downloaded from the Gene Expression Omnibus.

1 Introduction

Elucidating the structure and dynamics of chromatin in cell nuclei is essential for understanding numerous cellular processes such as DNA transcription and replication [1]. Advances in whole-genome analyses, e.g. chromosome conformation capture techniques such as Hi-C, have provided important insights into long-ranged chromatin interactions and hierarchical chromatin organization [2]. Hi-C experiments provide chromatin contact maps, often represented as a symmetric matrix $\mathcal{A}$, where $\mathcal{A}_{ij}$ gives the number of times that loci i and j of chromatin come into close proximity. Bulk Hi-C contact maps provide information on chromatin fragment interactions averaged over millions of cells. In contrast, *single-cell* Hi-C maps give the frequency of chromatin contacts in each individual cell.

It is now well established that chromatin structure and organization can differ significantly across cell populations [1, 3]. Transcription analyses and imaging studies have shown that gene expression profiles and cell morphology can differ even between genetically identical cells [3, 4, 5].

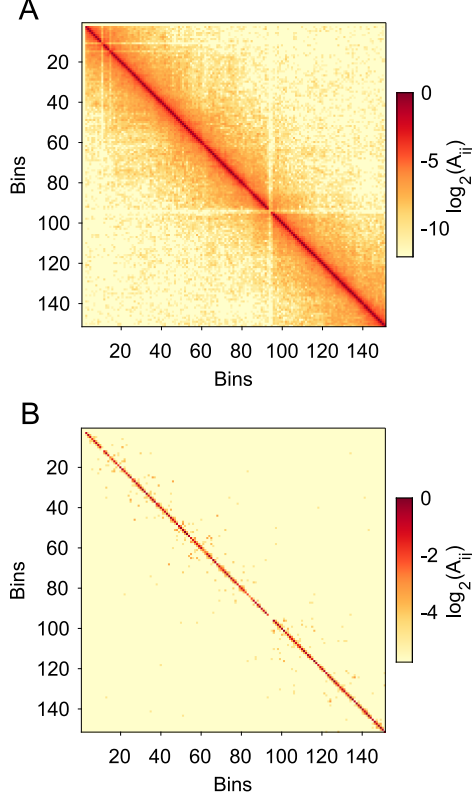


Figure 1: (A) A pseudo-bulk Hi-C map ($\log_2 \mathcal{A}_{ij}$) for chromosome 5 using pooled mouse oocyte cells before division from the Collombet [25] dataset (normalized so that $\max(\mathcal{A}_{ij}) = 1$). (B) An example single-cell Hi-C map from the Collombet dataset for chromosome 5 of a mouse oocyte cell using the same normalization.

In addition, the size and location of chromatin loops and topologically associating domains (TADs) can vary between the Hi-C maps of individual cells for a given organism [6, 7, 8, 9]. As a result, the loci that possess high contact frequencies in bulk Hi-C maps can differ from those that are in close spatial proximity in fluorescence in situ hybridization (FISH) experiments, in part due to the heterogeneity in chromatin structure across individual cells [10, 11, 12, 13, 14]. Thus, bulk Hi-C maps cannot be used to capture the structure and organization of chromatin in individual cells.

Several single-cell Hi-C technologies have been developed to capture chromatin interactions for large numbers of individual cells [15, 16, 17, 18, 19]. Single-cell Hi-C techniques enable studies of genome organization in individual cells, as well as comparisons of chromatin structure and organization across different cell types. Using data from single-cell Hi-C experiments, computational studies have focused on TAD, loop, and compartment identification for individual cells within a population [18, 19]. However, despite the rapid advances in genome-wide assays, single-cell Hi-C maps are still sparse, only capturing a fraction of the interactions that are obtained in bulk Hi-C experiments [20, 21]. For example, in Fig. 1, we show a pseudo-bulk Hi-C map for chromosome 5 in mouse oocyte cells [25] and compare it to a single-cell Hi-C map for the same chromosome and cell type. The single-cell Hi-C map shows significant sparsity, with most of the off-diagonal elements having a count of 0, as well as large variability for elements near the diagonal.

While techniques like fluorescence-activated cell sorting can be used to label single cells during chromosome conformation capture methods, these techniques are more expensive, lower throughput, and not as widely available as single-cell Hi-C experiments. Thus, the development

of classification algorithms for single-cell Hi-C maps may enable researchers to identify the key chromatin interactions that distinguish different cell types. Algorithms developed for bulk Hi-C analysis, including topologically associating domain callers [33], often work with limited efficacy on raw single-cell Hi-C data. Thus, due to their inherent variability and sparsity, specialized algorithms must be developed to identify robust features in single-cell Hi-C maps. In this article, we focus on the specific task of classifying single-cell Hi-C maps based on the cell labels that have been provided by the experimental studies. Most algorithms for clustering single-cell Hi-C maps use dimensionality reduction, treating each single-cell Hi-C map as a point in high-dimensional space and then mapping each point to a lower-dimensional space to cluster the data [20, 21, 22]. Despite the fact that there are more than a dozen algorithms to date for clustering single-cell Hi-C maps, there are many single-cell Hi-C datasets for which these methods achieve a maximum adjusted Rand index $\text{ARI} \lesssim 0.4$ [21, 36]. Moreover, there are many cases where one clustering method performs well on one single-cell Hi-C dataset, but then performs poorly on another dataset [20, 21], suggesting that current methods have trouble identifying features that generalize across multiple single-cell Hi-C datasets for clustering.

We develop a novel algorithm, SCUDDO (single-cell clustering using diagonal diffusion operators), which is fully unsupervised, fast, and easy to interpret to separate single-cell Hi-C maps into distinct groups. We then compare the predicted labels of the single-cell Hi-C maps to the cell types that are provided by experimental studies. To quantify the accuracy of the clustering, we calculate the ARI and normalized mutual information (NMI) using the predicted and ground truth labels. We find that SCUDDO outperforms current state of the art methods on three difficult-to-cluster single-cell Hi-C datasets, achieving an ARI and NMI greater than those for all of the tested methods on each of the datasets. We also find that SCUDDO achieves higher accuracy for clustering single-cell Hi-C maps compared to other algorithms when using only a fraction of the number of intrachromosomal maps and a fraction of the diagonals in each map.

The remainder of the manuscript is organized as follows. In the Materials and Methods section, we describe the key elements and hyperparameters of SCUDDO for clustering single-cell Hi-C maps. We then describe the three difficult-to-cluster datasets for benchmarking the new algorithm and the two metrics (ARI and NMI) for quantifying the clustering accuracy. In the Results section, we provide the ARI and NMI scores for SCUDDO and three current methods for clustering single-cell Hi-C maps on each of the three difficult-to-cluster Hi-C datasets. We also show SCUDDO’s performance across different hyperparameter regimes and when limiting the number of diagonals and intrachromosomal maps sampled. In the Discussion, we include some interpretations of the results, our conclusions, and promising future research directions.

2 Materials and Methods

The Materials and Methods is organized into three subsections. We first define the necessary notation and summarize the steps of the SCUDDO method to cluster single-cell Hi-C maps. Second, we describe three difficult-to-cluster single-cell Hi-C datasets that will be used to benchmark SCUDDO alongside three other current algorithms. Finally, we define the two metrics, ARI and NMI, which are used to quantify the unsupervised clustering accuracy.

2.1 SCUDDO algorithm

The SCUDDO algorithm takes as input a set of intrachromosomal Hi-C maps with nonnegative integer entries for a cells, each with b chromosomes, totaling $a \times b$ intrachromosomal Hi-C maps. As for bulk Hi-C maps, single-cell Hi-C maps are represented as symmetric matrices with elements $\mathcal{A}_{ij}$ that represent the number of contacts between loci i and j on chromatin. To distinguish between the cell and chromosome indices, we define $\mathcal{A}_{s,ij}^k$ as the ij th element of the $n^k \times n^k$ Hi-C map for chromosome k of cell s . n^k only depends on k since the dimensions of the Hi-C map only vary across different chromosomes. Given a set of intrachromosomal matrices,

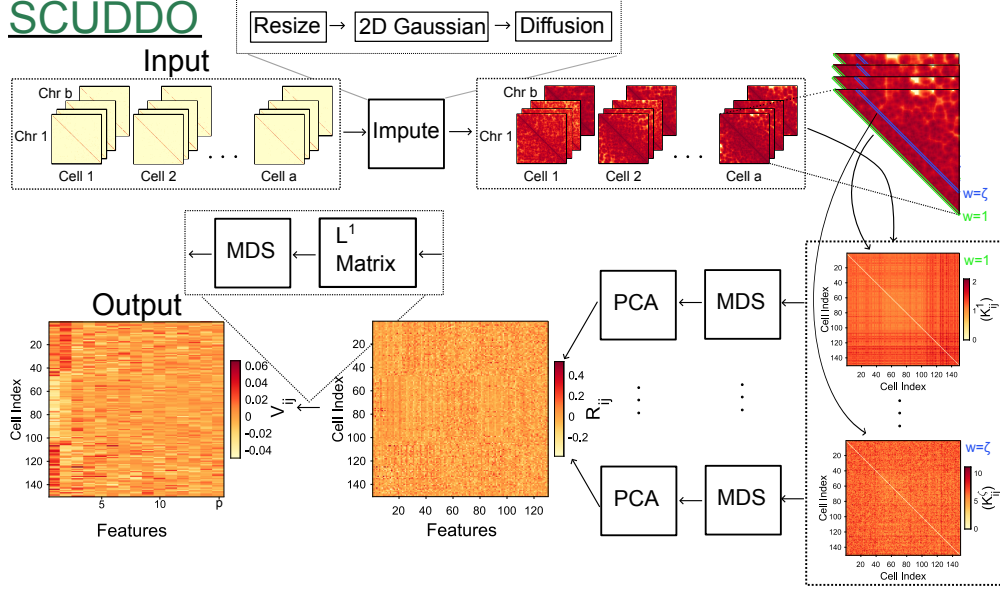


Figure 2: A schematic of the SCUDDO algorithm for clustering single-cell Hi-C maps. We illustrate the method using intrachromosomal Hi-C maps from the Li, et al. dataset [25]. SCUDDO first imputes the set of intrachromosomal Hi-C matrices (indexed by $\mathcal{A}_s^k$) for each cell and then samples each diagonal (indexed by w) from each Hi-C matrix to form a feature matrix $\mathcal{K}^w$ for each sampled diagonal. Principal component analysis (PCA) and nonmetric multi-dimensional scaling (MDS) are then applied to each feature matrix to form the matrix $\mathcal{R}$, which is then embedded in a lower dimensional latent space using the L^1 norm to form the embedding, $\mathcal{V}$.

SCUDDO returns a low-dimensional embedding of the Hi-C matrices. This embedding is then used as an input into a clustering algorithm, for example K-means++ [26], where each cell is assigned to one of l predicted labels.

SCUDDO starts by pre-processing and performing imputation on each intrachromosomal matrix $\mathcal{A}_s^k$. First, each $\mathcal{A}_s^k$ is reshaped into the same size $r \times r$ matrix, $\mathcal{A}_s'^k$, using a bicubic interpolation kernel, where $r = \sum_{k=1}^b n^k / b$. SCUDDO then convolves each intrachromosomal matrix with a Gaussian kernel:

$$\mathcal{A}_{s,ij}'^k = \sum_{\chi=1}^9 \sum_{\omega=1}^9 \mathcal{G}_{\chi\omega} \mathcal{A}_{s,(i-4+\chi)(j-4+\omega)}^k, \quad (1)$$

where $\mathcal{G}$ is a two-dimensional 9×9 Gaussian kernel with standard deviation $\sigma = 0.5$ that uses replicate padding, where values outside of the bounds of the original Hi-C map are set to the values of the nearest border entry. $\mathcal{G}$ smooths local regions in each individual intrachromosomal matrix. The final pre-processing step is to normalize each intrachromosomal matrix and apply a diffusion kernel via a matrix exponential:

$$\mathcal{B}_s^k = \exp \left(- \frac{\mathcal{A}_s'^k}{\sum_{ij} \mathcal{A}_{s,ij}'^k} \right), \quad (2)$$

which represents backwards diffusion over $\mathcal{A}_s'^k$. Next, we construct high-dimensional embeddings

of the intrachromosomal Hi-C maps for each cell. Let $d^w(\mathcal{B}_s^k)$ be the ordered set of Hi-C map entries on the w th superdiagonal of the $r \times r$ matrix $\mathcal{B}_s^k$:

$$d^w(\mathcal{B}_s^k) = \{\mathcal{B}_{s,1,1+w}^k, \mathcal{B}_{s,2,2+w}^k, \dots, \mathcal{B}_{s,r-w,r}^k\}. \quad (3)$$

For a given w , $d^w(\mathcal{B}_s^k)$ for each chromosome k for cell s is concatenated to form the embedding vector:

$$\vec{e}_s^w = \{d^w(\mathcal{B}_s^1), d^w(\mathcal{B}_s^2), \dots, d^w(\mathcal{B}_s^b)\}, \quad (4)$$

where $\alpha = 1, \dots, b(r-w)$ indexes the entries in $\vec{e}_s^w$. Every embedding vector, $\vec{e}_s^w$, is z-score normalized such that

$$\vec{e}_s'^w = \frac{\vec{e}_s^w - \mu}{\sqrt{\frac{1}{b(r-w)-1} \sum_{\alpha=1}^{b(r-w)} |\vec{e}_s^w - \mu|^2}}, \quad (5)$$

where $\mu = \sum_{\alpha=1}^{b(r-w)} (\vec{e}_s^w)_\alpha / [b(r-w)]$. This pooling approach is similar to previous work [36, 37] that employs band normalization for Hi-C matrices. Next, each embedding vector is transformed into a signed difference vector:

$$\vec{f}_s^w = \text{sgn}(\nabla(\vec{e}_s'^w)), \quad (6)$$

where $\nabla(\vec{e}_s'^w)_\alpha = (\vec{e}_s'^w)_\alpha - (\vec{e}_s'^w)_{\alpha+1}$ and sgn is the sign function. (Note that we set the last entry of the chromosome difference vector $(\vec{f}_s^w)_{b(r-w)} = (\vec{e}_s'^w)_{b(r-w)}$.) Eq. 6 transforms $\vec{e}_s'^w$ into a ternary vector with values 1, 0, or -1 . For a given w , each cell's difference vector, $\vec{f}_1^w, \vec{f}_2^w, \dots, \vec{f}_a^w$ is used to calculate the distance matrix between cells i and j using cosine similarity:

$$D_{ij}^w = 1 - \frac{\vec{f}_i^w \cdot \vec{f}_j^w}{|\vec{f}_i^w| |\vec{f}_j^w|}, \quad (7)$$

where $|\vec{X}|$ indicates the magnitude of $\vec{X}$. A separate distance matrix is calculated for $\vec{e}_s'^w$:

$$D'_{ij}^w = 1 - \frac{\vec{e}_i'^w \cdot \vec{e}_j'^w}{|\vec{e}_i'^w| |\vec{e}_j'^w|}, \quad (8)$$

and combined to form a final exponentiated distance matrix using element-wise exponentiation:

$$\mathcal{K}_{ij}^w = e^{(D_{ij}^w + D'_{ij}^w)(D_{ij}^w D'_{ij}^w)}. \quad (9)$$

Finally, SCUDDO uses nonmetric multidimensional scaling (MDS) [32] to transform the $a \times a$ matrix $\mathcal{K}^w$ into a lower dimensional representation, i.e. an $a \times p$ matrix where $p < a$, which preserves the distances in $\mathcal{K}^w$. The multidimensional scaling is followed by principal component analysis to further reduce the dimension to an $a \times q$ matrix $\mathcal{U}^w$, where $q < p$ ($p = 30$ and $q = 5$). This procedure is performed for the diagonal ($w = 0$) and a given number of superdiagonals ($w = \zeta > 0$), and each set of dimensionality-reduced representations are concatenated, forming the $a \times (q(\zeta + 1))$ matrix $\mathcal{R} = \mathcal{U}^0, \mathcal{U}^1, \dots, \mathcal{U}^\zeta$. $\mathcal{R}$ is then normalized feature-wise using the softmax function, $\mathcal{R}'_{ij} = e^{\mathcal{R}_{ij}} / \sum_{\theta=1}^a e^{\mathcal{R}_{\theta j}}$, and a distance matrix is constructed using the L^1 metric:

$$\mathcal{S}_{ij} = \sum_{\lambda=1}^{qw} |\mathcal{R}'_{i\lambda} - \mathcal{R}'_{j\lambda}|. \quad (10)$$

Another round of dimensionality reduction is performed using multidimensional scaling to reduce the dimension of $\mathcal{S}$ to the embedding size ϵ , which gives the $a \times \epsilon$ matrix, $\mathcal{V}$. Because there is no guarantee of convexity associated with each cluster when clustering single cell Hi-C matrices, we use spectral decomposition before performing the clustering. In particular, SCUDDO transforms $\mathcal{V}$ into the similarity matrix, $\mathfrak{A}_{ij} = e^{-\mathcal{Z}_{ij}^2}$, where $\mathcal{Z}_{ij} = |\vec{\mathcal{V}}_{i*} - \vec{\mathcal{V}}_{j*}|$ and $\mathcal{V}_{i*}$ is the vector consisting of all elements in the i^{th} row of $\mathcal{V}$. Next, we calculate the final $a \times l$ spectral embedding $\mathcal{C}$, where the columns of $\mathcal{C}$ are the smallest l eigenvectors of the random-walk Laplacian matrix constructed from $\mathfrak{A}$ using the Shi-Malik algorithm [38] with $\log(a)$ nearest neighbors. We then

input $\mathcal{C}$ and the number of labels l into a clustering algorithm, such as K-means++ [26], which returns the predicted labels for each single-cell Hi-C map.

SCUDDO includes two tunable hyperparameters: ζ , the set of (super)diagonals, $w = 0, 1, \dots, \zeta$ used to construct the embedding vectors, and the dimension ϵ , to which $\mathcal{V}$ is reduced. By default, SCUDDO outputs two embeddings: $\mathcal{C}$ and $\mathcal{V}$. Both ζ and ϵ are varied in the Results section to study their effects on SCUDDO’s performance for each dataset. For all results in this study unless otherwise noted, we use $\mathcal{C}$ as the input into K-means++ and set $\zeta = 25$ and $\epsilon = 5$. Our results are not sensitive to the values of the dimensions p and q .

2.2 Benchmarking of single-cell Hi-C clustering algorithms

We focus our studies on three difficult-to-cluster datasets of single-cell Hi-C maps from recent benchmarking studies [21, 31]. In particular, we consider the Li, et al. [16] dataset (GEO ID: GSE119171) consisting of $a = 150$ mouse embryonic stem cells that are separated into $l = 3$ labels: “2i”, “Serum1”, and “Serum2”, the Flyamer, et al. [18] dataset (GEO ID: GSE80006) consisting of $a = 134$ cells from developing mouse zygotes and oocytes with $l = 3$ cell types: “Oocyte”, “ZygP”, and “ZygM” as labels, and the Collombet, et al. [25] dataset (GEO ID: GSE129029) consisting of $a = 648$ mouse embryo cells with labels that represent $l = 5$ different cell stages: 1-cell, 2-cell, 4-cell, 8-cell, and 64-cell stages. In previous benchmarking studies [21], none of the eight tested methods for single-cell Hi-C map clustering achieved ARI or NMI ≥ 0.6 on the Collombet, et al. dataset and in another study [36] none of the eight methods tested achieved an ARI > 0.45 on the Li, et al. dataset across any clustering algorithm (not just k-means). For each dataset, we use 1 Mb bin sizes for the single-cell Hi-C maps, and re-bin those with higher resolution, as discussed in Zhou, et al. [30]. If the sum of all non-diagonal nonzero pairs of elements in the intrachromosomal Hi-C maps for a given cell is less than 5000, the data for this cell was not included in the analysis. Also, for each individual chromosome of size x for a cell, if the intrachromosomal Hi-C map for that chromosome has a sum of non-diagonal contacts that is less than x , all intrachromosomal Hi-C maps are not considered for that cell.

After considering previous single-cell Hi-C map clustering studies [21, 29, 31, 36], we selected consistent top performers across several datasets to compare with SCUDDO: i.e. the Higashi [29], HiCRep/MDS [35, 22], and scHiCluster [30] algorithms. While HiCRep/MDS is not as accurate as Higashi and scHiCluster, we include it in our analysis since it is the most widely used and best performing method that uses MDS similar to SCUDDO to the best of our knowledge. Importantly, all algorithms that we tested are unsupervised or self-supervised, and do not require labels for training. Other algorithms that require labels or significant pretraining are unable to cluster unlabeled single-cell Hi-C datasets, and thus they are not included in this manuscript. For each algorithm, we used the default hyperparameters and used the final embeddings (with no further processing) as input into K-means++ clustering to benchmark our calculations.

2.3 Metrics for clustering accuracy

To assess the accuracy of the predicted labels, we calculate the adjusted rand index (ARI) [27] and normalized mutual information (NMI) [28]. Let $\Omega_T(s)$ and $\Omega_G(s)$ be functions that map each cell index s (from 1 to a) to the integers l' and l'' respectively, where l' is the ground truth label for cell s and l'' is the predicted label for cell s . We then define $P_T = \{X_1, X_2, \dots, X_l\}$ as the “ground-truth” label set, where $X_{l'}$ denotes the set of cells such that $\Omega_T(s) = l'$, and $P_G = \{Y_1, Y_2, \dots, Y_l\}$ as the “predicted” label set, where $Y_{l''}$ is the set of cells such that $\Omega_G(s) = l''$.

The adjusted Rand index determines the similarity between the sets of cells with given ground truth and predicted labels:

$$\text{ARI} = \frac{\sum_{i=1}^l \sum_{j=1}^l \binom{\beta_{ij}}{2} - (\sum_{i=1}^l \binom{\Gamma_i}{2}) \sum_{j=1}^l \binom{\Delta_j}{2} / \binom{a}{2}}{\frac{1}{2} (\sum_{i=1}^l \binom{\Gamma_i}{2} + \sum_{j=1}^l \binom{\Delta_j}{2}) - (\sum_{i=1}^l \binom{\Gamma_i}{2}) \sum_{j=1}^l \binom{\Delta_j}{2} / \binom{a}{2}}, \quad (11)$$

where $\beta_{ij} = |X_i \cap Y_j|$, $\cap$ is the intersection between two sets, $|X|$ is the number of elements in set X , $\Gamma_k = \sum_{i=1}^l \beta_{ki}$, $\Delta_k = \sum_{j=1}^l \beta_{jk}$, and $\binom{m}{n} = \frac{m!}{n!(m-n)!}$. ARI = 1 indicates a perfect match

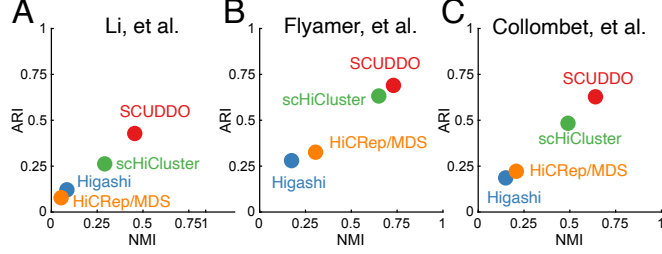


Figure 3: Accuracy of the four single-cell Hi-C map clustering algorithms (Higashi [29] (blue), HiCRep/MDS [35, 22] (orange), scHiCluster [30] (green), and SCUDDO (red)) on three difficult-to-cluster single-cell Hi-C datasets. We plot the adjusted Rand index (ARI) versus the normalized mutual information (NMI) for each algorithm on the (A) Li, et al. [16], (B) Flyamer, et al. [18], and (C) Collombet, et al. [25] datasets.

between P_T and P_G , whereas $\text{ARI} = 0$ indicates the match between P_T and P_G is no better than that achieved by random assignments in P_G .

We also quantify the accuracy of the clustering of the single-cell Hi-C maps using the normalized mutual information (NMI). NMI measures how much information can be learned about a given clustering by observing a different, but related clustering. NMI is defined as:

$$\text{NMI} = \frac{\sum_{i=1}^l \sum_{j=1}^l \mathcal{H}(i, j) \log_2 \frac{\mathcal{H}(i, j)}{\mathcal{H}(i)\mathcal{H}(j)}}{\sqrt{(-\sum_{i=1}^l \mathcal{H}(i) \log_2 \mathcal{H}(i))(-\sum_{j=1}^l \mathcal{H}(j) \log_2 \mathcal{H}(j))}}, \quad (12)$$

where $\mathcal{H}(i) = \frac{|X_i|}{a}$, $\mathcal{H}(j) = \frac{|Y_j|}{a}$, and $\mathcal{H}(i, j) = \frac{|Y_i \cap X_j|}{a}$. $0 < \text{NMI} < 1$, where $\text{NMI} = 1$ indicates that $P_T = P_G$ and $\text{NMI} = 0$ indicates that there is no correlation between P_T and P_G . We calculate both ARI and NMI since they can differ for different sized clusters: ARI is preferable when the sets in P_T are similar in size, whereas NMI is preferable when the sets in P_T are unbalanced. For all datasets and algorithms, we calculate the ARI and NMI after using the native embedding and K-means++ clustering.

3 Results

We carried out single-cell Hi-C map clustering on three difficult-to-cluster datasets (Collombet, et al. [25], Flyamer, et al. [18], and Li, et al. [16]) using three current algorithms (Higashi [29], HiCRep/MDS [21], and scHiCluster [30]) and compared the results to those obtained from SCUDDO. We plot ARI versus NMI for each dataset and algorithm in Fig. 3 (A)-(C). Overall, SCUDDO outperforms the other three methods for all datasets tested. For the Li, et al. dataset in Fig. 3 (A), we find a significant separation in accuracy between SCUDDO and the next most accurate method (scHiCluster). SCUDDO achieves an $\text{ARI} \sim 0.45$ and $\text{NMI} \approx 0.42$, while scHiCluster achieves $\text{ARI} \approx 0.29$ and $\text{NMI} \approx 0.25$. For the Flyamer, et al. dataset in Fig. 3 (B), we find that SCUDDO has $\text{ARI} \sim \text{NMI} \approx 0.73$, whereas the next most accurate method, again SciHiCluster, has $\text{ARI} \sim \text{NMI} \approx 0.65$. We also find that on some runs of K-means++ for this dataset, SCUDDO can achieve $\text{ARI} \geq 0.90$. Lastly, for the Collombet, et al. dataset in Fig. 3 (C), we find that SCUDDO has $\text{ARI} \sim \text{NMI} \approx 0.64$, whereas the next most accurate method, again SciHiCluster, has $\text{ARI} \sim \text{NMI} < 0.5$.

We next show that SCUDDO can accurately embed single-cell Hi-C maps using a reduced amount of information for already highly sparse single-cell Hi-C maps, surpassing the accuracy of previous algorithms using fewer intrachromosomal matrices for each cell, as well as fewer sampled superdiagonals for each matrix. We study the performance of SCUDDO after restricting the single-cell Hi-C data available to it in two ways: first by varying the hyperparameter ζ for the number of superdiagonals to sample for each intrachromosomal Hi-C map, as well as varying

the hyperparameter ϵ for the embedding dimension of $\mathcal{V}$. In Fig. 4, we show heatmaps of the ARI and NMI for the SCUDDO algorithm, while varying $0 \leq \zeta \leq 40$ and $1 \leq \epsilon \leq 40$. The pixels in the ζ - ϵ plane outlined in black indicate ARI or NMI values for the SCUDDO algorithm that exceed those for all other methods (when they use all of the available single-cell Hi-C data).

For the Li, et al. dataset, we show in the left column of Fig. 4 (A) and (B) that for $\approx 84\%$ and 82% of the ζ - ϵ plane SCUDDO outperforms all methods in ARI and NMI. In particular, when $\zeta > 2$, SCUDDO gives mean ARI and NMI values over all ϵ (including low-dimensional embeddings) that match the ARI and NMI for the next best method (when they use all available single-cell Hi-C data). Similarly, when $\epsilon > 1$, SCUDDO gives mean ARI and NMI values over all ζ that exceed the values for all other methods. Even in regimes with low diagonal sampling and embedding dimension, SCUDDO can obtain ARI values that surpass the next best method (e.g. $\zeta = 3, \epsilon = 2$). The maximum ARI and NMI for clustering the Li, et al. dataset using SCUDDO in the sampled hyperparameter space were ≈ 0.48 and ≈ 0.47 respectively.

On the Flyamer, et al. dataset, SCUDDO outperforms the other methods over a more restricted region of the hyperparameters ζ and ϵ , as shown in the middle column of Fig. 4 (A) and (B), with $\approx 50\%$ and $\approx 43\%$ of (ζ, ϵ) input pairs into SCUDDO resulting in ARI and NMI scores that surpassed the next best method’s ARI and NMI scores respectively. We find that unlike the other two datasets, SCUDDO requires generally higher ζ values (more sampled diagonals) to perform state of the art for the Flyamer, et al. dataset. We find that when $\zeta > 16$ (across all ϵ) and when $\epsilon > 16$ (across all ζ), SCUDDO achieves a larger mean ARI and NMI than the next best method. SCUDDO also achieves exceptional accuracy at $\epsilon = 5, \zeta = 3$ and $\epsilon = 7, \zeta = 6$ with ARI ≈ 0.93 and 0.94 and NMI ≈ 0.80 and 0.81 respectively.

For the Collombet, et al. dataset in the right column of Fig. 4 (A) and (B), SCUDDO outperforms the next best method in ARI and NMI over $\approx 87\%$ and $\approx 90\%$ of the ζ - ϵ plane. For $\zeta > 3$, the mean ARI and NMI across all ϵ values for SCUDDO is larger than the other tested methods. Similarly, SCUDDO outperforms the other methods in mean ARI and NMI when $\epsilon > 3$ (across all ζ). Across the sampled hyperparameters, we find that the maximum ARI and NMI are ≈ 0.66 and ≈ 0.67 respectively.

Previous single-cell clustering algorithms often require a large number of dimensions ($\epsilon \gtrsim 100$) to achieve reasonable clustering accuracy on single-cell Hi-C maps [30]. In addition, the ARI and NMI can possess large fluctuations as a function of the embedding dimension and depend strongly on the specific dimensionality reduction technique that is implemented, for instance with some methods requiring specific dimensionality reduction techniques to be competitive [29]. In contrast, we have shown that the ARI and NMI scores for the SCUDDO algorithm are large at both very low embedding dimensions and when sampling only a few superdiagonals. This result is true even when we treat ϵ as the final embedding dimension of the output for SCUDDO, despite the fact that SCUDDO always outputs a l dimensional embedding, where $l \leq \epsilon$ for all datasets studied. In addition, we find that SCUDDO does not depend sensitively on the specific dimensionality reduction technique. For instance on the Collombet, et al. dataset, SCUDDO performs roughly equivalently when $\mathcal{V}$ is embedded spectrally (i.e. the default embedding) (ARI $\approx$ NMI ≈ 0.64), embedded using UMAP [24] (ARI $\approx$ NMI ≈ 0.60), embedded using t-SNE [23] (ARI $\approx$ NMI ≈ 0.57), and when there is no further dimensionality reduction and using $\mathcal{V}$ directly (ARI $\approx$ NMI ≈ 0.56). While the default values for ζ and ϵ for SCUDDO were not optimized to the three selected datasets, the default parameters give excellent results for ARI and NMI for these datasets. However, there are (ζ, ϵ) pairs, e.g. $\zeta = 26, \epsilon = 7$, that give superior performance across all datasets in this manuscript.

In Fig. 5, we calculate ARI and NMI for the three datasets versus the number of intrachromosomal maps b that we sample. For these calculations, we sample all chromosomes with an index less than or equal to b , for instance if we set $b = 4$, the SCUDDO algorithm samples only chromosomes 1, 2, 3, and 4. For the Li, et al. and Collombet, et al. datasets in the left and right panels, we find that the ARI and NMI for the SCUDDO algorithm first exceed those for the next best method when $b \gtrsim 4$ and 2 , respectively. However, for the Flyamer, et al. dataset in the center panel, most of the chromosomes are needed to achieve high accuracy, with comparable performance with the next best method at $b = 11$. We also note that the ARI and NMI for the

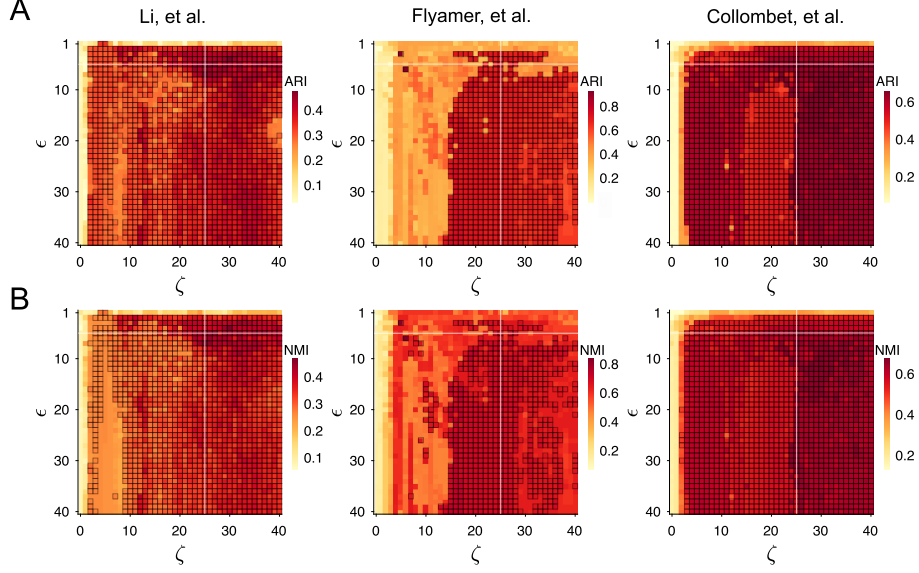


Figure 4: The clustering accuracy (ARI in (A) and NMI in (B)) for the SCUDDO algorithm for each of the three datasets, (left) Li, et al. [16], (middle) Flyamer, et al. [18], and (right) Collombet, et al. [25], plotted as a function of the number of sampled superdiagonals ζ and the embedding dimension ϵ . The pixels in the heatmap are outlined when the ARI or NMI for SCUDDO exceed those of the next best performing method, which is scHiCluster in all cases. The faint horizontal and vertical white lines in each heatmap indicate the row and column for the default values for the SCUDDO algorithm, $\zeta = 25$ and $\epsilon = 5$.

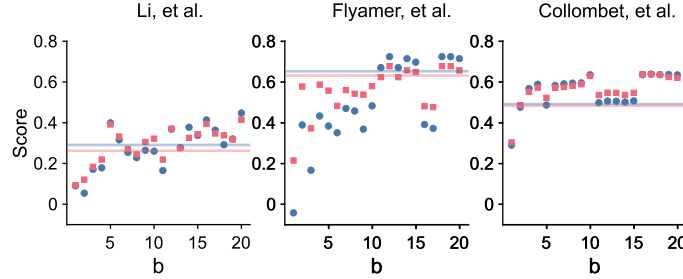


Figure 5: The clustering accuracy, ARI (red squares) and NMI (blue circles), for the SCUDDO algorithm plotted as a function of the number of sampled chromosomes b for the three datasets: (left) Li, et al. [16], (middle) Flyamer, et al. [18], and (right) Collombet, et al. [25]. The faint horizontal red and blue lines represent the values of ARI and NMI for the scHiCluster method, which is the next best performing method for these datasets.

Flyamer, et al. dataset fluctuates more than the values for the other datasets. For instance, there are large step changes in the ARI and NMI in Fig. 5 (B) for $b = 16$ and 17 .

4 Discussion

In this article, we develop a novel algorithm, SCUDDO, to determine a low-dimensional representation and then cluster single-cell Hi-C maps. We focused on three difficult-to-cluster single-cell Hi-C map datasets, where the datasets include ground-truth labels for each single-cell Hi-C map. We compared the ARI and NMI metrics for clustering accuracy from the SCUDDO algorithm to

those from three other clustering algorithms that were the most accurate in previous single-cell Hi-C map clustering benchmarking studies [21, 31]. The SCUDDO algorithm is for all cases more accurate in terms of both ARI and NMI compared to the other three methods for all datasets. We also find that the SCUDDO algorithm can accurately cluster single-cell Hi-C maps using a fraction of the intrachromosomal Hi-C maps and their diagonals, as well as fewer embedding dimensions.

The SCUDDO algorithm has several advantages over other existing methods for clustering single-cell Hi-C maps. First, SCUDDO uses, to our knowledge, a new and relatively simple imputation technique for single-cell Hi-C maps, smoothing over local neighborhood features in single-cell Hi-C map and then using backwards diffusion using a 2D Gaussian kernel followed by a diffusion operator. This technique improves the accuracy for datasets where there are few cells (e.g. the Li, et al. and Flyamer, et al. datasets), since it both short-range and long-range information transfer within an intrachromosomal Hi-C map. Additionally, SCUDDO mainly uses PCA and MDS for dimensionality reduction, both of which are much more interpretable than dimensionality reduction techniques like UMAP and t-SNE or using complex inscrutable networks that require training to find features.

The SCUDDO algorithm combines two key features: the diffused (normalized) diagonals of each single-cell Hi-C map and the trinarized differences along the diffused diagonals. The algorithm then calculates the angles between these features for each diagonal (using cosine similarity) and performs several steps of dimensionality reduction to achieve a final embedding. We find that for some datasets both features are necessary to achieve the best clustering accuracy, e.g. for the Flyamer et al. dataset using only one feature scores at best an ARI of only 0.5. While the details of the features and SCUDDO algorithm are easy to interpret mathematically, the biophysical interpretation of these features is less clear. For example, it is unclear whether the diffusion and smoothing steps used by SCUDDO have a clear biophysical interpretation.

Interesting future studies involve coupling molecular dynamics simulations of polymer models of chromosomes [34] with the SCUDDO algorithm to further improve clustering accuracy and to better understand the biophysical mechanisms that support the efficacy of the methods used in the SCUDDO algorithm. In addition, the SCUDDO algorithm can be applied to synthetic datasets with labels with tunable noise and sparsity, as well as to experimental datasets without labels to predict cell fate.

5 Competing interests

No competing interest is declared.

6 Author contributions statement

L.M developed and implemented the SCUDDO algorithm to embed and cluster single-cell Hi-C maps and wrote the first draft of the manuscript. M.D.S and C.S.O edited the manuscript and provided input on the methodology.

7 Funding

This work was supported by the National Science Foundation Grant Nos. 1830904 (L.M. and C.S.O.) and 2124558 (L.M. and C.S.O.).

8 Acknowledgments

We thank Parisa A. Vaziri for insightful comments.

References

- [1] Misteli T. The self-organizing genome: principles of genome architecture and function. *Cell*. 2020; 183(1): 28–45.
- [2] 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(5950):289–93.
- [3] Finn EH, Pegoraro G, Brandao HB, Valton AL, Oomen ME, Dekker J, Mirny L, and Misteli T. Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell*. 2019; 176: 1502–1515.
- [4] Shah S, Takei Y, Zhou W, Lubeck E, Yun J, Eng CL, Koulina N, Cronin C, Karp C, Liaw EJ, Amin M, Cai L. Dynamics and spatial genomics of the nascent transcriptome by intron seqFISH. *Cell*. 2018; 174: 363–376.
- [5] Rodriguez J, Ren G, Day CR, Zhao K, Chow CC, and Larson DR. Intrinsic dynamics of a human gene reveal the basis of expression heterogeneity. *Cell*. 2019; 176: 213–226.
- [6] Bintu B, Mateo LJ, Su JH, Sinnott-Armstrong NA, Parker M, Kinrot S, Yamaya K, Boettiger AN, and Zhuang X. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*. 2018; 362:6413.
- [7] Finn EH, and Misteli T. Molecular basis and biological function of variability in spatial genome organization. *Science*. 2018; 365:6457.
- [8] Cattoni DI, Cardozo Gizzi AM, Georgieva M, Di Stefano M, Valeri A, Chamousset D, Houbroun C, Dejardin S, Fiche JB, Gonzalez I, Chang J, Sexton T, Marti-Renom MA, Bantignies F, Cavalli G, Nollmann M. Single-cell absolute contact probability detection reveals chromosomes are organized by multiple low-frequency yet specific interactions. *Nat Commun*. 2017; 8:1753.
- [9] Mateo LJ, Murphy SE, Hafner A, Cinquini IS, Walker CA, and Boettiger AN. Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature*. 2019; 568:49–54.
- [10] Shi, G., Thirumalai, D. Conformational heterogeneity in human interphase chromosome organization reconciles the FISH and Hi-C paradox. *Nat Commun*. 2019; 10(1):3894.
- [11] Giorgetti L, Heard E. Closing the loop: 3C versus DNA FISH. *Genome Biol*. 2016; 17:215.
- [12] Fudenberg G, Imakaev M. FISHing for captured contacts: towards reconciling FISH and 3C. *Nat Methods*. 2017; 14(7):673–678.
- [13] Bickmore WA, van Steensel B. Genome architecture: domain organization of interphase chromosomes. *Cell*. 2013; 152(6):1270–1284.
- [14] Williamson I, Berlivet S, Eskeland R, Boyle S, Illingworth RS, Paquette D, Dostie J, Bickmore WA. Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes Dev*. 2014; 28(24):2778–2791.
- [15] 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.
- [16] Li G, Liu Y, Zhang Y, Kubo N, Yu M, Fang R, Kellis M, Ren B. Joint profiling of DNA methylation and chromatin architecture in single cells. *Nat. Methods*. 2019; 16: 991–993 .
- [17] Nagano T, Lubling Y, Várnai C, Dudley C, Leung W, Baran Y, Mendelson Cohen N, Wingett S, Fraser P, Tanay A. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*. 2017; 547: 61–67.
- [18] Flyamer IM, Gassler J, Imakaev M, Brandão HB, Ulianov SV, Abdennur N, Razin SV, Mirny LA, Tachibana-Konwalsi K. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*. 2017; 544(7648): 110–4.

- [19] Stevens TJ, Lando D, Basu S, Atkinson LP, Cao Y, Lee SF, Leeb M, Wohlfahrt KJ, Boucher W, O’Shaughnessy-Kirwan A, Cramard J, Faure AJ, Ralser M, Blanco E, Morey L, Sansó M, Palayret MGS, Lehner B, Di Croce L, Wutz A, Hendrich B, Klennerman D, Laue ED. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*. 2017; 544: 59–64.
- [20] Galitsyna AA, Gelfand MS. Single-cell Hi-C data analysis: safety in numbers. *Brief. Bioinform.* 2021; 22(6): bbab316.
- [21] Zhen C, Wang Y, Geng J, Han L, Li J, Peng J, Wang T, Hao J, Shang X, Wei Z, Zhu P, Peng J. A review and performance evaluation of clustering frameworks for single-cell Hi-C data. *Brief. Bioinform.* 2022; 23(6):bbac385.
- [22] Liu J, Lin D, Yardimci GG, Noble WS. Unsupervised embedding of single-cell Hi-C data. *Bioinformatics*. 2018; 34(13):i96–i104.
- [23] Hinton G, Roweis S. Stochastic neighbor embedding. *Advances in Neural Information Processing Systems*. 2002; 15.
- [24] McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection. *J. Open Source Softw.* 2018; 3(29): 861.
- [25] Collombet S, Ranisavljevic N, Nagano T, Varnai C, Shisode T, Leung W, Piolot T, Galupa R, Borensztein M, Servant N, Fraser P, Ancelin K, Heard E. Parental-to-embryo switch of chromosome organization in early embryogenesis. *Nature*. 2020; 580(7801): 142–146.
- [26] Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. 2007; 1027–1035.
- [27] Hubert L, Arabie P. Comparing partitions. *J Classif.* 1985; 2(1): 193–218.
- [28] Vinh NX, Epps J, Bailey J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J Mach Learn Res.* 2010; 11: 2837–2854.
- [29] Zhang R, Zhou T, Ma J. Multiscale and integrative single-cell Hi-C analysis with Higashi. *Nat Biotechnol.* 2022; 40: 254–261.
- [30] Zhou J, Ma J, Chen Y, Cheng C, Bao B, Peng J, Sejnowski TJ, Dixon JR, Ecker JR. Robust single-cell Hi-C clustering by convolution- and random-walk-based imputation. *Proc Natl Acad Sci U S A.* 2019; 116(28): 14011–14018.
- [31] Ma R, Huang J, Jiang T, Ma W. A mini-review of single-cell Hi-C embedding methods. *Comput. Struct. Biotechnol. J.* 2024; 23:4027–4035.
- [32] Kruskal JB. Nonmetric Multidimensional Scaling: A Numerical Method. *Psychometrika*. 1964; 29(2):115–129.
- [33] Maisuradze L, King MC, Surovtsev IV, Mochrie SGJ, Shattuck MD, O’Hern C. Identifying topologically associating domains using differential kernels. *PLOS Comput. Biol.* 2024; 20(7): e1012221.
- [34] Sun Q, Perez-Rathke A, Czajkowsky DM, Shao Z, Liang J. High-resolution single-cell 3D-models of chromatin ensembles during *Drosophila* embryogenesis. *Nat Commun. Biol.* 2021; 12(205).
- [35] Yang T, Zhang F, Yardimci GG, Song F, Hardison RC, Noble WS, Yue F, Li Q. Hicrep: assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient. *Genome Res. Biol.* 2017; 27(11):1939–1949.
- [36] Zheng Y, Shen S, Keleş S. Normalization and de-noising of single-cell Hi-C data with BandNorm and scVI-3D. *Genome Biol.* 2022; 23(1): 222.
- [37] Fletez-Brant K, Qiu Y, Gorkin DU, Hu M, Hansen KD. Removing unwanted variation between samples in Hi-C experiments. *Brief Bioinform.* 2024; 25(3): bbae217.
- [38] Shi J, Malik J. Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell.* 2000; 22: 888–905.