Virginia Commonwealth University
VCU Scholars Compass

Theses and Dissertations | Graduate School

2019

Spectral methods for the detection and characterization of Topologically Associated Domains

Kellen Garrison Cresswell
Virginia Commonwealth University

Follow this and additional works at: https://scholarscompass.vcu.edu/etd

 Part of the Biostatistics Commons

© The Author

Downloaded from
https://scholarscompass.vcu.edu/etd/6100

This Dissertation is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.

©Kellen G. Cresswell 2019

All Rights Reserved

Spectral methods for the detection and characterization of Topologically Associated Domains

By

Kellen Cresswell

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Biostatistics

Virginia Commonwealth University

Advisor: Dr. Mikhail Dozmorov

Virginia Commonwealth University

Richmond, Virginia

October, 2019

Contents

1 Acknowledgments

I would like to thank my advisor Dr. Mikhail Dozmorov for his mentoring and input on my research. Our frequent meetings and his expertise in HiC data has made my graduate school experience a smooth one. I would also like to thank Dr. Nitai Mukhopadhyay, Dr. Shanshan Chen and Dr. Yongyun Shin for advising me on past SSTP projects. I would like to thank my committee members Dr. Joseph McClay, Dr. Ekaterina Smirnova and Dr. Silviu-Alin Bacanu. Finally, thanks to Spiro and John for your support and collaboration.

2 Abstract

The three-dimensional (3D) structure of the genome plays a crucial role in gene expression regulation. Chromatin conformation capture technologies (Hi-C) have revealed that the genome is organized in a hierarchy of topologically associated domains (TADs), sub-TADs, and chromatin loops which is relatively stable across cell-lines and even across species. These TADs dynamically reorganize during development of disease, and exhibit cell- and condition-specific differences. Identifying such hierarchical structures and how they change between conditions is a critical step in understanding genome regulation and disease development. Despite their importance, there are relatively few tools for identification of TADs and even fewer for identification of hierarchies. Additionally, there are no publicly available tools for comparison of TADs across datasets. These tools are necessary to conduct large-scale genome-wide analysis and comparison of 3D structure.

To address the challenge of TAD identification, we developed a novel sliding window-based spectral clustering framework that uses gaps between consecutive eigenvectors for TAD boundary identification. Our method, implemented in an R package, SpectralTAD, has automatic parameter selection, is robust to sequencing depth, resolution and sparsity of Hi-C data, and detects hierarchical, biologically relevant TADs. SpectralTAD outperforms four state-of-the-art TAD callers in simulated and experimental settings. We demonstrate that

TAD boundaries shared among multiple levels of the TAD hierarchy were more enriched in classical boundary marks and more conserved across cell lines and tissues. SpectralTAD is available at http://bioconductor.org/packages/SpectralTAD/.

To address the problem of TAD comparison, we developed TADCompare. TADCompare is based on a spectral clustering-derived measure called the eigenvector gap, which enables a loci-by-loci comparison of TAD boundary differences between datasets. Using this measure, we introduce methods for identifying differential and consensus TAD boundaries and tracking TAD boundary changes over time. We further propose a novel framework for the systematic classification of TAD boundary changes. Colocalization- and gene enrichment analysis of different types of TAD boundary changes revealed distinct biological functionality associated with them. TADCompare is available on https://github.com/dozmorovlab/TADCompare.

3 Chapter 1: Introduction

3.1 Motivation

The introduction of chromatin conformation capture technology and its high-throughput derivative Hi-C enabled researchers to accurately model chromatin interactions across the genome and uncover the non-random 3D structures formed by folded genomic DNA [1–5]. The structure and interactions of the DNA in 3D space inside the nucleus has been shown to shape cell type-specific gene expression [3,6–11], replication [12], DNA repair and chromosome translocation, orchestrate the assembly of antigen receptors [13], guide X chromosome inactivation [14,15], and regulate the expression of tumor suppressors and oncogenes [16,17].

Topologically Associated Domains (TADs) refer to a common structure uncovered by Hi-C technology, characterized by groups of genomic loci that have high levels of interaction within the group and minimal levels of interaction outside of the group [1,14,18–23]. TAD boundaries were found to be enriched in CTCF (considering the directionality of its binding) and other architectural proteins of cohesin and mediator complex (e.g., STAG2, SMC3, SMC1A, RAD21, MED12) [3,8,18,23–31], marks of transcriptionally active chromatin (e.g., DNAse hypersensitive sites, H3K4me3, H3K27ac, H3K36me3 histone modifications) [[19];[14]; [18]; [32];[33]; [34]; [35]; [16]; [21]; [36] and actively transcribed and housekeeping genes [18,36–39] 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 [18,19,21,23,32,38]

TADs organize themselves into hierarchical sets of domains [3,19,23,39–45]. These hierarchies are characterized by large "meta-TADs" that contain smaller sub-TADs and chromatin loops. To date, most methods were developed to find these single meta-TADs instead of focusing on the hierarchy of the TAD structures [35,46,47]. While interesting insights can

be gleaned from the meta-TADs, work has shown that smaller sub-TADs are specifically associated with gene regulation [38,44,48]. For example, it has been found that genes associated with limb malformation in rats are specifically controlled through interactions within sub-TADs [48]. 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 [18,23,32,39,49,50].

The first two aims of this work focus on developing a method for calling biologically relevant, hierarchical TADs. Currently the choice of hierarchical TAD callers is limited. Several previous methods have been designed to call hierarchical TADs. However, most algorithms require tunable parameters [34,35,51] 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 [49]). 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 [51,52], with some methods even lacking publicly available code [41]. Furthermore, the choice of TAD callers in R/Bioconductor ecosystem remains limited (For a complete overview of previous TAD callers see section 3.2.1).

The third and final aim involves methods for detecting changes in TAD boundaries across cell and tissue types. Many 3D structures are largely invariant between different cell types, and even conserved between mammalian species [3,8,12,14,18,53], indicating their high biological importance during genome evolution. Despite the high level of conservation, recent research uncovered the dynamic nature of the 3D genomic structures, and this plasticity accompanies various biological functions and phenomena [54]. In Drosophila (fruit flies), exposure to

heat-shock caused local changes in certain TAD boundaries resulting in TAD merging [55]. Another recent study showed that during motor neuron (MN) differentiation in mammals, TAD, and sub-TAD boundaries in the Hox cluster are not rigid and their plasticity is linked to changes in gene expression during differentiation [56]. The global organization of the 3D genomic structure is found in mitosis stage of cell cycle [57], fertilization [58,59], earliest stages of mammalian lineage development [21,60–62], and somatic cell reprogramming of pluripotent stem cells [63,64]. Fusion of TADs [9,14,26,65–68], creation or destruction of sub-TADs within existing TAD boundaries [16,17], and/or switching TAD states between active and inactive conformations [2,18] has been associated with a variety of phenotypes [69–71], ranging from limb malformation [17], congenital disorders [72], to cancer [17,30,65,69,73–78]. These observations highlight the importance of studying changes in TADs as a means to understand genomic regulation. However, methods for identifying changes in TAD boundaries remain underdeveloped.

Currently, there are only three methods that for differential TAD boundary detection (See Section 3.2.2 for a more in-depth discussion of previous methods). As Hi-C sequencing technologies improve, the number of replicates for a given experiment continue to rise, requiring methods for defining and comparing TAD boundaries across replicates of Hi-C data. Traditionally, two approaches have been developed to identify TAD boundaries across replicates. The first is to call TADs on individual replicates and aggregate them. The second approach involves combining all replicates into a consensus contact matrix and then calling TADs [3,79]. To date, only one method has been created that is specifically designed to identify TAD boundaries across multiple datasets [80]. Methods for comparing TAD boundaries between groups of Hi-C replicates remain undeveloped.

3.1.1 Brief overview of HiC data and terms

HiC data is stored in a **contact matrix** where entry ij is the number of terms region i of the genome contacts region j. The size of the regions are controlled by a parameter called

resolution. **Basepairs** (bp) are used to denote distance on the linear genome. Basepairs are commonly measured in increments of 1000 referred to as **kilobases** (kb) and 1 million referred to as **megabases** (mb). For instance, a resolution of 50kb means the genome is binned into units of 50000 basepairs. Basepairs can also be used as a measure of location on a chromosome. For instance, a TAD boundary located at 5mb is located 5 million basepairs from the beginning of the chromosome. An additional term that is commonly used in this manuscript is **TAD caller**. TAD callers are simply any tool that takes a contact matrix as an input and either returns the coordinates of TADs or the location of their boundaries.

3.2 Overview of current methods

3.2.1 Previous TAD callers

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 [18]. 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` [34], `HiCseg` [35], `TADLib` [81], `TopDom` [46], `Arrowhead` [47], `TADbit` [82] and `RHiCDB` [83]. Another intuitive metric, the insulation index [15], 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 [84], 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. [41] 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` [40] worked by creating TAD "forests" containing hierarchical "trees" of TADs. `TADtool` similarly provides hierarchical TAD detection and visualization [84]. Other hierarchical TAD finders include `HiTAD` [85] and `IC-Finder` [86] which take dynamic programming and probabilistic approaches, respectively. `ClusterTAD` [87] introduced a traditional hierarchical clustering-based approach to TAD classification. Another method, `rGMAP` [54] 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. `CaTCH` [88] is another approach that uses a novel measure called reciprocal insulation (RI) and iteratively partitions TADs into a hierarchy based on different thresholds of this value. More recently, a method called `OnTAD` was proposed [89]. 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 [46]. Of these methods, `TADtool`, `TADtree` and `TADLib` are Python-based. `IC-Finder` and `ClusterTAD` are MATLAB based with `ClusterTAD` also including Java implementation while `rGMAP` and `CaTCH` are available as R packages. Although some comparison of TAD detection tools has been performed [90–92], 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 [51,93,94]. `Arboretum-HiC` first introduced the idea of using Laplacian-graph segmentation to find structures in Hi-C data [80]. This method used a spectral clustering approach too simultaneously find 3D structures between multiple matrices. Chen et al. [52] 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` [95] and `HiTAD` [85] 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 [96]. 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.

R/Bioconductor is the de facto gold standard programming language for the genomics community [97]. Currently, the number of TAD callers implemented in the R programming language are limited and include `HiCseg` [35], `TopDom` [46], `rGMAP`[54], `CaTCH` [88] and `HiCDB` [83]. `HiCseg` is the only TAD-calling specific R package available on CRAN or Bioconductor, while `TopDom` is a downloadable R script. `HiCDB`, `rGMAP` and `CaTCH` 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 [98]. Additionally, the `HiTC` R package [99] 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.

3.2.2 Previous differential TAD detection methods

Traditionally, TADs have been compared by overlap-based metrics, such as Jaccard index and Venn diagrams [3,18]. These methods provide important information but suffer from inability to capture the dynamics of individual TAD boundaries, instead providing a global measure of TAD similarity. Additionally, they are heavily reliant on the method used to call TADs. This reliance exposes them to the large variation in TAD caller quality [49,91,100]. To date, no method exists that can identify and classify differential TAD based on their comparative structure.

DiffTAD [101] was the first publicly available method for direct comparison of TADs, at a TAD-by-TAD level. Their method works by taking two contact maps with pre-defined TADs, creating a differential contact map by subtracting the matrices from each other, subsetting the contact matrix based on TAD locations and either doing a parametric or non-parametric test on the differential contact map. The idea is that large values in the differential contact map correspond to differential regions. DiffTAD require users to specifically run the Armatus [34] TAD caller before analysis, or at the very least format their data to match the output of Armatus. Armatus is known to be sensitive to different parameters, resolution and sequencing depth [49], thus hindering unambigious TAD detection. Consequently, DiffTAD performance and user experience may be unsatisfactory.

Sauerwald et al. developed a method for quantifying the similarity of two sets of TADs using variation of information (VI) metric [102]. VI is a general information theoretic metric for computing distance between two clusterings; in terms of Hi-C data, a cluster is a set of Hi-C bins placed in the same TAD. This approach is significant in that it provided the first continuous measure of TAD similarity at the boundary level. In their method, VI is calculated at TAD boundaries and a permutation test is used to determine cutoffs for differential regions. This method, referred to as TADsim was followed by an updated version called localTADsim [103]. The new method is designed to be fast and introduces the concept of "hanging TADs". Hanging TADs refer to TAD boundaries which start within other TADs. `localTADSim` ignores

these TADs despite previous research highlighting the biological importance of such TADs [40–42,45,104]. Like DiffTAD, `localTADsim` requires users to use Armatus or to manually format their inputs to look like Armatus output. Additionally, this method has been shown to be exceptionally slow with speed depending on the number of TADs analyzed. Examples show run times of around an hour for certain chromsomes at 100kb resolution [102], creating problems with reproducibility and user experience like for DiffTAD.

The latest method to be introduced, HiCDB [83], uses a metric called relative local insulation (RI). This metric is similar to insulation score [15] but includes terms for correcting for background noise. The method works by detecting TADs between two contact matrices and then calculating the difference in relative local insulation (RI). The differential TAD boundaries are detected as any boundary where the difference in RI values is above the 90% quantile of RI differences. However, this approach is flawed in that it artificially forces 10% of TAD boundary pairs to be detected differential irrespectively of the data properties.

Another criticial issue with these tools are a lack of upkeep. As the resolution (and the corresponding size) of Hi-C data continue to increase, users are often interested in comparing data in individual chromosomes. HiCDB forces users to provide information for all chromosomes simultaneously, which, combined with slow runtime, makes it unsuitable for the analysis of modern Hi-C datasets. DiffTAD has been in a short pre-print form since 2016 and has not been updated since January 2017. In general, the previous methods are slow, require complex data inputs and output results that are difficult to interpret. Thus, a fast, flexible, and user-friendly R package for the detection of differential TADs is needed.

The uniqueness of `TADCompare` in terms of differential detection comes from the complete integration of TAD calling with the differential detection itself. Of the methods listed, only HiCDB is integrated with the TAD caller itself, using its relative insulation score to find differences. By removing the requirement of outside TAD calling, we remove an extra source of noise caused by inconsistent TAD callers and provide a completely data-driven approach.

Another advantage of `TADcompare` is its ability to quantify TAD boundary strength,

enabling statistical comparison of them. This is in contrast to methods like `localTADsim`, which does not use data from the contact matrix at all, thus lacking the ability to use statistical methods or quantify the degree of similarity at a given TAD boundary. This limitation requires complete reliance on the chosen TAD caller to account for statistical properties of contact matrices. `diffTAD` differs from our method in that it directly compares contact frequencies within regions bounded by TADs instead of directly analyzing the structure of TADs or their boundaries. In their own words, their approach analyses "differential contact frequency in topological domains" while TADCompare looks directly at differences in boundaries. Compared to other methods,`TADCompare` provides a unique middle-ground, being a data-driven method that has the ability to recognize TAD boundaries without the need for outside TAD callers.

3.3 Aims

We develop a set of methods for detecting and analyzing topologically associatied domains (TADs). First, we develop a spectral clustering-based approach for detection of TADs. We then extend these methods to detect hierarchical TADs, implementing them in a publicly available package (SpectralTAD). Finally, we create a package (TADCompare) for analysis of differential, time-varying and consensus TADs across datasets. For each method, we perform validation using real biological and simulated data. In the end, we provide a set of tools for end-to-end analysis of HiC data. At each step, we use the tools to provide novel insights into TADs and their hierarchy.

3.3.1 Aim 1: Develop and benchmark a method for detecting Topologically Associated Domains using Spectral Clustering

We will develop an approach to detect TADs using a novel, parameterless, spectral clustering approach. This method will take advantage of the natural graph-like nature of the genome and provide an efficient method for identifying and comparing TAD boundaries.

The method will be benchmarked against existing methods based on speed, consistency, robustness to sequencing depth and sparsity, and biologically relevant enrichments.

3.3.2 Aim 2: Extend the method to find hierarchical TADs and benchmark against hierarchical TAD callers and implement in an R package

We will use an iterative spectral clustering based approach to partition TADs into smaller units called sub-TADs. The method will be benchmarked at each level using the same criteria as the single-level algorithm along with hierarchy-based measures such as the consistency of sub-TADs across levels. The method will be implemented in an easy-to-use R package.

3.3.3 Aim 3: Develop a method and R package for finding differential TADs between experiements

Using the same framework as TAD detection we aim to extend our method to be able to find differences in TADs between datasets. This method will take advantage of the natural structure of the eigenvectors of the graph spectrum. The method will be applicable to any situation where we need to compare TADs across multiple replicates or contact matrices (differential analysis, time-course and consensus TAD calling). The method will be benchmarked for robustness to noise, sequencing depth and sparsity. The approach will be validating using publicly available datasets, confirming the biological relevance of the features it finds. The method has been released in an R package called `TADCompare` (https://github.com/dozmorovlab/TADCompare).

4 Chapter 2: Aim 1 - Develop and benchmark a method for detecting Topologically Associated Domains using Spectral Clustering

4.1 Introduction

The introduction of chromatin conformation capture technology and its high-throughput derivative Hi-C enabled researchers to accurately model chromatin interactions across the genome and uncover the non-random 3D structures formed by folded genomic DNA [1–3]. The structure and interactions of the DNA in 3D space inside the nucleus has been shown to shape cell type-specific gene expression [3,6], replication [12], guide X chromosome inactivation [14], and regulate the expression of tumor suppressors and oncogenes [16,17].

Topologically Associated Domains (TADs) refer to a common structure uncovered by Hi-C technology, characterized by groups of genomic loci that have high levels of interaction within the group and minimal levels of interaction outside of the group [1,14,18,19]. TAD boundaries were found to be enriched in CTCF (considering the directionality of its binding) and other architectural proteins of cohesin and mediator complex (e.g., STAG2, SMC3, SMC1A, RAD21, MED12) [3,18,24,25], marks of transcriptionally active chromatin (e.g., DNAse hypersensitive sites, H3K4me3, H3K27ac, H3K36me3 histone modifications) [14,18,19,32]. 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 [32,38].

Our goal was to develop a simple data-driven method to detect TADs and uncover hierarchical sub-structures (See Aim 2 for complete explanation of hierarchy and hierarchical benchmarking) within these TADs. We propose a novel method that exploits the graph-like structure of the chromatin contact matrix and extend it to find the full hierarchy of sub-TADs,

limited only by the resolution of Hi-C data. Our approach employs a modified version of the multiclass spectral clustering algorithm [105] and uses a sliding window based on the commonly used 2 megabase biologically maximum TAD size [18,106]. We introduce a novel method for automatically choosing the number of clusters (TADs) based on maximizing the average silhouette score [107]. We show that this approach finds TAD boundaries with more significant enrichment of known boundary marks than those called by other TAD callers. We then extend the method to find hierarchies of TADs and demonstrate their biological relevance. Our method provides a parameterless approach, efficiently operating on matrices in text format with consistent results regardless of the level of noise, sparsity, and resolution of Hi-C data. The method is fast and scales linearly with the increasing amount of data. Our method is implemented in the `SpectralTAD` R package, freely available on GitHub (https://github.com/dozmorovlab/SpectralTAD) and Bioconductor (http://bioconductor.org/packages/SpectralTAD/).

4.2 Methods

4.2.1 Data Sources

Experimental Hi-C matrices from the GM12878 cell line ([3] at 50kb, 25kb, and 10kb) and 35 different cell line and tissue samples ([108], 40kb resolution) were downloaded from Gene Expression Omnibus (GEO, Supplementary Table S1). The GM12878 data is the "primary+replicate" data from [3] and was generated by combining results from two experiments by counting total contacts and binning the genome, using increasing sizes, until 80% of bins contained over 1000 contacts. 25 simulated matrices with manually annotated TADs ([91], 40kb resolution) were downloaded from the `HiCToolsCompare` repository (Supplementary Table S1). Data for chromatin states, histone modification and transcription factor binding sites (TFBS) were downloaded from the UCSC genome browser database [109]. Given the fact that some transcription factors have been profiled by different institutions (e.g., CTCF-Broad, CTCF-Uw, and CTCF-Uta), we selected annotations most frequently enriched

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

4.2.2 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.

4.2.3 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 [18,23]. 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 [18,96,110,111].

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 [18]. 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.

4.2.4 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 [112]; but since our method builds upon the multiclass spectral clustering algorithm [105], 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.

4.2.5 Projection onto the unit circle

Our method builds on the approach to spectral clustering first introduced in [105], 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 S1). 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.

4.2.6 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 [107]:

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

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, allowing us to identify poorly organized regions (gaps, Supplementary Methods). Cluster (TAD) boundaries are mapped to genomic coordinates based on their location in the contact matrix. If a TAD is detected and found to be less than 5 bins wide it is ignored due to previous evidence

suggesting these are not biologically relevant [18,96,110,111]. 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.

4.2.7 Identification and removal of gaps

In addition to regions of the genome belong to TADs, there is an additional type of region called a gap. Gaps refer to an area where there are no TADs present either due to a lack of sequencing depth, a centromere, or simply a lack of organization (Supplementary Methods). The percentage of non-centromeric gaps varies across chromosomes and resolutions (Supplementary Table S3), being 19.9% 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.

For practical purposes, we consider two types of gaps (Centromeric and unsequenced). The first category of gaps is removed by simply getting rid of loci with more than 95% zero contacts. This works because centromeres show up as columns in the contact matrix which are entirely made up of zeros. 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.

4.2.8 Simulating levels of noise, sparsity, and sequencing depth

Simulated contact matrices [91] (Supplementary Table S1) 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 [113]. 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.

4.2.9 Normalization of Hi-C data

Normalization of Hi-C data matrices is a common step in Hi-C data analysis [2,114–119]. To test for the effect of normalization on the detection of TADs, we applied iterative correction and eigenvector decomposition (ICE) [114], Knight-Ruiz (KR) [3,115] normalization, and the Square Root Vanilla Coverage (sqrtVC) [3] to the simulated and experimental Hi-C matrices. ICE was implemented using the `ICE` function from the `dryHiC` R package version 0.0.0.9000. KR-normalization [120] was performed using Juicer [47], and sqrtVC normalization was performed using a function written by the authors.

4.2.10 Measuring association of TAD boundaries with genomic annotations

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 of overlap of TAD boundaries with genomic annotations. Briefly, the difference in the mean number of genomic annotations within 50kb of each boundary and that of for all other regions was calculated (observed enrichment). 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 (expected enrichment). This procedure was repeated 10000 times. We determined the permutation p-value by taking the number of 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. For all hierarchical TAD callers, only the first level of TAD boundaries was used.

4.2.11 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 S2). 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.

4.2.12 Modified Jaccard with a flank

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.

4.3 Results

4.3.1 An overview of the `SpectralTAD` algorithm

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

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

Sequence- and technology-driven biases may be present in Hi-C matrices [114,116,121,122]. Consequently, numerous normalization methods have been developed [2,114–119]. However, their effect on the quality of TAD detection has not been explored.

We investigated the effect of three normalization methods, Knight-Ruiz (KR), iterative correction and eigenvector decomposition (ICE), and square root vanilla coverage (sqrtVC) on TAD detection using `SpectralTAD`. Using simulated matrices with the ground-truth TADs, we found that all normalization methods marginally degraded the performance of `SpectralTAD` under different levels of noise, sparsity, and downsampling (Figure 1A-C). These results suggest that the use of raw Hi-C data is appropriate for the detection of TAD boundaries.

Using the experimental Hi-C data from GM12878 cell line, we found that ICE normalization only marginally affected the average number and width of TADs, and these results were consistent across resolutions (Supplementary Figure S3A). In contrast, KR, and sqrtVC normalization resulted in a larger variability in TAD widths across chromosomes and between

resolutions (Supplementary Figure S3B). We also assessed the average number and the enrichment (permutation test) of genomic annotations at TAD boundaries detected from unnormalized, KR-, ICE-, and sqrtVC-normalized data. The average number of genomic annotations was not significantly different in TAD boundaries detected from raw and ICE-normalized data as compared with those from KR- and sqrtVC-normalized, where the number of annotations was significantly less (Figure 1D). We found CTCF, RAD21, "Insulator," and "Heterochromatin" states to be significantly enriched in TAD boundaries, and this enrichment was frequently more significant in TAD boundaries detected from the ICE-normalized data (Figure 1F). Similarly, "enhancer"-like chromatin states were significantly depleted at TAD boundaries, and this depletion was more pronounced in boundaries detected from raw data (Figure 1G). The enrichment results were consistent across resolutions (Supplementary Figure S3C-F). These results suggest that both ICE-normalized and raw Hi-C data are suitable for the robust detection of biologically relevant TADs. Based on these results, and the fact that previous studies showed graph-based TAD identification methods work well un-normalized HiC data [80,123], consequent results are presented with the use of raw Hi-C data.

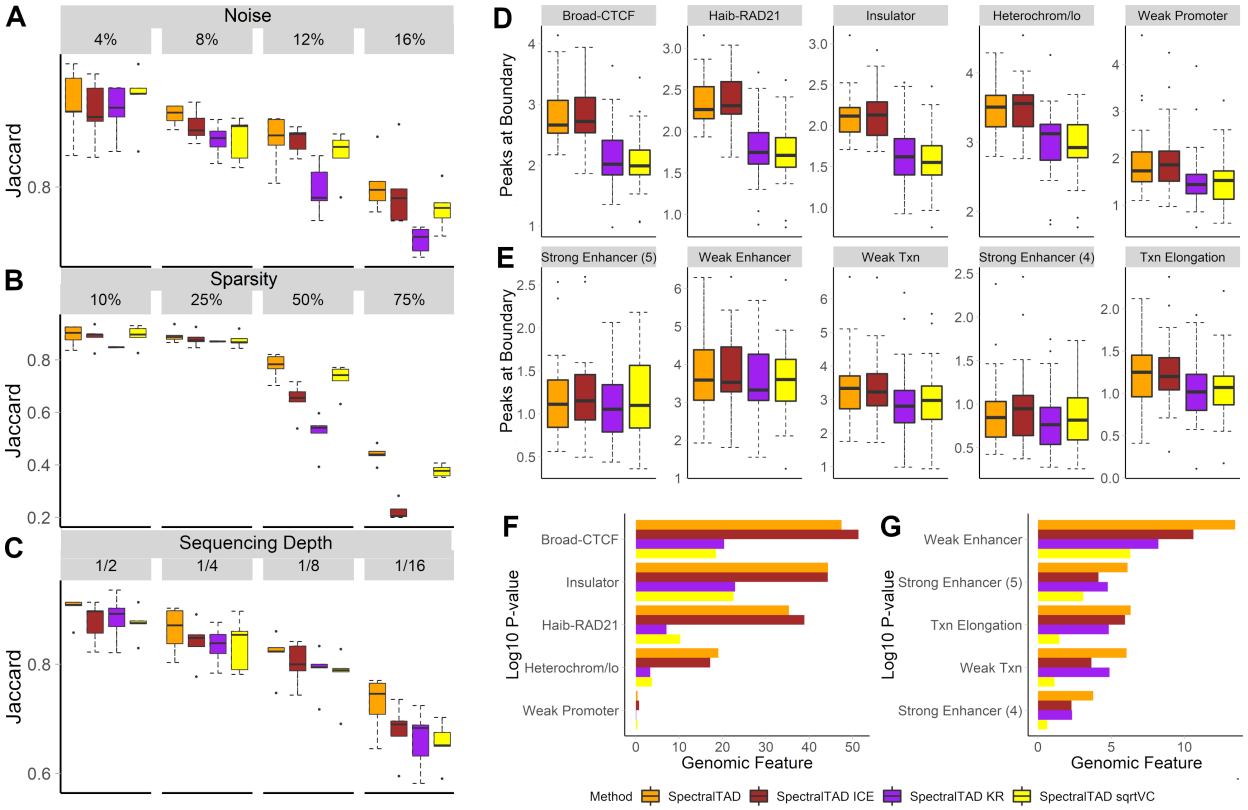


Figure 1. The effect of normalization on TAD consistency and enrichment. To test for robustness to noise, sparsity, and sequencing depth, simulated Hi-C matrices were used as-is, KR-, ICE- and square root VC-normalized. TADs detected by SpectralTAD were compared with the ground-truth TADs using the Jaccard similarity metric. The effect of normalization was assessed at different levels of noise (A, the percentage of the original contact matrix modified by adding a constant), sparsity (B, the percentage of the original contact matrix replaced with zero), and downsampling (C, the fraction of contacts kept, see Methods). Using the raw and normalized data from 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 for enriched (D) and depleted (E) genomic features and the permutation p-values corresponding to enrichment (F), and depletion (G) for the top five most enriched/depleted genomic annotations are shown. Results averaged across chromosome 1-22 are shown.

4.3.3 SpectralTAD identifies more consistent TADs than other methods

Using simulated matrices, we compared the performance of `SpectralTAD` with `rGMAP`, `TopDom`, `OnTAD` and `HiCSeg` at different noise levels. We found that both `SpectralTAD` and `TopDom` had a significantly higher agreement with the ground truth TADs than `rGMAP` across the range of noise levels (Figure 2A). To better understand the poor performance of `rGMAP`,

we hypothesized that inconsistencies might arise due to the "off-by-one" errors that occur when, by chance, a TAD boundary may be detected adjacent to the true boundary location. We analyzed the same data using TAD boundaries flanked by 50kb regions. Expectedly, the performance of all TAD callers, including `rGMAP`, increased; yet, the performance of `rGMAP` remained significantly low (Supplementary Figure S4A). At low level of noise, `HiCseg` detected highly consistent TAD boundaries; however, these TAD boundaries were the least biologically relevant, detailed below. In summary, these results suggest that, with the presence of high noise levels, a situation frequent in experimental Hi-C data, `SpectralTAD` performs better than other TAD callers in detecting true TAD boundaries.

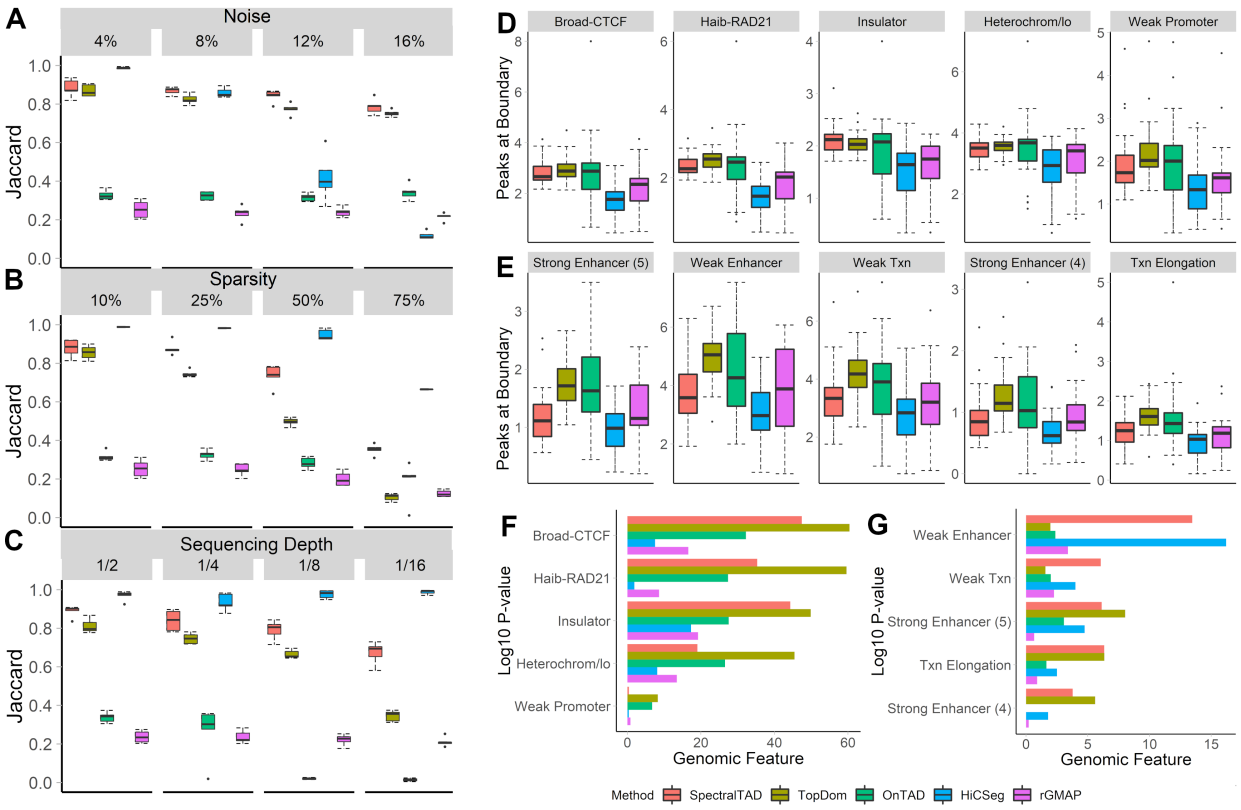


Figure 2. The comparison of SpectralTAD and other TAD callers regarding TAD consistency and biological significance. To test for robustness to noise, sparsity, and sequencing depth, TADs were called from simulated Hi-C matrices using SpectralTAD and four other TAD callers. They were compared with the ground-truth TADs using the Jaccard similarity metric. The performance was assessed at different levels of noise (A), sparsity (B), and downsampling (C, see Methods). Using the raw data from GM12878 at 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 for enriched (D) and depleted (E) genomic features and the permutation p-values corresponding to enrichment (F), and depletion (G) for the top five most enriched/depleted genomic annotations are shown. Results averaged across chromosome 1-22 are shown.

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 2B). `SpectralTAD` outperformed all TAD callers except `HiCseg` at all sparsity levels. We further tested whether accounting for the "off-by-one" error improves the performance; the performance of `SpectralTAD` remained superior (Supplementary Figure S4B). These results demonstrate the robustness of `SpectralTAD` to sparsity.

TAD callers should be robust to changes in sequencing depth. We introduced four levels

of downsampling into simulated matrices and compared the detected TADs with the ground truth TADs. Downsampling involves removing contacts at random, simulating sequencing depth. Expectedly, the average Jaccard similarity degraded for all TAD callers with the increased level of downsampling (Figure 2C). Notably, the performance of `SpectralTAD` was consistently higher than that of for other TAD callers except `HiCseg`. Similar observations were true when accounting for the "off-by-one" error (Supplementary Figure S4C). Despite seemingly good performance of `HiCseg`, the biological relevance of TAD boundaries it detects is low, as shown below (Figure 2). 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 better than other TAD callers.

4.3.4 SpectralTAD outperforms other TAD callers in finding biologically relevant TAD boundaries

To evaluate the biological relevance of TAD boundaries detected by `SpectralTAD` and the other TAD callers, we evaluated their enrichment in genomic annotations known to be associated with TAD boundaries. We found that the TAD boundaries called by `SpectralTAD`, `TopDom`, and `OnTAD` had a significantly higher number of CTCF and RAD21, "Insulator," and "Heterochromatin" annotations than those called by `HiCseg` and `rGMAP` (Figure 2D). Consequently, these marks were more enriched at TAD boundaries detected by `SpectralTAD` and `TopDom` as compared with the other TAD callers (Figure 2F). In terms of depleted genomic annotations, "enhancer"-like chromatin states were underrepresented at TAD boundaries, and this depletion was highly significant for boundaries detected by `SpectralTAD` (Figure 2E, G). Notably, the TAD boundaries detected by `HiCseg` had the lowest number of these genomic annotations. They also exhibited the lowest level of enrichment and depletion (Figure 2E, G). These results suggest that, despite robustness to noise, sparsity, and sequencing depth, `HiCseq` detects boundaries that are less biologically relevant in terms of known TAD biology. The performance of `SpectralTAD` and other callers was consistent at different resolutions

(Supplementary Figure S4D-G, Supplementary Table S4). In summary, these results suggest that `SpectralTAD` outperforms other TAD callers in detecting biologically relevant TAD boundaries.

4.3.5 SpectralTAD identifies consistent TADs across resolutions of Hi-C data

If TAD boundaries called at different resolutions of Hi-C data are inconsistent, one risks receiving vastly different results despite the data being the same. Using the GM12878 Hi-C data at 10kb, 25kb, and 50kb resolutions, we estimated the average number and width of TADs called by `SpectralTAD`, `TopDom`, `HiCseg`, `OnTAD`, and `rGMAP`. As resolution of Hi-C data increased, the average number of TADs decreased for all but `SpectralTAD` (Supplementary Figure S5A). Similarly, the average width of TADs increased for all but `SpectralTAD` TAD callers (Supplementary Figure S5B). We further compared the consistency of TADs detected in 50kb vs. 25kb, 50kb vs. 10kb, 25kb vs. 10kb resolution comparisons. We found that, for nearly all comparisons, `SpectralTAD` and `HiCseg` had significantly higher consistency quantified by modified Jaccard statistics than the other TAD callers (Supplementary Figure S5C). These results show that `SpectralTAD` identifies consistent TADs at different resolutions of Hi-C data.

4.3.6 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 [3]. We compared TAD boundaries called by `SpectralTAD` across various tissues and cell types (Supplementary Table S1, [108]). The Jaccard for all TADs, ignoring hierarchy, between cell-line samples ranged from 0.33 to 0.73 with a mean of 0.45 (SD = 0.08). The Jaccard between tissues ranged from 0.21 to 0.38, with a mean of 0.27 (SD = 0.03), significantly lower than that of cell lines (Wilcoxon p-value < 0.0001). The lower conservation of TADs called from tissue samples is expected as cell lines come from a "pure" single source while tissues

are a mixture of different cells. These results were summarized in heatmaps, comparing different cell lines (Supplementary Figure S6A) and tissues (Supplementary Figure S6B) according to Jaccard similarity. Hierarchical clustering of cell type-specific samples by the Jaccard similarity of their TADs, ignoring hierarchy, identified the expected associations between cell-type-specific data, with replicates clustering together and cell types being distinct (Supplementary Figure S6A). To a lesser extent, these results were similar in tissue-specific samples (Supplementary Figure S6B). In summary, these results show conservation of TAD boundaries called by SpectralTAD across tissues and cell lines similar to previously reported results.

5 Chapter 3: Aim 2 - Extend the method to find hierarchical TADs, benchmark against hierarchical TAD callers and implement in an R package

5.1 Introduction

TADs organize themselves into hierarchical sets of domains [40–42]. These hierarchies are characterized by large "meta-TADs" that contain smaller sub-TADs and chromatin loops. To date, most methods were developed to find these single meta-TADs instead of focusing on the hierarchy of the TAD structures [35,46,47]. While interesting insights can be gleaned from the meta-TADs, work has shown that smaller sub-TADs are specifically associated with gene regulation [38,44,48]. For example, it has been found that genes associated with limb malformation in rats are specifically controlled through interactions within sub-TADs [48]. These results highlight the importance of identifying the full hierarchy of TADs.

Several methods have been designed to call hierarchical TADs (Supplementary Material). However, most algorithms require tunable parameters [34,35,51] 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 [49]). 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 [51,52], with some methods even lacking publicly available code [41]. Furthermore, the choice of TAD callers in R/Bioconductor ecosystem remains limited (Supplementary Material). Aim 2 introduces the hierarchy and demonstrates how we incorporate hierarchical TAD identification into the SpectralTAD framework as explained in Aim 1.

5.2 Methods

5.2.1 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. The Z-scores are referred to as boundary scores. This is done following our empirical observation of the log-normality of eigenvector gaps (Supplementary Figure S7). We then label any distance with a Z-score greater than 2 as a sub-boundary. If significant sub-boundaries are detected, we partition the TAD with each sub-boundary indicating the end of a given sub-TAD. This procedure is then repeated for each sub-TAD until either the TAD is too small to be partitioned into two sub-TADs or no significant boundaries are found. The TADs detected during the initial run of the algorithm are considered primary TADs, and the TADs detected after partitioning are considered secondary, tertiary, etc., sub-TADs. In practice, this approach can also be used for the first iteration of the algorithm (non sub-TADs) and is an option in the `SpectralTAD` R package.

5.2.2 Log-normality of eigenvector gaps allows for calculation of boundary score

Our method relies on the assumption that the distance between eigenvectors is lognormally distributed. We tested this assumption by collecting all distances across all chromosomes and 10kb, 25kb and 50kb. We then fit six potential distributions (gamma, Weibull, lognormal, normal, logistic, Cauchy) with maximum likelihood estimation using `fitdistr` function from the `MASS` R package (Version 7.3-51.4). This process was performed on [3] data. The fit of each distribution was compared using the log-likelihood. Out of the 131 real datasets, 67 were fit best by a lognormal distribution and 64 were fit best by a Weibull. Empirically, we find that in general, the log-likelihoods of these values are similar. Due to the interchangeability of the methods, we choose to model these values as lognormal as this gives us the ability to calculate boundary scores and make statistical inferences on boundary strength.

To allow for direct comparison of eigenvector gaps between contact matrices it is important that their distribution is relatively consistent. If we have large variation in mean or variance between contact matrices' eigenvector gaps then the magnitude of boundary scores can be inflated or deflated making differential detection inaccurate. We assess this by plotting the distributions separated by resolution (Supplementary Figure S7). It is clear that within resolutions, there is little variation in distributions. However, as resolution increases we have many more smaller eigenvector gaps. Between resolutions, the distribution is consistent but as the resolution becomes finer the proportion of non-TADs increase and increase the number of boundary scores around zero. Based on these results, we see that boundary scores behave as expected and can be used as intended.

5.2.3 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 3A). 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 3A). 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. [89], 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 3B).

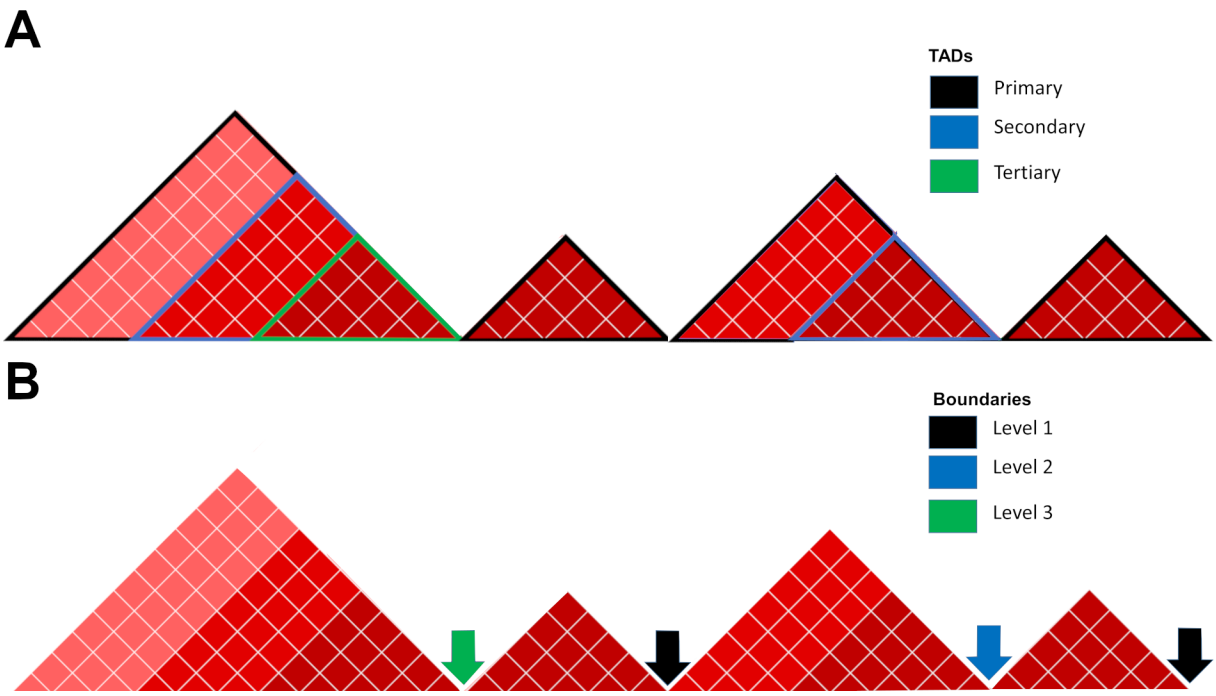


Figure 3.Hierarchy of TADs and boundaries. A) TADs not enclosed within other TADs are defined as primary, while TADs contained within other TADs are defined as secondary, tertiary, etc. B) Boundaries are defined as the rightmost point of a given TAD. Boundaries belonging to a single TADs are defined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

5.3 Results

5.3.1 The hierarchical structure of TADs is associated with biological relevance

Having established the strong performance of `SpectralTAD`, we investigated the biological importance of the hierarchy of TAD boundaries detected by it. We tested the relationship between the number of times a TAD boundary occurs in a hierarchy (Figure 4B) and enrichment of genomic annotations. We hypothesized that TAD boundaries shared by two or more TADs (Level 2 and 3 boundaries) would be more biologically important, hence, harbor a larger number of key markers such as CTCF and RAD21. We found that this is indeed the case, as illustrated by a significant increase in the average number of CTCF and RAD21 annotations, and "Insulator," and "Heterochromatin" states around Level 2 and 3 boundaries as compared with Level 1 boundaries (Figure 4A). A similar trend was observed in the stronger enrichment of Level 2 and 3 TAD boundaries in those annotations (Figure 4C). TAD boundaries at all levels of the hierarchy were similarly depleted in the "enhancer"-like annotations (Figure 4B, D), although these depletions were more significant for the Level 3 TAD boundaries. These observations were consistent across resolutions (Supplementary Figure S8). 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 [54,89] and confirm that `SpectralTAD` identifies a biologically relevant hierarchy of TAD boundaries.

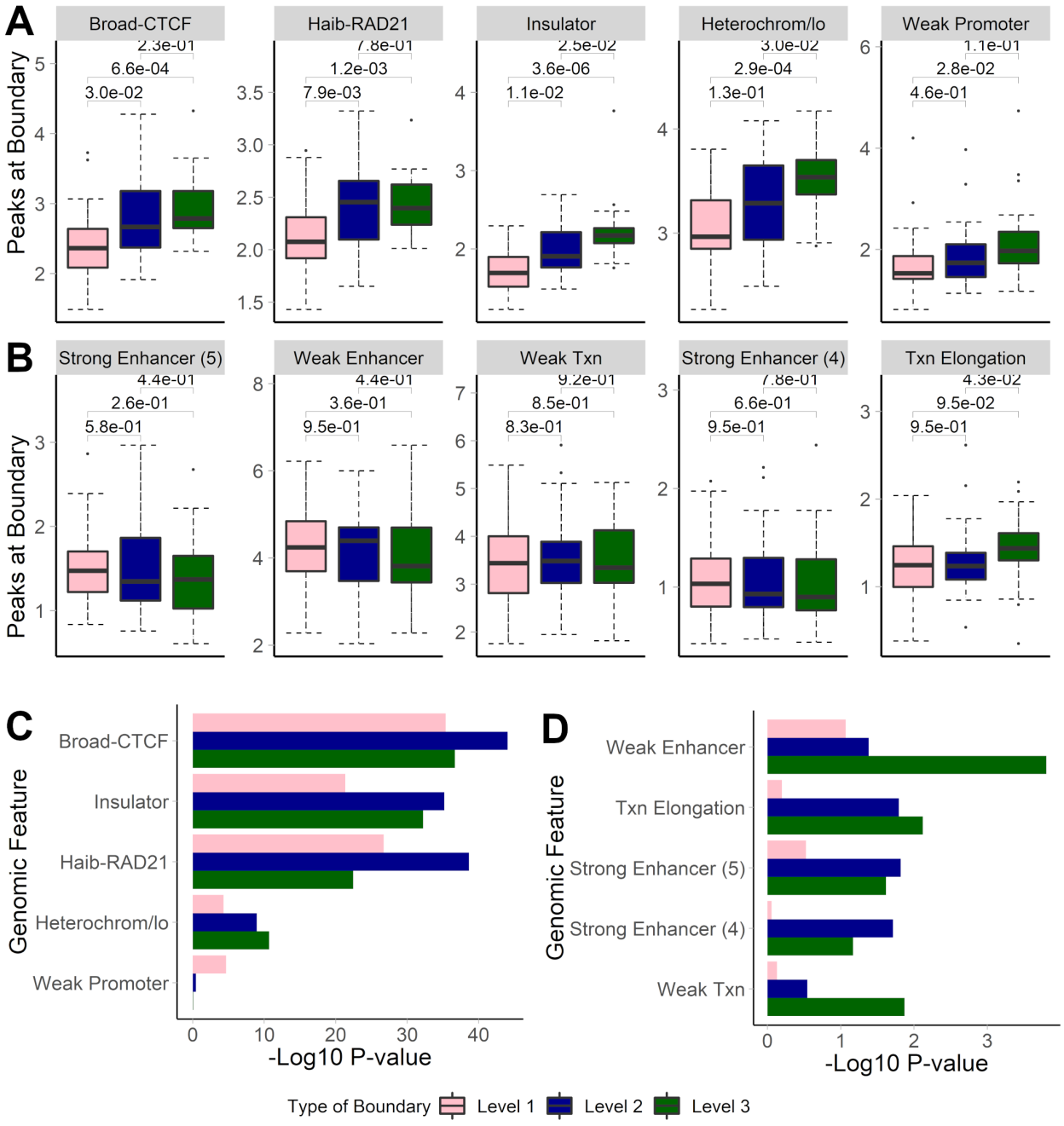


Figure 4. The effect of the hierarchy of TAD boundaries detected by SpectralTAD on the average number of annotations in enriched (A) & depleted (B) genomic markers and on enrichment (C) and depletion (D) for different genomic annotations. Results for TAD boundaries detected as level 1, 2, and 3 boundaries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

5.3.2 Hierarchy of TADs and boundaries affect conservation of TADs

Following our definition of TADs (Primary, Secondary, and Tertiary, Figure 3A), we hypothesized that primary TADs would be better conserved than Secondary or Tertiary TADs. The primary TADs are detected during the first pass of the algorithm; hence, they are robustly supported by the underlying data and expected to reproduce across different datasets. Indeed, the average Jaccard for Primary, Secondary, and Tertiary TADs across cell types was 0.42, 0.40, and 0.35, respectively (Supplementary Table S5), and this decrease was significant (Wilcoxon p-value < 0.0001). These observations were consistent when analyzing TADs called from tissue samples, although the average Jaccard coefficients for Primary, Secondary, and Tertiary TADs were significantly lower (Supplementary Table S5). These results demonstrate that Primary TADs are the most conserved across cell types and tissues.

We hypothesized that Level 3 TAD boundaries (Figure 3B, 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, Supplementary Table S5). These results were also observed in TAD boundaries called in tissue types. One possibility of the lower Jaccard coefficient for Level 1 and 2 boundaries is that they may change their assignment due to higher probability of detection of sub-TADs in different datasets. In summary, these results demonstrate that boundaries shared by several TADs have high biological significance and are better conserved across cell types and tissues.

5.3.3 SpectralTAD is the fastest TAD caller for high-resolution data

We evaluated the runtime performance of `SpectralTAD`, `TopDom`, `OnTAD`, `rGMAP` and `HiCSeg`. `SpectralTAD` showed comparable performance with `TopDom` and was faster than `rGMAP` at all resolutions (Supplementary Figure S9A). 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. `OnTAD` takes ~4 minutes to run on 25kb data and ~30 minutes on 10kb data. `rGMAP` takes ~12 minutes on 25kb data and ~47 minutes on 10kb data. We find that HiCSeg is prohibitively slow, taking ~609 minutes on 25kb data and multiple days to run on 10kb data with chromosome 1 taking over 24 hours alone. Importantly, our method scales nearly linearly with the size of the data (see Methods), making it amenable for fast processing of data at higher resolutions. Furthermore, when parallelized, `SpectralTAD` is several orders of magnitude faster than other TAD callers (Supplementary Figure S9A), 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 $O(n)$ (Supplementary Methods), making it scalable for large Hi-C datasets. In summary, these results demonstrate that `SpectralTAD` is significantly faster than `TopDom`, `rGMAP` and `HiCSeg`, providing near-instant results when running on multiple cores.

5.4 Discussion

In aims 1 and 2, we introduce the `SpectralTAD` R package implementing a spectral clustering-based approach that allows for fast TAD calling and scales well to high-dimensional data. The method was benchmarked against four TAD callers - `TopDom` and `HiCseg` that detect single-level TADs, and `OnTAD` and `rGMAP` that detect hierarchical TADs. We show better performance of `SpectralTAD` vs. the other TAD callers in nearly all conditions. We also demonstrate that `SpectralTAD` is more robust to sparsity, sequencing depth, and resolution. We show that `SpectralTAD` can robustly detect hierarchical TAD boundaries. Furthermore, we demonstrate different levels of TAD hierarchy to be differentially associated with known marks of TAD boundaries, highlighting their distinct biological roles and the importance of TAD hierarchy in general. The clear superiority of `SpectralTAD` regarding running speed and robustness to data irregularities suggests its use as the new gold-standard of hierarchical TAD callers in the R ecosystem.

The performance of SpectralTAD was frequently better, but not always superior to that of HiCseg. The better performance of HiCseg under different levels of noise, sparsity, and sequencing depth in some cases may be explained by the fact that HiCseg identifies non-hierarchical TAD boundaries at once, thus detecting the maximum number of them in one run. SpectralTAD, on the other hand, defines a hierarchy of primary, secondary, etc., TADs, restricted to the first three levels in the current analysis. Thus, TADs at a deeper level may have been detected by HiCseg inflating its performance. However, TAD boundaries identified by HiCseg were significantly less associated with known marks of TAD boundaries, undermining their biological relevance. We suggest that the tradeoff between robustness and biological relevance of TAD boundaries should be made for the latter, with SpectralTAD providing the optimal balance.

One overarching limitation with non-hierarchical TAD callers like TopDom and HiCSeg is their inability to capture all TADs in a dataset. While methods like TopDom may find biologically relevant TADs, they cannot account for the common situation where sub-TADs occur within a TAD. In the case of TADs enclosing sub-TADs, non-hierarchical callers are forced to make a choice that is often far from optimal (Supplementary Figure S10). We suggest that even when the hierarchy of TADs is not essential, hierarchical TAD callers like SpectralTAD should be used for maximally accurate reconstruction of TADs at first level of the hierarchy.

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

6 Chapter 4: Aim 3 - Develop a method and R package for finding differential TADs between experiements

6.1 Introduction

Recent research indisputably proves importance of the 3D genome organization in regulating gene expression and other genomic processes [124–137]. The 3D genomic structures consists of chromosome territories [138], A/B compartments corresponding to active/repressed chromatin [2,3], topologically associated domains (TADs) [1,14,18–22], smaller sub-TADs [3,6] and chromatin loops [3,7,9,139]. These structures help to regulate global gene expression [124–137]. Consequently, coordinated changes in the 3D structures [121,127,140] determine cell type-specific gene expression and identity [3,6–10,50,136], guide recombination [13], X chromosome inactivation [14,15]. Many 3D structures are largely invariant between different cell types, and even conserved between mammalian species [3,8,12,14,18,53], indicating their high biological importance during genome evolution.

Despite the high level of conservation, recent research uncovered the dynamic nature of the 3D genomic structures, and this plasticity accompanies various biological functions and phenomena [54]. In Drosophila, exposure to heat-shock caused local changes in certain TAD boundaries resulting in TAD merging [55]. Another recent study showed that during motor neuron (MN) differentiation in mammals, TAD, and sub-TAD boundaries in the Hox cluster are not rigid and their plasticity is linked to changes in gene expression during differentiation [56]. The global organization of the 3D genomic structure is found in mitosis [57], earliest stages of mammalian lineage development [21,60–62], and somatic cell reprogramming of pluripotent stem cells [63,64]. Fusion of TADs [9,14,26,65–68], creation or destruction of sub-TADs within existing TAD boundaries [16,17], and/or switching TAD states between active and inactive conformations [2,18] has been associated with a variety of phenotypes [69–71], ranging from limb malformation [17], congenital disorders [72], to cancer [17,30,65,69,73–78]. These observations highlight the importance of studying changes in TADs as a means to

understand genomic regulation. However, methods for identifying changes in TAD boundaries remain underdeveloped.

To our knowledge, there are only three methods that can be adapted for detecting changes in TAD boundaries: `localtadsim` [103], `HiCDB` [83] and `DiffTAD` [101]. Of the three methods, none provide an intuitive, easy to use way of calling differentialTADs. Both `localtadsim` and `DiffTAD` are two-step procedures requiring separately defined TADs and comparing them using a command line utility. `HiCDB` has a built-in TAD caller but does not allow for comparisons of chromosome-specific contact matrices. All three methods require highly specific data types and file names to be able to run. The lack of useability is compounded with issues such as a lack of upkeep, slow runtimes and lack of statistical rigor (Supplementary methods).

As costs of Hi-C data continue to drop, several studies started to investigate the dynamics of 3D changes over time. The most notable applications include cell differentiation studies [21], embryonic development [58,61,62], cancer progression [79]. Typically, TAD boundary changes over time are quantified by overlap [58,61] and classified into distinct patterns [79]. However, general-purpose methods for systematic analysis of TAD boundary changes over time do not exist. Furthermore, the number of replicates for a given experiment continue to rise, requiring methods for defining TAD boundaries consistently detected across replicates of Hi-C data. Two approaches have been developed to identify TAD boundaries across replicates. The first approach involves merging all replicates into a consensus contact matrix and then calling TADs (e.g., `Arrowhead` [3]). The second is to call TADs on individual replicates and aggregate them. Altogether, methods for detecting consensus TADs across Hi-C datasets remain underdeveloped.

We developed `TADCompare`, an R package aimed at providing a fast, accurate, user-friendly and well-documented approach to differential TAD analysis. We introduce a method based on the boundary score statistic [104] and use it to identify five types of TAD boundary changes. The method is extented to allow for calling consensus TAD boundaries and comparing them

between groups of Hi-C replicates. We further demonstrate how the TAD boundary score statistic may be used to analyze TAD dynamics over time course. For both differential TAD boundary detection and time course analysis, we provide novel terminology for classification of TAD boundary changes. We demonstrated robustness of `TADCompare` using simulated data with pre-defined TADs [91] and its ability to reveal distinct biological roles of different TAD boundary changes. In summary, `TADCompare` provides an all-in-one pipeline from TAD calling to differential boundary detection, including time course, that supports replicated Hi-C data analysis. The output is formatted in a commonly used BED format that allows for flexible downstream analyses and visualization. The `TADCompare` R package is freely available on GitHub (https://github.com/dozmorovlab/TADCompare).

6.2 Methods

6.2.1 Representation of Hi-C data as a graph

For a given Hi-C experiment, Hi-C data is represented by a chromosome-specific contact matrix C of non-overlapping regions (aka bins) of size r (resolution of the data). Each entry C_{ij} corresponds to the number of contacts between region i and region j. Previous work has shown that this contact matrix is essentially an analog of the adjacency matrix found in graph theory and Hi-C data can be thought of as a naturally occuring graph where edges are contacts and vertices are genomic regions [51,93,94,104]. The graph representation of Hi-C data is the foundation of our method and allows us to use a graph-clustering based approach to identify and analyze TADs.

6.2.2 Calculating the graph spectrum

The first step of our method is to calculate the graph spectrum, defined as the eigenvectors of the Laplacian of an adjacency matrix. Using the interpretation of the contact matrix as a naturally occuring adjacency matrix, we calculate the Laplacian directly from the contact data. Briefly, the graph spectrum for a given contact matrix is calculated as follows:

1. Calculate the normalized Laplacian $\bar{L}$:

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

where $D = diag(\mathbf{1}^T C)$, where $\mathbf{1}$ is a column vector of size C where each entry is 1. D can be thought of as a vector containing the sum of the degrees for a given node.

2. Perform an eigendecompositon of the Laplacian:

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

In practice, we calculate the first two eigenvectors with the largest absolute values of eigenvalues and organize them into a matrix $\bar{V}$ with dimensions $i \times 2$, where i is the number of regions in the contact matrix. $\bar{V}$ is referred to the graph spectrum of the contact matrix.

6.2.3 Eigenvector gap as a measure of pattern change

We can think of each row of the matrix $\bar{V}$ as a quantification of the pattern of contacts in each region of the contact matrix. Previous work [104] has demonstrated that by taking the Euclidean distance between row $V_{i.}$ and its neighboring row $V_{(i+1).}$, one can measure the similarity in the pattern of contacts between region i and region $i+1$ of the chromosome, termed "eigenvector gap". A TAD boundary manifests itself as a sudden break in the pattern of contacts. This pattern is reflected in the eigenvector gap by a spike in gap size followed by and preceded by smaller gaps (Figure 5). The eigenvector gap quantifies the degree of this break, acting as a proxy for TAD boundary likelihood. To calculate the eigenvector gaps, we perform the following procedure:

1. Normalize columns of $\bar{V}$ to sum to 1:

$$\hat{V_{ij}} = \frac{\bar{V_{ij}}}{\|\bar{V_{.j}}\|}$$

where the subscript $.j$ corresponds to column j.

2. Normalize $\hat{V}$ and project onto a unit circle:

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

3. Calculate the distance between neighboring regions (rows i and $i-1$ of $\tilde{Z}$) and store in a vector D_i:

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

We refer to D as the vector where each entry D_i is referred to as an eigenvector gap. Formally, an eigenvector gap is the Euclidean distance between each succesive row of the first two eigenvectors. In practical terms, the eigenvector gap for a given loci is a measure of how likely that loci is a TAD boundary.

To maintain the associaton of each entry of the vector with its corresponding matrix region, a placeholder is used in the first entry of the vector. This is necessary because we cannot calculate an eigenvector gap for the first entry of the contact matrix due to a lack of a left-bound neighbor. In mathematical terms, this means that for a matrix of size n the total number of eigenvector gaps is $n-1$.

6.2.4 Converting eigenvector gaps to boundary scores

We showed that the distribution of eigenvector gaps can be approximated by a log-normal distribution (Supplementary Figure S7, Supplementary Material). The log-normality allows us to convert the eigenvector gap values into boundary scores:

$$B_i = \frac{(ln(D_i) - \mu)}{\sigma^2}$$

where $ln(D) \sim N(\mu, \sigma^2)$ where μ and σ^2 are the mean and variance of the distribution of the natural log of the eigenvector gaps, respectively, and B is a vector of boundary scores with a N(0,1) distribution. In practice, this value is simply the Z-score for the natural log of eigenvector gaps.

6.2.5 Sliding window eigenvector gap calculation

Frequency of interactions decays following power law as the distance between interacting region in a linear genome increases [141]. This decay leads to noisy and non-informative interactions farther off-diagonal of the contact matrix. To alleviate the effect of noisy distant interactions, we perform spectral decomposition within a fixed-size window that moves along the diagonal of the matrix. For instance, a window size of 15 bins (Supplementary Material) means that only values within 15 bins of the diagonal will be used to calculate the eigenvector gap. The sliding window approach improves the performance of eigenvector gap calculation [104]. Additionally, it provides for faster calculations, operating on many small matrices instead of one large matrix.

6.2.6 Handling of non-informative bins

Non-informative bins refer to bins with less than 20% of non-zero interactions. This percentage is calculated based on regions within our sliding window. Such bins can introduce instability in the algorithm and lack important information. To counter this, we remove these bins before the analysis. This is done for both contact matrices such that, if one contact matrix contains a non-informative bin at a given location and the other does not, we remove it from both. This allows us to make one-to-one comparison of bins.

6.2.7 Differential analysis using boundary scores

To define the differences between two contact matrices, P and R, we compare their eigenvector gaps D_P and D_R, respectively. Given that $ln(D_P) \sim N(\mu_P, \sigma_P^2)$ and $ln(D_R) \sim N(\mu_R, \sigma_R^2)$, it follows that $ln(D_P) - ln(D_R) \sim N(\mu_P - \mu_R, \sigma_P^2 + \sigma_R^2)$. These results allow us to calculate a vector of differential boundary scores:

$$DB_i = \frac{(ln(D_{Pi}) - ln(D_{Ri})) - (\mu_P - \mu_R)}{\sigma_P^2 + \sigma_R^2}$$

or more simply,

$$DB_i = \frac{\sigma_P^2 B_P - \sigma_R^2 B_R}{\sigma_P^2 + \sigma_R^2}$$

where B_P and B_R are the boundary scores for the P and R matrices, respectively. This score can be thought of as the difference in TAD boundary likelihood for a given loci in two data sets. Due to the aforemnetioned normality of the difference in log eigenvector gaps, DB_i can be thought of as a simple z-score where $DB \sim N(0, 1)$.

6.2.8 Time course boundary changes

Boundary scores provide a convenient method for modeling the change of TAD boundaries over time. For a given TAD boundary, or, any region of the genome, we can monitor the trajectory of the boundary score. Over time, we can define boundary score changes based on their deviation from a baseline level (typically, the first time point). It is expected that these scores will be relatively constant over time except in regions where a boundary appears or disappears. The trend across time points can be recorded and the pattern of change classified accordingly. Our implementation of time course boundary analysis allows for the usage of multiple replicates for a given time point. Briefly, at each region of the genome the consensus boundary score is calculated, defined as the median of consensus scores across all replicates, and is then used to identify boundaries.

6.2.9 Data sources

All simulated data were downloaded from `HiCToolsCompare` repository [91]. In total, we used 25 simulated matrices with varying levels of noise. For sparsity and downsampling analysis matrices were manually created based on matrices from `HiCToolsCompare` matrices with the mininum noise level (see [104] for methods description). Data for comparisons across cell lines, replicates and tissues were taken from [108], generated at 40kb resolution (Supplementary Table S6). Time course data was taken from [142], HCT-116 human colon cancer cell-line at four time points after auxin-treatment withdrawal (20, 40, 60, 180 minutes). Contact matrices were generated at 25kb, 50kb and 100kb using the `straw` tool from `Juicer` [47]. Chromatin state data were taken from chromHMM [143]. Histone modifications and transcription factor binding sites were downloaded from the Encyclopedia of DNA Elements (ENCODE) [144] (Supplementary Table S7).

6.2.10 Gene enrichment testing

All gene enrichment testing was performed using the `rGREAT` (Version 2.0) R package. Briefly, we detect genes within 5kb upstream and 1kb downstream of each type of TAD boundary changes, similar to work of others [83]. For each Gene Ontology (GO) and pathways, a hypergeometric test is then performed to determine over-representation of TAD boundary-associated genes. For all figures, we report results for GO Biological Processes. Results for GO Molecular Function, GO Cellular Component, MSigDB and PANTHER pathways are reported in tables.

6.2.11 Colocalization enrichment testing

A permutation test was used to quantify the enrichment of colocalization of TAD boundaries of interest with genomic annotations. Briefly, we flank each type of TAD boundary changes (differential or time course) by 50kb on each side and calculate the mean number of genomic annotations across those regions (observed enrichment). Next, we generate two sets

of bins, one the size of the TAD boundaries which we are testing (considering the flanking) and another the size of all other bins. The difference in the mean number of genomic annotations colocalized with TAD boundaries of interest was calculated for each set (expected enrichment). We repeat this procedure 10,000 times. We calculate the permutation p-value by taking the number of times the expected enrichment was greater than the observed enrichment, and dividing by 10000. $\alpha = 0.05$ was set to assess statistical significance.

6.3 Results

6.3.1 A modified spectral clustering approach is better suited for TAD boundary detection than other approaches

Our previous work on TAD detection using spectral clustering, implemented as a `SpectralTAD` R package [104], introduced the concept of the boundary score statistic, adapted here for differential boundary detection. Briefly, the boundary score is calculated for each bin by sliding a window across the diagonal of the contact matrix, calculating the eigenvectors of the Laplacian matrix, finding the distance between consecutive eigenvectors (eigenvector gap) and converting them into Z-scores (boundary score, see Methods). The boundary score is a continuous measure of TAD boundary likelihood.

In contrast to other metrics for boundary identification that rely on finding inflection points of monotonic functions, such as directionality index [18], insulation score [15], RobusTAD score [49] (Supplementary Material), our boundary score weakly oscillates within TADs and spikes at the TAD boundary (Figure 5). This unique behavior enables easy distinction between TAD boundaries and non-TAD boundaries. An additional advantage of the boundary score is that its magnitude is directly interpretable as a "boundary strength". This is in contrast to other methods which are only interpretable relative to neighboring points. We can use this interpetability for parametric modeling of TAD boundary behavior. Our previous work has shown that the boundary score is robust to noise, sparsity and changes in sequencing depth of Hi-C data [104]. Thus, the boundary score is ideal when finding differences in TAD

boundaries.

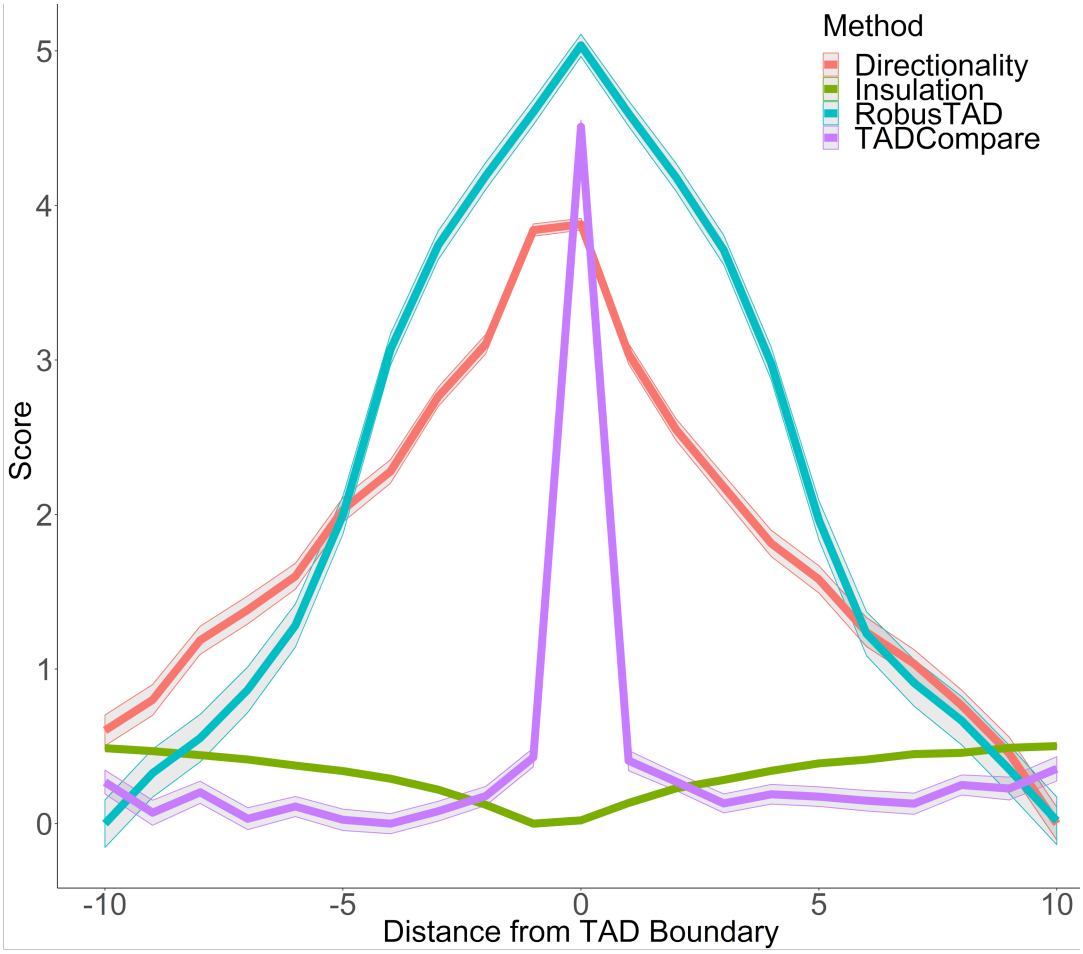


Figure 5. Boundary score distinguishes TAD boundaries better than monotonic metrics. TAD boundary scores calculated with four methods: directionality index, insulation score, RobusTAD and TADCompare boundary scores are shown. X-axis - Distance from TAD boundary, measured in bins (40kb each), Y-axis - Score (signed log10 values centered at zero). Results from five simulated contact matrices, 40kb resolution, with manually annotated TAD boundaries [91] are shown.

6.3.2 Differential boundary scores translate to five types of TAD boundary changes

Differential boundary score is a measure of the difference between TAD boundaries between two samples. This score follows a standard normal centered at 0 (see Methods, Supplementary Figure S7). Differential TADs are detected by finding regions with the differential boundary score is greater than 2 (Supplementary Figure S11), which intuitively corresponds to differences with a p-value smaller than 0.05.

We divide TAD boundary changes into five categories (complex, split, merge, shifted, strength change; Figure 6, Supplementary Figure S12). A similar strategy was used in Ke et al. [62]. A TAD can be **split** between the datasets meaning it exists as a continuous TAD in one and is split into two or more TADs in another. In practice, this situation requires two shared TAD boundaries and a differential TAD between them. **Merging** is the opposite of splitting and arises when a TAD boundary surrounded by two non-differential TAD boundaries disappears in one of the contact matrices. Classificiation of TAD boundary change as merged and split depends on the reference contact matrix being compared to. Finally, a TAD can be split in a **complex** way meaning they are neither split or merged but instead taking on an entirely new structure. Merged and split boundaries represent the structural change (splitting or merging) of the same TAD as opposed to complex boundaries which we consider to be part of a completely different TAD. The "complex", "merge", and "split" TAD boundaries are considered to be the most disruptive changes in the 3D structure of the genome.

A **shifted** TAD boundary is defined as the non-overlapping boundary that lies within five bins (or other user-defined threshold) of a boundary in the contact matrix in which it is being compared to. A **strength change** occurs when a TAD boundary is present in both contact matrices but its differential boundary score magnitude is greater than the differential threshold of 2. The other cases are considered to be non-differential TAD boundaries. This framework allows us to systematically compare and classify TAD boundary changes.

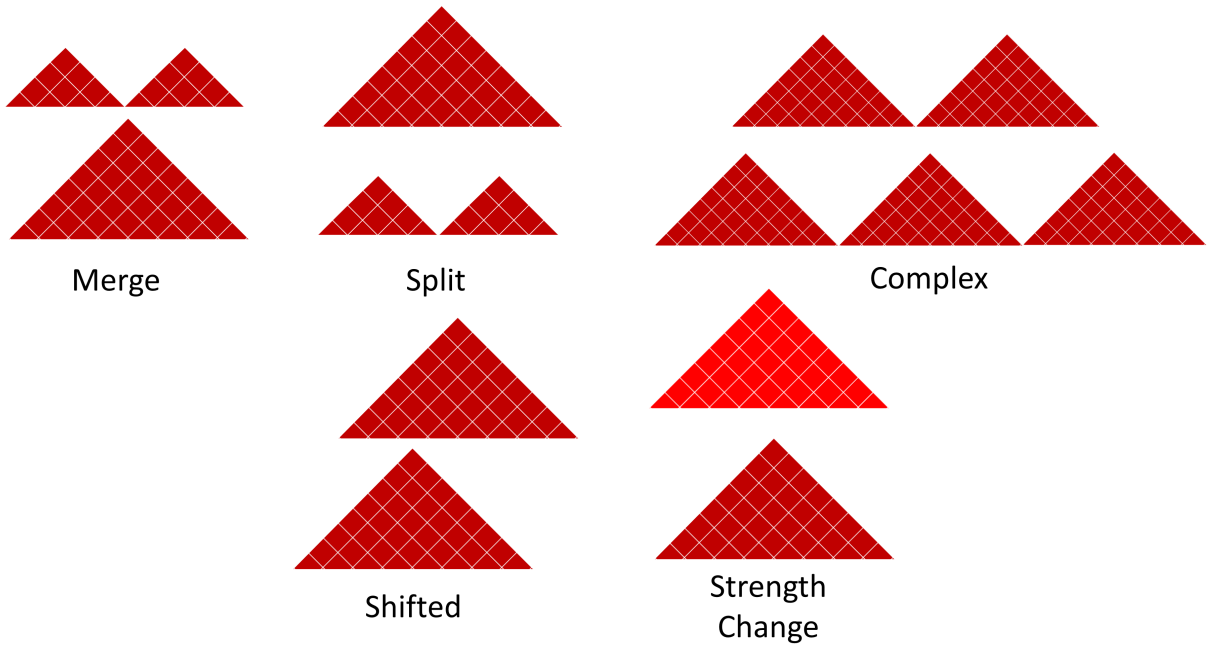


Figure 6. Five types of TAD boundary changes. Complex, split, and merge TAD boundary changes are considered as the major differences, while shifted and strength changes are considered as the minor differences.

6.3.3 TAD boundaries are highly consistent in both technical and biological replicates

Previous studies have shown that the overlap between TAD boundaries in replicate data ranges from around 60% to 70% [3,18,103]. Additionally, technical replicates have been shown to have a slightly higher proportion of shared TAD boundaries (~65%) than biological replicates (~60%) [103]. We have tested and confirmed these observations by showing that significantly more TADs were non-differential in technical replicates than in biological replicates (73% vs. 65.7%). Similarly, 9.3%/8.1% of boundaries showed significant strength change, while 7.8%/6.1% were shifted in the biological/techincal replicates, respectively. A similar trend was observed for complex and merge-split boundaries. In summary, only 17.2%/12.8%

of TAD boundaries were differential in biological/technical replicates, respectively (Figure 7A), confirming the higher stability of TAD structures in technical replicates.

6.3.4 TAD boundaries are more similar within cells than tissues

Previous research showed that TADs are largely invariant across cell lines and to a lesser extent tissue types [3,12,108]. However, the types of TAD boundary changes remained undefined. We compared Hi-C matrices of 7 different cell-lines and 18 different tissue types [108] (Supplementary Table S8). In total, the average percentage of differential TAD boundaries was significantly less in cell lines (22.5%) than tissue samples (39.7%, Figure 7B). As expected, these percentages were higher than those for biological (17.2%) and technical replicates (12.8%). These results suggest that the variability of TAD boundaries, mirrors the homogeneity of data types (technical replicates, biological replicates, cell lines and tissues, in that order).

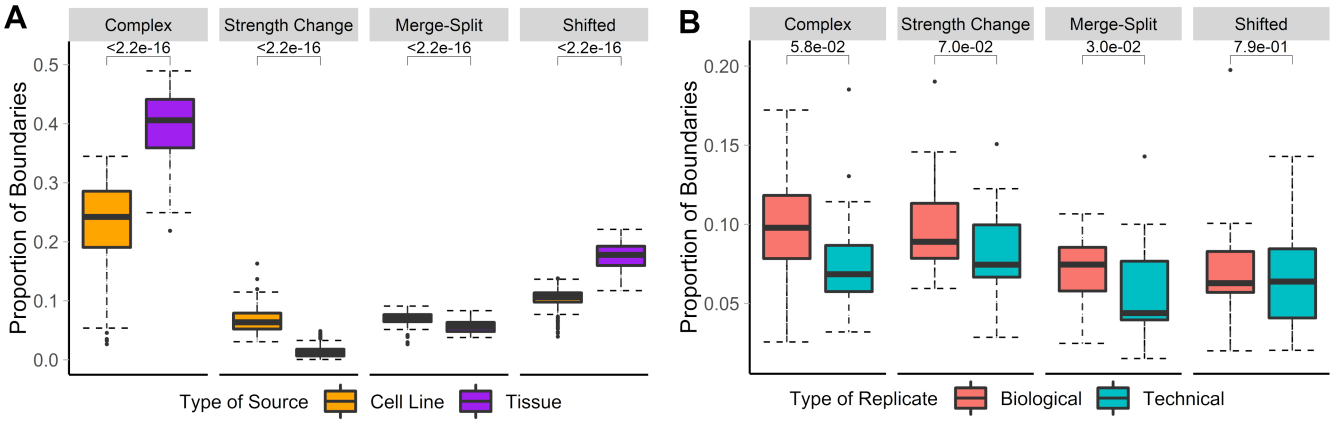


Figure 7. Biological replicates and cell lines have more differential TADs than technical replicates and tissues, respectively. Differential TADs were calculated between Hi-C datasets of biological and technical replicates (Panel A, HCT-116 cell line, 50kb resolution, chr 1-22 [142]) and between cell lines and tissues (Panel B, various cell lines, 40kb resolution, chr1-22, [108]). Types of TAD boundary changes were recorded and the proportions of TAD differences for each type were summarized across chromosomes.

6.3.5 Each type of differential TAD boundaries is associated with different levels of epigenomic enrichment

To understand biological relevance of types of TAD boundary changes, we identified changes between the GM12878 and IMR90 cell lines (chr1-22, 40kb resolution, [108]) and categorized them according to the type of change. For each change type, we assessed the number of overlapping peaks and calculated the enrichment of four genome annotation marks known to co-locate with TAD boundaries - CTCF, RAD21, insulators and heterochromatin states.

We found that non-differential boundaries had a higher average number of overlapping peaks for all four marks, followed by "strength change" TAD boundaries (Figure 8A). Similarly, enrichment of non-differential TAD boundaries was the most significant (Figure 8B). Notably, number of peaks for each mark was highly variable in strength change TAD boundaries (Figure 8A), suggesting their biological relevance is less certain. Similarly, "shifted" TAD boundaries had the lowest average number of peaks, suggesting that they may be detected due to noise and, consequently, be less biologically significant. In contrast, "complex" and "merge-split" TAD boundaries had a moderate number of overlapping peaks and were moderately enriched in them (Figure 8). These results highlight the varied biological relevance of different types of TAD boundary changes and suggests "complex" and "merge-split" changes are biologically important alterations of TAD structure.

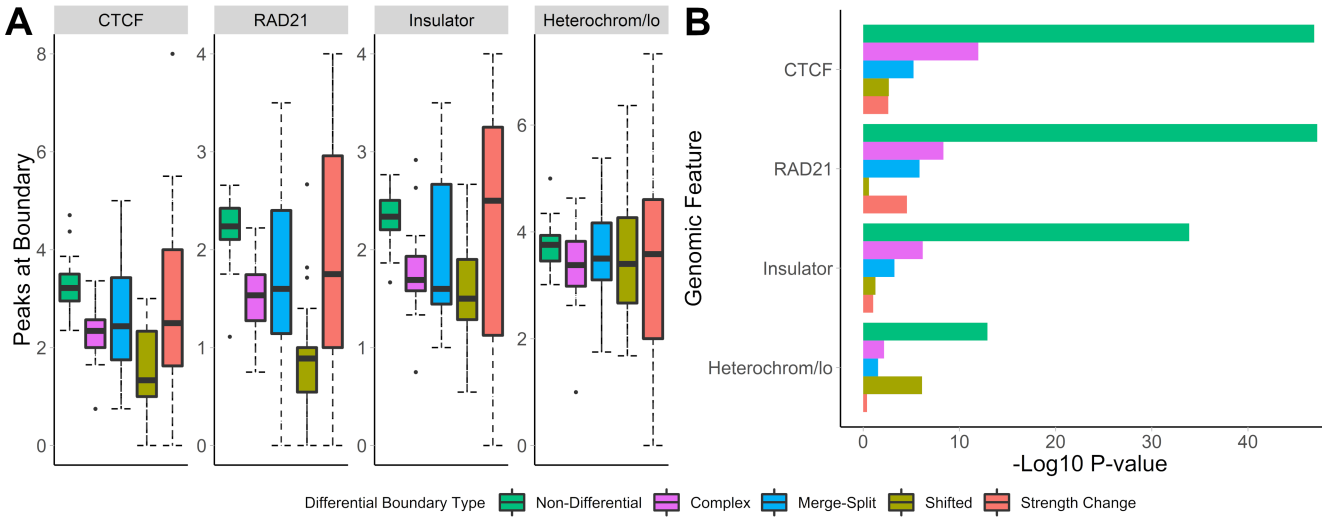


Figure 8. Non-differential TAD boundaries are more enriched for selected genome annotation marks than other types of differential TAD boundaries. Differential TAD boundariess were called between GM12878 and IMR90 cell lines and categorized based on differential boundary type. (A) Number of peaks at TAD boundaries and (B) permutation p-values (-log10) are shown. Data from [108], 40kb resolution, chr 1-22.

6.3.6 Each type of differential TAD boundaries is associated with distinct biological functionality

To test biological significance of different types of TAD boundary changes, we compared mesenchymal stem cells (MSC) against neural progenitor cells (NPC) [108]. Altogether, we found that the vast majority of boundaries are either complex (38.6%) or non-differential (32.6%). Shifted (17.5%), merge-split (7.7%) and strength change (3.5%) were less common (Figure 9A). We investigated enrichment of genes in proximity of each type of differential TAD boundary in biological processes and other gene ontology- and pathway types using GREAT [145] (see Methods). As NPCs are more advanced on differentiation path than MSCs, we expected that TAD boundaries changed between them would be associated with genes responsible for neural development-related processes. Indeed, genes around "merge" and "complex" TAD boundary changes, as well as the "non-differential" TAD boundaries

were enriched in a variety of developmental processes (e.g., "cellular developmental process", etc.), including neural-specific ("nervous system development", Figure 9A). Notably, "split" TAD boundary changes were not enriched in these processes, indicating the importance of directionality of TAD boundary changes. Genes around "merge" and "non-differential", but not "complex", TAD boundaries were enriched in differentiation-related processes (e.g., "positive regulation of cell differentiation"), while "forebrain radial glial cell differentiation" and "neural tube development" processes were exclusively enriched in genes around "merged" TAD boundaries (Figure 9A). In this case, "merge" indicates boundaries enriched in the NPC cell-line causing a seperation of TADs in MSC and "split" indicates a split in NPC caused by a boundary enriched in MSC. As expected, genes around "noisy" TAD boundary changes ("shifted" and "strength change") lacked enrichment in any biological processes (Figure 9A, Supplementary Table S9). These results emphasize importance of classifying TAD boundary changes into distinct patterns which tend to be associated with distinct biological functionality.

To further test whether different types of TAD boundary changes reflect biology of an experimental system, we used post-auxin treatment time course experiment from Rao et al. 2017 study (HCT-116 cell line, 40kb resolution, 20, 40, 60, and 180 minutes following auxin withdrawal, 4 replicates at each time point) [142]. Auxin treatment eliminates CTCF binding genome-wide; consequently, the majority of TAD boundaries should be absent and gradually re-appear following auxin withdrawal. To identify biological processes associated with re-appearing of TAD boundaries, we compared first and last time points (20 and 180 minutes) following auxin withdrawal. As TAD boundaries were reported to be enriched in housekeeping genes [10], we expected genes around appearing TAD boundaries to be enriched in general cellular processes. Indeed, the vast majority of differential TADs were complex (41.4%) and non-differential (34.7%) (Supplementary Figure S13). We found that only genes around "non-differential" and "complex" TAD boundary changes showed some level of enrichment (Supplementary Figure S13, Supplementary Table S10). As expected,

"metabolic processes" and various developmental and housekeeping processes were specifically enriched in genes around complex TAD boundary changes, while cyclic AMP syntesis and metabolic processes were enriched in genes around "non-differential" TADs. From these results, we show that `TADCompare` is able to correctly classify non-essential TAD boundary changes ("shifted", "strength change") and detect distinct TAD boundary changes associated with shared and unique biological processes.

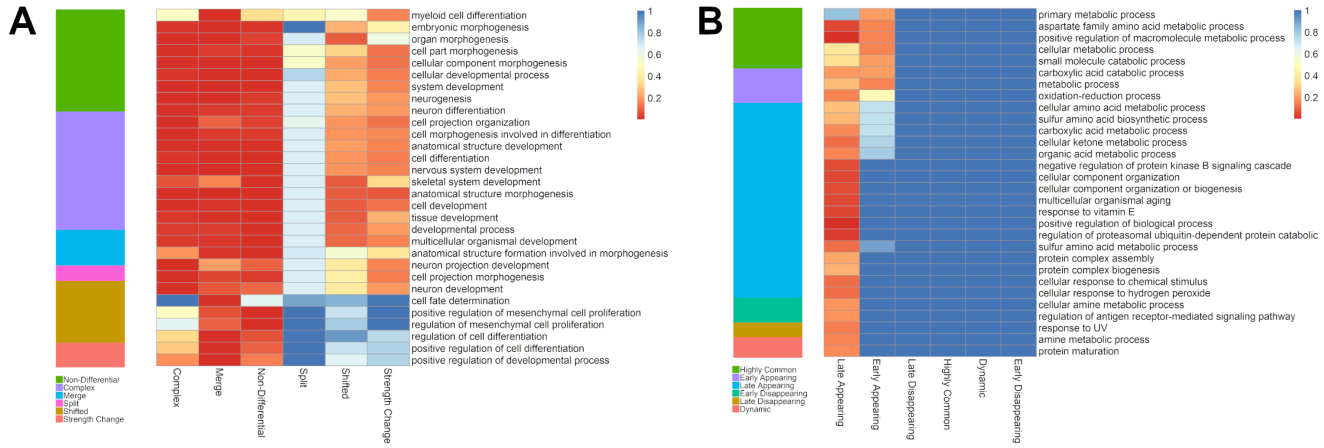


Figure 9. Gene enrichment analysis of differential TADs. Differential TADs were called between A) neural progenitor cell (NPC) and mesenchymal stem cells (MSC) (chr1-22, 50kb resolution, [142]), and B) across time-course in auxin-treated cells from the HCT-116 cell-line (chr1-22, 40kb resolution, [108]. For each type of TAD boundary change, -log10-transformed enrichment p-values (`rGREAT`, see Methods) are shown as heatmaps. The top 30 gene ontology biological processes, in terms of average enrichment, are shown.

6.3.7 Time course analysis framework

Time course analysis of TADs refers to the analysis of TAD boundary dynamics over time. The quantitative nature of boundary score allows us to monitor its changes at TAD boundaries across any number of time points. We recommend taking a union of TAD boundaries detected at each time point and monitor boundary score changes for each TAD boundary. Monitoring TAD boundary scores across time points provides an opportunity to quantify patterns of

TAD boundary changes.

Using the boundary score cutoff of 3 for TAD boundary definition, we define six patterns of temporal TAD boundary changes (adapted from [79], Table 2, Figure 10). *Highly common* TAD boundaries refer to boundaries present across all time points or in three out of four time points. *Early appearing* TAD boundaries switch from non-boundary to boundary at second time points and stay as boundaries for the rest of the time points. Conversely, *early disappearing* TAD boundaries switch from boundary to non-boundary at the second time point and stay as non-boundaries. *Late appearing* TAD boundaries switch from non-boundaries to boundaries at the last or the second to last time point. Conversely, *late disappearing* TADs switch from boundaries to non-boundaries at the last of the second to last time point. Finally, *dynamic* TAD boundaries are those which have inconsistent boundary status and do not follow any of the aforementioned patterns (Figure 10). These six patterns of temporal changes can be easily adopted for larger number of time points.

Table 1. Six patterns of temporal TAD boundary changes. Each column corresponds to a point in time. "1" refers to the presence of a TAD and "0" refers to the abscence of a TAD at considered time point. Total column shows percentage of occurences in CTCF degradation-recovery time course, HCT-116 cell line, chr1-22 [142].

Temporal TAD Type	Time Point 1	Time Point 2	Time Point 3	Time Point 4	Total (% Occurrence)
Highly Common	1	1	1	1	326 (17.35%)
	1	0	1	1	
Early Appearing	0	1	1	1	184 (9.79%)
Early Disappearing	1	0	0	0	133 (7.08%)
Late Appearing	0	0	1	1	1047 (55.72%)

Temporal TAD Type	Time Point 1	Time Point 2	Time Point 3	Time Point 4	Total (% Occurrence)
	0	0	0	1	
Late	1	1	0	0	79 (4.20%)
Disappearing					
	1	1	1	0	
Dynamic	1	0	1	0	110 (5.86%)
	1	0	0	1	

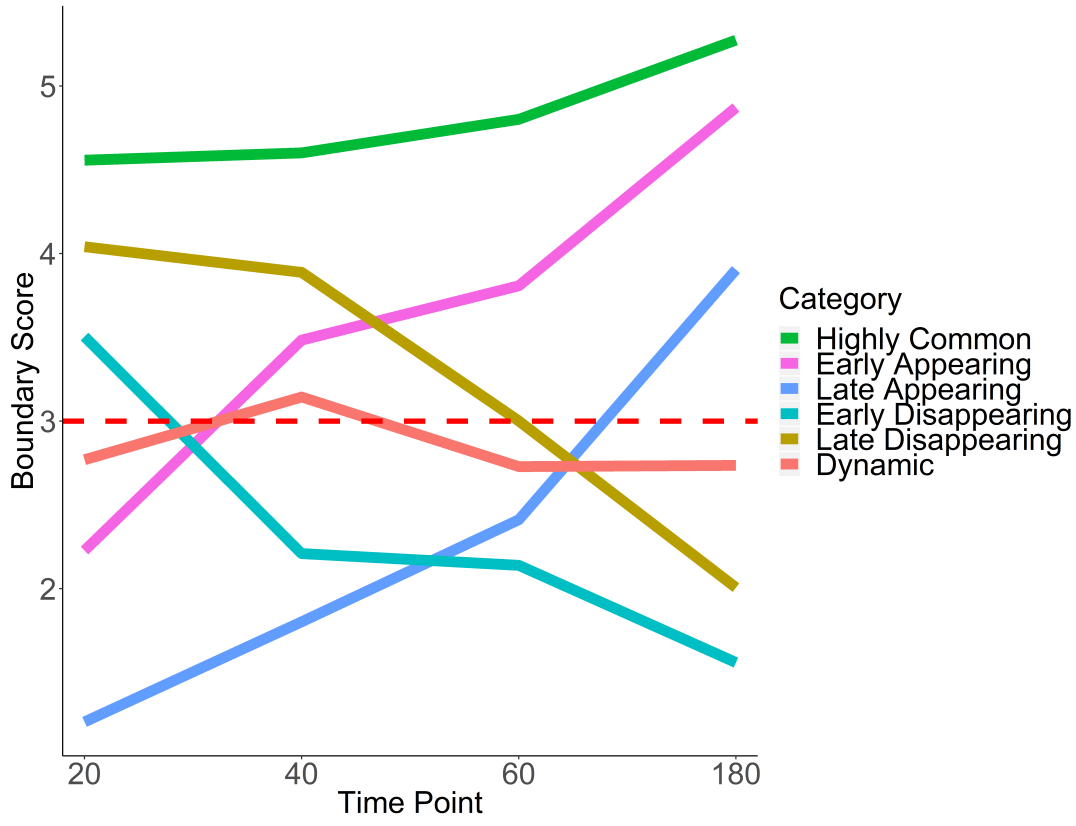


Figure 10. Six patterns of TAD boundary score changes across time. Average trajectories for each pattern of boundary score change are shown. The red horizontal line indicates the cutoff for TAD boundary detection. HCT-116 cell line, 40kb resolution, chr 1-22.

6.3.8 Temporal TAD boundary types are associated with different levels of epigenomic enrichment

To evaluate biological relevance of temporal patterns of TAD boundaries, we used post-auxin treatment time course experiment introduced above. Briefly, HCT-116 cells were treated with auxin to eliminate TAD boundaries, and Hi-C measures were obtained at 20, 40, 60, and 180 minutes following auxin withdrawal and subsequent TAD boundary reappearance [142]. Accordingly, we expected to detect some number of highly common TAD boundaries (already existing at 20 minutes) and boundaries appearing at different stages of post-auxin withdrawal (early/late appearing). Conversely, dynamic and early/late disappearing TAD boundaries should be rare and may potentially constitute noise in TAD boundary detection.

Boundary scores were calculated for auxin-treated cells 20, 40, 60 and 180 minutes after withdrawal. Taking the union of TAD boundaries (boundaries detected at one or more time points), we calculated temporal patterns for each boundary. We found that the vast majority of boundaries were late appearing TADs (55.7%) (Table 2, Figure 11B). Early appearing TADs (9.8%) and highly common TADs (17.3%) made up most of the other TADs present at the end of the time course. ~20% of TAD boundaries were highly common, i.e., resistant to auxin treatment, a number similar to previous works [146]. Meanwhile, 5.9% of TAD boundaries were dynamic, 7.1% were early disappearing and 4.2% were late disappearing, highlighting potential errors in TAD boundary detection. In summary, some TAD boundaries can be detected at 20 minutes post-auxin treatment and remain present through all time points; however, the timing of TAD boundary restoration varies by TAD.

To test whether TAD boundaries associated with different temporal patterns have different functional roles, we investigated their overlap with and enrichment in the common marks of TAD boundaries (CTCF, RAD21, insulators, heterochromatin, Figure 11A). For highly common, early- and late-appearing TAD boundaries, we observed more overlaps with CTCF and RAD21 sites, insulator and heterochromatin states (Supplementary Table S11). Similarly, these types of TAD boundaries were highly enriched in the aforementioned genomic

annotations (Figure 11B). Conversely, dynamic, early and late disappearing TAD boundaries showed less overlap with CTCF, RAD21, insulator and heterochromatin marks, and were less enrihed in them. These observations suggest that disappearing and dynamic TAD boundaries are likely detected due to noise in the data, while TADs appearing after auxin treatment expectedly represent biologically relevant signal.

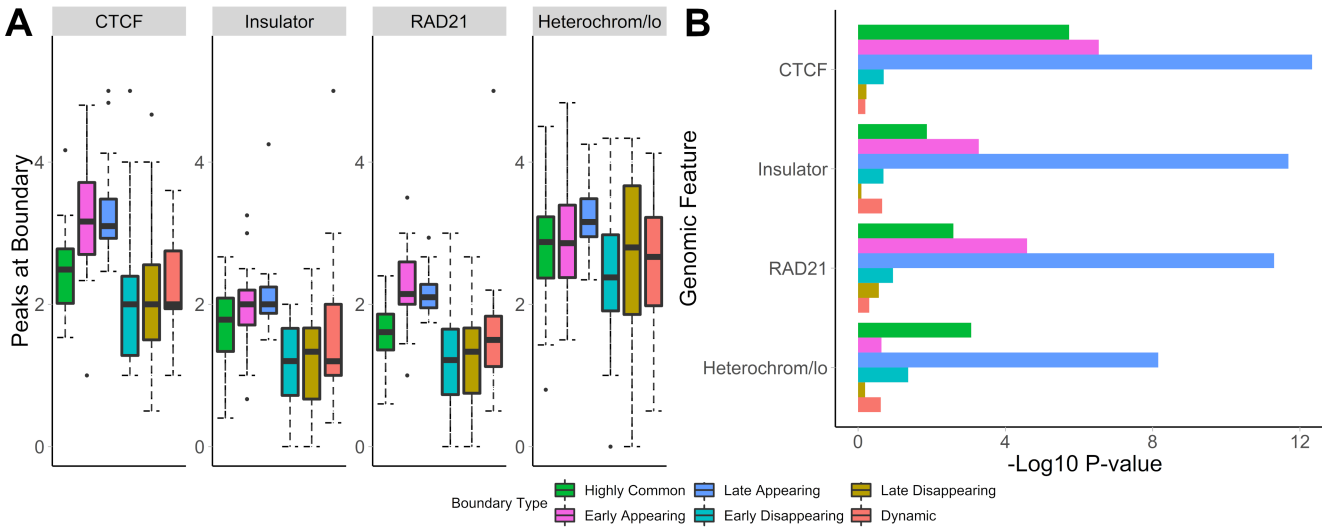


Figure 11. Common and appearing TAD boundaries show stronger enrichment in known epigenomic marks. The number of peaks at TAD boundaries (A), and permutation p-values (B) within 50kb of boundaries in each temporal classification are shown. Hi-C data from [142], 50kb resolution, HCT-116 cell-line, chr 1-22.

6.3.9 Temporal TAD boundary types are associated with distinct biological functionality

We further investigated these results using gene enrichment analysis (Supplementary Table S12) for temporal boundary types. We found that, with a few exceptions, all significant GO Biological pathways were enriched in late appearing TADs or early appearing TADs (Figure 9B, Supplementary Table S12), which make up the majority of TAD boundaries (Table 2, Figure 9B). Both early and late appearing TAD boundaries were enriched in metabolism-related processes, such as “cellular metabolic process”, “oxidation-reduction process”. Late

appearing TADs, on the other hand, were enriched in "cellular component organization", "protein complex biogenesis" and the like processes (Figure 9B). These results are expected as cells may be activating metabolic and biogenesis pathways to recover after destruction of TAD boundaries by auxin. These results confirm that `TADCompare` is able to accurately classify biologically relevant temporal TAD boundary changes and discern them from noisy changes.

6.3.10 Consensus boundary score for defining robust TAD boundaries across multiple Hi-C datasets

The sizeable proportion of noisy "shifted" and "strength change" TAD boundary changes across Hi-C datasets (Figure 9) highlights the need to identify TAD boundaries that are robustly detected. The consensus boundary score, defined as median of boundary scores across replicates, addresses this challenge. Intuitively, higher consensus boundary scores correspond to TAD boundaries supported by evidence from multiple replicates (Table 1). This is in contrast to a union of TAD boundaries, where TAD boundaries detected in at least one Hi-C dataset are pooled together. Consensus boundary scores allow us to filter out boundaries with insufficient support from multiple replicates, thus "denoise" the detected TAD boundaries. Given the fact that boundary scores are log-normally distributed (Supplementary Figure S7, Supplementary Methods), the consensus boundary scores will also be asymptotically normal. The consensus boundary score can be used as a proxy for the normal boundary score for the analysis of replicated Hi-C datasets. Consequently, the consensus boundary scores may be compared to define TAD boundary changes between groups of replicated Hi-C datasets.

Table 2. Consensus (aka median) boundary score is supported by high boundary scores from multiple replicates. Examples of boundary scores across five regions in three replicates, and the corresponding consensus boundary score. Both union and consensus TAD boundaries are calculated using a cutoff of 3.

Boundary Score 1	Boundary Score 2	Boundary Score 3	Consensus Boundary Score	Union TAD Boundary?	Consensus TAD Boundary?
1	2	1	1	No	No
3	2	1	2	Yes	No
5	5	4	5	Yes	Yes
3	3	3	3	Yes	Yes
6	0	0	0	Yes	No

6.3.11 Consensus TAD boundaries are supported by strong biological evidence

To investigate the biological relevance of TAD boundaries defined using consensus boundary score, we defined consensus TAD boundaries across 7 cell-lines (17 matrices total) [108]. These boundaries represent cell type-invariant TAD boundaries supported by evidence from multiple datasets. Bins of the genome were separated into three categories based on the level of their consensus boundary score (<2, 2-4 and >4). In total, there were 65,336 bins (40kb resolution). Expectedly, the majority (62,791 bins, 96.1% of all bins) were in the <2 category, 2,032 (3.1%) bins were in the 2-4 category, and 513 (0.8%) bins were in the >4 category. We assessed the number of overlapping peaks and the enrichment of CTCF, RAD21, insulators and heterochromatin states in different categories of bins. Expectedly, we observed increasing average number of peaks overlapping bins selected at more stringent consensus boundary score thresholds (Figure 12, Supplementary Table S13). Similarly, bins with higher consensus boundary scores have stronger enrichment in genome annotations. These results suggest that bins with higher consensus boundary scores (i.e., supported by evidence from multiple Hi-C datasets) are more biologically relevant. Therefore, to define consenus TAD boundaries, we use a consensus boundary score cutoff of 3.

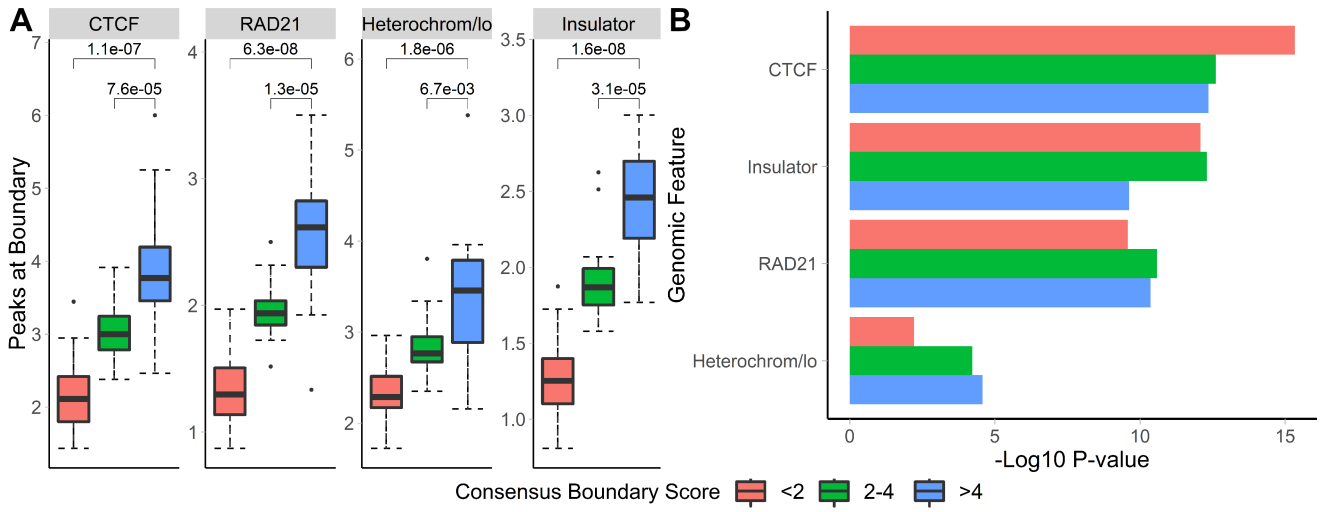


Figure 12. TAD boundaries defined at higher consensus boundary score thresholds show stronger overlap with and enrichment in known epigenomic marks. TAD boundaries were classified based on the range of their consensus boundary score. Enrichment of genomic factors known to occur near TAD boundaries were calculated. (A) The number of TAD peaks within 40kb of TAD boundaries with the corresponding consensus score range and (B) the permutation p-values for each score range are shown. Data from seven cell lines, chr1-22, 40kb resolution, [108].

6.3.12 The union of TAD boundaries is supported by weaker biological evidence than consensus boundaries

The union of TAD boundaries called in individual Hi-C datasets represent an alternative method of defining TAD boundaries across multiple datasets (Table 1). The union method may be useful for analysis of time course data, where TAD boundaries are expected to change across individual datasets. We hypothesized that the union method would select for less biologically relevant set of TAD boundaries because many may be detected due to noise in Hi-C data.

To evaluate the biological relevance of TAD boundaries called using both methods, we call consensus and union TADs on a set of replicates (four cell lines, 40kb resolution, 3 replicates

each, data from [108]). Consensus scores were calculated separately for each cell line among the 3 replicates. Expectedly, the consensus method filtered out 38% of TAD boundaries (4906 vs. 3059, Supplementary Figure S14), suggesting that many TAD boundaries are detected in single datasets. We found that TAD boundaries called using consensus boundary score overlapped significantly more with CTCF sites (P = 0.0006) and RAD21 (P = 0.0002) than those called using the union method (Figure 13A). While the enrichment results were similar for consensus- and union-defined TAD boundaries, consensus TAD boundaries were more significantly enriched in "heterochromatin" (Figure 13B). Together with previous observations (Figure 12), these results strengthen our conclusion that consensus boundary scores are more effective in removing "noisy" TAD boundaries that otherwise would be captured using the union method.

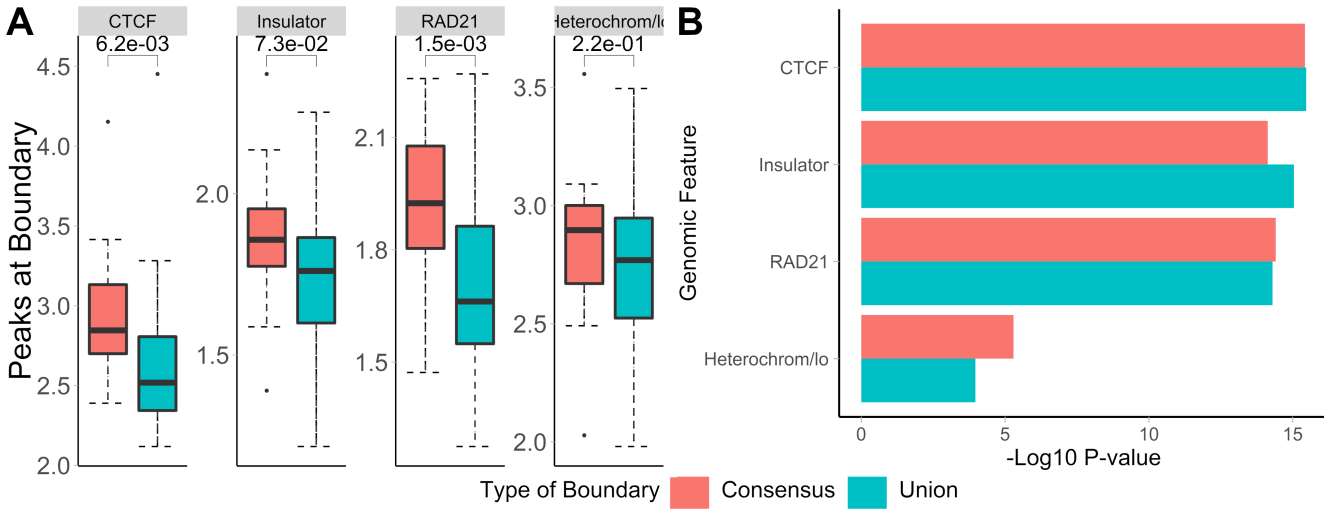


Figure 13. Consensus TAD boundaries show stronger overlap with and enrichment in known epigenomic marks than union of TAD boundaries. (A) Number of peaks at TAD boundaries and (B) permutation p-values (-log10) are shown. Data from [108], four cell lines, 40kb resolution, chr 1-22.

6.3.13 Runtime performance of TADcompare

When ran on data from [3], without parallelization, both consensus TAD calling and differential TAD detection were exceptionally fast. In total, for the entire genome, differential TAD detection took ~6 seconds on 100kb data, ~9 seconds on 50kb data, ~17 seconds on 25kb data and ~312 seconds on 10kb data. In the case of consensus TAD calling, `TADCompare` took ~17 seconds to run on 50kb data for 4 matrices, ~32 seconds for 8 matrices and ~45 seconds for 12 matrices. On 10kb data, it took ~611 seconds to run for 4 matrices, ~1152 seconds for 8 matrices and ~1680 seconds for 12 matrices. For a full summary of runtimes across all resolutions see Supplementary Figure S15.

6.4 Discussion

The initial development of Hi-C technologies focused on investigating individual genomes. While several key properties have been discovered (chromosome territories, A/B compartments, TADs, chromatin loops), the next steps include investigating changes in the 3D structure across multiple conditions. We [147,148] and others [149,150] started to develop methods for comparative analysis of the 3D structures. However, to our knowledge no methods are available for differential analysis of TAD boundaries. In this work, we introduce a method for differential TAD boundary analysis, including time course, that support replicated Hi-C data. The method is based on a novel boundary score metric that provides continuous measure of TAD boundary likelihood [104]. We introduce unqiue terminology for classifying differential and temporal TAD boundary changes. We show that our approach is robust and effective at identifying distinct biology associated with different types of TAD boundary changes. Our method is implemented in the `TADCompare` R package available on Bioconductor, filling a vital gap in intuitive R-based software for TAD detection and comparison.

The boundary score concept developed in our work addresses three main problems: differential TAD boundary detection, time course analysis of TAD boundary changes, and consensus TAD boundary calling. Yet, it has a wider scope of applications. Future work

will expand the utility of boundary score by developing a similarity/reproducibility score to measure the agreement between (multiple) Hi-C matrices, in the same vein as HiCRep [151], Selfish [152], GenomeDISCO [153], HiC-Spector [154], QuASAR-Rep [155]. Furthermore, for differential TAD boundary detection, our method is still limited to the comparison of two profiles of (consensus) TAD boundary scores. This approach will eventually be expanded to include comparisons of many contact matrices, similar to the concept of comparing groups of multiple replicates in RNA-seq data. Finally, there is still room for expansion of time course boundary analysis. The continuous nature of boundary score allows for adopting time course analysis methods developed for gene expression studies [156]. More flexible classification of temporal trends may be considered, such as 24 temporal patterns proposed by Zhou et al. 2019 [79], or fuzzy clustering techniques that do not require a pattern to belong to a specific cluster [157]. In summary, our work enables further development of various aspects of 3D genome analysis.

One difficulty in our work is how to properly quantify the biological relevance of TAD boundaries (differential, time-varying and consensus) that we detect. There is no natural gold standard for TAD boundaries but there are known genomic features that form the building blocks of TADs (CTCF, RAD21). In practice, we can use enrichment near boundaries of these as a proxy for "true boundaries". To test whether enrichment is different than random (non-boundaries), we use a permutation test and present these p-values. However, we can not compare these values between groups due to the fact that the variance of random samples are affected by sample sizes. As a result, all p-values presented in the manuscript must be considered only within the context of the boundary type itself and not compared between boundary types or resolutions.

Our results in this manuscript demonstrate the ability of `TADCompare` to provide accurate, biologically relevant, results. The methods implemented span differential, time-course and consensus analysis. To date, `TADCompare` is the only actively maintained and publicly available tool to provide any of this functionality. We intend for `TADCompare` to be a one-stop tool for

comparison of HiC datasets, providing simple, easy-to-interpet, results in a timely manner. As a one-of-a-kind tool, `TADCompare` will increase the ability of researchers to extract important biological insights from the structure of TAD boundaries.

7 Chapter 5: Discussion

7.1 Conclusion

In this work, we introduce a range of novel approaches for identifying and analyzing TADs. We justify the usage of these methods by demonstrating how the HiC contact matrix is actually a naturally occuring adjacency matrix of a weighted graph. We show how these approaches can be used to glean important biological insights from 3D genomic data. Each method introduced in this dissertation are available in either the `SpectralTAD` or `TADCompare` R packages.

In Chapter 2, we introduce the graph-representation of the contact matrix and the corresponding eigenvector gap metric. We show how a sliding window can be used to take advantage of the ordered nature of HiC data. Additionally, we introduce the silhouette score metric which can be used to automatically choose the number of TADs in a given dataset. We show that the eigenvector gap metric is a natural measure of TAD boundary likelihood. The method is applied to simulated to demonstrate its robustness to common sources of HiC bias (noise, sparsity and sequencing depth). We find that boundaries detected by SpectralTAD are associated with known genomic markers such as CTCF and RAD21, confirming their biological relevance. Finally, we show that our windowed spectral clustering method is exceptionally steps. At each step, we show that our method is superior to previous methods.

In Chapter 3, we demonstrate how the boundary score, derived from the eigenvector gap, can be used to automatically detect hierarchical TADs. Using this approach, we introduce novel results showing that the level of TAD hierarchy is associated with enrichment of

genomic markers. Additionally, we show that the hierarchy is preserved across cell-lines and tissue types demonstrating consistency of 3D structure. The complete hierarchical method is included in the `SpectralTAD` package.

In Chapter 4, we introduce the problem of differential TAD detection. We show how eigenvector gaps can be compared between datasets using the differential bpoundary score. We introduce the three problems of differential TAD analysis: differential boundary detection, time-course TAD analysis and consensus boundary detection. We classify differential TADS. In the case of differential TAD and time-course analysis, we introduce novel terminology for categorization of changes. We show these categories correspond to different levels of genomic enrichment. Additionally, we show that differential regions are associated with changes in gene expression levels. Finally, we introduce the consensus TAD score and show that higher levels directly correspond to stronger TAD boundaries and higher levels of enrichment, confirming its ability to linearly measure TAD likelihood. The three methods are freely available in the `TADCompare` R package.

In general, the methods in this work provide a comprehensive set of tools for end-to-end analysis of Topologically Associated Domains (TADs). Each of the tools are designed to run on the highly complex, high resolution, data that will become more-and-more available as HiC sequencing technology improves. By publicly releasing these tools, we will make analysis of TADs exceptionally easier for the average scientist. To this end, we aimed to develop fast, user-friendly, tools that provide simple to interpret and accurate results.

7.2 Future Work

7.2.1 SpectralRep

Oftentimes, it is of interest to measure the similarity of entire contact matrices. This is useful for comparing technical or biological replicates, and assessing the consistency of Hi-C sequencing. Calculating similarity is complicated by the need to capture the topological similarity of contact matrices and certain issues such as distance decay and

noise. Previous methods have attempted to quantify this similarity, such as Quasar-Rep [155], Hi-C-Spector[113], Hi-CRep [151] and GenomeDISCO [153]. Both the Quasar-Rep and approach takes two contact matrices, transforms them by stratifying based on distance and then takes a pearson correlation. GenomeDISCO and Hi-C-Spector are graph-based methods, with GenomeDISCO measuring similarity based on the distance between graph-difussion smoothed contact matrices and Hi-C-Spector measuring the distance between eigenvectors of the Laplacians of each contact matrix. We propose a method based on cross-correlation of boundaries to calculate similarity.

SpectralRep will build off of results in this work which showed shifted, defined as boundaries that shift less than 5 bins TAD, boundaries are biologically identical to non-differential boundaries. Shifted boundaries can essentially be thought of as sequencing noise. At its basis, SpectralRep is the cross-correlation between eigenvector gaps. However, we decompose it such that at each point the value is the minimum distance among the 5 closest loci. We define the function below:

Given two contact matrices P and R with boundary scores of B_P and B_R having means of μ_P and μ_R respectively, we can calculate matrix similarity using a modified form of cross-correlation that takes into account small shifts in TAD boundaries. For this example, we first define the basic cross-correlation of two sets of boundary scores with lag l:

$$S = \frac{\sum[(B_P(i) - \mu_P) * (B_R(i-l) - \mu_R)]}{\sqrt{\sum(B_P(i) - \mu_P)^2}\sqrt{\sum(B_R(i-l) - \mu_R)^2}}$$

where i is the corresponding vector entry and l is the lag. Using this framework, we incorporate the fact that shifts of less than 5 are biologically insignificant by replacing all instances of $B_R(i-l)$ with $B_{min}(i) = min(|(B_P(i) - \mu_P) - (B_R(i-l) - \mu_R)|)$ for l = 1,...,5. The resulting formula is the following:

$$S_{adj} = \frac{\sum_i[(B_P(i) - \mu_P) * (B_{min}(i) - \mu_R)]}{\sqrt{\sum_i(B_P(i) - \mu_P)^2}\sqrt{\sum_i(B_{min}(i) - \mu_R)^2}}$$

S_{adj} gives us a measure of similarity between contact matrices that gives equal weighting to lagged and non-lagged TAD boundaries, finding the region of maximum similarity. This score effectively provides a measure of similarity between contact matrices purely in terms of TAD similarity. The adjustment allows us to treat shifted boundaries the same as non-shifted boundaries and more accurately capture contact matrix similarity.

Using this formula, we can provide a single score for matrix similarity based purely on TAD configuration. This can be used for a variety purposes such as quality control (Determining how replicate similarity and identifying outliers) or quantification of boundary change between conditions (cancer vs. non-cancer).

7.2.2 Extensions of windowed spectral clustering

Future work will involve developing a more general package for windowed spectral clustering and one that implements simple eigenvector gap spectral clustering without a sliding window. These methods can be applied to any data that has a structure that can be thought of as coordinates such as image data or spatial data. It can be used to directly seperate geographic regions or detect edges in images, defined by abrupt changes in pixel intensity. To date, we have begun creating packages for some of the methods for fast eigendecompositon are implemented in an additional R package called FastPCA (https://github.com/cresswellkg/fastPCA). This package allows for much faster principal component analysis than other PCA packages in R with comparable results. In the future, we intend to submit this package to CRAN.

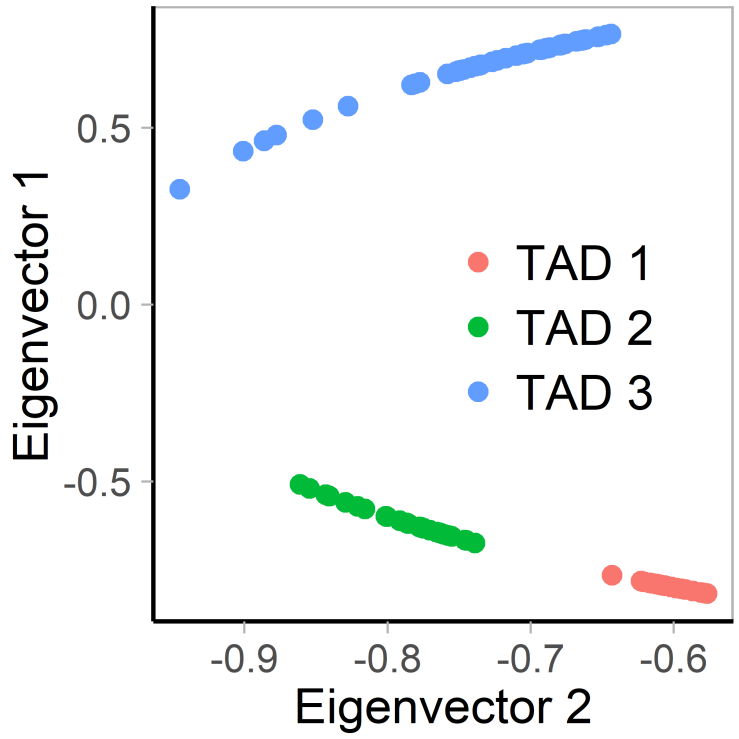
7.2.3 TAD Plotting

In the course of this work, we have developed methods for TAD visualization. To date, there is exactly one R-based function for plotting TADs and it is embedded in rGMAP [54], a TAD caller package. This function has many drawbacks and only works on their particular data form, making it essentially unsuable with results from other R packages. We intend to release our TAD plotting methods as a standalone R package called `TADPlotter`.

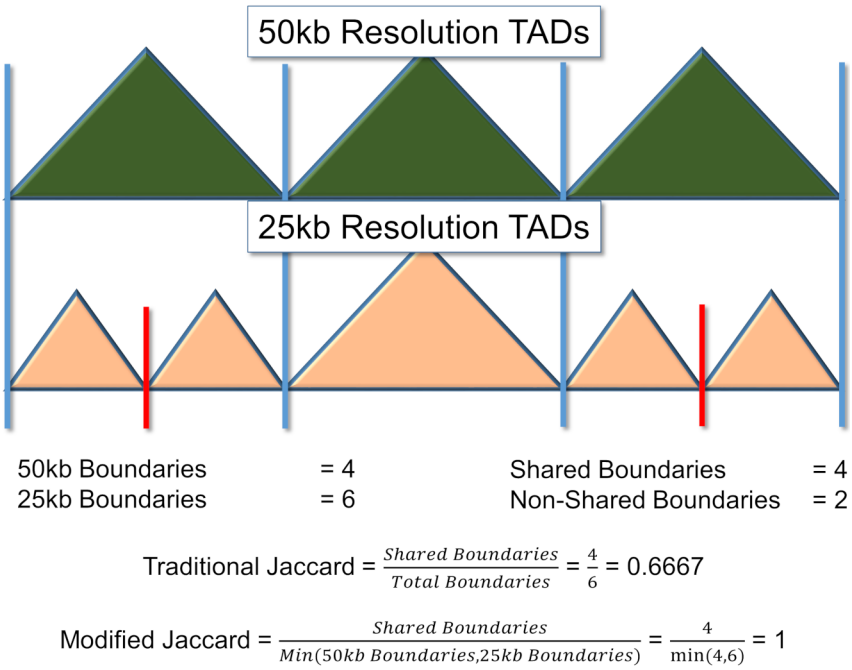
`TADPlotter` will be ggplot2 based, using a novel method that allows us to rotate heatmaps matching the conventional triangular visualization of TADs. To date, there are no packages allowing for customizable, triangular, heatmaps in R. Once completed, this package will be released on Bioconductor. Supplementary Figure 10 is an example `TADPlotter` output.

8 Appendix

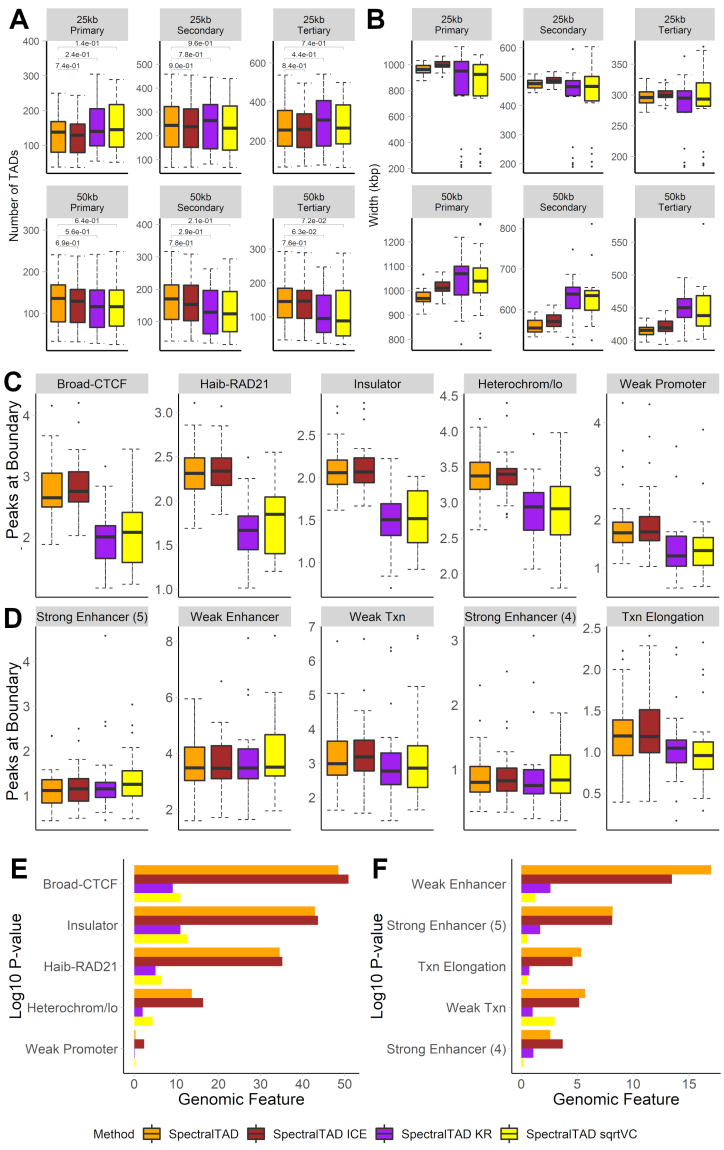
8.1 Supplementary Figures



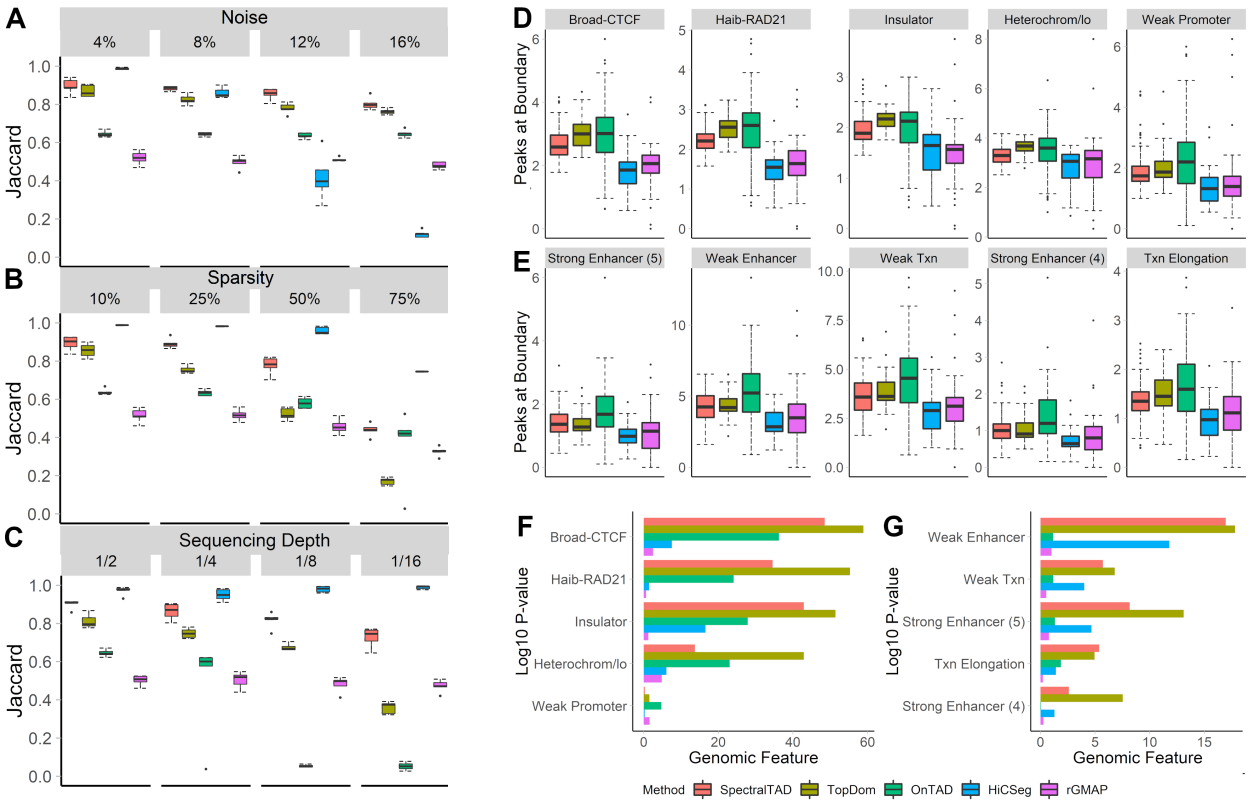
Supplementary Figure S1. Projection of eigenvectors on the unit circle. This projection allows us to identify TADs based on the distance between eigenvectors. The two largest gaps are used to seperate TAD 1, TAD 2 and TAD 3. We can also see the difference between a strongly organized group with close together points (TAD 1 and TAD 2) and a weaker group with more spread out points (TAD 3). Simulated data from [91].



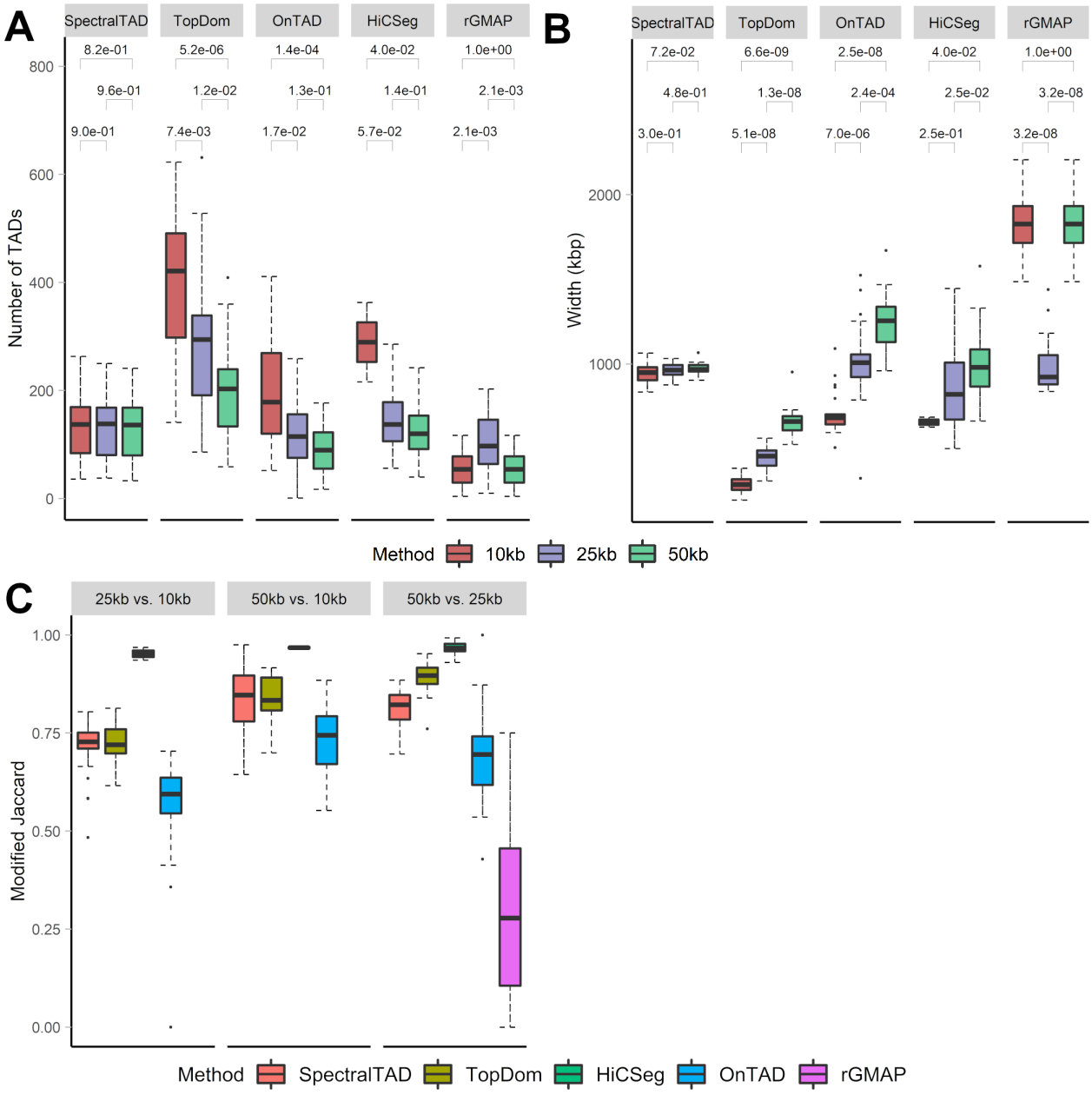
Supplementary Figure S2. Example of modified Jaccard statistics to measure agreement between TAD boundaries detected at different resolutions. The top triangles indicate TADs detected at 50kb resolution, while the bottom triangles indicate those detected at 25kb resolution. There are four shared boundaries (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.



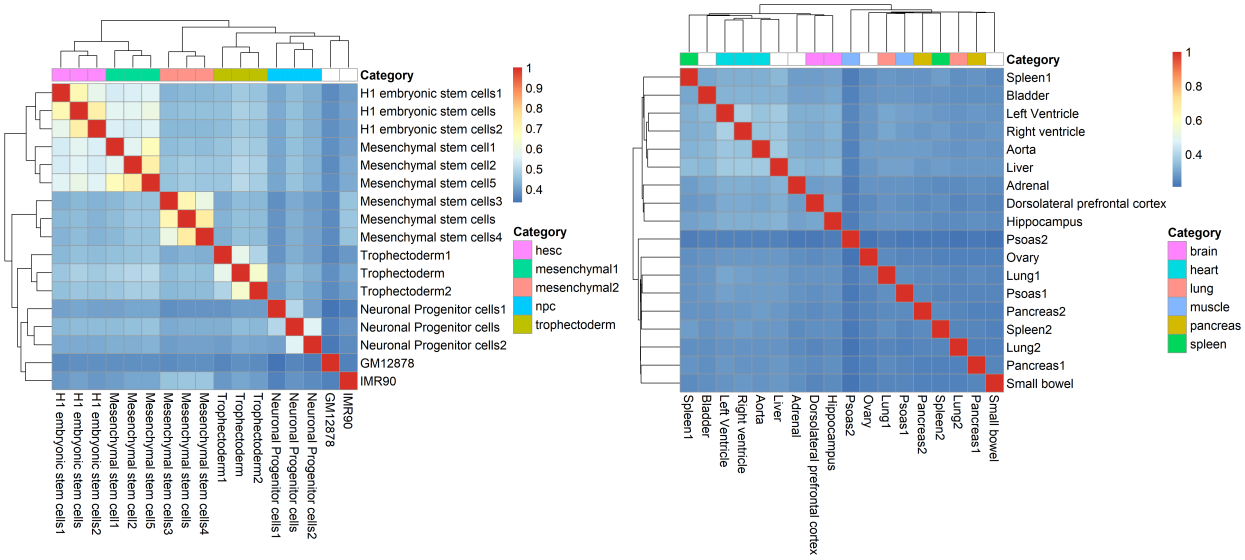
Supplementary Figure S3. The effect of data normalization on the average number (A) and width in kilobases (B) of TADs and the average number of peaks in enriched markers (C) & depleted markers (D), enrichment (F) and depletion (G) for different genomic annotations. Counts (A) and widths (B) for raw, KR-, ICE- and sqrtVC-normalized GM12878 data at 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 for enriched (D) and depleted (E) genomic features and the permutation p-values corresponding to enrichment (F), and depletion (G) for the top most enriched/depleted genomic annotations (permutation test) at TAD boundaries for GM12878 data at 50kb resolution are shown.



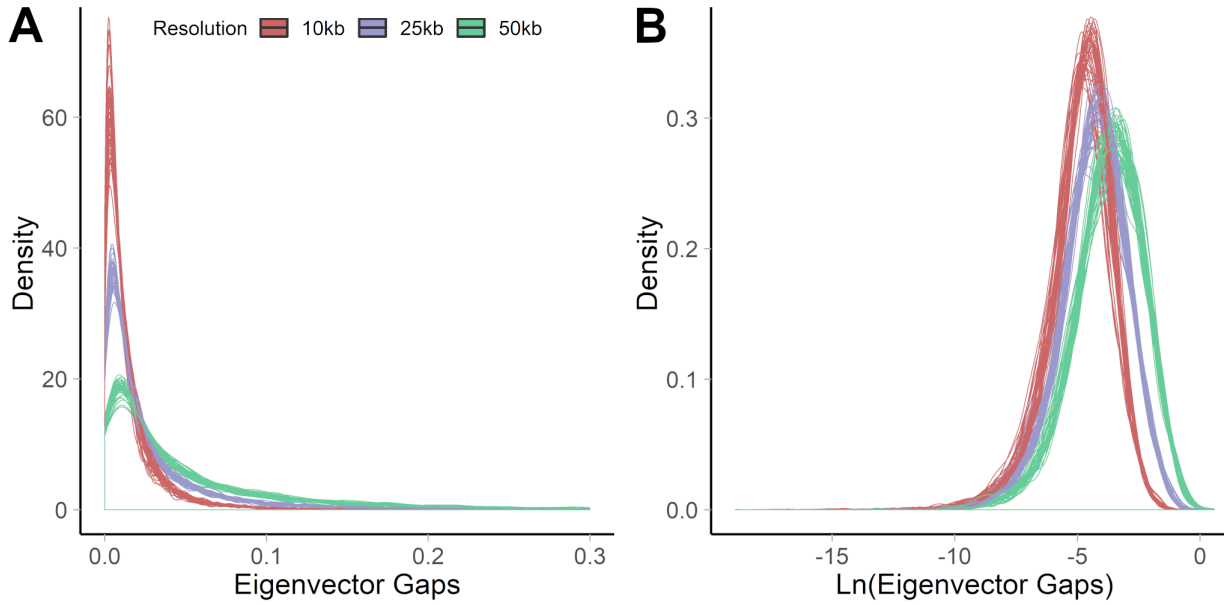
Supplementary Figure S4. The comparison of SpectralTAD and other TAD callers regarding TAD consistency and biological significance. To test for robustness to noise, sparsity, and downsampling, TADs were called from simulated Hi-C matrices using SpectralTAD and other TAD callers. The TAD boundaries were extended by 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 for enriched (D) and depleted (E) genomic features and the permutation p-values corresponding to enrichment (F), and depletion (G) for the top five most enriched/depleted genomic annotations are shown. Results averaged across chromosome 1-22 are shown.



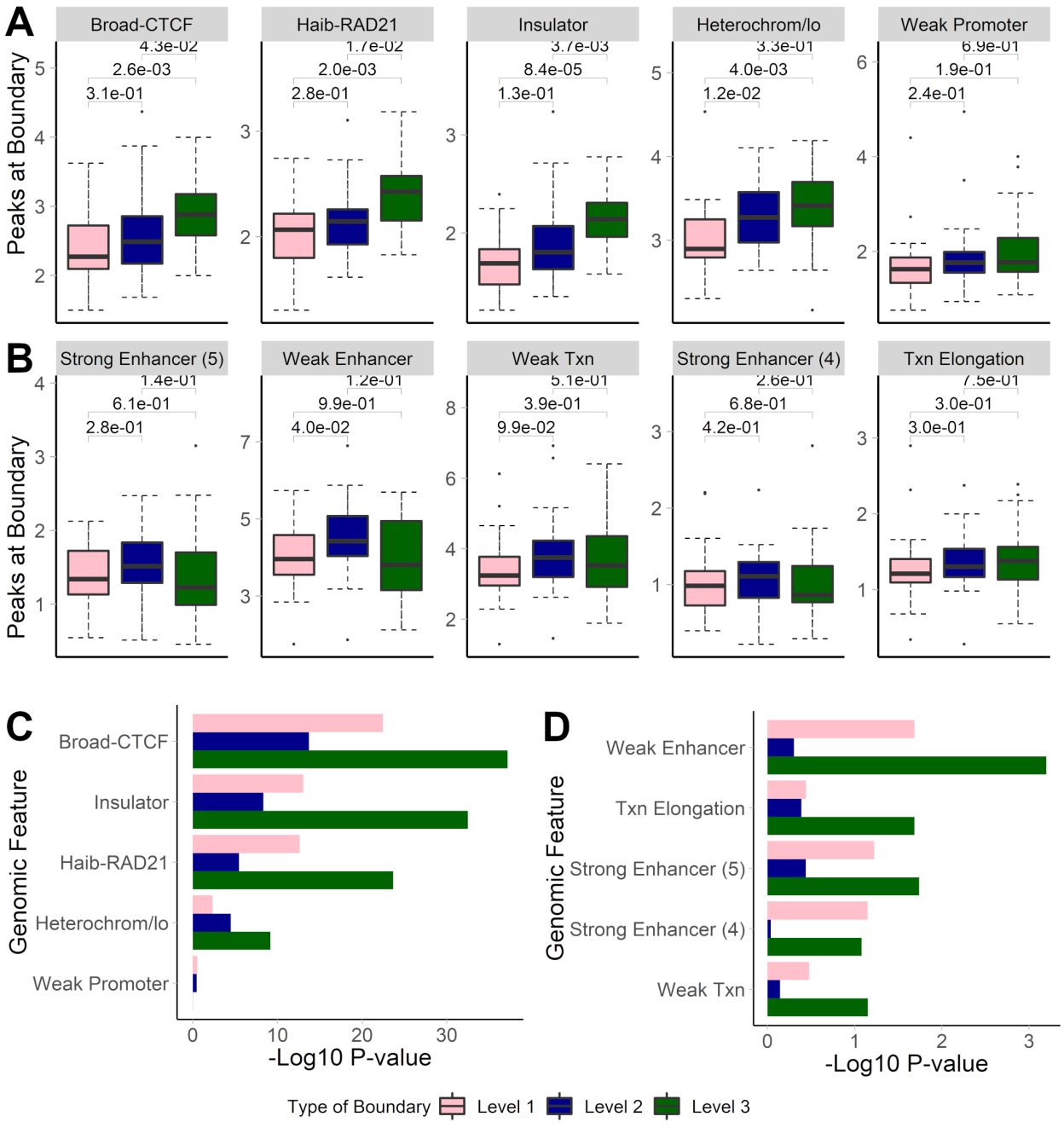
Supplementary Figure S5. The number, width, and consistency of TADs called across resolutions and primary vs. replicate for different methods. The average number (A) and width (B) of TADs across resolutions, Jaccard similarity between TAD boundaries detected from primary and replicate data and modified Jaccard similarity between TAD boundaries detected from data at 10kb, 25kb and 50kb resolutions (C). Wilcoxon test p-values are shown. Data from GM12878 cell line, chromosomes 1-22.



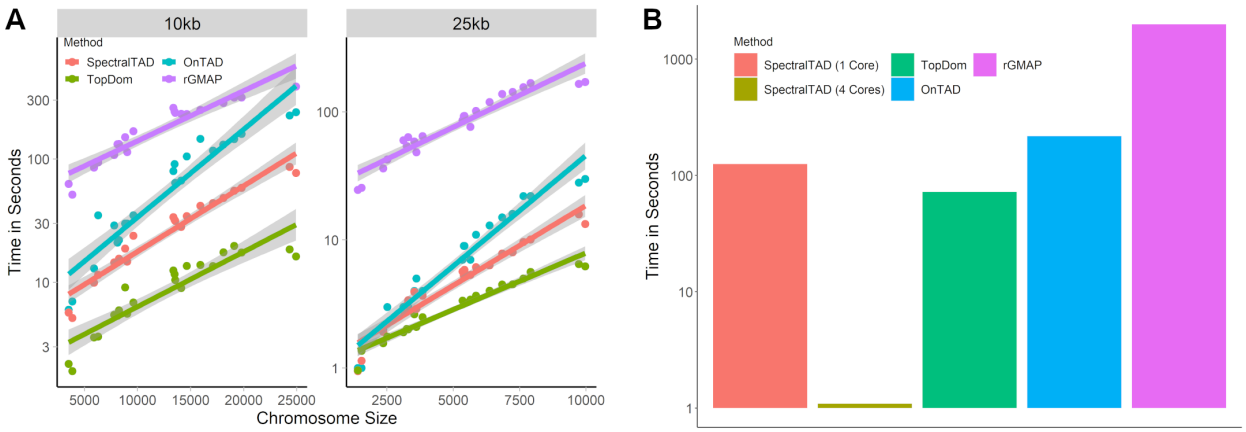
Supplementary Figure S6. Jaccard similarity of TAD boundaries across cell types (A) and tissues (B). TADs were called using SpectralTAD. Clustering was performed using Ward clustering applied to a Jaccard distance matrix. All TADs were called on raw 40kb data from [108]. Various cell-lines and tissues are used.



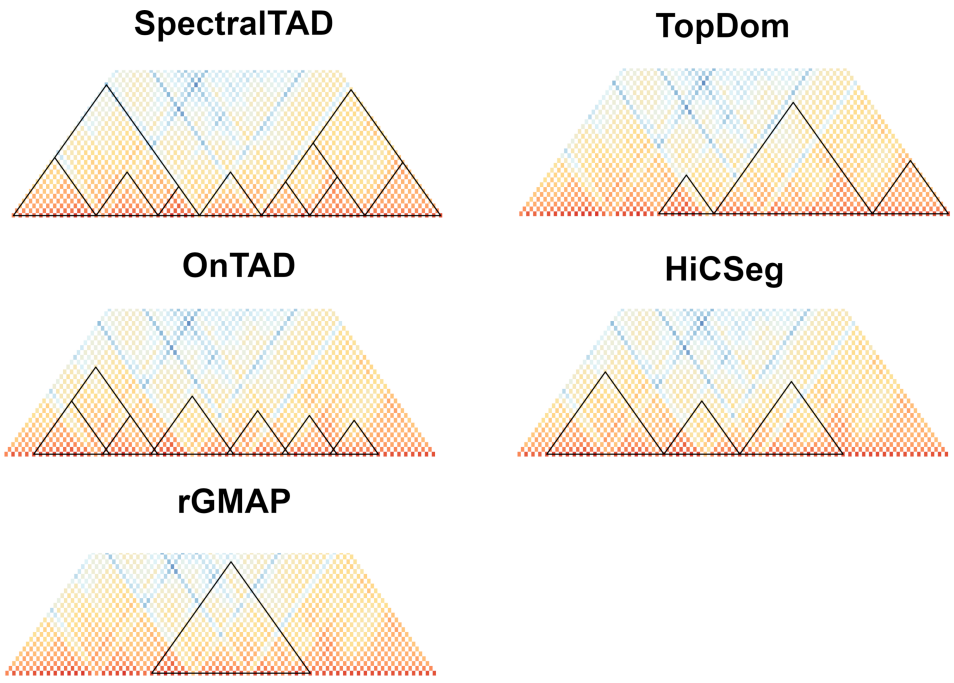
Supplementary Figure S7. Distribution of eigenvector gaps. The distributions of eigenvector gaps are plotted separately for each 10kb, 25kb and 50kb contact matrix from [3], 131 chromosome-specific datasets total. Results are colored by resolution. Higher resolution data shows smaller overall gaps due to a larger number of regions of high sparsity. The untransformed eigenvector gaps (A) and the natural log eigenvector gaps (B) are shown. `MASS::fitdistr()` function was used to establish the best fit by a lognormal (67 datasets) or a Weibull (64 datasets) distributions with similar log-likelihoods. The lognormal fit was chosen to model the distribution of log eigenvector gaps.



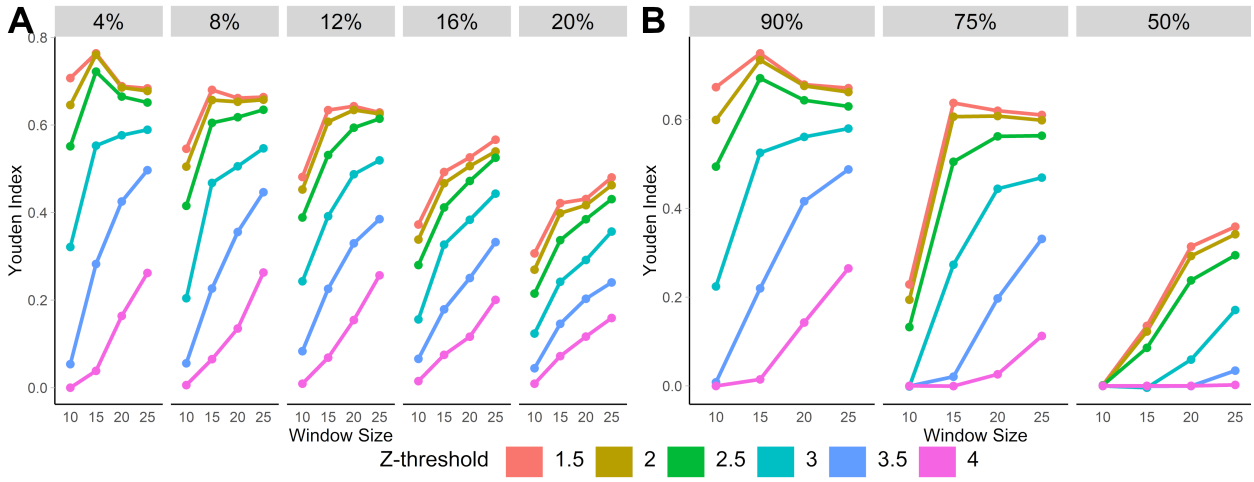
Supplementary Figure S8. The effect of the hierarchy of TAD boundaries detected by SpectralTAD on the average number of annotations in enriched (A) & depleted (B) genomic markers and on enrichment (C) and depletion (D) for different genomic annotations. Results for TAD boundaries detected as level 1, 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 & B and aggregated p-values, using the Fisher's method, are shown for panels C & D. Raw data from GM12878 cell line, chromosome 1-22, 50kb resolution.



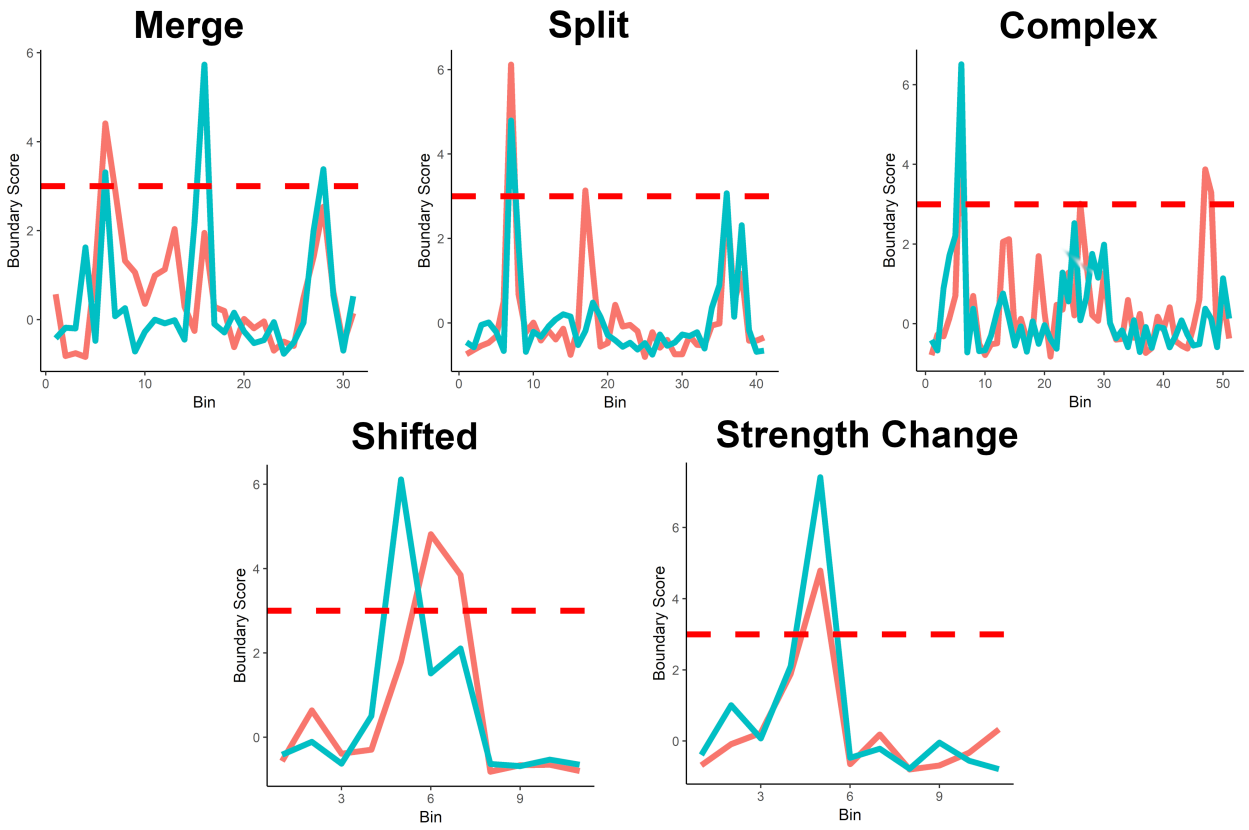
Supplementary Figure S9. Runtime performance of various TAD callers. TADs were called using data from GM12878 cell line at 10kb and 25kb resolution, and runtimes recorded. (A) Runtimes were summarized across different chromosomes. Each dot represents chromosome-specific run time averaged across three 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. Results for HiCSeg are excluded due to exceptionally slow runtimes (24+ hours for one 10kb chromosome).



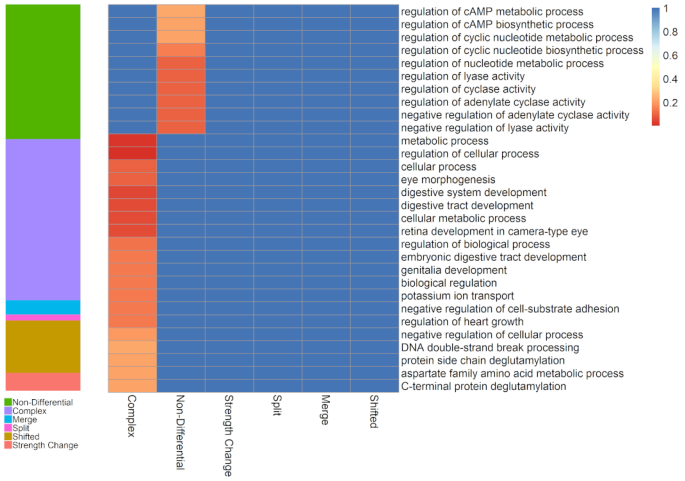
Supplementary Figure S10. Examples of TADs detected by different TAD callers. TADs detected by SpectralTAD, TopDom, OnTAD, HiCSeg and rGMAP. Darker red colors indicate a higher level of connectivity while lighter blues indicate less connectivity; triangles indicate TADs. Data are shown from [3], resolution 50kb, chr22 (Coordinates 23650000:26750000). All parameters were set according to the instructions of each package for analyzing 50kb resolution data with no normalization performed.



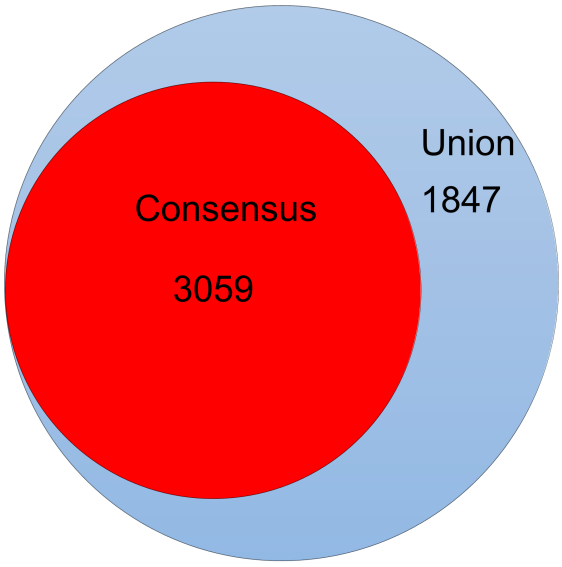
Supplementary Figure S11. Window size of 15 units of Hi-C data resolution and TAD boundary score cutoff of 2 yield consistent TAD boundary detection. Differential TAD boundaries were compared between two simulated data sets with window size sizes ranging from 10 to 25 and boundary score cutoff ranging from 1.5 to 4. Youden index (balanced sensitivity and specificity metric) was calculated for each combination and plotted to show agreement with ground-truth annotations. Results are shown for noise-injected matrices (A) and sparsity-injected matrices (B).



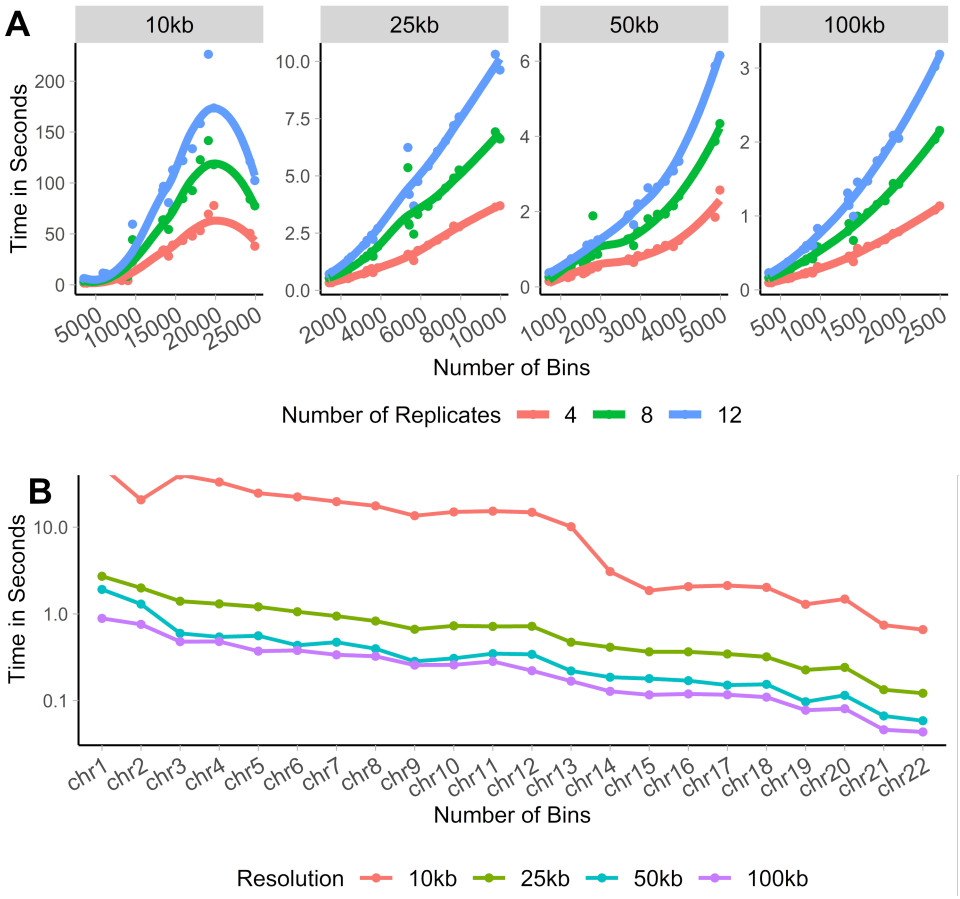
Supplementary Figure S12. Visualization of different types of boundary score patterns. Patterns of raw boundary scores are shown for 5 different types of differential boundaries (Merge, split, complex, shifted and strength change). The red horizontal line corresponds to the minimum cutoff for a TAD boundary (40kb resolution, human neural progenitor cell line, data from [108]). Data from chromosome 22 with the most representative examples chosen.



Supplementary Figure S13. Heatmap of gene ontology enrichment at the first and last time point in auxin-treated data. Differential boundary identification was performed on auxin-treated data at the time of application (first time point) and complete withdrawal (last time point) (HCT-116 cell line, chr1-22, 40kb resolution, [108]). A barplot of the proportion of each boundary type (A) and FDR-adjusted hypergeomtric p-values (B) obtained from gene ontology enrichment analysis using `rGREAT` (See methods) are shown. The top 30 pathways, in terms of average enrichment, are shown and clustered using Ward clustering.



Supplementary Figure S14. Venn diagram of union and consensus TAD counts. Consensus and union TADs were called across four different cell lines (hesc, mesynchymal, npc, trophectoderm) and the number of union and consensus TADs were recorded. The venn diagram shows the complete overlap of consensus TADs within union TADs. (40kb resolution, data from [108])



Supplementary Figure S15. Runtime of TADCompare. Plot containing the runtime of two-way comparison (A) and consensus TADs called on 4, 8, 12, and 16 replicates (B). Each point represent the runtime for a specific chromosome. X-axis - Chromosome, Y-axis - Runtime in seconds. Hi-C data from [3], chr 1-22, 10kb, 25kb, 50kb, and 100kb resolution.

8.2 Supplementary Tables

Supplementary Table S1. Hi-C Data sources. Information about experimental [3,108] and simulated [91] Hi-C data.

Supplementary Table S2. Experimental Data sources. Genome annotation (hg19/GRCh37) [109] data for GM12878 cell line used in the analysis, sorted by category, then by data type.

Supplementary Table S3. Summary of gaps. The percentage of gaps is summarized for all chromosomes at 10kb, 25kb, and 50kb resolution using raw GM12878 data [3]. Gaps are separated based on whether they are centromeric or other (unsequenced, or poorly organized chromatin).

Supplementary Table S4. Enrichment by Method. Enrichment/Depletion results are provided for all genomic annotations tested. Permutation p-values summarized using Fisher's method are shown. Data is sorted alphabetically by category and then by genomic annotation.

Supplementary 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.

Supplementary Table S6. Contact matrix data sources. The source of all contact matrices, experimental and simulated, used in this paper are provided. Experimental data are seperated based on study and cell line.

Supplementary Table S7. Genomic annotation data sources. The sources, with download links, for all genomic annotation used in this paper are included.

Supplementary Table S8. Summary of differential boundary types across tissues and cell-lines. The percentage of each type of differential boundary for all tissue-tissue and

cell line-cell line comparisons is reported. Results are aggregated over all chromosomes. Hi-C data from Schmitt et al. [108], 40kb resolution, chr 1-22.

Supplementary Table S9. Gene ontology enrichment for differential boundary types. Differential boundaries were identified between the neural progenitor cells (NPC) and mesenchymal stem cells (MSC) [108]. Pathway analysis was performed using `rGREAT` (Methods) and results are seperated by ontology. Boundaries with an FDR adjusted p-value of <0.3 are shown. 40kb resolution, chr1-22.

Supplementary Table S10. Gene ontology enrichment between the first and last time point in auxin-treated data . Differential boundaries were identified between the first and last time point of auxin-treated data [142]. Pathway analysis was performed using `rGREAT` (Methods) and results are seperated by ontology. Boundaries with an FDR adjusted p-value of <0.3 are shown. 50kb resolution, chr1-22.

Supplementary Table S11. Enrichment across different temporal boundary types. Temporal boundary types were identified across four time points in auxin-treated data [142]. Results are shown for four types of temporal TAD (Early Appearing, Late Appearing, Highly Common, Dynamic). Permutation p-values, along with enrichment or depletion designations, are reported. HCT-116 cell line, 40kb resolution, chr 1-22.

Supplementary Table S12. Gene ontology enrichment for different temporal boundary types. Temporal boundary types were identified across four time points in auxin-treated data [142]. For each temporal boundary type, pathway analysis was performed using `rGREAT` (Methods) and results are seperated by ontology. Boundaries with an FDR adjusted p-value of <0.3 are shown. HCT-116 cell line, 50kb resolution, chr1-22.

Supplementary Table S13. Enrichment across different consensus scores. Consensus scores were called across 17 contact matrices representing 7 different cell lines. Results were dichotomized into three groups (<2, 2-4, >4) based on consensus boundary scores. Permutation p-values, along with enrichment or depletion designations, are reported. Hi-C data from Schmitt et al. [108], 40kb resolution, chr 1-22.

References

1. Dekker J, Rippe K, Dekker M, Kleckner N: **Capturing chromosome conformation**. *Science* 2002, **295**:1306–1110.1126/science.1067799.

2. Lieberman-Aiden E, Berkum NL van, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome**. *Science* 2009, **326**:289–9310.1126/science.1181369.

3. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping**. *Cell* 2014, **159**:1665–8010.1016/j.cell.2014.11.021.

4. Schmitt AD, Hu M, Ren B: **Genome-wide mapping and analysis of chromosome architecture**. *Nat Rev Mol Cell Biol* 2016, **17**:743–75510.1038/nrm.2016.104.

5. Berkum NL van, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES: **Hi-c: A method to study the three-dimensional architecture of genomes**. *J Vis Exp* 2010, 10.3791/1869.

6. Phillips-Cremins JE, Corces VG: **Chromatin insulators: Linking genome organization to cellular function**. *Mol Cell* 2013, **50**:461–7410.1016/j.molcel.2013.04.018.

7. Ji X, Dadon DB, Powell BE, Fan ZP, Borges-Rivera D, Shachar S, Weintraub AS, Hnisz D, Pegoraro G, Lee TI, Misteli T, Jaenisch R, Young RA: **3D chromosome regulatory landscape of human pluripotent cells**. *Cell Stem Cell* 2016, **18**:262–7510.1016/j.stem.2015.11.007.

8. Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S: **Comparative hi-c reveals that ctcf underlies evolution of chromosomal domain architecture**. *Cell Rep* 2015, **10**:1297–30910.1016/j.celrep.2015.02.004.

9. Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, Weintraub AS,

Schuijers J, Lee TI, Zhao K, Young RA: **Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes**. *Cell* 2014, **159**:374–8710.1016/j.cell.2014.09.030.

10. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen C-A, Schmitt AD, Espinoza CA, Ren B: **A high-resolution map of the three-dimensional chromatin interactome in human cells**. *Nature* 2013, **503**:290–410.1038/nature12644.

11. Gierman HJ, Indemans MHG, Koster J, Goetze S, Seppen J, Geerts D, Driel R van, Versteeg R: **Domain-wide regulation of gene expression in the human genome.** *Genome research* 2007, **17**:1286–129510.1101/gr.6276007.

12. Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, Vera DL, Wang Y, Hansen RS, Canfield TK, Thurman RE, Cheng Y, Gülsoy G, Dennis JH, Snyder MP, Stamatoyannopoulos JA, Taylor J, Hardison RC, Kahveci T, Ren B, Gilbert DM: **Topologically associating domains are stable units of replication-timing regulation**. *Nature* 2014, **515**:402–510.1038/nature13986.

13. Jhunjhunwala S, Zelm MC van, Peak MM, Murre C: **Chromatin architecture and the generation of antigen receptor diversity**. *Cell* 2009, **138**:435–4810.1016/j.cell.2009.07.016.

14. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, Berkum NL van, Meisig J, Sedat J, Gribnau J, Barillot E, Blüthgen N, Dekker J, Heard E: **Spatial partitioning of the regulatory landscape of the x-inactivation centre**. *Nature* 2012, **485**:381–510.1038/nature11049.

15. Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, Meyer BJ: **Condensin-driven remodelling of x chromosome topology during dosage compensation**. *Nature* 2015, **523**:240–410.1038/nature14450.

16. Taberlay PC, Achinger-Kawecka J, Lun ATL, Buske FA, Sabir K, Gould CM, Zotenko E, Bert SA, Giles KA, Bauer DC, Smyth GK, Stirzaker C, O'Donoghue SI, Clark SJ: **Three-dimensional disorganization of the cancer genome occurs co-**

incident with long-range genetic and epigenetic alterations. *Genome Res* 2016, **26**:719–3110.1101/gr.201517.115.

17. Lupiáñez DG, Spielmann M, Mundlos S: **Breaking tads: How alterations of chromatin domains result in disease**. *Trends Genet* 2016, **32**:225–3710.1016/j.tig.2016.01.003.

18. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: **Topological domains in mammalian genomes identified by analysis of chromatin interactions**. *Nature* 2012, **485**:376–8010.1038/nature11082.

19. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G: **Three-dimensional folding and functional organization principles of the drosophila genome**. *Cell* 2012, **148**:458–7210.1016/j.cell.2012.01.010.

20. Jackson DA, Pombo A: **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* 1998, **140**:1285–95.

21. Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot J-P, Tanay A, Cavalli G: **Multiscale 3D genome rewiring during mouse neural development**. *Cell* 2017, **171**:557–572.e2410.1016/j.cell.2017.09.043.

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

23. Sofueva S, Yaffe E, Chan W-C, Georgopoulou D, Vietri Rudan M, Mira-Bontenbal H, Pollard SM, Schroth GP, Tanay A, Hadjur S: **Cohesin-mediated interactions organize chromosomal domain architecture**. *EMBO J* 2013, **32**:3119–2910.1038/emboj.2013.237.

24. Splinter E, Heath H, Kooren J, Palstra R-J, Klous P, Grosveld F, Galjart N, Laat W de: **CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus**. *Genes Dev* 2006, **20**:2349–5410.1101/gad.399506.

25. Phillips JE, Corces VG: **CTCF: Master weaver of the genome**. *Cell* 2009, **137**:1194–21110.1016/j.cell.2009.06.001.

26. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J, Ruszczycki B, Michalski P, Piecuch E, Wang P, Wang D, Tian SZ, Penrad-Mobayed M, Sachs LM, Ruan X, Wei C-L, Liu ET, Wilczynski GM, Plewczynski D, Li G, Ruan Y: **CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription**. *Cell* 2015, **163**:1611–2710.1016/j.cell.2015.11.024.

27. Beagan JA, Duong MT, Titus KR, Zhou L, Cao Z, Ma J, Lachanski CV, Gillis DR, Phillips-Cremins JE: **YY1 and ctcf orchestrate a 3D chromatin looping switch during early neural lineage commitment**. *Genome Res* 2017, **27**:1139–115210.1101/gr.215160.116.

28. Sanyal A, Lajoie BR, Jain G, Dekker J: **The long-range interaction landscape of gene promoters**. *Nature* 2012, **489**:109–1310.1038/nature11279.

29. Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CWH, Ye C, Ping JLH, Mulawadi F, Wong E, Sheng J, Zhang Y, Poh T, Chan CS, Kunarso G, Shahab A, Bourque G, Cacheux-Rataboul V, Sung W-K, Ruan Y, Wei C-L: **CTCF-mediated functional chromatin interactome in pluripotent cells**. *Nat Genet* 2011, **43**:630–810.1038/ng.857.

30. Corces MR, Corces VG: **The three-dimensional cancer genome**. *Curr Opin Genet Dev* 2016, **36**:1–710.1016/j.gde.2016.01.002.

31. 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, Grosveld FG, Ren B, Wendt KS: **Cohesin and ctcf differentially affect chromatin architecture and gene expression in human cells**. *Proc Natl Acad Sci U S A* 2014, **111**:996–100110.1073/pnas.1317788111.

32. Ciabrelli F, Cavalli G: **Chromatin-driven behavior of topologically associating domains**. *Journal of Molecular Biology* 2015, **427**:608–62510.1016/j.jmb.2014.09.013.

33. Hong S, Kim D: **Computational characterization of chromatin domain boundary-associated genomic elements**. *Nucleic Acids Research* 2017, **45**:10403–1041410.1093/nar/gkx738Available: https://doi.org/10.1093/nar/gkx738.

34. Filippova D, Patro R, Duggal G, Kingsford C: **Identification of alternative topological domains in chromatin**. *Algorithms Mol Biol* 2014, **9**:1410.1186/1748-7188-9-14.

35. Lévy-Leduc C, Delattre M, Mary-Huard T, Robin S: **Two-dimensional segmentation for analyzing hi-c data**. *Bioinformatics* 2014, **30**:i386–9210.1093/bioinformatics/btu443.

36. Bonev B, Cavalli G: **Organization and function of the 3D genome**. *Nat Rev Genet* 2016, **17**:661–67810.1038/nrg.2016.112.

37. Oti M, Falck J, Huynen MA, Zhou H: **CTCF-mediated chromatin loops enclose inducible gene regulatory domains**. *BMC Genomics* 2016, **17**10.1186/s12864-016-2516-6.

38. Dixon JR, Gorkin DU, Ren B: **Chromatin domains: The unit of chromosome organization**. *Mol Cell* 2016, **62**:668–8010.1016/j.molcel.2016.05.018.

39. Luzhin AV, Flyamer IM, Khrameeva EE, Ulianov SV, Razin SV, Gavrilov AA: **Quantitative differences in tad border strength underly the tad hierarchy in drosophila chromosomes**. *J Cell Biochem* 2018, 10.1002/jcb.27737.

40. Weinreb C, Raphael BJ: **Identification of hierarchical chromatin domains**. *Bioinformatics* 2016, **32**:1601–910.1093/bioinformatics/btv485.

41. Fraser J, Ferrai C, Chiariello AM, Schueler M, Rito T, Laudanno G, Barbieri M, Moore BL, Kraemer DCA, Aitken S, Xie SQ, Morris KJ, Itoh M, Kawaji H, Jaeger I, Hayashizaki Y, Carninci P, Forrest ARR, FANTOM Consortium, Semple CA, Dostie J, Pombo A, Nicodemi M: **Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation**. *Mol Syst Biol* 2015, **11**:852.

42. Gibcus JH, Dekker J: **The hierarchy of the 3D genome**. *Mol Cell* 2013, **49**:773–8210.1016/j.molcel.2013.02.011.

43. Hansen AS, Cattoglio C, Darzacq X, Tjian R: **Recent evidence that TADs and chromatin loops are dynamic structures**. *Nucleus* 2017, **9**:20–3210.1080/19491034.2017.1389365.

44. Phillips-Cremins JE, Sauria ME, Sanyal A, Gerasimova TI, Lajoie BR, Bell JS, Ong C-T, Hookway TA, Guo C, Sun Y, Bland MJ, Wagstaff W, Dalton S, McDevitt TC, Sen R, Dekker J, Taylor J, Corces VG: **Architectural protein subclasses shape 3D organization of genomes during lineage commitment**. *Cell* 2013, **153**:1281–129510.1016/j.cell.2013.04.053.

45. Dong Q, Li N, Li X, Yuan Z, Xie D, Wang X, Li J, Yu Y, Wang J, Ding B, Zhang Z, Li C, Bian Y, Zhang A, Wu Y, Liu B, Gong L: **Genome-wide hi-c analysis reveals extensive hierarchical chromatin interactions in rice**. *Plant J* 2018, **94**:1141–115610.1111/tpj.13925.

46. Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, Zhou XJ: **TopDom: An efficient and deterministic method for identifying topological domains in genomes**. *Nucleic Acids Res* 2016, **44**:e7010.1093/nar/gkv1505.

47. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL: **Juicer provides a one-click system for analyzing loop-resolution hi-c experiments**. *Cell Syst* 2016, **3**:95–810.1016/j.cels.2016.07.002.

48. Berlivet S, Paquette D, Dumouchel A, Langlais D, Dostie J, Kmita M: **Clustering of tissue-specific sub-TADs accompanies the regulation of HoxA genes in developing limbs**. *PLoS Genetics* 2013, **9**:e100401810.1371/journal.pgen.1004018.

49. Dali R, Blanchette M: **A critical assessment of topologically associating domain prediction tools**. *Nucleic Acids Res* 2017, **45**:2994–300510.1093/nar/gkx145.

50. Dekker J, Marti-Renom MA, Mirny LA: **Exploring the three-dimensional organization of genomes: Interpreting chromatin interaction data**. *Nat Rev Genet* 2013, **14**:390–40310.1038/nrg3454.

51. Wang Y, Sarkar P, Ursu O, Kundaje A, Bickel PJ: **Network modelling of topological domains using hi-c data**. *arXiv preprint arXiv:1707.09587* 2017,.

52. Chen J, Hero AO 3rd, Rajapakse I: **Spectral identification of topological domains**. *Bioinformatics* 2016, **32**:2151–810.1093/bioinformatics/btw221.

53. Naumova N, Imakaev M, Fudenberg G, Zhan Y, Lajoie BR, Mirny LA, Dekker J: **Organization of the mitotic chromosome**. *Science* 2013, **342**:948–5310.1126/science.1236083.

54. Yu M, Ren B: **The three-dimensional organization of mammalian genomes**. *Annu Rev Cell Dev Biol* 2017, **33**:265–28910.1146/annurev-cellbio-100616-060531.

55. Li L, Lyu X, Hou C, Takenaka N, Nguyen HQ, Ong C-T, Cubeñas-Potts C, Hu M,

Lei EP, Bosco G, Qin ZS, Corces VG: **Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing**. *Mol Cell* 2015, **58**:216–3110.1016/j.molcel.2015.02.023.

56. Narendra V, Bulajić M, Dekker J, Mazzoni EO, Reinberg D: **CTCF-mediated topological boundaries during development foster appropriate gene regulation**. *Genes Dev* 2016, **30**:2657–266210.1101/gad.288324.116.

57. Nagano T, Lubling Y, Várnai C, Dudley C, Leung W, Baran Y, Mendelson Cohen N, Wingett S, Fraser P, Tanay A: **Cell-cycle dynamics of chromosomal organization at single-cell resolution**. *Nature* 2017, **547**:61–6710.1038/nature23001.

58. Hug CB, Grimaldi AG, Kruse K, Vaquerizas JM: **Chromatin architecture emerges during zygotic genome activation independent of transcription**. *Cell* 2017, **169**:216–228.e1910.1016/j.cell.2017.03.024.

59. Flyamer IM, Gassler J, Imakaev M, Brandão HB, Ulianov SV, Abdennur N, Razin SV, Mirny LA, Tachibana-Konwalski K: **Single-nucleus hi-c reveals unique chromatin reorganization at oocyte-to-zygote transition**. *Nature* 2017, **544**:110–11410.1038/nature21711.

60. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, Diao Y, Liang J, Zhao H, Lobanenkov VV, Ecker JR, Thomson JA, Ren B: **Chromatin architecture reorganization during stem cell differentiation**. *Nature* 2015, **518**:331–610.1038/nature14222.

61. Du Z, Zheng H, Huang B, Ma R, Wu J, Zhang X, He J, Xiang Y, Wang Q, Li Y, Ma J, Zhang X, Zhang K, Wang Y, Zhang MQ, Gao J, Dixon JR, Wang X, Zeng J, Xie W: **Allelic reprogramming of 3D chromatin architecture during early mammalian development**. *Nature* 2017, **547**:232–23510.1038/nature23263.

62. Ke Y, Xu Y, Chen X, Feng S, Liu Z, Sun Y, Yao X, Li F, Zhu W, Gao L, Chen H, Du Z, Xie W, Xu X, Huang X, Liu J: **3D chromatin structures of mature gametes and structural reprogramming during mammalian embryogenesis**. *Cell* 2017, **170**:367–

381.e2010.1016/j.cell.2017.06.029.

63. Zhang Y, Xiang Y, Yin Q, Du Z, Peng X, Wang Q, Fidalgo M, Xia W, Li Y, Zhao Z-A, Zhang W, Ma J, Xu F, Wang J, Li L, Xie W: **Dynamic epigenomic landscapes during early lineage specification in mouse embryos**. *Nat Genet* 2018, **50**:96–10510.1038/s41588-017-0003-x.

64. Novo CL, Javierre B-M, Cairns J, Segonds-Pichon A, Wingett SW, Freire-Pritchett P, Furlan-Magaril M, Schoenfelder S, Fraser P, Rugg-Gunn PJ: **Long-range enhancer interactions are prevalent in mouse embryonic stem cells and are reorganized upon pluripotent state transition**. *Cell Rep* 2018, **22**:2615–262710.1016/j.celrep.2018.02.040.

65. Flavahan WA, Drier Y, Liau BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov AO, Suvà ML, Bernstein BE: **Insulator dysfunction and oncogene activation in idh mutant gliomas**. *Nature* 2016, **529**:110–410.1038/nature16490.

66. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA: **Formation of chromosomal domains by loop extrusion**. *Cell Rep* 2016, **15**:2038–4910.1016/j.celrep.2016.04.085.

67. Sanborn AL, Rao SSP, Huang S-C, Durand NC, Huntley MH, Jewett AI, Bochkov ID, Chinnappan D, Cutkosky A, Li J, Geeting KP, Gnirke A, Melnikov A, McKenna D, Stamenova EK, Lander ES, Aiden EL: **Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes**. *Proc Natl Acad Sci U S A* 2015, **112**:E6456–6510.1073/pnas.1518552112.

68. Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y, Lu Y, Wu Y, Jia Z, Li W, Zhang MQ, Ren B, Krainer AR, Maniatis T, Wu Q: **CRISPR inversion of ctcf sites alters genome topology and enhancer/promoter function**. *Cell* 2015, **162**:900–1010.1016/j.cell.2015.07.038.

69. Krijger PHL, Laat W de: **Regulation of disease-associated gene expression in the 3D genome**. *Nat Rev Mol Cell Biol* 2016, **17**:771–78210.1038/nrm.2016.138.

70. Misteli T: **Higher-order genome organization in human disease**. *Cold Spring*

Harb Perspect Biol 2010, **2**:a00079410.1101/cshperspect.a000794.

71. Spielmann M, Lupiáñez DG, Mundlos S: **Structural variation in the 3D genome**. *Nat Rev Genet* 2018, 10.1038/s41576-018-0007-0.

72. Ibn-Salem J, Köhler S, Love MI, Chung H-R, Huang N, Hurles ME, Haendel M, Washington NL, Smedley D, Mungall CJ, Lewis SE, Ott C-E, Bauer S, Schofield PN, Mundlos S, Spielmann M, Robinson PN: **Deletions of chromosomal regulatory boundaries are associated with congenital disease**. *Genome Biol* 2014, **15**:42310.1186/s13059-014-0423-1.

73. Mitelman F: **Recurrent chromosome aberrations in cancer**. *Mutat Res* 2000, **462**:247–53.

74. Valton A-L, Dekker J: **TAD disruption as oncogenic driver**. *Curr Opin Genet Dev* 2016, **36**:34–4010.1016/j.gde.2016.03.008.

75. Rickman DS, Soong TD, Moss B, Mosquera JM, Dlabal J, Terry S, MacDonald TY, Tripodi J, Bunting K, Najfeld V, Demichelis F, Melnick AM, Elemento O, Rubin MA: **Oncogene-mediated alterations in chromatin conformation**. *Proc Natl Acad Sci U S A* 2012, **109**:9083–810.1073/pnas.1112570109.

76. Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, Goldmann J, Lajoie BR, Fan ZP, Sigova AA, Reddy J, Borges-Rivera D, Lee TI, Jaenisch R, Porteus MH, Dekker J, Young RA: **Activation of proto-oncogenes by disruption of chromosome neighborhoods**. *Science* 2016, **351**:1454–810.1126/science.aad9024.

77. Gröschel S, Sanders MA, Hoogenboezem R, Wit E de, Bouwman BAM, Erpelinck C, Velden VHJ van der, Havermans M, Avellino R, Lom K van, Rombouts EJ, Duin M van, Döhner K, Beverloo HB, Bradner JE, Döhner H, Löwenberg B, Valk PJM, Bindels EMJ, Laat W de, Delwel R: **A single oncogenic enhancer rearrangement causes concomitant evi1 and gata2 deregulation in leukemia**. *Cell* 2014, **157**:369–38110.1016/j.cell.2014.02.019.

78. Barutcu AR, Lajoie BR, McCord RP, Tye CE, Hong D, Messier TL, Browne G, Wijnen AJ van, Lian JB, Stein JL, Dekker J, Imbalzano AN, Stein GS: **Chromatin interaction**

analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome Biol* 2015, **16**:21410.1186/s13059-015-0768-0.

79. Zhou Y, Gerrard DL, Wang J, Li T, Yang Y, Fritz AJ, Rajendran M, Fu X, Schiff R, Lin S, Frietze S, Jin VX: **Temporal dynamic reorganization of 3D chromatin architecture in hormone-induced breast cancer and endocrine resistance**. *Nat Commun* 2019, **10**:152210.1038/s41467-019-09320-9.

80. Fotuhi Siahpirani A, Ay F, Roy S: **A multi-task graph-clustering approach for chromosome conformation capture data sets identifies conserved modules of chromosomal interactions**. *Genome Biol* 2016, **17**:11410.1186/s13059-016-0962-8.

81. Wang X-T, Dong P-F, Zhang H-Y, Peng C: **Structural heterogeneity and functional diversity of topologically associating domains in mammalian genomes**. *Nucleic Acids Research* 2015, **43**:7237–724610.1093/nar/gkv684.

82. Serra F, Baù D, Goodstadt M, Castillo D, Filion GJ, Marti-Renom MA: **Automatic analysis and 3D-modelling of hi-c data using tadbit reveals structural features of the fly chromatin colors**. *PLoS Comput Biol* 2017, **13**:e100566510.1371/journal.pcbi.1005665.

83. Chen F, Li G, Zhang MQ, Chen Y: **HiCDB: A sensitive and robust method for detecting contact domain boundaries**. *Nucleic Acids Research* 2018, 10.1093/nar/gky789.

84. Kruse K, Hug CB, Hernández-Rodríguez B, Vaquerizas JM: **TADtool: Visual parameter identification for tad-calling algorithms**. *Bioinformatics* 2016, **32**:3190–319210.1093/bioinformatics/btw368.

85. Wang X-T, Cui W, Peng C: **HiTAD: Detecting the structural and functional hierarchies of topologically associating domains from chromatin interactions**. *Nucleic Acids Res* 2017, **45**:e16310.1093/nar/gkx735.

86. Haddad N, Vaillant C, Jost D: **IC-finder: Inferring robustly the hierarchical organization of chromatin folding**. *Nucleic Acids Res* 2017, **45**:e8110.1093/nar/gkx036.

87. Oluwadare O, Cheng J: **ClusterTAD: An unsupervised machine learning approach to detecting topologically associated domains of chromosomes from hi-c**

data. *BMC Bioinformatics* 2017, **18**:48010.1186/s12859-017-1931-2.

88. Zhan Y, Mariani L, Barozzi I, Schulz EG, Blüthgen N, Stadler M, Tiana G, Giorgetti L: **Reciprocal insulation analysis of hi-c data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes**. *Genome Research* 2017, **27**:479–49010.1101/gr.212803.116.

89. An L, Yang T, Yang J, Nuebler J, Li Q, Zhang Y: **Hierarchical domain structure reveals the divergence of activity among TADs and boundaries**. 2018, 10.1101/361147.

90. Ay F, Noble WS: **Analysis methods for studying the 3D architecture of the genome**. *Genome Biol* 2015, **16**:18310.1186/s13059-015-0745-7.

91. Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S: **Comparison of computational methods for hi-c data analysis**. *Nat Methods* 2017, **14**:679–68510.1038/nmeth.4325.

92. Nicoletti C, Forcato M, Bicciato S: **Computational methods for analyzing genome-wide chromosome conformation capture data**. *Curr Opin Biotechnol* 2018, **54**:98–10510.1016/j.copbio.2018.01.023.

93. Boulos RE, Arneodo A, Jensen P, Audit B: **Revealing long-range interconnected hubs in human chromatin interaction data using graph theory**. *Phys Rev Lett* 2013, **111**:11810210.1103/PhysRevLett.111.118102.

94. Wang H, Duggal G, Patro R, Girvan M, Hannenhalli S, Kingsford C: **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. New York, NY, USA: ACM; 2013:306:306–306:315. Available: http://doi.acm.org/10.1145/2506583.2506633.

95. Yan K-K, Lou S, Gerstein M: **MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions**. *PLOS Computational Biology* 2017, **13**:e100564710.1371/journal.pcbi.1005647.

96. Norton HK, Emerson DJ, Huang H, Kim J, Titus KR, Gu S, Bassett DS, Phillips-Cremins

JE: **Detecting hierarchical genome folding with network modularity**. *Nat Methods* 2018, **15**:119–12210.1038/nmeth.4560.

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

98. Dali R, Bourque G, Blanchette M: **RobusTAD: A tool for robust annotation of topologically associating domain boundaries**. *bioRxiv* 2018, 10.1101/293175Available: https://www.biorxiv.org/content/early/2018/04/02/293175.

99. Servant N, Lajoie BR, Nora EP, Giorgetti L, Chen C-J, Heard E, Dekker J, Barillot E: **HiTC: Exploration of high-throughput 'c' experiments**. *Bioinformatics* 2012, **28**:2843–410.1093/bioinformatics/bts521.

100. Zufferey M, Tavernari D, Oricchio E, Ciriello G: **Comparison of computational methods for the identification of topologically associating domains**. *Genome Biology* 2018, **19**10.1186/s13059-018-1596-9.

101. Zaborowski R, Wilczynski B: **DiffTAD: Detecting differential contact frequency in topologically associating domains hi-c experiments between conditions**. 2016, 10.1101/093625.

102. Sauerwald N, Kingsford C: **Quantifying the similarity of topological domains across normal and cancer human cell types**. *Bioinformatics* 2018, **34**:i475–i48310.1093/bioinformatics/bty265.

103. Sauerwald N, Singhal A, Kingsford C: **Analysis of the structural variability of topologically associated domains as revealed by hi-c:** 2018, 10.1101/498972.

104. Cresswell KG, Stansfield JC, Dozmorov MG: **SpectralTAD: An r package for defining a hierarchy of topologically associated domains using spectral clustering**. *bioRxiv,*:54917010.1101/549170Available: http://biorxiv.org/content/early/2019/02/13/549170.abstract.

105. Yu SX, Shi J: **Multiclass spectral clustering**. In *Proceedings of the ninth ieee international conference on computer vision - volume 2*. ICCV '03. Washington, DC, USA:

IEEE Computer Society; 2003:313. Available: http://dl.acm.org/citation.cfm?id=946247.946658.

106. Dekker J, Heard E: **Structural and functional diversity of topologically associating domains**. *FEBS Lett* 2015, **589**:2877–8410.1016/j.febslet.2015.08.044.

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

108. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, Li Y, Lin S, Lin Y, Barr CL, Ren B: **A compendium of chromatin contact maps reveals spatially active regions in the human genome**. *Cell Rep* 2016, **17**:2042–205910.1016/j.celrep.2016.10.061.

109. Rosenbloom KR, Dreszer TR, Long JC, Malladi VS, Sloan CA, Raney BJ, Cline MS, Karolchik D, Barber GP, Clawson H, Diekhans M, Fujita PA, Goldman M, Gravell RC, Harte RA, Hinrichs AS, Kirkup VM, Kuhn RM, Learned K, Maddren M, Meyer LR, Pohl A, Rhead B, Wong MC, Zweig AS, Haussler D, Kent WJ: **ENCODE whole-genome data in the ucsc genome browser: Update 2012.** *Nucleic acids research* 2012, **40**:D912–D91710.1093/nar/gkr1012.

110. Won H, Torre-Ubieta L de la, Stein JL, Parikshak NN, Huang J, Opland CK, Gandal MJ, Sutton GJ, Hormozdiari F, Lu D, Lee C, Eskin E, Voineagu I, Ernst J, Geschwind DH: **Chromosome conformation elucidates regulatory relationships in developing human brain**. *Nature* 2016, **538**:523–52710.1038/nature19847.

111. Jiang Y, Loh Y-HE, Rajarajan P, Hirayama T, Liao W, Kassim BS, Javidfar B, Hartley BJ, Kleofas L, Park RB, Labonte B, Ho S-M, Chandrasekaran S, Do C, Ramirez BR, Peter CJ, W JTC, Safaie BM, Morishita H, Roussos P, Nestler EJ, Schaefer A, Tycko B, Brennand KJ, Yagi T, Shen L, Akbarian S: **The methyltransferase SETDB1 regulates a large neuron-specific topological chromatin domain**. *Nature Genetics* 2017, **49**:1239–125010.1038/ng.3906.

112. Luxburg U von: **A tutorial on spectral clustering**. *Statistics and Computing 17(4),*

2007 2007, Available: http://arxiv.org/abs/0711.0189v1.

113. Yardimci G, Ozadam H, Sauria MEG, Ursu O, Yan K-K, Yang T, Chakraborty A, Kaul A, Lajoie BR, Song F, Zhan Y, Ay F, Gerstein M, Kundaje A, Li Q, Taylor J, Yue F, Dekker J, Noble WS: **Measuring the reproducibility and quality of hi-c data**. *bioRxiv*, Available: http://biorxiv.org/content/early/2017/09/14/188755.abstract.

114. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA: **Iterative correction of hi-c data reveals hallmarks of chromosome organization**. *Nat Methods* 2012, **9**:999–100310.1038/nmeth.2148.

115. Knight PA, Ruiz D: **A fast algorithm for matrix balancing**. *IMA Journal of Numerical Analysis* 2012,:drs019.

116. Cournac A, Marie-Nelly H, Marbouty M, Koszul R, Mozziconacci J: **Normalization of a chromosomal contact map**. *BMC Genomics* 2012, **13**:43610.1186/1471-2164-13-436.

117. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS: **HiCNorm: Removing biases in hi-c data via poisson regression**. *Bioinformatics* 2012, **28**:3131–310.1093/bioinformatics/bts570.

118. Li W, Gong K, Li Q, Alber F, Zhou XJ: **Hi-corrector: A fast, scalable and memory-efficient package for normalizing large-scale hi-c data**. *Bioinformatics* 2015, **31**:960–210.1093/bioinformatics/btu747.

119. Ay F, Bailey TL, Noble WS: **Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts**. *Genome Res* 2014, **24**:999–101110.1101/gr.160374.113.

120. Vidal E, Dily F le, Quilez J, Stadhouders R, Cuartero Y, Graf T, Marti-Renom MA, Beato M, Filion GJ: **OneD: Increasing reproducibility of hi-c samples with abnormal karyotypes**. *Nucleic Acids Res* 2018, 10.1093/nar/gky064.

121. Yaffe E, Tanay A: **Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture**. *Nat Genet* 2011, **43**:1059–6510.1038/ng.947.

122. O'Sullivan JM, Hendy MD, Pichugina T, Wake GC, Langowski J: **The statistical-**

mechanics of chromosome conformation capture. *Nucleus*, **4**:390–810.4161/nucl.26513.

123. Li T, Jia L, Cao Y, Chen Q, Li C: **OCEAN-c: Mapping hubs of open chromatin interactions across the genome reveals gene regulatory networks**. *Genome Biol* 2018, **19**:5410.1186/s13059-018-1430-4.

124. Schoenfelder S, Clay I, Fraser P: **The transcriptional interactome: Gene expression in 3D**. *Curr Opin Genet Dev* 2010, **20**:127–3310.1016/j.gde.2010.02.002.

125. Steensel B van: **Chromatin: Constructing the big picture**. *EMBO J* 2011, **30**:1885–9510.1038/emboj.2011.135.

126. Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, Kraft K, Kempfer R, Jerković I, Chan W-L, Spielmann M, Timmermann B, Wittler L, Kurth I, Cambiaso P, Zuffardi O, Houge G, Lambie L, Brancati F, Pombo A, Vingron M, Spitz F, Mundlos S: **Formation of new chromatin domains determines pathogenicity of genomic duplications**. *Nature* 2016, **538**:265–26910.1038/nature19800.

127. Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, Ettwiller L, Spitz F: **Functional and topological characteristics of mammalian regulatory domains**. *Genome Res* 2014, **24**:390–40010.1101/gr.163519.113.

128. Sexton T, Cavalli G: **The role of chromosome domains in shaping the functional genome**. *Cell* 2015, **160**:1049–5910.1016/j.cell.2015.02.040.

129. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, Sim HS, Peh SQ, Mulawadi FH, Ong CT, Orlov YL, Hong S, Zhang Z, Landt S, Raha D, Euskirchen G, Wei C-L, Ge W, Wang H, Davis C, Fisher-Aylor KI, Mortazavi A, Gerstein M, Gingeras T, Wold B, Sun Y, et al.: **Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation**. *Cell* 2012, **148**:84–9810.1016/j.cell.2011.12.014.

130. Papantonis A, Cook PR: **Transcription factories: Genome organization and gene regulation**. *Chem Rev* 2013, **113**:8683–70510.1021/cr300513p.

131. Laat W de, Grosveld F: **Spatial organization of gene expression: The active**

chromatin hub. *Chromosome Res* 2003, **11**:447–59.

132. Mora A, Sandve GK, Gabrielsen OS, Eskeland R: **In the loop: Promoter-enhancer interactions and bioinformatics**. *Brief Bioinform* 2016, **17**:980–99510.1093/bib/bbv097.

133. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA, Herman B, Happe S, Higgs A, LeProust E, Follows GA, Fraser P, Luscombe NM, Osborne CS: **Mapping long-range promoter contacts in human cells with high-resolution capture hi-c**. *Nat Genet* 2015, **47**:598–60610.1038/ng.3286.

134. Shavit Y, Lio' P: **Combining a wavelet change point and the bayes factor for analysing chromosomal interaction data**. *Mol Biosyst* 2014, **10**:1576–8510.1039/c4mb00142g.

135. Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E, Goyenechea B, Mitchell JA, Lopes S, Reik W, Fraser P: **Active genes dynamically colocalize to shared sites of ongoing transcription**. *Nat Genet* 2004, **36**:1065–7110.1038/ng1423.

136. Schoenfelder S, Sexton T, Chakalova L, Cope NF, Horton A, Andrews S, Kurukuti S, Mitchell JA, Umlauf D, Dimitrova DS, Eskiw CH, Luo Y, Wei C-L, Ruan Y, Bieker JJ, Fraser P: **Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells**. *Nat Genet* 2010, **42**:53–6110.1038/ng.496.

137. Tanizawa H, Iwasaki O, Tanaka A, Capizzi JR, Wickramasinghe P, Lee M, Fu Z, Noma K-i: **Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation**. *Nucleic Acids Res* 2010, **38**:8164–7710.1093/nar/gkq955.

138. Cremer T, Cremer M: **Chromosome territories**. *Cold Spring Harb Perspect Biol* 2010, **2**:a00388910.1101/cshperspect.a003889.

139. Denker A, Laat W de: **The second decade of 3C technologies: Detailed insights into nuclear organization**. *Genes Dev* 2016, **30**:1357–8210.1101/gad.281964.116.

140. Dai Z, Dai X: **Nuclear colocalization of transcription factor target genes strengthens coregulation in yeast**. *Nucleic Acids Res* 2012, **40**:27–3610.1093/nar/gkr689.

141. Lajoie BR, Dekker J, Kaplan N: **The hitchhiker's guide to hi-c analysis: Practical guidelines**. *Methods* 2015, **72**:65–7510.1016/j.ymeth.2014.10.031.

142. Rao SSP, Huang S-C, Glenn St Hilaire B, Engreitz JM, Perez EM, Kieffer-Kwon K-R, Sanborn AL, Johnstone SE, Bascom GD, Bochkov ID, Huang X, Shamim MS, Shin J, Turner D, Ye Z, Omer AD, Robinson JT, Schlick T, Bernstein BE, Casellas R, Lander ES, Aiden EL: **Cohesin loss eliminates all loop domains**. *Cell* 2017, **171**:305–320.e2410.1016/j.cell.2017.09.026Available: http://dx.doi.org/10.1016/j.cell.2017.09.026.

143. Ernst J, Kellis M: **Discovery and characterization of chromatin states for systematic annotation of the human genome**. *Nat Biotechnol* 2010, **28**:817–2510.1038/nbt.1662.

144. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, Onate KC, Graham K, Miyasato SR, Dreszer TR, Strattan JS, Jolanki O, Tanaka FY, Cherry JM: **The encyclopedia of DNA elements (ENCODE): Data portal update**. *Nucleic Acids Research* 2017, **46**:D794–D80110.1093/nar/gkx1081.

145. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G: **GREAT improves functional interpretation of cis-regulatory regions.** *Nature biotechnology* 2010, **28**:495–50110.1038/nbt.1630.

146. Nora EP, Goloborodko A, Valton A-L, Gibcus JH, Uebersohn A, Abdennur N, Dekker J, Mirny LA, Bruneau BG: **Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization**. *Cell* 2017, **169**:930–944.e2210.1016/j.cell.2017.05.004.

147. Stansfield JC, Cresswell KG, Vladimirov VI, Dozmorov MG: **HiCcompare: An r-package for joint normalization and comparison of hi-c datasets**. *BMC Bioinformatics* 2018, **19**:27910.1186/s12859-018-2288-x.

148. Stansfield JC, Cresswell KG, Dozmorov MG: **MultiHiCcompare: Joint normalization and comparative analysis of complex hi-c experiments**. *Bioinformatics* 2019,

10.1093/bioinformatics/btz048.

149. Djekidel MN, Chen Y, Zhang MQ: **FIND: DifFerential chromatin interactions detection using a spatial poisson process**. *Genome Res* 2018, 10.1101/gr.212241.116.

150. Lun ATL, Smyth GK: **DiffHic: A bioconductor package to detect differential genomic interactions in hi-c data**. *BMC Bioinformatics* 2015, **16**:25810.1186/s12859-015-0683-0.

151. Yang T, Zhang F, Yardımcı GG, Song F, Hardison RC, Noble WS, Yue F, Li Q: **HiCRep: Assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient**. *Genome Res* 2017, 10.1101/gr.220640.117.

152. Ardakany AR, Ay F, Lonardi S: **Selfish: Discovery of differential chromatin interactions via a self-similarity measure**. 2019, 10.1101/540708.

153. Ursu O, Boley N, Taranova M, Wang YXR, Yardimci GG, Noble WS, Kundaje A: **GenomeDISCO: A concordance score for chromosome conformation capture experiments using random walks on contact map graphs**. *bioRxiv*, Available: http://biorxiv.org/content/early/2017/08/29/181842.abstract.

154. Yan K-K, Yardimci GG, Yan C, Noble WS, Gerstein M: **HiC-spector: A matrix library for spectral and reproducibility analysis of hi-c contact maps**. *Bioinformatics* 2017, **33**:2199–220110.1093/bioinformatics/btx152.

155. Sauria ME, Phillips-Cremins JE, Corces VG, Taylor J: **HiFive: A tool suite for easy and efficient hic and 5C data analysis**. *bioRxiv*, Available: http://biorxiv.org/content/early/2015/10/09/009951.abstract.

156. Bar-Joseph Z, Gitter A, Simon I: **Studying and modelling dynamic biological processes using time-series gene expression data**. *Nat Rev Genet* 2012, **13**:552–6410.1038/nrg3244.

157. Abu-Jamous B, Kelly S: **Clust: Automatic extraction of optimal co-expressed gene clusters from gene expression data**. *Genome Biol* 2018, **19**:17210.1186/s13059-018-1536-8.