Computational and Structural Biotechnology Journal 23 (2024) 1584–1593

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Method Article

A method for chromatin domain partitioning based on hypergraph clustering ☆

Haiyan Gong [a,c], Sichen Zhang [b], Xiaotong Zhang [b,c,*], Yang Chen [d,**]

[a] Beijing Advanced Innovation Center for Materials Genome Engineering, Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing, 100083, China
[b] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, 100083, China
[c] Shunde Innovation School, University of Science and Technology Beijing, Foshan, 528399, Guangdong, China
[d] The State Key Laboratory of Common Mechanism Research for Major Diseases, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100005, China

ARTICLE INFO

Dataset link: http://nmdms.ustb.edu.cn/

Dataset link: http://mged.nmdms.ustb.edu.cn/storage/data/28658183

Dataset link: https://github.com/ghaiyan/TORNADOES

Keywords:
Hi-C
A/B compartment
Sub-compartment
Hypergraph generation
Hypergraph clustering

ABSTRACT

For many years, multi-scale models of chromatin domains, such as A/B compartments, sub-compartments, topologically associated domains (TADs), sub-TADs, and loops have been popular. However, existing methods can only identify structures at a single scale and cannot partition multi-scale structures. In this paper, we proposed a method (TORNADOES) for chromatin domain partitioning based on hypergraph clustering. First, we use a density clustering algorithm to identify TADs at different scales based on Hi-C data with different resolutions. Then, by combining ChIP-seq data features and TAD results at different scales, we generate a hypergraph based on these TADs. Finally, we partition the chromatin domain structure at different scales, including A/B, A1, A2, B1, B2, and B3 based on the Laplacian matrix feature of the hypergraph. Similarity comparison experiments and ChIP-seq signal enrichment analysis are performed on the A/B region and sub-TAD levels, respectively, demonstrating that our method can identify chromatin domains with distinct features and provide a deeper understanding of the organizational patterns and functional differences in TADs at the genomic hierarchical structure. Comparative analysis of multiple cell line data shows that TORNADOES can better classify different numbers and types of compartments by changing the factors ChIP-seq data and clustering number used to characterize TAD compared to other methods. Source code for the TORNADOES method can be found at https://github.com/ghaiyan/TORNADOES.

1. Introduction

Accurately identifying chromatin structures is crucial in fields such as biomedical research, bioinformatics, and molecular biology, as it can help us better understand gene expression, genetic variation, and the mechanisms that underlie many diseases. The hierarchical structure of chromatin includes chromosome territories (CT), A/B compartments, sub-compartments, and topologically associated domains (TADs) [1], in order of larger to smaller scales, and researchers have developed many different methods based on Hi-C [2], micro-C [3] and other 3C-based [4,5] technologies to study chromatin structure.

In 2009, Erez Lieberman-Aiden et al. [6] discovered the A/B compartments of chromatin using principal component analysis. By analyzing the first principal component that presented a bimodal distribution and dividing the chromosome into A (open) and B (closed) compartments based on gene density, the authors found that regions with positive eigenvalues displayed characteristics such as a higher number of genes, higher gene expression levels, stronger signals for DNase-sensitive sites, and higher GC content, indicating that these regions were more open and accessible. In 2012, Lin et al. [7] proposed HOMER for performing PCA on Hi-C data, and in 2015, Fortin et al. [8] predicted the A/B compartments of chromatin in multiple cell lines. The

☆ This work has been supported by the National Key R&D Program of China (2023YFB3812901), China Postdoctoral Science Foundation (2023M740219), Postdoctoral Fellowship Program of CPSF (GZC20230239), the Foshan Higher Education Foundation (No. BKBS202203).
* Corresponding author at: School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, 100083, China.
** Principal corresponding author.
E-mail addresses: zxt@ies.ustb.edu.cn (X. Zhang), yc@ibms.pumc.edu.cn (Y. Chen).

https://doi.org/10.1016/j.csbj.2024.04.008
Received 28 January 2024; Received in revised form 29 March 2024; Accepted 4 April 2024
Available online 16 April 2024

authors used DNA methylation data, DNase hypersensitivity data, and single-cell epigenetic data to validate the structural and functional characteristics of the A/B compartments. In the same year, Servant et al. [9] introduced the Hi-C-Pro software package, a tool for processing Hi-C data that provides several methods for identifying the A/B compartments of chromatin, of which the Insulation Score [10], Directionality Index [1], TopDom [11] and TADbit [12] are also used to define A/B compartments. In 2016, Durand et al. [13] subsequently proposed the Juicer method, which uses PCA to reduce the high-dimensional interaction frequency matrix to a two- or three-dimensional space and then applies the k-means clustering algorithm to cluster each region of chromatin. Similarly, the Miura et al. [14], CscoreTool [15], FAN-C [16], POSSUMM [17], and PENtad [18] methods all follow the PCA based idea to identify the A/B compartment. In 2017, Dong et al. [19] applied constrained clustering to partition the chromatin interaction matrix into blocks with similar contact probabilities and computed the feature vectors of each block to infer its local A/B compartment. This enabled the display of typical euchromatin and heterochromatin features. Rao et al. [20] performed a more detailed analysis of A/B compartments using high-coverage Hi-C data from the GM12878 cell line, further dividing them into five major sub-compartments: A1, A2, B1, B2, and B3. Importantly, these sub-compartments exhibit unique and more complete associations with various genomic and epigenomic features.

In recent years, several new methods have been developed to understand the three-dimensional structure of chromatin. For example, the Calder method [21] infers a complete hierarchical structure of partition domains solely from chromatin interactions using a similarity measure between genome loci and a hierarchical clustering method to cluster block domains, ignoring their continuity in the genome sequence. Moreover, SCI algorithm proposed by Ashoor et al. [22] predicts genome sub-compartments using graph embedding and K-means clustering. It starts from a normalized full-genome Hi-C chromatin interaction matrix, constructs a Hi-C interaction graph, and projects the interaction graph onto a low-dimensional vector space for K-means clustering to predict sub-compartments. These new methods provide a more comprehensive and accurate approach to understanding the three-dimensional structure and function of chromatin. In addition, the SNIPER [23] method uses a Gaussian Hidden Markov model (HMM) to cluster the rows of the inter-chromosome matrix. Clusters are then classified into A1, A2, B1, B2 or B3 sub-compartments according to the Spearman correlation between clusters.

With the development of high-throughput sequencing technologies, more and more studies [20–23] have begun to focus on investigating the spatial and hierarchical structure of chromatin to gain a deeper understanding of the three-dimensional structure and function of the genome. However, the spatial structure and organization of chromatin are complex areas that remain unexplored. The classification of TADs structures, however, is conducive to a more detailed understanding of chromatin structure and function. Analysis of chromatin hierarchical structure relies primarily on Hi-C technology and ChIP-seq technology, which identify chromatin structures at different scales by calculating Hi-C data with different resolutions and analyzing various histone modifications and transcription factor binding information in ChIP-seq data to explore chromatin function and regulation mechanisms. With this in mind, this paper combines the analysis of Hi-C and ChIP-seq signal data to classify different types of TAD structures.

Graph models are effective tools for discovering hidden correlations and inherent structures in data, and hypergraphs, as a generalization of graph models, can better capture higher-order information in data and provide great assistance for learning tasks. Compared to ordinary graphs, each edge in a hypergraph can link multiple vertices, thus better representing complex relationships in data. Additionally, hypergraphs can also surround vertices with similar features with hyperedges in an elegant way, making it easier to understand and process data. What's more, hypergraph learning is related to graph learning because hypergraphs are derived from graphs. Similar to graph learning, learning on hypergraphs can be seen as a process of information propagation along the hypergraph structure when analyzing structured data and solving problems such as node classification, link prediction, and community detection. From this perspective, graph learning is a special case of hypergraph learning that only considers pairwise connections between data. Unlike graph learning, hypergraph learning models explore higher-order associations between data, extending graph learning models to a high-dimensional, more complete nonlinear space, resulting in higher modeling capacity for correlations and better practical performance [24]. In the face of challenges representations of learning data, especially when dealing with complex data, hypergraphs exhibit more flexibility in data modeling as well.

Due to these advantages, there have already been various hypergraph-based applications, such as the application of hypergraphs to social networks [25], where communities can be regarded as hyperedges, that capture higher-order interactions in social and communication networks beyond simple pairwise relations. In bioinformatics, hypergraphs can be used to represent network data such as single-cell Hi-C data [26] and brain data [27]. Currently, hypergraphs have also been proposed in various machine learning methods, among which hypergraph spectral clustering, hypergraph semi-supervised learning, and hypergraph neural networks utilize hypergraph structures for relevant tasks processing and learning. Furthermore, hypergraph spectral clustering implements the extension of spectral graph theory through Laplacian hypergraphs, and hypergraph semi-supervised learning can constrain results by introducing hypergraph structures. In addition, hypergraph neural networks (HGNN) [28] use hyperedge convolutional operations to handle data correlations, which is well-suited to traditional hypergraph learning.

Considering the many advantages of hypergraph learning, in this paper we propose a chromatin domain partitioning algorithm called TORNADOES (a method for chromaTin dOmain paRtitioNing bAseD On hypErgraph cluStering). We first introduce the model and algorithm, focusing on hypergraph construction, feature generation, and hypergraph learning modules. Then, experimental results are analyzed at different scales. By comparing the similarity between A/B compartments, and sub-compartments, we find that TORNADOES can yield similar chromatin partitions with both of these chromatin domains. What's more, by setting the number of clustering, we can obtain useful biological chromatin domains.

2. Related work

2.1. Definition of a hypergraph

The basic definition of a hypergraph is as follows: let a hypergraph be denoted as G and consist of a set of vertices V and a set of hyperedges ε. In a weighted hypergraph, each hyperedge $e \in \varepsilon$ is assigned a weight to indicate its importance in the connectivity of the hypergraph. Let $diag(W) = [w(e_1), w(e_2), ...w(e_{|\varepsilon|})]$ be the diagonal matrix of the hyperedge weights. For a hypergraph represented as $G = (V, \varepsilon, W)$, its structure is typically represented by the adjacency matrix $H(v, e)$ which indicates whether a vertex v is in a hyperedge e. The degrees of hyperedges and vertices are defined by Equation (1) and (2), respectively.

$$\delta(e) = \sum_{v \in V} H(v, e) \tag{1}$$

$$d(v) = \sum_{e \in \varepsilon} w(e) * H(v, e) \tag{2}$$

The Laplacian matrix plays a crucial role in graph theory, particularly in spectral analysis of graphs by means of spectral clustering and spectral partitioning. For a simple graph, the Laplacian matrix can be defined using the diagonal matrix of vertex degrees and the adjacency matrix as $\Delta = D - A$, where D is the diagonal matrix of vertex degrees, and A is the adjacency matrix of the graph. However, in hypergraphs, the Laplacian matrix needs to be defined based on the relationship be-

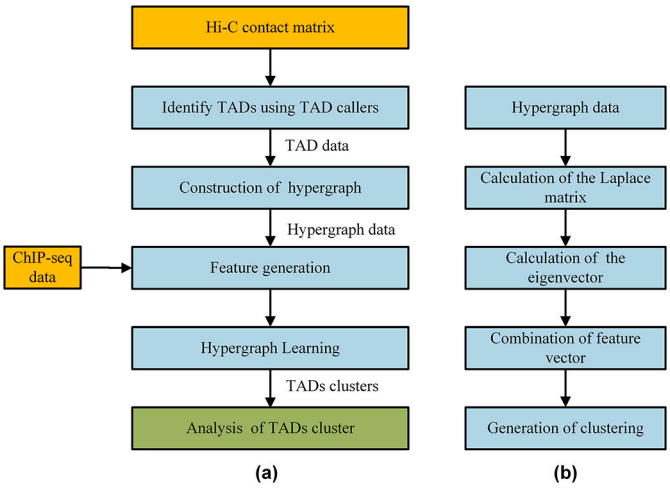

Fig. 1. Workflow of the TORNADOES method. (a) Workflow of analyzing the chromatin domains. (b) Workflow of generating the clusters based on hypergraph learning.

tween hyperedges and hypernodes. Specifically, in a hypergraph, the Laplacian matrix is defined as in Equation (3).

$$\Delta = D_v - HWD_e^{-1}H^T \tag{3}$$

The standardized Laplacian matrix can be expressed by Equation (4).

$$\Delta = I - D_v^{-\frac{1}{2}}HWD_e^{-1}H^TD_v^{-\frac{1}{2}} \tag{4}$$

2.2. Generation of hypergraph

The quality of a generated hypergraph structure directly affects the effectiveness of data correlation modeling. Hypergraph generation methods can generally be classified into four categories: distance-based methods [29], representation-based methods [30], attribute-based methods [31], and network-based methods [32].

In this paper, we employ an attribute-based approach to construct a hypergraph. The attribute-based hypergraph generation method utilizes attribute information to construct hyperedges. Each hyperedge in the attribute-based hypergraph is viewed as a clique, and the average heat kernel weight between any two edges in the clique is taken as the weight of the hyperedge. Since attributes can be hierarchical, the generated hyperedges can have different levels, resulting in different hyperedges that represent multi-scale attribute connections. Although attribute information has significant advantages in representing data, in some cases, such information may not be available. One possible solution is to define a set of attributes that can be learned from existing data. Therefore, we add ChIP-seq attributes to the hypergraph.

3. Materials and methods

Fig. 1 (a) depicts the workflow of our methods (TORNADOES) for chromatin domain structure partitioning and analysis. This method consists of three main modules: hypergraph construction, feature generation, and hypergraph learning. The hypergraph construction module is used to combine different TADs to form a hypergraph; the feature generation module generates corresponding feature vectors for the nodes in the hypergraph; and the hypergraph learning module is used to identify chromatin domain clusters based on the constructed hypergraph for further analysis. The next three sections provide detailed descriptions of each module.

3.1. The hypergraph construction module

The hypergraph construction module is based on the CASPIAN [33] method we previously proposed to combine domain structures identified at different scales to generate a nested hypergraph of nodes. Specifically, it includes the following steps:

(1). Determination of nodes in the hypergraph: Each TAD structure is considered to be a node in the hypergraph. Therefore, TAD partitioning is performed before constructing the hypergraph. Using the CASPIAN method [33], TAD structures are identified using a density-based clustering method and each TAD is labeled as a node. The center of each TAD can be used to represent the node.

(2). Defining hyperedges: Hyperedges define the interaction between TADs. Sub-compartments located between A/B compartments and topologically associating domain hierarchical structures are typically at the Mb level. Therefore, by setting parameters such as MinPts, the CASPIAN method is used to obtain domain structure partitioning results. After identifying domain structures at different scales, each genomic interval of the TADs included in each sub-compartment is used as a hyperedge. Then, each domain structure partition is traversed, and TAD partitions with overlapping regions are connected by a hyperedge. Specifically, if two TADs belong to the same domain structure, a hyperedge is created between them. When a TAD overlaps with multiple domain structures, the corresponding hypergraph node will exist in multiple hyperedges simultaneously.

(3). Hypergraph construction and representation: Based on the above steps, all nodes and hyperedges are combined to construct the hypergraph. The hypergraph can be represented using a graph, with each node represented by a circle and each hyperedge represented by a line. Alternatively, the hypergraph can be represented using an adjacency matrix, where each row represents a node and each column represents a hyperedge. The elements in the matrix indicate whether the node belongs to the corresponding hyperedge.

After hypergraph construction, each node in the hypergraph represents a TAD, and each hyperedge represents interactions between multiple TADs. This hypergraph representation can thus effectively describe the interactions and correlations between chromatin domains, thereby helping us to understand the 3D structure and function of the genome further. Additionally, this representation method provides powerful tools and ideas for further analyzing the structure and function of the genome.

3.2. The feature generation module

To understand the three-dimensional structure and function of the genome better, researchers often need to perform feature analysis on chromatin TADs. ChIP-seq signal data can provide information on different components of chromatin, such as transcription factor binding sites and histone modifications, that can be used to describe the features of chromatin TADs. Therefore, in this study, ChIP-seq signal data from CTCF, H3K4me3, and H3K27ac were selected, and the signal values of each genomic bin were calculated. The mean signal value of all bins within each TAD was then used as the signal feature of that TAD, forming an 3×3 feature matrix, where n is the number of TADs. Since each node in the hypergraph represents a TAD, this feature matrix is the node feature matrix of the hypergraph. We therefore combine ChIP-seq signal data with the structural features of TADs, in order to provide more comprehensive and accurate information for subsequent analysis. As Fig. 1 (b) shows, the feature extraction is described by the following steps:

(1). Calculation of the Laplacian matrix. The Laplacian matrix of the hypergraph is used to describe its structure, with diagonal elements representing the sum of the degrees of hyperedges, and off-diagonal elements representing the number of overlapping nodes between hyperedges.

(2). Calculation of eigenvalues and eigenvectors. After obtaining the Laplacian matrix of the hypergraph, double QR decomposition is used to transform the matrix into a Hessenberg matrix. The Hessenberg matrix is an upper triangular matrix with zero elements below the diagonal except for the element immediately below the diagonal. QR decomposition is a method for decomposing a matrix into an orthogonal matrix and an upper triangular matrix. By repeatedly applying QR decomposition, any matrix can be transformed into a Hessenberg matrix, with specific implementation methods including Householder transformation or Givens rotation. An iterative process based on implicit QR decomposition is then applied to the Hessenberg matrix to obtain the eigenvalues and eigenvectors. In the iterative process, the Hessenberg matrix is first transformed into a tridiagonal matrix, which is then subjected to QR decomposition to obtain a new tridiagonal matrix, and a series of iterations are then performed until this converges to a diagonal matrix. The elements on this diagonal represent the eigenvalues of the matrix, and the eigenvectors can be obtained through backward iteration.

(3). Combination of eigenvectors: The final features of each hypergraph node can be obtained by combining the features from the ChIP-seq signal data and the Laplacian matrix eigenvectors. This method has the advantage of integrating information from different sources to describe the structure and function of chromatin TADs more accurately. In addition, the hypergraph model can well reflect the interactions and connections between chromatin TADs, providing more comprehensive and accurate information for subsequent analysis. Furthermore, this method has good scalability and adaptability, making it applicable for analyzing chromatin TADs in different biological systems.

3.3. Hypergraph learning module

Clustering is one of the most commonly used methods for studying the three-dimensional (3D) structure and function of genomes. Prior to conducting clustering analysis, it is necessary to construct a hypergraph in order to extract features from and generate features for the hypergraph nodes. Once this node feature matrix is obtained, clustering algorithms such as k-means [34] or spectral clustering [35] can be used to classify hypergraph nodes into different types, in order to study the differences in 3D structure and function of different types of TADs. Specifically, the feature matrix of hypergraph nodes can be used as input, and clustering algorithms such as k-means or spectral clustering can be run on these vectors, with the expectation that vertices can be well separated in k-dimensional Euclidean space. In addition to k-means [34] and spectral clustering [35], there are other clustering algorithms that can be used for hypergraph clustering, such as hierarchical clustering [36] and density clustering [37]. These algorithms can be selected based on the specific research question at hand. In this way, TADs can be classified into different categories and the similarities and differences in 3D structure and function of different categories of TADs can be compared. Additionally, this method can help researchers discover TADs with special functions, such as regulatory elements and boundaries of chromatin higher-order structures.

3.4. Datasets

The real Hi-C data of IMR90, GM12878, H1-hESC, HepG2, and K562 cell lines were obtained from https://data.4dnucleome.org/files-processed/4DNFIH7TH4MF/, and the Hi-C contact matrix data were extracted from the original .hic file using the Juicer tool [38]. ChIP-Seq data from the Encyclopedia of DNA Elements (ENCODE) project [39] (https://www.encodeproject.org/) were used to analyze the enrichment of genome loci for CTCF and other histone modifications. The ChIP-Seq data files were in the bigWig format, which describes the signal p-value on contiguous genome loci. All datasets used in this paper are listed in Table S1.

3.5. Evaluation metrics

To evaluate the similarity between chromatin structure clusters generated by TORNADOES as well as A/B compartments and sub-compartments at genomic locations, several metrics based on existing research are used for comparison [40]:

(1). Pearson correlation coefficient: This metric is used to measure the linear correlation between two datasets. The value of the Pearson correlation coefficient ranges from -1 to 1, indicating the strength and direction of the correlation between the two datasets.

(2). Jaccard similarity coefficient: This metric calculates the ratio of the intersection to the union of two sets, $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$. It can be used to measure the number of common elements between two datasets.

(3). Cosine similarity: This metric is used to measure the similarity between two vectors. For our purposes we treat the datasets as a vector space, and calculating the cosine value of the angle between them, $Cosine\ Similarity(A,B) = \frac{A \bullet B}{||A|| * ||B||}$.

(4). Euclidean distance: This metric can calculate the distance between two vectors and can also be used to calculate the difference between two datasets. The Euclidean distance formula is $d(A,B) = sqrt(sum((A[i] - B[i])^2))$, where i represents each element in the datasets.

(5). Manhattan distance: This metric can also be used to calculate the distance between two vectors, with the formula $d(A,B) = sum(abs(A[i] - B[i]))$, where i represents each element in the datasets.

3.6. Benchmark models

To validate the effectiveness of the proposed TORNADOES method, we need to generate the A/B compartment and sub-compartment partitioning results as groundwork for subsequent experimental comparisons. The HOMER [7], CscoreTool [15], FAN-C [16], POSSUMM [17], and PENtad [18] methods all being PCA based A/B compartment recognition methods, but the FAN-C tool integrates PCA based A/B compartment recognition methods. Therefore, we select the FAN-C tool for A/B compartment recognition. SNIPER and Calder are both suitable for the identification of sub-compartments, so we selected Calder in this paper for the identification of A1/A2/B1/B2 compartments, and SNIPER method is selected for the identification and comparison of A1/B1/A2/B2/B3 compartments. The recognition and acquisition methods for these structures are described as follows:

(1). A/B Compartment Recognition: The recognition of A/B compartments is based on Hi-C matrix data generation. First, the correlation matrix is calculated, where each entry i, j of the correlation matrix corresponds to the Pearson correlation between row i and column j of the Hi-C matrix. Then, the inter-chromosomal type and strength of each genomic bin in the matrix are deduced using the eigenvectors of the correlation matrix. These are assigned using the sign of the eigenvectors (EV) of the correlation matrix. Generally, regions with positive values are assigned as “A” compartments, and regions with negative values are assigned as “B” compartments. Continuous bins with the same eigenvector sign are considered part of a “domain”. FAN-C tool [16] was used to detect A/B compartments.

(2). Sub-compartment Recognition: we use the Calder method proposed by Liu et al. [21] to recognize sub-compartments. First, the repository is cloned and installed from https://github.com/CSOgroup/CALDER and the input is in triple format, $(pos\_x, pos\_y, contact\_value)$. This method includes three modules: (i) calculation of chromatin structure domains; (ii) inference of hierarchical organization and obtainment of sub-compartments; and (iii) computation of nested sub-compartments within each domain. Finally, a bed format file is obtained that contains information about the subinterval to which each genomic region belongs.

(3). A1/A2/B1/B2/B3 compartment Recognition: SNIPER [23] is used to detect A1/A2/B1/B2/B3 compartments after cloning and installing the source code from https://github.com/ma-compbio/SNIPER.

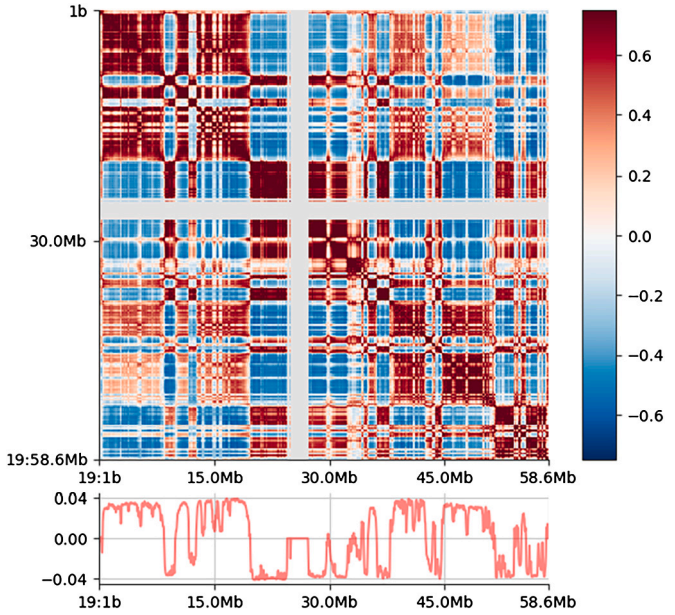

Fig. 2. Visualization of A/B compartments and eigenvector values.

The type of each genomic bin can be obtained using the A/B compartment recognition method, and sequence position comparison can give the compartment type of each TAD. Similarly, after sub-compartment recognition, the sub-compartment type of each continuous domain can be obtained, including A1, A2, B1, and B2, and the subinterval type of each TAD can be obtained through overlapping intervals calculation.

4. Results and discussion

4.1. Analysis of two clusters

First, we obtained the A/B compartment scores of each genomic bin through the feature decomposition of the correlation matrix of the Hi-C matrix. Fig. 2 shows the visualization of the A/B compartment division results of chromosome 19 in the IMR90 cell line. The red area corresponds to the A compartment, and the blue area corresponds to the B compartment. The line chart shows the feature vector values corresponding to the genomic loci. Those with values greater than or equal to 0 are assigned to the A compartment, and those with values less than 0 are assigned to the B compartment.

To analyze the features of the two domains clustered by TORNADOES, we compared the chromatin structural domains identified by the TORNADOES method with the A/B compartment structure. First, we set the number of clusters to 2 and clustered the nodes in the hypergraph into two clusters. Then, we used the evaluation metrics to verify the similarity between the TAD clustering sequence and the A/B compartment partition sequence. Fig. 3 (a) shows the Jaccard, cosine, and Pearson similarities between the 122 chromatin domains partitioned by TORNADOES into two types and the A/B compartment sequence. The positive correlations between the two types of partition sequences on each chromosome, indicating that the different structural domains identified by this method correspond to A/B compartments at the compartment level, demonstrating the similarities and differences between different TADs and laying a foundation for inferring the similarities and differences of these TADs in chromatin structure and function.

Table 1 presents the similarity metrics between chromatin domains obtained using different clustering methods with IMR90 cell line data from chromosome 19 as input. The k-means and spectral clustering (SC) methods achieved similar results, indicating the stability of the methods. Moreover, all similarity metrics had high positive values,

Table 1
Similarity results between two chromatin domains under different clustering methods.

Evaluation metrics	Cluster methods	Value
Jaccard similarity	K-Means	0.325
	Spectral Clustering	0.333
Cosine similarity	K-Means	0.564
	Spectral Clustering	0.603
Pearson correlation	K-Means	0.368
	Spectral Clustering	0.322
Euclidean distance	K-Means	7.681
	Spectral Clustering	7.615
Manhattan distance	K-Means	59
	Spectral Clustering	58

with the cosine similarity reaching 0.603, indicating a high positive correlation between the two chromatin domains obtained by the TORNADOES method and the A/B compartment structure. In addition, the enrichment of ChIP-seq signals within TADs and boundaries helps to measure the correlation between the three-dimensional structure of chromatin and gene expression even better. TADs and boundaries are often enriched with certain histone modifications, such as H3K4me1, H3K4me3, and H3K27ac, and these histone modifications are usually associated with the binding of gene promoters, enhancers, and transcription factors. In this section, ChIP-seq data related to transcription-promoting factors including H3K27ac, H3K4me3, CTCF, POLR2A, and transcription-inhibiting factor H3K9me3 were selected to validate the distribution of ChIP-seq signals in different chromatin domains identified by the TORNADOES method. As shown in Fig. 3 (b), the first type of domains contains more ChIP-seq signals related to transcription promotion, and the second type contains more signals related to transcription inhibition. These results indicate that the TORNADOES method can effectively identify chromatin domains with different features. They also demonstrate the interaction between different types of TADs, which are organized into corresponding A/B compartments to achieve specific gene expression.

Next, we visualized and analyzed the identified chromatin domains and related ChIP-seq signals. As shown in Fig. 3 (c), the different chromatin domains obtained by applying the TORNADOES method to the genomic loci in chromosome 19 (2450000-9900000) of the IMR90 cell line are displayed, with each triangle representing a TAD region and different colors representing different types of identified regions. The A/B compartment distribution is also visualized, with red representing the A compartment and blue representing the B compartment., and the two distributions exhibit a strong similarity. Additionally, the figure displays the distribution of different ChIP-seq signals in this genomic region, including H3K27ac, H3K4me3, CTCF, POLR2A, and H3K9me3. From the figure we can see that the ChIP-seq signals associated with transcriptional activation are mostly enriched at the TAD boundaries and exhibit a high level of enrichment in the A compartment and corresponding structural clusters and that the H3K9me3 signal associated with transcriptional inhibition is mainly distributed in the B compartment. Therefore, we conclude that the two types of chromatin domains clustered by TORNADOES method have similar biological properties to the A/B compartments.

4.2. Analysis of three clusters

We next clustered the nodes of the hypergraph into three types and used evaluation metrics to examine the similarity of the two node clustering results with the A/B compartmentalization sequence. Fig. 4 (a) shows the Jaccard similarity, cosine similarity, and Pearson correlation coefficient between the chromatin domains identified by TORNADOES and the A/B compartmentalization sequence for chromosomes 122 when two types of chromatin domains are identified. The results show that there is still a positive correlation between the two types of

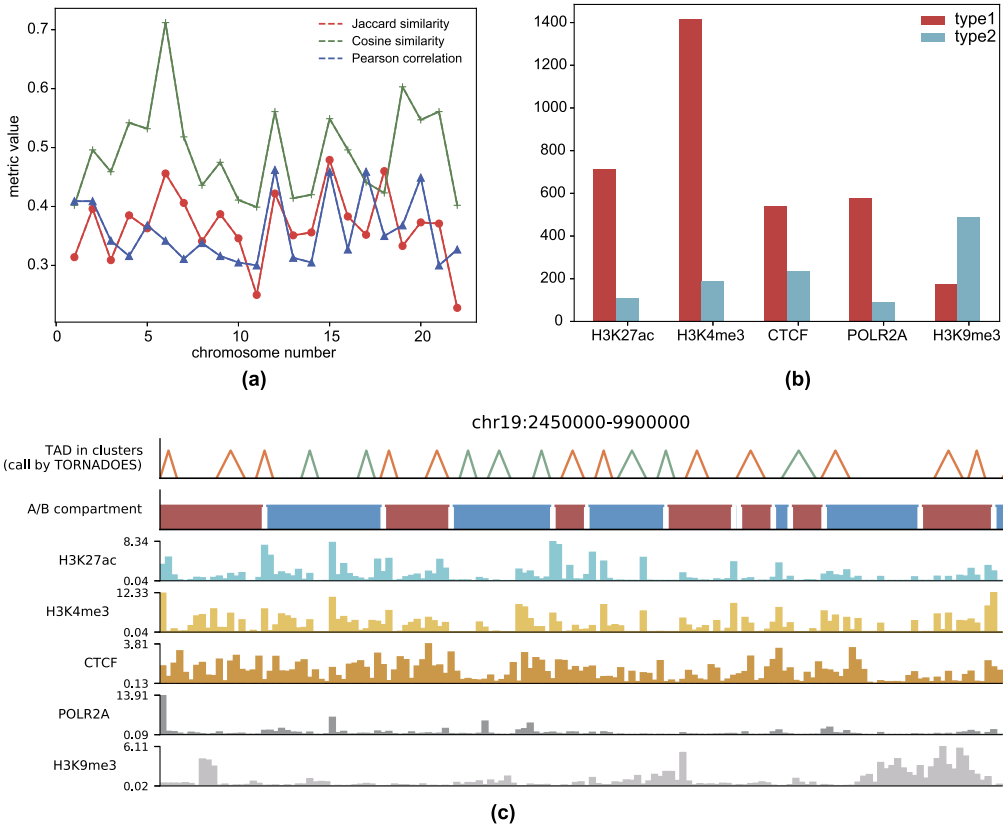

Fig. 3. Evaluation of two clusters. (a) the Similarity between two types of chromatin domains and A/B compartment sequences. (b) comparison of ChIP-seq signal values of two domains. (c) visualization of chromatin domains and ChIP-seq tracks.

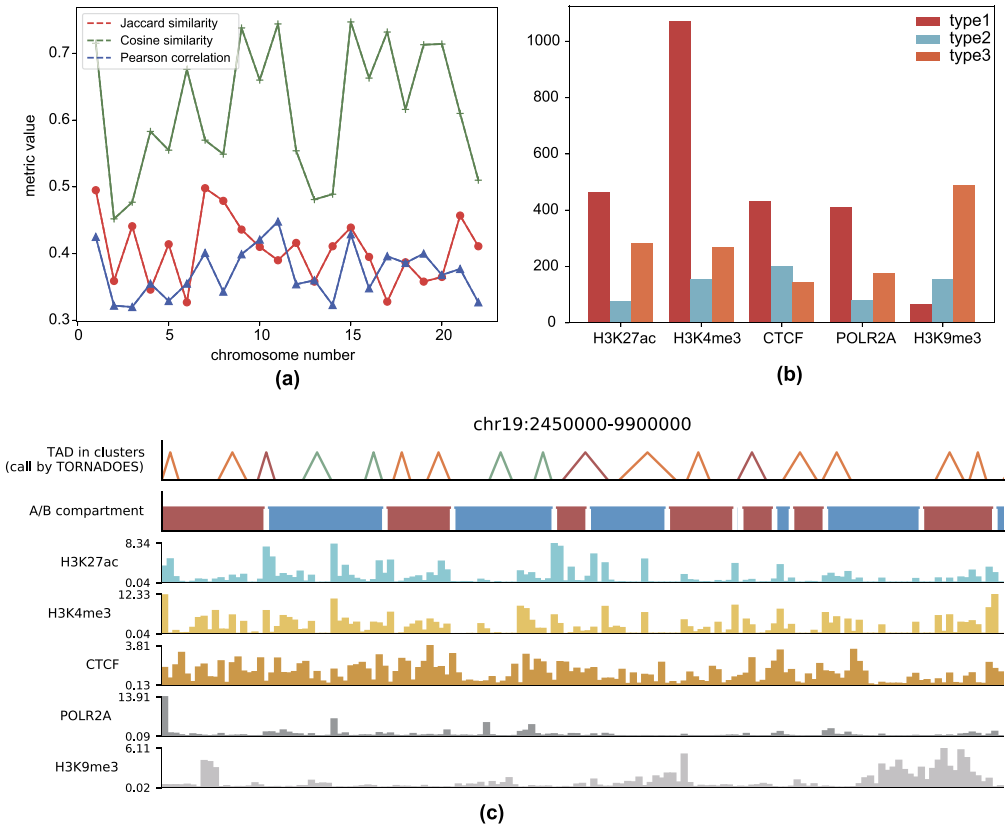

Fig. 4. Evaluation of three clusters. (a) The similarity between three types of chromatin domains and A/B compartment sequences. (b) Comparison of ChIP-seq signal values of three domains. (c) Visualization of chromatin domains and ChIP-seq tracks.

Table 2
Similarity results between three chromatin domains under different clustering methods.

Evaluation metrics	Cluster methods	Value
Jaccard similarity	K-Means	0.350
	Spectral Clustering	0.358
Cosine similarity	K-Means	0.674
	Spectral Clustering	0.713
Pearson correlation	K-Means	0.400
	Spectral Clustering	0.382
Euclidean distance	K-Means	6.48
	Spectral Clustering	6.78
Manhattan distance	K-Means	42
	Spectral Clustering	46

partitioning sequences on each chromosome, and the similarity results are better compared to the two cluster results.

Table 2 shows the similarity metrics for different clustering methods with three clusters on the IMR90 cell line chromosome 19 data. A ll similarity metrics achieve high positive values, with cosine similarity reaching 0.713, indicating a high positive correlation between the two chromatin domain structure partitions. Compared to the two clusters, there is an improvement, suggesting that a more detailed clustering is more in line with the complexity of chromatin domain structure distribution.

Similarly, the ChIP-seq signal values for each type of chromatin domain were also calculated, and the results are shown in Fig. 4 (b), which indicates that the first type of chromatin cluster contains more ChIP-seq signals associated with transcriptional activation, that the second type of chromatin cluster has low values for all signals, and that the third type of chromatin domain contains more ChIP-seq signals associated with transcriptional repression. Finally, the distribution of chromatin structural domains and the ChIPseq signal factor distribution were visualized. As shown in Fig. 4 (c), the TORNADOES method identified domains with different colors that still had a certain correlation with the A/B compartment, and the three types of structural clusters were mostly distributed in the A compartment, B compartment, and A/B compartment boundaries. These results demonstrate the effectiveness of the TORNADOES method in identifying chromatin domains with different features and also suggest that some gene loci do not exhibit obvious features of transcriptional activation or repression.

4.3. Analysis of four clusters

We also conducted a correlation analysis between the four types of chromatin domains identified by TORNADOES and the sub-compartment hierarchy using the Calder method [18] to identify sub-compartments. Hi-C data from chromosomes 1-22 of the IMR90 cell line were used as input, with the resolution set to 50 kb. Fig. 5 (a) shows the visualization of the sub-compartment regions identified by Calder on chromosome 19 of the IMR90 cell line, with different colors representing different sub-compartment categories. The visualization clearly demonstrates that this method can identify multiple sub-compartments.

The Calder method assigns a normalized level between 0 and 1 to each partition within each chromosome, with 0 being the least active partition and 1 being the most active. The method then calculates the Spearman correlation coefficient (SCC) between the partition levels and the reference genome's raw sequence, represented by ρ (Rho in Fig. 5 (b)). When ρ is less than 0.4, it indicates inaccurate compartmentalization. Fig. 5 (b) shows the SCC values between the partition levels and the reference genome for chromosomes 1-22 during the calculation of sub-compartments using the Calder method. The curve in Fig. 5 (b) demonstrates that all ρ values are greater than 0.4, with a maximum of 0.79, indicating that the method achieved accurate sub-compartmentalization.

In order to evaluate the correlation between the four types of chromatin domains identified by TORNADOES and sub-compartments, as

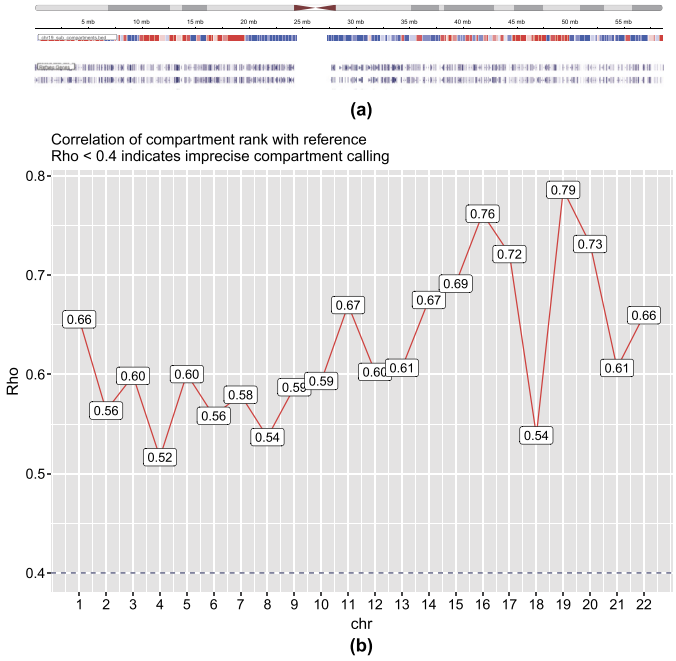

Fig. 5. Visualization of sub-compartments identified by Calder method. (a) the visualization of sub-compartments sequences. (b) the SCC values between the partition levels and the reference genome.

Table 3
Similarity results between four chromatin domains under different clustering methods.

Evaluation metrics	Cluster methods	Value
Jaccard similarity	K-Means	0.358
	Spectral Clustering	0.402
Cosine similarity	K-Means	0.556
	Spectral Clustering	0.738
Pearson correlation	K-Means	0.390
	Spectral Clustering	0.365
Euclidean distance	K-Means	8.520
	Spectral Clustering	6.190
Manhattan distance	K-Means	51
	Spectral Clustering	44

Fig. 6 (a) depicts, we calculated the Jaccard similarity, cosine similarity, and Pearson correlation coefficients between the four types of chromatin domain partitioning and the sub-compartment sequence, and found a positive correlation between the two partitioning sequences on each chromosome.

Table 3 shows the similarity metrics results between chromatin domains obtained using different clustering methods with a cluster number of 4 as input using IMR90 cell line chromosome 19 data. All similarity metrics achieved high positive values, with cosine similarity reaching 0.738, indicating a high positive correlation between the two chromatin domain structure partitions. Compared to the A/B compartment level with three clusters, there was an improvement, indicating that a more detailed classification better conforms to the inherent complexity of the chromatin domain structure distribution.

Next, the ChIP-seq signal values contained in the four types of chromatin domains were calculated, and the results are shown in Fig. 6 (b). Here we see that the first and second types of chromatin domains contain more ChIP-seq signals related to transcription promotion, which is more in line with the characteristics of the A sub-compartment. However, the third and fourth types of chromatin domains have lower levels of factors related to transcription promotion, and contain more enrichment of signals related to transcriptional inhibition, which is more in line with the characteristics of the B sub-compartment. This result

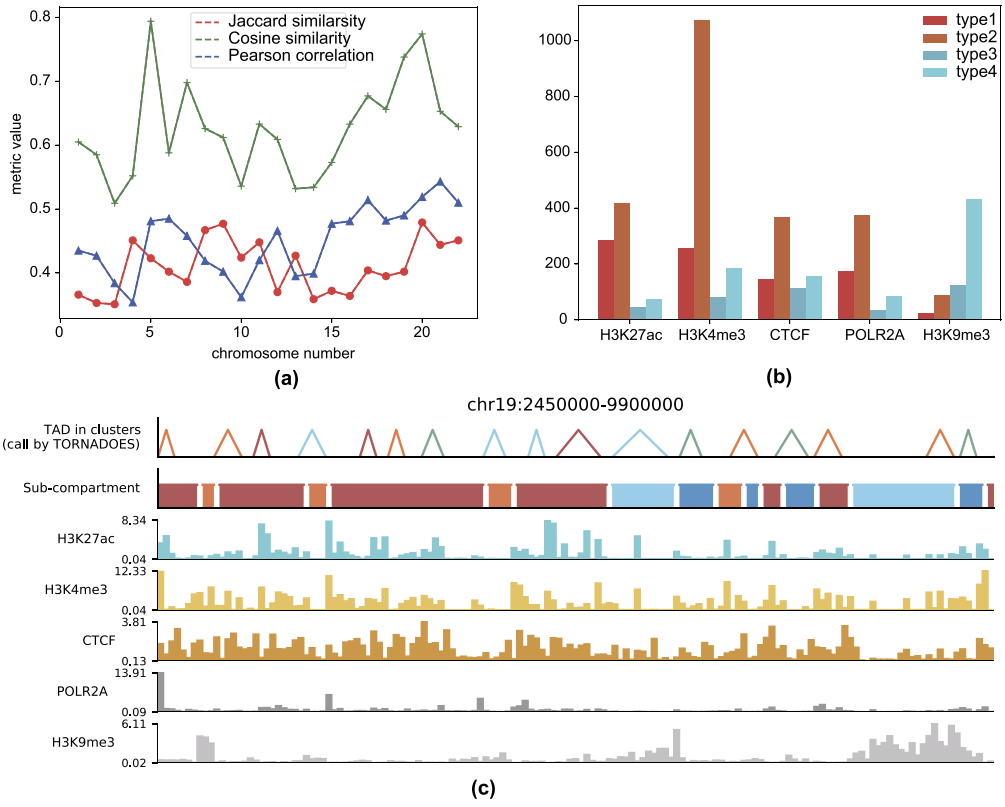

Fig. 6. Visualization of sub-compartments identified by the Calder method. (a) The visualization of sub-compartments sequences. (b) The SCC values between the partition levels and the reference genome.

demonstrates that the TORNADOES method can effectively distinguish different types of chromatin domains.

Furthermore, as shown in Fig. 6 (c), the distribution of chromatin domains and ChIP-seq factor signals was also visualized. The different types of TADs identified by TORNADOES were then displayed using different colors, and they still exhibited a certain correlation with the sub-compartmentalization chromatin sequence. The four types correspond to the A1, A2, B1, and B2 sub-compartments.

4.4. Analysis of five clusters

Finally, we conducted a correlation analysis between the five types of chromatin domains identified by TORNADOES and the A1/A2/B1/B2/B3 sub-compartments identified by the SNIPER [23] method. Hi-C data from chromosomes 1-22 of the IMR90 cell line were used as input, with a resolution of 50 kb. Similar to sections 4.1-4.3, we calculated the cosine similarity between the four types of chromatin domain partitioning and the sub-compartment sequence to evaluate the correlation between the five types of TADs identified by TORNADOES and the sub-compartments identified by SNIPER.

As shown in Fig. 7(a), the cosine similarity metric achieved the highest positive value yet (0.84), indicating a high positive correlation between the two chromatin domain structure partitions. The ChIP-seq signal values contained in the five types of TADs were then calculated, and as shown in Fig. 7 (b) the density distribution of the different factors can be divided into five types. To observe the distribution of different factors at the TAD boundaries in these five TAD types, we calculated the proportion of the peaks of various ChIP-seq factors anchored within the 20kb range of the TAD boundaries. As shown in Fig. 7(c), taking IMR90 cell line 1 chromatin as an example, we found that the third and fifth types have a higher ratio of anchoring factors associated with promoting transcription (CTCF, POLR2A, rad21, SMC3, H3K27ac, H3K36me3) and that the first, second and fourth types have a higher ratio of anchoring factors associated with inhibiting transcription (EZH2, H3K9me3). Therefore, we suppose that the third and fifth type TADs are more related to the function of promoting transcription, to the A1 and B1 domains and that the first, second, and fourth type TADs are more related to the function of inhibiting transcription, the B1, B2, and B3 domains.

4.5. TORNADOES on other cell lines

In order to test the performance of TORNADOES in other cell line data, we selected the Hi-C data of GM12878, H1-hESC, HepG2, and K562 cell lines at 50kb resolution for all chromatins. Hypergraphs were then constructed based on CTCF, POLR2A, H3K4me3, H3K27ac, and H3K9me3 ChIP-seq signal features. For the GM12878 and K562 cell lines, type = 2,3,4,5 was set for clustering, and the TADs were divided. For the H1-hESC and HepG2 cell lines, type = 2,3,4 was set for clustering, and the TADs were divided here as well.

First, we calculated the cosine similarity value of the chamber division result obtained when type was set to different values and the chamber cosine similarity value of A/B, A1/A2/B1/B2, A1/A2/B1/B2/B3, as shown in Figure S1-S3. When type = 2, the cosine similarity with the A/B area was generally the highest; when type = 3 and 4, the chamber cosine similarity with A1/A2/B1/B2 was generally the highest; and when type = 5, the chamber cosine similarity with A1/A2/B1/B2/B3 was generally the highest. Therefore, we conclude that by clustering TADs by setting type = 2,3,4,5 they can be mapped to different types of compartments.

Next, we observed the density distribution of different factors when the cluster number was 2, 3, 4, or 5 in different cell lines. As shown in Figures S4, S6, S8, and S10, the TORNADOES results for different types have highly differentiated ChIP-seq signal distributions. Therefore, we believe that the TORNADOES method can be used to classify TADs according to different ChIP-seq signal distribution ranges.

Last, we observed the count ratio for different factors when the cluster number was 2, 3, 4, or 5 and the chromosome number was 1 in the different cell lines. As shown in Figures S5, S7, S9, S11, when

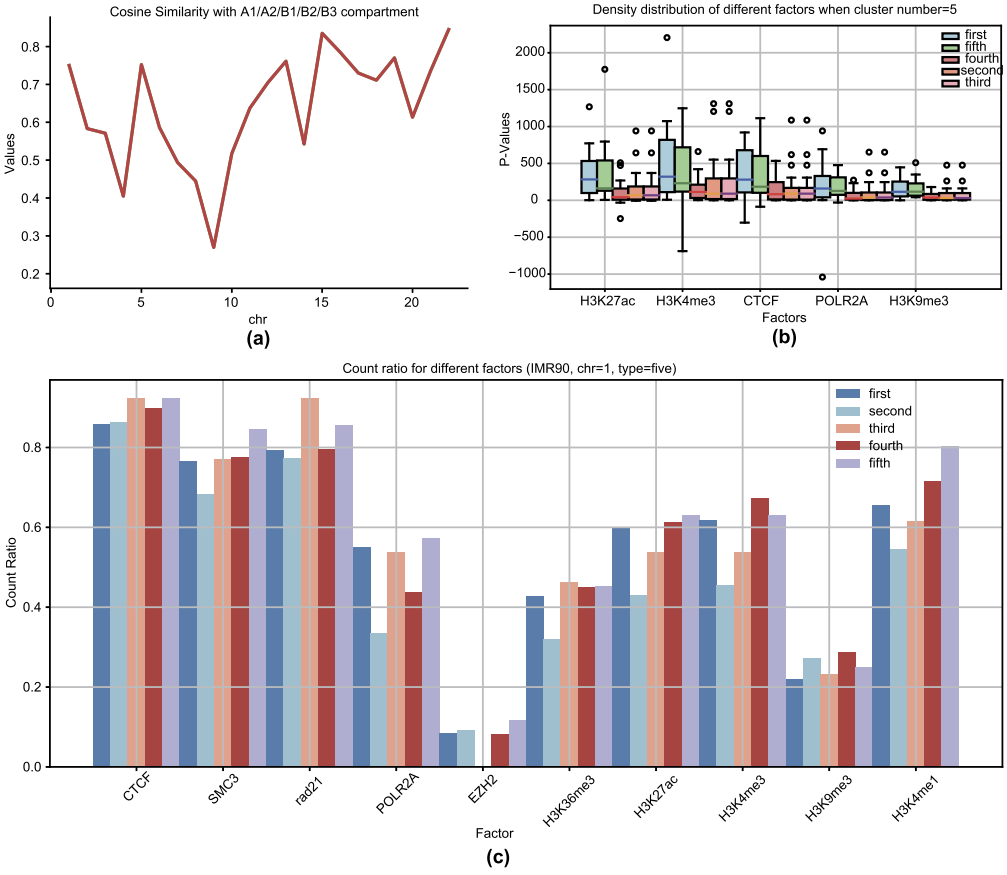

Fig. 7. Evaluation of five clusters. (a) Cosine similarity with A1/A2/B1/B2/B3 sub-compartment in the IMR90 cell line. (b) Density distribution of different factors when the cluster number equals 5 in the IMR90 cell line. (c) Count ratio for different factors when the cluster number equals 5 and the chromosome number equals 1 in the IMR90 cell line.

type = 2, the first type has a higher ratio of anchoring factors associated with promoting transcription, and the second type has a higher ratio of anchoring factors associated with inhibiting transcription (H3K9me3). Therefore, the first type corresponds to compartment A, and the second type corresponds to compartment B. When type = 3, the first type has a higher ratio of anchoring factors associated with promoting transcription, but the second and third type have a higher ratio of anchoring factors associated with inhibiting transcription. Therefore, the first type corresponds to compartment A, and the second and third types correspond to compartment B. We thus conclude that TORNADOES can also be used with other cell line data.

5. Conclusion

In this paper, we proposed a chromatin domain partitioning algorithm, TORNADOES, based on hypergraph partitioning. First, we used the CASPIAN algorithm, which is based on spatial density, to identify TADs. Then, based on these resulting TADs, we combined corresponding ChIP-seq data of histone modifications and transcription factors for the specific cell line, and generate hypergraphs and their corresponding features. Finally, by using different clustering algorithms, we performed hypergraph learning to cluster the TADs. The experimental results showed that by comparing the similarity of different numbers of clustering results with A/B compartments and sub-compartments, and the enrichment levels of different types of ChIP-seq corresponding to different types, TORNADOES obtained chromatin domains with different biological meanings, such as the A compartment associated with gene expression and the B compartment associated with gene repression. Although we only tested our method using the TAD identification method CASPIAN and two clustering methods (k-means and spectral clustering), users can change the TAD identification method and clustering methods when performing the hypergraph learning to get a potentially better result.

The TORNADOES method can be used to cluster different types of TADs successfully, but it does have some limitations. For example, obtaining accurate classification of TAD types requires a combination of TAD identification method and ChIP-seq factor selection. Therefore, we recommend that users choose a TAD identification method with high classification accuracy. In terms of ChIP-seq factor selection, if users want to divide TADs with different levels of promoting transcription, they can choose more common factors related to promoting transcription, such as CTCF, POLR2A, H3K4me3, and H3K27ac. If users want to differentiate TADs with different levels of transcriptional inhibition, they can choose more common factors associated with transcriptional inhibition, such as H3K9me3. All of these ChIP-seq data are available on the ENCODE platform.

Declaration of competing interest

The authors declare no conflicts of interest.

Data and code availability

All source data are listed in Table S1.

All processed data are stored in NMDMS (http://nmdms.ustb.edu.cn/) with accession linker http://mged.nmdms.ustb.edu.cn/storage/data/28658183.

Source code for the TORNADOES method can be found at https://github.com/ghaiyan/TORNADOES.

Acknowledgements

The authors thank Dr. Jin Yabin of Foshan First People's Hospital for his assistance in data analysis. The authors thank AiMi Academic Services (www.aimieditor.com) for English language editing and review services.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csbj.2024.04.008.

References

[1] Dixon JR, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 2012;485(7398):376.
[2] Van Berkum NL, et al. Hi-C: a method to study the three-dimensional architecture of genomes. J Vis Exp 2010;39.
[3] Hsieh T-HS, Weiner A, Lajoie B, Dekker J, Friedman N, Rando OJ. Mapping nucleosome resolution chromosome folding in yeast by micro-C. Cell 2015;162(1):108–19.
[4] Liang Z, et al. BL-Hi-C is an efficient and sensitive approach for capturing structural and regulatory chromatin interactions. Nat Commun 2017;8(1):1622.
[5] Ramani V, et al. Mapping 3D genome architecture through in situ DNase Hi-C. Nat Protoc 2016;11(11):2104–21.
[6] Lieberman-Aiden E, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009;326(5950):289–93.
[7] Lin YC, et al. Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. Nat Immunol 2012;13(12):1196–204.
[8] Fortin JP, Hansen KD. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. Genome Biol 2015;16(1):180.
[9] Servant N, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol 2015;16:259.
[10] Crane E, et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature 2015;523(7559):240.
[11] Shin H, et al. TopDom: an efficient and deterministic method for identifying topological domains in genomes. Nucleic Acids Res 2016;44(7):e70.
[12] Serra F, Baù D, Goodstadt M, Castillo D, Filion GJ, Marti-Renom MA. Automatic analysis and 3D modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLoS Comput Biol 2017;13(7):e1005665.
[13] Durand NC, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst 2016;3(1):95–8.
[14] Miura H, Poonperm R, Takahashi S, Hiratani I. Practical analysis of Hi-C data: generating A/B compartment profiles. (in eng), Methods Mol Biol 2018;1861:221–45.
[15] Zheng X, Zheng Y. CscoreTool: fast Hi-C compartment analysis at high resolution. Bioinformatics 2018;34(9):1568–70.
[16] Kruse K, Hug CB, Vaquerizas JM. FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data. Genome Biol 2020;21:1–19.
[17] Gu H, Harris H, Olshansky M, Eliaz Y, Krishna A, Kalluchi A, et al. Fine-mapping of nuclear compartments using ultra-deep Hi-C shows that active promoter and enhancer elements localize in the active A compartment even when adjacent sequences do not. BioRxiv 2021;2021-10.
[18] Magnitov MD, Garaev AK, Tyakht AV, et al. Pentad: a tool for distance-dependent analysis of Hi-C interactions within and between chromatin compartments. BMC Bioinform 2022;23(1):116.
[19] Dong P, et al. 3D chromatin architecture of large plant genomes determined by local A/B compartments. Mol Plant 2017;10(12):1497–509;
Chen F, Li G, Zhang MQ, Chen Y. HiCDB: a sensitive and robust method for detecting contact domain boundaries. Nucleic Acids Res 2018;46(21):11239–50.
[20] Rao SS, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 2014;159(7):1665–80.
[21] Liu Y, et al. Systematic inference and comparison of multi-scale chromatin subcompartments connects spatial organization to cell phenotypes. Nat Commun 2021;12(1):2439.
[22] Ashoor H, et al. Graph embedding and unsupervised learning predict genomic subcompartments from HiC chromatin interaction data. Nat Commun 2020;11(1):1173.
[23] Xiong K, Ma J. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. Nat Commun 2019;10(1):5069.
[24] Gao Y, Zhang Z, Lin H, Zhao X, Du S, Zou C. Hypergraph learning: methods and practices. IEEE Trans Pattern Anal Mach Intell 2022;44(5):25482566.
[25] Yang D, Qu B, Yang J, Cudre-Mauroux P. Revisiting user mobility and social relationships in lbsns: a hypergraph embedding approach. In: The world wide web conference; 2019. p. 2147–57.
[26] Zhang Ruochi, Zhou Tianming, Ma Jian. Multiscale and integrative single-cell Hi-C analysis with Higashi. Nat Biotechnol 2022;40(2):254–61.
[27] Zu C, Gao Y, Munsell B, et al. Identifying high order brain connectome biomarkers via learning on hypergraph. In: Machine learning in medical imaging: 7th international workshop. Springer; 2016. p. 1–9.
[28] Feng Y, You H, Zhang Z, Ji R, Gao Y. Hypergraph neural networks. Proc AAAI Conf Artif Intell 2019;33(01):35583565.
[29] Gao Y, Wang M, Tao D, Ji R, Dai Q. 3-D object retrieval and recognition with hypergraph analysis. IEEE Trans Image Process 2012;21(9):4290–303.
[30] Liu Q, Sun Y, Wang C, Liu T, Tao D. Elastic net hypergraph learning for image clustering and SemiSupervised classification. IEEE Trans Image Process 2017;26(1):452–63.
[31] Huang S, Elhoseiny M, Elgammal A, Yang D. Learning hypergraph-regularized attribute predictors. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR); 2015. p. 409–17.
[32] Moscato V, Picariello A, Sperlí G. An hypergraph data model for expert finding in multimedia social networks. In: Conte D, Ramel J-Y, Foggia P, editors. Graph-based representations in pattern recognition. Cham: Springer International Publishing; 2019. p. 110–20.
[33] Gong H, Yang Y, Zhang X, et al. CASPIAN: a method to identify chromatin topological associated domains based on spatial density cluster. Comput Struct Biotechnol J 2022;20:4816–24.
[34] Hamerly G, Elkan C. Learning the k in k-means. Adv Neural Inf Process Syst 2003;16.
[35] Liu J, Han J. Spectral clustering. In: Data clustering. Chapman and Hall/CRC; 2018. p. 177–200.
[36] Nielsen F, Nielsen F. Hierarchical clustering. In: Introduction to HPC with MPI for data science; 2016. p. 195–211.
[37] Kriegel HP, Kröger P, Sander J, Zimek A. Density-based clustering. Wiley Interdiscip Rev Data Min Knowl Discov 2011;1(3):231240.
[38] Durand NC, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst 2016;3(1):99–101.
[39] Dunham I, et al. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489(7414):57–74.
[40] Li Y, Wu A, Liu G, Liu L. A review of methods to quantify the genomic similarity of topological associating domains. J Comput Biol 2019;26(11):1326–38.