TADBIT: A PACKAGE FOR THE DETECTION AND ANALYSIS OF TOPOLOGICALLY ASSOCIATING DOMAINS

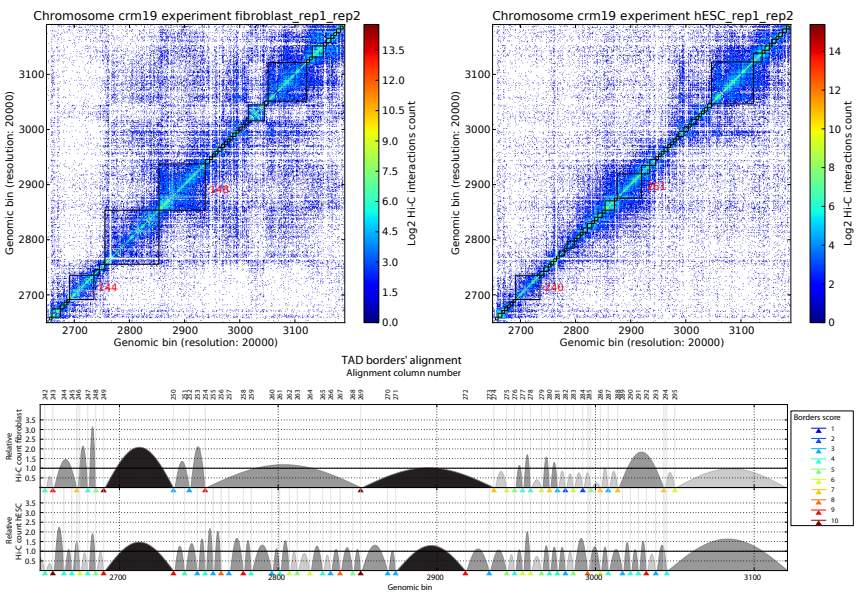
François Serra[1,2], Davide Baù[1,2], Guillaume Filion[2] and Marc A. Marti-Renom[1,2]

[1]Genome Biology Group. National Center for Genomic Analysis (CNAG), Barcelona, Spain. [2]Gene Regulation, Stem Cells and Cancer Program, Centre de Regulació Genòmica (CRG), Barcelona, Spain.

In the last years important advances have been done in the characterization of chromatin structure, especially due to the development of the Chromosome Conformation Capture (3C) technology. 3C-based technologies allow generating contact maps of a genome, at resolutions ranging from few Kb to few Mb. These technologies have revealed that chromatin is organized into the so-called Topologically Associating Domains (TADs), regions of the genome that are enriched in interactions [1,2]. Moreover, it has been shown that TAD localization in the genome is conserved between cell types and during evolution [2]. Here we present TADBit, a computational package that deals with 3C-based data to detect TAD borders and to further analyze TADs by: i) the identification of conserved TAD borders and ii) the definition of sets of TADs sharing structural features.

A Python library for analyzing Hi-C experiments

Hi-C experiments generate genomic interaction data that is usually summarized at the chromosome level. TADBit is built around the concept of chromosome, and uses it as a central item to compare different Hi-C experiments. The library has been designed to be used by researchers with minimal level of expertise in computer science; the *all-in-one* scripts provided in TADBit allow to run a full analysis in one single command line, while advanced users may produce their own programs using TADBit as a complementary tool.



Top panel. Filtered and normalized Hi-C data for the end of Chr19 in human fibroblast and hESC interaction matrices. The data shows interaction counts merged from two independent replicas. TAD are detected by TADBit and shown as black boxes within the matrix. Numbers highlighted in red show internally similar TADs (see next figure). **Bottom panel**. Alignment of TAD borders detected for the two cell types (fibroblast and hESC). TAD borders are highlighted by a vertical grey line and the score of the border. Hight of the arch is proportional to the relative internal Hi-C counts of the TAD. Highlighted arches in black indeicate TADs compared in next figure.

Alignment of TAD borders

Once identified, TAD borders can be aligned to characterize their conservation between two datasets of interactions. Two algorithms are proposed here, one based on the Needleman-Wunsch algorithm and the other based on a reciprocal best hit approach.

Comparison of TADs

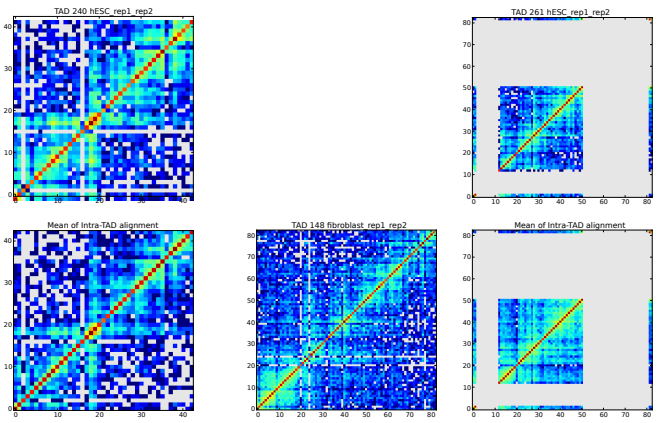
TADs can also be compared directly from their Hi-C data matrices, in a similar fashion as how it is done for protein contact maps [4]. TADBit aligns TAD internal matrices by calculating their principal eigenvectors and using a dynamic programming algorithm (Needleman-Wunsch) that will produce a final alignment of the matrices.

Interaction data filtering and normalization

TADBit has implemented a series of functions that aim at filtering and normalizing binned 3C-based data. For example, TADBit will automatically remove columns from a Hi-C matrix that contain a biased distribution of Hi-C counts toward low values (*e.g.,* centromeric columns will be removed from the analysis). Then, with a single line of code, the Hi-C data matrix can be normalized according to the bin's "visibility" [3].

TAD identification

TADBit can detect the position of TAD border's through a breakpoint detection algorithm that returns an optimal segmentation of a given region (*e.g.,* a chromosome) according to an estimation of likelihood (BIC-penalized). The model assumes that Hi-C counts have a *Poisson* distribution and that the expected value of the counts decreases like a power-law with the distance of the chromosome. Finally TADBit calculates a confidence score of the detection by penalized dynamic programming.



Left panel. Alignment of Hi-C internal interaction matrix for TADs 144 and 240 for chromosome 19 in the fibroblast and hESC experiments, respectively. The compared TAD are conserved between the two experiments with a r^2 between the matrices of 0.83. **Right panel**. Alignment of Hi-C internal interaction matrix for TADs 148 and 261 for chromosome 19 in the fibroblast and hESC experiments, respectively. Even though a large gap needs to be introduced, the compared TAD are still conserved (r^2 = 0.43).

TADBit for the identification and comparison TADs

TADBit introduces a set of computational tools for the identification, comparison and 3D modeling of TADs. TADBit, as a python library, can be used or "called" from other programs. This tool, designed to be extended with new features, represents the first framework for the analysis of TADs and sets conventions for the representation of the results. Please, write to mmarti@pcb.ub.cat if your are interested in getting TADBit.

Bibliography

1. *Nora, E. P. et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature, 485(7398), 381–5.*
2. *Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., … Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 485(7398), 376–80.*
3. *Imakaev, M. et al. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Meth 9, 999–1003.*
4. *Di Lena, P., et al. (2010). Fast overlapping of protein contact maps by alignment of eigenvectors. Bioinformatics, 26(18), 2250–8.*

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

CRG Centre de Regulació Genòmica

HFSP

>> more information at http://marciuslab.org <<