Computational and Structural Biotechnology Journal 20 (2022) 4816–4824

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

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

Method article

CASPIAN: A method to identify chromatin topological associated domains based on spatial density cluster ☆

Haiyan Gong [a], Yi Yang [a], Xiaotong Zhang [a,c,*], Minghong Li [a], Sichen Zhang [a], Yang Chen [b,*]

[a] School of Computer and Communication Engineering, Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing Advanced Innovation Center for Materials Genome Engineering, University of Science and Technology Beijing, Beijing 100083, China
[b] The State Key Laboratory of Medical Molecular Biology, Department of Biochemistry and Molecular Biology, Institute of Basic Medical Sciences, School of Basic Medicine, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing 100005, China
[c] Shunde Graduate School, University of Science and Technology Beijing, Foshan 528399, Guangdong, China

ARTICLE INFO

Article history:
Received 26 May 2022
Received in revised form 10 August 2022
Accepted 27 August 2022
Available online 5 September 2022

Keywords:
Hi-C
Chromatin structure
TAD
Spatial distance
Cluster

ABSTRACT

With the development of Hi-C technology, the detection of topologically associated domains (TADs) boundaries plays an important role in exploring the relationship between gene structure and expression. However, a method that can identify accurate TAD boundaries from the Hi-C contact matrix with different resolutions is currently lacking. We proposed a method named CASPIAN that can identify chromatin TAD boundaries based on the spatial density clustering algorithm. CASPIAN requires few parameters to call TADs. This method is realized using the hierarchical density-based clustering method HDBSCAN, where the distance of pairwise bins is calculated based on three distance metrics (Euclidean, Manhattan, and Chebyshev distance metric) to adapt to the characteristics of the Hi-C contact matrix generated from simulation experiments or normalized methods. Our results show that, same as standard methods (e.g., Insulation Score, TopDom), CASPIAN can enrich factors related to promoting the gene expression, such as CTCF, H3K4me1, H3K4me3, RAD21, POLR2A, and SMC3. We also calculated the approximate proportion of various factors anchored at the TAD boundaries to observe the distribution of these factors surrounding the TAD boundaries. In conclusion, CASPIAN is an easy method to explore the relationship between transcription factors and TAD boundaries. CASPIAN is available online (https://gitee.com/ghaiyan/caspian).

1. Introduction

Chromosome structure plays an important role in the gene expression and transcription, and available research has shown that chromosome structure is hierarchical with A/B compartments [1], topologically associated domains (TADs) [2], sub-TADs [3], and Loops [1]. An "A" compartment is defined as an open and expression-active chromatin domain with high gene density, and a "B" compartment is defined as a closed and expression-inactive domain with low gene density. Girelli et.al. [4] revealed that "A" compartments were more central than "B" compartments in relation to their radial distribution in the cell nucleus, the features of active chromatin, gene density, and expression increased globally toward the nuclear interior in parallel. This proved that gene function is related to the chromatin structure. TADs are thought to be the basic unit of chromatin organization and play significant roles in gene transcription, regulation and replication [5]. Increasing evidence indicates that when TAD boundaries are disrupted, gene

☆ This work has been supported by the National Key R&D Program of China [2018YFB0704301, 2018YFB0704304, 2018YFA0801402], the Scientific and Technological Innovation Foundation of Shunde Graduate School, USTB [BK20BF009], the National Natural Science Foundation of China [31871343] and the CAMS Innovation Fund for Medical Sciences (2020-RC310-009). Funding for open access charge: Department of Computer Science and Technology, Advanced Innovation Center for Materials Genome Engineering, University of Science and Technology Beijing.

* Corresponding authors at: School of Computer and Communication Engineering, Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing Advanced Innovation Center for Materials Genome Engineering, University of Science and Technology Beijing, Beijing 100083, China (X. Zhang). The State Key Laboratory of Medical Molecular Biology, Department of Biochemistry and Molecular Biology, Institute of Basic Medical Sciences, School of Basic Medicine, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing 100005, China (Y. Chen).

E-mail addresses: ghaiyan@xs.ustb.edu.cn (H. Gong), zxt@ies.ustb.edu.cn (X. Zhang), yc@ibms.pumc.edu.cn (Y. Chen).

https://doi.org/10.1016/j.csbj.2022.08.059

expression is adversely affected, leading to diseases such as dysmorphic syndromes and cancer [2,6]. Chromatin Loops are reported to be squeezed to form chromatin domains [7], resulting in the spatial proximity [8] of chromatin, such as enhancer-promoter interactions, that are real and biological chromatin interactions. The methods of identifying significant chromatin interactions from chromatin contact matrices are usually used to identify chromatin Loops. Existing studies show that some TADs are related to the chromatin Loops mediated by the CCCTC-binding factor (CTCF). Most of the two boundaries of a TAD coincide with the two anchor sites of a chromatin Loop [9–11]. Therefore, the identification of TADs can better facilitate our understanding of the mechanisms of chromatin formation.

Observation of the Hi-C contact heatmap has shown that TAD areas appear to be "triangles" on the diagonal of the heat map when the Hi-C resolution is less than 100 kb and that the intra-TAD interactions are merely $\sim 2X$ those of interactions between nearby TADs [2,7]. This indicates two fundamental features of TAD organization: the "self-association" property within TAD regions and the "insulation" property between adjacent TADs. The boundaries of TADs are enriched with the CTCF, cohesin complex, housekeeping genes, histone mark H3K4me3, as well as other factors [2,12], and all of these properties are essential features upon which various computational algorithms for identifying TADs rely.

Over the past decade, most researchers identified TADs by extracting 1-dimensional (1D) features from contact matrices for segmentation or by using clustering algorithms. The former includes methods such as Directionality Index (DI) [2], Insulation score [10], Arrowhead [13], TopDom [14], HiCSeg [15], TADtree [17], rGAMP [12], PSYCHIC [18], HiCDB [19], EAST [20], TADBD [21], and TADreg [22], convert the identification of TADs to be the local changes associated with topological features in the 1D signal. The latter methods including DHDF [23], TAD_Identification [24], IC-Finder [25], MrTADFinder [26], 3DnetMod [27], SpectralTAD [28], ClusterTAD [29], and TADpole [30] identified TADs by clustering algorithms. However, most of these methods cannot identify TADs from high-resolution Hi-C contact matrices (Dali et al. [31] have proved that only Arrowhead and DomainCaller could be run at 5 kb resolution on a server with 23 GB of RAM.), or the identified TAD boundaries cannot anchor factors with a high ratio (i.e., the ratio of TAD boundaries anchoring factors or transciption elements, such as CTCF, enhancer, promoter), or the parameters of TAD callers are too many for researchers to understand how to set suitable parameters to identify TADs.

Hence, we propose a chromatin topological associated domains (TADs) identification algorithm named CASPIAN based on the spatial density cluster without setting too many parameters. This method considers three distance metrics (Euclidean, Manhattan, and Chebyshev distance metric) to do clustering from the Hi-C contact matrix generated by simulated experiments or by different Hi-C normalized methods to detect TADs. Results show that CASPIAN using Euclidean distance metric could always perform well for simulated Hi-C contact matrix or real Hi-C data at different resolutions (5 kb, 25 kb, and 50 kb). By observing the average P-value and anchor ratio of different factors target ChIP-seq surrounding the TAD boundaries, we conclude that TAD boundaries are enriched with factors related to promoting the gene expression (CTCF, H3K4me1, H3K4me3, RAD21, POLR2A, and SMC3), in particular, the detected TAD boundaries anchor about 70% CTCF, 29% enhancer, 14% promoters,30% H3K36me3, 40% H3K4me1, 30% H3K4me3, 30% POLR2A, 60% RAD21, and 50% SMC3. By assessing the recalling ratio of TADs called by other methods, CASPIAN can recall more TADs that are detected from Insulation score [10] and TopDom [14]. Overall, our CASPIAN method can detect TAD boundaries from simulated Hi-C and real Hi-C generated with different resolutions to help explore the distribution of different factors surrounding the TAD boundaries.

2. Methods

2.1. Overview of the CASPIAN Framework

Fig. 1(a) illustrates the data analysis pipeline of CASPI-AN, including three modules for identifying chromatin TAD boundaries: the data preprocessing module, the cluster module, and the TADs generation module. The data preprocessing module is designed for obtaining a raw contact matrix or a normalized contact matrix from raw sequence Hi-C data. The cluster module applies hierarchical density-based clustering methods [32] to classify these Hi-C bins' features with different labels. The TADs generation module divides the boundaries of TAD according to these clustering results.

2.2. The Data Preprocessing Module

By processing the raw Hi-C sequence data with mapping and alignment, we could get the raw Hi-C matrix. The Hi-C experiment measures the probability of physical proximity between pairs of chromosomal loci at the genome scale, but there are some systematic biases during the experiment that seriously affect the experimental results [33]. Therefore, it's essential to normalize Hi-C data to eliminate system bias. In this paper, we chose the vanilla coverage (VC) and Knight & Ruiz methods (KR) [34] to do Hi-C data normalization. The VC algorithm divided each bin of the matrix by its row sum and column sum to remove different sequencing coverage of each loci. The KR algorithm [34] aimed to balance nonnegative square contact matrices to eliminate this bias, which is widely used for correcting Hi-C contact matrices.

2.3. The Cluster Module

The CASPIAN method is designed to cluster bins based on the chromosome space distance density. Importantly, the DBSCAN [35] (density-based spatial clustering of applications with noise) clustering algorithm does not need to set the number of clusters and can be used to divide clusters with complex shapes. The principle of DBSCAN is to identify points in dense regions of the feature space where many data points are close together, and this is similar to the characteristics of TADs in chromatin. Therefore, we can apply DBSCAN to identify TADs without considering the number of TADs.

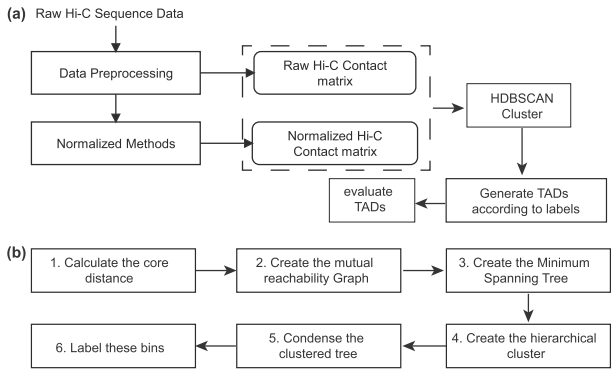

Fig. 1. Overview of the CASPIAN Framework. (a) The data pipeline for CASPIAN to identify and evaluate chromatin TAD boundaries. (b) The workflow for HDBSCAN.

2.3.1. The Description of DBSCAN

The DBSCAN algorithm has two parameters: the minimum number of points in the radius, m_{pts} and the neighborhood radius, *eps*. Given a sample, let $B_{eps}(x_i)$ represent a sphere within the radius of the neighborhood centered on the sample. The function $N(\cdot)$ given below counts the number of samples contained in the set. A sample $x_i \in X$ is called a core sample if it satisfies Eq. (1):

$$N(B_{eps}(x_i)) \geqslant m_{pts} \tag{1}$$

The DBSCAN algorithm puts core samples that are less than *eps* away from each other into a cluster. The algorithm first randomly selects a sample x_i, and then finds all samples whose distance is less than or equal to *eps*. If the number of samples within *eps* from x_i is less than m_{pts}, then the sample is marked as noise, that is, it does not belong to any cluster. If this sample is a core sample, it is assigned a new cluster label. After this, all neighbors within $B_{eps}(x_i)$ are visited. If these neighbors have not been assigned a cluster label, and assign them the cluster label just created; if these neighbors are core samples, visit their neighbors. The cluster grows until there are no more core samples within the cluster's neighborhood radius. The above steps are then repeated until all samples have been visited. Altogether there are three types of samples: core samples, boundary samples (samples whose distance from the core samples is within *eps*), and noise.

However, the parameters (*eps* and m_{pts}) of DBSCAN algorithm determine the quality of clustering. In this paper, we chose the extended algorithm HDBSCAN to do clustering without manual selection of parameter *eps* and m_{pts}.

2.3.2. The Description of HDBSCAN

Given a sample set $X = \{x_1, x_2, \ldots, x_m\}$, as Fig. 1(b) shows, we can cluster the samples based on HDBSCAN within six steps:

(1) Choose the Minkowski distance metric (Manhattan, Euclidean, or Chebyshev) to calculate the core distance $d_{core}(x_p)$ of sample $x_p \in X$ from x_p to its m_{pts}-nearest neighbor including x_p, where the Manhattan distance, Euclidean distance, and Chebyshev distance can be described as Eq. (2), the core distance satisfies the Eq. (1).

$$P(A,B) = \left(\sum (a[i] - b[i])^p\right)^{1/p} \tag{2}$$

where A, B are two data sets with same number n of data, $i \in [1, 2, 3, \ldots, n]$. When p is equal to be 1, the Minkowski distance is the Manhattan distance. When p is equal to be 2, the Minkowski distance is the Euclidean distance. When p is equal to be ∞, the Minkowski distance is the Chebyshev distance.

(2) Caculate the mutual reachability distance to generate the Mutual Reachability Graph $G_{m_{pts}}$, where the mutual reachability distance ($d_{mreach}(x_p, x_q)$) between sample x_p and x_q is defined as Eq. (3) shows. The Mutual Reachability Graph $G_{m_{pts}}$ is a complete graph, where the vertices are the samples of X, the weight of each edge is the mutual reachability distance between pairwise samples.

$$d_{mreach}(x_p, x_q) = \max(d_{core}(x_p), d_{core}(x_q), d(x_p, x_q)) \tag{3}$$

(3) Create the minimum spanning tree based on the minimum spanning tree algorithm prime [37].

(4) Create a hierarchical cluster similar to the preocess of Huffman tree construction. Firstly, sort all the edges in the tree by increasing distance. Secondly, select each edge in turn and merge the two subgraphs of the edge link.

(5) Condense the clustered tree. Firstly, determine the minimum cluster size n. Secondly, traverse the clustering tree from top to bottom to see if the number of the two sample subsets generated by the split is greater than n. If the number of samples of a child node in the left or right sons is $< n$, delete the node directly, and the other child node retains the identity of the parent node. If both of the sample size of the two child nodes is $< n$, then delete both of its child nodes. If the number of samples in both child nodes is $>= n$, then the original clustering tree keeps unchanged.

(6) Label these samples based on the cluster stability. Let $\lambda = 1/dist$, where *dist* is the mutual reachability distance. For the vertices of a tree, define λ_{birth} and λ_{death} to be the λ value when the cluster splits and becomes its own cluster or the λ value when a cluster splits into smaller clusters. For a sample p in a given cluster, we define λ_p to be λ value of the point "outlier", where $\lambda_{birth} \leqslant \lambda_p < \lambda_{death}$. The cluster stability is define to be $\sum_{p \in cluster}(\lambda_p - \lambda_{birth})$. Firstly, initialize the cluster by assigning each leaf node of the condensed tree to be a cluster. Secondly, walk through the tree from the bottom up with the following process. If the stability of the current node is less than the sum of the stability of the two children nodes, then we set the stability of this node to be the sum of the stability of its children nodes. If the stability of the current node is greater than the sum of the stability of the two children nodes, the current node is defined as a cluster.

2.4. The TADs Generation Module

As described above, each bin is labeled based on the clustering module. Research has shown that the length of TAD is between 100 kb and 5 Mb [36]. Therefore, for a continuous area formed by bins with a same label is identified to be a TAD, when the number of $bins \times resolution > 100kb$. For example, when the resolution of Hi-C data is equal to be 50 kb, then the continuous area must contain at least 2 bins with a same label.

2.5. Simulated Hi-C Data

We used the simulated Hi-C data proposed by Mattia etc. [38] for TAD callers. The simulation process is described as the following. These bins coordinates inside a TAD are randomly simulated by sampling TAD sizes from a uniform distribution with minimum TAD bins size value of 3 and a maximum TAD bins size value of 50. The simulated contact matrix contains a fixed number of TADs and a random sampling of TAD sizes. The target size of the simulated contact matrix is defined based on the expected average of the uniformly distributed TAD sizes. Using this strategy, 171 TADs were simulated at 40 kb resolution with a target size similar to the size of the human chromosome 5 (180.92 Mb), i.e., the same used to estimate the power-law decay parameters. Among the 171 TADs, we used the simulated 40 kb resolution Hi-C contact matrix when noise levels 4, 8, 12, and 16 to compare with other TAD callers. To assess the performance of CASPIAN at different resolutions, we also downloaded the synthetic Hi-C dataset provided by Trussart etc. [39]. This synthetic Hi-C dataset was generated by a worm-like chain (WLC) model, which was characterized by the physical thickness (the diameter, unit: nm), the stiffness (the persistence length, unit: nm), the level of DNA compaction of the chain (the linear density, unit: bp/nm), the experimental noise level to select pairwise interactions with a Gaussian probability model. In this paper, we used the synthetic Hi-C dataset with a linear density of 40, 75, 150 bp/nm and a noise level of 50, 100, 150, and 200.

2.6. Real Hi-C Data

To obtain the Hi-C contact matrix, we downloaded the Hi-C data (486.85 million read pairs of sequencing coverage) in the GM12878 cell line with the accession number 4DNFI1UEG1HD from the 4dnucleome platform (https://data.4dnucleome.org/). By executing the jar package juicer_tools_1.22.01.jar, we could get an N*N raw,

KR-normalized, or VC-normalized matrix at 5 kb, 25 kb or 50 kb resolution, where N is the number of bins.

To run the Directionality Index (DI) [2] and Insulation Score (IS) [10] algorithms, we downloaded the.mcool format Hi-C data in the GM12878 cell line with the accession number 4DNFIXP4QG5B from the 4dnucleome platform.

2.7. ChIP-seq Data

The CTCF, POLR2A, RAD21, SMC3, H3K4me3, H3K3-6me3, and H3K9me3 target ChIP-seq data were downloaded from ENCODE platform [16] (www.encodeproject.org [40]) with accession number ENCFF749HDD, ENCFF002GST, ENCF-F822QJA, ENCFF775OOS, ENCFF480KNX, ENCFF537K-DM, ENCFF174RRQ, ENCFF218YZR, respectively.

We used the replicated peaks files (bed narrow peak file type) to calculate the number of peaks where the identified TAD boundaries anchor. We used the signal p-value files (bigWig file type) to plot the average p-value of ChIP-seq data around the genomic bins of identified TAD boundaries within 40 kb genomic distance.

2.8. Evaluating the Quality of TADs

To evaluate the quality of TADs, we chose the following metrics. First, we chose the average length of identified TADs and the number of identified TADs to be the metrics.

Second, to compare the similarity of TADs identified by CASPIAN with the simulated TADs (the true TADs) from simulation Hi-C data or TADs identified from other existing methods, bins are divided into three clusters according to the TAD results: bins in TAD boundaries, bins within TADs, and bins between TADs. The Fowlkes-Mallows score (FMS), and Rand index (RI) provided by the scikit-learn tool [41] are then used to evaluate the extent of the agreement between two clusters. The FMS is defined as the mean of the precision and recall as shown in Eq. (4).

$$FMI = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}} \tag{4}$$

In Eq. (4), TP is the number of true-positive sample pairs, FP is the number of false-positive sample pairs, and FN is the number of false-negative sample pairs. The FMI ranges from 0 to 1, and a higher value indicates more similarity between the two clusters.

The Rand Index is defined as $RI = N_1/N$ to compute the similarity between two clusters, where N_1 is the number of agreeing pairs, N is the number of pairs. A higher RI value indicates more similarity between two clusters.

Third, one existing study [2] has suggested that TAD boundaries are enriched with CTCF, H3K4me3, H3K36me3, and transcription start sites (TSSs). Because CTCF is related to some active transcriptional activity, the RAD21 and SMC3 are all transcription factors that promote transcription. Hence, both the ratio of anchoring CTCF binding sites and histone modification mark H3K4me3, H3K36me3, and transcription factors signal (POLR2A, RAD21, and SMC3) are applied to evaluate the TAD quality.

Last, reasearches [9–11] have shown that some of the two boundaries of a TAD coincide with the two anchor sites of a chromatin Loop (such as enhancer-promoter interaction). Therefore, we also explore the ratio of TAD boundaries identified by different TAD callers anchoring enhancer or promoter elements to validate the correctness of identified TAD boundaries. We define $ratio_e = N_e/N_{TAD}, ratio_p = N_p/N_{TAD}$, where N_e is the number of TAD boundaries anchoring an enhancer element, N_p is the number of TAD boundaries anchoring an promoter element, N_{TAD} is the number of TAD boundaries identified by TAD callers.

2.9. Baselines

In this paper, we chose the following methods to compare with CASPIAN: Insulation Score (IS) [10], TopDom [14], Directionality Index (DI) [2], IC-Finder [25], HiCSeg [15], and ClusterTAD [29]. We chose the above methods by considering two factors: 1) whether TADs identified by these methods are hierarchical. Because we do not consider identifying hierarchical TADs structures in this paper. 2) the only input data of TAD callers is Hi-C contact matrix. Among these methods, Insulation Score, Directionality Index (DI), TopDom, IC-Finder, and HiCSeg are five standard TAD callers, ClusterTAD is a TAD caller based on a clustering algorithm. Since the TADs identified in this paper do not consider the hierarchical structure, the above-selected methods all can not detect hierarchical TADs. All scripts for calling TADs by the above methods can be found in "Scripts_for_TAD_caller.txt", which can be downloaded from https://gitee.com/ghaiyan/caspian.

3. Results and Discussion

3.1. Determination of the Distance Metric

For the first step of HDBSCAN, considering the characteristic of TADs (i.e., the triangular domain on the diagonal of the Hi-C map) in the Hi-C contact map, we used the Minkowski distance including Manhattan, Euclidean, and Chebyshev distance to measure the distance of pairwise bins. For simulated and real Hi-C contact matrix, the three distance metrics had a different performance.

Firstly, we explored the distance metrics how to affect the TAD quality of TADs identified from simulated 40 kb resolution Hi-C contact matrices with noise leveling 4, 8, 12, 16 provided by an existing study [38]. As Fig. 2 and Figure S1-S3 show, CASPIAN using

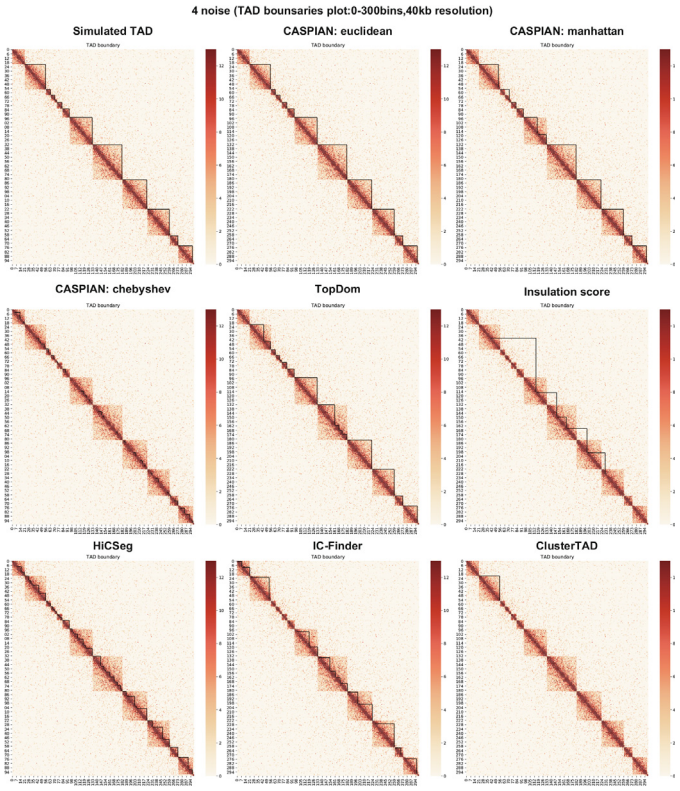

Fig. 2. Heatmaps with detected TAD boundaries based on a simulated 40 kb resolution Hi-C contact matrix at noise level 4, where the identified TADs are outlined by black lines.

the Euclidean distance metric gets similar TADs segmentation results with the simulated TADs provided by Mattia et al. [38] from the aspect of heatmap visualization, CASPIAN using the Manhattan distance metric not only detects the simulated TADs provided by Mattia, etc [38], but also detects more TADs within these TADs, CASPIAN using the Chebyshev distance metric detects tiny TADs, but obviously differs from the simulated TADs. As Fig. 3 shows, we used the rand index (RI) and Fowlkes-Mallows score (FMS) to measure the similarity between identified TADs and simulated TADs. These results show that CASPIAN using the Euclidean distance metric performs much better than the other two distance metrics on simulated Hi-C data with noise leveling 4, 8, 12, 16, where RI and FMS are both close to 1.

To assess the performance of CASPIAN under different simulation conditions, we used the synthetic Hi-C dataset provided by Trussart et al. [39] with a linear density of 40, 75, 150 bp/nm and a noise level of 50, 100, 150, 200 to detect TAD boundaries using CASPIAN (FigureS4, S5, S6). The heatmaps under different conditions show that the Hi-C contact matrix at a lower noise level is sparser than the Hi-C contact matrix at a higher noise level, and the Hi-C contact matrix with higher linear density is sparser than the Hi-C contact matrix with lower linear density. Therefore, we think the Hi-C contact matrix with high linear density at a low noise level is similar to the real Hi-C contact matrix at high resolution, the Hi-C contact matrix with low linear density at a high noise level is similar to the real Hi-C contact matrix at low resolution. The number of TADs detected using CASPIAN with different distance metrics (Table S1) shows that the Euclidean metric performs more robust than Manhattan and Chebyshev metrics. FigureS4, S5 and S6 show that CASPIAN using the Euclidean metric performs better than using the Manhattan and Chebyshev metrics on detecting TAD boundaries under different simulation conditions. Therefore, we recommend users run CASPIAN using Euclidean metric in most conditions.

We next explored the distance metrics how to affect the TAD quality of TADs identified from the real Hi-C contact matrix of all chromosomes (from chromosome 1 to chromosome 22) in the GM12878 cell line at 50 kb resolution. By comparing the distribution of mean P-value of the CTCF, POLR2A, RAD21, SMC3, H3K4me3, H3K36me3, and H3K9me3 target ChIP-seq data near the locus of identified TAD boundaries ranging from −40 kb to + 40 kb genomic distance. As Fig. 4 shows, Near the boundaries of identified TADs by CASPIAN using the three distance metrics, CTCF, POLR2A, RAD21, SMC3, H3K4me3, and H3K36me3 factors are enriched, chromatin modifications H3K9me3 is unenriched. The results are consistent with the results that boundaries of TADs are enriched for the insulator binding protein CTCF and housekeeping genes as Dixon, et al. [2] reported. This indicates that CASPIAN using all the three distance metrics could detect TAD boundaries correctly. However, from the aspect of factors enrichment at the TAD boundaries (the genomic distance from boundary = 0), the Euclidean and Manhattan distance metrics perform better than the Chebyshev distance metric. We also compared the distributions at 25 kb (FigureS7) and 5 kb (FigureS15) resolution, and get a similar conclusion with the results at 50 kb resolution. We detected TADs from 5 kb resolution KR-normalized Hi-C contact matrix (chromosome 9, 11, 12, 13, 14, 15) in the GM12878 cell line by CASPIAN using the Euclidean distance metric. As FigureS15 shows, we plotted the mean P-value of CTCF target ChIP-seq surrounding the TAD boundaries within 400 kb genomic distance. The results show that TADs called from 5 kb resolution Hi-C contact matrix also show enrichment for factors related to promoting the gene expression (CTCF, H3K4me1, H3K4me3, RAD21, POLR2A, and SMC3). This means CASPIAN could also detect TADs accurately from Hi-C data at a high resolution (e.g., 5 kb).

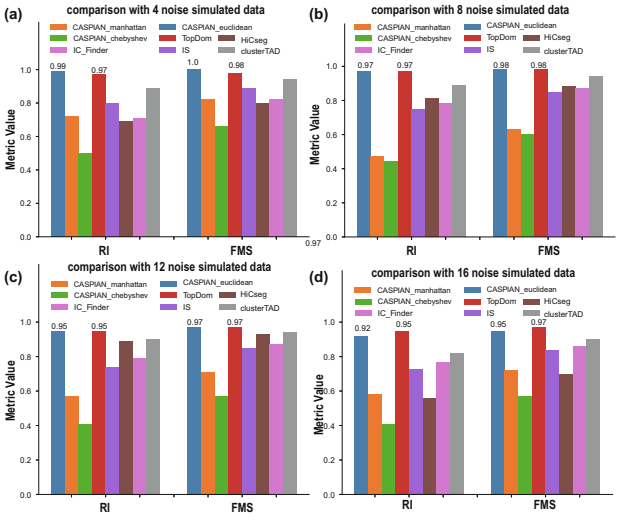

Fig. 3. Histogram plot of RI and FMS to show the similarity of TADs called from different methods compared to the simulated TADs with different noise value. Figure (a), (b), (c), (d) represent the comparison with 4, 8, 12, 16 noise simulated data, respectively.

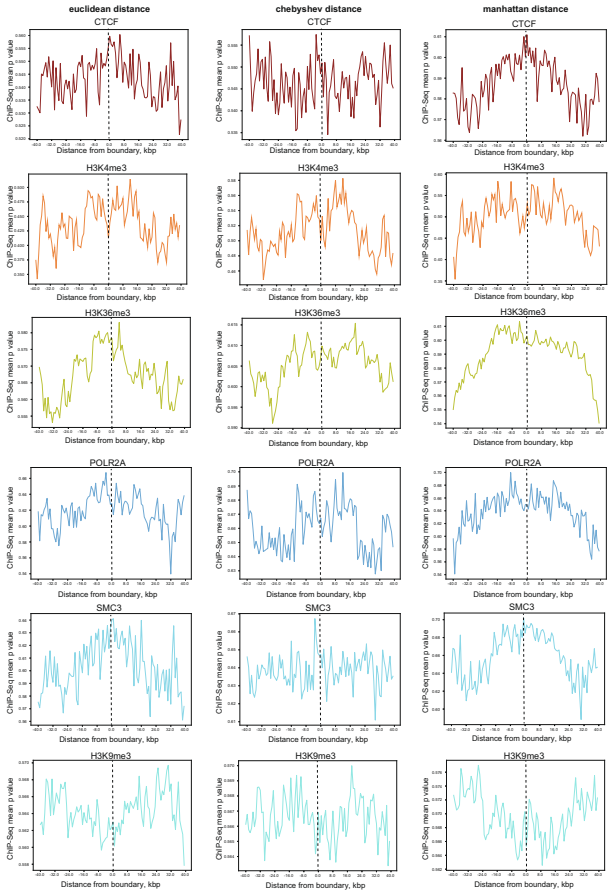

Fig. 4. The plots of mean P-value calculated from different target factor ChIP-seq surrounding the genomic locus of identified TAD boundaries from −40 kb to + 40 kb genomic distance. From left to right, represents the identified TADs results by using the Euclidean, Chebyshev, and Manhattan distance metrics, respectively. From top to bottom, represent the CTCF, H3K4me3, H3K36me3, POLR2A, SMC3, and H3K9me3 target factors ChIP-seq. The Hi-C contact matrices used in the plots are from all chromosomes (from chromosome 1 to chromosome 22) in the GM12878 cell line at 50 kb resolution.

To assess the performance of CASPIAN using different distance metrics on different normalization datasets, as Fig. 5 shows, we compared the number of detected TADs, the anchor ratio of factors including the CTCF, POLR2A, RAD21, SMC3, H3K4me3, H3K36me3, and H3K9me3 within 20 kb genomic distance. Fig. 5a shows that CASPIAN using the Manhattan distance metric detects more TADs than using the other two distance metrics for raw or VC-normalized Hi-C contact matrix, CASPIAN using the Chebyshev distance metric detects more TADs than using the other two distance metrics for KR-normalized Hi-C contact matrix. However, from the aspect of the anchor ratio of the CTCF factor, Fig. 5b shows that CASPIAN using the three distance metric have a similar anchor ratio (all close to 0.6) of CTCF for VC-normalized Hi-C contact matrix, CASPIAN using the Euclidean and Manhattan distance metric performs better than the Chebyshev distance metrics for the raw and KR-normalized Hi-C contact matrix. Hence, from the aspect of TADs count and CTCF anchor ratio, for the raw and VC-normalized Hi-C contact matrix, CASPIAN using the Manhattan distance metric performs better than the other two distance metrics. For the KR–normalized Hi-C contact matrix, the Chebyshev distance metric performs better.

We also performed the comparison anchor ratio of H3K36me3, H3K9mes (Fig. 5c, d), H3K4me1, H3K4me3, RAD21, POLR2A, and SMC3 (Figure S8). The results show that CASPIAN using the Manhattan distance metric could always anchor more factors related to promoting the gene expression (H3K4me1, H3K4me3, RAD21, POLR2A, and SMC3) for raw and KR-normalized Hi-C, CASPIAN using the Euclidean distance metric could always anchor more factors related to promoting the gene expression for VC-normalized Hi-C (Fig. 5c, Figure S8), CASPIAN using the Chebyshev distance metrics anchor more factors (H3K9me3) related to inhibiting the gene expression (Fig. 5d) for the three types of Hi-C contact matrices.

3.2. Assessment of CASPIAN on Simulated Hi-C datasets

We next evaluated the performance of CASPIAN by comparing the results on simulated 40 kb resolution Hi-C data at 4, 8, 12, and 16 noise levels with other five methods: Insulation Score [10], TopDom [14], IC-Finder [25], HiCSeg [15], and ClusterTAD [29]. We didn't compare Directionality Index (DI) [2] with CASPIAN on simulated Hi-C contact matrix due to the input Hi-C data format (.cool or.hic format) requirment of DI method. The true TADs and TADs identified by the six methods are outlined on the heatmap. The heatmaps at 4 noise level (Fig. 2) show that CASPIAN using the Euclidean distance metric is more similar to the true TADs than the other methods. The heatmaps at 8, 12, and 16 noise levels (Figure S1-S3) show that CASPIAN using the Euclidean or Manhattan distance metric, TopDom, Insulation score, HiCSeg, and IC-Finder are all similar to the true TADs.

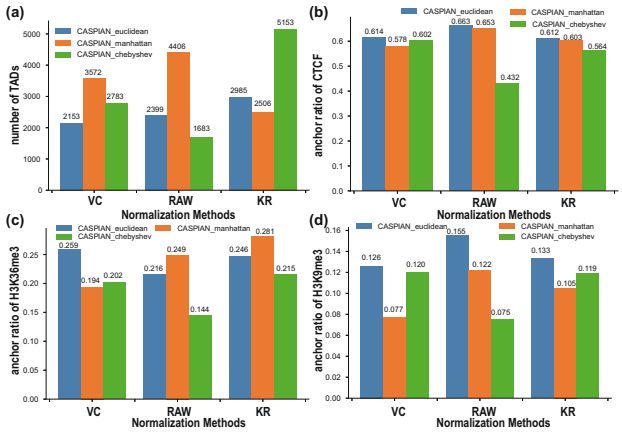

Fig. 5. The comparison of TAD boundary detection results from raw, VC-normalized, and KR-normalized Hi-C contact matrix of all chromosomes (from chr1 to chr22) in GM12878 cell on 50 kb resolution using different methods using three different distance metrics. (a). Comparison of the number of identified TADs. (b - d). Comparison of the anchor ratio of CTCF, H3K36me3, and H3K9me3 within 20 kb genomic distance. Blue represents the results of CASPIAN using the Euclidean distance metric. Orange represents the results of CASPIAN using the Manhattan distance metric. Green represents the results of CASPIAN using the Chebyshev distance metric.

When we compared the identified TAD boundaries with the simulated TAD boundaries from the aspect of cluster similarity, values of the RI and FMI metrics show that CASPIAN using the Euclidean distance metric had the highest cluster similarity compared with the true TADs when noise levels 4, the same RI and FMI values with TopDom when noise levels 8 and 12. Therefore, CASPIAN using the Euclidean distance metric performs better than the other existing methods in terms of cluster quality compared with the simulated TADs.

3.3. Assessment of CASPIAN on Real Hi-C Datasets

To assess the performance of CASPIAN on real Hi-C datasets even further, we tested CASPIAN, Insulation Score [10], TopDom [14], IC-Finder [25], HiCSeg [15], Directionality Index (DI) [2] ClusterTAD [29] on the KR-normalized Hi-C contact matrics (from chromosome 1 to chromosome 22) of the GM12878 cell line at 50 kb resolution. We tested Directionality Index (DI) [2] on the cool file of Hi-C data (from chromosome 1 to chromosome 22) of the GM12878 cell line at 50 kb resolution. The results of Insulation Score, TopDom, ClusterTAD, IC-Finder, and HiCSeg were calculated using the recommended parameters in their published code. The result of DI was calculated using the fanc directionality tool (https://vaquerizaslab.github.io/fanc/) and HMM_calls.m provided by PSYCHIC [16]. The results of CASPIAN were generated by choosing the Euclidean, Manhattan, and Chebyshev distance metrics, respectively.

Firstly, we plotted the KR-normalized Hi-C contact heatmaps with detected TAD boundaries of chromosome 1 from bin 140 to bin 250 in the GM12878 cell line at 50 kb resolution. The plots (Fig. 6) show that CASPIAN using the Euclidean distance metric has a more accurate visualization performance than the other methods. By comparing the visualization performance of TADs identified by CASPIAN using different distance metrics, we found that CASPIAN using the Euclidean metric and Chebyshev distance metric is more likely to identify tiny TADs, CASPIAN using the Manhattan metric is more likely to identify TADs with big size. By overlapping these TADs identified by CASPIAN using the three distance metrics (the top three heatmaps of Fig. 6), we found some TADs identified by CASPIAN using the Chebyshev or Euclidean metric are sub-TADs of TADs identified by CASPIAN using the Manhattan metric. Therefore, we can also identify hierarchical TADs by combining the TAD results identified by CASPIAN using the three distance metrics.

Secondly, we observed the distribution of the mean P-value of CTCF target ChIP-seq surrounding the TAD boundaries of TADs detected by the above methods. The results (Fig. 7) show that CASPIAN using the Manhattan distance metric gets the highest mean P-value (0.61), CASPIAN using the Euclidean distance metric gets a similar mean P-value with HiCSeg, TopDom, Insulation Score, and DI. This means that CASPIAN using the Manhattan distance metric could enrich more CTCF factors than other methods in the domains of the TAD boundaries. We also observed the distribution of the mean P-value of H3K4me3 (Figure S9), H3K36me3 (Figure S10), POLR2A (Figure S11), SMC3 (Figure S12), H3K9me3 (Figure S13), H3K4me3 (Figure S9). The results show that the TAD

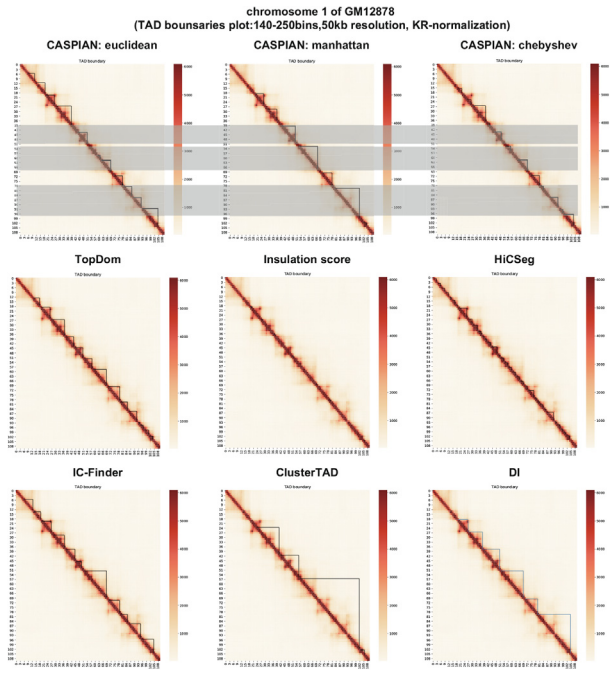

Fig. 6. Heatmaps with detected TAD boundaries using different methods based on a real 50 kb resolution Hi-C contact matrix, where the identified TADs are outlined by black lines.

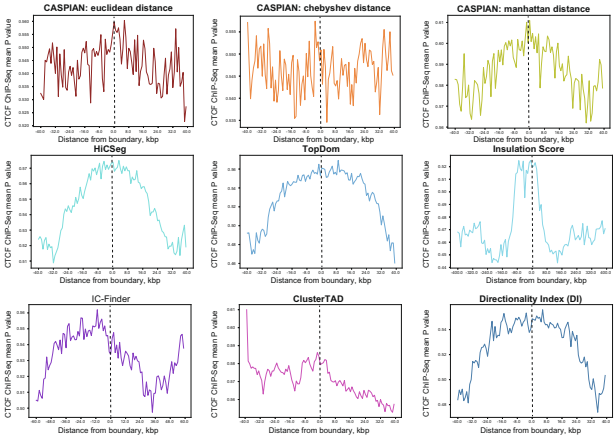

Fig. 7. The plots of mean P-value calculated from CTCF target ChIP-seq around the genomic locus of identified TAD boundaries of all chromosomes (from chr1 to chr22) in GM12878 cell on 50 kb resolution using different methods from −40 kb to + 40 kb genomic distance. The plot of the Insulation Score is calculated from −400 kb to + 400 kb.

boundaries identified by CASPIAN are enriched with factors related to promoting the gene expression.

Thirdly, we compared CASPIAN with other methods on the number of TADs, the average length of TADs, and the anchor ratio of different factors. The TADs were detected from Hi-C data of all chromosomes (from chromosome 1 to chromosome 22) in the GM12878 cell line at 50 kb resolution. The results show that CASPIAN obtains about 2000–5000 TADs with TAD lengths ranging from 0.4 Mb to 2 Mb (Fig. 8a,b), where the average TAD size (0.4 Mb - 2 Mb) is reasonable as study [36] showed. CASPIAN using the Euclidean distance metric identified about 3000 TADs with avergae TAD size of 0.81 MB. This result is similar to the TADs identified by DI method. As article [31] described, the number and size of identified TADs are correlated with the resolution, sequencing

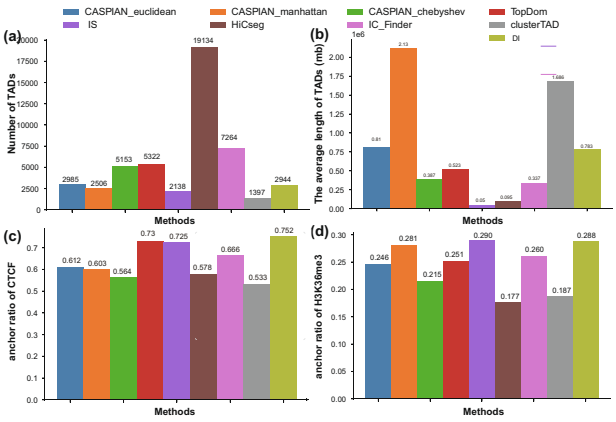

Fig. 8. The comparison of TAD boundary detection results of all chromosomes (from chr1 to chr22) in GM12878 cell at 50 kb resolution using different methods. (a). Comparison of the number of detected TADs. (b) Comparison of the average length of TADs. (c, d) Comparison of the anchor ratio of CTCF and H3K36me3 target factors.

depth of the Hi-C data and TAD callers. At 50 KB resolution, the average TAD size ranged from 215 kb to 1.2 Mb. Most tools identified more tiny TADs at a higher resolution. Fig. 8a,b show that CASPIAN using the Euclidean or Chebyshev distance metric tends to identify tiny TADs, while CASPIAN using the Chebyshev distance metric identified more tiny TADs (about 5000) than CASPIAN using the Euclidean metric (about 3000). CASPIAN using the Manhattan metric tends to identify TADs with a big size. Therefore, we can choose CASPIAN using the Chebyshev distance metric to identify more tiny TADs. These TAD callers including CASPIAN have a anchor ratio of CTCF about 0.6–0.7 (Fig. 8c), H3K36me3 0.2 (Fig. 8d), H3K4me1 0.4, H3K4me3 0.3, H3K9me3 0.1, POLR2A 0.3, RAD21 0.4–0.6 and SMC3 0.3–0.5 (Figure S14). Though CASPIAN doesn't always anchor the highest ratio of the above factors, we could get an approximate anchoring ratio for various factors by observing the distribution of these anchor ratios provided by different TAD callers. These results further indicate that a large number of transcription factors related to gene transcription or histone modified ChIP-seq signals are enriched near the TADs boundary.

Fourthly, to examine the similarity between TADs called by CASPIAN and other methods, we calculated the RI and FMS values between the TADs called by different methods. As Fig. 9 shows, we calculated the RI and FMS values between TADs separately called by CASPIAN using the Euclidean, Manhattan, and Chebyshev distance metric and TADs called by other existing methods. The comparison results among TADs separately called by CASPIAN using the Euclidean, Manhattan, and Chebyshev distance metrics show that CASPIAN using the Euclidean metric could always recall the TADs called by the other two metrics, with the highest RI and FMS values. The comparison results between TADs called by CASPIAN and other methods show that CASPIAN could always recall the TADs called by Insulation score and TopDom. Particularly, CASPIAN using the Manhattan distance metric has the highest RI value of 0.9 and the highest FMS value of 0.93 compared with the Insulation score, and CASPIAN using the Euclidean distance metric has the highest RI value of 0.72 and the highest FMS value of 0.83 compared with the Insulation score. By comparing the p values between RI (or FMS) calculated by CASPIAN and other methods, we found that the similarity (RI and FMS) between CASPIAN using the Euclidean metric and CASPIAN using the Manhattan metric is similar with the similarity between CASPIAN using the Euclidean metric and IS, the similarity (FMS) between CASPIAN using the Manhattan metric and TopDom (or clusterTAD, or DI), where p value = 1. The p values were calculated by the Mann–Whitney-

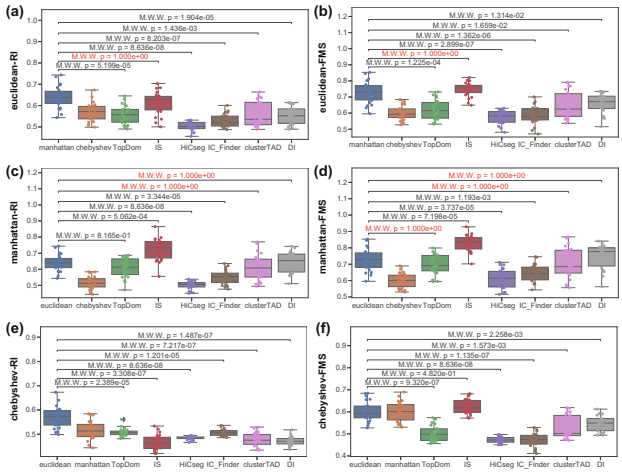

Fig. 9. The similarity between the TAD boundaries identified from CASPIAN and other methods. (a). The box plot describes the Rand Index (RI) between TAD boundaries identified from CASPIAN using the Euclidean distance metric and other methods. (b). The box plot describes the Fowlkes-Mallows score (FMS) between TAD boundaries identified from CASPIAN using the Euclidean distance metric and other methods. (c). The box plot describes the RI value between TAD boundaries identified from CASPIAN using the Manhattan distance metric and other methods. (d). The box plot describes the FMS value between TAD boundaries identified from CASPIAN using the Manhattan distance metric and other methods. (e). The box plot describes the RI value between TAD boundaries identified from CASPIAN using the Chebyshev distance metric and other methods. (f). The box plot describes the FMS value between TAD boundaries identified from CASPIAN using the Chebyshev distance metric and other methods. The p value statistics were calculated by the Mann–Whitney-Wilcoxon test two-sided with Bonferroni correction.

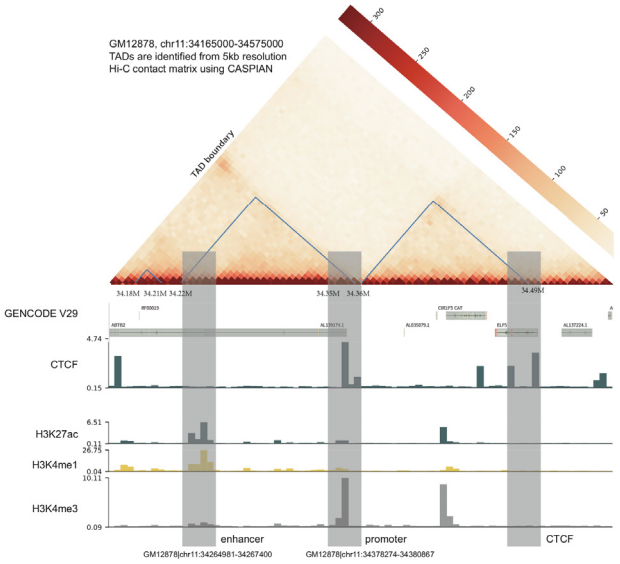

Fig. 10. A plot of Hi-C heatmap with TAD boundaries and CTCF, H3K27ac, H3K4me1, H3K4me3 ChIP-seq track of the area of chr11:34165000–34575000 in GM12878 cell. TADs are identified from 5 kb resolution Hi-C contact matrix using CASPIAN- Euclidian.

Wilcoxon test two-sided with Bonferroni correction. Hence, we could get the conclusion that CASPIAN using the Manhattan distance metric can recall more TADs called by other methods compared with the other two distance metrics.

Lastly, to verify the correctness of TAD boundaries, firstly, we identified the TAD boundaries by running CASPI-AN using different distance metrics from the Hi-C contact matrix of GM12878 cell line at 5 kb, 25 kb, and 50 kb resolution. Then, we downloaded the enhancer-promoter interactions (EPIs) dataset of the GM12878 cell line provided by https://github.com/wgmao/EPIANN/tree/master/GM12878, where the EPIs dataset is always used for training and verifying the accuracy of identifying EPIs. In this paper, we used the EPIs dataset of the GM12878 cell line to verify the correctness and function of TAD boundaries. For example, we plotted the Hi-C heatmap with TAD boundaries and CTCF, H3K27ac, H3K4me1, H3K4me3 ChIP-seq tracks of the area of chr11:34165000–34575000 in GM12878 cell, where GM12878—chr11:34264981–34267400 is enhancer domain, GM12878—chr11:34378274–34380867 is promoter domain, GM12878—chr11:34264981–34267400 and GM12878—chr11: 34378274–34380867 is labeled to be a pair of enhancer-promoter interaction (Fig. 10). We extracted the promoters and enhancers domains of GM12878 from the EPIs datasets, and analyzed the ratio of identified TADs anchoring promoters or enhancers (Table S3 and Table S4). By comparing the anchor ratio of promoters and enhancers, we found these TAD callers including our method (CASPIAN) obtained a $ratio_e$ values about 0.29, a $ratio_p$ values about 0.14. Table S2 and Table S3 futher validated that CASPIAN performed better than other TAD callers in anchoring enhancers and promoters.

4. Conclusion

In this study, we proposed a method named CASPIAN that includes three modules to detect TADs based on the spatial density cluster algorithm. Our CASPIAN method considers three different pairwise distance metrics (Euclidean, Manhattan, and Chebyshev distance metric) to cluster the bins in the Hi-C contact matrix generated by different Hi-C normalization methods. By assessing the CASPIAN methods on the simulated and real Hi-C data, we think the Euclidean distance metric is more suitable for simulated Hi-C data. Fig. 5 and Fig. 7 show that CASPIAN using the Euclidian/Manhattan distance metrics performs better than using the Chebyshev distance metric on the number of detecting TADs and ratio of anchoring CTCF, H3K36me3 from raw and VC-normalized Hi-C. Therefore, we recommend that we can choose CASPIAN using the Manhattan distance metric to detect more TADs with a higher anchoring ratio from raw Hi-C contact matrix without normalization steps, choose CASPIAN using Chebyshev distance metric from KR-normalized Hi-C contact matrix. We do not recommend using CASPIAN to detect TADs from VC-normalized Hi-C contact matrix. In this paper, we provide CASPIAN with three distance metrics to detect TADs at the same time. Therefore, we can also choose the best TAD results by screening from the quality file (number of TAD, anchor ratio of CTCF).

By observing the mean P-value of ChIP-seq and TAD boundaries' anchor ratio of CTCF, POLR2A, RAD21, SMC3, H3K4me3, H3K36me3, and H3K9me3, we conclude that the boundaries of TAD could always anchor more factors related to promoting the gene expression (CTCF, H3K4me1, H3K4me3, RAD21, POLR2A, and SMC3). The comparison of TADs similarity between CASPIAN and other methods shows that CASPIAN could recall the TADs called by Insulation score and TopDom. By comparing the ratio of TAD boundaries anchoring enhancer and promoter elements, we found that 29% of TAD boundaries identifed by most TAD callers anchored enhancers, and 14% anchored promoters. CASPIAN performed better than other methods in anchoring enhancers and promoters.

With the development of bioinformatics research, many TAD callers are developed to identify TAD boudaries. We listed a table (Table S4) to compare these TAD callers on input, parameters, output and resolution that tools could support. Though CASPIAN has the advantages of identifying TADs from high-resolution Hi-C data, anchoring more enhancers and promoters than other TAD callers, few parameters, and providing the visualization of TADs. There still

exist some limitations to its capabilities. The time to detect TADs on high-resolution (e.g., 5 kb) Hi-C data is quite long. Future work may include exploration into improving identification efficiency at high-resolution Hi-C data by investigating algorithms for parallel computing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.csbj.2022.08.059.

References

[1] Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009;326:289–93.
[2] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 2012;485:376.
[3] Berlivet S, Paquette D, Dumouchel A, Langlais D, Dostie J, et al. Clustering of tissue-specific sub-TADs accompanies the regulation of HoxA genes in developing limbs. PLoS genetics 2013;9:e1004018.
[4] Girelli G, Custodio J, Kallas T, et al. GPSeq reveals the radial organization of chromatin in the cell nucleus. Nat Biotechnol 2020;38:1184–93.
[5] Dixon JR, Gorkin DU, Ren B. Chromatin domains: the unit of chromosome organization. Molecular cell 2016;62:668–80.
[6] Lupiáñez DG, Spielmann M, Mundlos S. Breaking TADs: how alterations of chromatin domains result in disease. Trends in Genetics 2016;32:225–37.
[7] Wit ED. TADs as the caller calls them. Journal of Molecular Biology 2019;432.
[8] Montefiori L, Wuerffel R, Roqueiro D, Lajoie B, Guo C, et al. Extremely long-range chromatin loops link topological domains to facilitate a diverse antibody repertoire. Cell reports 2016;14:896–906.
[9] Tang Z et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell 2015;163:1611–27.
[10] Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, et al. Condensin-driven remodeling of X chromosome topology during dosage compensation. Nature 2015;523:240.
[11] Anania C, Acemel RD, Jedamzick J, et al. In vivo dissection of a clustered-CTCF domain boundary reveals developmental principles of regulatory insulation. Nat Genet 2022;54:1026–36.
[12] Yu W, He B, Tan K. Identifying topologically associating domains and subdomains by Gaussian Mixture model And Proportion test. Nature communications 2017;8:535.
[13] Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 2014;159:1665–80.
[14] Shin H, Shi Y, Dai C, Tjong H, Gong K, et al. TopDom: an efficient and deterministic method for identifying topological domains in genomes. Nucleic acids research 2015;44:e70.
[15] Lévy-Leduc C, Delattre M, Mary-Huard T, Robin S. Two-dimensional segmentation for analyzing Hi-C data. Bioinformatics 2014;30:i386–92.
[16] Ron G, Globerson Y, Moran D, Kaplan T. Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. Nature. Communications 2017;8(1):2237.
[17] Weinreb C, Raphael BJ. Identification of hierarchical chromatin domains. Bioinformatics 2015;32:1601–9.
[18] Ron G, Globerson Y, Moran D, Kaplan T. Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. Nature Communications 2017;8:2237.
[19] Chen F, Li G, Zhang MQ, Chen Y. HiCDB: a sensitive and robust method for detecting contact domain boundaries. Nucleic acids research 2018;46:11239–50.
[20] Abbas Roayaei Ardakany, S.L. in 17th International Workshop on Algorithms in Bioinformatics (WABI 2017) Vol. 88 (ed Russell Schwartz and Knut Reinert) (Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2017).
[21] Lyu H, Li L, Wu Z, Wang T, Wang H. TADBD: a sensitive and fast method for detection of typologically associated domain boundaries. BioTechniques 2020;69.
[22] Mourad R. TADreg: a versatile regression framework for TAD identification, differential analysis and rearranged 3D genome prediction. BMC bioinformatics 2022;23:1–14.
[23] Wang Y, Li Y, Gao J, Zhang MQ. A novel method to identify topological domains using Hi-C data. Quantitative Biology 2015;3:81–9.
[24] Chen J, Hero III AO, Rajapakse I. Spectral identification of topological domains. Bioinformatics 2016;32:2151–8.
[25] Haddad N, Vaillant C, Jost D. IC-Finder: inferring robustly the hierarchical organization of chromatin folding. Nucleic acids research 2017;45:e81.
[26] Yan K-K, Lou S, Gerstein M. MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions. PLoS computational biology 2017;13:e1005647.
[27] Norton HK, Emerson DJ, Huang H, Kim J, Phillips-Cremins JE. Detecting hierarchical genome folding with network modularity. Nature Methods 2018;15.
[28] Cresswell KG, Stansfield JC, Dozmorov MG. SpectralTAD: an R package for defining a hierarchy of topologically associated domains using spectral clustering. Bmc Bioinformatics 2020;21. https://doi.org/10.1186/s12859-020-03652-w.
[29] Oluwadare O, Cheng J. ClusterTAD: an unsupervised machine learning approach to detecting topologically associated domains of chromosomes from Hi-C data. BMC bioinformatics 2017;18:480.
[30] Soler-Vila SP, Cuscó P, Farabella I, Di Stefano M, Marti-Renom Marc A. Hierarchical chromatin organization detected by TADpole. Nucleic Acids Research 2020;48:e39. https://doi.org/10.1093/nar/gkaa087.
[31] Dali R, Blanchette M. A critical assessment of topologically associating domain prediction tools. Nucleic acids research 2017;45(6):2994–3005.
[32] Campello RJ, Moulavi D, Zimek A, Sander J. Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Transactions on Knowledge Discovery from Data (TKDD) 2015;10:1–51.
[33] Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nature genetics 2011;43:1059.
[34] Knight PA, Ruiz D. A fast algorithm for matrix balancing. IMA Journal of Numerical Analysis 2013;33:1029–47.
[35] Bäcklund H, Hedblom A, Neijman N. A density-based spatial clustering of application with noise. Data Mining TNM033 2011:11–30.
[36] Rocha PP, Raviram R, Bonneau R, et al. Breaking TADs: insights into hierarchical genome organization[J]. Epigenomics 2015;7(4):523–6.
[37] Prim RC. Shortest connection networks and some generalizations. The Bell System Technical Journal 1957;36:1389–401.
[38] Forcato FM, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. Comparison of computational methods for Hi-C data analysis. Nature methods 2017;14:679–85.
[39] Trussart M, Serra F, Baù D, Junier I, Serrano L, Marti-Renom MA. Assessing the limits of restraint-based 3D modeling of genomes and genomic domains. Nucleic Acids Res 2015;43(7):3465–77.
[40] Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57 (2012).
[41] Pedregosa F et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research 2011;12:2825–30.