Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Applications Note

OXFORD

Genome analysis

optimalTAD: annotation of topologically associating domains based on chromatin marks enrichment

Dmitrii N. Smirnov [1,2,3], Debra Toiber [2,3], Mikhail S. Gelfand [1,4]* and Ekaterina E. Khrameeva [1,*]

[1] Center for Molecular and Cellular Biology, Skolkovo Institute of Science and Technology, Moscow 121205, Russia
[2] Department of Life Sciences, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel
[3] The Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, Beer Sheva 8410501, Israel
[4] Institute for Information Transmission Problems (Kharkevich Institute) RAS, Moscow 127051, Russia

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: In many eukaryotes, chromosomes are organized as strings of spatially segregated Topologically Associating Domains (TADs), characterized by a substantially increased frequency of interactions within them. Boundaries of TADs are highly enriched in histone acetylation chromatin marks and occupied binding sites of architectural proteins, highlighting the functional role of TADs in the regulation of gene expression. While many computational approaches for the identification of TADs have been developed, it remains challenging because of their nested structure, resulting in weakly overlapping sets of TADs at different scales.

Results: Here, we propose a novel algorithm `optimalTAD` for identifying the optimal set of TADs based on epigenetic marks enrichment. Assuming that the most dramatic enrichment corresponds to the best annotation of TAD boundaries, our algorithm optimizes TAD calling parameters by maximizing the difference in chromatin mark levels between TADs and their boundaries. Using this algorithm, we annotated TADs in several publicly available Hi-C datasets and identified a set of epigenetic marks that are best suited for TAD prediction.

Availability: `optimalTAD` is available at https://github.com/cosmoskaluga/optimalTAD

Contact: E.Khrameeva@skoltech.ru and M.Gelfand@skoltech.ru

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Chromatin conformation capture methods (3C and Hi-C) have uncovered a fundamental level of chromosome organization: Topologically Associating Domains (TADs) (Dixon *et al.*, 2012, Nora *et al.*, 2012). The average TAD size differs across species: 1000 kb for the human, 880 kb for the mouse, and 140 kb for the fruit fly genome. Besides different sizes, TAD structural properties also vary between species. While mammalian chromatin is partitioned into contiguous topological domains, TADs in *Drosophila* are separated by stretches of decompressed chromatin - interTADs. According to recent Hi-C experiments with sub-kb resolution, inter-TAD regions are composed of small weak TADs (median size 9 kb) enriched for active chromatin (Wang *et al.*, 2018). These small active TADs are indistinguishable at lower resolution levels and therefore cannot be resolved in most Hi-C studies traditionally separating *Drosophila* chromatin into TADs and inter-TADs.

TADs represent functional domains of gene expression regulation, defining promoter-enhancer interactions through chromatin loops (Doyle *et al.*, 2014). TAD boundaries are enriched in active transcription and housekeeping genes (Ulianov *et al.*, 2016), which creates a pronounced gradient of histone acetylation and other active chromatin marks at TAD boundaries. In mammals, TAD boundaries are occupied by

© The Author 2023. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

insulator proteins (Cohesin, CTCF), maintaining spatial segregation of TADs (Hansen *et al.*, 2017). In *Drosophila*, other architectural proteins (BEAF32, Chromator) are enriched at TAD boundaries (Szabo *et al.*, 2019). The reported gradient of chromatin marks between TADs and their boundaries offers valuable information for precise annotation of TADs, yet none of the existing TAD-calling techniques takes it into account.

Several TAD-calling algorithms have been introduced, including `Directionality Index` (Dixon *et al.*, 2012), `Armatus` (Filippova *et al.*, 2014), `Arrowhead` (Rao *et al.*, 2014), `CITD` (Chen *et al.*, 2016), `HiCseg` (Levy-Leduc *et al.*, 2014), `TopDom` (Shin *et al.*, 2016), `calTADs` (Wang, 2016), `TADtree` (Weinreb and Raphael, 2016). Among these methods, the `Armatus` algorithm deserves special recognition because it provides a multiscale approach for TAD calling and can be successfully applied to both mammalian and *Drosophila* genomes (Sauerwald *et al.*, 2020, Ulianov *et al.*, 2016). The multiscale nature of `Armatus` allows it to produce different TAD sets depending on the value of the resolution parameter γ. TADs in the resulting sets vary substantially in size and number, depending on γ. Such variability might reflect the biological properties of predicted domains but impedes their interpretation because it is not clear which TAD set better matches chromatin functional properties and is therefore suitable for further analysis.

Here, we present a user-friendly and accurate tool called `optimalTAD` to determine the optimal set of TADs based on a combination of Hi-C and epigenetic data for robust annotation of TADs as both structural and functional chromatin units. For a given Hi-C matrix, `optimalTAD` identifies a complete set of multiscale domains via the `Armatus` method and then evaluates each TAD set using ChIP-seq profiles of histone marks or architectural proteins to obtain the optimal TAD segmentation.

2 Methods

2.1 The `optimalTAD` algorithm

- **Input data.** The algorithm requires data collected from both Hi-C and ChIP-seq experiments (Fig. 1a, step 1). Hi-C contact maps should be normalized using the iterative correction method (Imakaev *et al.*, 2012). ChIP-seq profiles should be normalized using both the total number of reads and corresponding inputs, and further $\log_2$ and Z-score transformed. `optimalTAD` binarizes ChIP-seq coverage into intervals of the same length as the resolution of the Hi-C contact matrix.
- **Empty line imputation.** A typical Hi-C matrix contains a number of rows/columns consisting entirely of missing values. Such empty lines appear in poorly mapped genome regions and regions enriched in repeated DNA sequences. Such empty lines impede TAD identification. `optimalTAD` imputes empty bins within such lines based on neighboring bins using the linear regression method.
- **TAD calling.** The next step of TAD set identification is performed using the `Armatus` tool. Importantly, the `Armatus` average TAD size depends on the value of the resolution parameter γ. In our approach, `Armatus` determines multiple TAD sets: one TAD set per each γ varying from $\gamma_{\min} > 0$ to $\gamma_{\max}$. The step size s used to increment γ within this range is specified by the user.
- **TAD numeration.** Once all TAD sets for the input Hi-C map are determined, `optimalTAD` performs inter- and intra-domain bin indexing. Each index is calculated as the distance from a bin to the nearest TAD border (Fig. 1a, step 2). Bins corresponding to domain boundaries are indexed as 0, whereas inter-TAD bin indices are multiplied by -1. Next, each bin index is matched with the corresponding value from the binarized epigenetic profile (Fig. 1a, step 3). Finally, for each TAD set, `optimalTAD` calculates medians

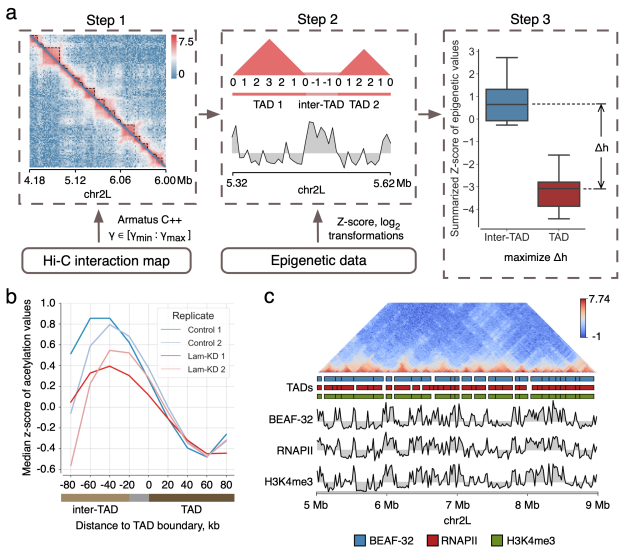

Fig. 1. Summary of the `optimalTAD` performance. (**a**) A general scheme of the algorithm. (**b**) The amplitude of histone acetylation levels in the TAD set predicted by `optimalTAD` in the Ulianov dataset for *Drosophila*. Control replicates (Control-1, Control-2) are shown by shades of blue and Lam-KD replicates are represented by shades of red. (**c**) An example of the Hi-C map with TAD boundaries predicted in the chr2L:5,000,000-9,000,000 genomic region using BEAF-32 (blue rectangles), RNAPII (red rectangles), H3K4me3 (green rectangles) profiles.

M^{TAD} and M^{interTAD} of ChIP-seq values belonging to TAD and inter-TAD regions, respectively.

- **Optimization** We optimize the TAD calling parameter γ by maximizing the difference Δh between the median ChIP-seq values in TADs and in inter-TADs:

$$\begin{aligned} \underset{\gamma}{\text{maximize}} \quad & \Delta h(\gamma) = M^{\mathrm{interTAD}}(\gamma) - M^{\mathrm{TAD}}(\gamma) \\ \text{subject to} \quad & \gamma \in [\gamma_{\min}, \gamma_{\max}] \\ & 0 \leq \gamma_{\min} < \gamma_{\max} \end{aligned} \tag{1}$$

where M is the median of ChIP-seq values and γ is the resolution parameter in `Armatus`.

2.2 Implementation

To accelerate the TAD calling step, we developed a parallel implementation of the `Armatus` algorithm that supports > 1 computational core. Parallel `Armatus` decomposes multiscale domain segmentation into N threads, where N is equal to the number of cores available on the machine. Therefore, it achieves maximal computational efficiency when using $\frac{\gamma_{\max}-\gamma_{\min}}{s}$ cores.

2.3 Algorithm validation

Three publicly available datasets (Ulianov *et al.*, 2019, Li *et al.*, 2015, Ing-Simmons *et al.*, 2021) were used for the algorithm validation. We used 20-kb Hi-C maps from the Ulianov dataset and Li dataset, and 16-kb Hi-C maps from the Ing-Simmons dataset. The Hi-C interaction frequency values were scaled from $x_{\min} = 1$ to $x_{\max} = 1024$ and then $\log_2$ transformed. Processed ChIP-seq profiles were also $\log_2$ transformed and standardized using Z-score. All `optimalTAD` validation runs were done with the step size parameter $s = 0.01$ using 4 cores on MacOS Catalina (Intel Core i5). Further details on data preprocessing are described in Supplementary Materials.

3 Results

3.1 Prediction of TADs based on histone acetylation levels

To examine the ability of our algorithm to identify TAD sets at differential conditions, we tested `optimalTAD` on Hi-C and ChIP-seq data from *Ulianov et al.*, in control S2 *Drosophila* cells and upon lamin Dm0 knock-down (Lam-KD). `optimalTAD` reported lower optimal γ values for control replicates (1.65 and 1.29) compared to Lam-KD replicates (2.15 and 1.73). Moreover, both Lam-KD replicates exhibited a specific shape of the acetylation amplitude curve, which clearly differed from the curves for both control replicates (Suppl. Fig. 1a,b). Such diverse shapes offer a new possibility for the cluster analysis of Hi-C samples because the acetylation amplitude curves reflect epigenetic properties of TADs at different hierarchical levels.

Next, we compared acetylation profiles near TAD boundaries obtained for the predicted optimal γ values in each replicate. As expected, we found a drastic difference in histone H3 acetylation levels between inter-TADs and TADs (Fig. 1b). In agreement with (Ulianov *et al.*, 2019), inter-TAD histone acetylation enrichment decreased in Lam-KD. The median size of predicted TADs varied from 80 to 120 kb (80-100 kb for Lam-KD replicates and 100-120 kb for control replicates, consistent with their optimal γ values), while the median inter-TAD size remained constant across replicates and was equal to 40 kb (Suppl. Fig. 1c). These observations agree with the *Drosophila* TAD properties reported in previous studies (Wang *et al.*, 2018; Sexton *et al.*, 2012).

3.2 Prediction of TADs based on the architectural proteins enrichment

Further, we aimed to determine *Drosophila* architectural proteins and chromatin marks most useful for predicting the optimal TAD sets. We selected ChIP-Seq profiles of architectural proteins (CTCF, CP190, RAD21, BEAF-32, TFIIIC, Z4, L(3)mbt, Chromator, CBP, CapH2, RNAPII) together with chromatin mark profiles (H3K4me3, H3K9me2) and profiles of the Polycomb-group proteins (PC-RJ, PC-VP) from the Li dataset. Consistent with previous studies on predictors of TAD boundaries in *Drosophila* (Sexton *et al.*, 2012; Wang *et al.*, 2018; Arzate-Mejía *et al.*, 2020), our analysis revealed that RNAPII, BEAF32, and H3K4me3 were highly enriched near the predicted TAD boundaries (Fig. 1c) and possessed maximal Δh (1.692, 1.532, and 1.527, respectively). On the contrary, CTCF was weakly enriched near TAD borders and thus was far less suitable for the identification of TADs, corroborating its limited role in TAD formation in flies (Ulianov *et al.*, 2016; Kaushal *et al.*, 2021).

3.3 Optimal TAD prediction in Toll$^{rm9/rm10}$ mutant embryos

To illustrate its general usability, we tested `optimalTAD` on the *Ing-Simmons et al.* dataset comprising the Toll$^{rm9/rm10}$ 16-kb Hi-C map and two H3K27ac ChIP-seq replicates. Our analysis resulted in two optimal TAD sets of the same median size of 64 kb, where 92% of borders were shared between the two domain sets (Suppl. Fig. 5a,b). Amplitude curves were also very similar (Suppl. Fig. 5c) with Pearson's R = 0.9942 and RMSD = 0.0018, highlighting the reproducibility of the `optimalTAD` segmentations between biological replicates.

4 Conclusion

In conclusion, `optimalTAD` provides an optimized identification of topological domains and can be successfully applied to Hi-C data obtained from various sources. While other TAD-calling algorithms are limited to using only structural information on contact frequencies and predict boundaries based on a predefined TAD size, `optimalTAD`, in addition, incorporates functional information about chromatin marks and architectural proteins to annotate TADs as both structural and functional units of chromatin. The extended results on algorithm validation are available in Supplementary Materials.

Funding

The research was supported by the Russian Foundation for Basic Research (project number 21-34-70051).

References

Arzate-Mejía, R. G. *et al.* (2020). In situ dissection of domain boundaries affect genome topology and gene transcription in drosophila. *Nat. Commun.*, **11**(1), 894.

Chen, Y. *et al.* (2016). *De novo* deciphering three-dimensional chromatin interaction and topological domains by wavelet transformation of epigenetic profiles. *Nucleic Acids Res*, **44**(11), e106–e106.

Dixon, J. R. *et al.* (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**(7398), 376–380.

Doyle, B. *et al.* (2014). Chromatin Loops as Allosteric Modulators of Enhancer-Promoter Interactions. *PLoS Comput Biol*, **10**(10), e1003867.

Filippova, D. *et al.* (2014). Identification of alternative topological domains in chromatin. *Algorithms Mol Biol*, **9**(1), 14.

Hansen, A. S. *et al.* (2017). CTCF and cohesin regulate chromatin loop stability with distinct dynamics.

Imakaev, M. *et al.* (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods*, **9**(10), 999–1003.

Ing-Simmons, E. *et al.* (2021). Independence of chromatin conformation and gene regulation during Drosophila dorsoventral patterning. *Nat Genet*, **53**(4), 487–499.

Kaushal, A. *et al.* (2021). CTCF loss has limited effects on global genome architecture in drosophila despite critical regulatory functions. *Nat. Commun.*, **12**(1), 1011.

Levy-Leduc, C. *et al.* (2014). Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics*, **30**(17), i386–i392.

Li, L. *et al.* (2015). Widespread Rearrangement of 3D Chromatin Organization Underlies Polycomb-Mediated Stress-Induced Silencing. *Molecular Cell*, **58**(2), 216–231.

Nora, E. P. *et al.* (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**(7398), 381–385.

Rao, S. *et al.* (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, **159**(7), 1665–1680.

Sauerwald, N. *et al.* (2020). Analysis of the structural variability of topologically associated domains as revealed by Hi-C. *NAR Genomics and Bioinformatics*, **2**(1), lqz008.

Sexton, T. *et al.* (2012). Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell*, **148**(3), 458–472.

Shin, H. *et al.* (2016). TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res*, **44**(7), e70–e70.

Szabo, Q. *et al.* (2019). Principles of genome folding into topologically associating domains. *eLife*, **5**(4), eaaw1668.

Ulianov, S. V. *et al.* (2016). Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res.*, **26**(1), 70–84.

Ulianov, S. V. *et al.* (2019). Nuclear lamina integrity is required for proper spatial organization of chromatin in Drosophila. *Nat Commun*, **10**(1),

1176.

Wang, Q. *et al.* (2018). Sub-kb Hi-C in d. melanogaster reveals conserved characteristics of TADs between insect and mammalian cells. *Nat. Commun.*, **9**(1), 188.

Wang, X. (2016). caltads: A hierarchical domain caller for hi-c data. *Zenodo*.

Weinreb, C. and Raphael, B. J. (2016). Identification of hierarchical chromatin domains. *Bioinformatics*, **32**(11), 1601–1609.