MrTADFinder: A network-based approach to identify topologically associating domains in multiple resolutions

Koon-Kiu Yan[1,2*] and Mark Gerstein[1,2,3*]

[1]Program of Computational Biology and Bioinformatics, [2]Department of Molecular Biophysics and Biochemistry, [3]Department of Computer Science, Yale University.

*To whom correspondence should be addressed.

Abstract

Genome-wide proximity ligation based assays such as Hi-C have revealed that eukaryotic genomes are organized into structural units called topological associating domains (TADs). From visual examination of the so-called chromosomal contact map, however, it is clear that the organization of the domains is not clear-cut. TADs appear to be overlapping and in many cases nested organization can also be found. Therefore it is important to develop new computational framework to understand the rich structures stored in a contact map. Here, based on the concept of modularity, we formulate TADs identification as an optimization problem and propose a network-based algorithm to identify TADs from intra-chromosomal contact maps. We introduce a matrix iteration procedure to derive a null model that preserves the coverage of each genomic bin as well as the distance dependence of contact frequencies for any observed contact map. In addition, by introducing a tunable parameter, our method, MrTADFinder, is able to identify TADs in different resolutions. In a low resolution, larger TADs are found whereas in a high resolution, smaller TADs are identified as the nucleome is viewed on a finer scale. We apply MrTADFinder to identify TADs in various Hi-C datasets. The identified TADs are confirmed by several downstream analyses, including the enrichment of HOT regions near TAD boundaries, and a distinctive pattern exhibited by the mutational load of cancer samples across boundaries. Overall, by facilitating the application of a variety of graph-theoretical tools, network-based approaches like MrTADFinder will be powerful for understanding the spatial organization of genome.

1. Introduction

The packing of a linear eukaryotic genome within a cell nucleus is tight and highly organized. Understanding the role of 3D genome in gene regulation is a major area of research [1][2][3][4]. Recently, genome-wide proximity ligation based assays such as Hi-C open a window to look at the intricate structure, and have revealed various structural features in terms of how genome is organized [5][6][7]. Perhaps, one of the most important discovery is the domain of self-interacting chromatin called topologically associating domain (TAD) [8][9]. Inside a TAD, genomic loci interact often but interactions between different TADs are less frequent. TAD emerges as a fundamental structural unit of chromatin organization; it plays an important role in mediating enhancer-promoter contacts and thus gene expression, and breaking or disruption of TADs can lead to diseases like cancers [10][11][12]. Therefore a deeper understanding of TADs from Hi-C data presents an important computational problem.

Results of a typical Hi-C experiment are usually summarized by a so-called chromosomal contact map [5]. By binning the genome into equally sized bins, the contact map is essentially a matrix whose element (i, j) reflects the population-averaged co-location frequencies of genomic loci originated from bins i and j. In this representation, TADs are essentially displayed as blocks along the diagonal of a contact map [8][9]. Despite TAD is a rather eye-catching feature in a contact map, computational identification is still tricky because of experimental factors such as noise and inadequate coverage. Moreover, it is clear from visual examination of the contact map that TADs appear to be overlapping and there are rich sub-structures within TADs.

Mathematically speaking, it is very natural to transform a contact matrix to a weighted network in which nodes are the genomic loci (or bins) whereas the interaction between two loci is quantified by a weighted edge. In network science, a widely studied problem is the identification of network module, also known as community detection problem [13]. A module refers to a set of nodes that are densely connected. In its simplest form, community detection problems concerns with whether nodes of a given network can be divided into non-overlapping groups such that connections within groups are relatively dense while those between groups are sparse. Therefore, by viewing the chromatin interactions as a network, the highly spatially localized TADs immediately resemble densely connecting modules.

Motivated by the resemblance, we formulate the identification of TADs as a global optimization problem based on the observational contact map and a background model. As a network-based approach, our method goes beyond a direct adaptation of standard community detection algorithm. We introduce a novel background model which takes into account the effect of genomic distance that are specific in the context of genome organization. The objective function is optimized using a heuristic and therefore efficient even the size of the input contact map is high. Furthermore, by introducing a tuning parameter, our network approach is able to identify TADs in different resolutions. In a low resolution, larger TADs are found whereas in a high resolution, smaller TADs are identified as the nucleome is viewed on a finer scale. We therefore name our method MrTADFinder for the acronym Mr stands for multiple resolutions.

In this paper, we present the methodology of MrTADFinder. We apply the method to various Hi-C datasets and validate the biological significance of the results by a few downstream analyses. Several methods have been developed for identifying TADs. The first method used by Dixon et al. is based on the so-called directionality index, a 1D statistics measuring whether the contacts have an upstream or downstream bias [8]. Later algorithms exploit the block diagonal nature of TADs in a contact map [14][15], however, the background contact frequencies are not explicitly

modeled. Using a novel matrix iteration procedure, MrTADFinder proposes a null model that preserves various properties of the empirical contact map. More recent efforts aim to investigate the hierarchical organization of TADs [16][17][18]. On the other hand, our method does not impose a hierarchical structure. A probabilistic framework and the tuning of a so-called resolution parameter are used to explore the rich structures stored in contact maps.

2. Methods

2.1. Identification of TADs as an optimization problem

The identification of modules in a network is generally formulated as a global optimization problem on the so-called modularity function over possible divisions of the network. For a given division of a network A, i.e. a mapping between the set of nodes to a set of defined modules, the modularity for a given division of a network is defined to be the fraction of edges within modules minus the expected fraction of such edges in a randomized null model of the network. Mathematically, the modularity is equal to

$$\frac{1}{2m}\sum_{i,j}\left(A_{ij}-\frac{k_i k_j}{2m}\right)\delta_{\sigma_i\sigma_j}. \tag{1}$$

Here, the summation sum over all possible pairs of nodes, the value of the Kronecker data $\delta_{\sigma_i\sigma_j}$ equals one if nodes i and j have the same label σ and zero otherwise, meaning only pairs of nodes within the same module are summed. In particular, m is the number of edges in the network whereas the expression $k_i k_j/2m$ represents the expected number of edges between i and j in a so-called configuration model. The configuration model is a randomized null model in which the degrees of nodes k_i are fixed to match those of the observed network but edges are in other respects placed at random. High values of the modularity correspond to good partitions of a network into modules and similarly low values to bad partitions. Optimizing the modularity function leads us to the best partition over all possible partitions.

Given a Hi-C contact map W, we define a similar optimization function Q as

$$\frac{1}{2N}\sum_{i,j}\left(W_{ij}-\gamma E_{ij}\right)\delta_{\sigma_i\sigma_j}. \tag{2}$$

Here, i, j refer to equally binned genomic loci. N is the total number of pair-end reads. γ is called the resolution parameter that could be used to tune the size of resultant TADs. Very much similar to the network setting, the identification of TADs aims to partition the loci into domains such that Q is optimized. Nevertheless, it is important to emphasize that, unlike nodes in a graph, the bins in a chromosome forms a linear structure. Therefore while nodes in a module could be separated in an arbitrary fashion, genomic loci in a TAD have to form a continuous chain. Moreover, the expected number of contacts between locus i and locus j depends on their genomic distance. Two loci that are close together in a 1-dimensional sense are expected to have a higher contact frequency as compared to two loci that are far apart. As a result, the null model E_{ij} in (2) has to be modified.

2.2. Expected null model of intra-chromosomal contact maps

Given an intra-chromosomal contact map W, the expected null model E is defined as

$$E_{ij} = c_i^* c_j^* f(|i-j|). \tag{3}$$

Here, f is the average number of contacts as a function of distance $d = |i-j|$. By considering all possible pairs of bins in W in terms of their distance apart and the contact frequency, we estimate f by a local smoothing with a window size equal to 1% of the data. For intermediate values of d, f follows pretty well with a power-law function d^{-1} (see Figure 1), which is a well-known observation first reported in [5].

As a null model, the resultant E matrix satisfies a set of constraints, namely

$$\begin{aligned} \sum_j E_{ij} &= \sum_j W_{ij} = c_i \quad \forall i, \\ \sum_{ij} E_{ij} &= \sum_{ij} W_{ij} = 2N. \end{aligned} \tag{3}$$

The first equation means that the coverage c_i, i.e. the total number of reads mapped to bin i, in the null model is the same as the coverage defined in the observed map. The second equation is a direct consequence of the first equation, where N is the number of reads mapped in the chromosome. As f has been estimated from the observed W, we then employ an iterative matrix procedure to numerically solve all the unknowns c_i^*. Mathematically, c_i^* can be regarded as an effective coverage because the correlation between c_i^* and the coverage c_i is extremely high (r=0.99). In comparison with (1), c_i^* is conceptually analogous to the degree k_i. The iterative procedure can be regarded as a generalization of a class of matrix balancing methods commonly used for normalizing Hi-C matrices [19]. As shown in Figure 1, given a particular matrix W, the contact frequencies of the resultant null model E are the highest in the diagonal and decrease gradually away from the diagonal.

2.3. Heuristic procedures for optimizing Q

To optimize the objective function Q, we employ a modified version of Louvain algorithm [20], which is widely used in identifying modules in networks. In a nutshell, the algorithm consists of two passes. The algorithm starts as every bin has its own label at the beginning. In the first pass, for each bin, the label was updated by either choosing the label of one of its neighboring bins or to remain unchanged based on whether or not the value of Q will be increased. When no more update is possible, the second pass is performed such that the adjacent bins with the same label were merged to form a new contact matrix. The two passes are repeated iteratively until there is no increase of modularity is possible.

The output of the modified Louvain algorithm is essentially a particular partition of the entire chromosome. As the result of the Louvain algorithm in general depends on the order of updates, multiple runs are performed to probe the fuzziness of the assignment. As the chromosome is binned into n equally sized bins, we examine, say after 10 trials, how likely the separator between bin i and bin $i+1$ is indeed a domain boundary, i.e. bin i and bin $i+1$ actually are called to belong to two different domains by the modified Louvain algorithm. Therefore, for each of the n+1 separators, we define a boundary score as the fraction of trials the location is called as a boundary. To define a set of consensus boundaries, we choose a cut-off of 0.9. In other words, the separation between two adjacent bins is defined as a confident boundary only if the two bins are

called to belong to two different domains in at least 9 out of 10 trials. The final output of MrTADFinder is a set of consensus TADs defined as regions between the consensus domains (see Figure 1 for an overall schematic).

The boundary score assigned to each bin separator is not merely an immediate, but serves as a proxy of the degree of insulation. A location with a high boundary score is more effective in forbidding the contacts between its left and right regions.

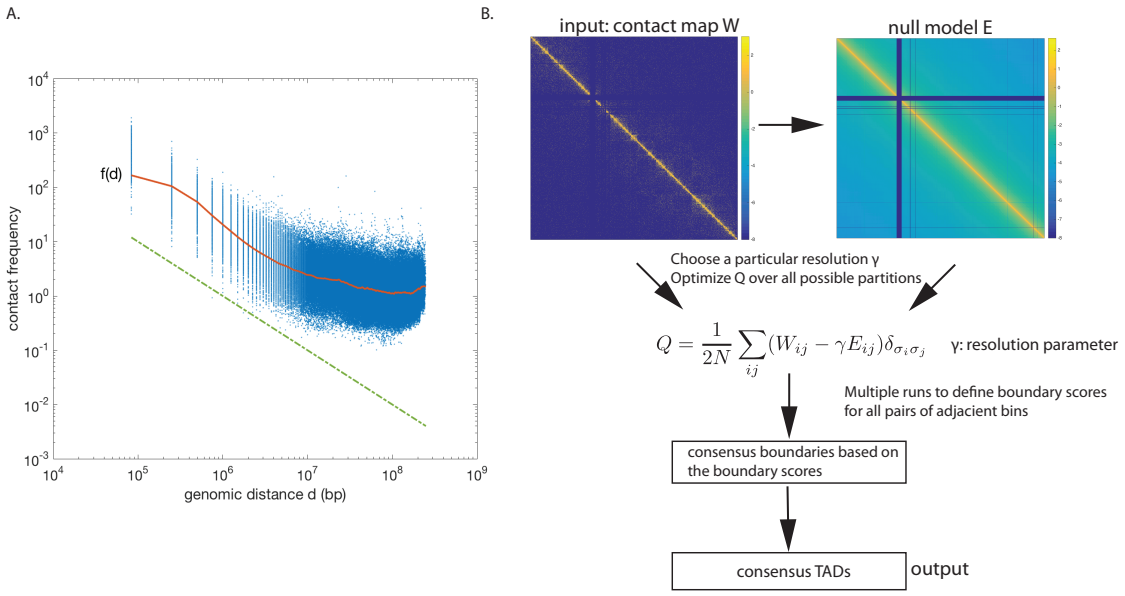

Figure 1: (A) Dependence between contact frequency and genomic distance. Analysis was performed using the contact map of the chromosome 1 of MCF7, binned in 250kb sized bins. The red line $f(d)$ is the average contact as a function of distance d obtained by smoothing all contacts. The green line shows a power-law function d^{-1}. (B) Overview of MrTADFinder.

2.4. Alternate approach using dynamic programming

Despite the similarity between equations (1) and (2), domains are continuous segments along the chromosome which is different from network modules that could be an arbitrary collection of nodes. In fact, the total number of possible partitions for a chromosome is much smaller than the total number of ways to divide a network into modules in community detection. As a result, while the optimization of (1) is a NP-hard problem, the optimization of (2) can be quite effectively solved using dynamic programming.

The idea is to extensively enumerate all the possible partitions of the chromosome. In a nutshell, a binned chromosome can be considered as a sequence $(1, 2, \cdots, N-1, N)$. Rather than partitioning the whole sequence at a first place, we look for the optimal partition for all the possible sub-sequences starting from sub-sequences with length 1. Let us denote the optimal value of modularity Q for a sequence $a_1 a_2 \ldots a_{l-1} a_l$ as $optQ(a_1 a_2 \ldots a_{l-1} a_l)$. The optimal value for the subsequence of length l is the maximum of the following l possibilities:

$$optO(a_1) + optO(a_2 \ldots a_{l-1} a_l) \tag{4}$$

$$optO(a_1a_2) + optO(a_3 \dots a_{l-1}a_l)$$

$$\vdots$$

$$optO(a_1a_2a_{3\dots}a_{l-1}) + optO(a_l)$$

$$\sum_{ij} Q_{ij}$$

Suppose the maximum correspond to the sum $optO(a_1a_2 \dots a_r) + optO(a_{r+1} \dots a_{l-1}a_l)$, where $1 \leq r < l$. The sum corresponds to the case that the optimal decomposition is to combine the optimal partitions of $a_1a_2 \cdots a_r$ and $a_{r+1} \dots a_{l-1}a_l$ (see Figure 2). It is not necessary that $a_1a_2 \cdots a_r$ forms a single domain. The key is that the expression $optQ(a_1a_2 \dots a_{l-1}a_l)$ can be found recursively because all possibilities depend on the optimal values of sub-sequences less than l, and the actual partition can be traced back using dynamics programming. The last summation in equation (4) sums Q over all positions from a_1 to a_l, meaning the l bins belong to the same domain. The procedure is analogous to the Nussinov algorithm in finding the optimal secondary structure of RNA [21].

The time complexity of this dynamic programming algorithm is in order of $O(n^3)$, where n is the size of the contact map. Nevertheless, given the time complexity, finding the optimal partition by binning the genome in a bin-size of 40kb is quite impractical. Therefore, though the connection between identifying TADs and problems like finding RNA secondary structure is of theoretical interest, MrTADFinder is developed based on the heuristic Louvain algorithm.

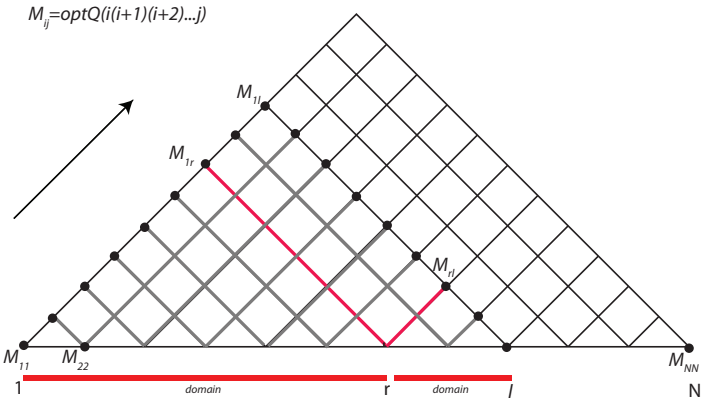

Figure 2: Identifying TADs by dynamic programming. The optimal value of Q for a chromosome segment running from i to j is stored in M_{ij}. The values of all elements in M can be enumerated using dynamic programming, starting from fragment of length 1 where $M_{ii} = Q_{ii}$. There are different ways to divide a fragment of length l (grey lines). Suppose the optimal way is marked by the red line, then $M_{1l} = M_{1r} + M_{rl}$.

2.5. Quantifying the consistency between two sets of TADs

Given two sets of TADs, say in different cell lines, or called by different algorithms, we employ the so-called normalized mutual information to quantify the consistency. Suppose X and Y are two random variables whose values x_i and y_i represent the corresponding domain labels of bin i. The normalized mutual information MI_{norm} is defined as

$$MI_{norm} = \frac{2I(X;Y)}{H(X) + H(Y)}, \quad (5)$$

where $H(X), H(Y)$ are the entropy of X and Y, and $I(X;Y)$ is the mutual information quantifying to what extent the domain labels in X predict the labels in Y. To have a fair comparison, bins that are not assigned to any TADs in both sets of partitions are not counted. If two sets of partitions are identical, the value of normalized mutual information is 1.

3. Results

3.1. Calling TADs in multiple resolutions

As a demonstration, we applied MrTADFinder to analyze Hi-C data of MCF7 cell as obtained from ref. [22]. The data have already been normalized by the so-called ICE algorithm in a whole-genome level [19][23]. Figure 3A shows a particular snapshot of the contact map (for chromosome 10) and its alignment with the identified TADs. In general, the TADs displayed agree well with the contact map. Of particular interest is the choice of the resolution parameter γ that capture the fine structures in domains organization. When γ increases, a large TAD is broken into a few small TADs. On the other hand, large TADs merge together to form even larger TADs as the value of γ is lowered. As estimated by MrTADFinder, when γ=4, there are about 80 TADs in the chromosome 10 of MCF7 with a median size of roughly 1.4Mb. Statistically speaking, γ is essentially quantifying to the ratio between the expected counts as compared to the observed counts. As γ increases, only elements close to the diagonal contribute positively to the modularity function. Therefore in general, the size of TADs decreases and the number of TADs increases (see Figure 3B).

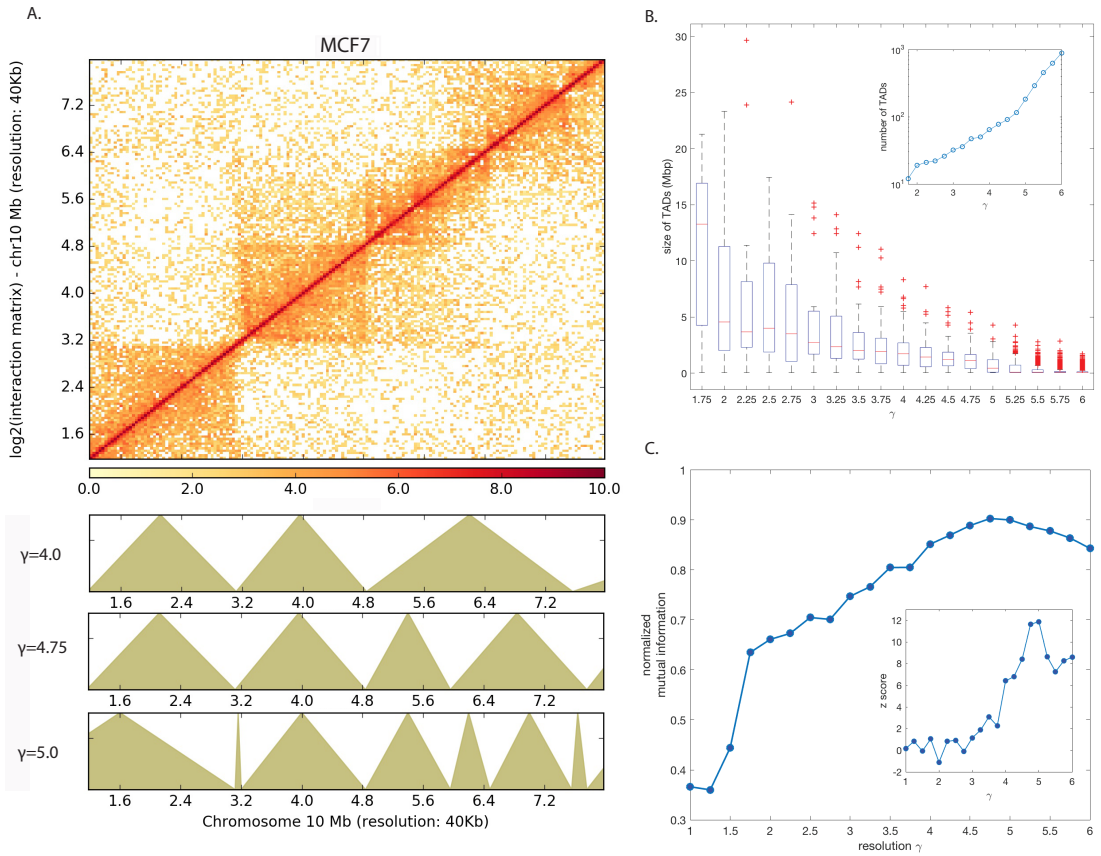

Figure 3: (A) The contact map and TADs of the chromosome in MCF7. The greenish triangles below represent TADs called by MrTADFinder in three different resolutions. The TADs called agree well

visually with the contact map. (B) The size of TADs in different resolutions. The inset shows the number of TADs in different resolutions. (C) Consistency between MrTADFinder and an existing method. The normalized mutual information between two sets of TADs peaks at a resolution $\gamma = 4.75$. The inset shows a similar result in which the empirical mutual information is normalized with respect to a null model. Panel A is generated using the tool HiCPlotter [24].

It is instructive to compare our results with TADs identified by an alternate method. Here, we report a comparison with the TADs called in ref. [22]. As quantified by the normalized mutual information, TADs identified by MrTAD Finder best match with alternate set of TADs when the resolution parameter is 4.75, with the normalized mutual information more than 90% (see Figure 3C). When the resolution is too small, there is a large discrepancy. As TADs exist in the form of continuous genomic segments, two sets of TADs tend to overlap to a certain extent and thus maintain a relatively high consistency. To provide a better estimate, we shuffled the TADs and generated a null distribution of mutual information. We then further calculated the z-score of the empirical mutual information with the null. As shown in the inset of Figure 2C, TADs called by MrTADFinder agrees the most of the existing method when $\gamma = 4.75$. Nevertheless, we want to emphasize that the introduction of the resolution parameter γ has broadened the previous work on domains identification that mostly focuses on a particular resolution instead.

3.2. Genomic features near TAD boundary

It is well known that certain chromatin features like histone marks and transcription factor binding sites are enriched near the boundary regions of TADs [8]. Instead of looking at individual factors, we further explored the location of the so-called HOT regions and XOT regions with respect to TADs. High-occupancy target (HOT) regions and extreme-occupancy target (XOT) regions are genomic regions that are bound by extensive amount of transcription factors [25]. As expected, we found a strong enrichment of HOT regions and an even stronger enrichment of XOT regions near the boundary regions (Figure 4A). The observation agrees with the idea that HOT regions are very accessible regions in open chromatin. Nevertheless, it is still widely unknown if transcription factors bind there simply because of thermodynamics, or it is driven by important biological functions.

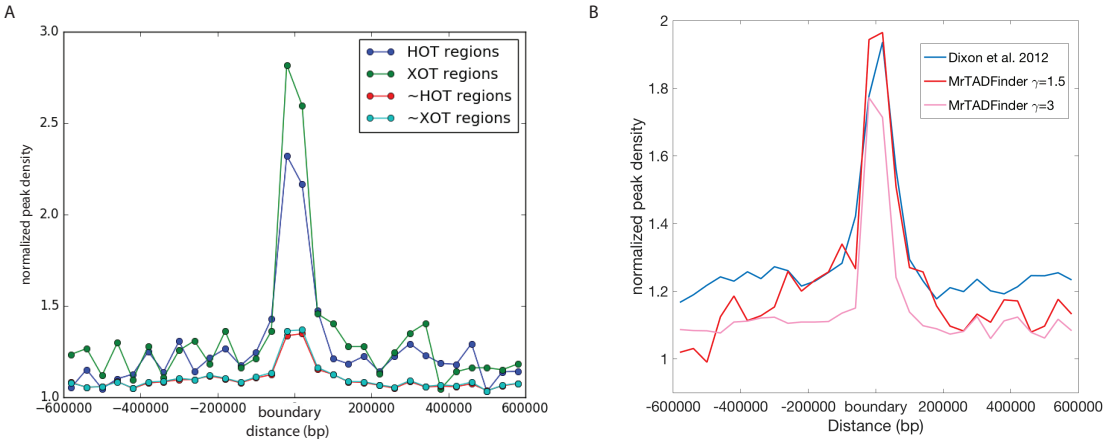

Figure 4 (A): Enrichment of HOT and XOT regions near TAD boundary in hES cell. Y-axis is normalized with respect to a null model that peaks are uniformly distributed in along the chromosome. (B): Enrichment of CTCF peaks near TAD boundary.

The enrichment of chromatin features has been used as a benchmark for various TAD calling algorithms. Dixon et al. identified TADs based on the directionality index using Hi-C data in hES cell and found an enrichment of CTCF binding sites at the boundary regions [8]. We performed the same analysis using TADs based on MrTADFinder. As shown in Figure 4B, the enrichment

of CTCF binding peaks using TADs called by two methods is similar. Nevertheless, MrTADFinder extends the observation by exploring the effects of TADs in different resolutions. In a low resolution, i.e. for larger TADs, the enrichment signal is stronger, and the signal tends to extend over a longer distance from the boundary. In a higher resolution, the signal is weaker and confines near the boundary. The observation suggests that the boundary regions separating two large domains tend to be bound by CTCFs more often, and points to the function of such an important architectural protein that plays an important role in mediating chromatin loops [26][27].

To provide further evidence on the validity TADs identified by MrTADFinder, we investigated the occurrence of somatic mutations at the TAD boundary regions. More specifically, using the ICGC data portal, we downloaded the set of somatic mutations called from the whole-genome sequencing of breast cancer samples from 676 donors, and mapped the mutations to the TAD boundaries identified by MrTADFinder. For each boundary region, we examined how the mutational load changes with respect to the distance from the boundary. As shown in Figure 5A, the about 100 boundary regions identified in chromosome 10 can be clustered in 3 groups based on their positional distribution of somatic mutations. Two of them exhibit a step-function behavior (blue and red in Figure 5B) in which the transition is essentially at the boundary. Because of the close relationship between TADs and replication-timing domains [28], the observation resonates with the high mutational load found in regions that are replicated later during DNA-replication [29]. For boundary regions in the green cluster, the mutational load exhibits no difference across the TAD boundary; the regions may be heterochromatin or other quiescent components.

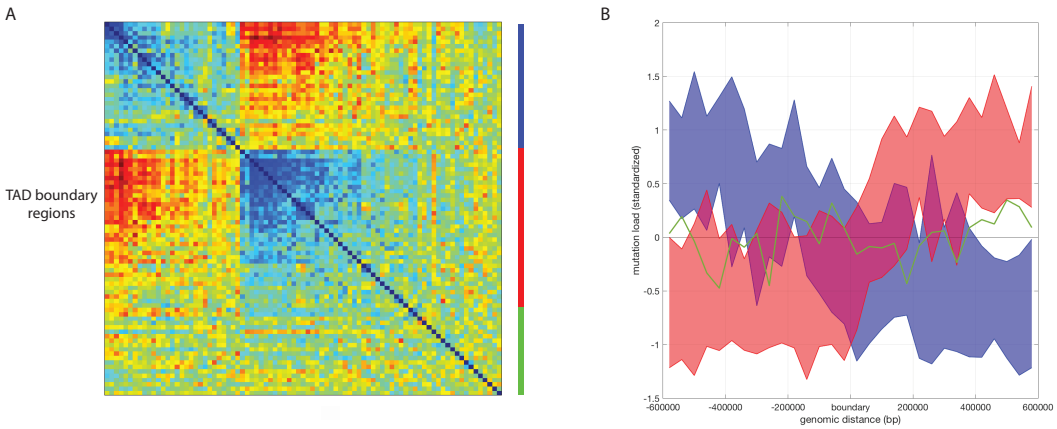

Figure 5 (A) TAD boundary regions (±600kb of boundary) of MCF7 (chromosome 10) are clustered based on the mutational load along the regions. The regions are clustered into 3 groups (blue, red, green). (B) The 3 clusters of boundary regions exhibit distinct patterns in terms of mutational load. For blue and red clusters, the area marks the first and the third quartiles. For the green cluster, only the mean values at different positions are shown for simplicity.

3.3. Robustness analysis

Because of the stochastic nature of Louvain algorithm, we explored the robustness of MrTADFinder. By calling TADs based on 10 runs of Louvain algorithm, we found the results of two independent callings highly robust. In fact, the normalized mutual information is higher than 0.99 (see Figure 6).

As MrTADFinder employs a heuristic to arrive at sub-optimal partitions of equation (2), it is important to explore the difference between the optimal partition and the sub-optimal partitions. Using a contact map of hES cell (chromosome 1) binned with a bin-size of 500kb, we found the sub-optimal partitions based on our modified Louvain algorithm are very close to the optimal partition based on dynamic programming. The normalized mutual information between optimal and sub-optimal values is 0.977±0.007. It is worthwhile to point out that the final partition defined by the confident boundaries, results at a lower modularity function as compared to any single sub-optimal solution. It is because the filtering of statistically significant boundaries removes a certain number of boundaries. Nevertheless, the importance of getting a set of consensus boundaries makes the choice justifiable.

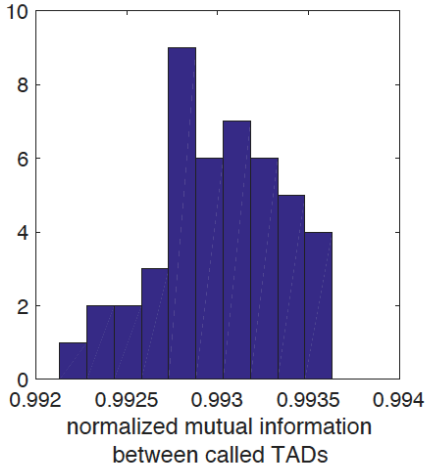

Figure 6 Robustness of MrTADFinder. Histogram for pairs of independently called TADs.

3.4. Implementation and benchmark

MrTADFinder is implemented in Julia. The source code can be downloaded in https://github.com/gersteinlab/MrTADFinder. Julia users can import MrTADFinder as a library. It can be also be run in command line if Julia and the required packages are installed (see the Github page for details).

We benchmarked the performance of MrTADFinder using Hi-C data reported by the Aiden lab [30]. The experimental data we tested on were from GM12878 cell. The reads were binned into contact maps in ref. [30] with size ranging from 500kb to 10kb. We used MrTADFinder to identify TADs in chromosome 10. Table 1 shows the time required for the calculation, including 10 multiple runs of Louvain algorithm.

Bin size	Size of contact map	Elapsed time (s)
500kb	272	10
250kb	543	60
100kb	1356	190
50kb	2712	410
25kb	5422	1080
10kb	13554	11260

Table 1 Performance of MrTADFinder as applied for the contact maps of chromosome 10.

4. Discussion

In this paper, we have introduced an intuitive algorithm based on network theory to identify TADs from Hi-C data, and performed several analysis to confirm the biological significance of the TADs identified. In particular, by introducing a single continuous parameter γ, we are able to further examine domain organization in multiple resolutions. It is important to emphasize that the idea of resolution we introduced in MrTADFinder is different from some other usages of the same term in Hi-C analysis. From an experimental standpoint, the resolution of a Hi-C experiment refers to the average fragment size as digested by restriction enzymes (~4kb to ~1kb)

[5][30] or more recently by micrococcal nuclease (~150bp) [31]. In terms of constructing contact maps, the term resolution has been used to refer to the bin size, where the proper choice usually depends on the number of reads in the stage of data processing. Both usages are essentially technical. What we mean by resolution, however, refers to the multiple length scale built inside the organization of genome. It is well known that there are structures in different length scales such as compartment, domains and sub-domains [32], and chromatin features like histone marks exhibit multiple length scales [33]. The concept of resolution introduced here points to the integration of these structures, and enables one to explore the rich structures hidden in contact maps.

A novel contribution of this work is the derivation of an expected model for any intra-chromosomal contact map using matrix iteration. The null model preserves the coverage of each genomic bin as well as the distance dependence of contact frequencies in the observed map. Apart from the identification of TADs, the expected model can be used for applications like finding compartments and identifying enhancer-target linkages. Recent studies employ various clustering methods to identify inter- chromosomal clusters using Hi-C data [35][36], similar expected models can also be derived to better separate signal and noise.

MrTADFinder is a motivated by the community detection problem in network studies. Although a network perspective of chromosomal interactions has previously been proposed [37][38], a lot of widely studied concepts in networks have rarely been explored in the context of chromosomal organization. A network representation is arguably more flexible, for instance, transcription factors binding and histone modifications can be easily incorporated into the network, forming a decorated network. Moreover, one could extend the framework by concatenating multiple Hi-C contact maps to form a multi-layer network. The same idea has been used for cross-species analysis [39]. By facilitating the application of a variety of graph-theoretical tools, we believe that network algorithms will be useful for future studies on the spatial organization of genome.

References

[1] J. Dekker, M. A. Marti-Renom, and L. A. Mirny, "Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data," *Nat. Rev. Genet.*, vol. 14, no. 6, pp. 390–403, Jun. 2013.

[2] V. I. Risca and W. J. Greenleaf, "Unraveling the 3D genome: genomics tools for multiscale exploration," *Trends Genet.*, vol. 31, no. 7, pp. 357–372, Jul. 2015.

[3] M. J. Rowley and V. G. Corces, "The three-dimensional genome: principles and roles of long-distance interactions," *Curr. Opin. Cell Biol.*, vol. 40, pp. 8–14, Jun. 2016.

[4] B. Bonev and G. Cavalli, "Organization and function of the 3D genome," *Nat. Rev. Genet.*, vol. 17, no. 11, pp. 661–678, Nov. 2016.

[5] E. Lieberman-Aiden *et al.*, "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome," *Science*, vol. 326, no. 5950, pp. 289–293, Oct. 2009.

[6] R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber, and L. Chen, "Genome architectures revealed by tethered chromosome conformation capture and population-based modeling," *Nat. Biotechnol.*, vol. 30, no. 1, pp. 90–98, Dec. 2011.

[7] M. J. Fullwood and Y. Ruan, "ChIP-based methods for the identification of long-range chromatin interactions," *J. Cell. Biochem.*, vol. 107, no. 1, pp. 30–39, May 2009.
[8] J. R. Dixon *et al.*, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, no. 7398, pp. 376–380, May 2012.
[9] T. Sexton *et al.*, "Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome," *Cell*, vol. 148, no. 3, pp. 458–472, Feb. 2012.
[10] J. Dekker and E. Heard, "Structural and functional diversity of Topologically Associating Domains," *FEBS Lett.*, vol. 589, no. 20, Part A, pp. 2877–2884, Oct. 2015.
[11] A.-L. Valton and J. Dekker, "TAD disruption as oncogenic driver," *Curr. Opin. Genet. Dev.*, vol. 36, pp. 34–40, Feb. 2016.
[12] D. G. Lupiáñez, M. Spielmann, and S. Mundlos, "Breaking TADs: How Alterations of Chromatin Domains Result in Disease," *Trends Genet.*, vol. 32, no. 4, pp. 225–237, Apr. 2016.
[13] M. E. J. Newman, "Modularity and Community Structure in Networks," *Proc. Natl. Acad. Sci.*, vol. 103, no. 23, pp. 8577–8582, Jun. 2006.
[14] D. Filippova, R. Patro, G. Duggal, and C. Kingsford, "Identification of alternative topological domains in chromatin," *Algorithms Mol. Biol.*, vol. 9, no. 1, p. 14, May 2014.
[15] C. Lévy-Leduc, M. Delattre, T. Mary-Huard, and S. Robin, "Two-dimensional segmentation for analyzing Hi-C data," *Bioinformatics*, vol. 30, no. 17, pp. i386–i392, Sep. 2014.
[16] C. Weinreb and B. J. Raphael, "Identification of hierarchical chromatin domains," *Bioinformatics*, p. btv485, Aug. 2015.
[17] L. I. Malik and R. Patro, "Rich chromatin structure prediction from Hi-C data," *bioRxiv*, p. 32953, Nov. 2015.
[18] J. Fraser *et al.*, "Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation," *Mol. Syst. Biol.*, vol. 11, no. 12, pp. 852–852, Dec. 2015.
[19] M. Imakaev *et al.*, "Iterative correction of Hi-C data reveals hallmarks of chromosome organization," *Nat. Methods*, vol. 9, no. 10, pp. 999–1003, Oct. 2012.
[20] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech. Theory Exp.*, vol. 2008, no. 10, p. P10008, Oct. 2008.
[21] R. Nussinov, G. Pieczenik, J. Griggs, and D. Kleitman, "Algorithms for Loop Matchings," *SIAM J. Appl. Math.*, vol. 35, no. 1, pp. 68–82, Jul. 1978.
[22] A. R. Barutcu *et al.*, "Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells," *Genome Biol.*, vol. 16, p. 214, 2015.
[23] N. Servant *et al.*, "HiC-Pro: an optimized and flexible pipeline for Hi-C data processing," *Genome Biol.*, vol. 16, no. 1, p. 259, Dec. 2015.

[24] K. C. Akdemir and L. Chin, "HiCPlotter integrates genomic data with interaction matrices," *Genome Biol.*, vol. 16, no. 1, p. 198, Sep. 2015.
[25] A. P. Boyle *et al.*, "Comparative analysis of regulatory information and circuits across distant species," *Nature*, vol. 512, no. 7515, pp. 453–456, Aug. 2014.
[26] C.-T. Ong and V. G. Corces, "CTCF: an architectural protein bridging genome topology and function," *Nat. Rev. Genet.*, vol. 15, no. 4, pp. 234–246, Apr. 2014.
[27] Z. Tang *et al.*, "CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription," *Cell*.
[28] B. D. Pope *et al.*, "Topologically associating domains are stable units of replication-timing regulation," *Nature*, vol. 515, no. 7527, pp. 402–405, Nov. 2014.
[29] M. S. Lawrence *et al.*, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, no. 7457, pp. 214–218, Jul. 2013.
[30] S. S. P. Rao *et al.*, "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping," *Cell*, vol. 159, no. 7, pp. 1665–1680, Dec. 2014.
[31] T.-H. S. Hsieh, A. Weiner, B. Lajoie, J. Dekker, N. Friedman, and O. J. Rando, "Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C," *Cell*, vol. 162, no. 1, pp. 108–119, Jul. 2015.
[32] B. A. Bouwman and W. de Laat, "Getting the genome in shape: the formation of loops, domains and compartments," *Genome Biol.*, vol. 16, no. 1, p. 154, Aug. 2015.
[33] A. Harmanci, J. Rozowsky, and M. Gerstein, "MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework," *Genome Biol.*, vol. 15, no. 10, p. 474, Oct. 2014.
[34] F. Ay and W. S. Noble, "Analysis methods for studying the 3D architecture of the genome," *Genome Biol.*, vol. 16, no. 1, p. 183, Sep. 2015.
[35] A. Fotuhi Siahpirani, F. Ay, and S. Roy, "A multi-task graph-clustering approach for chromosome conformation capture data sets identifies conserved modules of chromosomal interactions," *Genome Biol.*, vol. 17, p. 114, 2016.
[36] C. Dai *et al.*, "Mining 3D genome structure populations identifies major factors governing the stability of regulatory communities," *Nat. Commun.*, vol. 7, p. 11549, May 2016.
[37] I. Rajapakse, D. Scalzo, S. J. Tapscott, S. T. Kosak, and M. Groudine, "Networking the nucleus," *Mol. Syst. Biol.*, vol. 6, no. 1, Jan. 2010.
[38] K. Kruse, S. Sewitz, and M. M. Babu, "A complex network framework for unbiased statistical analyses of DNA–DNA contact maps," *Nucleic Acids Res.*, vol. 41, no. 2, pp. 701–710, Jan. 2013.
[39] K.-K. Yan, D. Wang, J. Rozowsky, H. Zheng, C. Cheng, and M. Gerstein, "OrthoClust: an orthology-based network framework for clustering data across multiple species," *Genome Biol.*, vol. 15, no. 8, p. R100, Aug. 2014.