SpectralTAD: an R package for defining a hierarchy of Topologically Associated Domains using spectral clustering

Kellen G. Cresswell[1,2] (cresswellkg@vcu.edu), John C. Stansfield[1] (stansfieldjc@vcu.edu), Mikhail G. Dozmorov[1,2*] (mikhail.dozmorov@vcuhealth.org)

[1] Dept. of Biostatistics, Virginia Commonwealth University, Richmond, VA, 23298, USA
[2] Equal contribution
[*] To whom correspondence should be addressed: Virginia Commonwealth University, Richmond, VA, 23298, 804-827-2055, mikhail.dozmorov@vcuhealth.org

Keywords: Hi-C, chromosome conformation capture, topologically associated domains, TADs, hierarchy, SpectralTAD

Running Title: Hierarchical TAD boundary detection using spectral clustering

Abstract

The three-dimensional (3D) structure of the genome plays a crucial role in regulating gene expression. Chromatin conformation capture technologies (Hi-C) have revealed that the genome is organized in a hierarchy of topologically associated domains (TADs), the fundamental building blocks of the genome. Identifying such hierarchical structures is a critical step in understanding regulatory interactions within the genome. Existing tools for TAD calling frequently require tunable parameters, are sensitive to biases such as sequencing depth, resolution, and sparsity of Hi-C data, and are computationally inefficient. Furthermore, the choice of TAD callers within the R/Bioconductor ecosystem is limited. To address these challenges, we frame the problem of TAD detection in a spectral clustering framework. Our `SpectralTAD` R package has automatic parameter selection, robust to sequencing depth, resolution and sparsity of Hi-C data, and detects hierarchical, biologically relevant TAD structure. Using simulated and real-life Hi-C data, we show that `SpectralTAD` outperforms `rGMAP` and `TopDom`, two state-of-the-art R-based TAD callers. TAD boundaries that are shared among multiple levels of the hierarchy were more enriched in relevant genomic annotations, e.g., CTCF binding sites, suggesting their higher biological importance. In contrast, boundaries of primary TADs, defined as TADs which cannot be split into sub-TADs, were found to be less enriched in genomic annotations, suggesting their more dynamic role in genome regulation. In summary, we present a simple, fast, and user-friendly R package for robust detection of TAD hierarchies supported by biological evidence. SpectralTAD is available on https://github.com/dozmorovlab/SpectralTAD and Bioconductor (submitted).

Introduction

The introduction of chromatin conformation capture technology and its high-throughput derivative Hi-C has given researchers the ability to accurately model chromatin interactions across the genome

and uncover the non-random 3D structures formed by folded genomic DNA (Dekker et al. 2002; Schmitt et al. 2016; Berkum et al. 2010; Lieberman-Aiden et al. 2009; Rao et al. 2014). The structure and interactions of the DNA in 3D space inside the nucleus has been shown to shape the cell type-specific gene expression (Dowen et al. 2014; Ji et al. 2016; Phillips-Cremins and Corces 2013; Rao et al. 2014; Vietri Rudan et al. 2015; Jin et al. 2013), orchestrate the assembly of antigen receptors (Jhunjhunwala et al. 2009), guide X chromosome inactivation (Nora et al. 2012; Crane et al. 2015), and regulate the expression of tumor suppressors and oncogenes (Valton and Dekker 2016; Rickman et al. 2012; Taberlay et al. 2016; Hnisz et al. 2016; Lupiáñez et al. 2016).

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 (Dixon et al. 2012; Jackson and Pombo 1998; Ma et al. 1998; Nora et al. 2012; Sexton et al. 2012; Bonev et al. 2017; Dekker et al. 2002; Sofueva et al. 2013). 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) (Splinter et al. 2006; Dixon et al. 2012; Rao et al. 2014, 2014; Beagan et al. 2017; Sanyal et al. 2012; Tang et al. 2015; Phillips and Corces 2009; Handoko et al. 2011; Vietri Rudan et al. 2015; Corces and Corces 2016; Zuin et al. 2014; Sofueva et al. 2013), marks of transcriptionally active chromatin (e.g., DNAse hypersensitive sites, H3K4me3, H3K27ac, H3K36me3 histone modifications) (Hong and Kim 2017; Dixon et al. 2012; Filippova et al. 2014; Lévy-Leduc et al. 2014; Nora et al. 2012; Sexton et al. 2012; Taberlay et al. 2016; Bonev et al. 2017; Ciabrelli and Cavalli 2015; Bonev and Cavalli 2016), actively transcribed and housekeeping genes (Oti et al. 2016; Bonev and Cavalli 2016; Dixon et al. 2016, 2012; Luzhin et al. 2018). 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 (Dixon et al. 2016, 2012; Sexton et al. 2012; Bonev et al. 2017; Ciabrelli and Cavalli 2015; Sofueva et al. 2013).

TADs were shown to organize themselves into hierarchical sets of domains (Hansen et al. 2017; Weinreb and Raphael 2015; Fraser et al. 2015; Rao et al. 2014; Phillips-Cremins et al. 2013; Gibcus and Dekker 2013; Sexton et al. 2012; Luzhin et al. 2018; Dong et al. 2018; Sofueva et al. 2013). These hierarchies are characterized by large "meta-TADs" that contain smaller sub-TADs. To date, most methods have focused on finding these single meta-TADs instead of focusing on the hierarchy of the structures (Shin et al. 2016; Lévy-Leduc et al. 2014; Wang et al. 2015; Durand et al. 2016; Serra et al. 2017). While interesting insights can be gleaned from the meta-TADs, work has shown that smaller sub-TADs are specifically associated with gene regulation (Phillips-Cremins et al. 2013; Berlivet et al. 2013; Dixon et al. 2016). For example, it has been found that genes associated with limb malformation in rats are specifically controlled through the interaction within sub-TADs (Berlivet et al. 2013). These results highlight the importance of identifying the full hierarchy of TADs.

While some important functions of TADs have been identified, their role in the genome remains to be fully understood. Besides the fact that TADs are a relatively recent discovery, we are also plagued by a lack of ground truth data and a clear determination of what exactly constitutes a TAD. The most basic definition of a TAD is a large kilobase- to megabase-sized group of loci which are contained within sharp boundaries and interact highly with each other when compared with loci outside of their group (Dixon et al. 2012; Ciabrelli and Cavalli 2015; Dali and Blanchette 2017; Dekker et al. 2013; Luzhin et al. 2018; Sofueva et al. 2013). Our goal was to develop a simple data-driven method

to detect such 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 hierarchicy of sub-TADs. Our method is implemented in the SpectralTAD R package, freely available on GitHub (https://github.com/dozmorovlab/SpectralTAD) and Biobonductor (submitted).

Previous Methods

Many of the first and most popular TAD detection methods were based on the directionality index which is a function of the average upstream and downstream interactions. This metric was then used as a parameter in a hidden Markov model, to establish the location of TAD boundaries (Dixon et al. 2012). The basis of this method was the fact that boundary regions are expected to interact with downstream regions more than upstream regions. This method was followed by a number of methods designed to calculate non-hierarchical TADs such as `Armatus` (Filippova et al. 2014), `HiCseg` (Lévy-Leduc et al. 2014), `TADLib` (Wang et al. 2015), `TopDom` (Shin et al. 2016), `Arrowhead` (Durand et al. 2016), `TADbit` (Serra et al. 2017), `CaTCH` (Zhan et al. 2017) and `RHiCDB` (Chen et al. 2018). Another intuitive metric, the insulation index (Crane et al. 2015), uses a sliding window approach to sum up contacts within a given region surrounding each locus. As TADs are regions of increased contacts, they can easily be identified via a contact count cutoff. Some tools, such as the `TADtool` Python package (Kruse et al. 2016), implement both metrics to call TADs. `HiCDB` uses an extension of the conventional insulation index that corrects for background noise.

The detection of hierarchical TAD structures was first introduced by Fraser J. et al. (Fraser et al. 2015) who used single-linkage clustering to create a hierarchical structure of meta-TADs which contain smaller sub-TADs. Since this discovery, there has been a lack of publicly available, user-friendly, hierarchical TAD callers. A hierarchical TAD caller refers to a tool that finds TADs and sub-TADs contained within them. To date, the choice of hierarchical TAD callers remains limited.

The first publicly available tool for hierarchical TAD calling, `TADtree` (Weinreb and Raphael 2016) worked by creating TAD "forests" containing hierarchical "trees" of TADs. `TADtool` similarly provides hierarchical TAD detection and visualization (Kruse et al. 2016). Other hierarchical TAD finders include `HiTAD` (Wang et al. 2017a) and `IC-Finder` (Haddad et al. 2017) which take dynamic programming, hidden Markov model-dynamic programming hybrid, and probabilistic approach, respectively. `ClusterTAD` (Oluwadare and Cheng 2017) introduced a traditional hierarchical clustering-based approach to TAD classification. Another method, `rGMAP` (Yu et al. 2017) has arisen as a potentially useful tool for TAD detection. This method utilizes a Gaussian Mixture model and a z-test of proportions to identify TADs. The model is then run iteratively to partition TADs into sub-TADs, but in practice is limited to two levels of TADs. More recently, a method called `onTAD` was proposed (An et al. 2018). This method uses `TopDom`, a single-level TAD caller that uses a statistical test on upstream and downstream contacts, to find all possible TAD boundaries and then to select a final configuration using a dynamic programming algorithm (Shin et al. 2016). Of these methods `TADtool`, `TADtree` and `TADLib` are Python based. `IC-Finder` and `ClusterTAD` are MATLAB based with `ClusterTAD` also including Java implementation and rGMAP is available as an R package. Although some comparison of TAD detection tools has been performed (Ay and Noble 2015; Forcato et al. 2017; Nicoletti et al. 2018), this diversity leaves the choice of the most appropriate method uncertain.

Hi-C data, represented as an adjacency matrix, naturally lends itself to the use of graph theory (Boulos et al. 2013; Wang et al. 2017b, 2013). `Arboretum-HiC` first introduced the idea of using Laplacian-graph segmentation to find structures in Hi-C data (Fotuhi Siahpirani et al. 2016). This method used a spectral clustering approach to simultaneously find 3D structures between multiple matrices. Chen et al. (Chen et al. 2016) proposed a method that framed the contact matrix as a weighted adjacency matrix and used recursive partitioning of the Fiedler vector to identify TADs. `MrTADFinder` (Yan et al. 2017) and `HiTAD` (Wang et al. 2017a) take a similar approach but address the question as a community detection problem. Most recently, `3DNetMod` was introduced which treats the Hi-C matrix as a network and uses network modularity to cluster the TADs (Norton et al. 2018). This method is also designed to find hierarchies of TADs. Network theoretical, or more broadly graph theoretical, approaches have a promise to provide us with a data-driven method of identifying TADs that takes the entire structure of loci-loci interactions into account.

One common issue with many of the previous methods is robustness. We are often interested in detecting TADs for chromosomes of different lengths and over a range of different resolutions and sequencing depth. To account for this, most algorithms require tunable parameters (Filippova et al. 2014; Lévy-Leduc et al. 2014; Wang et al. 2017b) that, if set incorrectly, can lead to a wide range of results. Other tools, such as `Arrowhead` and `Domain Caller`, have been shown to highly depend on sequencing depth (reviewed in (Dali and Blanchette 2017)). Furthermore, the time complexity of many algorithms is often prohibitive for detecting TADs on a genome-wide scale. The last, but important, issue is that many tools are not user-friendly and lack clear documentation (Chen et al. 2016; Wang et al. 2017b), with some methods even lacking publicly available code (Fraser et al. 2015)

R/Bioconductor is the de facto gold standard programming language for the genomics community (Gentleman et al. 2006). Currently, the number of TAD callers implemented in the R programming language are limited and include `HiCseg` (Lévy-Leduc et al. 2014), `TopDom` (Shin et al. 2016), `rGMAP`(Yu et al. 2017), and `HiCDB` (Chen et al. 2018). `HiCseg` is the only TAD-calling specific R package available on CRAN or Bioconductor, while `TopDom` is a downloadable R script. `HiCDB` and `rGMAP` are available on GitHub. Another tool, `RobustTad`, is in development and currently only provides a metric for TAD calling. It will potentially provide a new option for R users (Dali et al. 2018). Additionally, the `HiTC` R package (Servant et al. 2012) has functionality for calculating the directionality index but does not provide any tools for TAD identification. Neither `TopDom` nor `HiTC` can operate on the commonly used $n \times n$ contact matrices in text format. `TopDom` requires the data to be formatted as an $n \times n + 3$ matrix with the first three columns corresponding to the genomic coordinates. `HiTC` requires the user to transform the data into their package-specific `HTCexp` object. `HiCseg` can be used to analyze $n \times n$ text matrices but forces users to assume a distribution of contacts and estimate the number of TADs before running. These factors aren't always clearly apparent from the data and thus require constant tweaking to account for different levels of noise, sparsity, and resolution of Hi-C data. Although `HiCDB` can process $n \times n$ contact matrices, it requires matrices from multiple chromosomes thus limiting the analysis of single-chromosome data. Additionally, it requires data to have a resolution of 5kb, 10kb or 40kb. The aforementioned limitations limit the choice of R-based tools for direct comparison.

We developed an R package, `SpectralTAD`, for TAD identification using a spectral clustering-based approach. The method works by reframing the contact matrix as the adjacency matrix of a weighted graph where chromatin regions are vertices and interaction frequencies are weights of edges connecting them. Consequently, clustering is being performed on the eigenvalues of the normalized Laplacian of this graph. The method employs a modified version of the multiclass spectral clustering

algorithm (Yu and Shi 2003). We take advantage of the commonly used 2Mb biologically maximum TAD size (Dixon et al. 2012; Dekker and Heard 2015) to introduce a sliding-window approach that helps to increase the stability and speed of the algorithm. We introduce a novel method for automatically choosing the number of clusters (TADs) based on maximizing the average silhouette score (Rousseeuw 1987). We show that this approach finds TAD boundaries having more significant enrichment of genomic annotations known to mark the boundaries. We then extend the method to find hierarchies of TADs and demonstrate their biological relevance. Our method provides a parameterless approach efficiently operating on either sparse 3-column, $n \times n$ and $n \times n + 3$ matrices in text format with consistent results regardless of the level of noise, sparsity, and resolution of Hi-C data.

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 of the contact matrix allows us to use a spectral clustering-based approach, modified to use gaps between consecutive eigenvectors as a metric for gefining TAD boundaries. 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 mean silhouette score.

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 (Figure 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 (Figure 1A). 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. (An et al. 2018), 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 (Figure 1B).

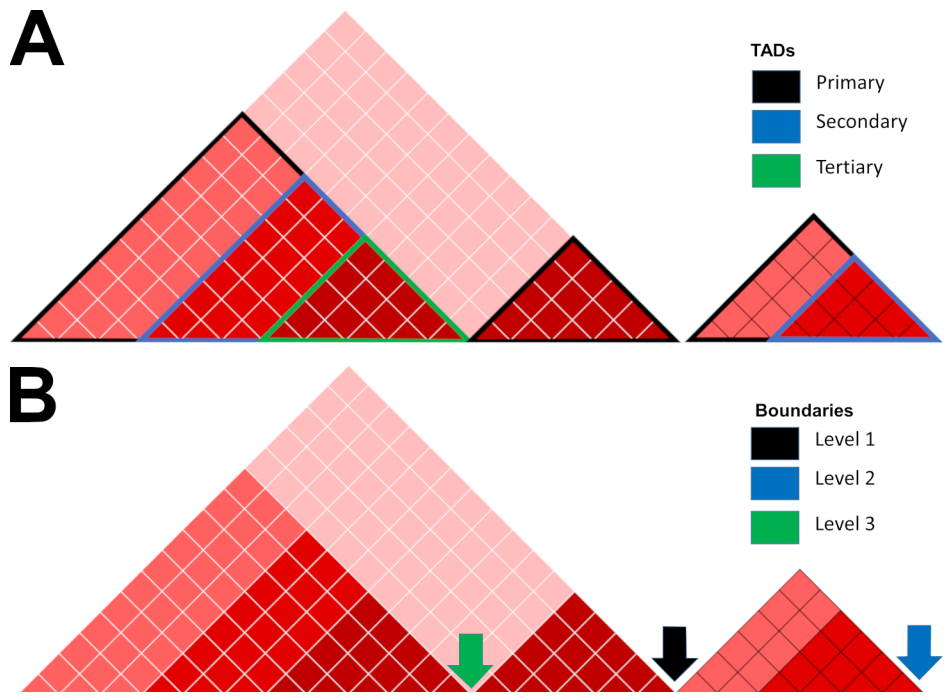


Figure 1. Hierarchy of TADs and boundaries. A) TADs not enclosed within other TADs are defined as primary, while TADs contained within other TADs are defined as secondary, tertiary, etc. B) Boundaries belonging to a single TADs are defined as level 1, while boundaries shared by two, three, TADs are defined as level 2, 3, etc. Boundaries are defined as the rightmost point of a given TAD.

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. The percentage of non-centromeric gaps varies across chromosomes and resolutions (Supplementary Table S1). Averaged across all chromosomes, the percentage of non-centromeric gaps is 21.8%. In general, we observe that data at higher resolution (e.g., 10kb) have the highest percentages of gaps due to sparsity. In our analysis, TADs are allowed to span the non-centromeric gaps.

Comparison of 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. For comparing TAD boundaries identified at different resolutions, we used a modified Jaccard similarity metric (Supplementary Figure S1, see Methods). Using simulated and real Hi-C data, we compared our method, `SpectralTAD`, with two other R-based methods for TAD detection, `rGMAP` and `TopDom`.

An important property of TAD detection methods is the ability to detect hierarchy of TAD structures (Hansen et al. 2017; Weinreb and Raphael 2015; Fraser et al. 2015; Rao et al. 2014; Phillips-Cremins et al. 2013; Gibcus and Dekker 2013; Sexton et al. 2012; Luzhin et al. 2018). Among R packages, only `rGMAP` allows for the detection of two levels of TAD hierarchy. Our method, `SpectralTAD`, can

detect deeper levels of hierarchy, limited to three in the current paper (Supplementary Figure S2). Using simulated and real-life 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 (Dixon et al. 2012; Shin et al. 2015; Fraser et al. 2015; Ciabrelli and Cavalli 2015). 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 Table S2).

Untransformed Hi-C data are better suited for TAD detection

We investigated the effect of log10- and Toeplitz transformation on the performance of `SpectralTAD`. Using the simulated matrices at a different level of noise, sparsity, and downsampling (see Methods), we tested the similarity of TAD boundaries called from untransformed and transformed data with the ground truth TADs using the Jaccard similarity statistics. We found that both log10- and Toeplitz transformation has a negative impact on the robustness of `SpectralTAD` (Figure 2A-C). The decrease in performance of TAD calling using log10- or Toeplitz-transformed data is in line with previous work showing spectral clustering work best on untransformed data (Fotuhi Siahpirani et al. 2016).

Using the real-life Hi-C data from the Gm12878 cell line, we found that the average number and width of TADs detected from the untransformed, log10-, and Toeplitz-transformed data were comparable, and these results were consistent across resolutions (Supplementary Figure S3A-B). We also assessed the average number and the enrichment of genomic annotations at TAD boundaries detected from untransformed, log10-, and Toeplitz-transformed data (Figure 2D-F). Expectedly, CTCF, RAD21, "Insulator," and "Heterochromatin" states were significantly enriched in TAD boundaries, and this enrichment was frequently more significant in TAD boundaries detected from the untransformed data (Figure 2E). Similarly, "enhancer"-like chromatin states were significantly depleted at TAD boundaries, and this depletion was more pronounced in boundaries detected from untransformed data (Figure 2F). The average number and enrichment of genomic annotations at boundaries detected from log10- and Toeplitz-transformed data was significantly less than from raw data (Figure 2D-E). Similarly, genomic annotations depleted at TAD boundaries were less significant and larger in number at TAD boundaries detected from transformed data (Figure 2D-F). The enrichment results were consistent across resolutions (Supplementary Figure S3C-E). In summary, these observations argue in favor of using untransformed data for the detection of biologically relevant TAD boundaries using spectral clustering.

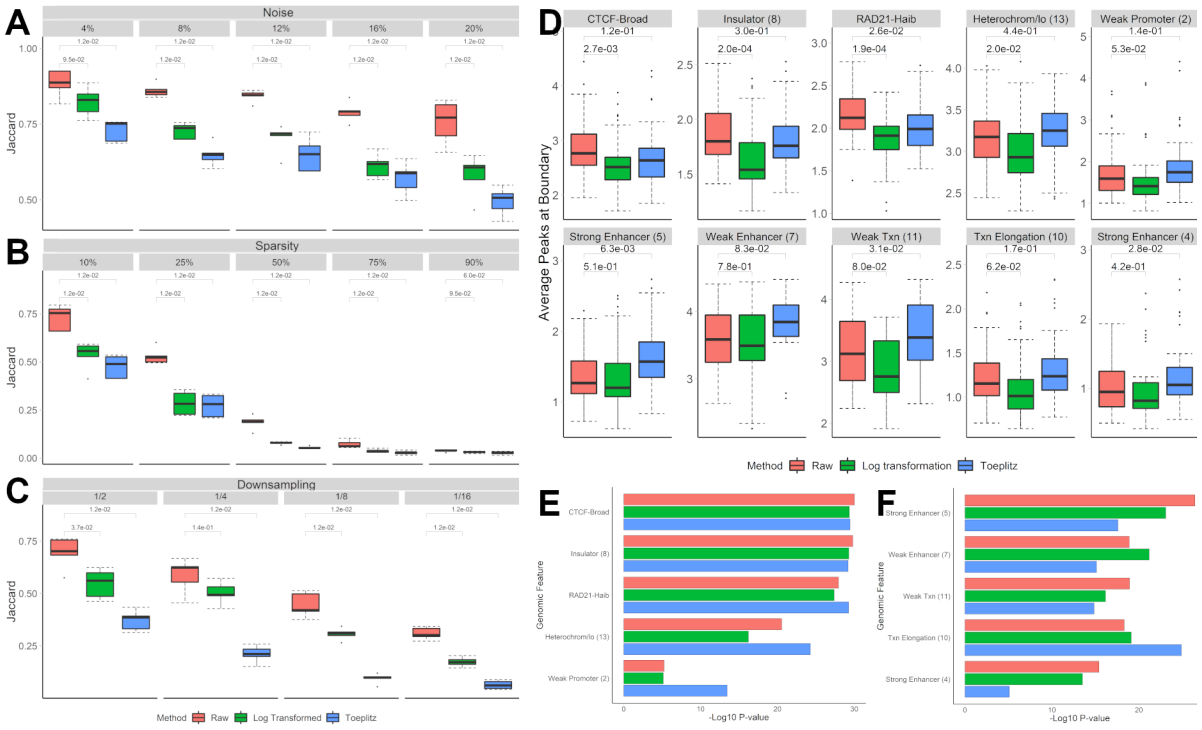


Figure 2. The effect of data transformation on TAD consistency and enrichment. To test for robustness to noise, sparsity, and downsampling, simulated Hi-C matrices were used as-is, log10-, or Toeplitz-transformed. TADs detected by SpectralTAD were compared with the ground-truth TADs using the Jaccard similarity metric. The effect of transformation was assessed at different levels 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 and transformed data from Gm12878 cell line at 25kb resolution, enrichment of genomic annotations within 50kb regions flanking a TAD boundary on both sides was assessed using a permutation test. The average number of annotations (D), enrichment (E), and depletion (F) for the top five most enriched/depleted genomic annotations are shown. Results averaged across chromosome 1-22 are shown. Wilcoxon test p-values are shown in panels A, B, C and D, and aggregated p-values, using the addCLT method, are shown for panels E and F.

Unnormalized Hi-C data are better suited for TAD detection

Sequence- and technology-driven biases may be present in Hi-C matrices (Yaffe and Tanay 2011; O'Sullivan et al.; Cournac et al. 2012; Imakaev et al. 2012). Numerous normalization methods have been developed to alleviate the presence of biases in Hi-C data (Lieberman-Aiden et al. 2009; Imakaev et al. 2012; Knight and Ruiz 2012; Cournac et al. 2012; Hu et al. 2012; Li et al. 2015; Ay et al. 2014). 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 sequential component normalization (SCN) on TAD detection using `SpectralTAD`. Using simulated matrices with the ground-truth TADs, we found that KR and SCN normalizations significantly degraded the performance of `SpectralTAD` under different levels

of noise, sparsity and downsampling (Figure 3A-C). Interestingly, ICE normalization marginally affected TAD detection (Figure 3A-C). These results suggest that, although ICE normalization may be suitable for TAD calling, unnormalized data is better for the detection of TAD boundaries.

Using the real-life Hi-C data from Gm12878 cell line, we found that all but SCN normalization methods marginally affected the average number and width of TADs, and these results were consistent across resolutions (Supplementary Figure S4A-B). We also assessed the average number and the enrichment of genomic annotations at TAD boundaries detected from unnormalized, KR-, ICE-, and SCN-normalized data (Figure 3D-F). As with the results obtained when testing the effect of transformation, 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 unnormalized data (Figure 3E). Similarly, "enhancer"-like chromatin states were significantly depleted at TAD boundaries, and this depletion was more pronounced in boundaries detected from unnormalized data (Figure 3F). The average number of genomic annotations was not significantly different in TAD boundaries detected from normalized data as compared with those from unnormalized data, except for SCN normalization, where the number of annotations was significantly less (Figure 3D). The enrichment results were consistent across resolutions (Supplementary Figure S4C-E). These results further strengthen our notion that unnormalized Hi-C data is perfectly suitable for the detection of biologically relevant TADs. Consequently, the following results are presented with the use of unnormalized Hi-C data.

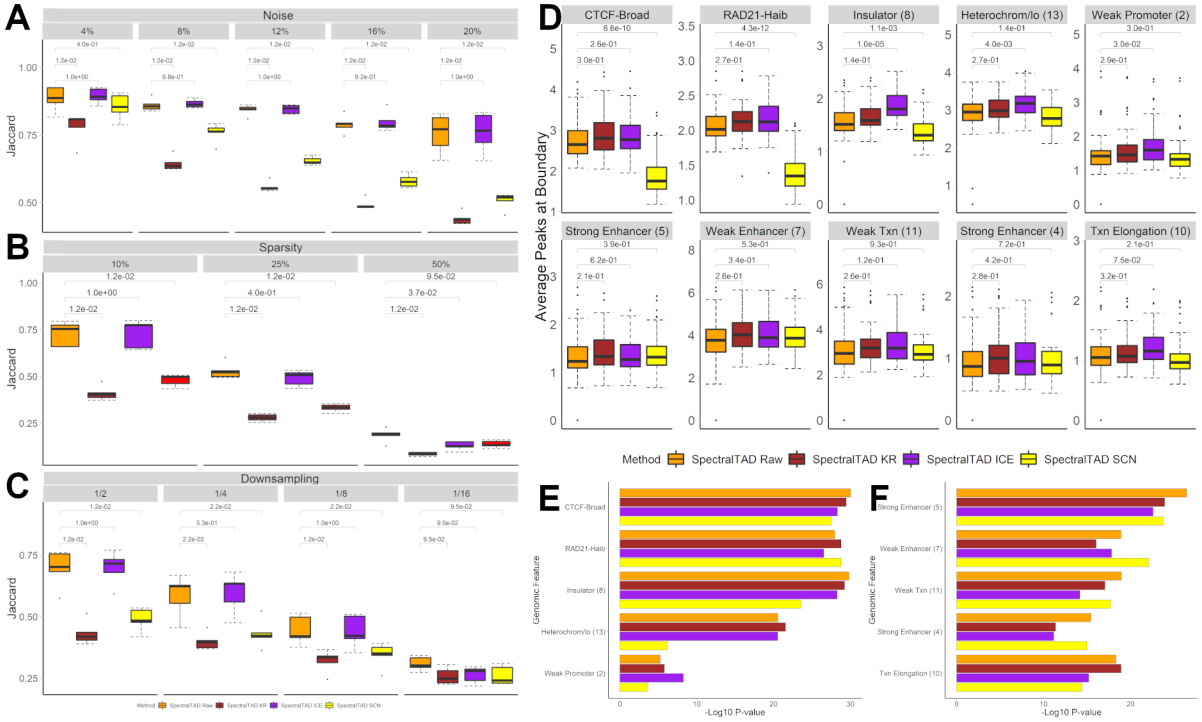


Figure 3. The effect of normalization on TAD consistency and enrichment. To test for robustness to noise, sparsity, and downsampling, simulated Hi-C matrices were used as-is, KR-, ICE- and SCN-normalized. TADs detected by SpectralTAD were compared with the ground-truth TADs using the Jaccard similarity 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 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 and normalized data from Gm12878 cell line at 25kb resolution, enrichment of genomic features within 50kb regions flanking a TAD boundary on both sides were assessed using a permutation test. The average number of annotations (D), enrichment (E), and depletion (F) for the top five most enriched/depleted genomic annotations are shown. Results averaged across chromosome 1-22 are shown. Wilcoxon test p-values are shown in panels A, B, C and D, and aggregated p-values, using the addCLT method, are shown for panels E and F.

`SpectralTAD` identifies more consistent TADs than `rGMAP` and `TopDom`

Using simulated matrices, we compared the performance of `SpectralTAD` with `rGMAP` and `TopDom` TAD callers 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 level (Figure 4A). To better understand the poor performance of `rGMAP`, we hypothesized that inconsistencies might arise due to the "off-by-one" errors, 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 50kb regions. Expectedly, the performance of all TAD callers, including `rGMAP`, increased; yet, the performance of `rGMAP` remained significantly low (Supplementary Figure S5A). In summary, these results suggest that, with the presence of high noise levels, a situation frequent in real-life Hi-C data, `SpectralTAD` and `TopDom` perform similarly and better than `rGMAP` in detecting true TAD boundaries.

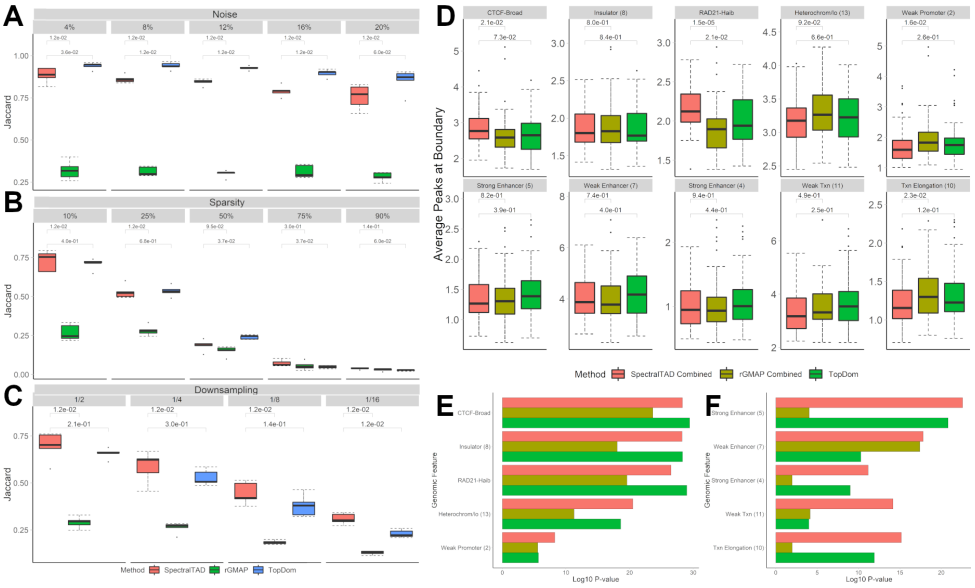


Figure 4. The comparison of SpectralTAD, rGMAP, and TopDom on TAD consistency and enrichment. To test for robustness to noise, sparsity, and downsampling, TADs were called from simulated Hi-C matrices using SpectralTAD, rGMAP, and TopDom. They were compared with the ground-truth TADs using the Jaccard similarity metric. The performance of the TAD callers was assessed at different levels 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 25kb resolution, enrichment of genomic annotations within 50kb regions flanking

a TAD boundary on both sides was assessed using a permutation test. The average number of annotations (D), enrichment (E), and depletion (F) for the top five most enriched/depleted genomic annotations are shown. Results averaged across chromosome 1-22 are shown. Wilcoxon test p-values are shown in panels A, B, C, and D and aggregated p-values, using the addCLT method, are shown for panels E and F.

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 the increased level of sparsity (Figure 4B). At high sparsity levels (>75%), `SpectralTAD` significantly outperformed `TopDom`. We further tested whether accounting for the "off-by-one" error improve the performance; however, the `rGMAP` results remained poor (Supplementary Figure S5B). These results demonstrate the robustness of `SpectralTAD` to sparsity.

TAD callers should also 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. Expectedly, the average Jaccard similarity degraded for all TAD callers with the increased level of downsampling (Figure 4C). Notably, the performance of `SpectralTAD` was consistently significantly higher than that of `rGMAP` and `TopDom`. Similar observations were true when accounting for the "off-by-one" error (Supplementary Figure S5C). 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` is similar or better than the other TAD callers.

SpectralTAD outperforms other TAD callers in finding biologically relevant TAD boundaries

The biological relevance of TAD boundaries detected by `SpectralTAD`, `rGMAP`, and `TopDom` was evaluated for the enrichment of various genomic annotations (Supplementary Table S3). We found that the boundaries of TADs called by `SpectralTAD` had a significantly higher number of CTCF and RAD21 annotations than those called by `rGMAP` (Figure 4D). Consequently, CTCF, RAD21, "Insulator," and "Heterochromatin" states were more enriched at TAD boundaries detected by `SpectralTAD` as compared with `rGMAP` (Figure 4E). `SpectralTAD` also identified more significant depletion of "enhancer"-like chromatin states at TAD boundaries than `rGMAP` (Figure 4F). The average number, enrichment, and depletion of genomic annotations in TAD boundaries called by `TopDom` were similar or less significant than that of for `SpectralTAD` (Figure 4D-F). These observations were consistent at different resolutions (Supplementary Figure S5D-F). In summary, these results suggest that `SpectralTAD` significantly outperforms `rGMAP` and performs similarly or better than `TopDom` in detecting biologically relevant TAD boundaries.

SpectralTAD identifies consistent TADs across resolutions and replicates of Hi-C data

It is imperative that TAD boundaries called at different resolutions of Hi-C data are consistent, or one risks finding vastly different TAD boundaries despite the data being the same. Using the Gm12878 Hi-C data at 10kb, 25kb, and 50kb resolutions, we estimated the average number and width of TADs called by `SpectralTAD`, `rGMAP`, and `TopDom`. Somewhat expectedly, the average number of TADs decreased, while width increased, with a lower resolution of Hi-C data for TADs detected by `rGMAP` and `TopDom`. However, the average number and width of TADs detected by `SpectralTAD` were similar across resolutions (Supplementary Figure S6A-B). We further compared the consistency

of TADs detected in 50kb vs. 25kb, 50kb vs. 10kb, 25kb vs. 10kb resolution comparisons. We found that, for all comparisons, `SpectralTAD` and `TopDom` had significantly higher consistency quantified by modified Jaccard statistics than `rGMAP` (Supplementary Figure S6E). These results show that, in contrast to `rGMAP` and `TopDom`, `SpectralTAD` identifies consistent TADs at different resolutions of Hi-C data.

An important issue in TAD identification is consistent performance across replicates of Hi-C data. We compared TADs called from replicates of Hi-C data from Gm12878 cell line at 10kb, 25kb, and 50kb resolutions. `SpectralTAD` and `TopDom` outperformed `rGMAP` in detecting consistent TAD boundaries from 25kb and 50kb resolution data (Supplementary Figure S6C), and these results were similar when accounting for the "off-by-one" error (Supplementary Figure S6D). Expectedly, as the data resolution becomes lower, hence, becoming less sparse, consistency of all TAD callers increased, with `SpectralTAD` and `TopDom` performing the best. Previous work using data at 100kb resolution has shown that around 60-70% of boundaries are shared between primary and replicates (Dixon et al. 2012; Rao et al. 2014; Sauerwald et al. 2018), and our results are in agreement with with these observations. In summary, these results demonstrate robust performance of `SpectralTAD` in detecting consistent TAD boundaries across resolutions and replicates of Hi-C data.

The hierarchical structure of TADs is associated with biological relevance

To demonstrate the biological importance of the hierarchy of TAD boundaries detected by `SpectralTAD`, a functionality lacking in `TopDom`, we tested whether there was a relationship between the number of times a TAD boundary occurs in a hierarchy (Figure 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, RAD21. We found that this is indeed the case, as illustrated by a significant increase in the average number and enrichment of CTCF and RAD21 annotations around Level 2 and 3 boundaries as compared with Level 1 boundaries (Figure 5A-B). A similar trend was observed in the enrichment of "Insulator," and "Heterochromatin" states, and in the depletion of "enhancer"-like chromatin states (Figure 5A-C), and these observations were consistent across resolutions (Supplementary Figure S7). 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 (Yu et al. 2017; An et al. 2018), and confirm that `SpectralTAD` identifies a biologically relevant hierarchy of TAD boundaries.

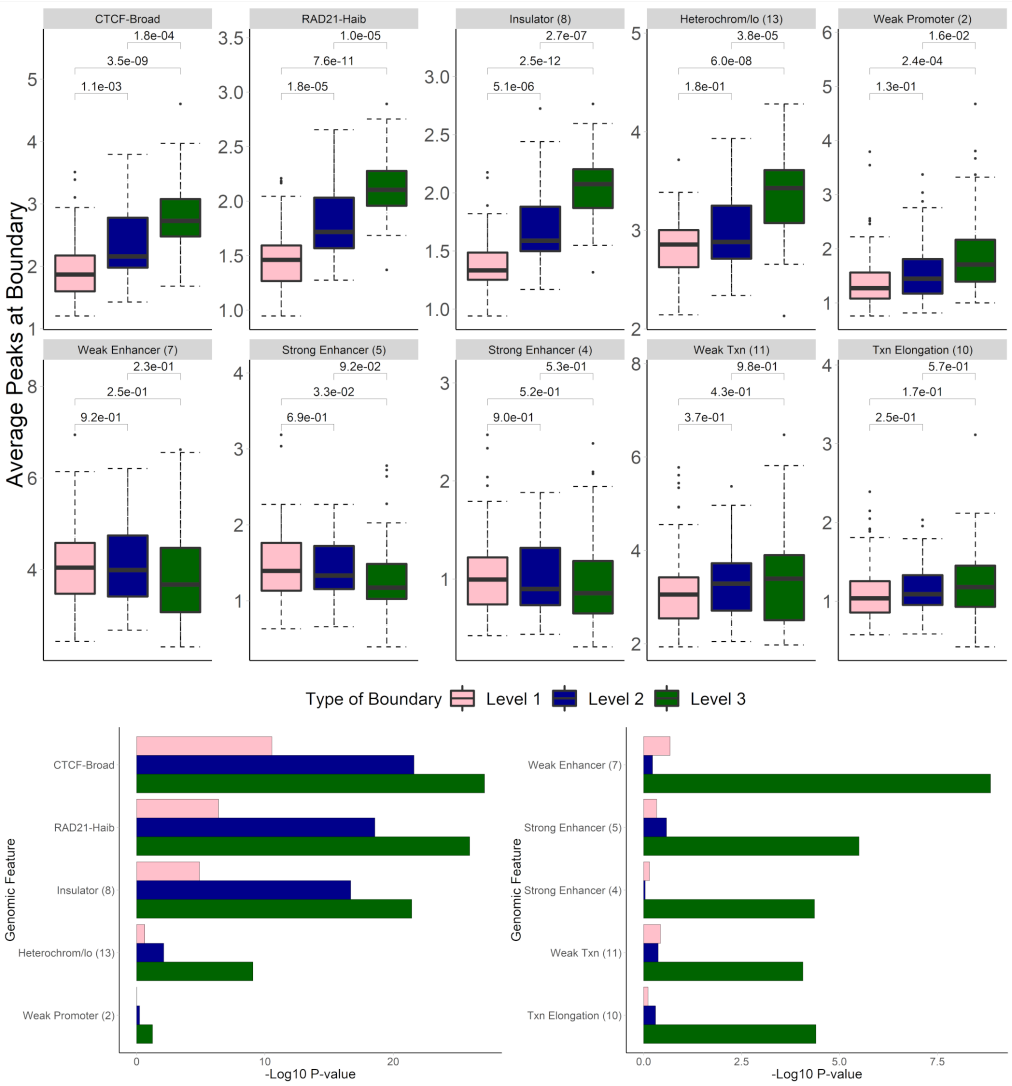


Figure 5. The effect of the hierarchy of TAD boundaries detected by SpectralTAD on the average number of annotations (A), enrichment (B) and depletion (C) for different genomic annotations. Results for TAD boundaries detected as level 1, 2, and 3 boundaries are shown. Genomic annotations were considered within 50kb regions flanking a boundary on both sides. Wilcoxon test p-values are shown in panel A and aggregated p-values, using the addCLT method, are shown for panels B and C. Data from GM12878 cell line, chromosome 1-22, 25kb resolution.

SpectralTAD is the fastest TAD caller for high-resolution data

We evaluated the runtime performance of `SpectralTAD`, `rGMAP`, and `TopDom`. `SpectralTAD` showed comparable performance with `TopDom` and was faster than `rGMAP` at all resolutions (Supplementary Figure S8A). Specifically, `SpectralTAD` takes ~45 seconds to run with 25kb data and ~4 minutes to run on 10kb data for the entire GM12878 genome. By comparison, `TopDom` takes ~1 minute to run on 25kb data but ~13 minutes on 10kb data. `rGMAP` takes ~12 minutes on 25kb data and ~47 minutes on 10kb data. Importantly, our method scales 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), e.g., with the entire genome taking 1 second to run for 25kb data when using four cores. In summary, these results demonstrate that SpectralTAD is significantly faster than rGMAP and TopDom, providing near-instant results when running on multiple cores.

Discussion

We introduce SpectralTAD, a spectral clustering-based approach that allows for fast TAD calling that scales well to high-dimensional data. The method was benchmarked against two TAD callers implemented in R - TopDom, the gold standard for detecting single-level TADs, and rGMAP, the only other two-level hierarchical TAD caller. We show similar or better performance of SpectralTAD vs. that of TopDom in all metrics while significantly outperforming rGMAP. We also demonstrate that SpectralTAD is more robust to sparsity, sequencing depth and resolution. We show the hierarchical TAD boundaries to be robustly detected and significantly associated with canonical genomic annotations of TAD boundaries, highlighting their biological importance. The clear superiority of SpectralTAD regarding running speed and robustness of TADs establishes 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 TopDom. The superior performance of TopDom in some cases may be explained by the fact that TopDom identifies non-hierarchical TAD boundaries at once, thus detecting the maximum number of them in one run. SpectralTAD, on the other hand, identifies a hierarchy of primary, secondary, etc., TADs. These TADs can vary in quality across levels making one-to-one comparison difficult. Additionally, primary TADs in one dataset can become secondary or tertiary TADs in another, thus adding another layer of complexity. Primary, secondary, and tertiary TADs detected by SpectralTAD were combined whenever possible and compared with TopDom (same was done for the two levels of TADs detected by rGMAP). This makes comparison possible but does not overcome some of the aforementioned issues. Still, SpectralTAD showed excellent performance in terms of TAD consistency, biological significance, and run time, providing an efficient tool for further dissection of TAD hierarchy.

One potential limitation of our method is that it cannot necessarily detect overlapping TAD boundaries of greater than five bins. Overlapping boundaries occur when a TAD begins within a neighboring TAD. In practice, there is no simple way to differentiate between an overlapping TAD boundary that is larger than the minimum TAD size (5 bins) and a sub-TAD other than finding all TAD boundary permutations and performing a visual inspection. To date, methods for distinguishing between a sub-TAD and an overlapping TAD remain undeveloped.

In summary, we have shown that SpectralTAD is a viable method for finding TADs at a single-level and also at multiple levels of hierarchy. This method builds 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 the method as an open source R package, we hope to provide an easy-to-use, accurate tool for hierarchical TAD detection.

Methods

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.

The contact matrix as an adjacency matrix

Our method relies on the fact that the 3D chromosome can be thought of as a naturally occurring graph. 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. 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 (Dixon et al. 2012; Sofueva et al. 2013). 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 10kb 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 (Dixon et al. 2012; Won et al. 2016; Jiang et al. 2017; Norton et al. 2018).

The restriction in window size means that the maximum resolution at which the algorithm can be run is 200kb. 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 100kb (Dixon et al. 2012). 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 (Luxburg 2007); but since our method builds upon the multiclass spectral clustering algorithm (Yu and Shi 2003), which uses the symmetric Laplacian, we use the normalized symmetric Laplacian as follows:

1. Calculating the normalized symmetric Laplacian:

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

where $D = diag(\mathbf{1}^T C)$

2. Solve the generalized eigenvalue problem:

$$\bar{L}\bar{V} = \lambda\bar{V}$$

The result is a matrix of eigenvectors $\bar{V}_{w\times k}$, 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 $\bar{L}$.

3. Normalize rows and columns to sum to 1:

$$\hat{V_{i.}} = \frac{\bar{V_{i.}}}{\|\bar{V_{i.}}\|}$$

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 (Yu and Shi 2003), 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). 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:

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

2. For $i = 2, ..., n$ 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:

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

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:

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

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:

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

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

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. Cluster 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 (Dixon et al. 2012; Won et al. 2016; Jiang et al. 2017; Norton et al. 2018). 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). 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 if 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.

Identification and removal of gaps

We define gaps as regions where there is no coherent connectivity structure. There are two types of gaps, those with no contacts at all (centromeres and unsequenced regions), and regions where contacts exist but TADs are not present. The first category of gaps is removed by simply getting rid of loci with more than 95% zero contacts. Since there are situations when TADs span unsequenced regions, we allow TADs to start on one side of a gap and end on another. This is done by essentially treating regions on either side of a gap as being adjacent. The second category of gaps is detected using the silhouette score. Regions where contacts are present but no TADs exist will frequently have low silhouette scores due to the poor similarity between each locus within the region. As a result, we can detect this type of gap by treating it as a potential TAD then filter it out if its silhouette score is lower than .25.

Data Sources

Hi-C matrices from Gm12878 cell line were downloaded from Gene Expression Omnibus (GEO accession number GSE63525) (Rao et al. 2014) at 50kb, 25kb, and 10kb resolution. For simulations, we used 25 simulated matrices downloaded from the HiCToolsCompare repository (Forcato et al. 2017). These matrices were generated in 40kb resolution with manually annotated non-hierarchical TADs. Data for chromatin states, histone modification and transcription factor binding sites (TFBS) were downloaded from the UCSC genome browser database (Rosenbloom et al. 2012). 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 was downloaded in Browser Extensible Data (BED) format using the hg19/GRCh37 genome coordinate system (Table 1, Supplementary Table S2).

Table 1. Summary of the data. All data, except simulated matrices, are Gm12878 cell line-specific. TFBS - transcription factor binding sites.

Data type	Data Source	Reference
Gm12878	GSE63525	(Rao et al. 2014)
Simulated Matrices	HiCToolsCompare	(Forcato et al. 2017)
Chromatin states	UCSC Browser	(Rosenbloom et al. 2012)
Histone Modifications	UCSC Browser	(Rosenbloom et al. 2012)
TFBS	UCSC Browser	(Rosenbloom et al. 2012)

Simulating levels of noise, sparsity, and sequencing depth

Simulated contact matrices (Forcato et al. 2017) (Supplementary Table S4) were modified to simulate various levels of noise, sparsity, and sequencing depth. The noise was added by randomly selecting a percentage (4%, 8%, 12%, 16%, and 20%) of entries in the matrix and adding a constant of two to these entries. Entries were sampled with replacement meaning certain entries may have received more noise than others. In summary, we used five replicates at each noise level, totaling 25 simulated matrices.

We created two extra sets of contact matrices for simulating sparsity and sequencing depth. To mimic sparsity, we took the five simulated matrices with the minimum level of noise (4%) and introduced 90%, 75%, 50%, 25%, or 10% of zeros uniformly at random, totaling 25 matrices. To simulate changes in sequencing depth, we took the same five minimum noise matrices and applied the downsampling procedure adapted from (Yardimci et al. 2017). Briefly, the full contact matrix was converted into a vector of pairwise individual intra-chromosomal contact counts. The vector was downsampled uniformly at random proportional to the level of downsampling (1/2, 1/4, 1/8, and 1/16). Following downsampling, the vector was re-binned to the original contact matrix. This procedure produced a set of 20 matrices (five matrices, each modified by four levels of downsampling) with varying levels of downsampling.

Transformation of Hi-C data

The fundamental property of Hi-C data is the power-law decay of chromatin interaction frequencies with the increasing linear distance between interacting regions (Dekker and Mirny 2016; Lieberman-Aiden et al. 2009; Dekker et al. 2013; Lajoie et al. 2015; Nagano et al. 2015; Naumova et al. 2013; Sanborn et al. 2015). Methods to account for the effect of distance-dependent decay have been developed (Yu et al. 2017; Sexton et al. 2012; Duan et al. 2010), including the log10-transformation and Toeplitz transformation (Chen et al. 2016). We investigated how the log10 and Toeplitz transformation of simulated and real Hi-C data affect TAD detection.

Normalization of Hi-C data

Normalization of Hi-C data matrices is a common step in Hi-C data analysis (Lieberman-Aiden et al. 2009; Imakaev et al. 2012; Knight and Ruiz 2012; Cournac et al. 2012; Hu et al. 2012; Li et al. 2015; Ay et al. 2014). To test for the effect of normalization on the detection of TADs, we applied iterative correction and eigenvector decomposition (ICE) (Imakaev et al. 2012), Knight-Ruiz (KR) (Rao et al. 2014; Knight and Ruiz 2012) normalization, and the Sequential Component Normalization (SCN) (Cournac et al. 2012) to the simulated and real Hi-C matrices. ICE was implemented using the `ICE` function from the `dryHiC` R package version 0.0.0.9000. KR-normalization (Vidal et al. 2018) was performed using Juicer (Durand et al. 2016), and SCN normalization was implemented using HiCcompare version 1.5.0 (Stansfield et al. 2018).

Jaccard and its modified version as a measure of similarity between TADs

Traditionally, Jaccard is used as a measure of overlap between sets. We use it as a measure of overlap between TAD boundaries. Given a set of TAD boundaries A and B, we define the Jaccard as:

$$J = \frac{A \cap B}{A \cup B}$$

In plain terms, this is the set of shared boundaries divided by the total number of unique boundaries.

While we expect TADs called at different resolutions of the same Hi-C data to be nearly identical, TADs called from a higher-resolution data may be finer partitioned than those called from lower-resolution data. The traditional Jaccard measure will penalize for these finer TADs even though the original boundaries were detected. We introduce a modified Jaccard score of J_a, which accounts for the difference between TADs detected across resolutions.

$$J_a = \frac{A \cap B}{min(|A|, |B|)}$$

Here, A and B are two sets of TAD boundaries and $|A|$ and $|B|$ indicates the size of sets (Supplementary Figure S1). This method is identical to the Jaccard statistic, but instead of dividing by the union of A and B we divide by the smallest size. A score of 1 indicates that all of set A is contained in a subset of B or vice-versa. This is in contrast to traditional Jaccard where a score of 1 indicates that all boundaries in A and B are identical.

The modified Jaccard can be extended to account for the impreciseness of TAD callers or differences in the resolution which make finding identical TADs impossible. For instance, half of the loci in a 25kb resolution contact matrix aren't actually in the 50kb resolution version of the same matrix but in one of the neighboring bins. This "off-by-one" error is accounted for by extending TAD boundary points at higher resolution by flanking regions of size f. Consequently, the modified Jaccard formula becomes:

$$J_a = \frac{\mathbf{A} \cap \mathbf{B}}{min(|A|, |B|)}$$

where $\mathbf{A} = \{A, A + f, A - f\}$ and $\mathbf{B} = \{B, B + f, B - f\}$.

When comparing the 50kb and 25kb resolution matrices, we can set $f = 25000$ to make up for any difference in resolution. Note that modified Jaccard is used only when comparing boundaries between different resolutions; otherwise, the traditional Jaccard statistic is used.

Testing for significant differences between methods

We assessed differences in TAD detection methods using several metrics. TADs were detected on multiple chromosomes and were evaluated using commonly used metrics (e.g., the number of TADs, width, the overlap of TAD boundaries, etc.). Unless otherwise stated, we use a Wilcoxon signed-rank

test to assess the statistical significance of the difference in those metrics between methods. P-values for multiple chromosomes were aggregated using the addCLT method that is robust to the presence of zero p-values (Nguyen et al. 2015). For simulated data, the metrics were summarized on a simulation-to-simulation level.

Measuring association of TAD boundaries with genomic features

To test for the enrichment of a genomic annotation at a given TAD boundary, we measure the total number of annotations within 50kb of the boundary point on either side. The flanking region accounts for the fact that a TAD boundary is a single point. Flanking also helps to correct for impreciseness due to issues like overlapping TAD boundaries and differences in resolutions. We specifically choose 50kb for comparison of genomic features because it allows for at least one bin of wiggle room for all resolutions used in this paper. By keeping the size of the flanking region consistent across resolutions, we can directly compare enrichment at TAD boundaries across resolutions.

A permutation test was used to quantify the enrichment. Briefly, the mean number of genomic annotations within 50kb of each boundary was calculated for the entire chromosome. Two sets of bins, one the size of the TAD boundaries and another the size of all other regions were sampled without replacement, and the difference in the mean number of genomic annotations in the corresponding sets was calculated. This procedure was repeated 10000 times. We determined the permutation p-value by taking the number of randomly sampled (expected) mean differences that were greater than the observed difference in means between TAD boundaries and all other regions of the chromosome and dividing by 10000. $\alpha = 0.05$ was set to assess statistical significance. Given the fact that `TopDom` detects all TADs in one run while `rGMAP` and `SpectralTAD` detect hierarchical TADs, hierarchial TAD boundaries were combined for `rGMAP` (two levels) and `SpectralTAD` (three levels).

Runtime analysis

One general drawback of spectral clustering is the fact that it scales poorly to large matrices. Traditionally, the three main bottlenecks are 1) the creation of the distance matrix, 2) the creation of the Laplacian matrix, and 3) the eigenvalue decomposition. Letting n equal the number of loci in the genome, the distance matrix creation has a complexity of $\mathrm{O}(n^3)$. The Laplacian matrix creation involves two matrix multiplication steps. The traditional multiplication step costs $\mathrm{O}(2n^3)$, and the eigendecomposition step costs $\mathrm{O}(n^3)$. In total, these bottlenecks result in computational complexity of $\mathrm{O}(4n^3)$ or more simply $\mathrm{O}(n^3)$, cubic complexity.

Our method manages to avoid all of these bottlenecks. The first bottleneck is avoided because the contact matrix is itself an adjacency matrix and doesn't require transformation to a distance matrix. The second bottleneck is solved by calculating relatively small Laplacian matrices for each window instead of calculating the Laplacian matrix of the entire contact map. The total number of windows for a given chromosome of size n with window size w is equal to $\sim \frac{n}{w}$ with some discrepancy for rounding and variable TAD size. Accordingly, the computational complexity of Laplacian matrix creation over a given number of windows is equal to $O(w^3 \frac{n}{w})$ or more simply $O(n)$. The third bottleneck is similarly addressed by the windowed approach. The window makes the computation complexity of the eigendecomposition equal to $O(Kw^2nm)$, where K is the number of eigenvalues and m is the number of iterations for convergence of the eigensolver. This reduces to $O(n)$. As

a result of these steps, the windowed spectral clustering algorithm reduces to the computational complexity of $O(2n)$, or simply linear complexity $O(n)$.

References

An L, Yang T, Yang J, Nuebler J, Li Q, Zhang Y. 2018. Hierarchical domain structure reveals the divergence of activity among TADs and boundaries.

Ay F, Bailey TL, Noble WS. 2014. Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts. *Genome Res* **24**: 999–1011.

Ay F, Noble WS. 2015. Analysis methods for studying the 3D architecture of the genome. *Genome Biol* **16**: 183.

Beagan JA, Duong MT, Titus KR, Zhou L, Cao Z, Ma J, Lachanski CV, Gillis DR, Phillips-Cremins JE. 2017. YY1 and ctcf orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res* **27**: 1139–1152.

Berkum NL van, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES. 2010. Hi-c: A method to study the three-dimensional architecture of genomes. *J Vis Exp*.

Berlivet S, Paquette D, Dumouchel A, Langlais D, Dostie J, Kmita M. 2013. Clustering of tissue-specific sub-TADs accompanies the regulation of HoxA genes in developing limbs ed. B. Ren. *PLoS Genetics* **9**: e1004018.

Bonev B, Cavalli G. 2016. Organization and function of the 3D genome. *Nat Rev Genet* **17**: 661–678.

Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot J-P, Tanay A, et al. 2017. Multiscale 3D genome rewiring during mouse neural development. *Cell* **171**: 557–572.e24.

Boulos RE, Arneodo A, Jensen P, Audit B. 2013. Revealing long-range interconnected hubs in human chromatin interaction data using graph theory. *Phys Rev Lett* **111**: 118102.

Chen F, Li G, Zhang MQ, Chen Y. 2018. HiCDB: A sensitive and robust method for detecting contact domain boundaries. *Nucleic Acids Research*.

Chen J, Hero AO 3rd, Rajapakse I. 2016. Spectral identification of topological domains. *Bioinformatics* **32**: 2151–8.

Ciabrelli F, Cavalli G. 2015. Chromatin-driven behavior of topologically associating domains. *Journal of Molecular Biology* **427**: 608–625.

Corces MR, Corces VG. 2016. The three-dimensional cancer genome. *Curr Opin Genet Dev* **36**: 1–7.

Cournac A, Marie-Nelly H, Marbouty M, Koszul R, Mozziconacci J. 2012. Normalization of a chromosomal contact map. *BMC Genomics* **13**: 436.

Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, Meyer BJ. 2015. Condensin-driven remodelling of x chromosome topology during dosage compensation.

Nature **523**: 240–4.

Dali R, Blanchette M. 2017. A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Res* **45**: 2994–3005.

Dali R, Bourque G, Blanchette M. 2018. RobusTAD: A tool for robust annotation of topologically associating domain boundaries. *bioRxiv*. https://www.biorxiv.org/content/early/2018/04/02/293175.

Dekker J, Heard E. 2015. Structural and functional diversity of topologically associating domains. *FEBS Lett* **589**: 2877–84.

Dekker J, Marti-Renom MA, Mirny LA. 2013. Exploring the three-dimensional organization of genomes: Interpreting chromatin interaction data. *Nat Rev Genet* **14**: 390–403.

Dekker J, Mirny L. 2016. The 3D genome as moderator of chromosomal communication. *Cell* **164**: 1110–21.

Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. *Science* **295**: 1306–11.

Dixon JR, Gorkin DU, Ren B. 2016. Chromatin domains: The unit of chromosome organization. *Mol Cell* **62**: 668–80.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376–80.

Dong Q, Li N, Li X, Yuan Z, Xie D, Wang X, Li J, Yu Y, Wang J, Ding B, et al. 2018. Genome-wide hi-c analysis reveals extensive hierarchical chromatin interactions in rice. *Plant J* **94**: 1141–1156.

Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, Weintraub AS, Schuijers J, Lee TI, Zhao K, et al. 2014. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**: 374–87.

Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS. 2010. A three-dimensional model of the yeast genome. *Nature* **465**: 363–367.

Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. 2016. Juicer provides a one-click system for analyzing loop-resolution hi-c experiments. *Cell Systems* **3**: 95–98.

Filippova D, Patro R, Duggal G, Kingsford C. 2014. Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology* **9**: 14.

Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. 2017. Comparison of computational methods for hi-c data analysis. *Nat Methods* **14**: 679–685.

Fotuhi Siahpirani A, Ay F, Roy S. 2016. A multi-task graph-clustering approach for chromosome conformation capture data sets identifies conserved modules of chromosomal interactions. *Genome Biol* **17**: 114.

Fraser J, Ferrai C, Chiariello AM, Schueler M, Rito T, Laudanno G, Barbieri M, Moore BL, Kraemer DCA, Aitken S, et al. 2015. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol Syst Biol* **11**: 852.

Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S. 2006. *Bioinformatics and computational biology solutions using r and bioconductor*. Springer Science & Business Media.

Gibcus JH, Dekker J. 2013. The hierarchy of the 3D genome. *Mol Cell* **49**: 773–82.

Haddad N, Vaillant C, Jost D. 2017. IC-finder: Inferring robustly the hierarchical organization of chromatin folding. *Nucleic Acids Res* **45**: e81.

Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CWH, Ye C, Ping JLH, Mulawadi F, et al. 2011. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* **43**: 630–8.

Hansen AS, Cattoglio C, Darzacq X, Tjian R. 2017. Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus* **9**: 20–32.

Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, Goldmann J, Lajoie BR, Fan ZP, Sigova AA, et al. 2016. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**: 1454–8.

Hong S, Kim D. 2017. Computational characterization of chromatin domain boundary-associated genomic elements. *Nucleic Acids Research* **45**: 10403–10414.

Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. 2012. HiCNorm: Removing biases in hi-c data via poisson regression. *Bioinformatics* **28**: 3131–3.

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nat Methods* **9**: 999–1003.

Jackson DA, Pombo A. 1998. Replicon clusters are stable units of chromosome structure: Evidence that nuclear organization contributes to the efficient activation and propagation of s phase in human cells. *J Cell Biol* **140**: 1285–95.

Jhunjhunwala S, Zelm MC van, Peak MM, Murre C. 2009. Chromatin architecture and the generation of antigen receptor diversity. *Cell* **138**: 435–48.

Ji X, Dadon DB, Powell BE, Fan ZP, Borges-Rivera D, Shachar S, Weintraub AS, Hnisz D, Pegoraro G, Lee TI, et al. 2016. 3D chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell* **18**: 262–75.

Jiang Y, Loh Y-HE, Rajarajan P, Hirayama T, Liao W, Kassim BS, Javidfar B, Hartley BJ, Kleofas L, Park RB, et al. 2017. The methyltransferase SETDB1 regulates a large neuron-specific topological chromatin domain. *Nature Genetics* **49**: 1239–1250.

Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen C-A, Schmitt AD, Espinoza CA, Ren B. 2013. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**: 290–4.

Knight PA, Ruiz D. 2012. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis* **33**: 1029–1047.

Kruse K, Hug CB, Hernández-Rodríguez B, Vaquerizas JM. 2016. TADtool: Visual parameter identification for tad-calling algorithms. *Bioinformatics* **32**: 3190–3192.

Lajoie BR, Dekker J, Kaplan N. 2015. The hitchhiker's guide to hi-c analysis: Practical guidelines. *Methods* **72**: 65–75.

Lévy-Leduc C, Delattre M, Mary-Huard T, Robin S. 2014. Two-dimensional segmentation for analyzing hi-c data. *Bioinformatics* **30**: i386–92.

Li W, Gong K, Li Q, Alber F, Zhou XJ. 2015. Hi-corrector: A fast, scalable and memory-efficient package for normalizing large-scale hi-c data. *Bioinformatics* **31**: 960–2.

Lieberman-Aiden E, Berkum NL van, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289–93.

Lupiáñez DG, Spielmann M, Mundlos S. 2016. Breaking tads: How alterations of chromatin domains result in disease. *Trends Genet* **32**: 225–37.

Luxburg U von. 2007. A tutorial on spectral clustering. *Statistics and Computing 17(4), 2007*. http://arxiv.org/abs/0711.0189v1.

Luzhin AV, Flyamer IM, Khrameeva EE, Ulianov SV, Razin SV, Gavrilov AA. 2018. Quantitative differences in tad border strength underly the tad hierarchy in drosophila chromosomes. *J Cell Biochem*.

Ma H, Samarabandu J, Devdhar RS, Acharya R, Cheng PC, Meng C, Berezney R. 1998. Spatial and temporal dynamics of dna replication sites in mammalian cells. *J Cell Biol* **143**: 1415–25.

Nagano T, Várnai C, Schoenfelder S, Javierre B-M, Wingett SW, Fraser P. 2015. Comparison of hi-c results using in-solution versus in-nucleus ligation. *Genome Biol* **16**: 175.

Naumova N, Imakaev M, Fudenberg G, Zhan Y, Lajoie BR, Mirny LA, Dekker J. 2013. Organization of the mitotic chromosome. *Science* **342**: 948–53.

Nguyen T, Tagett R, Donato M, Mitrea C, Draghici S. 2015. A novel bi-level meta-analysis approach: Applied to biological pathway analysis. *Bioinformatics* **32**: 409–416.

Nicoletti C, Forcato M, Bicciato S. 2018. Computational methods for analyzing genome-wide chromosome conformation capture data. *Curr Opin Biotechnol* **54**: 98–105.

Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, Berkum NL van, Meisig J, Sedat J, et al. 2012. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature* **485**: 381–5.

Norton HK, Emerson DJ, Huang H, Kim J, Titus KR, Gu S, Bassett DS, Phillips-Cremins JE. 2018. Detecting hierarchical genome folding with network modularity. *Nature Methods* **15**: 119–122.

Oluwadare O, Cheng J. 2017. ClusterTAD: An unsupervised machine learning approach to detecting topologically associated domains of chromosomes from hi-c data. *BMC Bioinformatics* **18**: 480.

O'Sullivan JM, Hendy MD, Pichugina T, Wake GC, Langowski J. The statistical-mechanics of chromosome conformation capture. *Nucleus* **4**: 390–8.

Oti M, Falck J, Huynen MA, Zhou H. 2016. CTCF-mediated chromatin loops enclose inducible gene regulatory domains. *BMC Genomics* **17**.

Phillips JE, Corces VG. 2009. CTCF: Master weaver of the genome. *Cell* **137**: 1194–211.

Phillips-Cremins JE, Corces VG. 2013. Chromatin insulators: Linking genome organization to cellular function. *Mol Cell* **50**: 461–74.

Phillips-Cremins JE, Sauria ME, Sanyal A, Gerasimova TI, Lajoie BR, Bell JS, Ong C-T, Hookway TA, Guo C, Sun Y, et al. 2013. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**: 1281–1295.

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**: 1665–1680.

Rickman DS, Soong TD, Moss B, Mosquera JM, Dlabal J, Terry S, MacDonald TY, Tripodi J, Bunting K, Najfeld V, et al. 2012. Oncogene-mediated alterations in chromatin conformation. *Proc Natl Acad Sci U S A* **109**: 9083–8.

Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG, et al. 2012. ENCODE data in the UCSC genome browser: Year 5 update. *Nucleic Acids Research* **41**: D56–D63.

Rousseeuw PJ. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**: 53–65.

Sanborn AL, Rao SSP, Huang S-C, Durand NC, Huntley MH, Jewett AI, Bochkov ID, Chinnappan D, Cutkosky A, Li J, et al. 2015. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A* **112**: E6456–65.

Sanyal A, Lajoie BR, Jain G, Dekker J. 2012. The long-range interaction landscape of gene promoters. *Nature* **489**: 109–13.

Sauerwald N, Singhal A, Kingsford C. 2018. Analysis of the structural variability of topologically associated domains as revealed by hi-c:

Schmitt AD, Hu M, Ren B. 2016. Genome-wide mapping and analysis of chromosome architecture. *Nat Rev Mol Cell Biol* **17**: 743–755.

Serra F, Baù D, Goodstadt M, Castillo D, Filion GJ, Marti-Renom MA. 2017. Automatic analysis and 3D-modelling of hi-c data using tadbit reveals structural features of the fly chromatin colors. *PLoS Comput Biol* **13**: e1005665.

Servant N, Lajoie BR, Nora EP, Giorgetti L, Chen C-J, Heard E, Dekker J, Barillot E. 2012. HiTC: Exploration of high-throughput 'c' experiments. *Bioinformatics* **28**: 2843–4.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-dimensional folding and functional organization principles of the drosophila genome. *Cell* **148**: 458–72.

Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, Zhou XJ. 2016. TopDom: An efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res* **44**: e70.

Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, Zhou XJ. 2015. TopDom: An efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Research* **44**: e70–e70.

Sofueva S, Yaffe E, Chan W-C, Georgopoulou D, Vietri Rudan M, Mira-Bontenbal H, Pollard SM, Schroth GP, Tanay A, Hadjur S. 2013. Cohesin-mediated interactions organize chromosomal domain architecture. *EMBO J* **32**: 3119–29.

Splinter E, Heath H, Kooren J, Palstra R-J, Klous P, Grosveld F, Galjart N, Laat W de. 2006. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev* **20**: 2349–54.

Stansfield JC, Cresswell KG, Vladimirov VI, Dozmorov MG. 2018. HiCcompare: An r-package for joint normalization and comparison of HI-c datasets. *BMC Bioinformatics* **19**.

Taberlay PC, Achinger-Kawecka J, Lun ATL, Buske FA, Sabir K, Gould CM, Zotenko E, Bert SA, Giles KA, Bauer DC, et al. 2016. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Res* **26**: 719–31.

Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J, Ruszczycki B, et al. 2015. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* **163**: 1611–27.

Valton A-L, Dekker J. 2016. TAD disruption as oncogenic driver. *Curr Opin Genet Dev* **36**: 34–40.

Vidal E, Dily F le, Quilez J, Stadhouders R, Cuartero Y, Graf T, Marti-Renom MA, Beato M, Filion GJ. 2018. OneD: Increasing reproducibility of hi-c samples with abnormal karyotypes. *Nucleic Acids Res*.

Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S. 2015. Comparative hi-c reveals that ctcf underlies evolution of chromosomal domain architecture. *Cell Rep* **10**: 1297–309.

Wang H, Duggal G, Patro R, Girvan M, Hannenhalli S, Kingsford C. 2013. Topological properties of chromosome conformation graphs reflect spatial proximities within chromatin. In *Proceedings of the international conference on bioinformatics, computational biology and biomedical informatics*, *BCB'13*, pp. 306:306–306:315, ACM, New York, NY, USA.

Wang X-T, Cui W, Peng C. 2017a. HiTAD: Detecting the structural and functional hierarchies of topologically associating domains from chromatin interactions. *Nucleic Acids Res* **45**: e163.

Wang X-T, Dong P-F, Zhang H-Y, Peng C. 2015. Structural heterogeneity and functional diversity of topologically associating domains in mammalian genomes. *Nucleic Acids Research* **43**: 7237–7246.

Wang Y, Sarkar P, Ursu O, Kundaje A, Bickel PJ. 2017b. Network modelling of topological domains using hi-c data. *arXiv preprint arXiv:170709587*.

Weinreb C, Raphael BJ. 2015. Identification of hierarchical chromatin domains. *Bioinformatics* **32**: 1601–1609.

Weinreb C, Raphael BJ. 2016. Identification of hierarchical chromatin domains. *Bioinformatics* **32**: 1601–9.

Won H, Torre-Ubieta L de la, Stein JL, Parikshak NN, Huang J, Opland CK, Gandal MJ, Sutton GJ, Hormozdiari F, Lu D, et al. 2016. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**: 523–527.

Yaffe E, Tanay A. 2011. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* **43**: 1059–65.

Yan K-K, Lou S, Gerstein M. 2017. MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions ed. S. Zhong. *PLOS Computational Biology* **13**: e1005647.

Yardimci GG, Ozadam H, Sauria ME, Ursu O, Yan K-K, Yang T, Chakraborty A, Kaul A, Lajoie BR, Song F, et al. 2017. Measuring the reproducibility and quality of hi-c data.

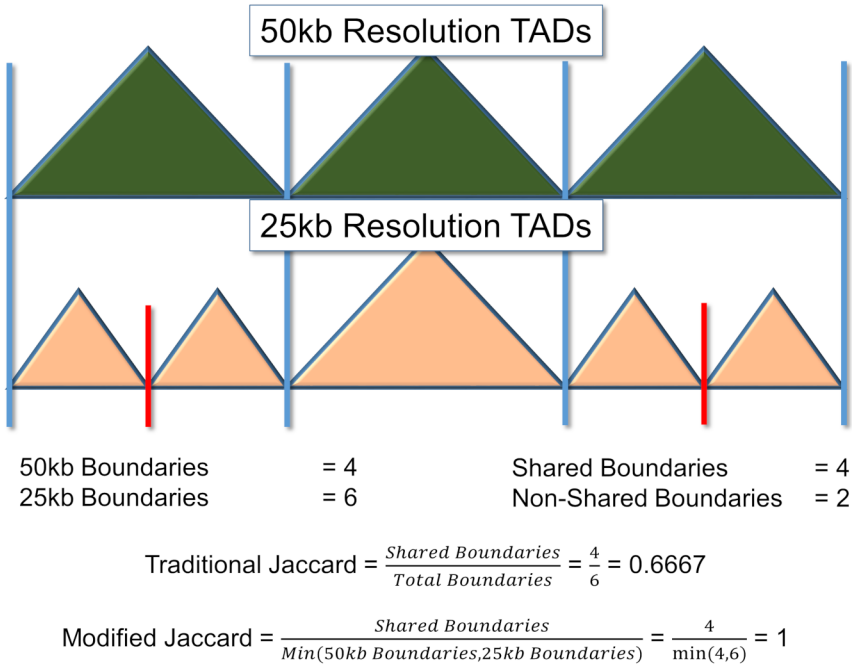
Yu SX, Shi J. 2003. Multiclass spectral clustering. In *Proceedings of the ninth ieee international conference on computer vision - volume 2*, *ICCV '03*, pp. 313, IEEE Computer Society, Washington, DC, USA.

Yu W, He B, Tan K. 2017. Identifying topologically associating domains and subdomains by gaussian mixture model and proportion test. *Nature Communications* **8**.

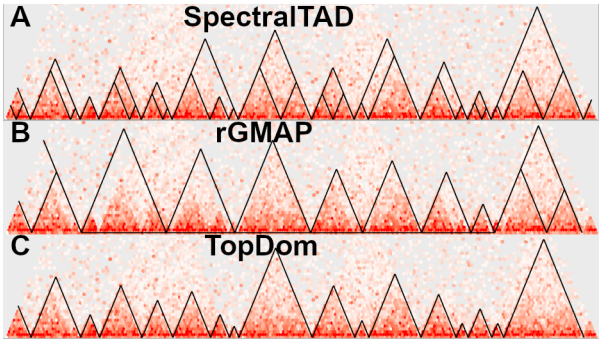
Zhan Y, Mariani L, Barozzi I, Schulz EG, Blüthgen N, Stadler M, Tiana G, Giorgetti L. 2017. Reciprocal insulation analysis of hi-c data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome Research* **27**: 479–490.

Zuin J, Dixon JR, Reijden MIJA van der, Ye Z, Kolovos P, Brouwer RWW, Corput MPC van de, Werken HJG van de, Knoch TA, IJcken WFJ van, et al. 2014. Cohesin and ctcf differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci U S A* **111**: 996–1001.

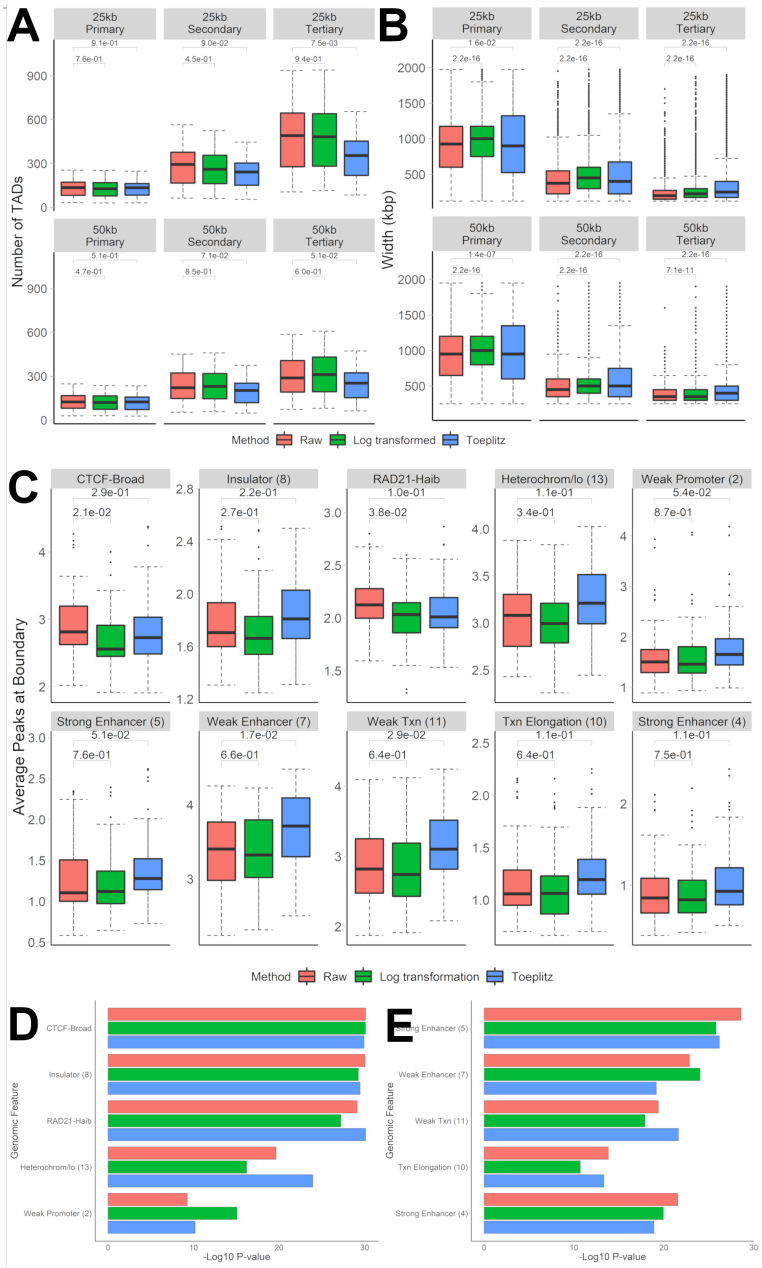
Supplementary Figures



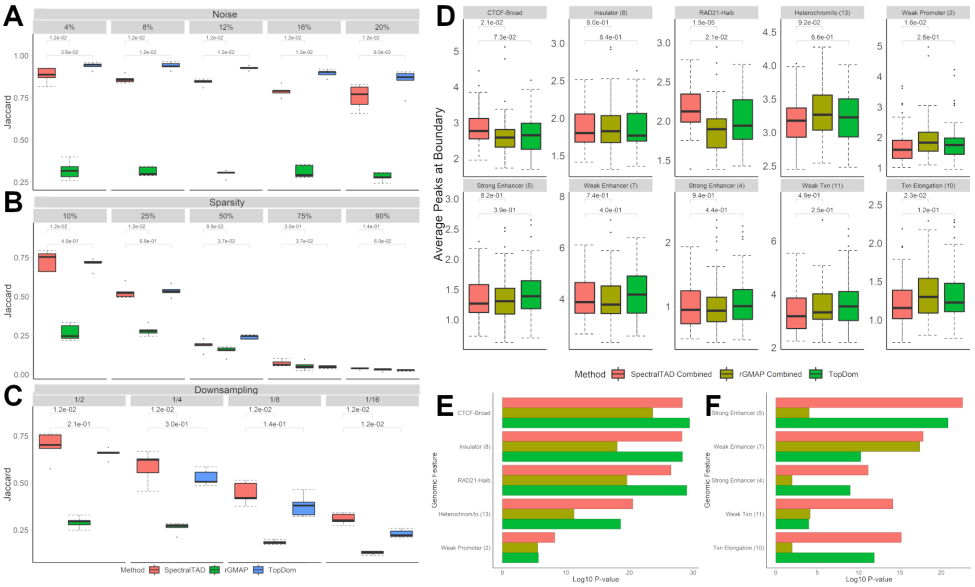
Supplementary Figure S1. Example of modified Jaccard statistics to measure agreement between TAD boundaries detected at different resolutions. The top triangles indicate TADs detected at 50kb resolution, while the bottom triangles indicate those detected at 25kb resolution. There are four shared boundaries (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.



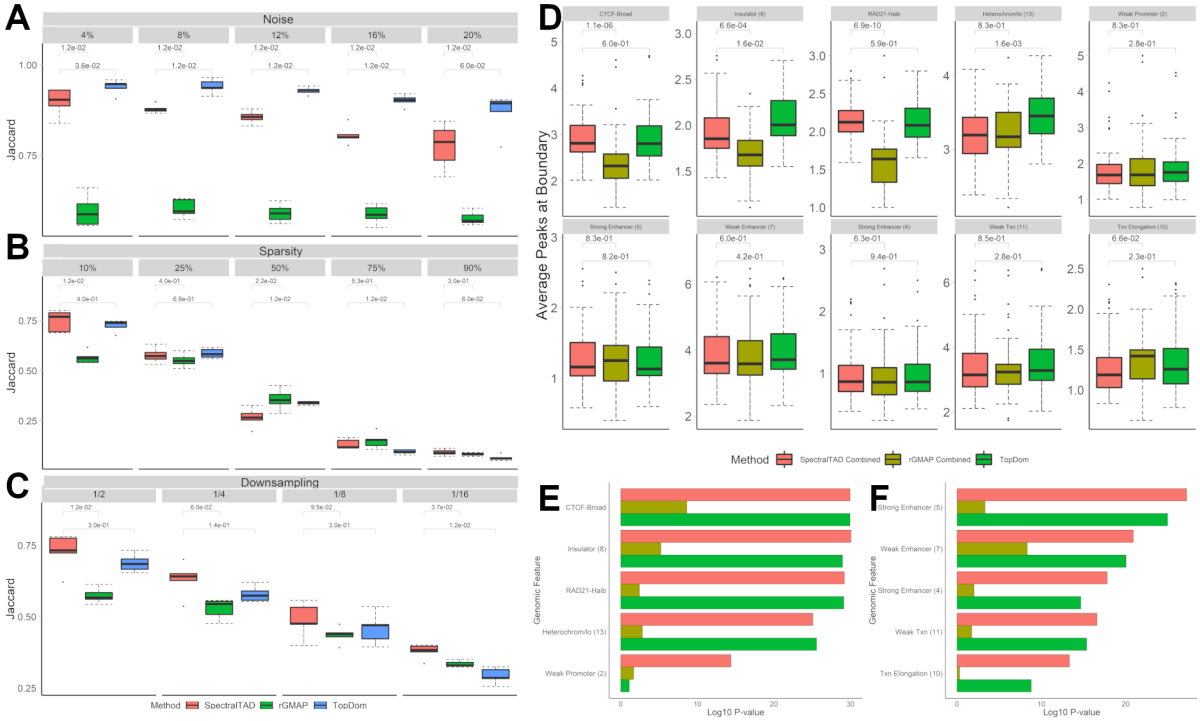
Supplementary Figure S2. Examples of TADs detected by different TAD callers. Hierarchical TADs detected by SpectralTAD (A) and rGMAP (B). Non-hierarchical TADs detected by TopDom (C). Darker colors indicate a higher level of connectivity; triangles indicate TADs. Data are shown from simulated matrices designed to emulate TADs at 40kb resolution (Supplementary Table S4). All parameters were set according to the instructions of each package for analyzing 40kb resolution data.



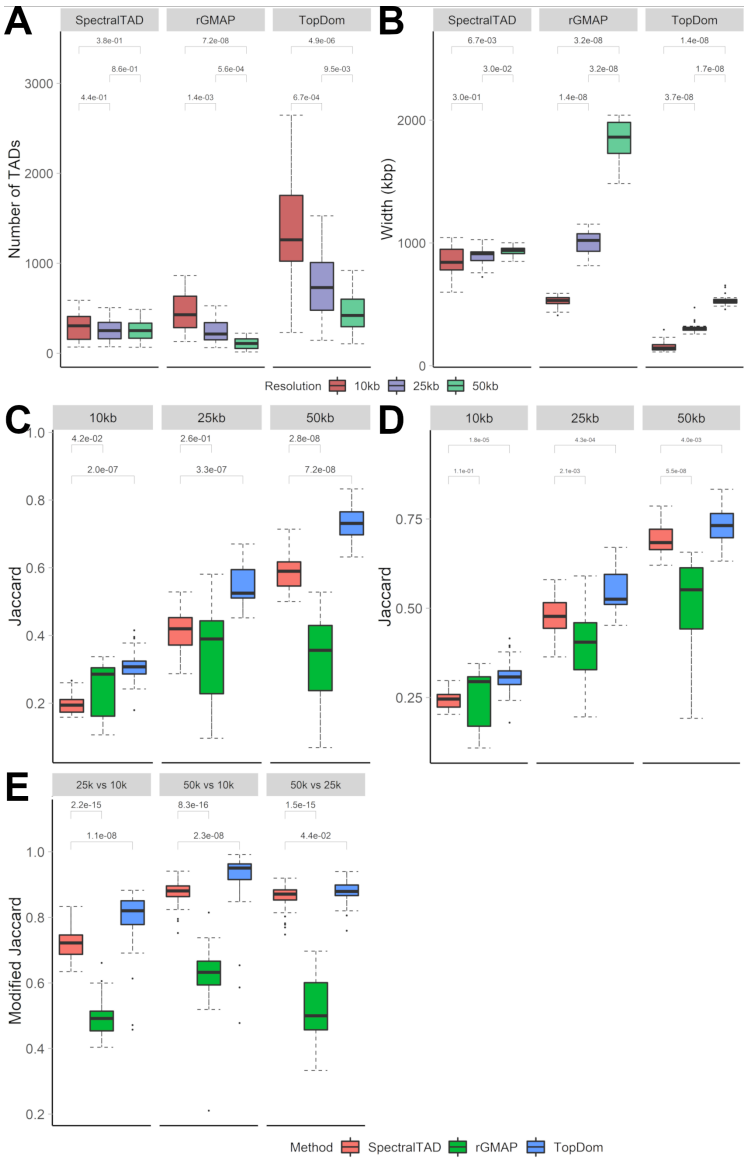
Supplementary Figure S3. The effect of data transformation on the average number (A) and width in kilobases (B) of TADs and the average number of annotations (C), enrichment (D) and depletion (E) for different genomic annotations. Counts (A) and widths (B) for raw, log10-, or Toeplitz-transformed Gm12878 data at 25kb and 50kb resolutions, averaged across chromosome 1-22, are shown for primary, secondary, and tertiary TADs detected by SpectralTAD. The average number of annotations (C), enrichment (D) and depletion (E) for the top five most enriched/depleted genomic annotations at TAD boundaries for Gm12878 data at 50kb resolution are shown. Wilcoxon test p-values are shown in panels A, B, and C, and aggregated p-values using the addCLT method are shown for panels D and E.



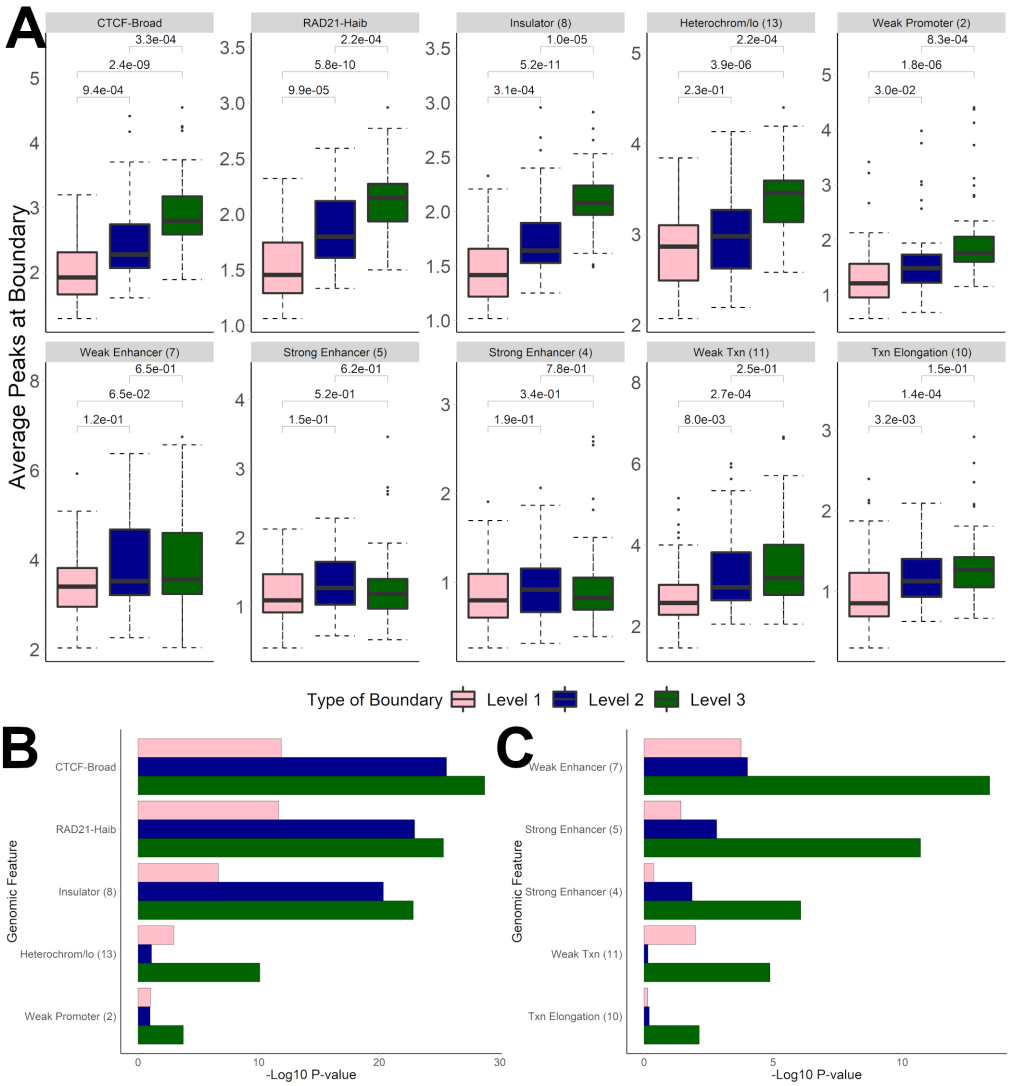
Supplementary Figure S4. The effect of data normalization on the average number (A) and width in kilobases (B) of TADs and the average number of annotations (C), enrichment (D) and depletion (E) for different genomic annotations. Counts (A) and widths (B) for as-is, KR-, ICE- and SCN-normalized Gm12878 data at 25kb and 50kb resolutions, averaged across chromosome 1-22, are shown for primary, secondary, and tertiary TADs detected by SpectralTAD. The average number of annotations (C), enrichment (D) and depletion (E) for the top most enriched/depleted genomic annotations at TAD boundaries for Gm12878 data at 50kb resolution are shown. Wilcoxon test p-values are shown in panels A, B, and C and aggregated p-values, using the addCLT method, are shown for panels D and E.



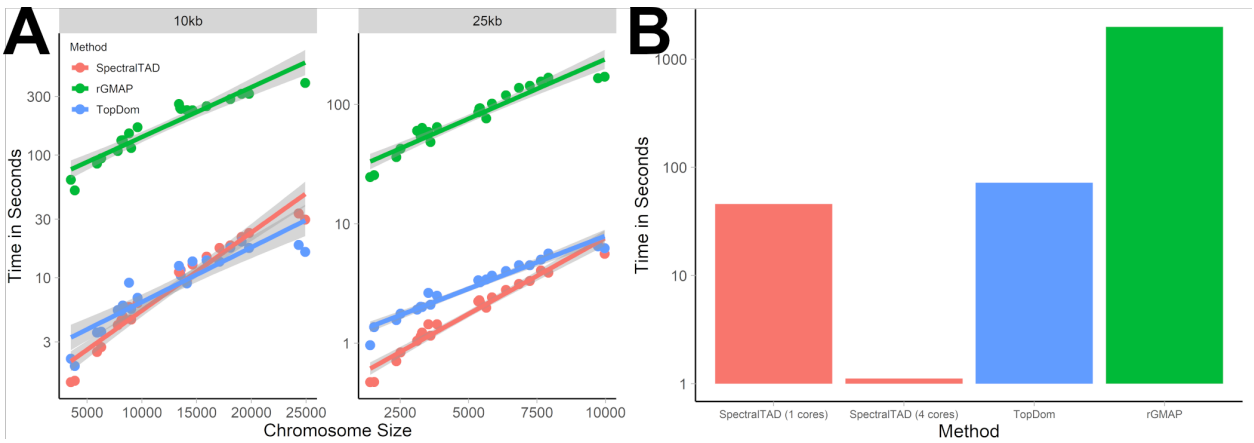
Supplementary Figure S5. The comparison of SpectralTAD, rGMAP, and TopDom on TAD consistency and enrichment. To test for robustness to noise, sparsity, and downsampling, TADs were called from simulated Hi-C matrices using SpectralTAD, rGMAP, and TopDom. The TAD boundaries were extended by 50kb regions 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 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 50kb resolution, enrichment of genomic annotations within 50kb regions flanking a TAD boundary on both sides was assessed using a permutation test. The average number of annotations (D), enrichment (E), and depletion (F) for the top five most enriched/depleted genomic annotations are shown. Results averaged across chromosome 1-22 are shown. Wilcoxon test p-values are shown in panels A, B, C, and D and aggregated p-values, using the addCLT method, are shown for panels E and F.



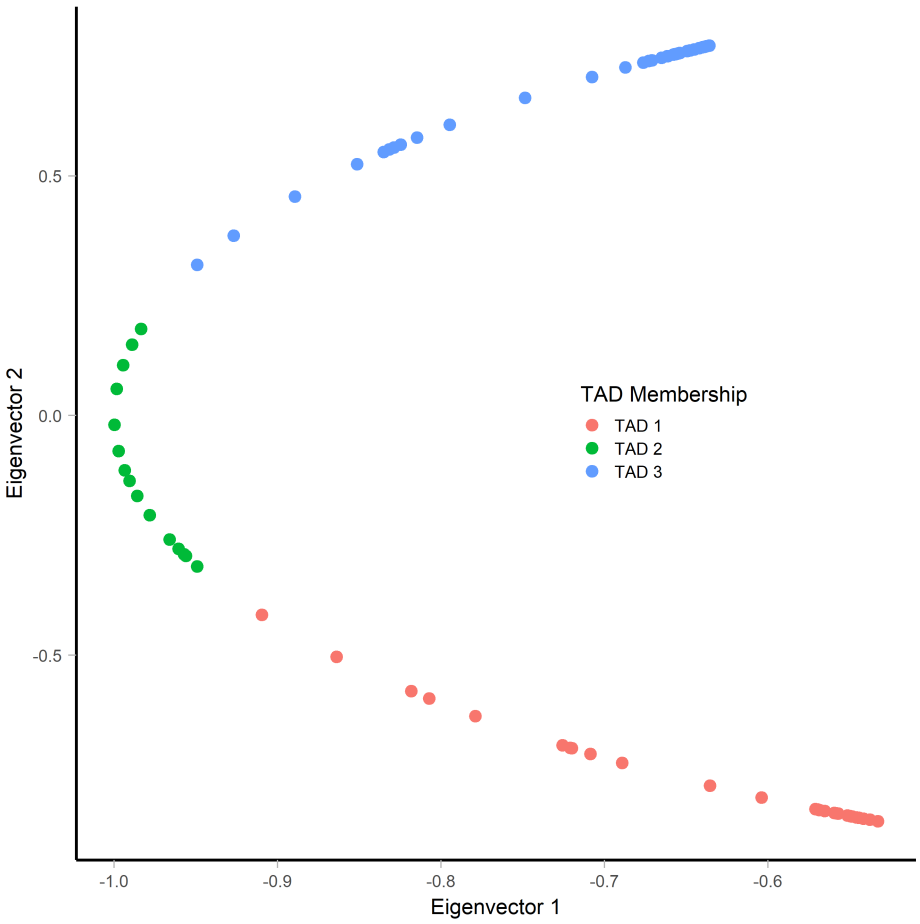
Supplementary Figure S6. The number, width, and consistency of TADs called across resolutions and primary vs. replicate for different methods. The average number (A) and width (B) of TADs across resolutions, Jaccard similarity between TAD boundaries detected from primary and replicate data (C), Jaccard similarity between TAD boundaries detected from primary and replicate data and extended by 50kb regions on both sides (D), and modified Jaccard similarity between TAD boundaries detected from data at 10kb, 25kb and 50kb resolutions (E). Wilcoxon test p-values are shown. Data from Gm12878 cell line, chromosomes 1-22.



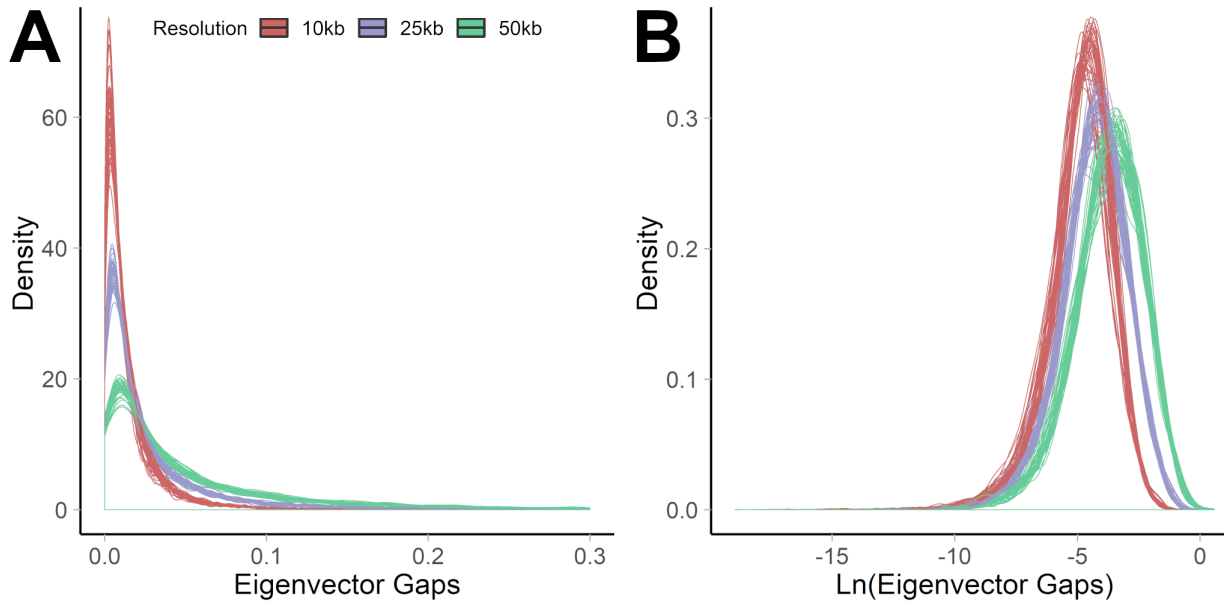
Supplementary Figure S7. The effect of the hierarchy of TAD boundaries detected by SpectralTAD on the average number of annotations (A), enrichment (B) and depletion (C) for different genomic annotations. Results for TAD boundaries detected as level 1, 2, and 3 boundaries are shown. Genomic annotations were considered within 50kb regions flanking a boundary on both sides. Wilcoxon test p-values are shown in panel A and aggregated p-values, using the addCLT method, are shown for panels B and C. Data from Gm12878 cell line, chromosome 1-22, 50kb resolution.



Supplementary Figure S8. Runtime performance of various TAD callers. TADs were called using data from Gm12878 cell line at 10kb and 25kb 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 25kb resolution data. X-axis – Method, Y-axis – time in seconds.



Supplementary Figure S9. Projection of eigenvectors on the unit circle. This projection allows us to identify TADs based on the distance between eigenvectors. A gap separates TAD 2 from both TAD 1 and TAD 3. We can also see the difference between a strongly organized group with close together points (TAD 2) and a weaker group with more spread out points (TAD 1 and TAD 3).



Supplementary Figure S10. Distribution of eigenvector gaps. The distributions of eigenvector gaps are plotted separately for each 10kb, 25kb and 50kb contact matrix from (Rao et al. 2014), 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.

Supplementary Table S1. Summary of gaps. The percentage of gaps is summarized for all chromosomes at 10kb, 25kb, and 50kb resolution. Gaps are separated based on whether they are centromeric or other (unsequenced, or poorly organized chromatin). The percentage listed is average between a primary and replicate for each resolution and chromosome.

Supplementary Table S2. Real Data sources. Hi-C (Rao et al. 2014) and genome annotation (hg19/GRCh37) (Rosenbloom et al. 2012) data for Gm12878 cell line used in the analysis, sorted by category, then by data type.

Supplementary Table S3. Enrichment by Method. Enrichment/Depletion results are provided for all genomic annotations tested. P-values and whether the genomic marker was enriched or depleted are shown for each resolution and method. All p-values were obtained from a permutation test (see Methods) and summarized using the addCLT method. Data is sorted alphabetically by category and then by genomic annotation.

Supplementary Table S4. Simulated Data sources. Information about simulated contact matrices is provided including file names, URL for downloading and the analysis in which they were used.