TADCompare: Differential and Temporal Analysis of Boundaries of Topologically Associated Domains and Loops

We developed TADCompare, an R package aimed at providing a fast, accurate, user-friendly, and well-documented approach to differential and temporal analysis of boundaries of TADs and chromatin loops. We introduce a method based on the boundary score statistic [1] and use it to identify five types of boundary changes. The method is extended to allow for calling consensus boundaries and comparing them between groups of Hi-C replicates. We further demonstrate how the boundary score statistic may be used to analyze the dynamics of boundaries of interacting domains over the time course. For both differential boundary detection and time course analysis, we provide novel terminology for the classification of boundary changes. We demonstrated the robustness of TADCompare using simulated data with pre-defined interacting domains [2] and its ability to reveal distinct biological roles of different boundary changes. In summary, TADCompare provides an all-in-one pipeline from consensus boundary calling to differential boundary detection, including time course. The output is formatted in a commonly used BED format that allows for flexible downstream analyses and visualization.

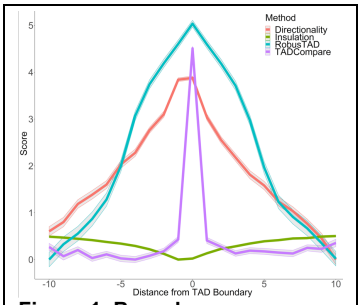

Figure 1. Boundary score distinguishes boundaries better than monotonic metrics. X-axis - Distance from the boundary, Y-axis - Score

Our previous work on TAD detection using spectral clustering, implemented as a SpectralTAD R package [1], introduced the concept of the boundary score statistic. Briefly, boundary score is calculated following eigenvector decomposition of the Laplacian matrix (Hi-C data), finding the distance between consecutive eigenvectors (eigenvector gap) and converting them into Z-scores. In contrast to other metrics for boundary identification, our boundary score spikes at the boundary (Figure 1), simplifying the distinction between boundaries and non-boundaries, and its magnitude is directly interpretable as a "boundary strength". Consequently, the consensus boundary score, defined as a median boundary score across multiple Hi-C datasets, allows for the robust detection of consistent boundaries and filtering out noisy boundaries detected in single datasets. The quantitative nature of the boundary score enables statistical comparison of TAD boundaries between two conditions and across timecourse.

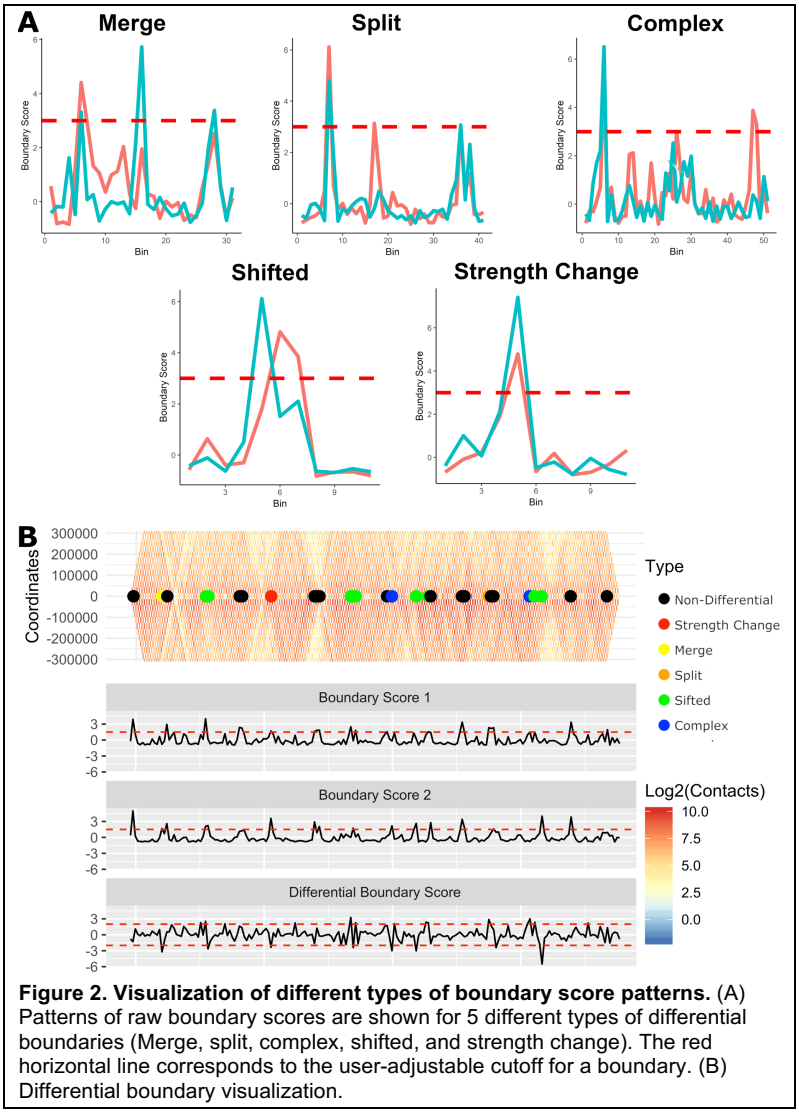

Figure 2. Visualization of different types of boundary score patterns. (A) Patterns of raw boundary scores are shown for 5 different types of differential boundaries (Merge, split, complex, shifted, and strength change). The red horizontal line corresponds to the user-adjustable cutoff for a boundary. (B) Differential boundary visualization.

Given that the Z-score-transformed boundary scores can be directly compared between the two datasets, we define five types of boundary changes (Figure 2). The “complex”, “merge”, and “split” boundaries are considered to be the most disruptive changes in the 3D structure of the genome.

Comparison of boundaries in technical and biological replicates of experimental Hi-C data identified higher proportion of stable boundaries in technical replicates. Similarly, replicates of Hi-C datasets obtained from cells shared more boundaries than those of tissues. We demonstrated that each type of differential boundary is associated with different levels of epigenomic enrichment (Figure 3). Similarly, gene enrichment analysis revealed distinct functional roles of genes near different types of boundary changes. In summary, our results suggest that “complex” and “merge-split” changes are biologically important alterations of the 3D structure of the genome.

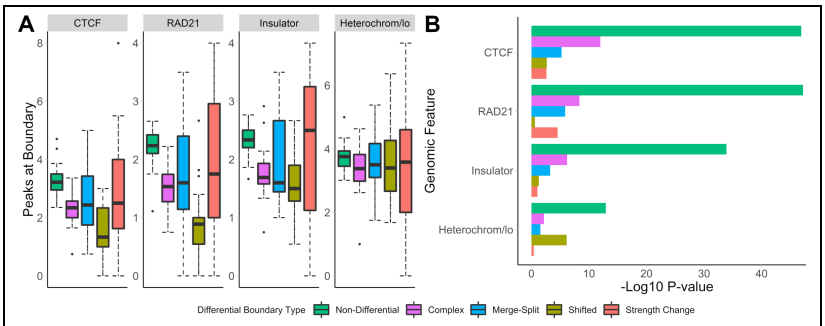

Figure 3. Non-differential boundaries are more enriched for selected genome annotation marks than other types of differential boundaries. Differential boundaries were called between GM12878 and IMR90 cell lines and categorized based on differential boundary type. (A) The number of peaks at boundaries and (B) permutation p-values (-log10) are shown.

The quantitative nature of boundary score allows us to monitor its changes at boundaries across any number of time points. Using the boundary score cutoff of 3 for boundary definition, we define six patterns of temporal boundary changes (Figure 4). Application of our method to the analysis of auxin-induced CTCF depletion and recovery experiment (HCT-116 cells, post-auxin-treated timecourse at 20, 40, 60, and 180 min after auxin treatment) revealed the expected pattern of early- and late-appearing TAD boundaries. Epigenomic and functional enrichment analyses suggest that disappearing and dynamic boundaries are likely detected due to noise in the data, while boundaries appearing after auxin treatment expectedly represent the biologically relevant signal.

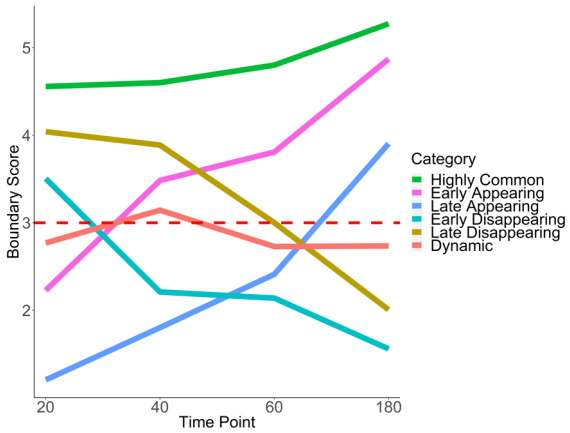

Figure 4. Six patterns of boundary score change across time. Average trajectories for each pattern of boundary score change are shown. The red horizontal line indicates the cutoff for boundary detection.

Our results demonstrate the ability of TADCompare to provide accurate, biologically relevant results. The methods implemented span differential, timecourse, and consensus analysis. To date, TADCompare is the only actively maintained and publicly available tool to provide this functionality. We intend for TADCompare to be a one-stop tool for comparison of HiC datasets, providing simple, easy-to-interpret results in a timely manner. TADCompare is pending Bioconductor approval and is available at https://github.com/dozmorovlab/TADCompare

References

1. Cresswell KG, Stansfield JC, Dozmorov MG: **SpectralTAD: An r package for defining a hierarchy of topologically associated domains using spectral clustering**. *bioRxiv*,:549170 10.1101/549170 Available: http://biorxiv.org/content/early/2019/02/13/549170.abstract.

2. 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**:679–685 10.1038/nmeth.4325.