Supplementary files to:

Hierarchical chromatin organization detected by TADpole.

Paula Soler-Vila[1,†], Pol Cuscó[2,†], Irene Farabella[1], Marco Di Stefano[1,*] and Marc A. Marti-Renom[1,3,4,5,*]

[1]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain.

[2]Gastrointestinal and Endocrine Tumors Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain.

[3]Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain.

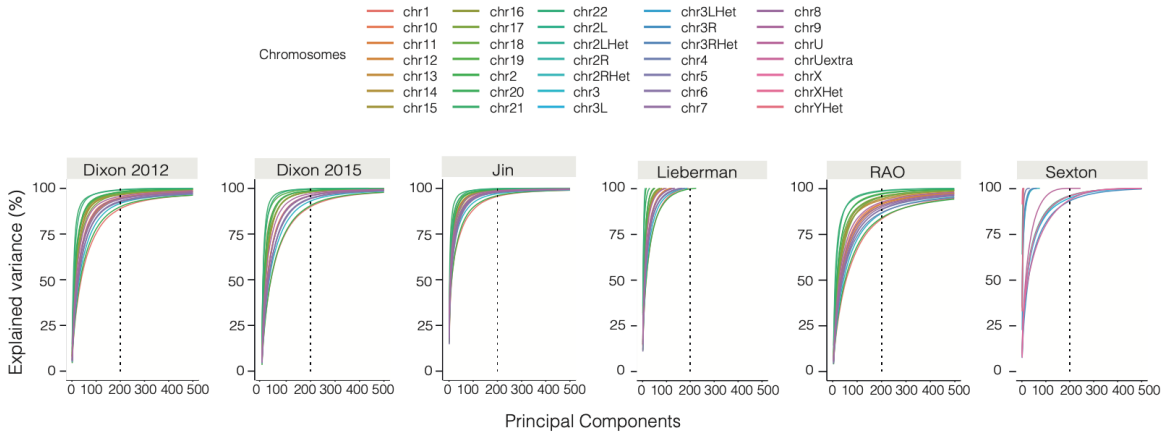
[4]Universitat Pompeu Fabra (UPF), Barcelona, Spain.

[5]ICREA, Barcelona, Spain.

†Joint first authors

*To whom correspondence should be addressed. Emails: martirenom@cnag.crg.eu & marco.distefano@cnag.crg.eu

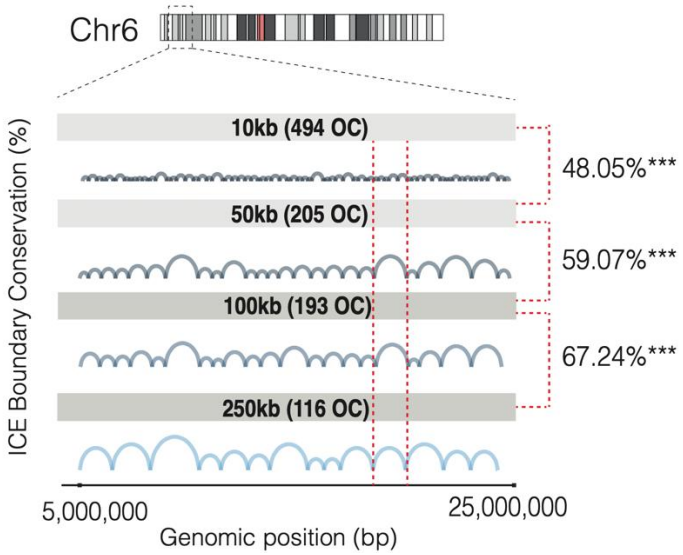
A



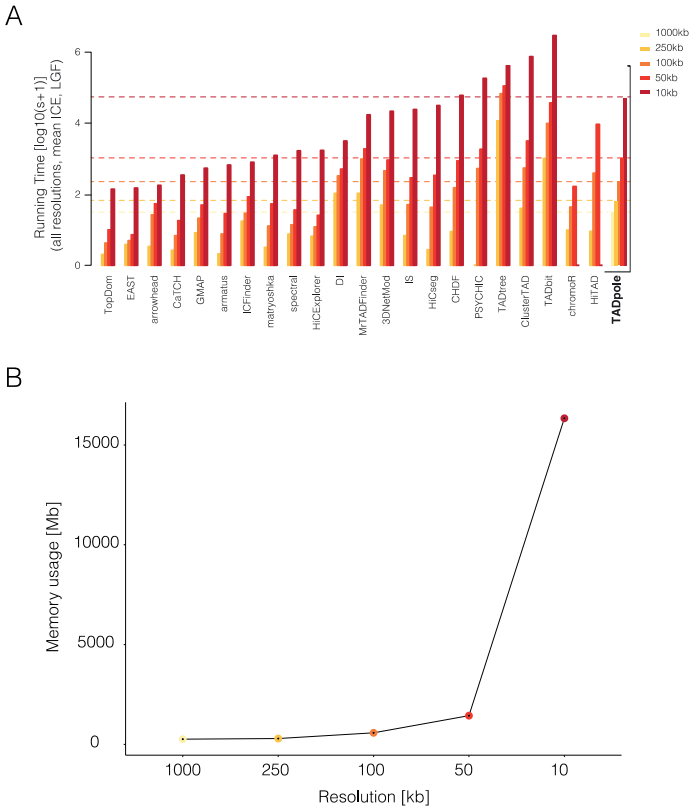
B

Dataset	Cell type	Enzyme	ID	Filtered_reads	Resolution
Lieberman-Aiden	GM06990	HindIII	SRR027956	3,653,331	1Mb
Sexton	Fly Embryo	DpnII	SRR389762, SRR389763,SRR389764,SRR389765,SRR389766,SRR389767,SRR389768	47,321,181	40kb
Dixon_2012	H1-hESC	HindIII	SRR400260,SRR400261,SRR400262,SRR400263	22,912,612	40kb
Jin	IMR90	HindIII	SRR639030,SRR639031,SRR639032,SRR639033	167,135,412	40kb
Rao	GM12879	MboI	SRR1658602	64,941,983	40kb
Dixon 2015	H1-hESC	HindIII	SRR1030718,SRR1030719,SRR1030720,SRR1030721	221,757,193	40kb

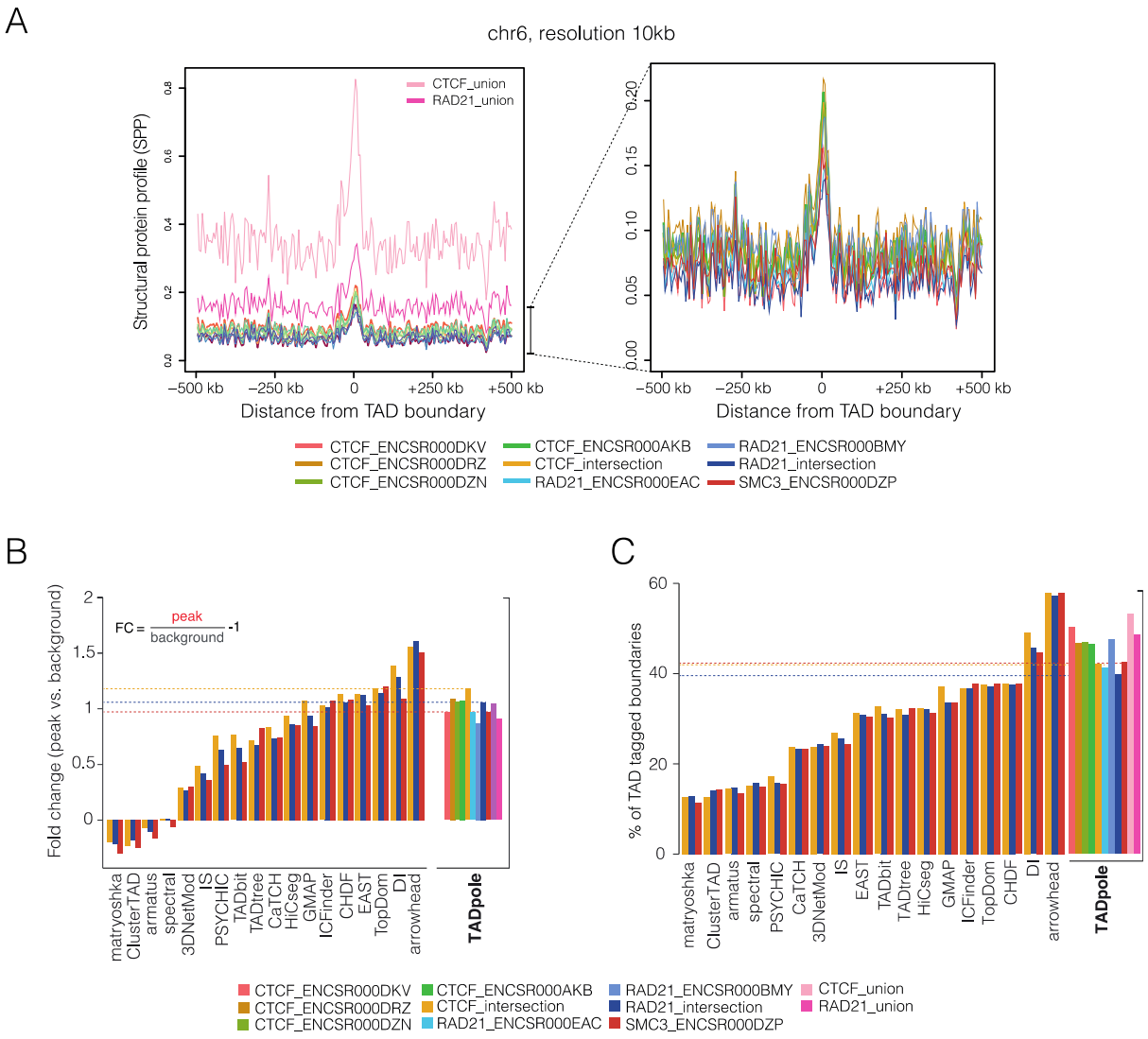
Supplementary Figure 1. Percentage of explained variance as a function of the number of retained principal components for various datasets. (A) Each continuous line represents a different chromosome, and the vertical dashed lines mark the default number (200) of first PCs (N_{PCs}) retained by TADpole. **(B)** The six Hi-C datasets used, identify by: cell type, restriction enzyme, the NCBI accession numbers, number of the valid reads retrieved after filtering using an in-house pipeline based on TADbit (56), and binning size. Datasets with multiple NCBI entries were merged and (after filtering) the resulting matrices were binned using an equal bin-width of 40kb, with the exception of Lieberman-Aiden dataset (13) which was binned at 1Mb.



Supplementary Figure 2. Percentage of conserved TADs boundaries across different resolutions on the entire chromosome 6. The diagram illustrates the analysis on a random *locus* from 5 to 25Mb. The p-value is computed using a shuffle test **(Material and Methods)**.

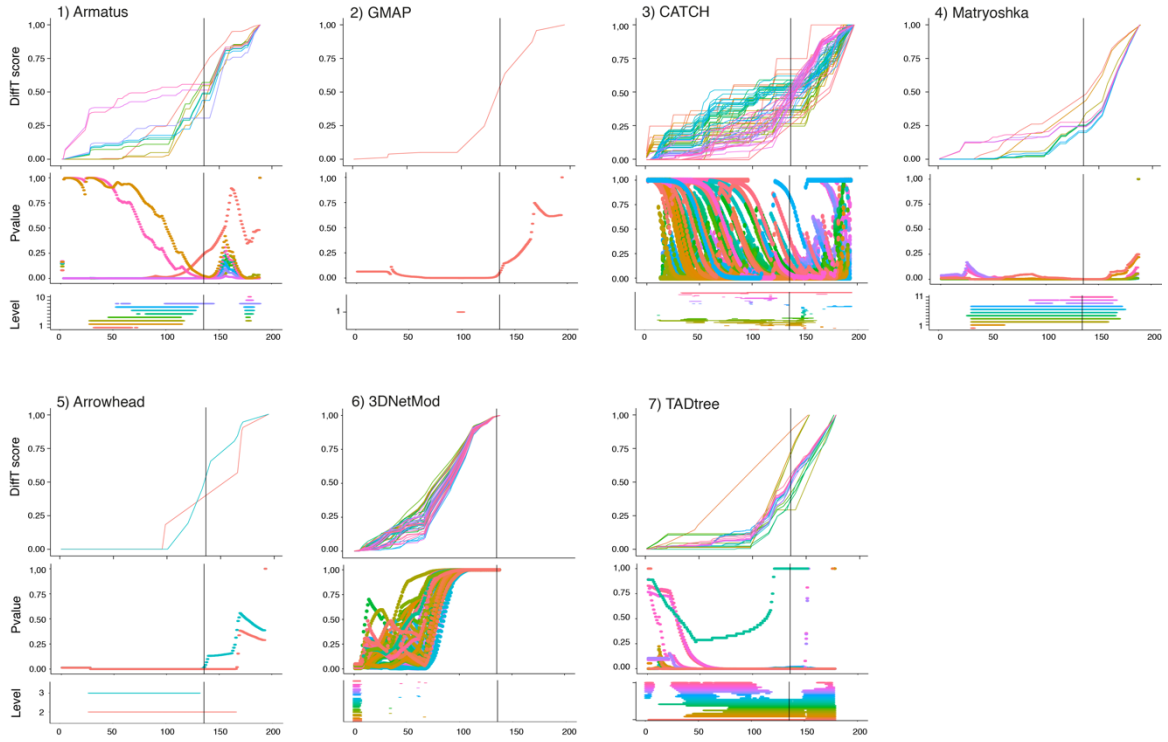


Supplementary Figure 3. Computational analysis of TADpole. (A) Execution time (in logarithmic scale) of the all TAD callers analyzed. The average value computed between the two normalization strategies (ICE and LGF) is shown across resolutions (1000kb, 250kb, 100kb, 50kb, 10kb). **(B)** Memory usage test of TADpole. Each dot represents the maximum memory usage computed for LGF normalization matrices across different resolutions (1000kb, 250kb, 100kb, 50kb, 10kb).



Supplementary Figure 4. Biological replicas benchmarking. (A) ***Right***: Structural protein profiles (SPPs) per sample type: individual replicas, union and intersection. ***Left***: Zooming on the SPPs of individual replicas and intersection profiles. **(B)** The fold-change of CTCF, RAD21 and SMC3 at domain borders and **(C)** The percentage of identified TADs boundaries occupied by CTCF, RAD21 and SMC3 per sample type in TADpole compared with other 22 TAD callers.

Supplementary Figure 5



Supplementary Figure 5. DiffT score profiles across 8 different hierarchical TAD callers. The DiffT score profiles as a function of the matrix bins for each tool are represented in rows 1 and 4. The p-value profiles per bin for automated detection of significant differences are represented in rows 2 and 5. The bin(s) associated with the minimum p-values per level are represented in rows 3 and 6. Note that only the levels containing at least one bin with a DiffT score associated p-value < 0.05 are shown. In all the panels, the different hierarchical levels recovered by each tool have a distinctive color, while the Inv1 breakpoint is highlighted with a solid black line.

Chromatin Marks	Encode ID
CTCF	ENCSR000DRZ, ENCSR000DKV, ENCSR000DZN, ENCSR000AKB
SMC3	ENCSR000DZP
RAD21	ENCSR000BMY, ENCSR000EAC
H3K4me3	ENCFF295GNH
H3K36me3	ENCSR000DRW
H3K27me3	ENCSR000DRX
H3K9me3	ENCFF138CTR, ENCFF331ODM, ENCFF782FRS

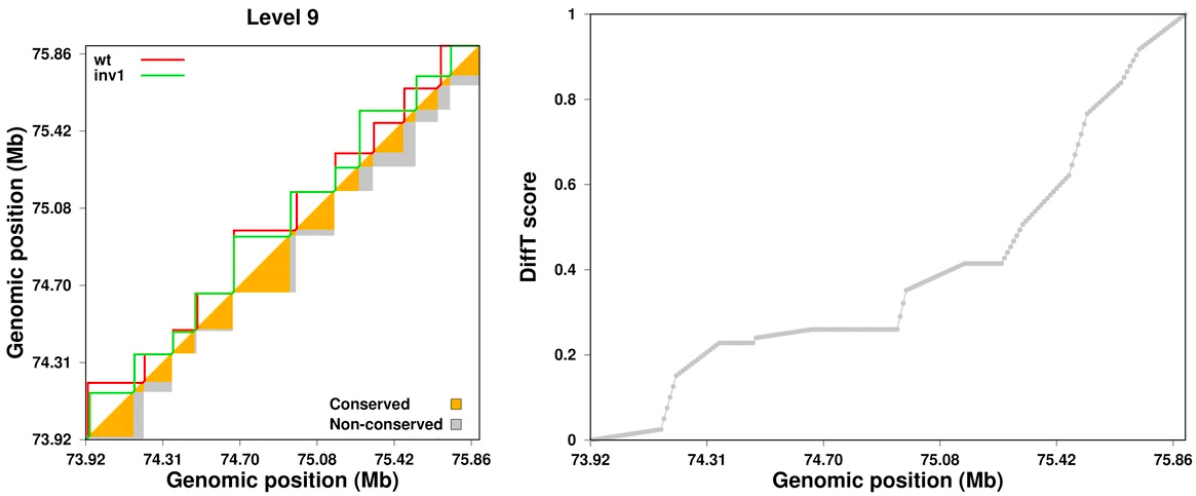
Supplementary Table 1. Encode IDs of the ChIP-seq experiments used in the biological benchmarking analysis.

TAD CALLER	PARAMETERS	LEVEL SELECTION
Armatus and Matryoshka (33,38)	Default parameters were used: -g 0.5 -s 0.05 -n 100 -m.	The multiscale domains (-m option) were used and the caller provided one partition per each value of Gamma from 0 to 0.5 in steps of 0.05.
Arrowhead (15)	Default parameters were used: -m 2000.	Levels were selected manually considering larger domains first.
TADtree (35)	Default parameters were used: S = 50, M = 25, p = 3, q = 12, and gamma = 500.	Levels were selected setting the maximum proportion of duplicate TADs to 2%.
CaTCH (36)	No parameters	We considered all the distinct partitions generated using reciprocal insulation values from 0 to 1.
GMAP (37)	Default parameters were used, but for maxDistInBin set to 195.	We considered the levels based on the provided domain order.
PSYCHIC (39)	The window size was set to 0.5Mb.	Levels were provided automatically in the *hierarchy.bed* file.
3DNetMod (34)	Overlap=0, region_size=194, badregionfile=None, badregionfilter=False, diagonal_density=0, consecutive_diagonal_zero=19, scale=chr1, plateau=3, chaosfilter=False, num_part=20, pctile_threshold=0, pct_value=0, size_threshold=4, variance_type=percent, size_s1=4000000, size_s2=12000000, var_thresh1=var_thresh2=100,	We grouped the identified domains in levels by applying four criteria: *(i)* the maximum gap between consecutive domains was <=2 bins, (*ii*) at fixed level domain overlap was not allowed, *(iii)* the hierarchy was build ordering the partitions by the total number of domains from low to high, and *(iv)* partitions with the same number of domains were placed in the hierarchy from high to low average domains size.

Supplementary Table 2. Description of the parameters used and the level selection process followed by each hierarchical TAD caller.

Resolution	Raw N° of TADs	Raw Size (kb)	Raw Size (bins)	ICE N° of TADs	ICE Size (kb)	ICE Size (bins)	LGF N° of TADs	LGF Size (kb)	LGF Size (bins)
250kb	118	1425,85kb	5.7	116	1465.52kb	5.86	120	1402.08kb	5.61
100kb	193	868,91kb	8.68	193	879.79kb	8.79	193	884.46kb	8.45
50kb	217	772,35kb	15.45	205	814.39kb	16.29	208	805.77kb	16.12
10kb	535	313.01	31	494	338.70kb	33.87	510	328.29kb	32.83

Supplementary Table 3. The total number of TADs and the corresponding average size detected in raw and normalized Hi-C matrices (by ICE and LGF) across different resolutions.



Supplementary video 1. Calculation of the DiffT score for the 9th level of the dendrogram (Figure 4B and C). The video displays two related synchronized panels. (*Left)* The upper triangle of the matrix shows the TADs borders identified by TADpole in WT and Inv1 matrices as red and green continuous lines, respectively. During the video, the matrix is scanned from the first to the last bin, and simultaneously the lower triangle gets filled with the areas of the TADs that are conserved (in orange) or non-conserved (in gray) between the two partitions. The DiffT score is computed as the normalized sum of the non-conserved (gray) areas. *(Right)* DiffT score profile *versus* the genomic position grows proportional to the gray areas appearing over time in the left panel.