Efficient and Accurate Detection of Topologically Associating Domains from Contact Maps*

Abbas Roayaei Ardakany[1] and Stefano Lonardi[2]

1 **Department of Computer Science, University of California, Riverside, CA, USA**

2 **Department of Computer Science, University of California, Riverside, CA, USA**

Abstract

Continuous improvements to high-throughput conformation capture (Hi-C) are revealing richer information about the spatial organization of the chromatin and its role in cellular functions. Several studies have confirmed the existence of structural features of the genome 3D organization that are stable across cell types and conserved across species, called *topological associating domains* (TADs). The detection of TADs has become a critical step in the analysis of Hi-C data, e.g., to identify enhancer-promoter associations. Here we present EAST, a novel TAD identification algorithm based on fast 2D convolution of Haar-like features, that is as accurate as the state-of-the-art method based on the directionality index, but 75-80× faster. EAST is available in the public domain at `https://github.com/ucrbioinfo/EAST`.

1998 ACM Subject Classification F.2.1 Numerical Algorithms and Problems I.4.6 Image Segmentation J.3 Life and Medical Sciences

Keywords and phrases Chromatin, TADs, 3D genome, Hi-C, contact maps

Digital Object Identifier 10.4230/LIPIcs.WABI.2017.22

1 Introduction

Recent studies have revealed that genomic DNA is not arbitrarily packed into the nucleus. The chromatin has a well organized and regulated structure in accordance to the stage of the cell cycle and environmental changes [15, 16]. The structure of chromatin in the nucleus plays a critical role in gene expression, epigenetic organization, and DNA replication, among others [7, 6, 18, 17].

With the advent of genome-wide DNA proximity ligation (Hi-C), life scientist have shed new light on the way that chromatin folds and its relation to cellular functions [13, 1, 2, 10]. The analysis of Hi-C data has revealed surprising new findings including the discovery of new structural features of chromosomes such as topologically associating domains [7] and chromatin looping [17].

Topological associating domains (TADs) are large, megabase-sized contiguous local chromatin interaction domains that have a high average interaction within and a low average interaction with their surrounding regions. Because of the role that TADs play in cellular functions they have been widely explored since their discovery. TADs are stable across different cell types and highly conserved across species [7]. TADs tend to interact with each other in a tree-like structure and form a hierarchy of domains-within-domains (metaTAD),

* This work was partially supported by NSF IIS-1526742.

17th International Workshop on Algorithms in Bioinformatics (WABI 2017).
Editors: Russell Schwartz and Knut Reinert; Article No. 22; pp. 22:1–22:11
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

which can scale up to the size of chromosomes [9]. metaTADs show correlation with genetic and epigenomic features. TAD boundaries are enriched for the insulator binding protein CTCF, housekeeping genes, transfer RNAs and histone modifications [7, 8]. More importantly, enhancers tend to interact with gene promoters within the same TAD [11]. Disruption of TAD boundaries can affect the expression of nearby genes and lead to developmental disorders or cancer [14].

Several methods have been developed to identify TADs genome-wide. Dixon *et al.* were the first group to define and identify TADs [7]. In their seminal work, they proposed an identification method based on the *directionality index* (DI) which measures the frequency of interaction of a genomic locus with a fixed-sized neighborhood. Drastic changes of the DI score are expected at TAD boundaries where the region tends to have a high rate of both upstream and downstream interactions.

Filippova *et al.* [8] introduced a single parameter, two-step dynamic programming method. Assuming that there exist a few characteristic resolutions across which TADs are similar, they identify a set of non-overlapping domains that are persistent across the resolutions.

Crane *et al.* [4] proposed a method based on the *insulation index* (IS). For each chromosome segment, IS score is the average number of interactions that cross the segment in a pre-specified size neighborhood. Given that interactions tend to be isolated within TADs, IS local minima are expected to occur at TAD boundaries. The IS score can be computed efficiently by sliding a window across the diagonal of the contact matrix and computing the average number of interactions that fall inside the window.

Chen *et al.* [3] translated the TAD identification problem into a graph segmentation/-clustering problem. In this method, domains at different scales are identified by running the spectral graph cuts algorithm recursively until the connectivity of the graph reaches some predefined threshold.

In this paper, we present a novel TAD calling algorithm called East (for "Efficient and Accurate Summed-area-table-based TAD calling") that takes advantage of fast 2D convolution. Experimental results show that East is as accurate in detecting TADs as the DI method [7], which is accepted as the state-of-the art algorithm. East is however, 75-80× faster than DI.

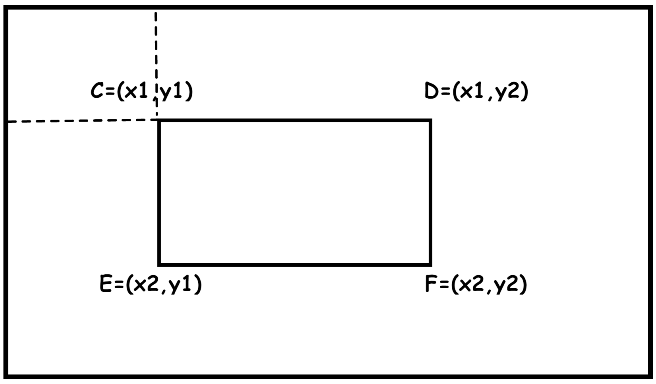
2 Methods

Each chromosome is segmented into evenly sized fragments, where the size of the segment is based on the resolution of the data. In a Hi-C *contact map* (or *interaction matrix*) A, each entry $A[i,j]$ represents the number of times segments i and j are observed together in a DNA proximity ligation experiment. Larger values of $A[i,j]$ indicate closer loci i and j in 3D space inside the nucleus. Segments that are close in genomic 1D distance tend to form dense areas which can be seen as isolated high frequency blocks along the matrix diagonal, namely, TADs. TADs have high intra-frequency within and low inter-frequency with their neighboring blocks. The aim is to identify TADs efficiently and accurately.

We propose an algorithm called East that utilizes rectangular Haar-like features [21] and dynamic programming to identify TADs. Genomic regions are scored based on an objective function that measures their likelihood of containing a TAD with respect to the characteristics mentioned above. We use Haar-like features to describe such a scoring function.

2.1 Summed area table and Haar-like features

A *Haar-like feature* is a set of adjacent rectangular regions each of which has a certain weight. Weights of rectangular regions indicate certain characteristics of a particular area of



$$\sum_{\substack{x_1 \leq x \leq x_2 \\ y_1 \leq y \leq y_2}} A(x,y) = A_{\text{SAT}}(C) + A_{\text{SAT}}(F) - A_{\text{SAT}}(D) - A_{\text{SAT}}(E)$$

Figure 1 If the summed area table A_{SAT} is available, computing the sum of values in any rectangular region takes $O(1)$ time.

the image. By convolving Haar-like features, i.e., by computing the weighted sum of pixel values for a particular location, we obtain a value that represents how well a region (window) satisfies the characteristics we are looking for. To compute the weighted sum efficiently we use the summed area table.

A *summed area table* (SAT), also known as *integral image* in computer vision, is a data structure used for efficiently calculating the sum of values in a rectangular region. By precomputing the summed area table one can obtain the sum of values in any arbitrary rectangular region using only a constant number of operations. SAT was first introduced to computer graphics in 1984 by Frank Crow [5] and later to computer vision in 2001 by Lewis [21] in a popular face detection framework called Viola-Jones. The value of a point (x, y) in a summed area table A_{SAT} is the sum of all pixels above and to the left of that point in the original grid A, including the (x, y) point itself.

$$A_{\text{SAT}}(x,y) = \sum_{x' \leq x,\ y' \leq y} A(x', y') \,.$$

Since the value of each point in the SAT can be computed based on the values of neighboring points, the formula can be rewritten as

$$A_{\text{SAT}}(x,y) = A(x,y) + A_{\text{SAT}}(x-1,y) + A_{\text{SAT}}(x,y-1) - A_{\text{SAT}}(x-1,y-1)\,.$$

Given the summed area table, computing the sum of values in an arbitrary size rectangular region can be done in $O(1)$ time (see Figure 1).

2.2 TAD objective function

To score TADs we need to define a function f that quantifies the quality of an arbitrary region along the matrix diagonal with respect to the following properties:

1. The average frequency inside the region must be “high”.
2. The average frequency with the neighborhood must be “low”.
3. The average frequency between start and end segments of the region must be higher than the average frequency inside the region.

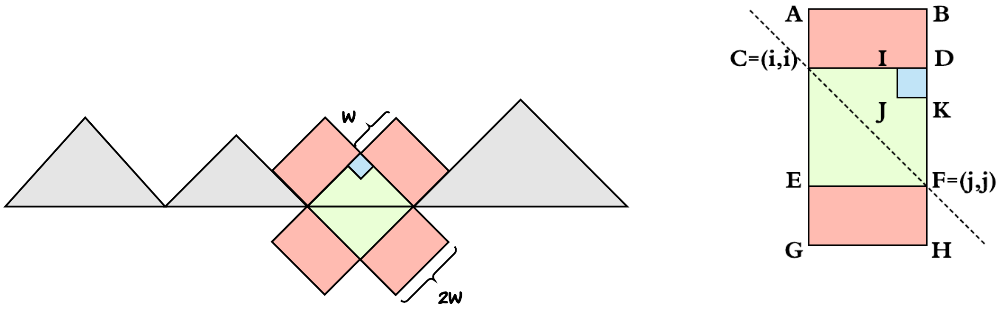


Figure 2 Objective function f. (LEFT) Representation of a TAD of size $2w$. High interaction frequency expected inside the TAD's domain (green) while low interaction frequency is expected between the TAD and surrounding domains (red) (RIGHT) Coordinates of Haar-like representation of a TAD.

The last property derives from the fact that TADs are the result of a compact locality or loop formation in the chromatin. To explain the design of the objective function f we refer to Figure 2, where different colors indicates different weighting. The area in green color is the region we expect to have a high frequency of interaction (intra-frequency), as opposed to the area in red where lower frequency is expected (inter-frequency). The corner region which is colored in blue in Figure 2 has a higher weight in order to account for the last property in the list above. Using the SAT data structure, function f can be computed as follows.

$$f([i,j]) = \frac{CDEF^{\diamond} - \alpha \cdot (ABGH^{\diamond} - CDEF^{\diamond}) + \beta \cdot IDJK^{\diamond}}{\mathcal{N}}$$

where $CDEF^{\diamond}$, $ABGH^{\diamond}$ and $IDJK^{\diamond}$ represent the sum of pixel values inside the rectangular regions $CDEF$ (defined by interval $[i,j]$), $ABGH$ and $IDJK$ respectively, which can be computed in $O(1)$ time from the SAT of the interaction matrix A. Parameters α and β are dataset-independent, and they can be determined experimentally. Parameter $\mathcal{N}$ is a normalization factor discussed in Subsection 3.1.

2.3 Finding the optimal set of domains

Given a $n \times n$ interaction matrix A, the problem of TAD identification is an optimization problem aimed at identifying the set of contiguous non-overlapping domains for which the

$$\sum_{d_i \in D} f(d_i)$$

is maximized, where $D = \{d_i | d_i = [s_i, e_i]\}$ is a set of non-overlapping intervals, i.e., $e_j < s_i$ or $e_i < s_j$ for $i \neq j$.

We use dynamic programming to solve this optimization problem. The optimal solution $OPT(i)$ for the sub-problem $[1, i]$ can be expressed by following recurrence relation

$$OPT(i) = \max_{0 \le k \le i-1} \{OPT(k) + f([k+1, i])\}.$$

By gradually increasing the size of the sub-problem and keeping track of the set of extracted domains, the optimal set of TADs for the entire interaction matrix can be computed. As we grow the size of the sub-problem, for each bin i, we need to find the optimal location to break the sub-problem $[1, i]$ into a sub-problem $[1, k]$ and a domain $d = [k + 1, i]$. The overall time-complexity is $O(n^2)$, where n is the number of bins/segments.

If we do not allow TADs to be larger than L, the optimal break point for a sub-problem $[1, i]$ can always be found in the interval $[max\{i - L, 0\}, i - 1]$. Therefore, the overall time complexity decreases to $O(nL)$.

▶ **Theorem 1.** *Let $D^* = \{[a_1, a_2], [a_2, a_3], \ldots, [a_{s-1}, a_s]\}$ be an optimal set of domains for the interaction matrix A for which*

$$\sum_{d_i \in D^*} f(d_i)$$

is maximized. Then,

$$OPT^*(n) = OPT(n)$$

where

$$OPT^*(i) = \max_{max\{i-L,0\} \leq k \leq i-1} \{OPT^*(k) + f([k + 1, i])\} .$$

Proof. We prove the theorem by induction. For the base case $OPT^*(a_1) = OPT(a_1) = 0$. Now, suppose $OPT^*(a_{i-1}) = OPT(a_{i-1})$ then we have

$$\begin{aligned} OPT^*(a_i) &= OPT^*(a_{i-1}) + f([a_{i-1} + 1, a_i]) \\ &= OPT(a_{i-1}) + f([a_{i-1} + 1, a_i]) \\ &= OPT(a_i) \text{ for } k = a_{i-1} \end{aligned}$$

where k satisfies the inequality $\max\{a_i - L, 0\} \leq k \leq a_i - 1$. ◀

3 Experimental results

We performed the analysis on Hi-C data for two mouse cell types (cortex and embryonic stem cell), and one human cell type (embryonic stem cell) at bin resolution of 40kb. The Hi-C data was obtained from [7].

3.1 Parameter settings

In addition to α, β and L, East relies on two additional parameters. The first is the minimum quality threshold τ that is used to filter out low-quality TADs. If we assume that TAD quality scores are distributed according to a Gaussian distribution, we define the threshold $\tau = \mu - \sigma$ where μ and σ are the mean and standard deviation of the distribution of scores. Observe that parameter τ can be computed from the analysis of the dataset.

The second parameter is the normalization parameter $\mathcal{N}$ for the function f. Since the quality measure f is proportional to the sum of interactions inside the domains, f grows as the TAD size increases. Figure 3 illustrates how the sum of interactions inside a domain grows as the size of the TAD increases for the three datasets used in the experimental results below and for a synthetic interaction matrix. In the synthetic data, the number of interactions was set to be inversely proportional to the genomic distance. For purpose of comparison, the sum of interactions is normalized by the sum of the largest domain.

Observe that the curve for the mouse embryonic data roughly matches the curve for the synthetic data. This suggests that the average interaction frequency of two loci in the

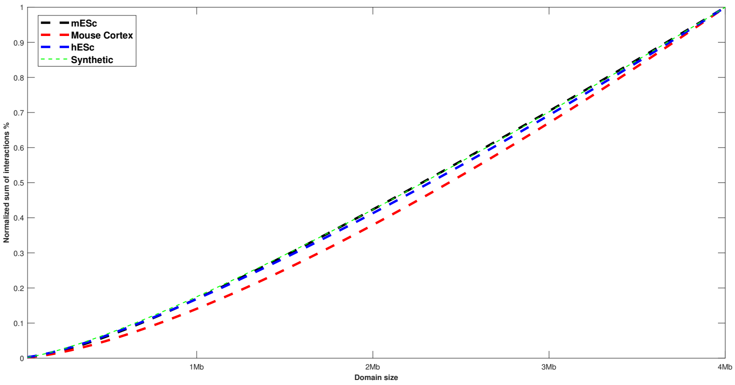


Figure 3 Growth of the quality measure f as the size of the TAD increases on the three datasets used in the experimental results and for a synthetic interaction matrix (see text).

mESC dataset is inversely proportional to their genomic distance. The growth function of the synthetic data can be estimated by $(n/L)^{1.2}$ where L is the largest domain size we are evaluating.

Also observe the hESC and mouse cortex curves are slightly different from the curve for the synthetic data, and they can be estimated by $(n/L)^{1.36}$ and $(n/L)^{1.4}$ respectively. We experimentally determined that as the curves diverge from the curve for the synthetic data, the normalization factor needs to adjusted accordingly. We set $\mathcal{N} = n^{0.4}$, $\mathcal{N} = n^{0.43}$ and $\mathcal{N} = n^{0.38}$ for hESC, mESC and mouse cortex, respectively. Parameters α and β were optimized experimentally to values $\alpha = 0.2$ and $\beta = 0.2$, and they are dataset-independent.

3.2 Comparison with existing methods

Based on the availability and popularity of TAD calling methods, we decided to compare East with the directionality index method [7], insulation score method [4] and multiscale method in [8]. We hereafter refer to these methods as DI, INS and MR respectively.

East, DI, INS and MR were ran on an Intel Core-i7 2.7GHz CPU with 16GB of memory. For the DI method we ran the experiments with posterior marginal probability threshold 0.99 and up/downstream span size of 2Mb (default parameters according to [7]). For the INS method, we set the insulation delta span to 200kb and the insulation square size to 500kb. For the MS method, we set the highest resolution parameter to 0.5.

In our experiment we investigated the enrichment of epigenetic characteristics of chromatin near the TAD boundaries. Although the mechanism behind the formation of TADs and their role in gene regulation are not fully understood, multiple studies have shown that some proteins and histone marks are enriched at the TAD boundary regions, implying that these boundaries play a role in gene transcription. As it was done in other studies [8, 20, 3], we can therefore use these genomic markers to evaluate the quality of the computed TADs.

To produce enrichment plots, we used each method to determine the boundary locations of TADs. Then, the frequency of each marker was calculated in 10kb bins in a window of 1Mb centered at the TAD boundaries. Each plot show the distribution of specific markers for each tool in the region centered at the TAD boundaries.

For mouse cortex and stem cells we evaluated the enrichment of transcription factor CTCF, promoter related marks RNA Polymerase II and H3K4me3, and enhancer-related histone modification H3K27ac. This marker data was collected from [19]. For human stem

Table 1 Running time of East, INS, MS and DI on the three datasets used in this work.

	hESC	mESC	Cortex
East	58s	50s	48s
INS	52s	44s	42s
DI	4,721s	3,845s	3,628s
MS	762s	545s	520s

cells we assessed the enrichment of CTCF near TAD boundaries. The CTCF data was obtained from [12].

Figure 4 shows that CTCF binding sites are almost twice as enriched near the TAD boundaries than the surrounding regions, suggesting that TAD boundaries are associated with insulator genomic regions and their mediator protein CTCF. Figure 5 and Figure 6 show that promoter marks RNA Polymerase II and H3K4me3 peak within the TAD boundaries for both mouse cortex and embryonic stem cells. Observe in Figure 7 that histone modification mark H3K27ac is highly enriched around TAD boundaries in mouse embryonic stem cells but not in mouse cortex cells. Also observe in Figure 8 that enhancer marks are highly depleted around TAD boundaries in mouse cortex cells but not in mouse embryonic stem cells.

Overall, observe in Figures 4–8 that the blue curve for East is almost always higher than the other three tools, suggesting that our tool generates TADs with very accurate boundaries. The closest competitor is DI (green curve), but East is significantly faster than DI.

We compared the running time of East with that of DI, MS and INS on Hi-C data for human embryonic stem cells, mouse embryonic stem cells and mouse cortex [7]. Table 1 shows that East and INS are comparable in speed, MS is 10–14× slower, DI is 75-80× slower.

Figure 9 illustrates the size distribution of TADs for all four methods for the human embryonic stem cells. The numbers of TADs extracted by East, DI, MS and INS are 2229, 2429, 12427 and 4708 respectively. Observe that East and DI roughly produce the same size distribution.

In summary, these experimental results show that while East can identify the TAD boundaries as accurately as the best method (DI), but it is much more time efficient.

4 Conclusion

In this paper, we introduced an efficient algorithm called East, to accurately identify topological associating domains in chromatin from interaction matrices obtained from high-throughput chromosome conformation capture (Hi-C). East can be downloaded from `https://github.com/ucrbioinfo/EAST`.

We performed a comparative evaluation of East on Hi-C data for human stem cells, mouse stem cells and mouse cortex cells. We showed that our algorithm extracts TADs as accurately as the state-of-the art. TADs identified by East show substantial enrichment of various epigenetic modification factors at their boundaries, confirming similar findings in previous studies. By comparing the running time of East with the other published methods, we showed that our method is very time efficient. For a given Hi-C dataset, the only parameter in East that might need to be tuned by the user is the normalization factor for which we have given some guidance in Subsection 3.1.

The framework we presented here for TAD identification is based on fast 2D-convolution of Haar-like features. We believe that this framework could be adapted to other chromatin feature detection problems such as chromatin loops [17]. We also plan to extend our work to efficiently identify chromatin features at arbitrary scales.

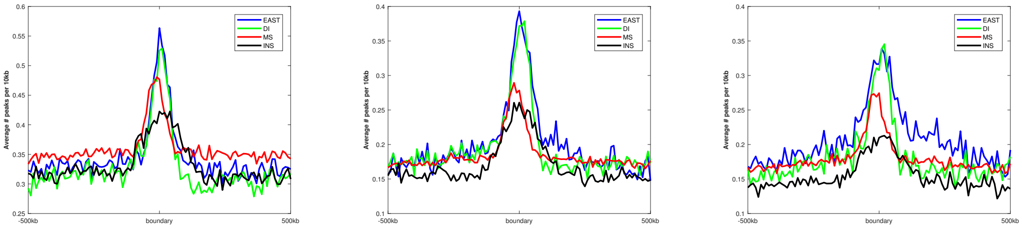


Figure 4 CTCF enrichment in human embryonic stem cells, (left) mouse embryonic stem cells (center) and mouse cortex cells (right).

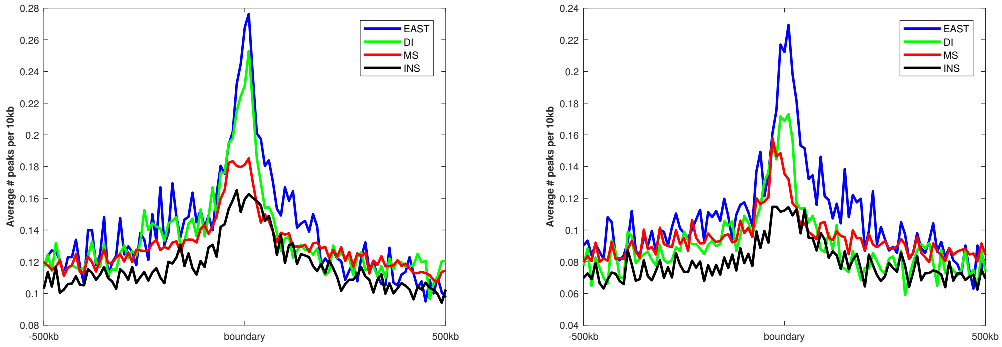


Figure 5 H3K4me3 enrichment in mouse embryonic stem cells (left) and mouse cortex cells (right).

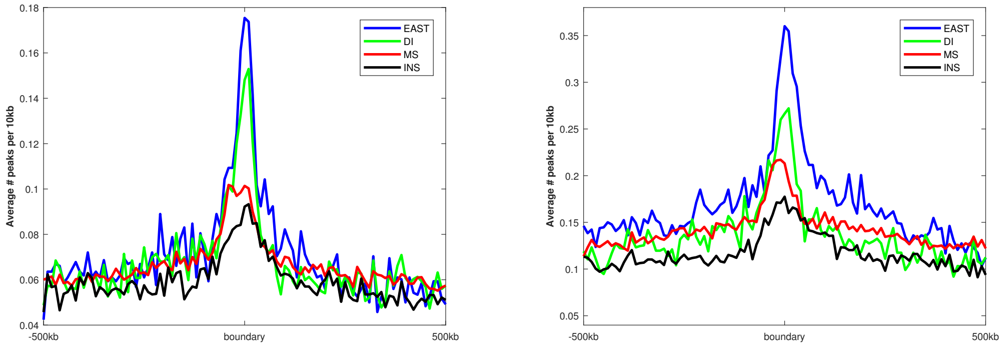


Figure 6 polII enrichment in mouse embryonic stem cells (left) and mouse cortex cells (right).

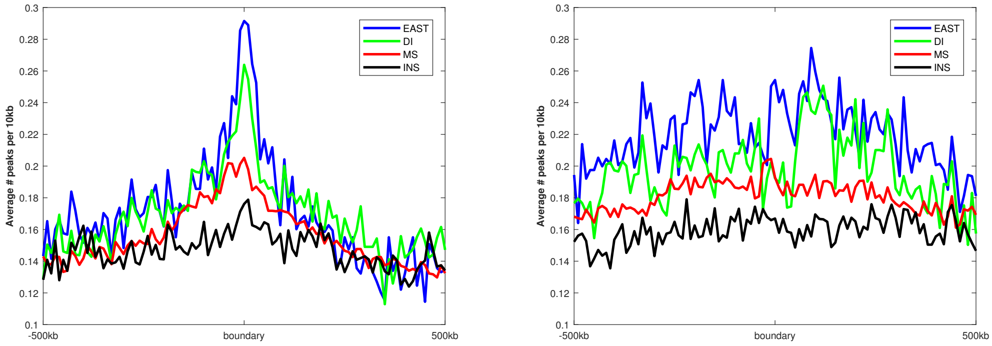


Figure 7 H3K27ac enrichment in mouse embryonic stem cells (left) and mouse cortex cells (right).

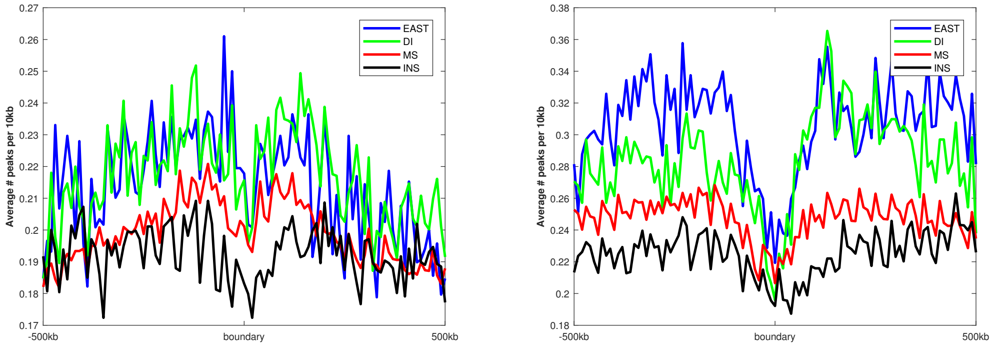


Figure 8 Enhancer enrichment in mouse embryonic stem cells (left) and mouse cortex cells (right).

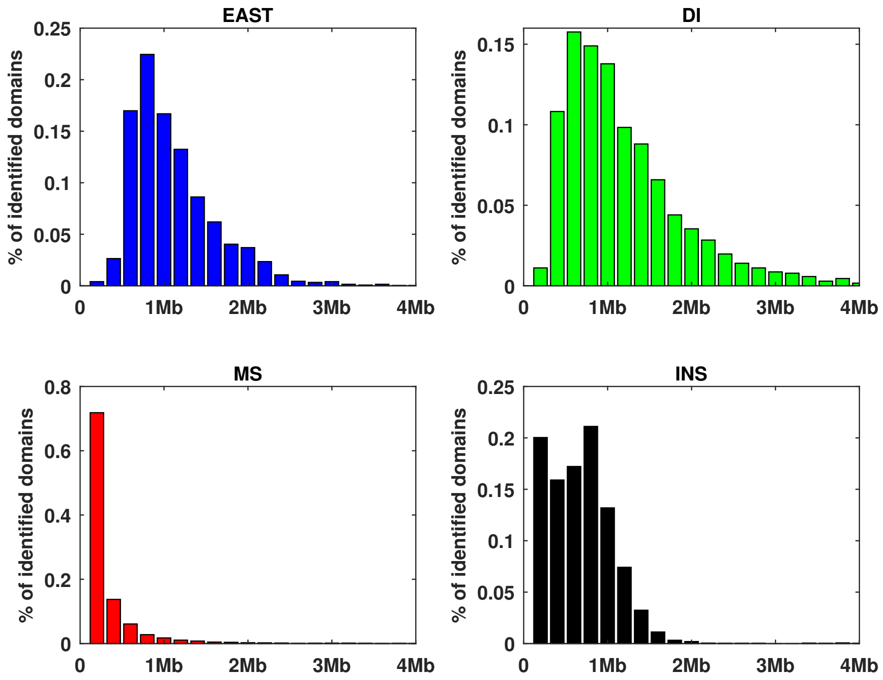


Figure 9 Comparison of the distribution of TAD size.

References

1 Giacomo Cavalli and Tom Misteli. Functional implications of genome topology. *Nat. Struct. Mol. Biol.*, 20(3):290–299, 5 March 2013.

2 Haiming Chen, Jie Chen, Lindsey A. Muir, Scott Ronquist, Walter Meixner, Mats Ljungman, Thomas Ried, Stephen Smale, and Indika Rajapakse. Functional organization of the human 4D nucleome. *Proc. Nat'l Acad. Sci. USA*, 112(26):8002–8007, 30 June 2015.

3 Jie Chen, Alfred O. Hero, 3rd, and Indika Rajapakse. Spectral identification of topological domains. *Bioinformatics*, 32(14):2151–2158, 15 July 2016.

4 Emily Crane, Qian Bian, Rachel Patton McCord, Bryan R. Lajoie, Bayly S. Wheeler, Edward J. Ralston, Satoru Uzawa, Job Dekker, and Barbara J. Meyer. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, 523(7559):240–244, 9 July 2015.

5 Franklin C. Crow. Summed-area tables for texture mapping. In *Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH'84, pages 207–212, New York, NY, USA, 1984. ACM.

6 Jesse R. Dixon, Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E. Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, Yarui Diao, Jing Liang, Huimin Zhao, Victor V. Lobanenkov, Joseph R. Ecker, James A. Thomson, and Bing Ren. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331–336, 19 February 2015.

7 Jesse R. Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 17 May 2012.

8 Darya Filippova, Rob Patro, Geet Duggal, and Carl Kingsford. Multiscale identification of topological domains in chromatin. In *Algorithms in Bioinformatics*, pages 300–312. Springer, Berlin, Heidelberg, 2 September 2013.

9 James Fraser, Carmelo Ferrai, Andrea M. Chiariello, Markus Schueler, Tiago Rito, Giovanni Laudanno, Mariano Barbieri, Benjamin L. Moore, Dorothee C. A. Kraemer, Stuart Aitken, Sheila Q. Xie, Kelly J. Morris, Masayoshi Itoh, Hideya Kawaji, Ines Jaeger, Yoshihide Hayashizaki, Piero Carninci, Alistair R. R. Forrest, FANTOM Consortium, Colin A. Semple, Josée Dostie, Ana Pombo, and Mario Nicodemi. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Syst. Biol.*, 11(12):852, 23 December 2015.

10 David U. Gorkin, Danny Leung, and Bing Ren. The 3D genome in transcriptional regulation and pluripotency. *Cell Stem Cell*, 14(6):762–775, 5 June 2014.

11 Daniel Jost, Cédric Vaillant, and Peter Meister. Coupling 1D modifications and 3D nuclear organization: data, models and function. *Curr. Opin. Cell Biol.*, 44:20–27, 2017.

12 Galih Kunarso, Na-Yu Chia, Justin Jeyakani, Catalina Hwang, Xinyi Lu, Yun-Shen Chan, Huck-Hui Ng, and Guillaume Bourque. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.*, 42(7):631–634, 6 June 2010.

13 Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S. Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 9 October 2009.

14 Darío G Lupiáñez, Malte Spielmann, and Stefan Mundlos. Breaking TADs: How alterations of chromatin domains result in disease. *Trends Genet.*, 32(4):225–237, 1 April 2016.

15 Yiqin Ma, Kiriaki Kanakousaki, and Laura Buttitta. How the cell cycle impacts chromatin architecture and influences cell fate. *Front. Genet.*, 6:19, 3 February 2015.

16 T. Pederson. Chromatin structure and the cell cycle. *Proc. Nat'l Acad. Sci. USA*, 69(8):2224–2228, August 1972.

17 Suhas S. P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and Erez Lieberman Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 18 December 2014.

18 Tom Sexton, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli. Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, 148(3):458–472, 3 February 2012.

19 Yin Shen, Feng Yue, David F. McCleary, Zhen Ye, Lee Edsall, Samantha Kuan, Ulrich Wagner, Jesse Dixon, Leonard Lee, Victor V. Lobanenkov, and Bing Ren. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–120, 2 August 2012.

20 Hanjun Shin, Yi Shi, Chao Dai, Harianto Tjong, Ke Gong, Frank Alber, and Xianghong Jasmine Zhou. TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.*, 44(7):e70, 20 April 2016.

21 P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–511–I–518 vol.1, 2001.