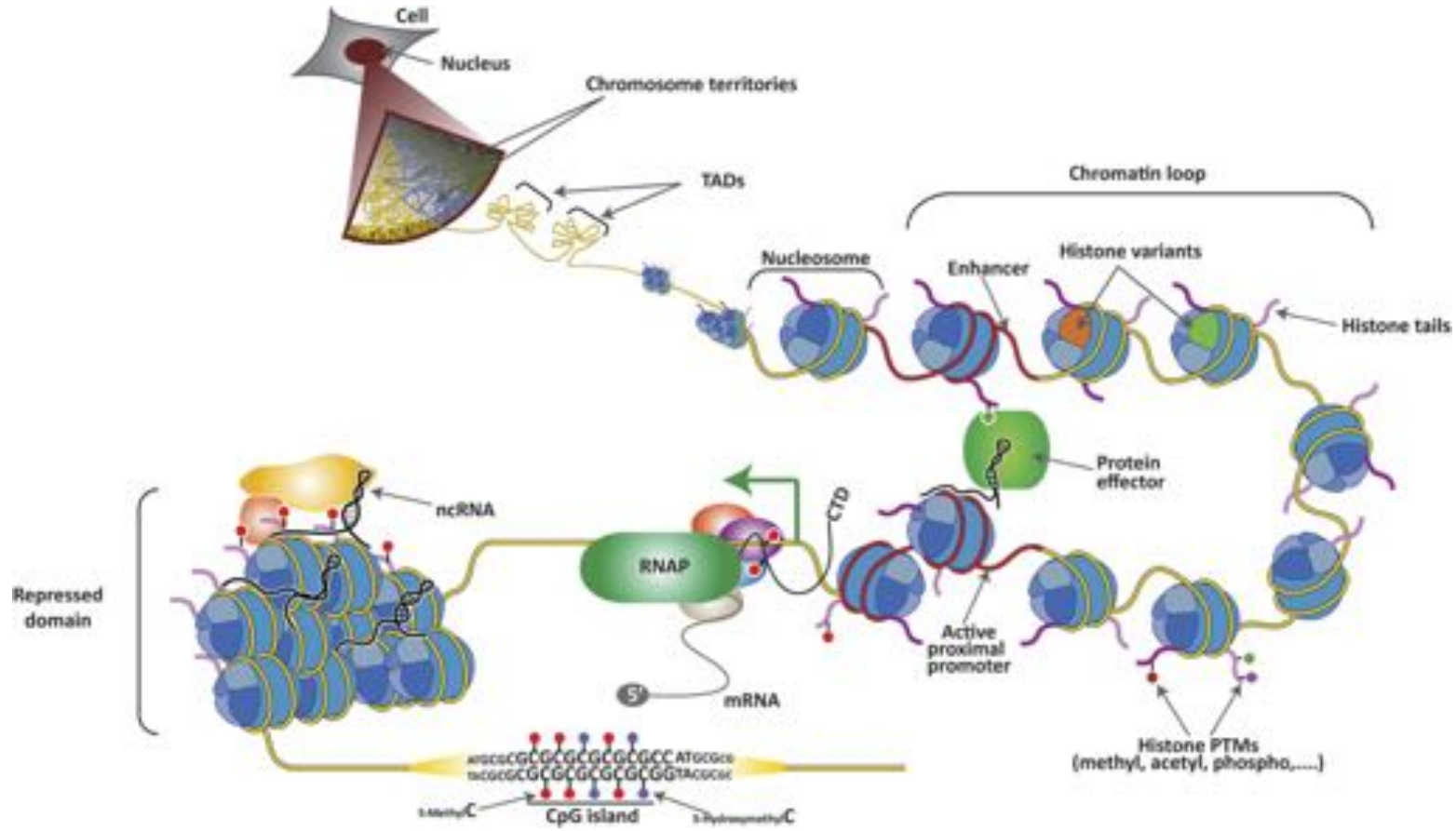
SpectralTAD: Defining Hierarchy of Topologically Associated Domains Using Graph Theoretical Clustering

Mikhail Dozmorov, Kellen Cresswell

July 29, 2019

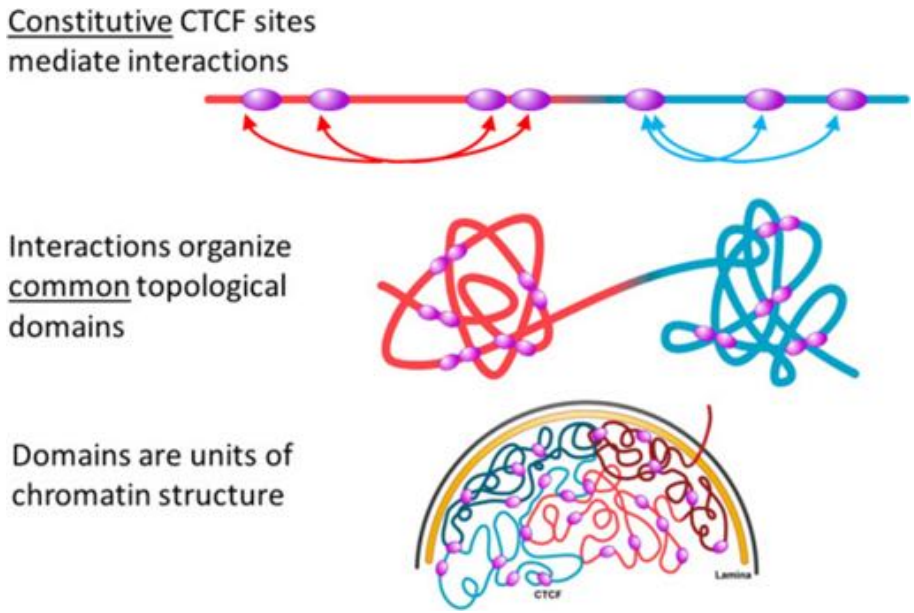
The Genome

- Human genome is big - ~3.2 billion base pairs
- ~2 meters (~6ft) of DNA in one cell packed into the ~10μm nucleus
- ~500 times distance from Earth to Sun in all cells from human body



3D Genomics

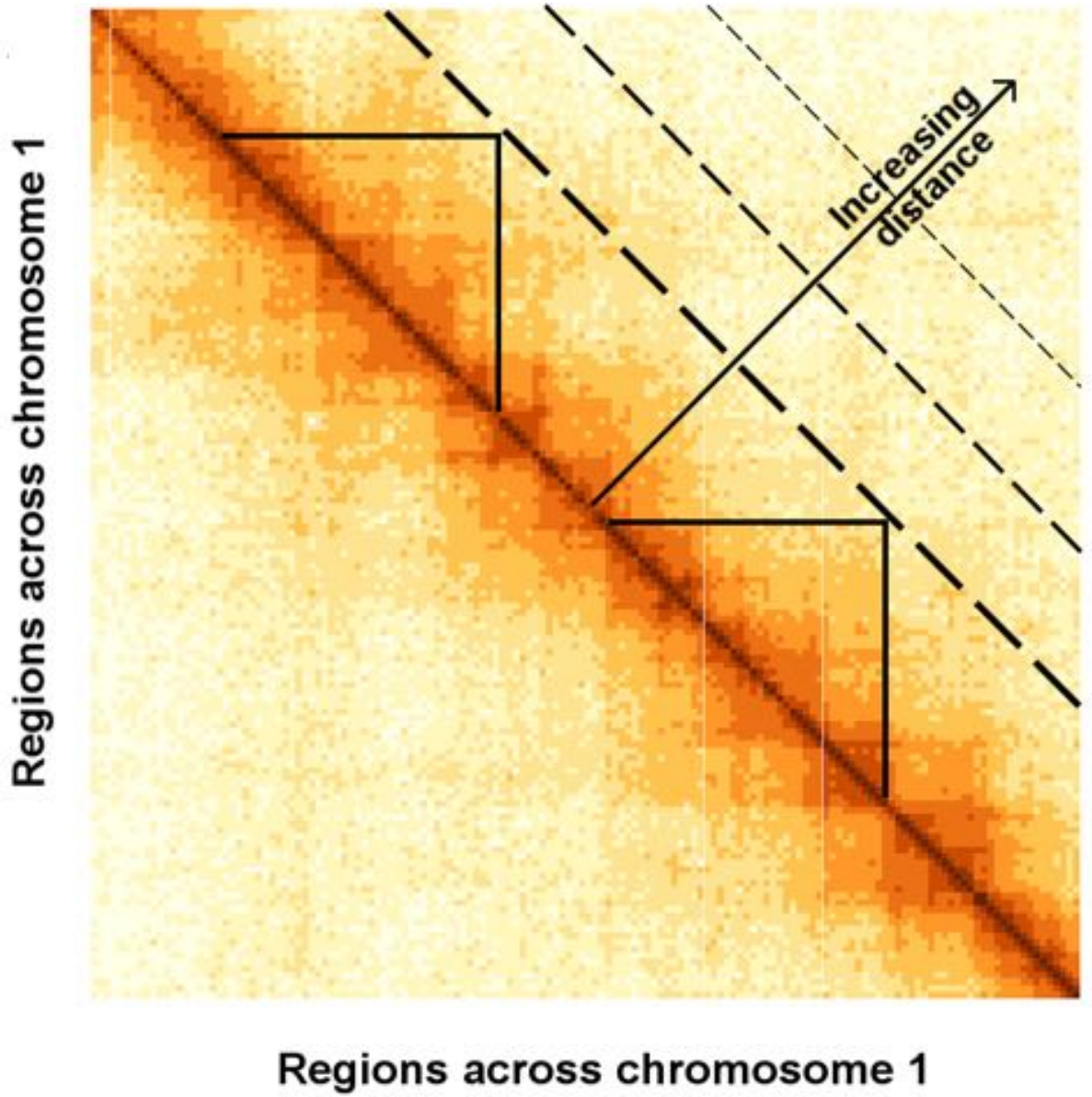
- **Genome folding** enables interaction between distant genomic regions
- **Hi-C sequencing** (Chromatin Conformation Capture technology) allows for identification of genomic interactions genome-wide



Hi-C Data as a matrix

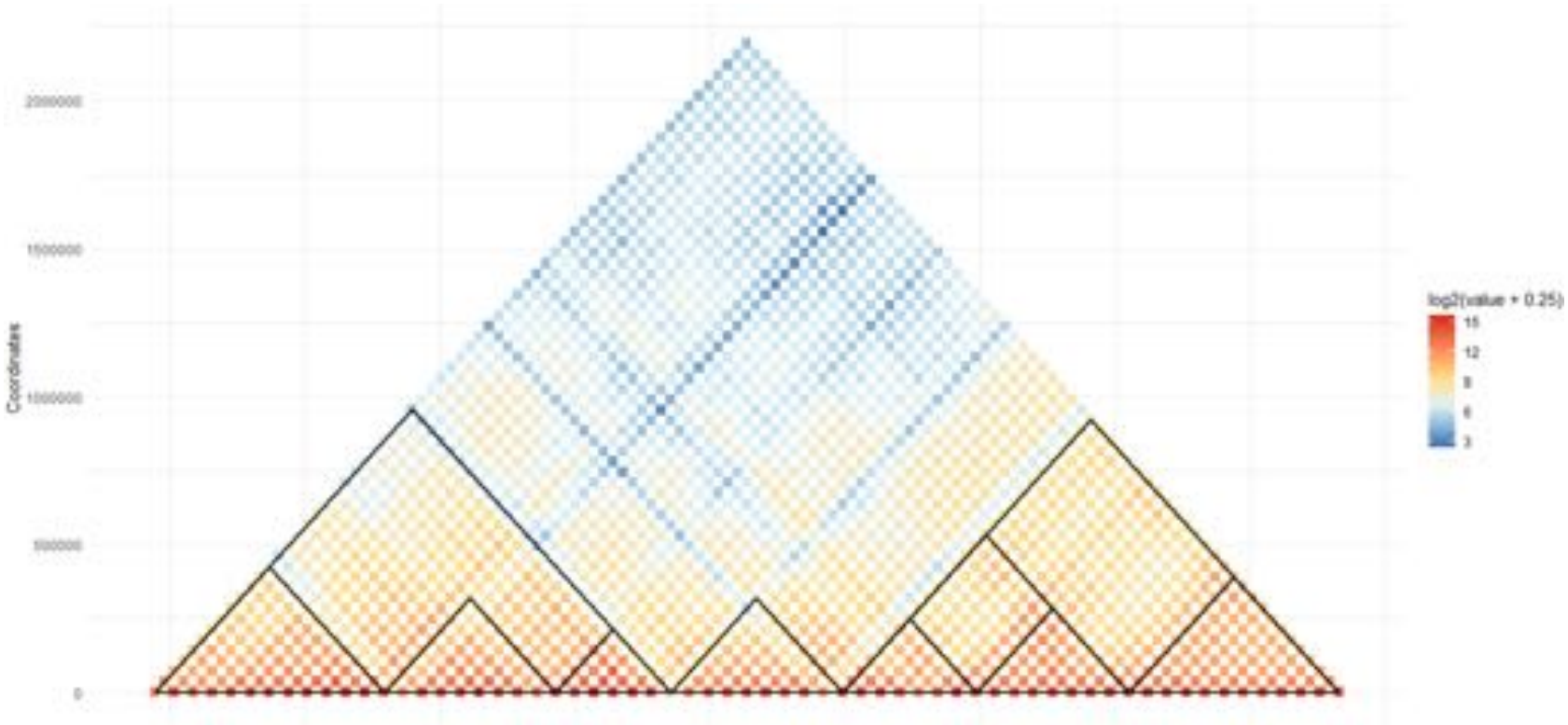
- The genome (chromosome) is split into equally sized regions
- Region size (resolution) is determined by sequencing depth
- Data is represented by a symmetric matrix of contacts C_{ij} where entry ij corresponds to the number of times region i comes into contact with region j

Hi-C Data as a matrix



Topologically Associated Domains (TADs)

- TADs are domains of frequent local interactions separated by boundaries across which interactions are less frequent
- Boundaries are associated with specific genomic features (CTCF, cohesin, mediator)
- Can be nested (TADs containing sub-TADs)

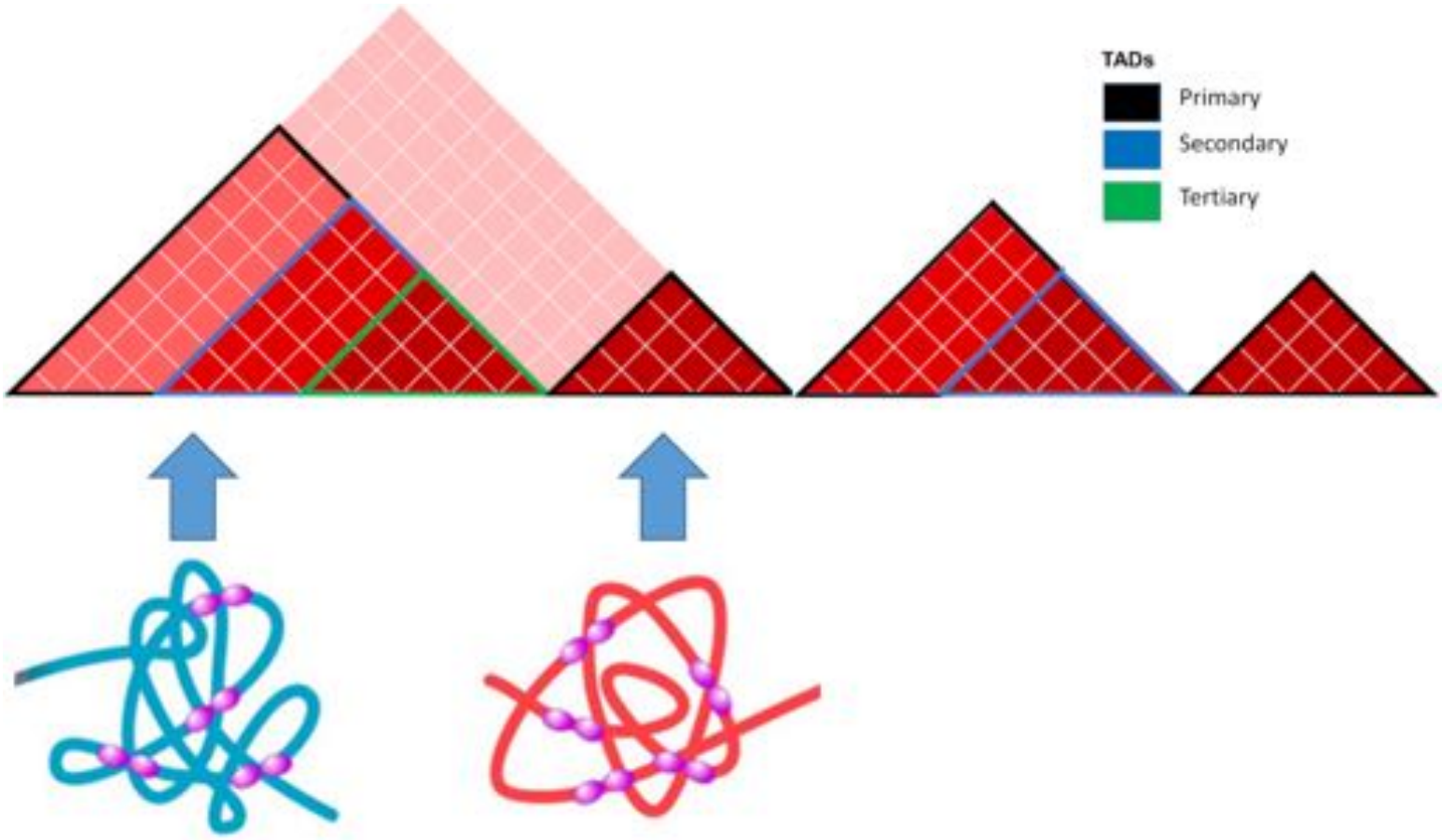


Why are TADs Important?

- Established early in development and highly conserved
- TADs create "autonomous gene-domains" essentially partitioning the genome into discrete functional regions
- Disruptions of TADs lead to de novo enhancer-promoter interactions and dysregulation of gene expression
- Can be altered using CRISPR

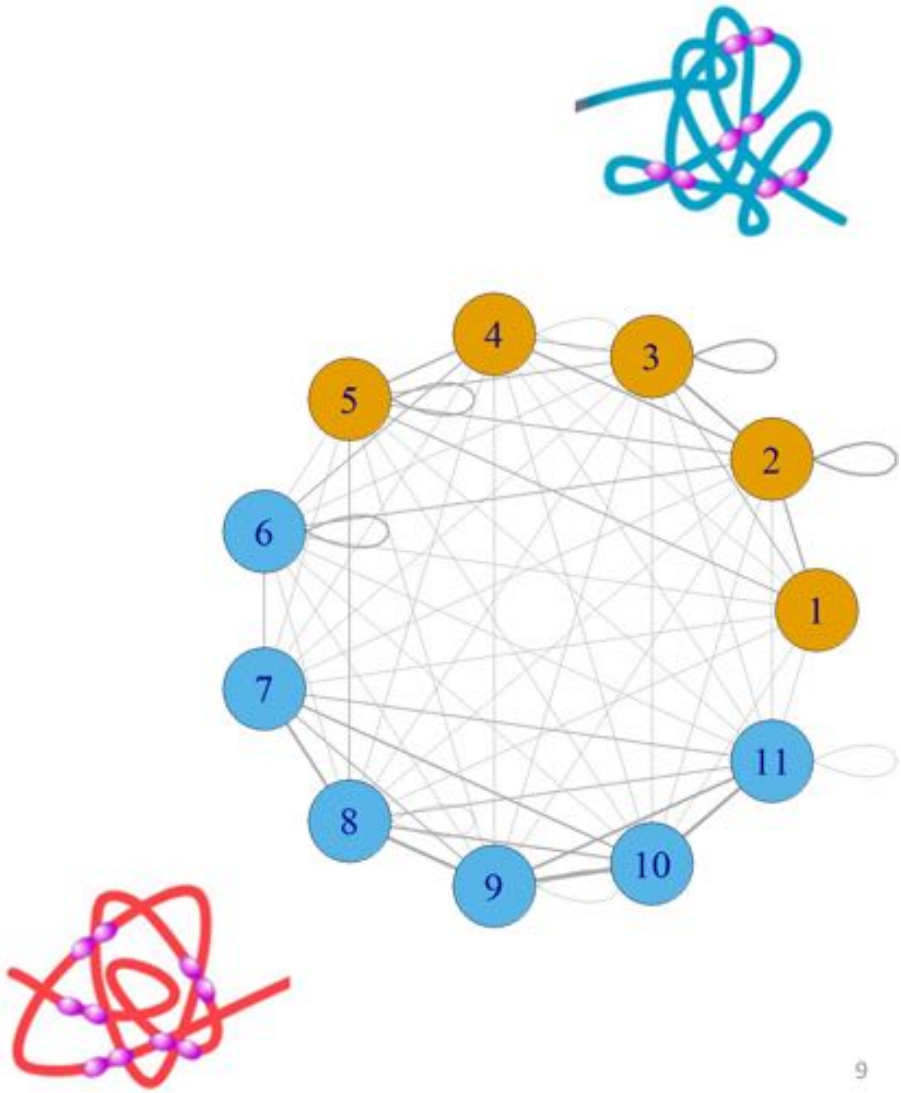
TADs are hierarchical

- Organized in a hierarchy
- Characterized by large "meta-TADs" containing small "sub-TADs"
- Level of hierarchy has an effect on biological relevance



Graph Representation of 3D Data

- Hi-C data has a natural graph structure, defined by vertices V and edges E
  - **Vertices** are genomic regions
  - **Edges** represent interaction strength between any pair of regions
- Vertices and edges are stored in an **adjacency matrix** A_{ij} where ij is the number of edges between a given set of vertices ij



Traditional Spectral Clustering

- Specifically designed to cluster graphs
- Works by projecting the data into a lower-dimensional space
- Excels on noisy and non-normally distributed data (Hi-C data)
- Clusters the adjacency matrix $A_{n \times n}$

How to perform spectral clustering

- Calculate the Laplacian:

$$D = diag(A\mathbf{1_n})$$

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

- Calculate the eigenvectors of the Laplacian matrix (graph spectrum):

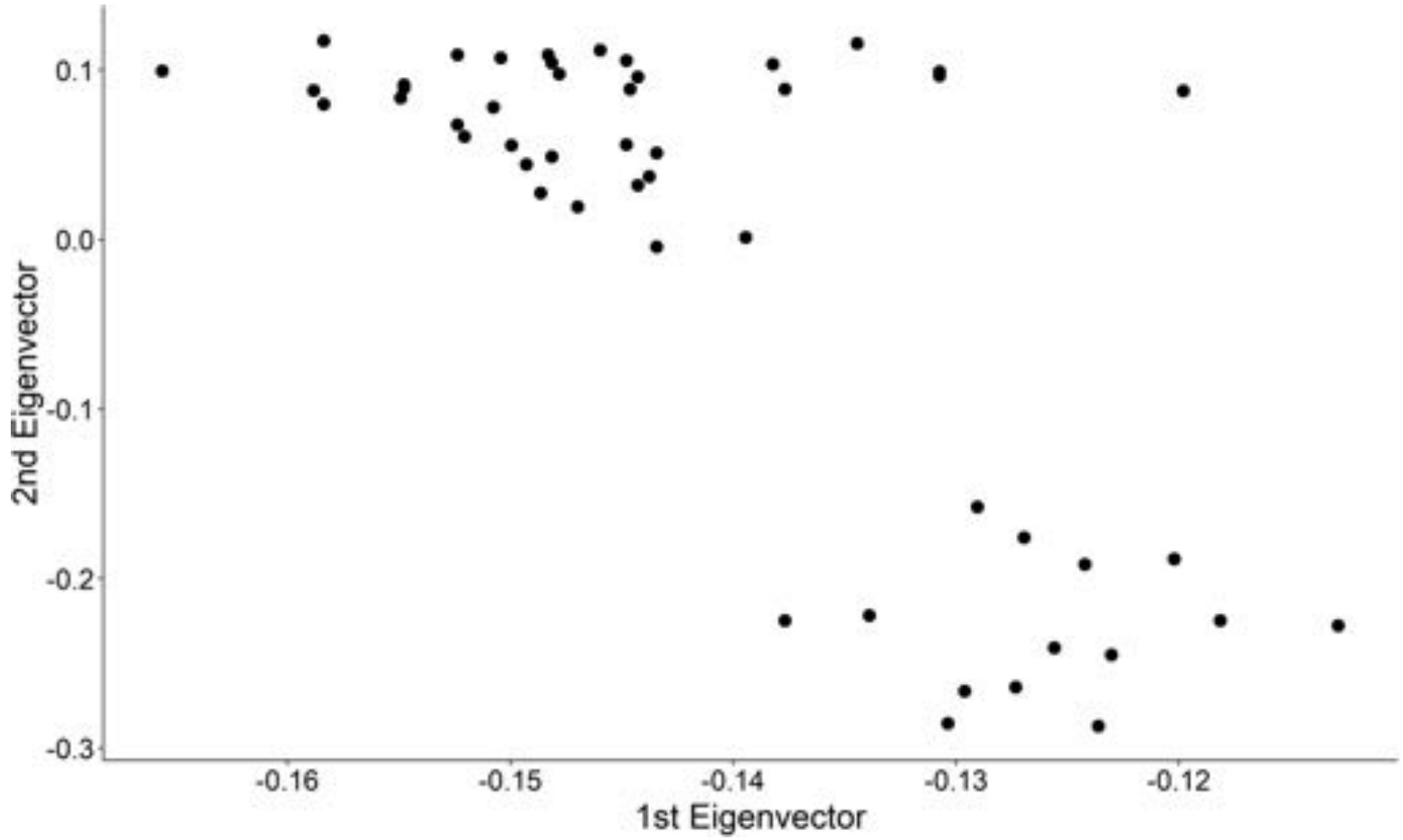
$$\bar{L}\mathbf{v} = \lambda\mathbf{v}$$

- Normalize the eigenvectors and cluster

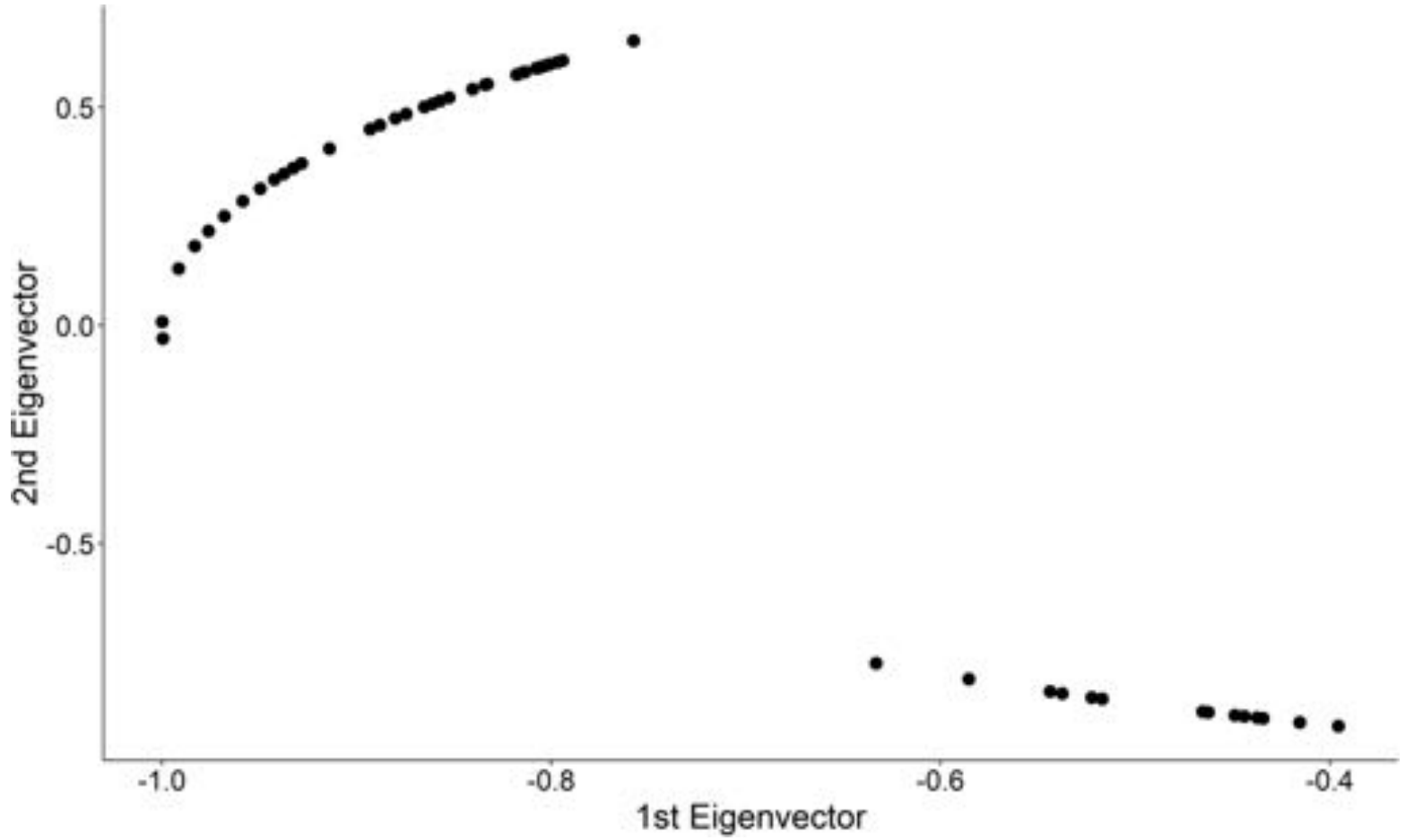
Spectral clustering with eigenvector gaps

- Rows and columns of contact matrices are naturally ordered
- TADs are continuous
- Ordering allows us to reframe clustering as finding cut points
- We propose a simple, novel, approach to clustering ordered data using gaps between consecutive eigenvectors

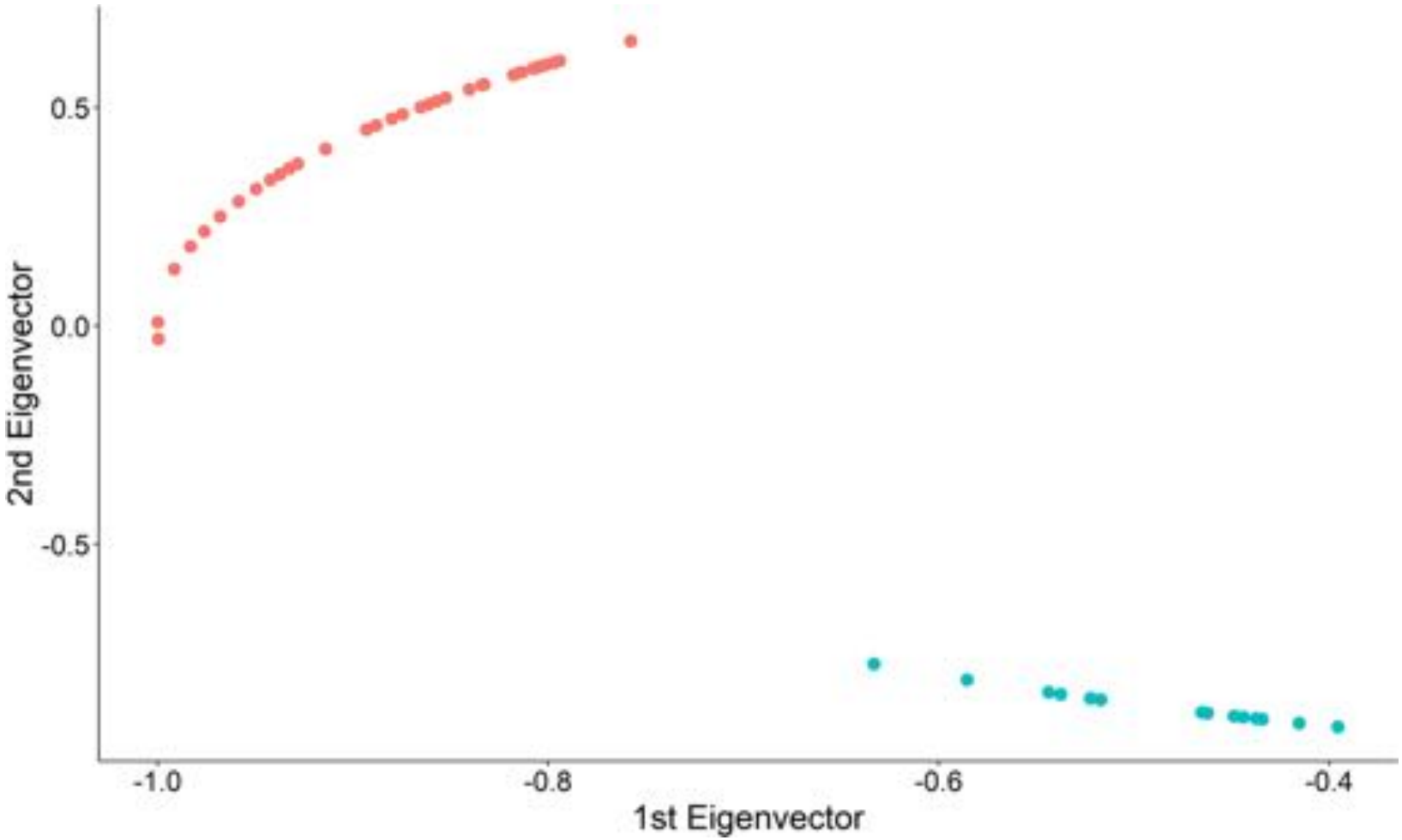
Step 1: Plot the non-normalized eigenvectors



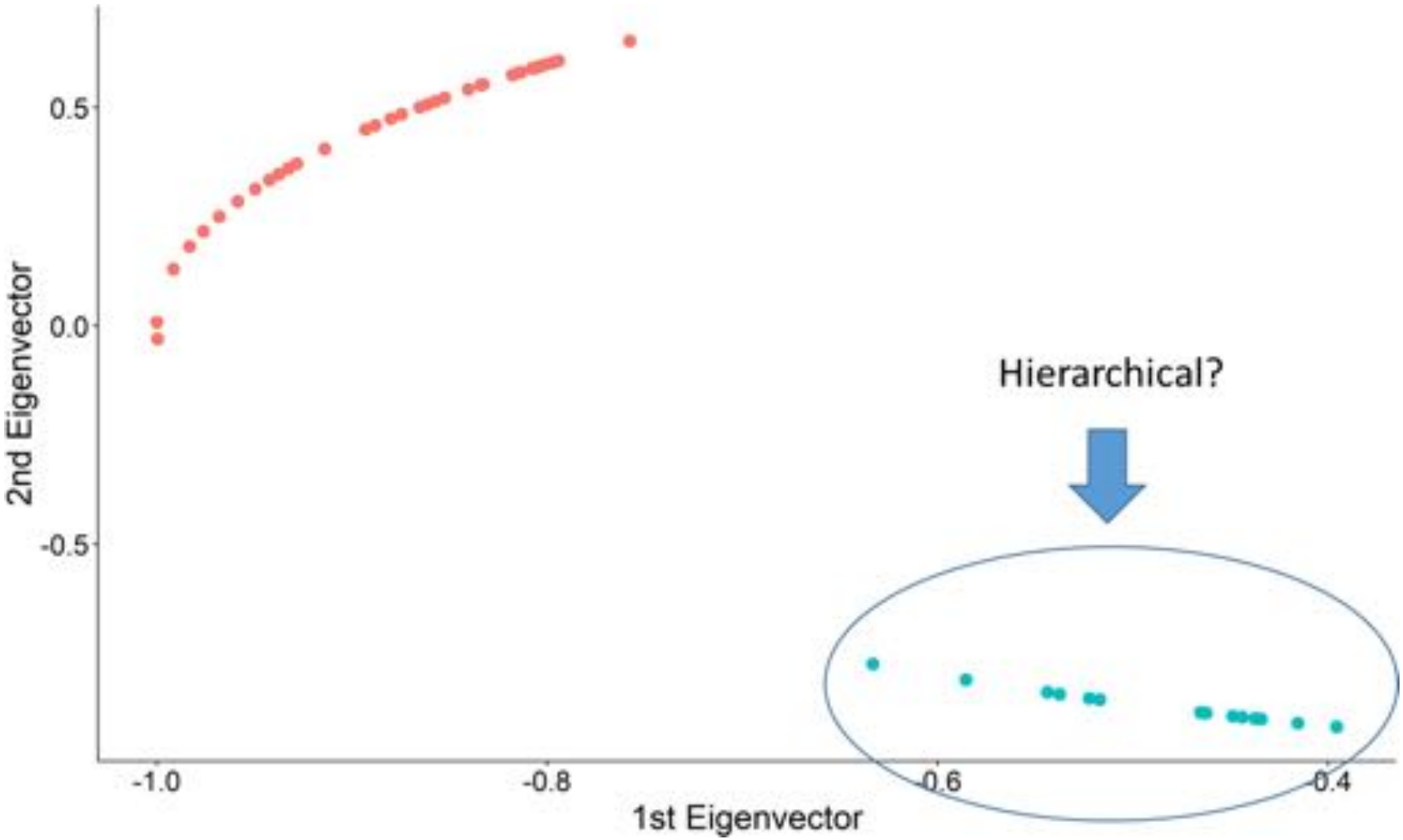
Step 2: Project on to Unit Circle



Step 3: Find the k-largest gaps and partition



Step 3: Find the k-largest gaps and partition



Windowed Spectral Clustering

- We know the biologically maximum TAD size (2 million bp)
- We can use a 2 million bp sliding window to perform spectral clustering and aggregate
- Advantages of the sliding window
  - Reduced cubic complexity of spectral clustering $O(n^3)$ to linear complexity $O(n)$
  - Naturally discards noisy interactions at large genomic distances

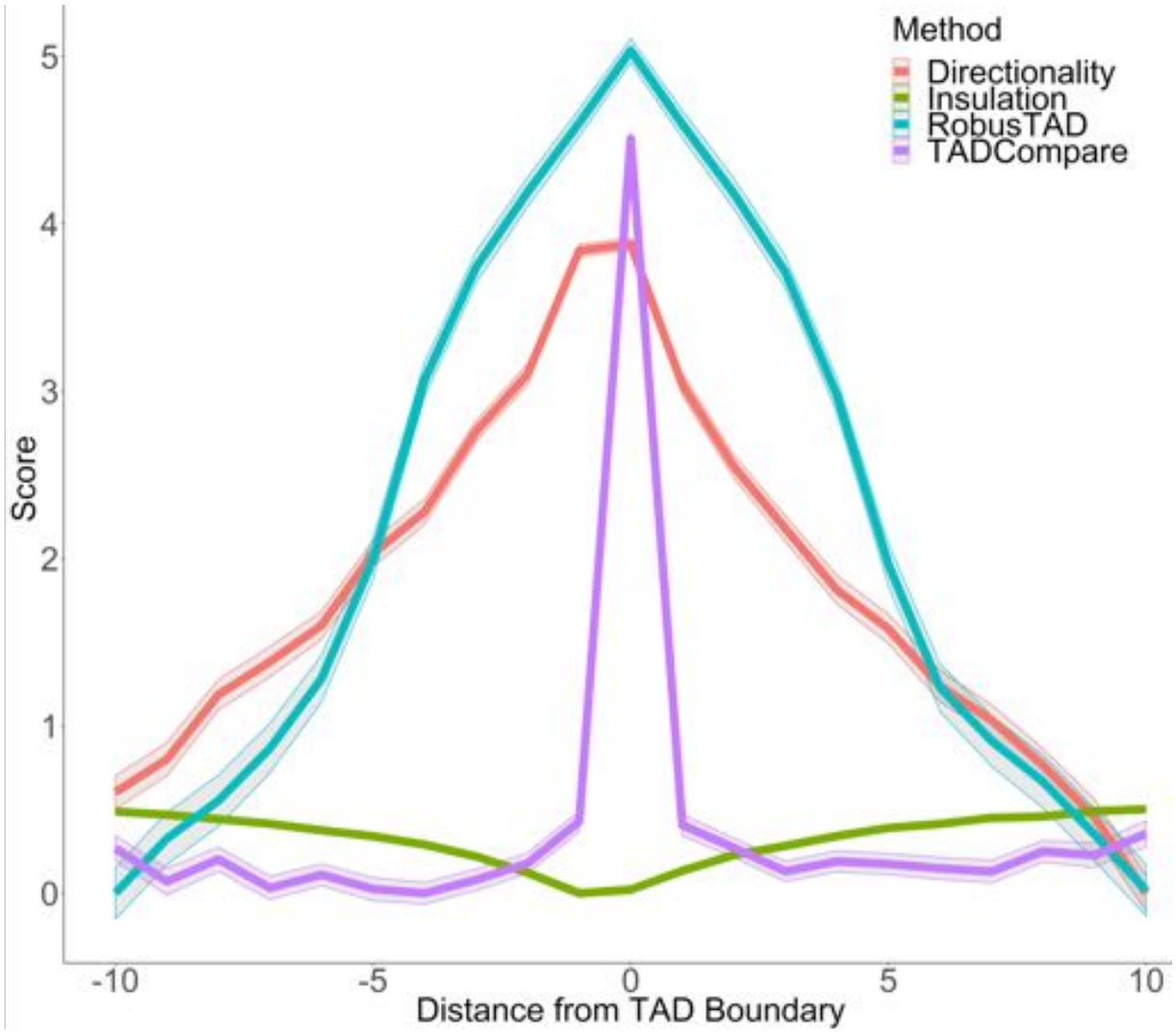
SpectralTAD algorithm

1. Cut a window from the matrix equal to the maximum TAD size (2Mb)
2. Find the graph spectrum of the window and calculate eigenvector gaps
3. Find n-largest gap values
4. Find the set of clusters that maximize the silhouette score
5. Slide the window to the next group of loci and repeat

Determining a hierarchy of TADs

- TADs are hierarchical in nature (organized into large meta-TADs with sub-TADs within them)
- Need to find sub-TADs within those detected by sliding window
- To find sub-TADs, we use a novel metric called boundary score
- **Boundary score** is just the z-score for each eigenvector gap

Boundary score as a metric for TAD boundary detection



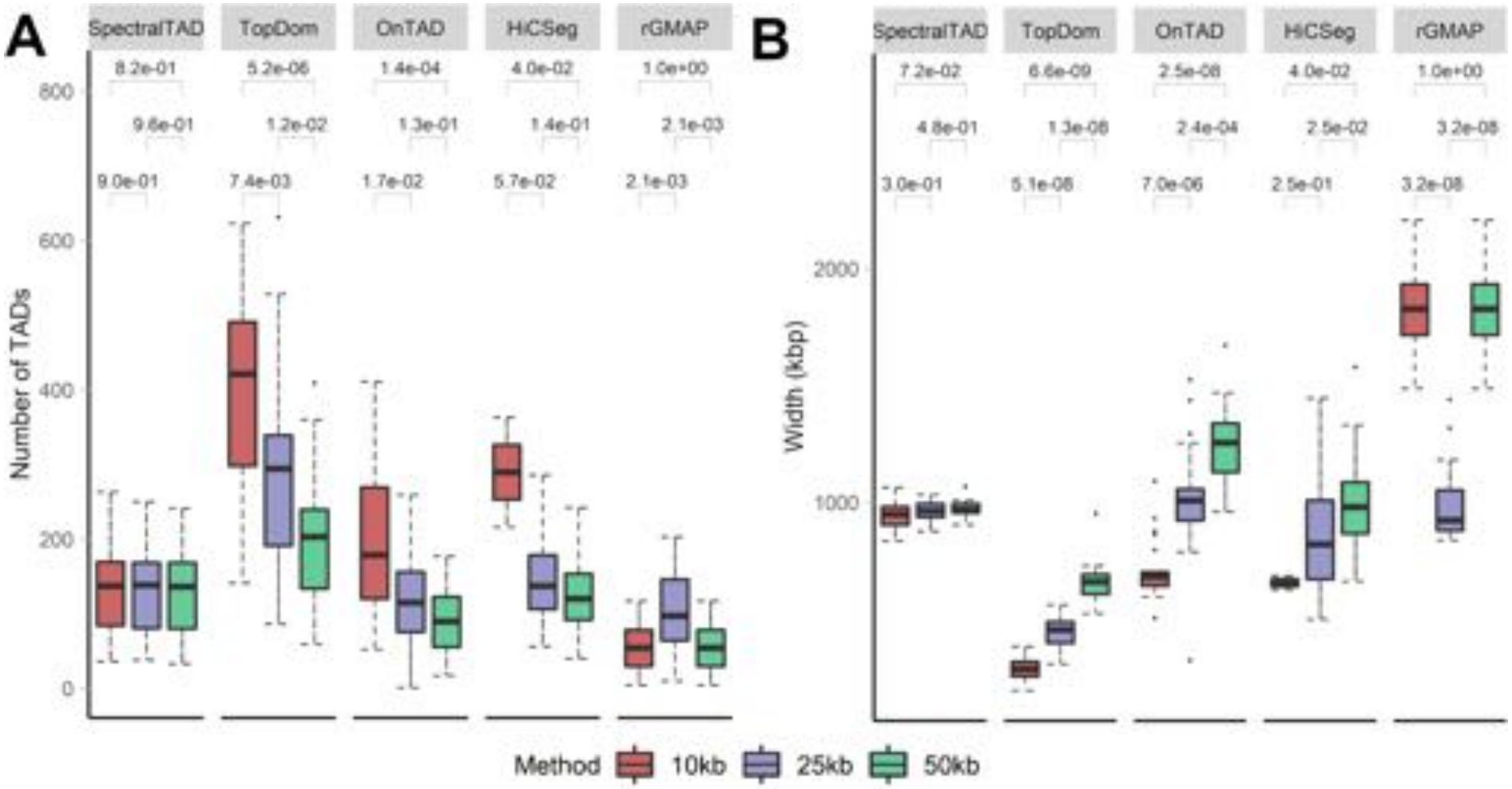
Determining a hierarchy of TADs

- For each initial TAD:
  - Perform spectral clustering on the submatrix defined by the initial TAD
  - Calculate the eigenvector gaps for each consecutive pair of regions
  - Convert eigenvector gaps to boundary scores
  - If any boundary score is greater than 1.96, this is a sub-TAD boundary
  - Repeat for all sub-TADs until no z-score is greater than 1.96

TAD Calling

- Good TAD callers must satisfy three criteria:
  - Be robust to Hi-C data imperfections (resolution, sparsity, sequencing depth)
  - Detect biologically significant, hierarchical TAD boundaries
  - Be fast
- We compared SpectralTAD against four TAD callers:
  - TopDom
  - HiCSeg
  - OnTAD
  - rGMAP

SpectralTAD is robust to resolution

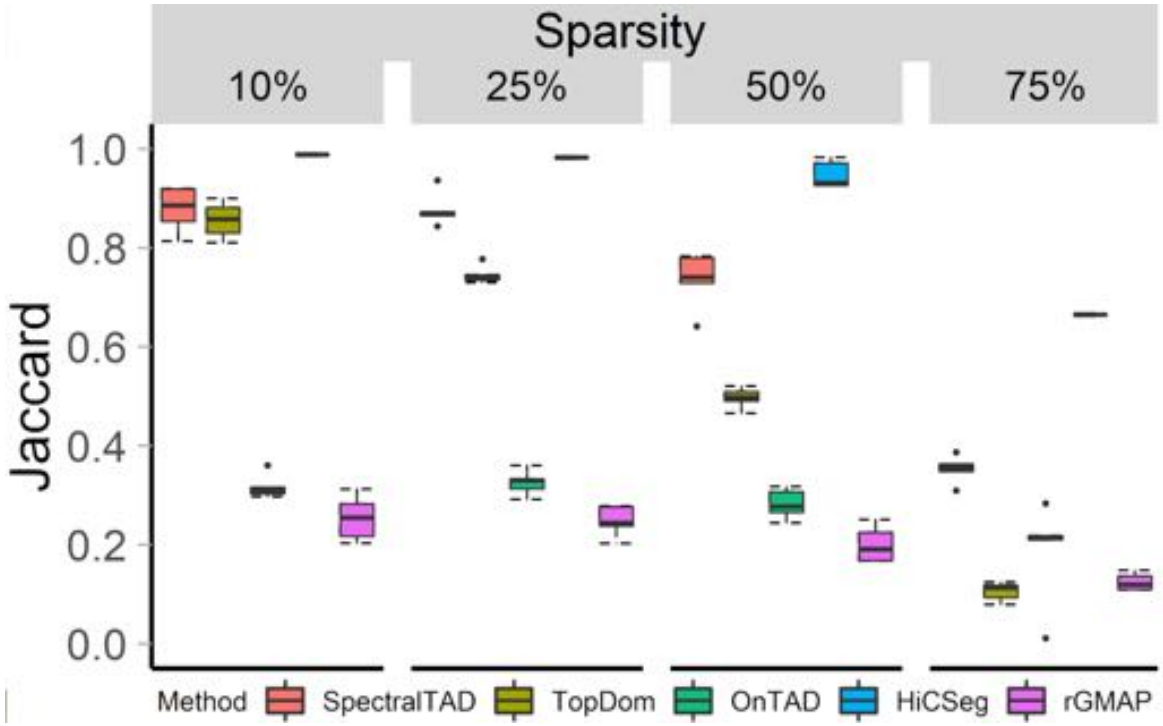


Sparsity

- One of the main biases in HiC data
- Characterized by random zeros in the contact matrix
- Simulated by replacing a certain percentage of the contact matrix with zeros

SpectralTAD is robust to sparsity

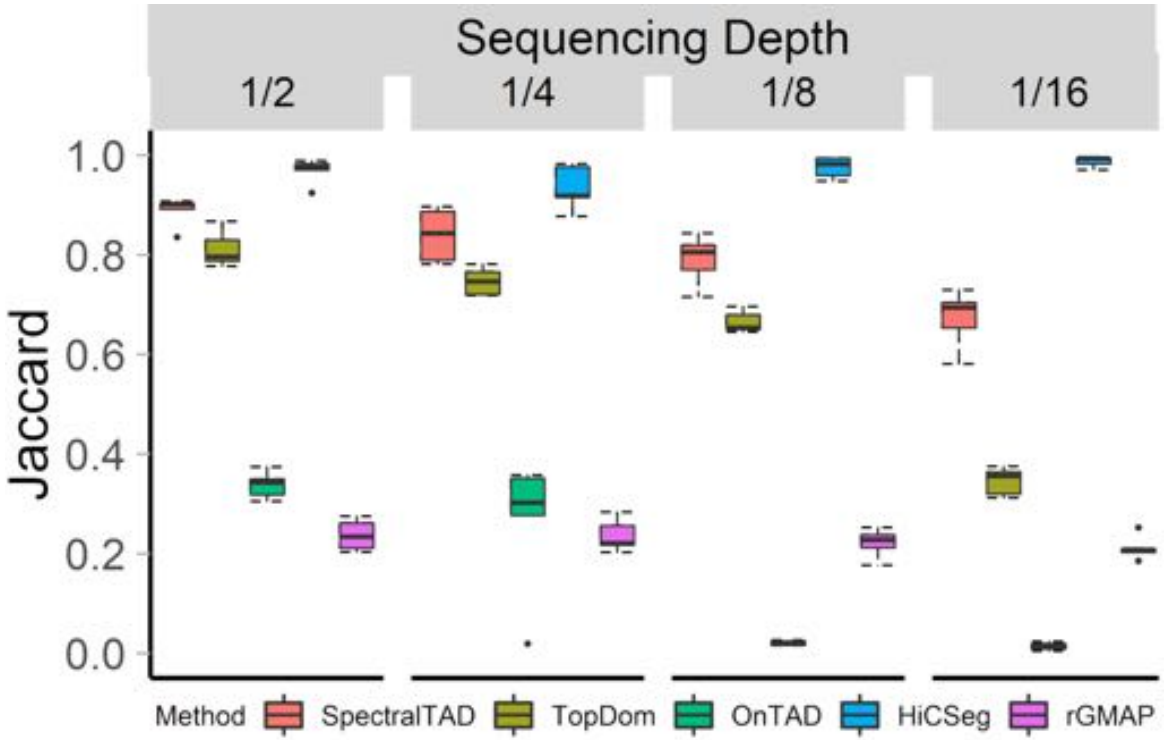
- 25 simulated matrices with pre-defined TADs (HiCToolsCompare)
- The percentage of the matrix replaced with zeros
- Jaccard similarity between the detected and pre-defined TADs



- Our method is better than other methods at all levels of sparsity (except HiCseg, which detects least biologically significant TADs)

SpectralTAD is robust to sequencing depth

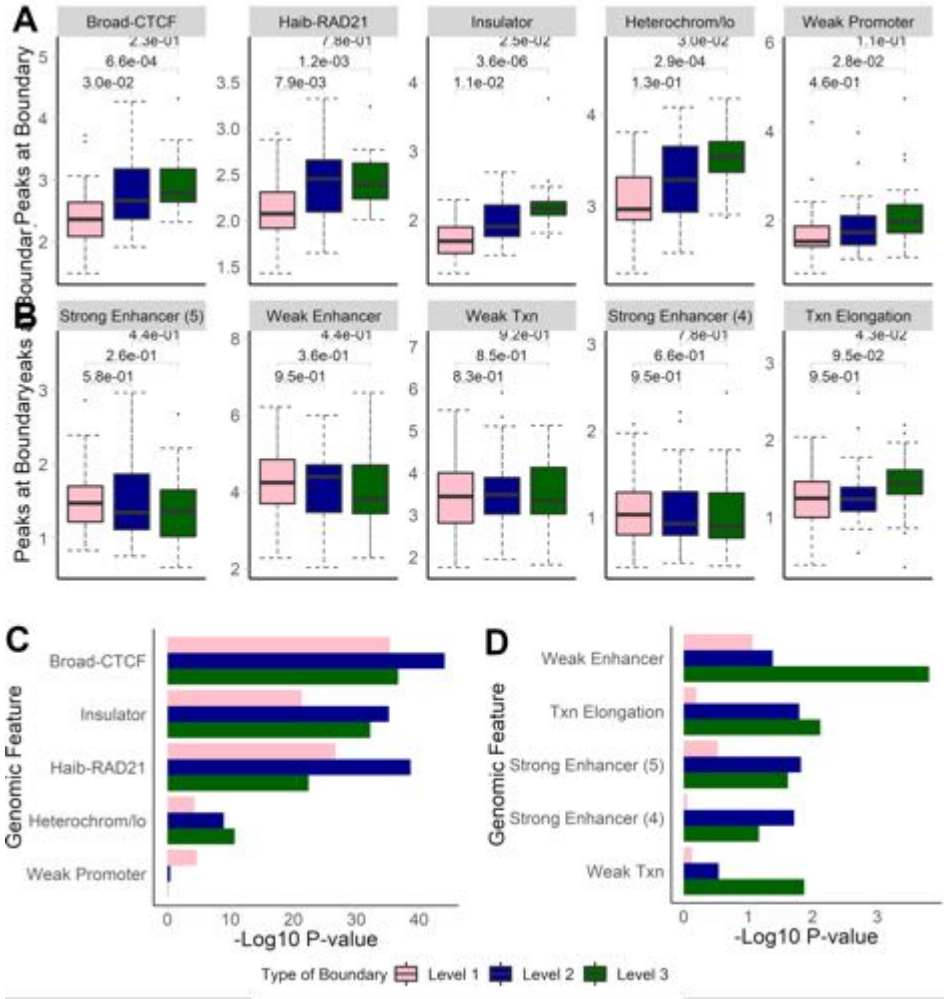
- The fraction indicates the proportion of contacts removed.



- Our method outperforms all other methods at all levels of downsampling (excluding HiCSeg, which detects least biologically significant TADs)

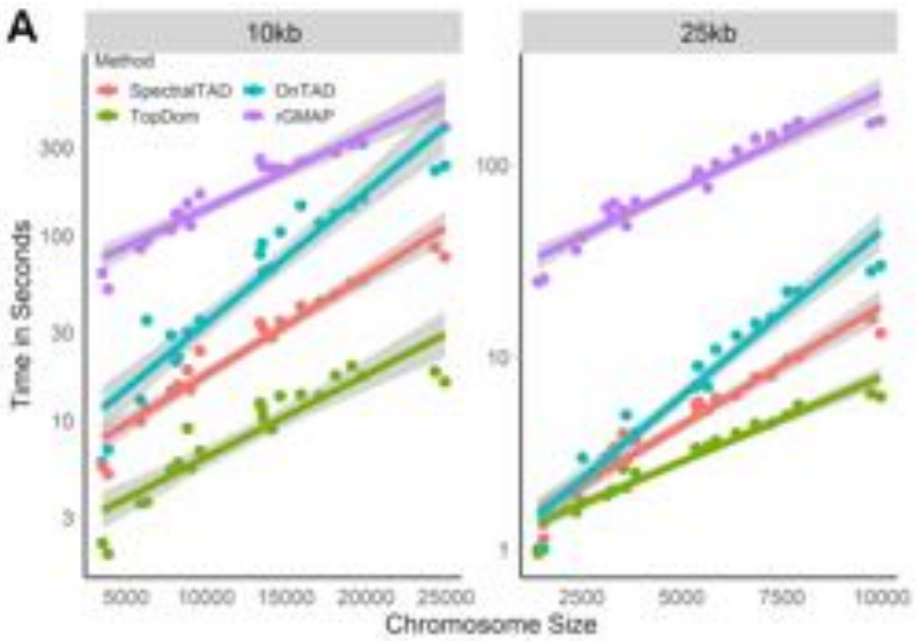
Hierarchical TAD boundaries differ

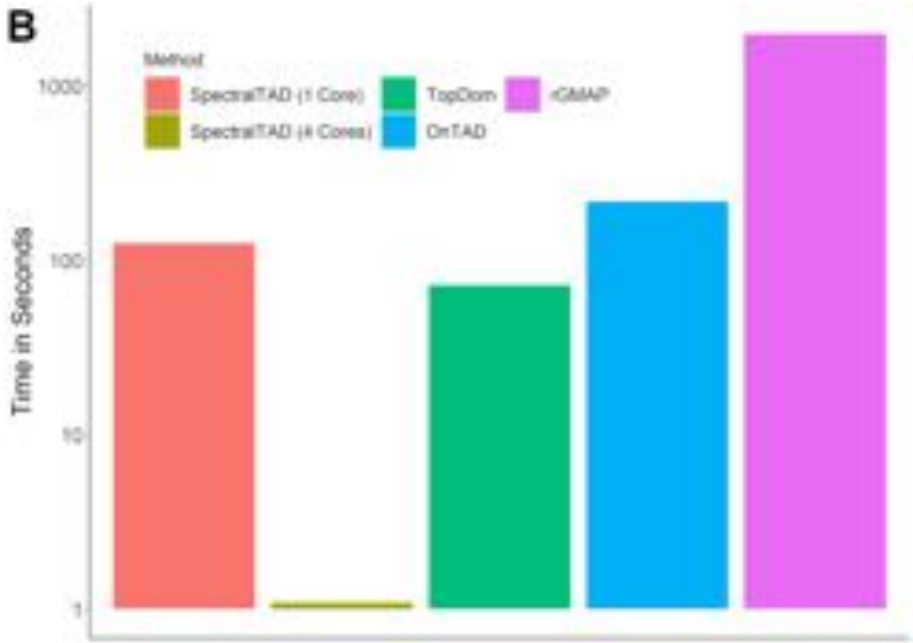
- Boundaries shared by two TADs (Level 2) or three TADs (Level 3) are more biologically significant



SpectralTAD is fast

A) Runtimes for various TAD callers at different chromosome sizes
B) Runtimes for various TAD callers across all chromosomes (25kb data)

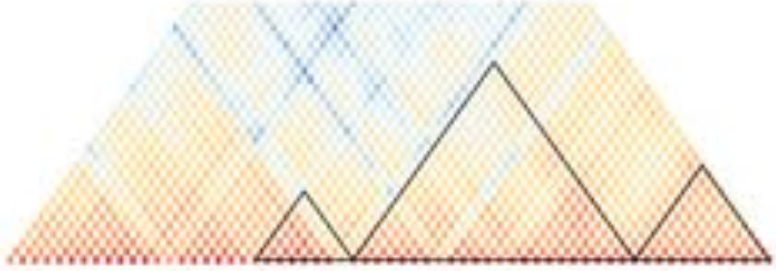




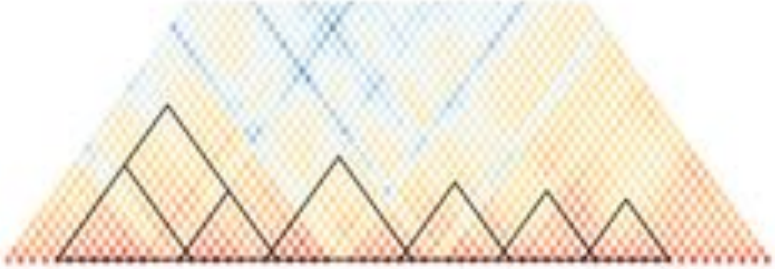
Side-by-side comparison

SpectralTAD

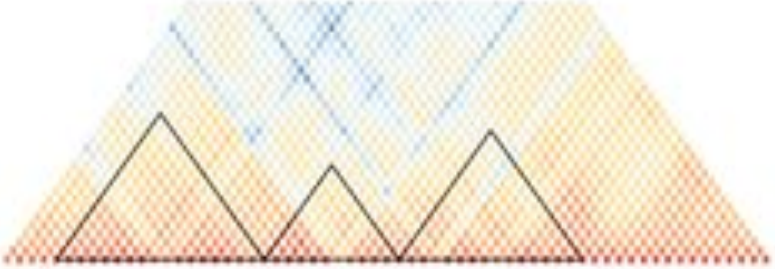
TopDom



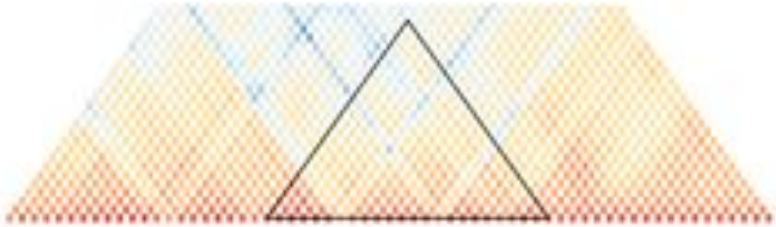
OnTAD



HiCSeg



rGMAP



SpectralTAD Package

- **Input**: three types of contact matrices ($n \times n$, sparse and $n \times (n+3)$) in text format, import from `.hic` and `.cool` files supported
- Two main functions: `SpectralTAD` and `SpectralTAD_Par` (parallelized)
- **Output**: A 3-column BED file for each hierarchy level
- Visualization options include output for `Juicebox`

Summary

- We propose a new approach for TAD detection based on spectral clustering, `SpectralTAD`
- `SpectralTAD` implements two novel methods (sliding window and eigenvector gap clustering) for improving clustering on ordered data with size restrictions
- Benchmarked against existing methods, `SpectralTAD` has shown a significant improvement on several criteria
- `SpectralTAD` has been released as an R package and is available on Bioconductor

Learn more

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

Kellen G. Cresswell, John C. Stansfield, Mikhail G. Dozmorov

doi: https://doi.org/10.1101/549170

Download PDF
Supplementary Material
Email
Share
Citation Tools
Tweet
Like 0

- `SpectralTAD` is available at http://bioconductor.org/packages/SpectralTAD/
- Slides are available at https://github.com/mdozmorov/Talk_JSM2019
- Preprint is available at https://www.biorxiv.org/content/10.1101/549170v2

PhRMA FOUNDATION
American Cancer Society®