

METHOD

Open Access



# SuperTAD: robust detection of hierarchical topologically associated domains with optimized structural information

Yu Wei Zhang, Meng Bo Wang and Shuai Cheng Li\* 

\*Correspondence:

shuaicli@cityu.edu.hk

Department of Computer Science,  
City University of Hong Kong, 83 Tat  
Chee Ave, Kowloon Tong, Hong  
Kong, China

## Abstract

Topologically associating domains (TADs) are the organizational units of chromosome structures. TADs can contain TADs, thus forming a hierarchy. TAD hierarchies can be inferred from Hi-C data through coding trees. However, the current method for computing coding trees is not optimal. In this paper, we propose optimal algorithms for this computation. In comparison with seven state-of-art methods using two public datasets, from GM12878 and IMR90 cells, SuperTAD shows a significant enrichment of structural proteins around detected boundaries and histone modifications within TADs and displays a high consistency between various resolutions of identical Hi-C matrices.

**Keywords:** Topologically associating domain, Hi-C, Structure information theory, Dynamic programming

## Background

The 3D architecture of chromatin plays vital roles in DNA replication and gene transcription process. Many techniques have been devised to capture the architectural information within whole genome [1–9], among which the High-throughput Chromosome Conformation Capture (Hi-C) technique has gained widespread adoption. Hi-C applies high-throughput sequencing to collect fragments that are ligated due to spatial proximity within the genome. Well-established procedures can aggregate and transform sequenced reads into a Hi-C matrix (called *contact map*) at a specific resolution. An element in the Hi-C matrix represents the contact frequency between two fixed-size genome regions (referred to as *bins* or *windows*) of which the indices correspond to the row and column indices in the matrix.

Contact maps have enabled the discovery of architectural units within a chromatin, called *topologically associating domains* (TADs) [10–12]. Genomic regions within a TAD



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

interact with each other more intensely than those between different TADs. TAD can be further subdivided into sub-topologies (sub-TADs), and proximal TADs can aggregate into a higher-order structural domain (meta-TAD), resulting in a hierarchy of the TADs. The size of a TAD varies from thousands of base pairs (kbps) to several million base pairs [10]. TADs are the basic unit of both nucleus conformation and gene regulation [11, 13–16]. The boundary between TADs can obstruct the spread of activity and has been shown to enrich inhibiting factors such as CTCF binding sites, cohesin complexes, and housekeeping gene TSSs, SINE retrotransposons [10, 13, 16–18]. More significantly, the alteration of certain TAD boundaries can lead to cancers or developmental disorders [19–23].

Determining TAD boundaries from Hi-C data remains a challenging task [24]. Some computational methods for detecting the hierarchical structure of TADs exist [25–27]. In 2014, Rao et al. [28] proposed Arrowhead algorithm to transform the square domain feature into an arrowhead-shaped feature and detect TAD through corner score. Each TAD is determined independently, and some of them show a hierarchy manner. TADtree [29], the first published algorithm to identify hierarchy, assumes that the signals of the background and TADs are linear. It models and captures the hierarchical structure of TADs with a forest. Haddad et al. [30] proposed a hierarchical clustering approach named IC-Finder for reconstructing TADs. Yu et al. [31] introduced a Gaussian Mixture model And Proportion test (GMAP) algorithm, which is iteratively applied to normalize the Hi-C matrix until no elements of the statistical test are significant, or the domain size is smaller than a pre-specified threshold. Norton et al. [32] proposed a graph-theory-based method named 3DNetMod to detect TADs by maximizing network modularity. Li et al. [33] proposed a method named deDoc that interprets Hi-C matrix as a weighted graph. Then, the problem is to find a partition with minimal structural information (entropy). They proposed a method which, through a top-down greedy recursion of partitioning and clustering, produces a hierarchical structure (called a *coding tree*) of TADs with the minimal *structural entropy*. The algorithm is heuristic and does not guarantee optimal results. As both the graph partitioning and clustering problems are NP-complete, it is hard to obtain a coding tree of minimal structural entropy. Nevertheless, the TAD boundaries inferred by deDoc demonstrated high consistency with Hi-C matrices at different resolutions. This shows structural information theory to hold promises in the discovery of the TADs. The recent proposed OnTAD algorithm [34] applied dynamic programming to identify the TADs from candidate boundaries, which recursively partitioned the genome while maximizing a score function that depicts the contact frequency inside the TAD hierarchy.

In this work, we design optimal algorithms for computing the coding tree of a contact map. While the problem of finding an optimal tree from a general graph is NP-hard, we observe that the graphs which correspond to the contact maps possess specific properties that allow efficient algorithms for finding their coding trees. One such property is that the vertices in a contact map are ordered. As a result, the leaf nodes of the coding tree form a partition of the bins according to the order. Here, we prove that the problem is polynomial-time solvable. Also, we prove a unique property that can significantly reduce the search space. We designed an optimal algorithm using dynamic programming with polynomial time for computing the coding tree of a Hi-C contact map with minimal structural information. We implemented the algorithms into a software package named SuperTAD.

We compare our method with seven existing methods that can infer the TAD hierarchy, namely Arrowhead, TADtree, IC-Finder, GMAP, 3DNetMod, deDoc, and OnTAD (Table 1). The results reveal that the TADs detected by SuperTAD have minimal average structure entropy and the highest average contact density, as well as the highest enrichment of structural proteins at boundaries and histone modifications within TADs. The results of SuperTAD under various resolution matrices give the highest agreement (the average overlapping ratio is 0.945 for GM12878 cell line, and 0.95 for IMR90 cell line across 25 kb vs. 50 kb and 50 kb vs. 100 kb).

## Results

### Overview of SuperTAD

SuperTAD implements two variants of our algorithms, one which requires a pre-specified threshold, and one without such a requirement. Both variants find optimal coding trees from a contact map. SuperTAD is an open-source, written in C++, and runs from the command line. It accepts either raw or normalized Hi-C matrix as input (Fig. 1). Given an input matrix, SuperTAD provides two modes for users, corresponding to the two implemented variants. If a user supplies an integer parameter  $h$ , it will construct the optimal coding tree of height at most  $h$ , as SuperTAD( $h$ ). Otherwise, it will construct the optimal tree among all the possible heights. Given an optimal coding tree, we provide a filter to the tree nodes, pruning away non-TAD ones, resulting only in TAD nodes.

To evaluate the similarity between two coding trees  $T$  and  $T'$  resulted from the same contact map, we propose a symmetry metric called *overlapping ratio*, which measures the maximum intersection between two results. In our work, we use the overlapping ratio, an asymmetry metric weighted similarity proposed by Li et al. [33] as well as the measure of concordance (MoC) proposed by Zufferey et al. [27] to quantify the level of agreement in the results called through different methods.

### Comparison with deDoc using simulation data with various noise ratios and sizes

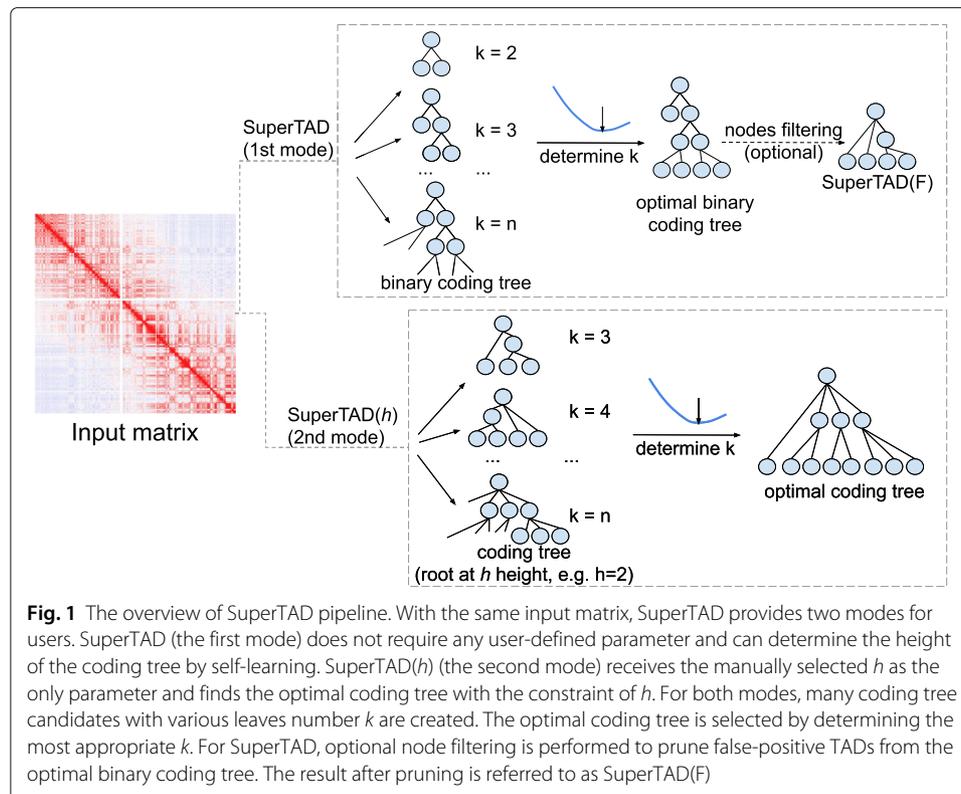
As deDoc and SuperTAD both apply structural information theory, it is interesting to know if they lead to divergent results. We first compared their relative accuracy and robustness. To better quantify the performance, we used simulated data with various noise ratios and TAD sizes and independently executed both approaches 100 times under each setting. As input, we generate an adjacent matrix  $A = \{a_{ij}\}_{N \times N}$ ,  $\forall a_{ij} \in \{0, 1\}$  representation for each graph  $G = \{V, E\}$ .  $G$  contains non-overlapped clusters, and the number of clusters is fixed. If an edge  $(v_i, v_j) \in E(G)$ , then  $a_{ij} = 1$ ; otherwise,  $a_{ij} = 0$ .

To quantify the accuracy and the quality of the results, we compared the overlapping ratio, the weighted similarities between the real structure (reflected in simulated data), and the inferred structures from deDoc or SuperTAD. As weighted similarity is asymmetry, we set  $X$  in  $ws_X^Y$  to the actual structure, and let  $Y$  be the result of deDoc or SuperTAD. We calculated the structural entropy of the coding tree for every result.

First, we examined the influence of noise in the performance of both deDoc and SuperTAD. We fixed the probability of intra-interaction in each cluster and set the size of all clusters (the number of vertices contained in the cluster) to be equal. Then, we tested the probability of inter-interaction from 5 to 50% by 5% of the probability of intra-interaction (denoted as noise ratio). A higher noise ratio implies more edges across different clusters. For each noise ratio, we simulated a matrix as input to both algorithms. This is repeated

**Table 1** Properties of methods for detecting hierarchies of TADs

	SuperTAD	OnTAD	deDoc	3DNetMod	GMAP	IC-Finder	TADtree	Arrowhead
Last updated	2020	2019	2018	2018	2017	2017	2016	2014
Journal	–	<i>Genome Biology</i>	<i>Nature Communications</i>	<i>Nature</i>	<i>Nature Communications</i>	<i>Nucleic Acids Research</i>	<i>Bioinformatics</i>	<i>Cell</i>
Input	Raw/normalized	Raw/normalized	Raw/normalized	Normalized	Normalized	Raw/normalized	Normalized	Raw/normalized
Model	Structural information theory, dynamic programming	Adaptive local minimum search, contact density signal maximization	Structure information theory, greedy merging and combining	Network modularity maximization, hierarchical spatial variance minimization	Gaussian mixture model, proportion test	Hierarchical clustering, local directionality index (DI)	Boundary index (BI), dynamic programming, weighted interval scheduling with multiplicities (WISM)	Arrowhead transformation
Levels of TAD	Not limited	Not limited	2	Not limited	Controlled by parameter	Controlled by parameter	Not limited	Not limited
# Parameters	0	2	0	5	4	1	6	1
Reference	–	[34]	[33]	[32]	[31]	[30]	[29]	[28]



100 times. The simulated structure is a set of non-overlapped TADs (i.e., all at one level). Hence, we only assessed the result of deDoc(E), which is the case where there is only one layer of clusters.

Statistical results of repeated experiments show that the interactions across the clusters have a greater influence on deDoc than SuperTAD. Before the noise ratio reaches 40%, the overlapping ratio and weighted similarity between actual structure and results of SuperTAD are close to 1 (Additional file 2: Figure S6a, blue boxes), suggesting a high level of noise tolerance in SuperTAD. For deDoc, the overlapping ratio and weighted similarity between true structure and result of deDoc began to decrease when noise ratio reaches 15% (Additional file 2: Figure S6a, orange boxes). Note that the median value (Additional file 2: Figure S6a, black solid line in the boxes) of deDoc boxes reaches 0 when the noise ratio is equal to 40%. That is, deDoc failed to discover any cluster from the input matrix. The structural entropy of coding tree detected by both algorithms demonstrated similar distributions (except for some outliers of deDoc) at 5% noise ratio. This shows that both algorithms can detect the clusters with high accuracy when there are few inter-domain interactions. However, at a higher noise ratio, deDoc's solutions have higher structural entropy than SuperTAD's in general.

Next, we examined how the methods perform under different standard deviations on the TAD sizes. We fixed the probability of intra-interaction in each cluster, with a noise ratio of 10%, and assumed the sizes to obey a Gaussian distribution with the same mean. We performed 100 tests with each standard deviation in {1, 2, 3, 4, 5}. For both algorithms, the solutions started to deviate from the true structure when the standard deviation is at or above 3, with deDoc deteriorating faster than SuperTAD (Additional file 2: Figure S6b).

As with the previous tests, deDoc's solutions demonstrated higher structural entropy than SuperTAD's.

### Comparing SuperTAD(2) with deDoc using real Hi-C matrix

In addition to the tests using simulated data, we tested the algorithms with real Hi-C matrices. As deDoc detects TADs with only two levels, we tested it against the second mode of SuperTAD (SuperTAD( $h$ )), setting  $h$  to 2 (referred to as SuperTAD(2)). We downloaded the two in situ Hi-C processed contact datasets (.hic format) from Rao et al. [28]. Both datasets are combined across replicates and filtered with  $\text{MAPQ} \geq 30$ . In the comparison between SuperTAD(2) and deDoc, we selected two bin resolutions 25 kb and 50 kb for assessing the robustness of the algorithms at various resolutions. The raw matrices were normalized with Juicer built-in Knight-Ruiz normalization into normalized Hi-C matrices (referred to as *KR matrix*).

First, we evaluated SuperTAD(2) and deDoc with identical matrix at bin resolution of 25 kb. We compared the distribution of length (size), structural entropy, and contact density of TADs inferred through both methods (Additional file 3: Figure S7a–c). The contact density is defined as the count of intra-TAD contacts divided by TAD length [33]. Compared to deDoc, TADs of SuperTAD(2) have a higher mean and median value in length, structural entropy, and contact density (the solid line in the box corresponds to the median while the dashed line and number in red corresponds to the mean). Next, we compare the structure entropy of the coding tree across various cell lines and bin resolutions. As shown in Additional file 1: Figure S3d, SuperTAD(2)'s solution always has less structure entropy for each comparison, which indicates SuperTAD(2) encoded the input data with lower uncertainty.

It has been previously reported that TAD boundaries are positively associated with the enrichment of the CCCTC-binding factor (CTCF) and members of the cohesin protein complex, such as RAD21 and SMC3 [10, 13]. We downloaded the IDR peaks data of transcription factor (TF) ChIP-seq from ENCODE and computed the fold change of peak enrichment between TAD boundaries and background for each structural protein (see the “Methods” section). We noticed that the boundary inferred by SuperTAD(2) has a greater fold change than deDoc for both cell lines (Additional file 3: Figure S7e)

Considering the fact that some Histone H3 modifications indicating the transcriptional activity are associated with TADs, we next evaluated the enrichment of Histone H3 modifications within detected TADs. We chose the repressing (H3K27me3) and activating (H3K36me3) marks as they exhibit a good mutual exclusion and can indicate either active or repressed transcriptional domains on a well mappable part of the genome [11, 12, 28, 35]. We downloaded the fold change over control data of ChIP-seq from ENCODE and calculated the observed average log10 ratio (LR) of H3K27me3/H3K36me3 and the empirical  $p$  value for each TAD. Based on the FDR-corrected  $p$  value, we identified the TADs that significantly enriched for either mark (FDR-corrected  $p$  value  $\leq 0.1$ ) from those with no significant enrichment (FDR-corrected  $p$  value  $> 0.1$ ). Then, based on the observed LR value, we further divided the TADs from the former group (FDR-corrected  $p$  value  $\leq 0.1$ ) into two sets, one is enriched in H3K27me3, the other is enriched in H3K36me3. The TADs inferred by SuperTAD(2) have a higher fraction of TADs enriched for histone modifications than deDoc (Additional file 3: Figure S7f).

We evaluated the robustness of the algorithms by comparing the similarity of the detected TADs for the same cell type across different bin resolutions. Resultant heatmaps for the 25-kb and 50-kb resolution matrix for both cell lines are shown with the detected TAD boundaries in Additional file 3: Figure S7g and h. To better quantify the similarity, we calculated the overlapping ratio, weighted similarity, and the MoC between the two results (Table 2). The result shows that the agreement of SuperTAD(2) is lower ( $-0.02$ ) for GM12878 cells and higher ( $+0.08$ ) for IMR90 cells than deDoc across all metrics.

As deDoc has been reported to work well with raw Hi-C matrices (i.e., without normalization), we performed further experiments to assess the similarity between results from raw and KR matrices. The heatmaps with the detected boundaries from both raw and KR matrices are shown in Additional file 3: Figure S7g, h. The overlapping ratio, weighted similarity, and MoC are as shown in Table 2. The comparison shows that compared to deDoc, SuperTAD(2) has higher consistency between its results from raw and KR matrices.

### Comparison of SuperTAD with existing methods for detecting hierarchies of TADs

The results thus far suggest that, under the two-layer constraint, SuperTAD(2) performs better than deDoc in terms of accuracy and robustness. We further investigate the performance of SuperTAD without the constraint. Many methods are able to determine the number of layers to use naturally from the input matrix. SuperTAD is able to do the same in the first mode. We compare SuperTAD (the first mode) with seven existing methods, namely OnTAD, deDoc, 3DNetMod, GMAP, IC-Finder, TADtree, and Arrowhead (Table 1). The analysis is performed on Hi-C data sets of two human cell lines (GM12878 and IMR90), the same as the last section. We selected a segment from chromosome 6 (Chr 6: 20000.0–30000.0 kb) for evaluation and comparison among all the methods.

### Comparison of length, structural entropy, and contact density of inferred TADs

We first compared the distribution of length (size), structural entropy, and density of the TADs from each method. As input, we use KR contact map at 25-kb bin resolution. Note that when constructing the coding tree for calculating structure entropy of the TADs for

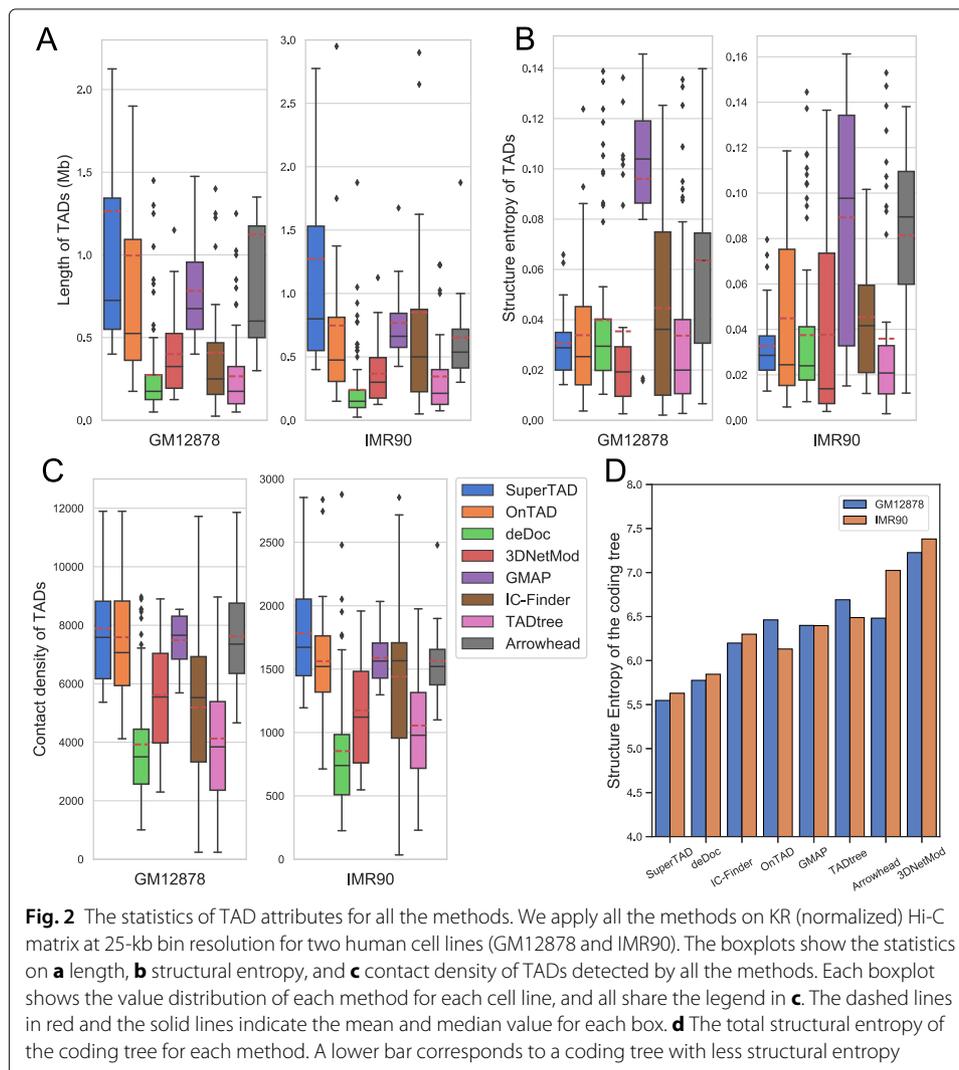
**Table 2** Assessment of similarity criteria between the results of various resolutions and raw/normalized matrix

			SuperTAD(2)	deDoc
GM12878	25 vs. 50 (kb)	OR	0.81	0.83
		WS	0.85	0.87
		MoC	0.71	0.73
	Raw vs. KR (average between 25 and 50 kb)	OR	0.92	0.83
		WS	0.93	0.88
		MoC	0.90	0.77
IMR90	25 vs. 50 (kb)	OR	0.88	0.79
		WS	0.90	0.83
		MoC	0.75	0.67
	Raw vs. KR (average between 25 and 50 kb)	OR	0.93	0.88
		WS	0.95	0.91
		MoC	0.89	0.83

The larger value is labeled in red for each line  
OR overlapping ratio, WS weighted similarity, MoC measure of concordance

Arrowhead and 3DNetMod, we discarded the small TADs that are incompatible with the formed coding tree (these methods allows for overlapping across TADs, which conflict with the definition of the coding tree). Additionally, we use all the identified TADs from Arrowhead and 3DNetMod for the other analysis. The length of TADs inferred by SuperTAD has a broader range for both cell lines (Fig. 2a), which agrees with the hierarchical property of TADs. The TADs inferred by SuperTAD have the minimal mean value (the dashed line in red in boxes) of structural entropy for both cell lines (Fig. 2b). OnTAD and deDoc have a median value of structural entropy similar to SuperTAD but also a higher variance. SuperTAD has the highest mean value of contact density for both cell lines and the highest median value for IMR90 cells. OnTAD, GMAP, and Arrowhead also have a higher mean value of contact density for both cell lines (Fig. 2c). The conserved performance of SuperTAD between both cell lines proves that the TADs inferred by our method are highly self-dense structures.

To assess the uncertainty embedded in the detected coding tree, we further computed the structural entropy of the whole coding tree for each method. SuperTAD gives the



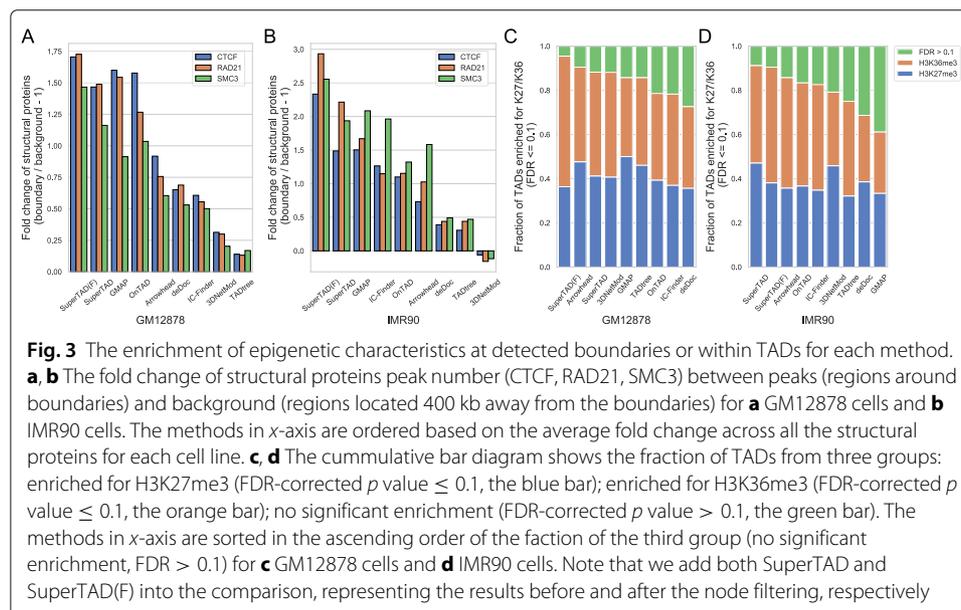
coding trees with the minimum structural entropy for both cell lines (Fig. 2d), with deDoc ranked in the second place.

### Significant enrichment of epigenetic characteristics at SuperTAD-detected boundaries or within TADs

Growing evidence shows that TAD may serve as the fundamental unit of gene regulation. The boundaries between TADs can obstruct the spread of activity within the genome. To validate the TAD boundaries inferred from each method, we downloaded the IDR peaks data of Transcription Factor (TF) ChIP-seq from ENCODE (<https://www.encodeproject.org/>) and computed the fold-change enrichment between detected boundaries and the background region. The data include the CCCTC-binding factor (CTCF), RAD21, and SMC3.

We computed the average number of peaks for each structural protein for every 5 kb. The region around the identified TAD boundary ( $\pm 1$  bin) are referred to as peaks while the 100-kb-length region located 400 kb away from the boundaries at both sides are referred to as background. We then calculated the ratio between the average number of peak and background and then let fold change as the ratio  $-1$ . We also added the selected TADs after pruning the optimal coding tree (referred to as SuperTAD(F)) into the comparison. The selected TADs have a higher structure entropy than their parent and a high probability of being self-dense from a random experiment with 1000 times simulations. As shown in Fig. 3a, SuperTAD(F) has the greatest fold change for all the structural proteins, while SuperTAD ranks the second. The results on the IMR90 dataset have the same trend (Fig. 3b).

To quantify the enrichment of Histone H3 methylation marks within TADs, we computed the observed average log10 ratio (LR) between H3K27me3/H3K36me3 as well as the empirical  $p$  value. Based on the FDR-corrected  $p$  value and observed LR value, we classified the TADs into three groups, TAD enriched for H3K27me3 (FDR-corrected  $p$  value  $\leq 0.1$ ), TAD enriched for H3K36me3 (FDR-corrected  $p$  value  $\leq 0.1$ ), and neither (FDR-corrected  $p$  value  $> 0.1$ ). SuperTAD(F) shows the minimal fraction of the TADs that



enriched for neither of the histone marks for GM12878 cells (Fig. 3c). SuperTAD ranks the second for GM12878 cells, while the first for IMR90 cells. Note that both SuperTAD and SuperTAD(F) exhibit a consistent performance in the overall enrichment comparisons, indicating the TAD inference's validity.

#### **High consistency between SuperTAD results at different resolutions**

To assess the consistency of the results' output by the algorithms at various resolutions of identical data, we tested them using Hi-C matrices at 25-kb, 50-kb, and 100-kb bin resolutions.

We show the Hi-C matrices' heatmap with the inferred boundaries of both results at 25 kb vs. 50 kb in Additional file 4: Figure S8 (for 50-kb vs. 100-kb heatmap, see Additional file 1: Figure S1). As can be observed, SuperTAD, SuperTAD(F), and deDoc are relatively consistent on both the GM12878 (the top line) and the IMR90 (the bottom line) cell lines. Arrowhead has a higher divergency for IMR90 than GM12878 cells, indicating it suffers from the relative lower depth of the data. TADtree has many duplications around the boundaries, resulting in vast disagreement between its results at different resolutions. OnTAD and IC-Finder show similar across 25 kb and 50 kb for GM12878 cells. However, their identified TADs at higher levels show high divergency in IMR90 cells, which implies that both OnTAD and IC-Finder are susceptible to the depth of data. For GMAP, the size of TADs inferred at 50-kb resolution is much larger than that at 25-kb resolution for both cell lines (GMAP fails to detect TADs from the identical input at 100-kb resolution). In this test, 3DNetMod performed the poorest. Its result at 25-kb resolution has as many duplications around the boundaries as TADtree, and it was unable to detect any boundary at 50-kb (as well as 100 kb) resolution (since 3DNetMod filters out all the regions when detecting "good regions," and the algorithm cannot determine the value of maximum gamma afterward).

We show in Table 3 the overlapping ratio, weighted similarity, and MoC for the similarity quantitative assessment. We compute the agreement of 25 kb vs. 50 kb and 50 kb vs. 100 kb. Among all the methods, SuperTAD shows the highest consistency in its results at various resolutions. Node filtering in SuperTAD(F) degrades the consistency to some extent for both cell lines (SuperTAD(F) ranks the second for GM12878 cells and the third for IMR90 cells). deDoc and IC-Finder also show high agreement across resolutions through overlapping ratio and weighted similarity for both cell lines. As both SuperTAD and deDoc are structural information theory-based algorithms, self-consistency at different resolutions may be advantageous for this approach.

## **Discussion**

We have demonstrated the usefulness of our proposed algorithm, SuperTAD. In our experiments, SuperTAD outperformed the existing methods in terms of our new metric (overlapping ratio), as well as in robustness and self-consistency.

For further work, we plan to improve SuperTAD in the following aspects. First of all, the complexity of the dynamic programming used in SuperTAD remains relatively high. Some heuristics can be employed to significantly reduce the running time and memory requirement. A relatively difficult obstacle is that the current definition of the coding tree does disallow for the TADs to overlap. A challenging future work is to devise novel strategies that will allow us to identify the overlap between TADs.

**Table 3** Assessment of similarity criteria between results of 25-kb vs. 50-kb and 50-kb vs. 100-kb resolutions for all methods

		SuperTAD	SuperTAD(F)	OntTAD	deDoc	3DNetMod	GMAP	IC-Finder	TADtree	Arrowhead
GM12878	25/50 <sup>1</sup>	OR	0.96	0.91	0.88	N/A <sup>2</sup>	0.68	0.92	0.64	0.64
		WS	0.97	0.94	0.90	N/A	0.78	0.94	0.75	0.78
		MoC	0.88	0.84	0.78	0.74	0.69	0.76	0.55	0.44
50/100		OR	0.93	0.74	0.82	N/A	N/A	0.77	0.56	0.67
		WS	0.95	0.86	0.85	N/A	N/A	0.84	0.79	0.65
		MoC	0.80	0.60	0.70	N/A	N/A	0.63	0.59	0.17
IMR90	25/50	OR	0.97	0.80	0.85	N/A	0.57	0.76	0.60	0.4
		WS	0.98	0.93	0.88	N/A	0.75	0.86	0.75	0.47
		MoC	0.91	0.75	0.70	N/A	0.66	0.66	0.53	0.34
50/100		OR	0.93	0.62	0.81	N/A	N/A	0.91	0.56	0.92
		WS	0.95	0.81	0.84	N/A	N/A	0.95	0.73	0.87
		MoC	0.81	0.59	0.69	N/A	N/A	0.80	0.61	0.54

The top 2 largest values are labeled in red for each line

OR overlapping ratio, WS weighted similarity, MoC measure of concordance

<sup>1</sup>25/50: agreement of 25 kb vs. 50 kb; 50/100: agreement of 50 kb vs. 100 kb

<sup>2</sup>N/A: lack of as least one result

## Conclusions

In this article, we proposed SuperTAD, a novel method to find the optimal coding tree with the minimum structural entropy from the Hi-C matrix. A coding tree represents a hierarchical structure of TADs. SuperTAD operates in two different modes in restricting the size of the coding tree, namely by the number of leaves, or by tree height. The first mode, SuperTAD, requires no user-defined parameters, while the second mode, SuperTAD( $h$ ), requires a parameter  $h$  to determine the number of layers. Both modes run in polynomial time and find a globally optimal solution to the coding tree problem. In our experiments, SuperTAD performed better than existing methods in our metrics, as well as in robustness and self-consistency. Furthermore, the coding trees computed are proved to be biologically meaningful.

## Methods

### Hi-C data collection

We download the processed Hi-C contacts (.hic format) of two human cell lines (GM12878 and IMR90) from NCBI with accession number GSE63525 [28], and both are in situ Hi-C protocol datasets. GM12878 dataset has 4.9 B contacts while IMR90 dataset has 1.1 B contacts. The contacts are merged across primary and replicates with a filtering  $\text{MAPQ} \geq 30$ . The raw matrix is further normalized by Juicer [36] built-in KR (Knight-Ruiz) normalization as normalized Hi-C matrix (referred to as *KR matrix*).

### ChIP-seq data collection and analysis

To obtain the enrichment information of epigenetics characteristics, we downloaded the Transcription Factor (TF) ChIP-seq from ENCODE (<https://www.encodeproject.org/>). For structural proteins like CCCTC-binding factor (CTCF), RAD21 and SMC3, we downloaded the optimal IDR thresholded peaks. And for histone modifications H3K27me3 and H3K36me3, we downloaded the fold change over control signals. The experiment accession numbers are summarized in Additional file 1: Table S1.

To assess the enrichment of structural proteins around TAD boundaries, we firstly summed the ChIP-seq peaks into 5-kb intervals around boundaries. Then, we calculated the average peak number of the intervals from two regions, one is the region surrounding the boundaries (the bin detected as boundary and  $\pm 1$  bin, referred to as *peak*), the other is the 100-kb region located 400 kb away from the boundaries at both sides (referred to as *background*). The TAD boundaries are defined as the ends of TADs. We computed the fold change between the average peak number of peak and background per TAD and took the average. Zero value of the average fold change stands by no enrichment around boundaries and a higher value means the boundaries are enriched for the structural proteins.

To assess the enrichment of two histone modifications, H3K27me3 (repressing) and H3K36me3 (activating) within TADs, we adopted the modified analysis from the work of Zufferey et al. [27]. We summed the ChIP-seq signals into intervals with fixed length (10% of the resolution). Next, we computed the log10 ratio between H3K27me3 and H3K36me3 for each interval (LR value) and computed the average LR values of intervals within each TAD as the observed LR values. Then, we performed 1000 times shuffling to calculate the empirical  $p$  value for each TAD and corrected the empirical  $p$  value through

false discovery rate (FDR) using the Benjamini-Hochberg (BH) method. With the constraint that FDR-corrected  $p$  value  $\leq 0.1$ , we classified the TADs into two groups, one is enriched for either H3K27me3 or H3K36me3, the other is enriched for neither (FDR-corrected  $p$  value  $> 0.1$ ). We further divided the former group (FDR-corrected  $p$  value  $\leq 0.1$ ) into two subgroups, TADs enriched for H3K27me3 and enriched for H3K36me3 based on each TAD's observed LR value. We reported the fraction of the three clusters. A higher fraction of TADs enriched for either H3K27me3 or H3K36me3 is considered to reflect a more biological meaningful result of the algorithm.

## The SuperTAD algorithm

### Notation and definition

To study TAD or loops, researchers often partition the genome into a sequence of bins or windows, where a bin contains a fix length segment of the genome. Denote the number of bins as  $n$ . Denote a Hi-C matrix as  $X = \{x_{ij}\}$ , where  $x_{ij}$ ,  $1 \leq i, j \leq n$ , is a non-negative real number which represents the interaction frequency between bins  $i$  and  $j$ ; it is often the normalized read count which hit both bin  $i$  and  $j$  simultaneously.  $X$  is symmetric. The diagonal elements are set to be zeros in the matrix. A symmetric matrix is equivalent to a undirected graph. In deDoc [33],  $X$  is interpreted as a weighted graph and the objective is to find a hierarchical structure of TADs where the structural information (entropy) is minimum.

### Coding tree

Structure information theory is proposed to measure the uncertainty embedded in the dynamics of a graph [37]. Finding a partition of the graph with the minimum structural entropy is akin to finding a partition which can best represent the original graph while reducing all the random variation and noise to a minimum. Here, we introduce the definitions in structure information theory that are relevant to our TAD finding problem.

A coding tree  $T$  of  $X$  forms a hierarchical partitioning of the bins of the Hi-C matrix. The coding tree can be multi-nary. Each node of the tree contains (or codes) a set of consecutive bins. The root  $\lambda_T$  represents, or codes, the entire genome. Each tree node codes a subset of consecutive bins along the genome. The children of each tree node partition the bins of their parent node. These partitions are used to define TAD boundaries and each node is a TAD candidate.

Denote the bins represented by a node  $u \in T$  as  $b_T(u)$ , and denote its volume as  $V(u)$ ; that is,  $V(u) = \sum_{i \in b_T(u), j \in b_T(\lambda_T)} x_{ij}$ . The structural entropy of  $u$  is then defined as

$$S_T(X; u) = -\frac{g(u)}{2m} \log_2 \frac{V(u)}{V(p_T(u))}, \quad (1)$$

where  $g(u) = \sum_{i \in b_T(u), j \in b_T(\lambda_T) - b_T(u)} x_{ij}$ ,  $p_T(u)$  is the parent node of  $u$ , and  $2m = \sum_{i, j \in b_T(\lambda_T)} x_{ij}$ . Clearly, if a node contains only one bin,  $g(u) = V(u)$ . Denote the leaf node in  $T$  a bin  $b_i$  belongs to as  $e_T(b_i)$ , let the structural entropy of bin  $b_i$  in  $T$  as

$$S_T(X; b_i) = -\frac{g(b_i)}{2m} \log_2 \frac{V(b_i)}{V(e_T(b_i))}, \quad (2)$$

where  $g(b_i) = \sum_{j \neq i} x_{ij}$ , and  $V(b_i) = \sum_j x_{ij}$

According to the definitions of  $g$  and volume  $V$ , it is clear that the following hold:

**Lemma 1** *If the bins coded by node  $v_1, \dots, v_\ell$  partition the bins coded by  $v$ , then  $\sum_{1 \leq i \leq \ell} g(v_i) \geq g(v)$ , and  $\sum_{1 \leq i \leq \ell} V(v_i) \geq V(v)$ .*

We write  $p_T(u)$  as  $p(u)$  and  $e_T(b_i)$  as  $e(i)$  when the context are clear.

The root  $\lambda_T$  has a structural entropy of 0. The structural entropy  $S_T(X)$  of a coding tree is the sum of the structural entropy of all its nodes and all the bins; that is,

$$S_T(X) = \sum_{u \in T} S_T(X; u) + \sum_{1 \leq i \leq n} S_T(X; i) \tag{3}$$

The *optimal coding tree* is for the matrix  $X$  a tree  $T_{opt}(X)$  with minimal structural entropy. The TAD finding task is then to find an optimal coding tree.

**Finding optimal coding trees**

First, we prove the following results:

**Lemma 2** *The structural entropy of an optimal coding tree with  $k + 1$  leaves is always no more than that of an optimal tree with  $k$  leaves, where  $k$  is an integer.*

*Proof* Assume  $T_a$  is a tree of  $k$  leaves. Without loss of generality, we assume its first leaf  $v$  contains bins 1 to  $\ell$ ,  $\ell \geq 2$ . We transform  $T_a$  into  $T_b$  by: (1) Creating new leaves  $v_1$ , and  $v_2$ , where  $v_1$  codes bins 1, ...,  $j$ ,  $j \leq \ell$  and  $v_2$  codes bins  $j + 1, \dots, \ell$ , (2)  $v_1$  and  $v_2$  are the children of  $v$ . We just need to prove that  $T_b$  has the same or lower structural entropy than  $T_a$ .

$$\begin{aligned} S_{T_a}(X) - S_{T_b}(X) &= -S_{T_b}(X; v_1) - S_{T_b}(X; v_2) + \sum_{1 \leq i \leq \ell} (S_{T_a}(X; i) - S_{T_b}(X; i)) \\ &= \frac{g(v_1) - \sum_{1 \leq i \leq j} g(b_i)}{2m} \log_2 \frac{V(v_1)}{V(v)} + \frac{g(v_2) - \sum_{j < i \leq \ell} g(b_i)}{2m} \log_2 \frac{V(v_2)}{V(v)} \\ &\geq 0. \end{aligned}$$

Therefore,  $S_{T_a}(X) \geq S_{T_b}(X)$ . The optimal tree with  $k + 1$  leaves will have an entropy no more than  $S_{T_b}(X)$ . Hence, our statement holds. □

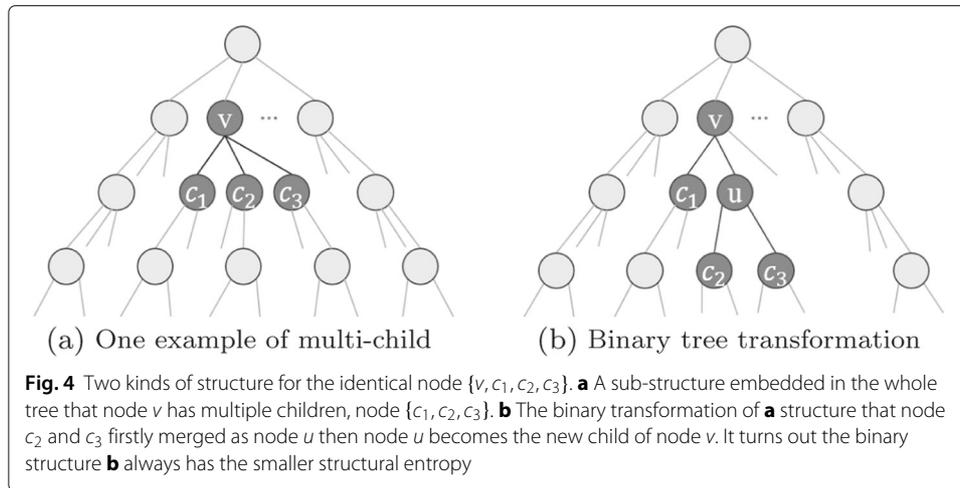
Clearly, due to Lemma 2, we need to restrict the number of leaves to have an optimal coding tree. We assume the number of leaves in the coding tree is  $k$ .

Without loss of generality, we assume that each internal node  $u$  has at least two children. If a node  $u$  is the only child of its parent  $p(u)$ , then  $V_T(u) = V_T(p(u))$ , and  $S_T(X; u) = 0$ , showing  $u$  to be redundant.

We next show that an optimal tree can be required to be binary without loss of generality.

**Lemma 3** *For every contact matrix  $M$ , there exists a binary coding tree of minimum structural entropy.*

*Proof* Given a node  $v$  in a tree  $T_a$  with more than two children,  $c_1, c_2, \dots, c_\ell$ ,  $\ell \geq 3$ , we can transform optimal coding tree  $T_a$  into  $T_b$  such that, (1) in  $T_b$ ,  $v$  has children  $c_1$  and  $u$ ,  $u$  has children  $c_2, \dots, c_\ell$ ; and (2) all the other parts of  $T_a$  and  $T_b$  are the same (see Fig. 4).



We just need to prove that  $T_b$  has no more structural entropy than  $T_a$ .

$$\begin{aligned}
 S_{T_a}(X) - S_{T_b}(X) &= -S_{T_b}(X; u) + \sum_{2 \leq i \leq \ell} S_{T_a}(X; c_i) - \sum_{2 \leq i \leq \ell} S_{T_b}(X; c_i) \\
 &= \frac{g(u) - \sum_{2 \leq i \leq \ell} g(c_i)}{2m} \log_2 \frac{V(u)}{V(v)} \\
 &\geq 0.
 \end{aligned}
 \tag{4}$$

Therefore, the statement holds. □

Note that both Lemmas 2 and 3 hold for general graphs. To find an optimal coding tree, we merely need to search the binary trees. Here, we adopted a dynamic programming approach to find the tree. Let  $S(i : j, k)$  be the structural entropy of the optimal binary coding tree when partitioning bins  $\{b_i, b_{i+1}, \dots, b_j\}$  with  $k$  leaves, denote  $H_l(i_l : i_r, j)$  as structural entropy of the node containing bins  $\{b_{i_l}, b_{i_l+1}, \dots, b_j\}$  with its parent containing bins  $\{b_{i_l}, b_{i_l+1}, \dots, b_{i_r}\}$ ,  $i_l \leq j \leq i_r$ ; and denote  $H_r(i_l : i_r, j)$  as structural entropy of the node containing bins  $\{b_{j+1}, b_{j+2}, \dots, b_{i_r}\}$  with its parent containing bins  $\{b_{i_l}, b_{i_l+1}, \dots, b_{i_r}\}$ ,  $i_l \leq j \leq i_r$ .

Then, we can write the recurrent relations to find the optimal binary coding tree with  $k$  leaf nodes for  $X$

$$S(1 : n, k) = \min_{1 \leq i < n, 1 \leq k_1 < k} \{S(1 : i, k_1) + S(i + 1 : n, k - k_1) + H_l(1 : n, i) + H_r(1 : n, i)\}
 \tag{5}$$

where  $S(1 : n, k)$  is the structural entropy of the optimal binary coding tree with partitioning bins  $\{b_1, b_2, \dots, b_n\}$  with  $k$  leaves.  $k_1$  and  $k - k_1$  are the number of leaves in the left subtree and right subtree, respectively (Additional file 1: Figure S2).

There are  $O(n^3)$  possible  $H_l(1 : n, j)$  and  $O(n^3)$  possible  $H_r(1 : n, j)$  terms, each can be calculated in  $O(1)$  time. Hence,  $O(n^3)$  time is necessary to compute these  $H$  terms. A table of size  $O(kn^2)$  can be created to store the values of  $S(1 : n, k)$ , and each value of  $S(1 : n, k)$  takes time  $O(kn)$ . Hence,

**Theorem 1** *There exists an algorithm that finds the optimal coding tree of  $k$  leaves with time complexity  $O(k^2n^3)$ .*

Also, we may restrict the height of the coding tree. A heuristic algorithm exists for the problem [33]. Here, we solve the problem exactly by a dynamic programming. We propose SuperTAD( $h$ ) which restricts the size of the coding tree by assuming the optimal coding tree is to be of height at most  $h$ . The tree may not be binary and a node can have more than two children. Our dynamic programming is as follows. Let  $T(l : r, p, k, h)$  store the structural entropy of a multi-nary optimal coding tree where (1) the root codes bins  $\{b_l, b_{l+1}, \dots, b_r\}$ ; (2) children nodes partition bins  $\{b_l, b_{l+1}, \dots, b_p\}$ ,  $p \leq r$ ; (3) there are a total of  $k$  leaves; and (4) the height is at most  $h$ . Then, we can write the recurrence relation as follows:

$$T(l : r, p, k, h) = \min_{l \leq i < p, 1 \leq k_1 < k} \{ \min\{T(l : r, i, k_1, h), T(l : i, i, k_1, h - 1) + H_l(l : r, i)\} \\ + T(i + 1 : p, p, k - k_1, h - 1) + H(l : r, i + 1 : p) \} \quad (6)$$

where  $H(l : r, i + 1 : p)$  is the structural entropy of a node which codes the bins  $\{b_{i+1}, b_{i+2}, \dots, b_p\}$ , where its parent node codes bins  $\{b_l, b_{l+1}, \dots, b_r\}$  (Additional file 1: Figure S3).

**Theorem 2** *There exists an algorithm that finds the optimal coding tree of at most height  $h$  and at most  $k$  leaves with time complexity  $O(n^A k^2 h)$ .*

The time complexity can be reduced by extracting a candidate set of TAD boundaries prior to applying the algorithm. This shows that *the optimal coding tree with restricted height problem is polynomial-time solvable*.

#### **Determine the number $k$ of leaves**

As mentioned, structural entropy decreases with an increase of the number of leaf nodes  $k$ . We consider the problem of determining a suitable  $k$ . First, we propose a Bayesian Information Criteria (BIC) approach. Second, we normalize the elbow point at the structural entropy vs.  $k$  curves. Third, we try to compare the structural entropy to a background model; that is, we try to derive the structural entropy in the ideal contact matrix and use it as a normalization factor. A fourth approach is based on our observation that as  $k$  increases, the sum of the structural entropy of the leaf nodes drops at first, but increases after a minimum is reached. We explore using the  $k$  value which corresponds to this minimal entropy for the leaves.

#### **Filter TADs**

Each node in a coding tree gives a potential candidate for defining TADs. We consider the task of filtering out the nodes which are unlikely to be TAD. A node is likely to be a TAD if the intra-interactions are much dense. To eliminate the influence of hierarchy and compute the inherent density for each TAD, we compute the average interaction frequency at three layers: the parent node, the node's children, and the node itself. Starting from the root, we iteratively deduct the influence of the parent and children for each node up to the leaf nodes. In this way, we calculate each node's inherent density from the top to the bottom of the coding tree.

Based on the empirical distributions of contact frequencies in the Hi-C matrix, the contact frequency decreases with the increase of distance. Next, we cluster all the nodes into

two sets based on their inherent density and size, repeating with 1000 times random initialization. We select the set of TADs that shows a strong negative relationship between their sizes and inherent density. We calculate the probability of being selected for each TAD candidate. The candidates that show low probability, lower structure entropy than their parent and close to equally split from their parent are discarded (Additional file 1: “More details in nodes filtering” section, Figure S4, S5).

### Assess the similarity between two coding trees

Given two coding trees  $X$  and  $Y$  of the same Hi-C matrix. Suppose that  $X = \{x_1, x_2, \dots, x_m\}$  and  $Y = \{y_1, y_2, \dots, y_n\}$  where each node  $X_i$  or  $Y_j$  is a consecutive set of bins. We consider evaluating the similarity between  $X$  and  $Y$ .

The work of deDoc [33] proposes the *weighted similarity* between  $X$  and  $Y$ , which is defined as  $ws_X^Y$

$$ws_X^Y = \frac{\sum_{j=1}^n |y_j| \cdot S_X^Y(j)}{\sum_{j=1}^n |y_j|} \quad (7)$$

$$S_X^Y(j) = \max_{i=1}^m \left\{ \frac{|x_i \cap y_j|}{\sqrt{|x_i| \cdot |y_j|}} \right\} \quad (8)$$

However, the definition shows that weighted similarity is an asymmetry metric, and it is hard to determine the similarity when there is a big difference between  $ws_X^Y$  and  $ws_Y^X$ .

Zufferey et al. [27] adopted the measure of concordance (*MoC*), a symmetric metric to compare clustering assignments, which is defined as  $MoC(X, Y)$

$$MoC(X, Y) = \begin{cases} 1, & \text{if } N_X = N_Y = 1 \\ \frac{1}{(\sqrt{N_X N_Y} - 1)} \left( \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \frac{\|F_{ij}\|^2}{\|X_i\| \|Y_j\|} - 1 \right), & \text{otherwise} \end{cases} \quad (9)$$

However, the MoC is upper and lower bounded only if the partitions are disjoint (any two TADs of  $X$  or  $Y$  do not have overlap). To adopt the MoC for assessing the agreement of two hierarchical TAD structures, we only selected the level one of hierarchy (the TAD can no further partition) and added the inter-TAD regions into the assessment as [27].

In this work, we use, in conjunction with the weighted similarity and MoC, a new symmetry metric we call *overlapping ratio* to measure coding trees similarity. First, we build a Bipartite Graph  $G = (V, E)$  in which the vertex set can be partitioned  $V = \{X, Y\}$ , and every edge  $e \in E$  links one node in  $X$  and the other node in  $Y$ . We define the weight of edges as the intersection between the two linked nodes, denoted as  $w(x_i; y_j)$ . Obviously, the graph  $G$  is complete.

Then, we apply Maximum Bipartite Matching to the graph with the goal of finding a maximum matching  $M$  that the summation of selected edges' weight is maximum. That is, we find the global optimal matching for every node in  $X$  and  $Y$ . The overlapping ratio between  $X$  and  $Y$  is defined as the function  $S(X, Y)$

$$S(X, Y) = \frac{\sum_{i=1}^M w(x_i; *) + \sum_{j=1}^N w(*; y_j)}{\sum_{i=1}^M |x_i| + \sum_{j=1}^N |y_j|} \quad (10)$$

where  $w'$  is the edge weight in the maximum matching  $M$  and  $w'(x_i; *)$  is defined as

$$w'(x_i; *) = \begin{cases} w(x_i, y_j), & \text{if edge } e(x_i, y_j) \text{ is selected in } M \\ 0, & \text{none of } e(x_i; *) \text{ is selected in } M \end{cases} \quad (11)$$

The overlapping ratio is symmetric,  $S(X, Y) = S(Y, X)$ . The value of overlapping ratio between any coding trees ranges from 0 to 1, where 1 indicates that the two coding trees are the same while 0 indicates the two coding trees contain no intersection between any pair of  $x_i$  and  $y_j$ .

### The SuperTAD C++ package

SuperTAD is implemented as a command line tool in C++. We compiled and tested our software on both local computers and a Linux server with CentOS 7.6 pre-installed that has 96 12-core processors and 598 GB memory. Our method and software guarantee accuracy while do not sacrifice computational performance. The source codes of SuperTAD package are available at <https://supertad.deepomics.org/>, where the example dataset is also deposited. The version used in the manuscript is permanently available at <https://doi.org/10.5281/zenodo.4314123>.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-020-02234-6>.

**Additional file 1:** Supplementary Information.

**Additional file 2: Figure S6.** The robustness comparison between SuperTAD and deDoc(E) under various noise ratios and sizes. **a** The influence of noise on the performance of both methods. The x-axis indicates the increase in noise ratio from 5 to 50% by 5%, while the boxes show the value distribution of certain metrics among 100 repeated experiments. **b** The influence of variance in TAD size (length) on the performances of both methods. The x-axis indicates the increase in standard deviation of TAD size while the boxes show the value distribution of certain metrics among 100 repeated experiments. The boxes of SuperTAD are colored in blue while deDoc are in orange. The colored line links the mean value (the green point in boxes) across boxes. For boxplots, centerline indicates the median, box limits indicate upper and lower quantiles, whiskers indicate the 1.5 interquartile range, and points indicate outliers.

**Additional file 3: Figure S7.** Comparison between SuperTAD(2) and deDoc using real Hi-C matrix for in situ Hi-C GM12878 and IMR90 cell lines. We first apply SuperTAD(2) and deDoc on KR (normalized) Hi-C matrix of two human cell lines (GM12878 and IMR90) at 25-kb bin resolution. The boxplots show the statistics on **a** length, **b** structural entropy, and **c** contact density of inferred TADs for both methods. The box shows the value distribution of each method for each cell line (blue boxes represent SuperTAD(2) while orange boxes represent deDoc). The marked numbers and dashed lines in red both indicate the mean value for each box. The contact density is defined as the number of intra-TAD contacts divided by TAD length. **d** The structure entropy of the coding tree detected by SuperTAD(2) and deDoc for both cell lines at 25-kb and 50-kb resolutions. **e** The fold change of structural proteins peak number (CTCF, RAD21, SMC3) between peaks (regions around boundaries) and background (regions located 400 kb away from the boundaries). The higher value indicates more enrichment of structural proteins around boundaries. **f** The cumulative bar diagram shows the fraction of TADs from three groups: enriched for H3K27me3 (FDR-corrected  $p$  value  $\leq 0.1$ , the blue bar); enriched for H3K36me3 (FDR-corrected  $p$  value  $\leq 0.1$ , the orange bar); no significant enrichment (FDR-corrected  $p$  value  $> 0.1$ , the green bar). **g, h** The heatmap and inferred boundaries with various inputs for GM12878 and IMR90 cell lines. Each heatmap exhibits different results with two inputs. Text in the upper/lower triangle indicates the input matrix's information, and the plotted boundaries on the same side present the corresponding result. The similarity between boundaries in different colors shows the robustness of performance between 25-kb and 50-kb bin resolution (or raw and KR) matrix for each method. Note that the heatmap is asymmetric when comparing two results from raw and KR matrices.

**Additional file 4: Figure S8.** Consistency comparison for the same cell line with 25-kb vs. 50-kb resolutions among all the methods. The heatmap and detected boundaries with 25-kb and 50-kb bin resolution input for GM12878 (the top line) and IMR90 (the bottom line) cell lines. The detected domains from 25-kb resolution are colored in blue at the upper triangle, and 50-kb resolution results are in pink at the lower triangle (as the texts indicate). The similarity between boundaries in different colors shows the robustness of performance between 25-kb and 50-kb bin resolution matrices for each method.

**Additional file 5:** Review history.

### Acknowledgments

We thank Dr. Yen Kaow Ng for constructive criticism of the manuscript.

**Peer review information**

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

**Review history**

The review history is available as Additional file 5.

**Authors' contributions**

SCL conceived the idea and supervised the work. SCL provided theoretical results. SCL and YWZ designed the SuperTAD package. YWZ and WMB implemented the package. YWZ performed downstream comparison and validation analysis and wrote the draft. YWZ, WMB, and SCL revised the manuscript. All authors read and approved the final manuscript.

**Authors' information**

Twitter handles: @VivianZhangyw (Yu Wei Zhang); @skyvagrant (Shuai Cheng Li).

**Funding**

Publication costs were funded by the GRF Research Projects 9042348 (CityU 11257316). The work described in this paper was also supported by the project.

**Availability of data and materials**

SuperTAD is available at <https://github.com/deepomicslab/SuperTAD> and <https://supertad.deepomics.org/> [38], under MIT license. The versions used in the manuscript are permanently available at <https://doi.org/10.5281/zenodo.4314123> [39].

Hi-C data: The processed Hi-C contacts (.hic format) of two human cell lines, GM12878 and IMR90, are downloaded from Rao et al. [28] (GEO accession number: GSE63525), and both were in situ Hi-C protocol datasets. The raw and KR normalized Hi-C contact maps at 25-kb, 50-kb, and 100-kb resolutions are included in this study.

Epigenomic data: The Transcription Factor (TF) ChIP-seq data of CTCF, cohesin protein complex RAD21, and SMC3 were downloaded from ENCODE project (<https://www.encodeproject.org/>). The optimal IDR thresholded peaks were downloaded in bigBed format. The Histone ChIP-seq data of H3K27me3 and H3K36me3 were also downloaded from the ENCODE project. The fold change over control signals was downloaded in bigWig format.

**Ethics approval and consent to participate**

Not applicable.

**Competing interests**

The authors declare that they have no competing interests.

Received: 14 December 2019 Accepted: 10 December 2020

Published online: 25 January 2021

**References**

- Berkum NLV, Lieberman-Aiden E, Williams L, Imakaev M, Lander ES. Hi-C: a method to study the three-dimensional architecture of genomes. *JoVE*. 2010;39(39):e1869.
- Duan Z, Andronescu M, Schutz K, Mclwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS. A three-dimensional model of the yeast genome. *Nature*. 2010;465(7296):363–7.
- Rodley CDM, Bertels F, Jones B, O'Sullivan JM. Global identification of yeast chromosome interactions using Genome conformation capture. *Fungal Genet Biol*. 2009;46(11):879–86.
- Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen C-A, Schmitt AD, Espinoza CA, Ren B. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*. 2013;503(7475):290–4.
- Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P. *Nature*. 2013;502:59–64.
- Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *2009;326(5950):289–93*.
- de Laat W, Dekker J. 3C-based technologies to study the shape of the genome. *Methods (San Diego, Calif)*. 2012;58(3):189–91.
- Han J, Zhang Z, Wang K. 3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering. *Mol Cytogenet*. 2018;11(1):1–10.
- Übelmesser N, Papantonis A. Technologies to study spatial genome organization: beyond 3C. *Brief Funct Genom*. 2019;18(6):395–401.
- Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. 2012;485(7398):381–5.
- Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*. 2012;485(7398):381–5.
- Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*. 2012;148(3):458–72.
- Dixon J, Gorkin D, Ren B. Chromatin domains: the unit of chromosome organization. *Mol Cell*. 2016;62(5):668–80.
- Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, Vera DL, Wang Y, Hansen RS, Canfield TK, et al. Topologically associating domains are stable units of replication-timing regulation. *Nature*. 2014;515(7527):402–5.
- Ramirez F, Bhardwaj V, Arrigoni L, Lam KC, Gruning B, Villavecies JM, Habermann B, Akhtar A, Manke T. High-resolution tads reveal dna sequences underlying genome organization in flies. *Nat Commun*. 2018;9(1):189.

16. Zhan Y, Mariani L, Barozzi I, Schulz EG, Bluthgen N, Stadler MB, Tiana G, Giorgetti L. Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome Res.* 2017;27(3):479–90.
17. Bintu B, Mateo LJ, Su J, Sinnott-Armstrong NA, Parker M, Kinrot S, Yamaya K, Boettiger AN, Zhuang X. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science.* 2018;362(6413):eaau1783.
18. Rocha PP, Raviram R, Bonneau R, Skok JA. Breaking TADs: insights into hierarchical genome organization. *Epigenomics.* 2015;7(4):523–6.
19. Harewood L, Kishore K, Eldridge MD, Wingett S. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biol.* 2017;18(1):1–11.
20. Lupianez DG, Spielmann M, Mundlos S. Breaking TADs: how alterations of chromatin domains result in disease. *Trends Genet.* 2016;32(4):225–37.
21. Peifer M, Hertwig F, Roels F, Dreidax D, Gartgruber M, Menon R, Kramer A, Roncaioli JL, Sand F, Heuckmann JM, et al. Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature.* 2015;526(7575):700–4.
22. Hnisz D, Weintraub AS, Day DS, Valton A, Bak RO, Li CH, Goldmann J, Lajoie BR, Fan ZP, Sigova AA, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science.* 2016;351(6280):1454–8.
23. Groschel S, Sanders MA, Hoogenboezem RM, De Wit E, Bouwman BAM, Erpelinck C, Der Velden VHJV, Havermans M, Avellino R, Van Lom K, et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell.* 2014;157(2):369–81.
24. Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically associating domains. *Sci Adv.* 2019;5(4):eaaw1668.
25. Dali R, Blanchette M. A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Res.* 2017;45(6):2994–3005.
26. Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. Comparison of computational methods for Hi-C data analysis. *Nat Methods.* 2017;14(7):679–85.
27. Zufferey M, Tavernari D, Oricchio E, Ciriello G. Comparison of computational methods for the identification of topologically associating domains. *Genome Biol.* 2018;19(1):217.
28. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell.* 2014;159(7):1665–80.
29. Weinreb C, Raphael BJ. Identification of hierarchical chromatin domains. *Bioinformatics.* 2016;32(11):1601–9.
30. Haddad N, Vaillant C, Jost D. IC-Finder: inferring robustly the hierarchical organization of chromatin folding. *Nucleic Acids Res.* 2017;45(10):81.
31. Yu W, He B, Tan K. Identifying topologically associating domains and subdomains by Gaussian mixture model and proportion test. *Nat Commun.* 2017;8(1):535.
32. Norton HK, Emerson DJ, Huang H, Kim J, Titus KR, Gu S, Bassett DS, Phillipscremins JE. Detecting hierarchical genome folding with network modularity. *Nat Methods.* 2018;15(2):119–22.
33. Li A, Yin X, Xu B, Wang D, Han J, Wei Y, Deng Y, Xiong Y, Zhang Z. Decoding topologically associating domains with ultra-low resolution Hi-C data by graph structural entropy. *Nat Commun.* 2018;9(1):3265.
34. An L, Yang T, Yang J, Nuebler J, Xiang G, Hardison RC, Li Q, Zhang Y. OnTAD: hierarchical domain structure reveals the divergence of activity among TADs and boundaries. *Genome Biol.* 2019;20(1):1–16.
35. Ea V, Baudement M-O, Lesne A, Forné T. Contribution of topological domains and loop formation to 3D chromatin organization. *Genes.* 2015;6(3):734–50.
36. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* 2016;3(1):95–8.
37. Li A, Pan Y. Structural information and dynamical complexity of networks. *IEEE Trans Info Theory.* 2016;62(6):3290–339.
38. Zhang YW, Wang MB, Li SC. SuperTAD: robust detection of hierarchical topologically associated domains with optimized structural information. *DeepOmics.* 2019. <https://supertad.deepomics.org/>.
39. Zhang YW, Wang MB, Li SC. SuperTAD: robust detection of hierarchical topologically associated domains with optimized structural information. *Github.* 2019. <https://doi.org/10.5281/zenodo.4314123>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.