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, and more conserved across cell lines and tissues, 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 and less conserved, 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 Bioconductor, https://bioconductor.org/packages/devel/bioc/html/SpectralTAD.html.

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; 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 cell type-specific gene expression (Phillips-Cremins and Corces 2013; Rao et al. 2014; Gierman et al. 2007), replication (Pope et al. 2014), guide X chromosome inactivation (Nora et al. 2012), and regulate the expression of tumor suppressors and oncogenes (Valton and Dekker 2016; Taberlay 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; Nora et al. 2012; Sexton et al. 2012; Dekker et al. 2002). 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; Tang et al. 2015; Phillips and Corces 2009), marks of transcriptionally active chromatin (e.g., DNAse hypersensitive sites, H3K4me3, H3K27ac, H3K36me3 histone modifications) (Hong and Kim 2017; Ciabrelli and Cavalli 2015), actively transcribed and housekeeping genes (Oti et al. 2016). 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; Ciabrelli and Cavalli 2015).

TADs were shown to organize themselves into hierarchical sets of domains (Weinreb and Raphael 2016; Fraser et al. 2015; Gibcus and Dekker 2013; Dong et al. 2018). These hierarchies are characterized by large "meta-TADs" that contain smaller sub-TADs. To date, most methods were developed to find these single meta-TADs instead of focusing on the hierarchy of the TAD structures (Shin et al. 2016; Lévy-Leduc et al. 2014; Durand et al. 2016). 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.

Several methods have been designed to call hierarchical TADs (Supplemental Material). However, most algorithms require tunable parameters (Filippova et al. 2014; Lévy-Leduc et al. 2014; Wang et al. 2017) 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 (Dali and Blanchette 2017)). 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 (Chen et al. 2016; Wang et al. 2017), with some methods even lacking publicly available code (Fraser et al. 2015). Furthermore, the choice of TAD callers in R/Bioconductor ecosystem remains limited (Supplemental Material).

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 hierarchy of sub-TADs. The method employs a modified version of the multiclass spectral clustering algorithm (Yu and Shi 2003), using a sliding window based on the commonly used 2Mb biologically maximum TAD size (Dixon et al. 2012; Dekker and Heard 2015). 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 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 (https://bioconductor.org/packages/devel/bioc/html/SpectralTAD.html).

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 defining TAD boundaries (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 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).

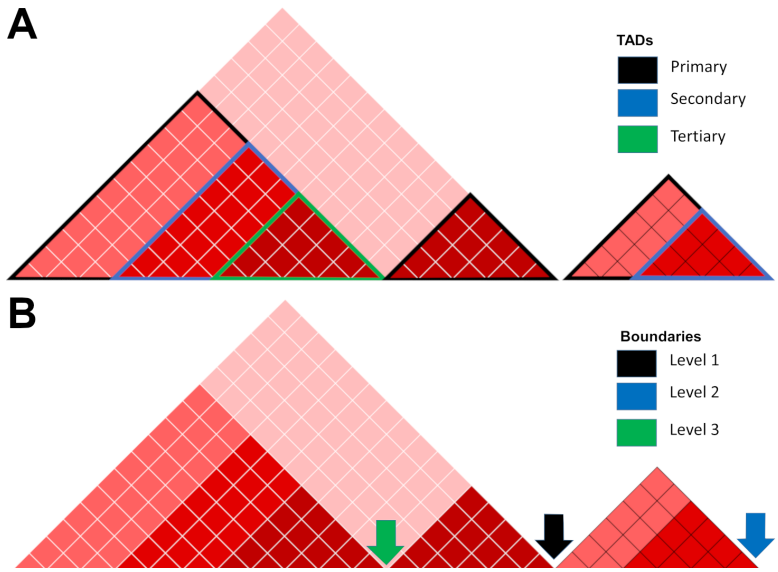

Figure 1. Hierarchy of TADs and boundaries. A) TADs not enclosed within other TADs are defined as primary, while TADs contained within other TADs are defined as secondary, tertiary, etc. B) Boundaries belonging to a single TADs are defined as level 1, while boundaries shared by two, three, TADs are defined 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 (Supplemental Methods). The percentage of non-centromeric gaps varies across chromosomes and resolutions (Supplemental Table S1), being 21.8% on average. 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 (Supplemental Figure S1, Supplemental 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 (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 (Supplemental 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. 2016; 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 (Supplemental Methods, Supplemental Table S2).

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 2A-C). Interestingly, ICE normalization marginally affected TAD detection (Figure 2A-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 (Supplemental Figure S3A-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 2D-F). 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 2E). Similarly, "enhancer"-like chromatin states were significantly depleted at TAD boundaries, and this depletion was more pronounced in boundaries detected from unnormalized data (Figure 2F). 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 2D). The enrichment results were consistent across resolutions (Supplemental Figure S3C-E). These results suggest that unnormalized Hi-C data is perfectly suitable for the robust detection of biologically relevant TADs. Consequently, the following results are presented with the use of unnormalized Hi-C data.

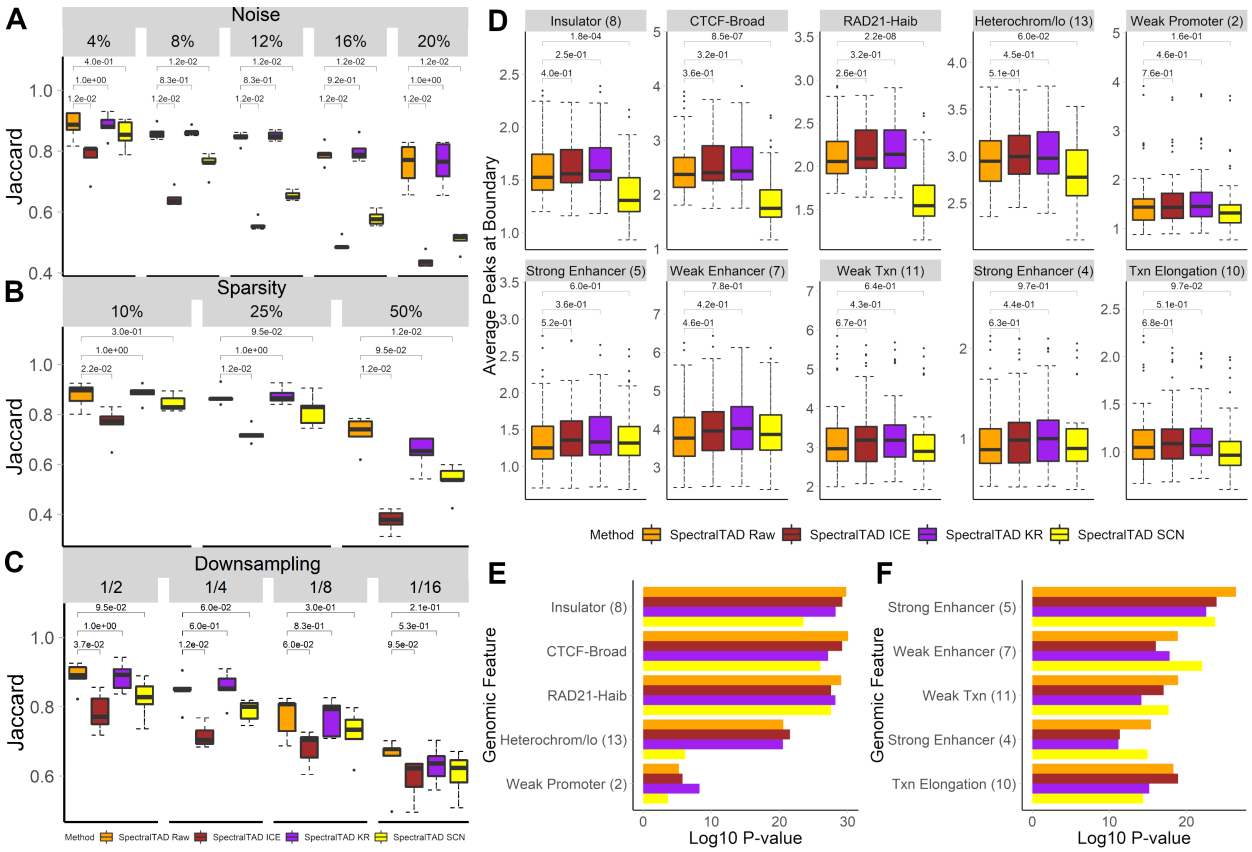

Figure 2. The effect of normalization on TAD consistency and enrichment. To test for robustness to noise, sparsity, and downsampling, simulated Hi-C matrices were used as-is, KR-, ICE- and SCN-normalized. TADs detected by SpectralTAD were compared with the ground-truth TADs using the Jaccard similarity 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 annotations 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 (Nguyen et al. 2015), 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 3A). 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 bound-

aries flanked by 50kb regions. Expectedly, the performance of all TAD callers, including `rGMAP`, increased; yet, the performance of `rGMAP` remained significantly low (Supplemental Figure S4A). 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.

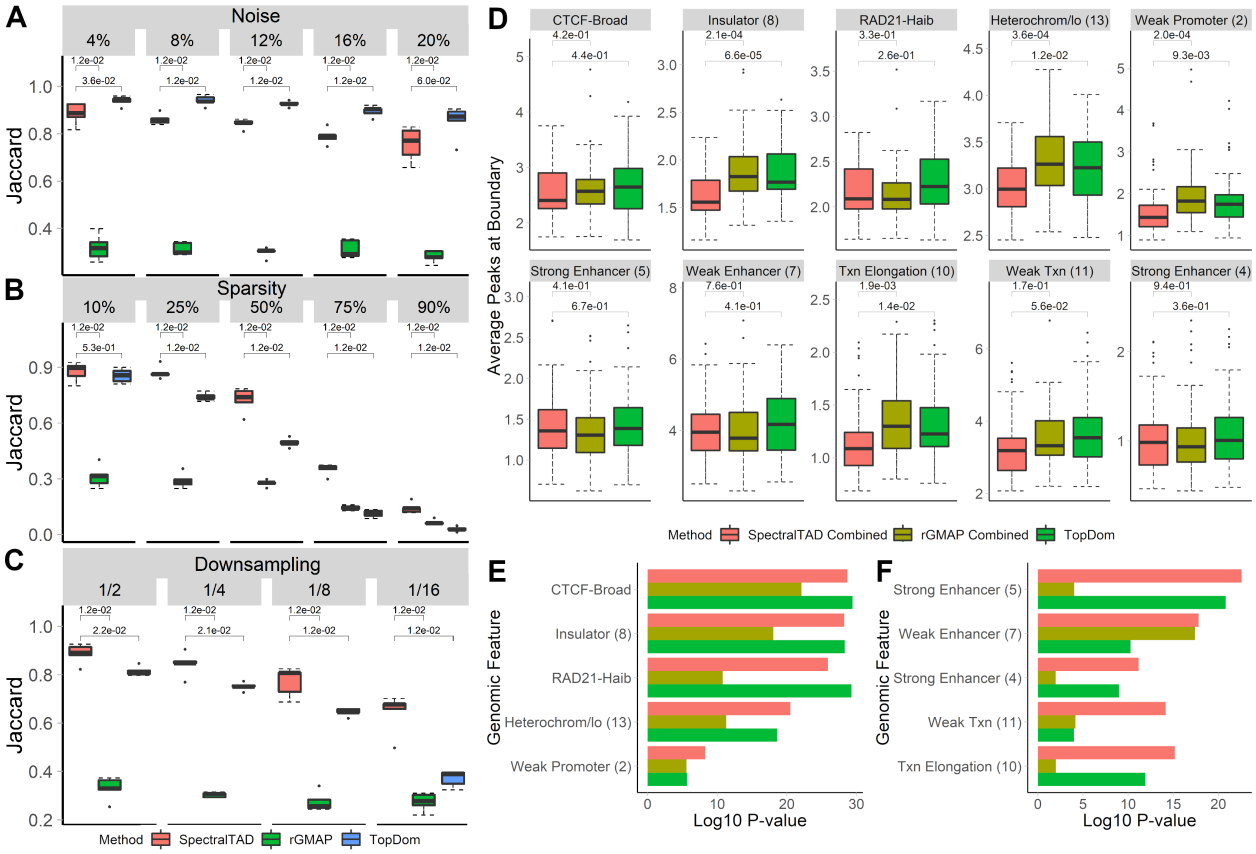

Figure 3. The comparison of SpectralTAD, rGMAP, and TopDom on TAD consistency and enrichment. To test for robustness to noise, sparsity, and downsampling, TADs were called from simulated Hi-C matrices using SpectralTAD, rGMAP, and TopDom. They were compared with the ground-truth TADs using the Jaccard 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 (Nguyen et al. 2015), 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 3B). 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 (Supplemental Figure S4B). 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 3C). 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 (Supplemental Figure S4C). 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 (Supplemental 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 3D). Consequently, CTCF, RAD21, "Insulator," and "Heterochromatin" states were more enriched at TAD boundaries detected by SpectralTAD as compared with rGMAP (Figure 3E). SpectralTAD also identified more significant depletion of "enhancer"-like chromatin states at TAD boundaries than rGMAP (Figure 3F). 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 3D-F). These observations were consistent at different resolutions (Supplemental Figure S4D-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 (Supplemental Figure S5A-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 (Supplemental Figure S5E). 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 (Supplemental Figure S5C), and these results were similar when accounting for the "off-by-one" error (Supplemental Figure S5D). 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 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 4A-B). A similar trend was observed in the enrichment of "Insulator," and "Heterochromatin" states, and in the depletion of "enhancer"-like chromatin states (Figure 4A-C), and these observations were consistent across resolutions (Supplemental Figure S6). 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.

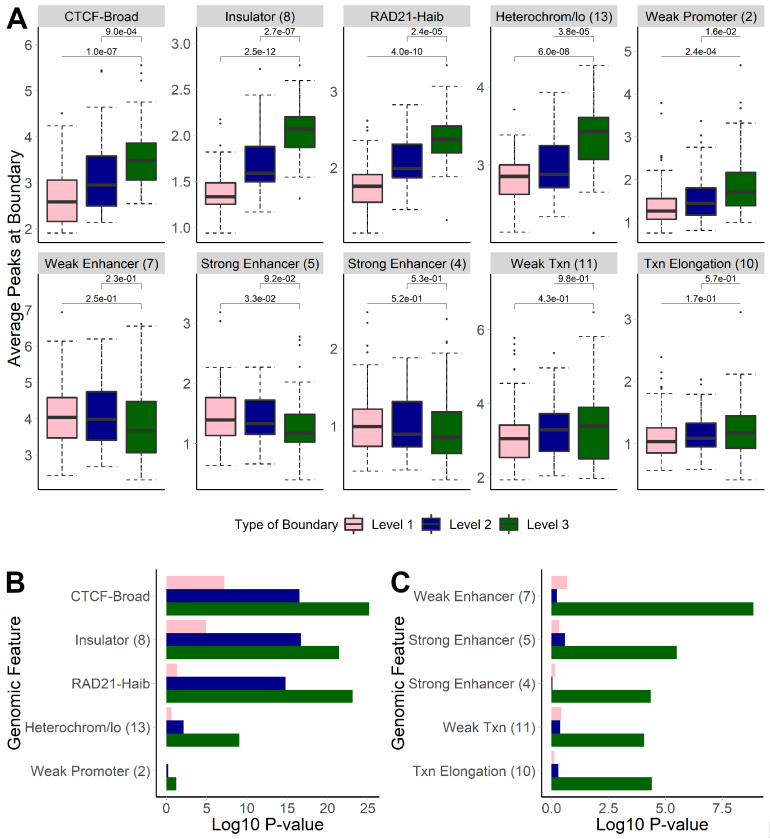

Figure 4. The effect of the hierarchy of TAD boundaries detected by SpectralTAD on the average number of annotations (A), enrichment (B) and depletion (C) for different genomic annotations. Results for TAD boundaries detected as level 1, 2, and 3 boundaries are shown. Genomic annotations were considered 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 the GM12878 cell line, chromosome 1-22, 25kb resolution.

TADs identified by SpectralTAD are conserved across cell-line 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 (Rao et al. 2014). We compared TAD boundaries called by SpectralTAD across various tissues and cell types (Supplemental Table S4, (Schmitt et al. 2016)). 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 for 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.

Hierarchical clustering of cell type-specific samples by their TADs, ignoring hierarchy, using the Jaccard statistics as a distance metric identified the expected similarity between replicates and differences between cell types (Supplemental Figure S7A). These results were similar in tissue-

specific samples, with replicated samples, or samples from the same tissue origin clustering together (Supplemental Figure S7B). In summary, these results show conservation of TAD boundaries called by SpectralTAD across tissues and cell lines similar to previously reported results.

Hierarchy of TADs and boundaries affect conservation of TADs

Following our definition of TADs (Primary, Secondary, and Tertiary, Figure 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 (Supplemental Table S5). These differences were significant in any pairwise comparison of the Jaccard coefficients (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 (Supplemental Table S5). These results demonstrate that Primary TADs are the most conserved across cell types and tissues.

We hypothesized that Level 3 TAD boundaries (Figure 1B, boundaries that are shared by three TADs), besides showing higher biological significance (Figure 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, Supplemental Table S5). These results were also observed in TAD boundaries called in tissue types. One possibility of the lower Jaccard coefficient for Level 2 and Level 1 boundaries is that they may change the level assignment due to less reliable detection of smaller (Secondary and Tertiary) TADs. 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`, `rGMAP`, and `TopDom`. `SpectralTAD` showed comparable performance with `TopDom` and was faster than `rGMAP` at all resolutions (Supplemental 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 (Supplemental Figure S8A), e.g., with the entire genome taking 1 second to run for 25kb data when using four cores. We demonstrate that our method has a linear complexity $O(n)$ (Supplemental Methods) making it scalable for large Hi-C datasets. 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` 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 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 comparison 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 completely alleviate 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 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

Data Sources

Real-life Hi-C matrices from Gm12878 cell line ((Rao et al. 2014) at 50kb, 25kb, and 10kb resolution) and 35 different cell line and tissue samples ((Schmitt et al. 2016), 40kb resolution) were downloaded from Gene Expression Omnibus (GEO, Supplemental Table S4). 25 simulated matrices with manually annotated TADs ((Forcato et al. 2017), 40kb resolution) were downloaded from the `HiCToolsCompare` repository (Supplemental Table S4). 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 (Supplemental Table S2).

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. 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 (Supplemental 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 (Rousseeuw 1987):

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

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, further allowing to identify poorly organized regions (gaps, Supplemental Methods). 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 (Supplemental 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.

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 gold-standard TADs was tested using Jaccard statistics; the cross-resolution TAD comparison was assessed using a modified version of the Jaccard coefficient (Supplemental Figure S1). The effect of Hi-C data normalization was tested using the iterative correction and eigenvector decomposition (ICE) (Imakaev et al. 2012), Knight-Ruiz (KR) (Rao et al. 2014; Knight and Ruiz 2012), and the Sequential Component Normalization (SCN) (Cournac et al. 2012) methods. The association of TAD boundaries with genomic annotations, such as CTCF, was assessed using a permutation test. See Supplemental Methods for details.

Software availability

The software is published under the MIT license. The source code of `SpectralTAD` is available at https://github.com/dozmorovlab/SpectralTAD.

Acknowledgements

Author contributions: MGD and KGC conceived the project, KGC implemented `SpectralTAD` and wrote the analysis scripts. JCS helped with the data analysis. MGD and KGC wrote the manuscript with JCS contributions.

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.

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.

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.

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

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

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

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.

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.

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.

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

Gierman HJ, Indemans MHG, Koster J, Goetze S, Seppen J, Geerts D, Driel R van, Versteeg R. 2007. Domain-wide regulation of gene expression in the human genome. *Genome Res* **17**: 1286–95.

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.

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.

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

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*.

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.

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.

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.

Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, Vera DL, Wang Y, Hansen RS, Canfield TK, et al. 2014. Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**: 402–5.

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.

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.

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, Jung I, Xu Z, Qiu Y, Tan CL, Li Y, Lin S, Lin Y, Barr CL, et al. 2016. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep* **17**: 2042–2059.

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 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.

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.

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

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.

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**.

Supplemental Legends

Supplemental 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.

Supplemental 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 (Supplemental Table S4). All parameters were set according to the instructions of each package for analyzing 40kb resolution data.

Supplemental Figure S3. 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.

Supplemental Figure S4. 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.

Supplemental 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 (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.

Supplemental Figure S6. 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.

Supplemental Figure S7. Heatmap of jaccard similarities TADs were called on tissue (A) and cell line (B) samples using SpectralTAD. Jaccard similarity of TAD boundaries was calculated between samples and are displayed in a heatmap. Clustering was performed using Ward clustering applied to a Jaccard distance matrix. All TADs were called on 40kb data from (Schmitt et al. 2016).

Supplemental 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.

Supplemental 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).

Supplemental 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.

Supplemental 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.

Supplemental Table S2. Real Data sources. Genome annotation (hg19/GRCh37) (Rosenbloom et al. 2012) data for Gm12878 cell line used in the analysis, sorted by category, then by data type.

Supplemental 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.

Supplemental Table S4. Hi-C Data sources. Information about real (Rao et al. 2014; Schmitt et al. 2016) and simulated (Forcato et al. 2017) Hi-C data.

Supplemental 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 hierarchy.