Cresswell et al. BMC Bioinformatics (2020) 21:319
https://doi.org/10.1186/s12859-020-03652-w

BMC Bioinformatics

SOFTWARE Open Access

SpectralTAD: an R package for defining a hierarchy of topologically associated domains using spectral clustering

Kellen G. Cresswell†, John C. Stansfield and Mikhail G. Dozmorov*†

* Correspondence: mikhail.dozmorov@vcuhealth.org
†Kellen G. Cresswell and Mikhail G. Dozmorov contributed equally to this work.
Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA

Abstract

Background: The three-dimensional (3D) structure of the genome plays a crucial role in gene expression regulation. Chromatin conformation capture technologies (Hi-C) have revealed that the genome is organized in a hierarchy of topologically associated domains (TADs), sub-TADs, and chromatin loops. Identifying such hierarchical structures is a critical step in understanding genome regulation. Existing tools for TAD calling are frequently sensitive to biases in Hi-C data, depend on tunable parameters, and are computationally inefficient.

Methods: To address these challenges, we developed a novel sliding window-based spectral clustering framework that uses gaps between consecutive eigenvectors for TAD boundary identification.

Results: Our method, implemented in an R package, SpectralTAD, detects hierarchical, biologically relevant TADs, has automatic parameter selection, is robust to sequencing depth, resolution, and sparsity of Hi-C data. SpectralTAD outperforms four state-of-the-art TAD callers in simulated and experimental settings. We demonstrate that TAD boundaries shared among multiple levels of the TAD hierarchy were more enriched in classical boundary marks and more conserved across cell lines and tissues. In contrast, boundaries of TADs that cannot be split into sub-TADs showed less enrichment and conservation, suggesting their more dynamic role in genome regulation.

Conclusion: SpectralTAD is available on Bioconductor, http://bioconductor.org/packages/SpectralTAD/.

Keywords: Hi-C, Chromosome conformation capture, Topologically associated domains, TADs, Hierarchy, SpectralTAD

Background

The introduction of chromatin conformation capture technology and its high-throughput derivative Hi-C enabled researchers to accurately model chromatin interactions across the genome and uncover the non-random 3D structures formed by folded genomic DNA [1–3]. The structure and interactions of the DNA in 3D space inside the nucleus has been shown to shape cell type-specific gene expression [3], replication

[4], guide X chromosome inactivation [5], and regulate the expression of tumor suppressors and oncogenes [6].

Topologically Associated Domains (TADs) refer to a common structure uncovered by Hi-C technology, characterized by groups of genomic loci that have high levels of interaction within the group and minimal levels of interaction outside of the group [1, 5, 7, 8]. TAD boundaries were found to be enriched in CTCF (considering the directionality of its binding) and other architectural proteins of cohesin and mediator complex (e.g., STAG2, SMC3, SMC1A, RAD21, MED12) [3, 7], marks of transcriptionally active chromatin (e.g., DNAse hypersensitive sites, H3K4me3, H3K27ac, H3K36me3 histone modifications) [5, 7–9]. From a regulatory perspective, TADs can be thought of as isolated structures that serve to confine genomic activity within their walls, and restrict activity across their walls. This confinement has been described as creating "autonomous gene-domains," essentially partitioning the genome into discrete functional regions [9, 10].

TADs organize themselves into hierarchical sets of domains [8, 11, 12]. These hierarchies are characterized by large "meta-TADs" that contain smaller sub-TADs and chromatin loops. To date, most methods were developed to find these single meta-TADs instead of focusing on the hierarchy of the TAD structures [13–15]. While interesting insights can be gleaned from the meta-TADs [16], work has shown that smaller sub-TADs are specifically associated with gene regulation [10, 17, 18]. For example, it has been found that genes associated with limb malformation in rats are specifically controlled through interactions within sub-TADs [18]. These results highlight the importance of identifying the full hierarchy of TADs.

Several methods have been designed to call hierarchical TADs (Supplementary Material, Additional file 1). However, most algorithms require tunable parameters [14, 19] that, if set incorrectly, can lead to a wide variety of results. Many tools have been shown to highly depend on sequencing depth and chromosome length (reviewed in [20]). Furthermore, the time complexity of many algorithms is often prohibitive for detecting TADs on a genome-wide scale. Also, many tools are not user-friendly and lack clear documentation [21], with some methods even lacking publicly available code [11]. Furthermore, the choice of TAD callers in the R/Bioconductor ecosystem remains limited (Supplementary Material, Additional file 1).

Our goal was to develop a simple data-driven method to detect TADs and uncover hierarchical sub-structures within these TADs. We propose a novel method that exploits the graph-like structure of the chromatin contact matrix and extend it to find the full hierarchy of sub-TADs, limited only by the resolution of Hi-C data. Our approach employs a modified version of the multiclass spectral clustering algorithm [22, 23] and uses a sliding window based on the commonly used 2 Mb biologically maximum TAD size [7, 24]. We introduce a novel method for automatically choosing the number of clusters (TADs) based on maximizing the average silhouette score [25]. We show that this approach finds TAD boundaries with more significant enrichment of known boundary marks than those called by other TAD callers. We then extend the method to find hierarchies of TADs and demonstrate their biological relevance. Our method provides a parameterless approach, efficiently operating on matrices in text format with consistent results regardless of the level of noise, sparsity, and resolution of Hi-C data. The method is fast and scales linearly with the increasing amount of data.

Our method is implemented in the SpectralTAD R package, freely available on GitHub (https://github.com/dozmorovlab/SpectralTAD) and Bioconductor (http://bioconductor.org/packages/SpectralTAD/).

Results

An overview of the SpectralTAD algorithm

SpectralTAD takes advantage of the natural graph-like structure of Hi-C data, allowing us to treat the Hi-C contact matrix as an adjacency matrix of a weighted graph. This interpretation allows us to use a spectral clustering-based approach, modified to use gaps between consecutive eigenvectors as a metric for defining TAD boundaries (Fig. 1a, Methods). We implement a sliding window approach that increases the stability of spectral clustering and reduces computation time. This approach detects the best number and quality of TADs in a data-driven manner by maximizing the number of internal contacts within TADs and minimizing those between TADs. To achieve this, we maximize a clustering metric called silhouette score that measures within TAD similarity and penalizes for the high similarity between TADs.

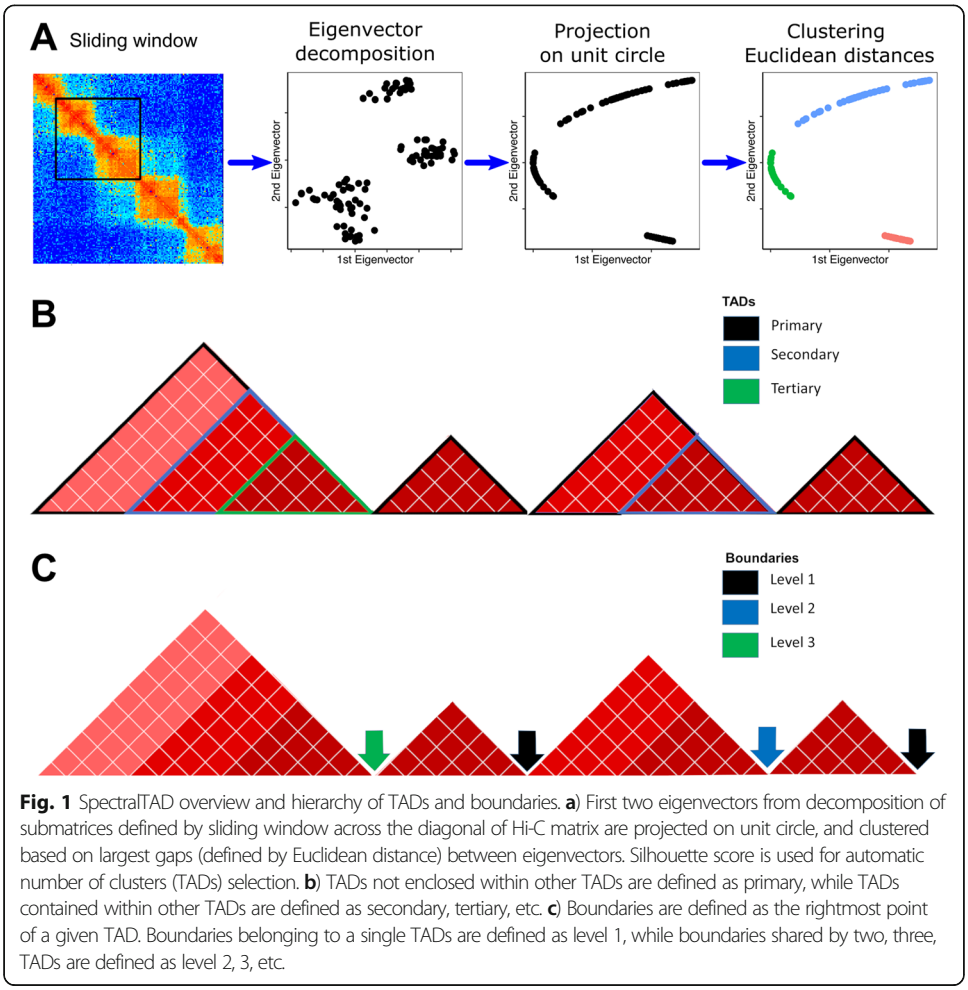

Fig. 1 SpectralTAD overview and hierarchy of TADs and boundaries. **a**) First two eigenvectors from decomposition of submatrices defined by sliding window across the diagonal of Hi-C matrix are projected on unit circle, and clustered based on largest gaps (defined by Euclidean distance) between eigenvectors. Silhouette score is used for automatic number of clusters (TADs) selection. **b**) TADs not enclosed within other TADs are defined as primary, while TADs contained within other TADs are defined as secondary, tertiary, etc. **c**) Boundaries are defined as the rightmost point of a given TAD. Boundaries belonging to a single TADs are defined as level 1, while boundaries shared by two, three, TADs are defined as level 2, 3, etc.

Defining hierarchical TADs and boundaries

We distinguish hierarchical types of TADs by their position with respect to other TADs. Primary TADs, or "meta-TADs," are defined as the top-level TADs that are not enclosed within other TADs (Fig. 1a). Conversely, we define "sub-TADs" as TADs detected within other TADs. We further refine the definition of sub-TADs to describe the level of hierarchy in which a sub-TAD is contained. Secondary TADs refer to sub-TADs which are contained within a primary TAD; tertiary TADs correspond to sub-TADs that are contained within two TADs and so on (Fig. 1b). Unless specified otherwise, we report results concerning primary TADs.

TAD boundaries represent another important element to be considered within the hierarchy. Using the terminology introduced in An et al. [26], we define a level 1 boundary as a TAD boundary belonging to a single TAD, irrespective of the TAD type. Level 2 and level 3 boundaries correspond to boundaries that are shared by two or three TADs, respectively (Fig. 1c).

An additional type of region is a gap, which refers to an area where there are no TADs present either due to a lack of sequencing depth, a centromere, or simply a lack of organization (Supplementary Methods, Additional file 1). The percentage of non-centromeric gaps varies across chromosomes and resolutions (Supplementary Table S1, Additional file 2), being 19.9% on average for the GM12878 data. In general, we observe that data at higher resolution (e.g., 10 kb) have the highest percentages of gaps due to sparsity. In our analysis, TADs are allowed to span the non-centromeric gaps.

Systematic approach for comparing TAD quality

A TAD boundary detection method ("TAD detection" hereafter) must be robust to sparsity and noise in Hi-C data, detect consistent TADs across sequencing depths and resolutions, and the TADs must be biologically and statistically meaningful. To compare the concordance of TAD boundaries identified by different TAD callers under different conditions, we used the Jaccard similarity metric. To compare TAD boundaries identified at different resolutions, we used a modified Jaccard similarity metric (Supplementary Material, Supplementary Figure S1, Additional files 1 and 3). Using simulated and experimental Hi-C data, we compared SpectralTAD with two single-level R-based TAD callers (TopDom [13] and HiCSeg [14]), and two hierarchical TAD callers (R-based rGMAP [27] and Python-based OnTAD [26]).

An important property of TAD detection methods is the ability to detect a hierarchy of TAD structures [3, 8, 11, 12, 17, 28]. Among R packages, rGMAP allows for the detection of two levels of the TAD hierarchy. Our method, SpectralTAD, and OnTAD can detect deeper levels of hierarchy, though we limit it to three in the current paper (Supplementary Figure S2, Additional file 4). Using simulated and experimental data, we compared the robustness of hierarchical TAD detection and defined properties of hierarchical TAD boundaries.

Multiple studies have demonstrated an enrichment of various genomic annotations at TAD boundaries [7, 9, 11, 13]. We quantified the biological relevance of TAD boundaries by using a permutation test to determine their enrichment in transcription factor binding sites, histone modification marks, and chromatin segmentation states (Supplementary Methods, Supplementary Table S2, Additional file 5).

ICE-normalized and raw Hi-C data are better suited for TAD detection

Sequence- and technology-driven biases may be present in Hi-C matrices [29–32]. Consequently, numerous normalization methods have been developed [2, 31–36]. However, their effect on the quality of TAD detection has not been explored.

We investigated the effect of three normalization methods, Knight-Ruiz (KR), iterative correction and eigenvector decomposition (ICE), and square root vanilla coverage (sqrtVC) on TAD detection using SpectralTAD. Using simulated matrices with the ground-truth TADs, we found that all normalization methods marginally degraded the performance of SpectralTAD under different levels of noise, sparsity, and downsampling (Fig. 2a-c). Based on these results, and the fact that previous studies showed graph-based TAD identification methods work well un-normalized Hi-C data [37, 38], consequent results are presented with the use of raw Hi-C data.

Using the experimental Hi-C data from the GM12878 cell line, we found that ICE normalization only marginally affected the average number and width of TADs, and these results were consistent across resolutions (Supplementary Figure S3A, Additional file 6). In contrast, KR, and sqrtVC normalization resulted in a larger variability in TAD widths across chromosomes and between resolutions (Supplementary Figure S3B, Additional file 6). We also assessed the average number and the enrichment (permutation test) of genomic annotations at TAD boundaries detected from unnormalized, KR-, ICE-, and sqrtVC-normalized data. The average number of genomic

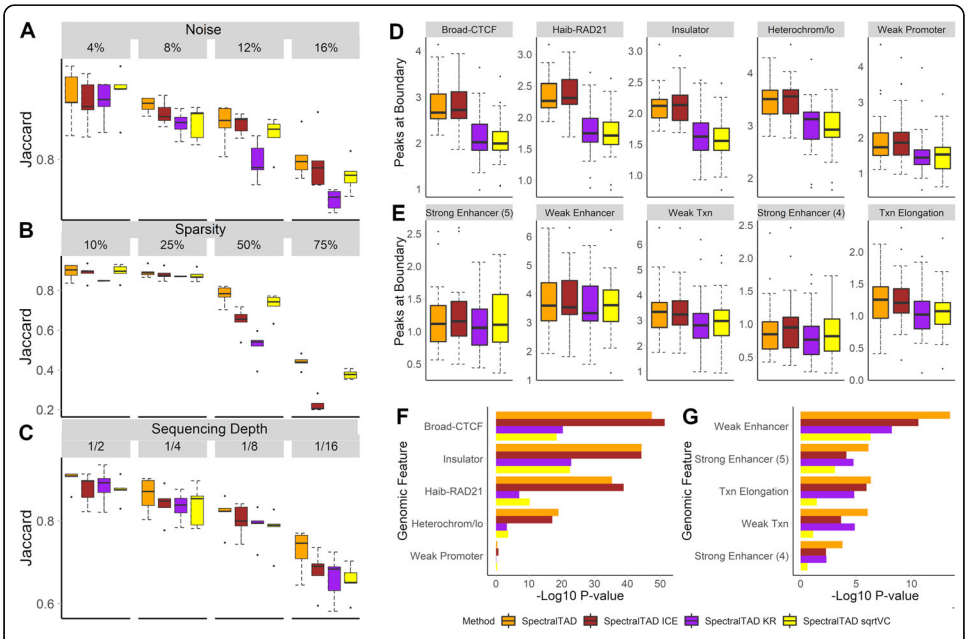

Fig. 2 The effect of normalization on TAD consistency and enrichment. To test for robustness to noise, sparsity, and sequencing depth, simulated Hi-C matrices were used as-is, KR-, ICE- and square root VC-normalized. TADs detected by SpectralTAD were compared with the ground-truth TADs using the Jaccard similarity metric. The effect of normalization was assessed at different levels of noise (**a**, the percentage of the original contact matrix modified by adding a constant), sparsity (**b**, the percentage of the original contact matrix replaced with zero), and downsampling (**c**, the fraction of contacts kept, see Methods). Using the raw and normalized data from the GM12878 cell line at 25 kb resolution, enrichment of genomic annotations within 50 kb regions flanking a TAD boundary on both sides was assessed using a permutation test. The average number of annotations for enriched (**d**) and depleted (**e**) genomic features and the permutation *p*-values corresponding to enrichment (**f**), and depletion (**g**) for the top five most enriched/depleted genomic annotations are shown. Results averaged across chromosome 1–22 are shown

annotations was not significantly different in TAD boundaries detected from raw and ICE-normalized data as compared with those from KR- and sqrtVC-normalized, where the number of annotations was significantly less (Fig. 2d). We found CTCF, RAD21, "Insulator," and "Heterochromatin" states to be significantly enriched in TAD boundaries, and this enrichment was frequently more significant in TAD boundaries detected from the ICE-normalized data (Fig. 2f). Similarly, "enhancer"-like chromatin states were significantly depleted at TAD boundaries, and this depletion was more pronounced in boundaries detected from raw data (Fig. 2g). The enrichment results were consistent across resolutions (Supplementary Figure S3C-F, Additional file 6). Furthermore, visual inspection of TAD boundaries detected from raw and ICE-normalized data demonstrated relatively good agreement, in contrast to those detected from KR-normalized data (Supplementary Figure S2, Additional file 4B). These results suggest that both ICE-normalized and raw Hi-C data are suitable for the robust detection of biologically relevant TADs.

SpectralTAD frequently identifies more consistent TADs than other methods

Using simulated matrices, we compared the performance of SpectralTAD with rGMAP, TopDom, OnTAD, and HiCSeg at different noise levels. We found that both SpectralTAD and TopDom had a significantly higher agreement with the ground truth TADs than rGMAP across the range of noise levels (Fig. 3a). To better understand the poor performance of rGMAP, we hypothesized that inconsistencies might arise due to the "off-by-one" errors that occur when, by chance, a TAD boundary may be detected adjacent to the true boundary location. We analyzed the same data using TAD boundaries flanked by 50 kb regions. Expectedly, the performance of all TAD callers, including rGMAP, increased; yet, the performance of rGMAP remained significantly low (Supplementary Figure S4A, Additional file 7). At low level of noise, HiCSeg detected highly consistent TAD boundaries; however, these TAD boundaries were the least biologically relevant, detailed below. In summary, these results suggest that, with the presence of high noise levels, a situation frequent in experimental Hi-C data, SpectralTAD performs better than other TAD callers in detecting true TAD boundaries.

We similarly investigated the effect of sparsity on the performance of the TAD callers. Expectedly, the average Jaccard similarity decreased for all TAD callers with an increased level of sparsity (Fig. 3b). SpectralTAD outperformed all TAD callers except HiCSeg at all sparsity levels. We further tested whether accounting for the "off-by-one" error improves the performance; the performance of SpectralTAD remained superior (Supplementary Figure S4B, Additional file 7). These results demonstrate the robustness of SpectralTAD to sparsity.

TAD callers should be robust to changes in sequencing depth. We introduced four levels of downsampling into simulated matrices and compared the detected TADs with the ground truth TADs. Downsampling involves removing contacts at random, simulating reduced sequencing depth. Expectedly, the average Jaccard similarity degraded for all TAD callers with an increased level of downsampling (Fig. 3c). Notably, the performance of SpectralTAD was consistently higher than that of other TAD callers except HiCSeg. Similar observations were true when accounting for the "off-by-one" error (Supplementary Figure S4C, Additional file 7). Despite the seemingly good performance

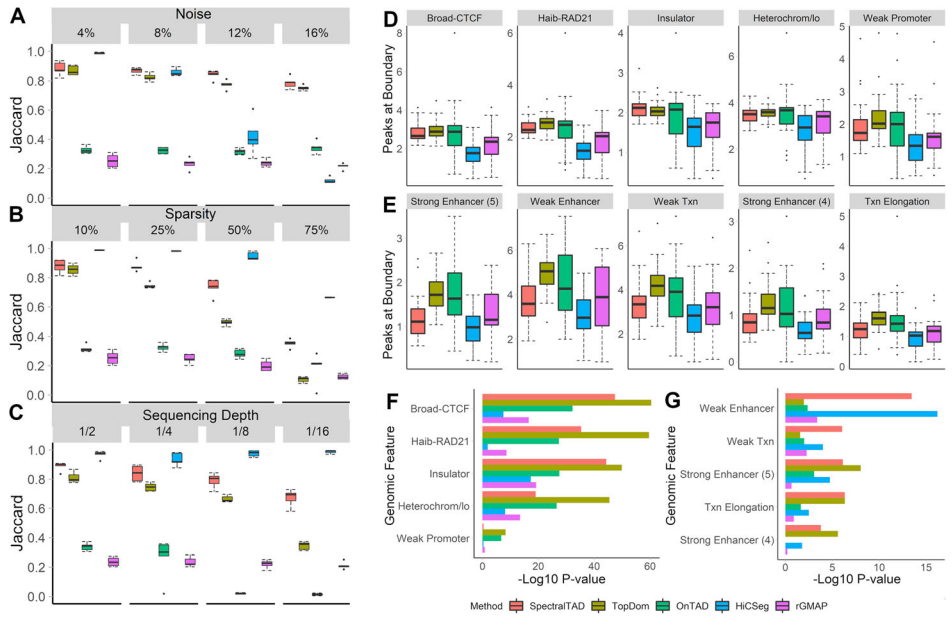

Fig. 3 The comparison of SpectralTAD and other TAD callers regarding TAD consistency and biological significance. To test for robustness to noise, sparsity, and sequencing depth, TADs were called from simulated Hi-C matrices using SpectralTAD and four other TAD callers. They were compared with the ground-truth TADs using the Jaccard similarity metric. The performance was assessed at different levels of noise (**a**), sparsity (**b**), and downsampling (**c**, see Methods). Using the raw data from GM12878 at 25 kb resolution, enrichment of genomic annotations within 50 kb regions flanking a TAD boundary on both sides was assessed using a permutation test. The average number of annotations for enriched (**d**) and depleted (**e**) genomic features and the permutation *p*-values corresponding to enrichment (**f**), and depletion (**g**) for the top five most enriched/depleted genomic annotations are shown. Results averaged across chromosome 1–22 are shown

of HiCSeg, further inspection showed it detects wide TADs that corresponds poorly to visually identifiable TADs while missing small-scale TADs detected by SpectralTAD (Supplementary Figure S2, Additional file 4A). Consequently, the biological relevance of TAD boundaries it detects is low, as discussed below (Fig. 3d-f). These observations, along with the results concerning sparsity and noise, suggest that with realistic levels of variation and noise in Hi-C data the performance of SpectralTAD in most cases is better than other TAD callers.

SpectralTAD outperforms other TAD callers in finding biologically relevant TAD boundaries

To evaluate the biological relevance of TAD boundaries detected by SpectralTAD and the other TAD callers, we evaluated their enrichment in genomic annotations known to be associated with TAD boundaries. We found that the TAD boundaries called by SpectralTAD, TopDom, and OnTAD had a significantly higher number of CTCF and RAD21, "Insulator," and "Heterochromatin" annotations than those called by HiCSeg and rGMAP (Fig. 3d). Consequently, these marks were more enriched at TAD boundaries detected by SpectralTAD and TopDom as compared with the other TAD callers (Fig. 3f). In terms of depleted genomic annotations, "enhancer"-like chromatin states were underrepresented at TAD boundaries, and this depletion was highly significant

for boundaries detected by SpectralTAD (Fig. 3e, g). Notably, the TAD boundaries detected by HiCSeg had the lowest number of these genomic annotations. They also exhibited the lowest level of enrichment and depletion (Fig. 3e, g). These results suggest that, despite robustness to noise, sparsity, and sequencing depth, HiCseg detects boundaries that are less biologically relevant in terms of known TAD biology. The performance of SpectralTAD and other callers was consistent at different resolutions (Supplementary Figure S4D-G, Supplementary Table S3, Additional files 7 and 8). In summary, these results suggest that SpectralTAD outperforms other TAD callers in detecting biologically relevant TAD boundaries.

SpectralTAD consistently identifies TADs across resolutions of Hi-C data

If TAD boundaries called at different resolutions of Hi-C data are inconsistent, one risks receiving vastly different results despite the data being the same. Using the GM12878 Hi-C data at 10 kb, 25 kb, and 50 kb resolutions, we estimated the average number and width of TADs called by SpectralTAD, TopDom, HiCSeg, OnTAD, and rGMAP. As the resolution of Hi-C data increased, the average number of TADs decreased for all but SpectralTAD (Supplementary Figure S5A, Additional file 9). Similarly, the average width of TADs increased for all but SpectralTAD TAD callers (Supplementary Figure S5B, Additional File 9). We further compared the consistency of TADs detected in 50 kb vs. 25 kb, 50 kb vs. 10 kb, 25 kb vs. 10 kb resolution comparisons. We found that, for nearly all comparisons, SpectralTAD and HiCSeg had significantly higher consistency quantified by modified Jaccard statistics than the other TAD callers (Supplementary Figure S5C, Additional File 9). When comparing the exact location of TADs detected by different TAD callers across four replicates of Hi-C data, SpectralTAD identified a higher proportion of TADs detected four times than OnTAD and rGMAP (Supplementary Figure S5D, Additional file 9). Consistent with previous observations (Fig. 3), TopDom and HiCSeg performed well; however, this analysis does not reflect the lower biological relevance of the detected TAD boundaries, as discussed above. These results show that SpectralTAD identifies consistent TADs at different resolutions of Hi-C data in most cases.

TAD hierarchy is associated with biological relevance

Having established the strong performance of SpectralTAD, we investigated the biological importance of the hierarchy of TAD boundaries detected by it. We tested the relationship between the number of times a TAD boundary occurs in a hierarchy (Fig. 1b) and the enrichment of genomic annotations. We hypothesized that TAD boundaries shared by two or more TADs (Level 2 and 3 boundaries) would be more biologically important, hence, harbor a larger number of key markers such as CTCF and RAD21. We found that this is indeed the case, as illustrated by a significant increase in the average number of CTCF and RAD21 annotations, and “Insulator”/“Heterochromatin” states around Level 2 and 3 boundaries as compared with Level 1 boundaries (Fig. 4a). A similar trend was observed in the stronger enrichment of Level 2 and 3 TAD boundaries in those annotations (Fig. 4c). TAD boundaries at all levels of the hierarchy were similarly depleted in the “enhancer”-like annotations (Fig. 4b, d), although these depletions were more significant for the Level 3 TAD boundaries. These observations were

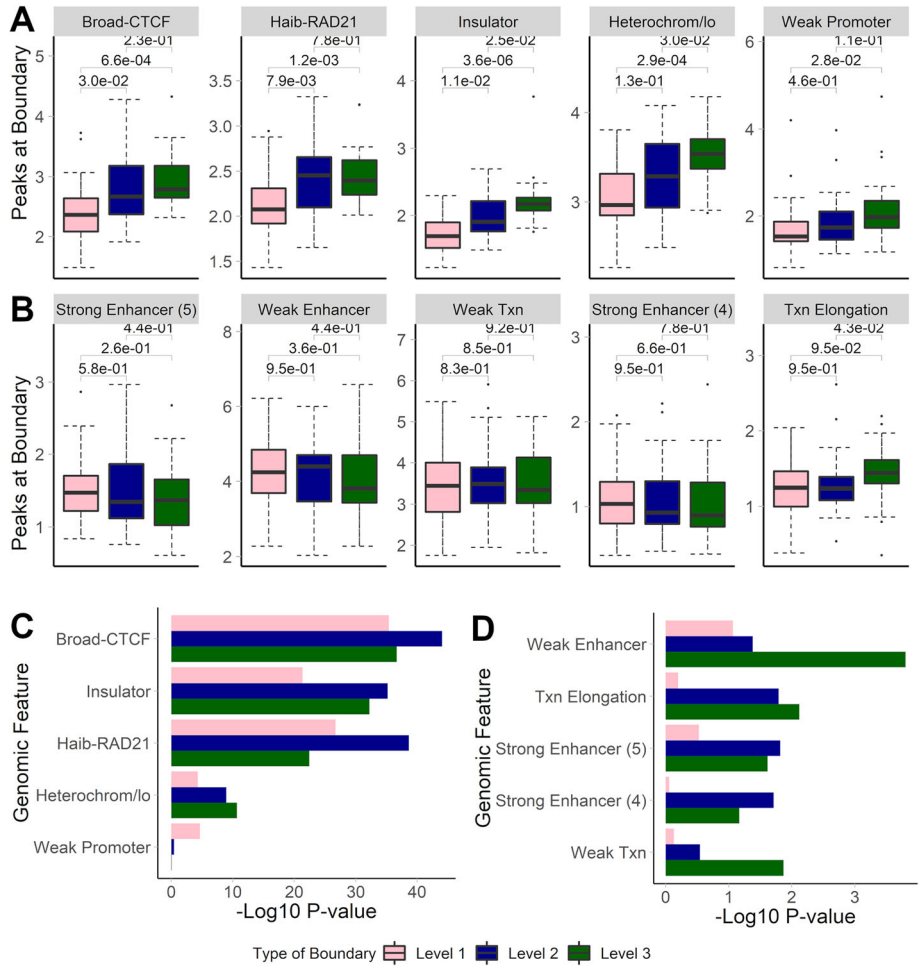

Fig. 4 The effect of the hierarchy of TAD boundaries detected by SpectralTAD on the average number of annotations in enriched (**a**) & depleted (**b**) genomic markers and on enrichment (**c**) and depletion (**d**) for different genomic annotations. Results for Level 1, 2, and 3 boundaries are shown. Genomic annotations were considered within 50 kb regions flanking a boundary on both sides. Wilcoxon test *p*-values are shown in panel **a & b**, and aggregated *p*-values, using the Fisher's method, are shown for panels **c & d**. Raw data from the GM12878 cell line, chromosome 1–22, 25 kb resolution

consistent across resolutions (Supplementary Figure S6, Additional file 10). Our results agree with previous research that has shown a positive correlation between the number of sub-TADs sharing a boundary and the number of biologically relevant genomic annotations at that boundary [26, 27] and confirm that SpectralTAD identifies a biologically relevant hierarchy of TADs.

TADs identified by SpectralTAD are conserved across cell lines and tissues

Previous studies reported relatively high conservation of TAD boundaries identified in different tissues and cell types, with the reported Jaccard statistics ranging from 0.21 to 0.30 [3]. We compared TAD boundaries called by SpectralTAD across various tissues and cell types (Supplementary Table S4, [39], Additional file 11). The Jaccard for all TADs, ignoring hierarchy, between cell-line samples ranged from 0.33 to 0.73 with a mean of 0.45 (SD = 0.08). The Jaccard between tissues ranged from 0.21 to 0.38, with a

mean of 0.27 (SD = 0.03), significantly lower than that of cell lines (Wilcoxon p-value < 0.0001). The lower conservation of TADs called from tissue samples is expected as cell lines come from a "pure" single source while tissues are a mixture of different cells. These results were summarized in heatmaps, comparing different cell lines (Supplementary Figure S7A, Additional file 12) and tissues (Supplementary Figure S7B, Additional file 12) according to Jaccard similarity. Hierarchical clustering of cell type-specific samples by the Jaccard similarity of their TADs, ignoring hierarchy, identified the expected associations between cell-type-specific data, with replicates clustering together and cell types being distinct (Supplementary Figure S7A, Additional file 12). To a lesser extent, these results were similar in tissue-specific samples (Supplementary Figure S7B, Additional file 12). In summary, these results show the conservation of TAD boundaries called using SpectralTAD across tissues and cell lines is similar to previously reported results [40].

Hierarchy of boundaries affect conservation of TADs

Following our definition of TADs (Primary, Secondary, and Tertiary, Fig. 1a), we hypothesized that primary TADs would be better conserved than Secondary or Tertiary TADs. The primary TADs are detected during the first pass of the algorithm; hence, they are robustly supported by the underlying data and expected to reproduce across different datasets. Indeed, the average Jaccard for Primary, Secondary, and Tertiary TADs across cell types was 0.42, 0.40, and 0.35, respectively (Supplementary Table S5, Additional file 13), and this decrease was significant (Wilcoxon p-value < 0.0001). These observations were consistent when analyzing TADs called from tissue samples, although the average Jaccard coefficients for Primary, Secondary, and Tertiary TADs were significantly lower (Supplementary Table S5, Additional file 13). These results demonstrate that Primary TADs are the most conserved across cell types and tissues.

We hypothesized that Level 3 TAD boundaries (Fig. 1b, boundaries that are shared by three TADs), besides showing higher biological significance (Fig. 4), will be better conserved. Indeed, the Jaccard coefficient of Level 3 TAD boundaries called in cell types was significantly higher (0.30) than that of Level 2 (0.23) and Level 1 (0.23) boundaries (Wilcoxon p-value ranging from 0.034 to < 0.0001, Supplementary Table S5, Additional file 13). These results were also observed in TAD boundaries called in tissue types. One possibility of the lower Jaccard coefficient for Level 1 and 2 boundaries is that they may change their assignment due to higher probability of detection of sub-TADs in different datasets. In summary, these results demonstrate that boundaries shared by several TADs have high biological significance and are better conserved across cell types and tissues.

SpectralTAD is the fastest TAD caller for high-resolution data

We evaluated the runtime performance of SpectralTAD, TopDom, OnTAD, rGMAP, and HiCSeg. SpectralTAD showed comparable performance with TopDom and was faster than rGMAP at all resolutions (Supplementary Figure S8A, Additional file 14). Specifically, SpectralTAD takes ~ 45 s to run with 25 kb data and ~ 4 min to run on 10 kb data for the entire GM12878 genome. By comparison, TopDom takes ~ 1 min to run on 25 kb data but ~ 13 min on 10 kb data. OnTAD takes ~ 4 min to run on 25 kb data

and ~ 30 min on 10 kb data. rGMAP takes ~ 12 min on 25 kb data and ~ 47 min on 10 kb data. We find that HiCSeg is prohibitively slow, taking ~ 609 min on 25 kb data and multiple days to run on 10 kb data with chromosome 1 taking over 24 h alone. Importantly, our method scales nearly linearly with the size of the data (see Methods), making it amenable for fast processing of data at higher resolutions. Furthermore, when parallelized, SpectralTAD is several orders of magnitude faster than other TAD callers (Supplementary Figure S8A, Additional file 14), e.g., with the entire genome taking 1 s to run for 25 kb data when using four cores. We demonstrate that our method has a linear complexity $O(n)$ (Supplementary Methods, Additional file 1), making it scalable for large Hi-C datasets. In summary, these results demonstrate that SpectralTAD is significantly faster than TopDom, rGMAP, and HiCSeg, providing near-instant results when running on multiple cores.

Discussion

We introduce the SpectralTAD R package implementing a spectral clustering-based approach that allows for fast TAD calling and scales well to high-dimensional data. The method was benchmarked against four TAD callers - TopDom and HiCSeg that detect single-level TADs, and OnTAD and rGMAP that detect hierarchical TADs. We show better performance of SpectralTAD vs. the other TAD callers in nearly all conditions. We also demonstrate that SpectralTAD is more robust to sparsity, sequencing depth, and resolution. We show that SpectralTAD can robustly detect hierarchical TAD boundaries. Furthermore, we demonstrate different levels of TAD hierarchy to be differentially associated with known marks of TAD boundaries, highlighting their distinct biological roles and the importance of the TAD hierarchy in general. The clear superiority of SpectralTAD regarding running speed and robustness to data irregularities suggests its use as the new gold-standard of hierarchical TAD callers in the R ecosystem.

The performance of SpectralTAD was frequently better, but not always superior to that of HiCSeg. The better performance of HiCSeg under different levels of noise, sparsity, and sequencing depth in some cases may be explained by the fact that HiCSeg identifies non-hierarchical TAD boundaries at once. Furthermore, HiCSeg uses complete Hi-C matrices to detect robust data-driven features [14]. However, due to the inherent noise and sparsity of Hi-C data at larger distances between interacting regions, those features may be detected due to consistent aberrations in the data. This may explain poor enrichment of HiCSeg-detected boundaries in CTCF, RAD21, and other functional signatures of TAD boundaries. SpectralTAD, on the other hand, defines a hierarchy of primary, secondary, etc., TADs, restricted to the first three levels in the current analysis. Furthermore, the sliding window approach implemented in SpectralTAD focuses on local, likely biology-driven chromatin interactions that define TAD boundaries. This may explain stronger enrichment of SpectralTAD-detected boundaries in genomic annotations previously associated with TAD boundaries. We suggest that the tradeoff between robustness and biological relevance of TAD boundaries should be made for the latter, with SpectralTAD providing the optimal balance.

One overarching limitation with non-hierarchical TAD callers like TopDom and HiCSeg is their inability to capture all TADs in a dataset. While methods like TopDom may find biologically relevant TADs, they cannot account for the common situation

where sub-TADs occur within a TAD. In the case of TADs enclosing sub-TADs, non-hierarchical callers are forced to make a choice that is often far from optimal (Supplementary Figure S2, Additional file 4). We suggest that even when the hierarchy of TADs is not essential, hierarchical TAD callers like SpectralTAD should be used for maximally accurate reconstruction of TADs at the first level of the hierarchy.

Our work focuses on R-based software for TAD boundary detection, motivating the selection of TAD callers implemented in R. However, many TAD callers, including those detecting hierarchical TADs, have been implemented in Python and/or as command-line tools (Additional file 1, reviewed in [41, 42]). One of our future goals includes a comparison of SpectralTAD with hierarchical TAD callers irrespectively of implementation, focusing on the quality of hierarchical TAD detection.

The goal of the SpectralTAD package is to provide the R-based implementation of our spectral clustering framework for TAD boundary detection. It outputs genomic coordinates of the detected TAD boundaries along with their level of hierarchy. These genomic coordinates provide flexibility for a range of downstream analyses and visualization. Examples include functional enrichment analysis of genes and genomic annotations in proximity of (different hierarchical) TAD boundaries using tools like rGREAT, GenomeRunner, LOLA [43–45]. Although SpectralTAD provides basic visualization functionality, external tools like HiCExplorer [46], Juicer [15], HiGlass [47], reviewed in [48, 49], provide greater flexibility in visualizing Hi-C contact maps and annotations. We continue exploring visualization options for SpectralTAD-detected boundaries.

Conclusion

In summary, we show that SpectralTAD is a robust method for defining the hierarchy of TAD boundaries. This method improves upon previous work showing the potential of spectral clustering for finding structures in Hi-C data while introducing modifications to make these methods practical for users. Specifically, we introduce two novel modifications to spectral clustering, the eigenvector gap and windowing, which can be used to quickly and accurately find changes in the pattern for ordered data. By releasing SpectralTAD as an open source R package, we aim to provide a user-friendly and accurate tool for hierarchical TAD detection.

Methods

Data sources

Experimental Hi-C matrices from the GM12878 cell line ([3] at 50 kb, 25 kb, and 10 kb, "primary+replicate", replicates (HIC001-HIC004)) and 35 different cell line and tissue samples ([39], 40 kb resolution) were downloaded from Gene Expression Omnibus (GEO, Supplementary Table S4, Additional file 11). 25 simulated matrices with manually annotated TADs ([42], 40 kb resolution) were downloaded from the HiCToolsCompare repository (Supplementary Table S4, Additional file 11). Data for chromatin states, histone modification and transcription factor binding sites (TFBS) were downloaded from the UCSC genome browser database [50]. Given the fact that some transcription factors have been profiled by different institutions (e.g., CTCF-Broad, CTCF-Uw, and CTCF-Uta), we selected annotations most frequently enriched at TAD boundaries

(typically, CTCF-Broad, RAD21-Haib). All genomic annotation data were downloaded in Browser Extensible Data (BED) format using the hg19/GRCh37 genome coordinate system (Supplementary Table S2, Additional file 5).

Windowed spectral clustering

Hi-C data representation

Chromosome-specific Hi-C data is typically represented by a chromatin interaction matrix C (referred hereafter as "contact matrix") binned into regions of size r (the resolution of the data). Entry C_{ij} of a contact matrix corresponds to the number of times region i interacts with region j. The matrix C is square and symmetric around the diagonal representing self-interacting regions. Our method relies on the fact that the 3D chromosome can be thought of as a naturally occurring graph [16, 51]. Traditionally, a graph $G(V, E)$ is represented by a series of nodes V connected by edges E. These graphs are summarized in an adjacency matrix A_{ij}, where entry ij indicates the number of edges between node i and node j. We can think of the contact matrix as a naturally occurring adjacency matrix (i.e., $C_{ij} = A_{ij}$) where each genomic locus is a node, and the edges are the number of contacts between these nodes [51]. This interpretation of the contact matrix allows us to proceed with spectral clustering.

Sliding window

To avoid performing spectral clustering on the entire matrix, which is highly computationally intensive, we apply the spectral clustering algorithm to submatrices defined by a sliding window across the diagonal of the entire matrix. The size of the window (the number of bins defining a submatrix) is based on the maximum possible TAD size of 2mb [7, 52]. In practice, the size of the window w is equal to $\frac{2mb}{r}$, where r is the resolution of the data. For example, at the 10 kb resolution, we would have a window size of $\frac{2mb}{10kb}$ or simply 200 bins. Following the guidelines of previous works on the minimum TAD size, we set a minimum window size of 5 bins [7, 53–55].

The restriction in window size means that the maximum resolution at which the algorithm can be run is 200 kb. At this resolution, the window can be partitioned into two separate TADs of 5 bin width. However, this is inappropriate as previous research indicated that TADs do not begin truly appearing until the resolution becomes less than 100 kb [7]. Therefore, our method is viable for all potential resolutions from which meaningful TADs can be called.

The algorithm starts at the beginning of the matrix and identifies the TADs in the first window. The window is then moved forward to the beginning of the last TAD detected, to account for the fact that the final TAD may overlap between windows. This is repeated until the end of the matrix. The result is a unique set of TADs.

Finding the graph spectrum

The first step of the algorithm is to find the graph spectrum. First, we calculate a Laplacian matrix - a matrix containing the spatial information of a graph. Multiple Laplacians exist [56]; but since our method builds upon the multiclass spectral clustering algorithm [22], which uses the symmetric Laplacian, we use the normalized symmetric Laplacian as follows:

1. Calculating the normalized symmetric Laplacian

$$L^{\boxtimes} = D^{-\frac{1}{2}} C D^{-\frac{1}{2}}$$

2. (where) $D = diag(1^T C)$
3. Solve the generalized eigenvalue problem

$$L^{\boxtimes} V^{\boxtimes} = \lambda V^{\boxtimes}$$

4. The result is a matrix of eigenvectors $V^{\boxtimes}{}_{w \times k}$, where w is the window size, and k is the number of eigenvectors used, and a vector of eigenvalues where each entry λ_i corresponds to the i_{th} eigenvalue of the normalized Laplacian $L^{\boxtimes}$.
5. Normalize rows and columns to sum to 1:

$$\widehat{V_{i.}} = \frac{V_{i.}^{\boxtimes}}{\boxtimes V^{\boxtimes}{}_{i.} \boxtimes}$$

6. where the subscript $i.$ Corresponds to column i

Projection onto the unit circle

Our method builds on the approach to spectral clustering first introduced in [22], which works by projecting the eigenvectors on a unit circle. Once we project these values on the circle, we can cluster regions of the genome by simply finding gaps in the circle (Supplementary Figure S9, Additional file 15). In the unit circle representation, a TAD boundary can be thought of as a region of discontinuity in the eigenvectors of adjacent values. Regions within the same TAD should have similar eigenvectors and have small distances between them. This approach takes advantage of the fact that eigenvectors are mapped to genomic coordinates which have a natural ordering. The steps for this portion of the algorithm are below:

1. Normalize the eigenvectors and project onto a unit circle

$$\tilde{Z} = diag\left(diag^{-\frac{1}{2}}\left(\widehat{V_{i.}}\widehat{V_{i.}}^{T}\right)\right)\widehat{V_{i.}}$$

2. For $i = 2, \ldots, n$ where n is the number of rows in $\tilde{Z}$ and k is the number of eigenvectors calculated (we suggest using two) to produce $\tilde{Z}$, calculate the Euclidean distance vector D

$$D_i = \sqrt{\left(\tilde{Z}_{i1} - \tilde{Z}_{(i-1)1}\right)^2 + \left(\tilde{Z}_{i2} - \tilde{Z}_{(i-1)2}\right)^2 \ldots + \left(\tilde{Z}_k - \tilde{Z}_{(i-1)k}\right)^2}$$

This step calculates the distance between the entries of the first two normalized eigenvectors that are associated with bin i and the bin to its left.

Choosing the number of TADs in each window

1. Find the location of the first $l = \frac{w}{5}$ largest values in D_i, where w is the window size, and $l + 1$ is the maximum number of TADs in a given window and partition the matrix into $l + 1$ sub-matrices with boundaries defined by the location of the l largest values.

2. For each sub-matrix calculate the silhouette statistic [25]:

$$s_i = \frac{b_i - a_i}{max[a_i, b_i]}$$

Here, a is the mean distance between each cluster entry and the nearest cluster, and b is the mean distance between points in the cluster. The distance between two given loci i and j is defined as $\frac{1}{C_{ij}+1}$, with "+1" added to avoid division by zero. C_{ij} corresponds to the number of contacts between loci i and loci j.

3. Find the mean silhouette score over all possible numbers of clusters m and organize into a vector of means

$$s\bar{}_m = \frac{\sum_{i=1}^{m} s_i}{m}$$

4. Find the value of m which maximizes $s\bar{}_m$

By taking the mean silhouette score, we can determine the number of eigenvectors, which allows us to maximize the similarity within clusters while minimizing the similarity between clusters. This translates into the number of clusters (i.e., TADs) that produces the most well-separated clusters. This procedure is performed within each window, allowing us to identify poorly organized regions (gaps, Supplementary Methods, Additional file 1). Cluster (TAD) boundaries are mapped to genomic coordinates based on their location in the contact matrix. If a TAD is detected and found to be less than 5 bins wide it is ignored due to previous evidence suggesting these are not biologically relevant [7, 53–55]. This step implies that, for a given window, the maximum number of TADs in a window is equal to the size of the window divided by 5.

Creating a hierarchy of TADs

We can find a hierarchy of TADs by iteratively partitioning the initial TADs. This is done by running a modified version of the main algorithm that includes an extra filtering step that tests for the presence of sub-TADs in each TAD. Briefly, each TAD is treated as an individual contact matrix, and a window is not used. To test for the existence of sub-TADs, we convert the distance vector D_i into a set of Z-scores by first taking the natural log of the distance vector before centering and scaling. This is done following our empirical observation of the log-normality of eigenvector gaps (Supplementary Figure S10, Additional file 16). We then label any distance with a Z-score greater than 2 as a sub-boundary. If significant sub-boundaries are detected, we partition the TAD with each sub-boundary indicating the end of a given sub-TAD. This procedure is then repeated for each sub-TAD until either the TAD is too small to be partitioned into two sub-TADs or no significant boundaries are found. The TADs detected during the initial run of the algorithm are considered primary TADs, and the TADs detected after partitioning are considered secondary, tertiary, etc., sub-TADs. In practice, this approach can also be used for the first iteration of the algorithm (non sub-TADs) and is an option in the SpectralTAD R package.

Benchmarking TAD callers

To evaluate the robustness of TAD callers, simulates Hi-C matrices were systematically modified to contain pre-defined levels of noise, sparsity, and sequencing depth. The overlap of TAD boundaries with the manually annotated TADs was tested using Jaccard statistics; the cross-resolution TAD comparison was assessed using a modified version of the Jaccard coefficient (Supplementary Figure S1, Additional file 3). The effect of Hi-C data normalization was tested using the iterative correction and eigenvector decomposition (ICE) [32], Knight-Ruiz (KR) [3, 33], and the Square Root Vanilla Coverage (sqrtVC) [3] methods. The association of TAD boundaries with genomic annotations, such as CTCF, was assessed using a permutation test. See Supplementary Methods (Additional file 1) for details.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-03652-w.

Additional file 1: Supplementary Material. An overview of previous methods, parameter selection, methods, and results not included in the main manuscript.

Additional file 2: Table S1. Summary of gaps. The percentage of gaps is summarized for all chromosomes at 10 kb, 25 kb, and 50 kb resolution using raw GM12878 data [3]. Gaps are separated based on whether they are centromeric or other (unsequenced, or poorly organized chromatin).

Additional file 3: Figure S1. Example of modified Jaccard statistics to measure agreement between TAD boundaries detected at different resolutions. The top triangles indicate TADs detected at 50 kb resolution, while the bottom triangles indicate those detected at 25 kb resolution. There are four shared boundaries (blue lines) and two non-shared boundaries (red lines). The traditional Jaccard statistic underestimates the fact that the four TAD boundaries agree at a different resolution, while the modified Jaccard statistics correctly identifies the perfect overlap between TAD boundaries by ignoring resolution differences.

Additional file 4: Figure S2. Examples of TADs detected under different conditions. A) TADs detected by SpectralTAD, TopDom, OnTAD, HiCSeg and rGMAP. B) TADs detected from raw, ICE-, and KR-normalized data. SqrtVC-normalized data could not be plotted due to format conversion issues. Red-yellow-blue color gradient indicate a high-medium low chromatin interaction strength; triangles indicate TADs. GM12878 data from [3], resolution 50 kb, chr1:60000000–80,000,000 (hg19). All parameters were set according to the instructions of each TAD caller. HiCExplorer v.3.0 [46] was used for visualization.

Additional file 5: Table S2. Experimental Data sources. Genome annotation (hg19/GRCh37) [50] data for GM12878 cell line used in the analysis, sorted by category, then by data type.

Additional file 6: Figure S3. The effect of data normalization on the average number (A) and width in kilobases (B) of TADs and the average number of peaks in enriched markers (C) and depleted markers (D), enrichment (F) and depletion (G) for different genomic annotations. Counts (A) and widths (B) for raw, KR-, ICE- and sqrtVC-normalized GM12878 data at 25 kb and 50 kb resolutions, averaged across chromosome 1–22, are shown for primary, secondary, and tertiary TADs detected by SpectralTAD. The average number of annotations for enriched (D) and depleted (E) genomic features and the permutation *p*-values corresponding to enrichment (F), and depletion (G) for the top most enriched/depleted genomic annotations (permutation test) at TAD boundaries for GM12878 data at 50 kb resolution are shown.

Additional file 7: Figure S4. The comparison of SpectralTAD and other TAD callers regarding TAD consistency and biological significance. To test for robustness to noise, sparsity, and downsampling, TADs were called from simulated Hi-C matrices using SpectralTAD and other TAD callers. The TAD boundaries were extended by 50 kb regions flanking a boundary on both sides. They were compared with the ground-truth TADs using the Jaccard similarity metric. The performance of the TAD callers was assessed at a different level of noise (A, the percentage of the original contact matrix modified by adding a constant of two), sparsity (B, the percentage of the original contact matrix replaced with zero), and downsampling (C, the fraction of contacts kept, see Methods). Using the raw data from GM12878 at 50 kb resolution, enrichment of genomic annotations within 50 kb regions flanking a TAD boundary on both sides was assessed using a permutation test. The average number of annotations for enriched (D) and depleted (E) genomic features and the permutation *p*-values corresponding to enrichment (F), and depletion (G) for the top five most enriched/depleted genomic annotations are shown. Results averaged across chromosome 1–22 are shown.

Additional file 8: Table S3 Enrichment by Method. Enrichment/Depletion results are provided for all genomic annotations tested. Permutation *p*-values summarized using Fisher's method are shown. Data is sorted alphabetically by category and then by genomic annotation.

Additional file 9: Figure S5. The number, width, and consistency of TADs called across resolutions and primary vs. replicate for different methods. The average number (A) and width (B) of TADs across resolutions, Jaccard similarity between TAD boundaries detected from primary and replicate data and modified Jaccard similarity between TAD boundaries detected from data at 10 kb, 25 kb and 50 kb resolutions (C) and the proportion of

shared boundaries across four replicates (D) are shown. HiCseg failed to run on some data due to sparsity, as indicated by gray bars on panel D. Wilcoxon test *p*-values are shown. Data from the GM12878 cell line, chromosomes 1–22.

Additional file 10: Figure S6. The effect of the hierarchy of TAD boundaries detected by SpectralTAD on the average number of annotations in enriched (A) and depleted (B) genomic markers and on enrichment (C) and depletion (D) for different genomic annotations. Results for TAD boundaries detected as Level 1, 2, and 3 boundaries are shown. Genomic annotations were considered within 50 kb regions flanking a boundary on both sides. Wilcoxon test *p*-values are shown in panel A & B, and aggregated *p*-values, using the Fisher's method, are shown for panels C & D. Raw data from GM12878 cell line, chromosome 1–22, 50 kb resolution.

Additional file 11: Table S4. Hi-C Data sources. Information about experimental [3, 39] and simulated [42] Hi-C data.

Additional file 12: Figure S7. Jaccard similarity of TAD boundaries across cell types (A) and tissues (B). TADs were called using SpectralTAD. Clustering was performed using Ward clustering applied to a Jaccard distance matrix. All TADs were called on raw 40 kb data from [39]. Various cell-lines and tissues are used.

Additional file 13: Table S5. Jaccard similarity across TAD hierarchy. Results for the corresponding comparison of Primary, Secondary, Tertiary TADs, and Level 1, 2, 3 TAD boundaries are shown. Jaccard similarity coefficients were compared using a Wilcoxon signed-rank test. Column *p*-values correspond to the comparison of Jaccard within levels between tissue samples and cell lines. Row *p*-values correspond to the comparisons within each type of data across the hierarchy.

Additional file 14: Figure S8. Runtime performance of various TAD callers. TADs were called using data from the GM12878 cell line at 10 kb and 25 kb resolution, and runtimes recorded. (A) Runtimes were summarized across different chromosomes. Each dot represents chromosome-specific run time averaged across three runs, with the regression line approximating the trend. X-axis – chromosome size in the number of bins, Y-axis – time in seconds. (B) The total time to analyze chromosomes 1–22 was calculated and summarized across methods and levels of parallelization for GM12878 25 kb resolution data. X-axis – Method, Y-axis – time in seconds. Results for HiCSeg are excluded due to exceptionally slow runtimes (24h hours for one 10 kb chromosome).

Additional file 15: Figure S9. Projection of eigenvectors on the unit circle. This projection allows us to identify TADs based on the distance between eigenvectors. The two largest gaps are used to separate TAD 1, TAD 2, and TAD 3. We can also see the difference between a strongly organized group with close together points (TAD 1 and TAD 2) and a weaker group with more spread out points (TAD 3). Simulated data from [42].

Additional file 16: Figure S10. Distribution of eigenvector gaps. The distributions of eigenvector gaps are plotted separately for each 10 kb, 25 kb, and 50 kb contact matrix from [3], 131 chromosome-specific datasets total. Results are colored by resolution. Higher-resolution data shows smaller overall gaps due to a larger number of regions of high sparsity. The untransformed eigenvector gaps (A) and the natural log eigenvector gaps (B) are shown. MASS::fitdistr() function was used to establish the best fit by a lognormal (67 datasets) or a Weibull (64 datasets) distributions with similar log-likelihoods. The lognormal fit was chosen to model the distribution of log eigenvector gaps.

Abbreviations

3D: Three-dimensional; Hi-C: Genome-wide chromatin conformation capture technology; ICE: Iterative correction and eigenvector decomposition normalization; KR: Knight-Ruiz normalization; sqrtVC: Square root vanilla coverage normalization; TADs: Topologically associated domains

Authors' contributions

MGD and KGC conceived the project, KGC implemented SpectralTAD. JCS helped with the data analysis. MGD and KGC wrote the manuscript with JCS contributions. All authors have read and approved the manuscript.

Funding

This work is supported in part by the PhRMA Foundation Research Informatics Award to MD.

Availability of data and materials

The datasets supporting the conclusions of this article are publicly available. For a comprehensive list of data sources and download links, see Supplementary Table S2 (Additional file 5) and Supplementary Table S4 (Additional file 11). The SpectralTAD R package can be downloaded from its Bioconductor page https://bioconductor.org/packages/SpectralTAD/. The source code is available under the MIT license at https://github.com/dozmorovlab/SpectralTAD.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 19 December 2019 Accepted: 10 July 2020
Published online: 20 July 2020

References

1. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002;295:1306–11.
2. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93.
3. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80.
4. Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, et al. Topologically associating domains are stable units of replication-timing regulation. Nature. 2014;515:402–5.
5. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the x-inactivation Centre. Nature. 2012;485:381–5.
6. Taberlay PC, Achinger-Kawecka J, Lun ATL, Buske FA, Sabir K, Gould CM, et al. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. Genome Res. 2016;26:719–31.
7. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485:376–80.
8. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the drosophila genome. Cell. 2012;148:458–72.
9. Ciabrelli F, Cavalli G. Chromatin-driven behavior of topologically associating domains. J Mol Biol. 2015;427:608–25.
10. Dixon JR, Gorkin DU, Ren B. Chromatin domains: the unit of chromosome organization. Mol Cell. 2016;62:668–80.
11. Fraser J, Ferrai C, Chiariello AM, Schueler M, Rito T, Laudanno G, et al. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol Syst Biol. 2015;11:852.
12. Gibcus JH, Dekker J. The hierarchy of the 3D genome. Mol Cell. 2013;49:773–82.
13. Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, et al. TopDom: An efficient and deterministic method for identifying topological domains in genomes. Nucleic Acids Res. 2016;44:e70.
14. Lévy-Leduc C, Delattre M, Mary-Huard T, Robin S. Two-dimensional segmentation for analyzing hi-c data. Bioinformatics. 2014;30:i386–92.
15. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al. Juicer provides a one-click system for analyzing loop-resolution hi-c experiments. Cell Systems. 2016;3:95–8.
16. Boulos RE, Tremblay N, Arneodo A, Borgnat P, Audit B. Multi-scale structural community organisation of the human genome. BMC Bioinformatics. 2017;18:209.
17. Phillips-Cremins JE, Sauria ME, Sanyal A, Gerasimova TI, Lajoie BR, Bell JS, et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell. 2013;153:1281–95.
18. Berlivet S, Paquette D, Dumouchel A, Langlais D, Dostie J, Kmita M. Clustering of tissue-specific sub-TADs accompanies the regulation of HoxA genes in developing limbs. PLoS Genet. 2013;9:e1004018.
19. Filippova D, Patro R, Duggal G, Kingsford C. Identification of alternative topological domains in chromatin. Algorithms for Molecular Biology. 2014;9:14.
20. Dali R, Blanchette M. A critical assessment of topologically associating domain prediction tools. Nucleic Acids Res. 2017; 45:2994–3005.
21. Chen J, Hero AO 3rd, Rajapakse I. Spectral identification of topological domains. Bioinformatics. 2016;32:2151–8.
22. Yu SX, Shi J. Multiclass spectral clustering. In: Proceedings of the ninth ieee international conference on computer vision - volume 2. Washington, DC: IEEE Computer Society; 2003. p. 313.
23. Chen H, Chen J, Muir LA, Ronquist S, Meixner W, Ljungman M, et al. Functional organization of the human 4D nucleome. Proc Natl Acad Sci U S A. 2015;112:8002–7.
24. Dekker J, Heard E. Structural and functional diversity of topologically associating domains. FEBS Lett. 2015;589(20 Pt A): 2877–84.
25. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
26. An L, Yang T, Yang J, Nuebler J, Xiang G, Hardison RC, et al. Hierarchical domain structure reveals the divergence of activity among tads and boundaries. bioRxiv:361147. https://doi.org/10.1101/361147.
27. Yu W, He B, Tan K. Identifying topologically associating domains and subdomains by gaussian mixture model and proportion test. Nat Commun. 2017;8.
28. Luzhin AV, Flyamer IM, Khrameeva EE, Ulianov SV, Razin SV, Gavrilov AA. Quantitative differences in tad border strength underly the tad hierarchy in drosophila chromosomes. J Cell Biochem. 2018;120(3):4494–503.
29. Yaffe E, Tanay A. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet. 2011;43:1059–65.
30. O'Sullivan JM, Hendy MD, Pichugina T, Wake GC, Langowski J. The statistical-mechanics of chromosome conformation capture. Nucleus. 4:390–8.
31. Cournac A, Marie-Nelly H, Marbouty M, Koszul R, Mozziconacci J. Normalization of a chromosomal contact map. BMC Genomics. 2012;13:436.
32. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of hi-c data reveals hallmarks of chromosome organization. Nat Methods. 2012;9:999–1003.
33. Knight PA, Ruiz D. A fast algorithm for matrix balancing. IMA J Numer Anal. 2012;33:1029–47.
34. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. HiCNorm: removing biases in hi-c data via poisson regression. Bioinformatics. 2012;28:3131–3.
35. Li W, Gong K, Li Q, Alber F, Zhou XJ. Hi-corrector: a fast, scalable and memory-efficient package for normalizing large-scale hi-c data. Bioinformatics. 2015;31:960–2.
36. Ay F, Bailey TL, Noble WS. Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts. Genome Res. 2014;24:999–1011.

37. Fotuhi Siahpirani A, Ay F, Roy S. A multi-task graph-clustering approach for chromosome conformation capture data sets identifies conserved modules of chromosomal interactions. Genome Biol. 2016;17:114.
38. Li T, Jia L, Cao Y, Chen Q, Li C. OCEAN-c: mapping hubs of open chromatin interactions across the genome reveals gene regulatory networks. Genome Biol. 2018;19:54.
39. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep. 2016;17:2042–59.
40. Sauerwald N, Kingsford C. Quantifying the similarity of topological domains across normal and cancer human cell types. Bioinformatics. 2018;34:i475–83.
41. Zufferey M, Tavernari D, Oricchio E, Ciriello G. Comparison of computational methods for the identification of topologically associating domains. Genome Biol. 2018;19.
42. Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. Comparison of computational methods for hi-c data analysis. Nat Methods. 2017;14:679–85.
43. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010;28:495–501.
44. Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in r and bioconductor. Bioinformatics. 2016;32:587–9.
45. Dozmorov MG, Cara LR, Giles CB, Wren JD. GenomeRunner web server: regulatory similarity and differences define the functional impact of snp sets. Bioinformatics. 2016;32:2256–63.
46. Ramirez F, Bhardwaj V, Arrigoni L, Lam KC, Gruning BA, Villaveces J, et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun. 2018;9.
47. Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, et al. HiGlass: Web-based visual exploration and analysis of genome interaction maps. bioRxiv. http://biorxiv.org/content/early/2017/10/30/121889.abstract.
48. Yardımcı GG, Noble WS. Software tools for visualizing hi-c data. Genome Biol. 2017;18:26.
49. Ing-Simmons E, Vaquerizas JM. Visualising three-dimensional genome organisation in two dimensions. Development. 2019;146.
50. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, et al. ENCODE data in the UCSC genome browser: year 5 update. Nucleic Acids Res. 2012;41:D56–63.
51. Boulos RE, Arneodo A, Jensen P, Audit B. Revealing long-range interconnected hubs in human chromatin interaction data using graph theory. Phys Rev Lett. 2013;111:118102.
52. Sofueva S, Yaffe E, Chan W-C, Georgopoulou D, Vietri Rudan M, Mira-Bontenbal H, et al. Cohesin-mediated interactions organize chromosomal domain architecture. EMBO J. 2013;32:3119–29.
53. Won H, de la Torre-Ubieta L, Stein JL, Parikshak NN, Huang J, Opland CK, et al. Chromosome conformation elucidates regulatory relationships in developing human brain. Nature. 2016;538:523–7.
54. Jiang Y, Loh Y-HE, Rajarajan P, Hirayama T, Liao W, Kassim BS, et al. The methyltransferase SETDB1 regulates a large neuron-specific topological chromatin domain. Nat Genet. 2017;49:1239–50.
55. Norton HK, Emerson DJ, Huang H, Kim J, Titus KR, Gu S, et al. Detecting hierarchical genome folding with network modularity. Nat Methods. 2018;15:119–22.
56. von Luxburg U. A tutorial on spectral clustering. Stat Comput. 2007;17(4):2007 http://arxiv.org/abs/0711.0189v1.

Publisher's Note