


METHOD

Open Access



RobusTAD: reference panel based annotation of nested topologically associating domains

Yanlin Zhang¹, Rola Dali¹ and Mathieu Blanchette^{1*} 

*Correspondence:
blanchem@cs.mcgill.ca

¹ School of Computer Science,
McGill University, Montréal,
Canada

Abstract

Topologically associating domains (TADs) are fundamental units of 3D genomes and play essential roles in gene regulation. Hi-C data suggests a hierarchical organization of TADs. Accurately annotating nested TADs from Hi-C data remains challenging, both in terms of the precise identification of boundaries and the correct inference of hierarchies. While domain boundary is relatively well conserved across cells, few approaches have taken advantage of this fact. Here, we present RobusTAD to annotate TAD hierarchies. It incorporates additional Hi-C data to refine boundaries annotated from the study sample. RobusTAD outperforms existing tools at boundary and domain annotation across several benchmarking tasks.

Keywords: Hi-C, TAD, Nonparametric test, Dynamic programming

Background

The hierarchical organization of mammalian chromosomes within the nucleus has been increasingly identified as an essential factor for cellular functions such as gene regulation, cell fate determination, and evolution [1]. Though the multi-scale chromosomal folding revealed by chromosome conformation capture techniques—such as Hi-C—are frequently studied [2–5], understanding its structural and functional roles is still in its infancy. Moreover, identifying spatial elements such as TADs and loops from Hi-C data is challenging, particularly due to the relatively low resolution permitted by Hi-C datasets of typical sequencing depth (200–500M valid read pairs) [6, 7].

TADs are self-interacting regions along the chromosome, manifesting as squares along the diagonal of Hi-C contact maps [8]. Researchers used to perform TAD annotations from Hi-C datasets at low resolution (i.e., 40 kb), and define TADs as megabase-scale structural elements [5]. Later, researchers observed that TADs can be much smaller (i.e., tens to hundreds kilobases) in mammalian chromosomes by investigating high-coverage and high-resolution Hi-C contact maps [3]. This observation led researchers to detect



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

TADs at high resolutions [1, 8, 9]. In addition, the study of sequence-encoded factors, such as CTCF (CCCTC-binding factor), that influence TAD formation also requires researchers to annotate TADs at much higher resolutions. Meanwhile, TAD-within-TAD (or subTAD) organization also gathered significant attention in recent years [8].

Many computational tools for TAD annotation have been proposed [3, 5, 10–23]. These approaches can be classified as either one-dimensional (1D) score-based and or matrix-based approaches. Score-based approaches, such as TopDom [10], Insulation Score (IS) [22], and OnTAD [15], assign each locus a score representing the strength of a potential TAD boundary and subsequently detect TAD boundaries by identifying local optima among the list of scores. Matrix-based approaches directly utilize two-dimensional (2D) data instead of transforming it into a 1D statistic. For example, Arrowhead [3] transforms the Hi-C contact map into an arrowhead-shaped feature map and subsequently identifies TADs by searching for corners in the transformed matrix. Chen and colleagues formulated TAD annotation as graph segmentation [23], viewing a Hi-C contact map as an adjacency matrix to model a chromosome as a graph and identified TADs through graph Laplacians.

Despite the numerous TAD annotation tools available, the identification of TAD hierarchy and the precise location of TAD boundaries at high resolution remains challenging. Most TAD annotation tools are designed for high-coverage contact maps, or operate at low-resolution. As reviewed in previous studies [7, 24–26], these tools are not robust to resolutions and sequencing coverages. Additionally, the predictions of TADs and domain boundaries show limited agreement across tools. Finally, most existing algorithms only use the contact map of the sample of interest to annotate TADs. Their performance is thus limited by the sequencing depth of the Hi-C data from that study. Most Hi-C data sets produced to date are in the range of 200M to 500M valid read pairs, with only a slow increase over time. We anticipate that this situation remains until sequencing costs drop dramatically.

The issue of insufficient data coverage also exists in many other biological data analysis tasks. Researchers often address this issue through introducing data from samples other than the sample under study [27, 28]. For example, though a SNP array typically covers only a few hundreds of thousands of loci, it is routine to infer unobserved genotypes through imputation with a reference panel containing a larger spectrum of genotyped variants [29, 30]. Similarly, homology modeling for protein structure prediction exploits a database of known structures [27]. Conversely, the vast amount of published Hi-C datasets are seldom employed to annotate TAD. While some research suggests that certain TADs are cell type or even replicate-specific [31], there are also observations demonstrating that many domain boundaries, or even entire TADs, are conserved across different cell types [32]. We hypothesize that since TADs are determined by genome sequence, epigenetics, and cellular dynamics, the vast majority of TADs present in a given cell-type/condition are also present in multiple other samples. Given the large number of Hi-C datasets in public depositories (e.g., [33]), the stage is set to develop TAD annotation approaches that better utilize existing Hi-C data sets. Recently, we introduced RefHiC [34], a reference panel enabled approach for TADs and chromatin loops annotation. Although we demonstrated that the introduction of a Hi-C reference panel enables RefHiC to significantly outperform alternatives in TAD annotation,

RefHiC suffers from several limitations: (i) RefHiC does not predict TAD hierarchies; (ii) the computationally intensive process of projecting Hi-C samples onto the latent space impedes including more samples into the reference panel.

Here, we introduce RobusTAD. RobusTAD is a TAD annotation algorithm that provides accurate and robust TAD annotation at high resolution. It improves TAD boundary annotation by leveraging publicly available Hi-C data and achieves superior performance by exploiting locally matched chromosome conformations (LMCC). Following TAD boundary annotation, it uses non-parametric tests and a dynamic programming algorithm to obtain the optimal nested TAD structure. RobusTAD outperforms existing TAD callers in a variety of contexts. We further demonstrated that RobusTAD is robust to low sequencing coverage and can produce high-resolution TAD annotations from Hi-C data of typical sequencing depth (250–300M reads). Finally, we show that applying RobusTAD to predict TAD at high resolution facilitates dissecting TADs according to transcription factor binding site profiles around TAD boundaries and consequently probe TAD formation.

Results

Overview of RobusTAD

RobusTAD takes a normalized Hi-C contact matrix as input and calls TADs in three steps (Fig. 1): (i) Low-accuracy TAD boundary identification based on the study sample; (ii) Refinement of TAD boundary locations based on locally-matched chromosome conformations from a reference panel of Hi-C data sets; (iii) Pairing of left and right boundaries into an optimal nested domain hierarchy.

Study sample based boundary identification is based on seeking local maxima in a vector of 1D nonparametric TAD boundary scores. RobusTAD assigns separate left and right TAD boundary scores to each locus by performing genomic distance stratified rank-sum test between upstream/downstream inter- and intra-domain interactions.

RobusTAD refines boundary calls made on the study sample by utilizing selected Hi-C samples from a reference panel. For a given candidate TAD boundary at position p , we define locally matched chromosome conformations $LMCC(p)$ as a collection of Hi-C samples in which a predicted TAD boundary occurs within 25 kb (i.e., five 5-kb bins) of p . It then computes refined boundary scores for the 50 kb region by combining the boundary scores from $LMCC(p)$ and the study sample itself; the position that reaches the maximum score is the final high-resolution boundary prediction.

In the third step, RobusTAD assembles a hierarchy of nested domains by pairing left and right boundary candidates using a dynamic programming algorithm, inspired by the Nussinov algorithm for RNA secondary structure prediction [35], that maximizes the full chromosomal TAD score (i.e., the sum of TADs scores). RobusTAD computes the TAD score with the distance-stratified rank-sum statistic of interactions between intra- and inter-domain in both upstream and downstream. To avoid the TAD score inflation caused by the presence of sub-TADs within a given region, lower-level domains previously identified during the execution of the algorithm are excluded from a region's TAD score calculation. The algorithm is guaranteed to produce the globally optimal nested TAD hierarchy.

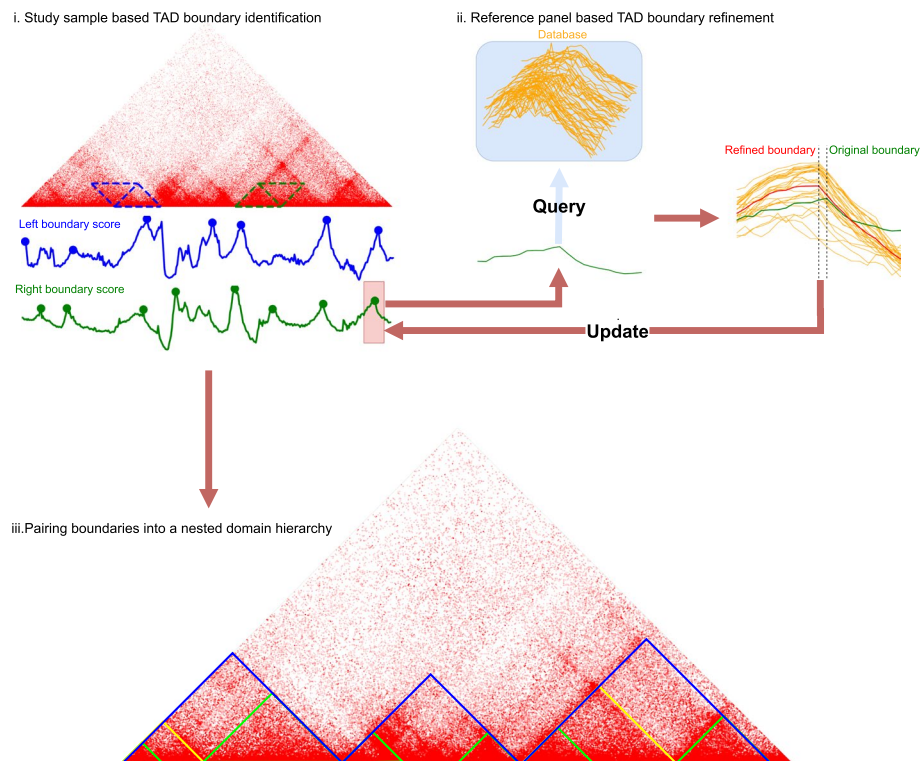


Fig. 1 Overview of the RobusTAD algorithm. RobusTAD detects TAD boundaries in three steps (i, ii, and iii). First, approximate left and right TAD boundaries are identified based on the study sample. Second, RobusTAD identifies locally matched chromosomal conformations (LMCCs) from a panel of reference data sets, and uses those LMCCs to refine the position of each TAD boundary. Finally (step iii), refined left and right boundaries are paired to form an optimal nested TAD hierarchy

Comparison with existing TAD callers

We compared the performance of RobusTAD to 14 other TAD callers: TopDom [10], Armatus [11], deDoc [12], Arrowhead [3], HiTAD [13], EAST [14], OnTAD [15], CaTCH [16], Grinch [17], Domaincall [5], GMAP [18], HiCseg [19], RefHiC [34], and IC-Finder [20]. We tried to include hierarchical TAD callers TADtree [36], TADpole [25], and SuperTAD [37] in our benchmarking, but they could not complete within a week of running time and hence we had to exclude them. Since RobusTAD and some other TAD callers detect nested TADs, we define TADs that do not contain any smaller TADs as level 0 TADs, and TADs that contain one or multiple smaller TADs as level 1+ TADs. We performed the benchmark evaluation experiments proposed by Zufferey et al. [24] on chromosomes 15–17 of Hi-C data for human GM12878 cells [3], down-sampled to 250 Million valid read pairs. We conducted all studies at 5 kb resolution and employed iterative normalized [38, 39] Hi-C contact maps as input. The running time varies significantly among tools in annotating TADs from all three chromosomes. OnTAD and callers that do not produce nested TADs are able to annotate the three chromosomes within 20 min. RobusTAD took a total of 1.5 hours to annotate the three chromosomes. Most of other nested TAD callers also require a similar amount of time.

We first compared the number and size of TADs identified by each tool. Interestingly, the number of TADs varies greatly, with Arrowhead identifying less than 500

TADs and OnTAD identifying around 3800. Most of the tools, including RobusTAD identified 1000 – 2000 TADs (Fig. 2a). Tools that identify more TADs naturally produce smaller TADs (Fig. 2b). RobusTAD identified TADs of a wide range of sizes, with a median size of 170 kb. UMAP embedding [40] on Measure of Concordance (MoC) [24] values among pairwise callers identifies three major caller groups. Among all callers, RobusTAD, RefHiC, Domaincall, GMAP, Armatus and HiTAD form a cluster with an average within-cluster MoC of 0.47 (Fig. 2c).

We then examined the quality of the TAD annotations produced by each tool. Additional file 1: Fig. S1 shows an example genomic region (chr16:10.2–12 Mb), with TADs annotated by RobusTAD and other TAD callers. TADs lack ground-truth annotation, so it is impossible to calculate the accuracy of TAD predictions. Thus, we used three metrics to evaluate each predicted TAD's quality. (i) RobusTAD's TAD

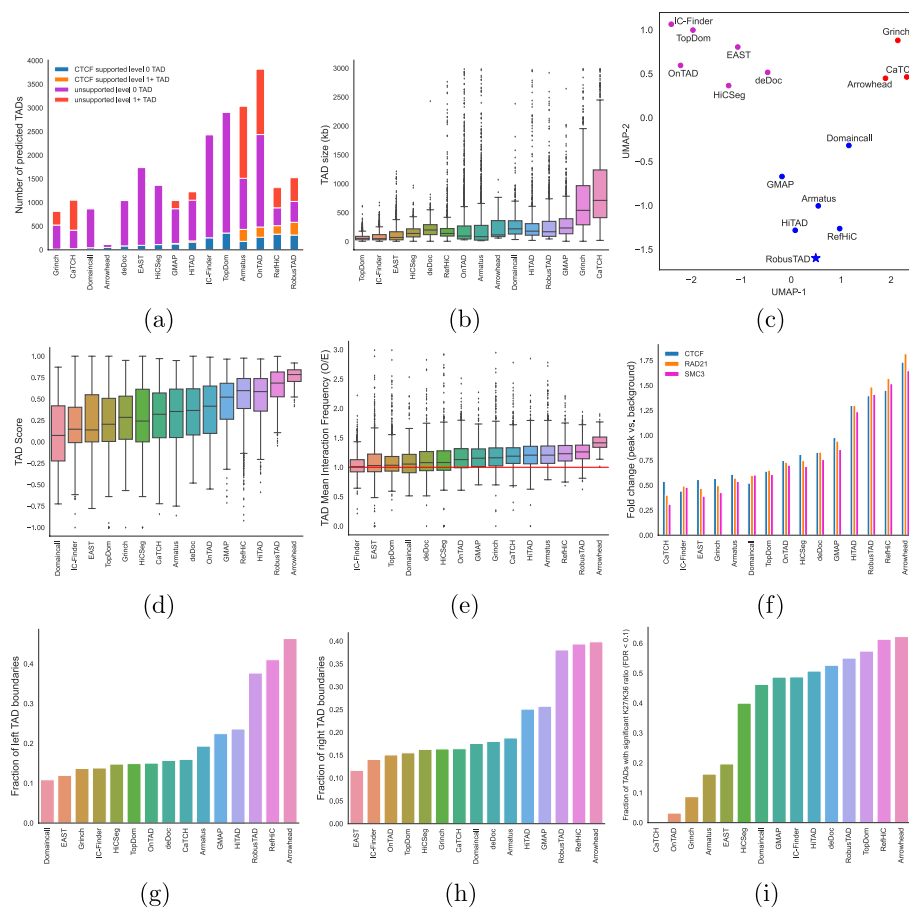


Fig. 2 Comparison of RobusTAD, and 14 other TAD callers on a GM12878 Hi-C data set of 250M valid read pairs. **a** Number of TADs predicted by different tools, and proportion of predicted TAD boundary pairs that are supported by CTCF ChIA-PET data. **b** Size distribution of predicted TADs. **c** U-MAP analysis performed on the Pearson's correlation matrix of the matrix of pairwise MoC between TADs identified by all callers. Comparison of the quality of TADs predicted by different tools using RobusTAD's TAD score (**d**), and TAD mean interaction frequency (observed/expected) (**e**). **f** Fold change of structural protein peak signals at TAD boundaries for CTCF, RAD21, and SMC3. Number of left (**g**) and right (**h**) boundaries that contain at least one CTCF ChIP-seq peak. **i** Fraction of TADs with significant log10 ratio between H3 K27 me3 and H3 K36 me3. Note: Panels **c**, **f**–**i** are generated with benchmarking code created by Zufferey et al. [24]

score (see the “[Methods](#)” section, on the full coverage data), (ii) mean interaction frequency (Observed over Expected, on the full coverage data) inside a TAD, and (iii) agreement with CTCF ChIA-PET data. The TAD score measures the enrichment of interaction frequencies inside a TAD by using its neighboring regions as the background. It ranges from -1 to 1 ; positive values indicate higher interactions within the TAD than across its boundaries. RobusTAD ranks second based on mean TAD score (Fig. 2d); only Arrowhead, a tool that predicts approximately 5 times fewer TADs, reaches a higher mean TAD score. Similar results are obtained when assessing predicted TADs based on average observed/expected ratios (Fig. 2e).

Loop TADs in mammalian genomes are TADs that exhibit a strong contact between their boundary loci [8]. We assessed TAD annotations by comparing predicted TAD boundary pairs with CTCF ChIA-PET data (allowing up to one 5-kb bin mismatch). Figure 2a shows that 582 (38%) TAD predictions made by RobusTAD match ChIA-PET data. This is both the largest number and the largest proportion of supported TAD predictions across all tools.

We also studied the performance of TAD annotations at the level of individual TAD boundaries. We first calculated the enrichment for ChIP-Seq signals of structural proteins (CTCF, RAD21, and SMC3) associated with predicted TAD boundaries (Fig. 2f and Additional file 1: Fig. S2). TAD boundaries predicted by most tools are enriched for these architectural proteins. RobusTAD ranked 3rd for the mean fold-change enrichment of the three structural proteins. Figure 2g, h compare left and right boundaries to CTCF ChIP-seq data (allowing 1-bin mismatch) separately; RobusTAD ranked 3rd for both left and right boundary predictions, slightly outperformed by Arrowhead, and the other reference panel enabled tool, RefHiC. Histone marks usually correlate with regulatory activity, and TADs are typically consistently enriched for either activating (H3 K36 me3) or repressive (H3 K27 me3) marks. We calculated the ratio between H3 K27 me3 and H3 K36 me3 within each TAD prediction and counted the fraction of TAD predictions where this ratio was particularly large or small (see the “[Methods](#)” section). Figure 2i shows RobusTAD ranked 4th, slightly outperformed by TopDom, RefHiC, and ArrowHead.

We then conducted a visual examination of TAD predictions made by all tools. We performed this analysis using rescaled pileup plots generated by Coolpup.py [41], with the “--local” and “--rescale” options. Figure 3 and S3 show most tools, including RobusTAD, identified TADs as square regions with increased interaction frequency, dot-corners, and well-defined domain borders. In contrast, tools such as Domaincall, deDoc, and ArmatuS yielded TAD predictions with less distinct domain borders. TAD predicted by Grinch and HiCseg are less enriched for Hi-C contacts. In addition, we observed clear vertical and horizontal stripes with increased interaction frequencies at the boundaries of TADs predicted by RobusTAD, HiCTAD, RefHiC, TopDom, and GMAP. The two stripes indicate these TAD callers can identify both TADs and sub-TADs.

Taken collectively, these results suggest that RobusTAD is the most accurate TAD caller, as it is the only tool ranking among the top four TAD callers in all accuracy metrics.

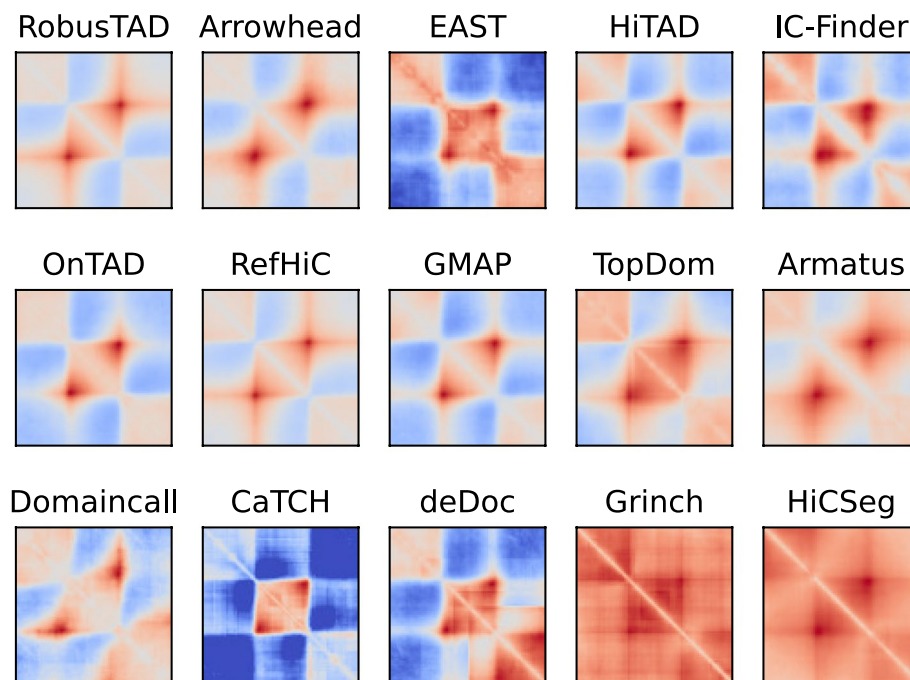


Fig. 3 Visual comparison of TADs predicted by RobusTAD and 14 other tools from GM12878 Hi-C data. The rescaled pileup plots aggregate areas around TAD predictions in the full-coverage Hi-C contact map. TAD predictions were annotated against a downsampled Hi-C contact map containing 250M valid read pairs

RobusTAD is robust to low sequencing coverage Hi-C data

TAD annotation is typically sensitive to sequencing depth. Many TAD callers do not perform well when the sequencing depth is low, and boundaries detected from contact maps of differing sequencing depths have been reported to lack reproducibility [6, 7]. We evaluated how RobusTAD and other tools performed on Hi-C contact maps of varying sequencing depths (generated from the combined Hi-C data set for GM12878 [3]), from data set of 4B valid read pairs down to downsampled versions with as few as 62.5M valid read pairs. As illustrated in Fig. 4a, different tools react differently to reduced coverage: some (including RobusTAD) conservatively reduce their predictions, while others are unaffected or even increase their number of predictions. These results suggest that tools like RobusTAD can mitigate false-positive identifications effectively.

We then assessed the tool's robustness by measuring the similarity between the predictions made on the highest coverage data set (i.e., 4 billion valid read pairs) to those made on downsampled data, both at the levels of TAD boundaries (Fig. 4b, using the Jaccard index) and TADs (Fig. 4c, using the Measure of Concordance). RobusTAD and RefHiC, the two reference panel based approaches, exhibit the highest levels of consistency at TAD boundaries level. Following Zufferey et al., we used the Measure of Concordance (MoC) [24] to compare two sets of TAD predictions. MoC does not handle overlapping TADs, thus we only included TADs that do not include any smaller TADs in this analysis. Figure 4c shows RefHiC, HiCseg, and RobusTAD outperformed other tools at most levels of coverage. Last, we evaluated TAD and domain boundary prediction accuracy by comparing predictions to CTCF ChIA-PET (Fig. 4d) and CTCF ChIP-Seq (Additional file 1: Fig. S4, S5) data. For contact maps containing more than 250M valid read pairs,

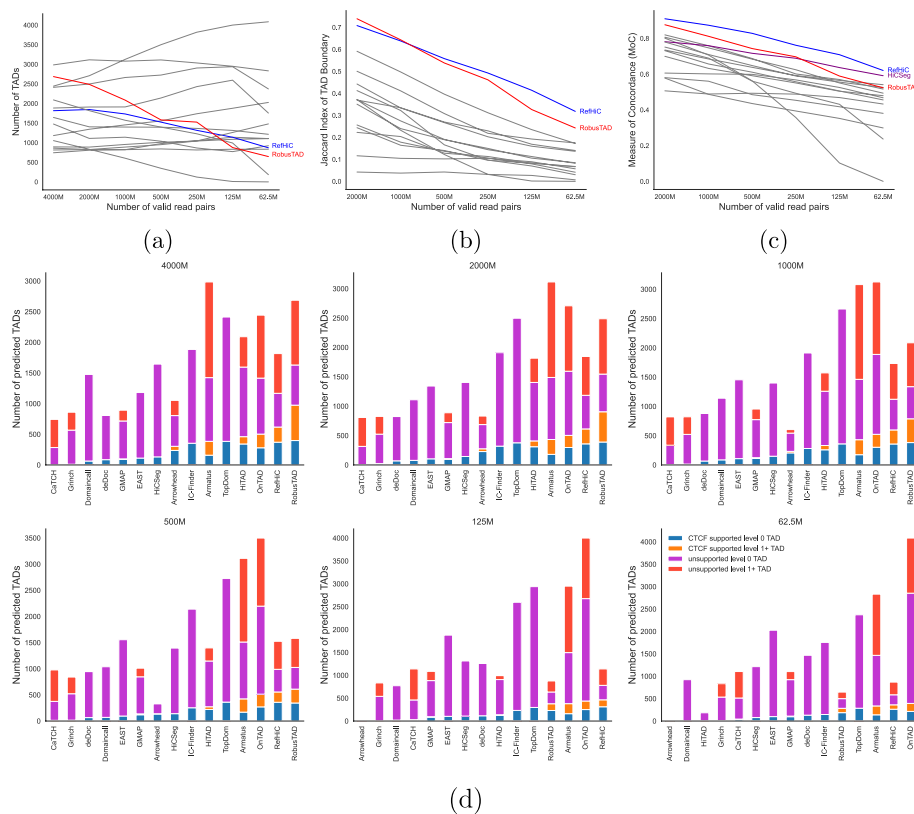


Fig. 4 Comparison of RobustTAD and other 14 TAD callers on downsampled GM12878 Hi-C data. **a** Number of TAD predicted from downsampled Hi-C data. Jaccard index of predicted TAD boundaries (**b**) and Concordance between TADs (**c**) predicted on full data (48 valid read pairs) compared to those predicted on the downsampled Hi-C data. **d** Number of TADs predicted from downsampled data, and proportion of predicted TAD boundary pairs that are supported by CTCF ChIA-PET data

RobustTAD performed the best. On data containing fewer valid contact pairs, RobustTAD is only slightly outmatched by RefHiC. Although OnTAD, Armitatus, and TopDom also identified more CTCF supported TADs from data containing few valid read pairs, they were less accurate than our tools, as they identified many more TADs that were not supported by CTCF data.

RobustTAD performs well across cell types

Here, we demonstrate that RobustTAD performs well across cell types. We applied RobustTAD and five other TAD callers (GMAP, HiTAD, Arrowhead, OnTAD, and RefHiC), to annotate TADs from Hi-C contact maps derived from IMR- 90 and K562 cell lines [3]. The rescaled pileup plots show that all tools successfully identified TADs as squares with increased interaction frequencies and dot-corners (Fig. 5a). In addition, TAD predictions made by all tools contain vertical and horizontal stripes with increased interaction frequency at its boundary locations. These stripes indicate these tools accurately detect TADs and subTADs from Hi-C contact maps. The number of TAD predicted by different tools from the two contact maps ranges from 500 to 3500, with RobustTAD detecting 2630 and 1710 TADs from the two contact maps (Fig. 5b). The mean expectation normalized interaction frequency (O/E) within a TAD further

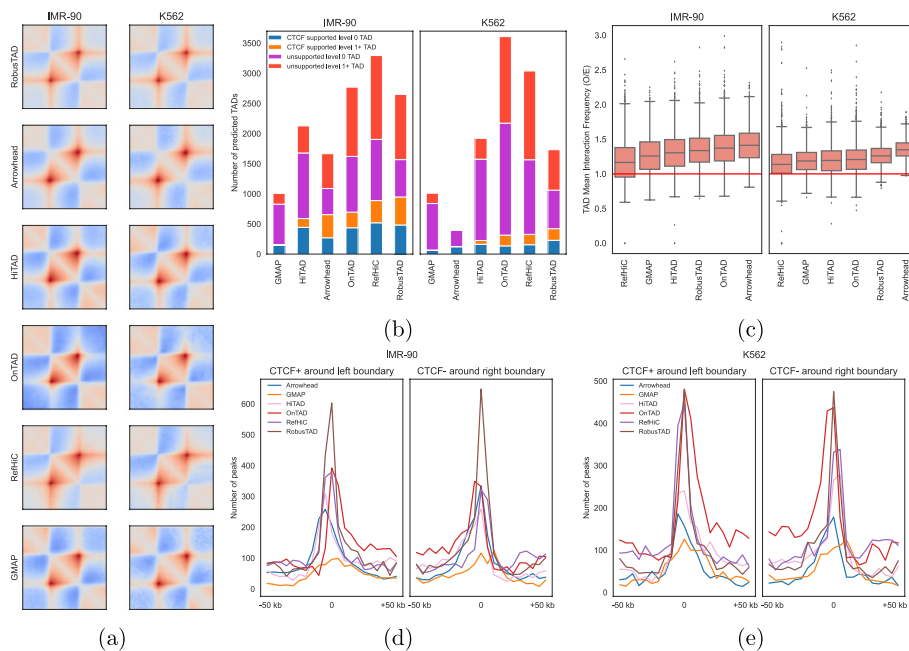


Fig. 5 Comparison of RobusTAD and other five TAD callers on Hi-C data derived from IMR-90 and K562 cell lines. **a** Rescaled pileup plots over predicted domains. **b** Number of TADs predicted by different tools, and proportion of predicted TAD boundary pairs that are supported by CTCF ChIA-PET data. **c** TAD mean interaction frequency (observed/expected). Occupancy of ChIP-seq identified forward and reverse CTCF binding sites as a function of distance to left **(d)** and right **(e)** boundary annotations

confirms that all tools, including RobusTAD, successfully identified TADs as a region with increased *cis*-contact pairs (Fig. 5c). Next, we compared boundary pairs to CTCF ChIA-PET data (Fig. 5b). The ChIA-PET data for IMR-90 contains 4957 contact pairs and the ChIA-PET data for K562 contains 2168 contact pairs. RobusTAD identified the most ChIA-PET supported TADs from both contact maps. Last, we evaluated boundary prediction accuracy by comparing predicted boundary to forward and reverse CTCF binding sites identified by ChIP-Seq experiment. Left and right boundaries predicted by RobusTAD are more enriched by CTCF binding sites than boundaries identified by alternative tools (Fig. 5d, e).

RobusTAD reveals multiple types of TADs

Building upon the high accuracy and resolution of RobusTAD, we used it to perform a study of TADs functionalities. We focused on TAD predictions made on the full set of autosomes for a combined Hi-C data set obtained from the GM12878 cell line [3]. We characterized a TAD as a binary vector of dimension $2 \times 116 = 232$, representing the ChIP-seq derived occupancy of 116 transcription factors [42] at its left and right domain boundaries. Unlike previous work that analyzed regulatory elements based on transcription factor binding, such as chromatin state annotations, this paired analysis of boundary regions offers a unique perspective often missing in earlier studies [43]. We identified six TAD groups by applying the UMAP algorithm [40] to project TADs onto a 2D space, followed by *K*-means clustering [44] (Fig. 6a). The symmetric pattern observed in the UMAP projection (Fig. 6a) and the group-averaged occupancy vectors (Fig. 6b)

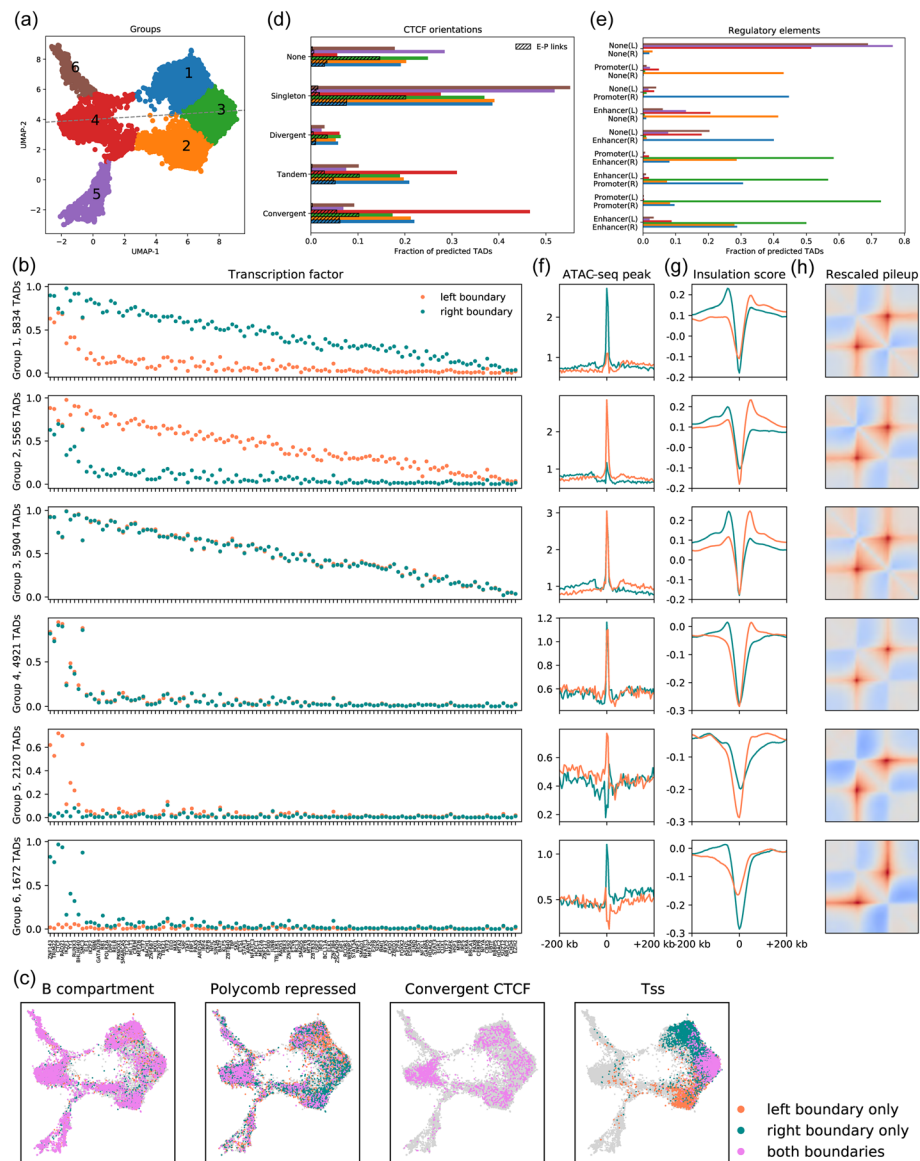


Fig. 6 Applying RobustTAD to Hi-C data for GM12878 cells reveals TAD groups. **a** A two-dimensional UMAP projection of TADs based on the occupancy of transcription factors at domain boundaries. **b** Occupancy of transcription factors in each group of TADs. **c** two-dimensional distributions in the UMAP projected space of TADs associated with different features. **d** Occupancy of different pairs of directional CTCF binding sites at domain boundaries. E-P links are domains supported by POLR2 A ChiA-PET data. **e** Proportion of annotated TADs with different regulatory element combinations at domain boundaries. Enrichment of ATAC-seq peaks (**f**) and insulation scores (**g**) at domain boundaries for each TAD group. **h** Rescaled pileup plots over TAD predictions for each TAD group

indicate that left and right domain boundaries play similar functional roles. Chromatin structural proteins such as ZNF143, CTCF, YY1, and subunits of cohesin complex (SMC3, and RAD21) are the most enriched proteins in all six groups. We also observed TRIM22 highly enriched at domain boundaries in all groups (Fig. 6b).

The distribution of transcription factors at TAD boundaries (Fig. 6b) and the chromatin accessibility quantified as the average count of ATAC-seq peaks at domain

boundaries (Fig. 6f) motivate us to interpret group assignments by investigating transcriptional activity and chromatin accessibility.

Group 3 is characterized by having both boundaries exhibiting evidence of transcriptional activity, with high TF occupancy (Fig. 6b) and chromatin accessibility (Fig. 6f), often involving pairs of regions annotated as active enhancers or promoters (as annotated by ENCODE's combined Segway ChromHMM segmentations [42]) (Fig. 6c, e). This is confirmed by comparing boundary pairs to Enhancer-Promoter links identified by a POLR2 A ChIA-PET experiment [42] (Fig. 6d). Additionally, we observe that Enhancer-Promoter pairs can mediate TAD formation even in the absence of CTCF binding sites, with 14.5% of Group- 3 TADs corresponding to such links but lacking CTCF occupancy.

Groups 1 and 2 only display evidence of transcriptional activity at one of their two boundaries, with the inactive boundary showing reduced TF occupancy. Although those three groups are quite different in terms of their activity profiles, their insulation profiles (Fig. 6g) and pile-up plots (Fig. 6h) are nearly identical.

Groups 4, 5, and 6 are characterized by TADs whose both boundaries are located in repressive chromatin (Fig. 6c) with low TF occupancy (Fig. 6b) and chromatin accessibility (Fig. 6f), and little overlap with active enhancers/promoters (Fig. 6e). Group- 4 TADs have both boundaries occupied by CTCF and associated structural proteins; these TADs' boundaries also display the highest level of convergent CTCF binding sites (Fig. 6d) and have sharper corner dots than domains in other groups (Additional file 1: Fig. S6), probably because of the reduced level of interactions in surrounding regions. On the contrary, Group 5 and 6 TADs lack CTCF at one or the other of their boundaries. They also exhibit weak insulation scores at the CTCF-free boundary (Fig. 6g, Additional file 1: Fig. S7, S8a), and weak dot corners (Additional file 1: Fig. S6).

Among all groups, domain boundaries in active regions are sharper than in repressed regions (Fig. 6g, Additional file 1: Fig. S7 and S8a). Domains in active regions are more enriched by Hi-C contacts than domains in repressive regions (Additional file 1: Fig. S8b).

We observe that domain boundaries shared by multiple TADs are more enriched for CTCF binding sites and activated promoters (Additional file 1: Fig. S9). We further studied the hierarchy structure of these TADs by classifying them into singleton TADs (isolated TADs that do not overlap with others), TADs (non-singleton TADs that do not reside within other TADs), and sub-TADs (non-singleton TADs found within larger TADs) (Additional file 1: Note S1 and Fig. S10). We found TADs frequently associated with boundaries marked by convergent CTCF motifs, and subTADs playing an important role in gene regulation, with a substantial portion being E-P links.

Discussion

Hi-C experiments and their derivatives have become routine in studying 3D genome organization at the genome-wide scale. Many Hi-C studies have been carried out in the past decade, and hundreds of Hi-C datasets have been published. Though this type of data has enabled the discovery of several key levels of 3D genome organizations (e.g., loops, TADs, and compartments), accurately identifying TAD boundaries and the ways in which they assemble to form a TAD hierarchy remain challenging with existing tools,

especially for Hi-C data with typical sequencing coverage. To deepen our understanding of 3D genome organization, high-resolution annotations of TAD hierarchy are required.

RobusTAD annotates high-resolution TAD boundaries and TAD hierarchy from Hi-C contact maps, taking advantage of a reference panel of high-quality Hi-C data sets. Including more reference samples improves annotation accuracy (Additional file 1: Fig. S11 and S12), so one can expect that RobusTAD will continue to get better as its reference panel grows. RobusTAD is based on a novel nonparametric statistic to score both domain boundaries and TADs. It is a distribution-free test to evaluate TAD and TAD boundaries. Thus, it is robust to changes in observation, such as those due to noise and sparsity. In addition, RobusTAD's separate scoring of left and right boundaries eases the dissection of domain boundaries. In contrast, most existing TAD callers do not quantify domain boundaries or amalgamate left and right boundaries as a single insulation locus.

RobusTAD overcomes the statistical challenges caused by high sparsity and signal-to-noise ratio limitation in a Hi-C contact map of typical coverage by identifying and combining many locally matched Hi-C data. At moderate sequencing depths, existing tools often fail at maintaining a low level of false identifications and identify many inaccurate TADs. Although a user can adjust parameters to limit the number of identified TADs in most tools, adjusting parameters is challenging and not necessarily effective at reducing false identification. Within RobusTAD, we use a simple and statistically sound target-decoy search strategy to select TAD boundaries from a list of candidate boundaries. A user only needs to specify the desired false discovery rate (FDR) threshold (α) to ensure that the final predictions to contain at most α expected false-positive boundaries.

Our tools outperformed many TAD callers in accuracy and reproducibility in identifying high-resolution TAD from multiple Hi-C contact maps. For instance, both CTCF ChIA-PET and CTCF occupancy data highlight the superiority of RobusTAD at both boundary detection and TAD assembly. The benefits of RobusTAD were shown to be particularly significant in typical moderate-to-low coverage Hi-C data. As demonstrated in Fig. 4a, d–i, RobusTAD can reduce false-positive identifications by identifying slightly fewer TAD boundaries from low-coverage Hi-C contact maps. In contrast, false-positive boundaries increased dramatically in predictions made by most other TAD callers when applied to low-coverage Hi-C contact maps. While RobusTAD was bested by RefHiC in terms of the accuracy of boundary annotation from very low coverage Hi-C data, it outmatched even that tool in terms of predicting CTCF ChIA-PET supported TADs. This advantage of RobusTAD over RefHiC is attributed to the newly developed dynamic programming algorithm for pairing TAD boundaries in RobusTAD. In addition, RobusTAD guarantees full interpretability without a loss of accuracy. When a boundary only appears in the study sample, RobusTAD will usually annotate it without using additional data from its reference panel. LMCC based boundary refinement makes mistake only if the boundary in the study sample is very close (i.e., within 50 kb) to another boundary in the reference panel. Given the superior performance achieved by RobusTAD, we believe this case rarely occurs.

Despite its advantages, RobusTAD has several limitations. First, given RefHiC slightly outperformed RobusTAD in some aspects of TAD boundary annotation, we believe RobusTAD does not fully capitalize on the advantages offered by the reference panel. It might be due to some weak boundaries that cannot be identified in the first step of

RobusTAD, and the heuristic approach for selecting reference samples is efficient but might be not as effective as RefHiC's data-driven approach. We can overcome this by replacing the first step with the deep learning model in RefHiC, at the cost of losing efficiency and interpretability. Second, the dynamic programming is time consuming, in part due to we need to compare two large sets of interaction frequencies to evaluate the score of a large candidate domain. We are planning to improve the running time by sampling a fraction of elements from the two sets to estimate the score of large candidate TADs.

Conclusions

RobusTAD allows for precise, high-resolution TAD annotation from Hi-C data of a wide range of sequencing depths, all the way down to only 62.5 million contact pairs. RobusTAD improves the performance of TAD boundary annotation by exploiting locally matched contact maps in a reference panel. By enabling high-resolution and robust analyses of topological domains from standard coverage Hi-C data, RobusTAD paves the way to gaining biological insights that had until now could only be possible from ultra-high coverage (and cost) data.

Methods

Notations

Consider an intra-chromosomal contact map $M = \{m_{ij}\}$, where m_{ij} represents (normalized) interaction frequency between bin i and j at fixed resolution r . RobusTAD aims to detect TAD boundaries $B = \{B^L, B^R\}$ and TADs $D = \{(B_i^L, B_i^R)\}$, where B^L and B^R are lists of left and right boundaries respectively, (B_i^L, B_i^R) indicates that the i^{th} left and i^{th} right boundaries form a TAD. Define $M_{[a,b]}$ as the submatrix corresponding genomic region $[a, b]$, and $S_{[a,b]}$, $S_{[a,b]}^L$ and $S_{[a,b]}^R$ as the domain score, left and right boundary scores for TAD (a, b) .

Boundary and domain scores

To calculate domain and boundary scores for (a, b) , we compared interactions for bins within $[a, b]$ and between bins in $[a, b]$ and its left and right flanking regions. Our null hypothesis in the nonparametric test assumes that, for each diagonal of the contact matrix, there are no differences between the distribution of within-TAD and across-TAD-boundary interactions (Additional file 1: Fig. S13). To quantify TAD, we performed a distance stratified (i.e., diagonal-wise) rank sum test between the two types of interactions. Define $D_k(a, b) = \{(i, i+k) : i \geq a, i+k \leq b\}$. We denote the TAD score evaluated from the k^{th} diagonal as $S_{[a,b]}^k$ and compute it using within-stratum ranks as follows:

$$S_{[a,b]}^k = \frac{\sum_{(i,j) \in D_k(a,b)} \sum_{(i',j') \in D_k(a-k,a+k-1) \cup D_k(b-k+1,b+k)} \mathbb{1}(m_{i,j} > \gamma m_{i',j'}) - \mathbb{1}(m_{i,j} < \frac{1}{\gamma} m_{i',j'})}{\sum_{(i,j) \in D_k(a,b)} \sum_{(i',j') \in D_k(a-k,a+k-1) \cup D_k(b-k+1,b+k)} \mathbb{1}(m_{i,j} > \gamma m_{i',j'}) + \mathbb{1}(m_{i,j} < \frac{1}{\gamma} m_{i',j'})}$$

where $\gamma \geq 1$ controls the minimum gap allowed between the two types of interactions. Setting $\gamma = 1$ is equivalent to the Wilcoxon rank sum test. We set $\gamma = 1$, and found that values in the range of $[1, 1.3]$ tend to produce similar results (Additional file 1: Fig. S14). The overall TAD score for region $[a, b]$ is a weighted sum of the per-stratum scores:

$$S_{[a,b]} = \frac{2}{3(1+b-a)(b-a)} \sum_{k=1}^{k=b-a} (b-a+k+1)S_{[a,b]}^k$$

$b-a+k+1$ is the number of Hi-C contact map entries used for comparison on the k^{th} diagonal, $\frac{3(1+b-a)(b-a)}{2}$ is the total number of Hi-C contact map entries used for comparison. $S_{[a,b]}$ falls between -1 and 1 , where $S_{[a,b]} = -1$ indicates all interactions inside the TAD are smaller than all interactions across TAD boundaries by a factor at least $\frac{1}{\gamma}$, $S_{[a,b]} = 0$ indicates no difference existed between the two types of interactions, $S_{[a,b]} = 1$ indicates all interactions inside the TAD exceed all interactions across TAD boundaries by a factor of at least γ . Note that if a nested TAD (a', b') was already determined to occur within (a, b) (with $a \leq a' < b' \leq b$), we exclude interactions belonging to (a', b') from the calculation.

We define per-stratum left and right boundary scores $S_{[a,b]}^{Lk}$ and $S_{[a,b]}^{Rk}$, left and right boundary scores $S_{[a,b]}^L$ and $S_{[a,b]}^R$ similarly but only using interactions across the corresponding boundary as the background:

$$S_{[a,b]}^{Lk} = \frac{\sum_{(i,j) \in D_k(a,b)} \sum_{(i',j') \in D_k(a-k,a+k-1)} \mathbb{1}(m_{i,j} > \gamma m_{i',j'}) - \mathbb{1}(m_{i,j} < \frac{1}{\gamma} m_{i',j'})}{\sum_{(i,j) \in D_k(a,b)} \sum_{(i',j') \in D_k(a-k,a+k-1)} \mathbb{1}(m_{i,j} > \gamma m_{i',j'}) + \mathbb{1}(m_{i,j} < \frac{1}{\gamma} m_{i',j'})}$$

$$S_{[a,b]}^{Rk} = \frac{\sum_{(i,j) \in D_k(a,b)} \sum_{(i',j') \in D_k(b-k+1,b+k)} \mathbb{1}(m_{i,j} > \gamma m_{i',j'}) - \mathbb{1}(m_{i,j} < \frac{1}{\gamma} m_{i',j'})}{\sum_{(i,j) \in D_k(a,b)} \sum_{(i',j') \in D_k(b-k+1,b+k)} \mathbb{1}(m_{i,j} > \gamma m_{i',j'}) + \mathbb{1}(m_{i,j} < \frac{1}{\gamma} m_{i',j'})}$$

$$S_{[a,b]}^L = \frac{1}{(b-a)} \sum_{k=1}^{k=b-a} S_{[a,b]}^{Lk}$$

$$S_{[a,b]}^R = \frac{1}{(b-a)} \sum_{k=1}^{k=b-a} S_{[a,b]}^{Rk}$$

Identifying candidate TAD boundaries

To identify domain boundaries from a normalized intra chromosomal contact map, we first compute a left boundary score $L_a = \max_{w \in \{w_{min}, \dots, w_{max}\}} S_{[a, a+w]}^L$ for each bin a along the whole chromosome. Right boundary scores are computed similarly: $R_b = \max_{w \in \{w_{min}, \dots, w_{max}\}} S_{[b-w, b]}^R$. On one hand, w should be as large as possible to include more interactions, thereby achieving a more robust estimation. On the other hand, w should not exceed the size of the TAD being evaluated, as interactions beyond the TAD size would distort the boundary strength calculation. For our analyses at 5 kb resolution, we use $w_{min} = 50$ kb, and $w_{max} = 250$ kb. We also conducted the calculation with alternative settings and found that setting w_{min} below 50 kb (i.e., 10 bins) often introduces false negative errors. In contrast, adjusting w_{max} to other larger values has a negligible impact on the results.

To identify left and right boundaries from boundary scores, we use `find_peak` function in SciPy [45] to identify local peaks. We assume the minimum distance between two domain

boundaries is 25 kb (i.e., five 5-kb bins) and set $distance=5$ in `find_peak`. The set of putative boundaries this identifies usually contains false positives. We use FDR control to select a subset of high confidence boundaries. Briefly, we produced a decoy contact map by shuffling interactions diagonal-wise. The shuffling strategy destroys all domains but maintains the interaction frequency decay pattern. We identify domain boundaries from this decoy contact map and compare scores for boundaries identified in the original and decoy Hi-C contact maps. We select the top boundaries at a FDR of α ($\alpha = 0.05$ for data containing more than 300M valid read pairs, $\alpha = 0.1$ for data containing less than 300M valid read pairs).

Refining boundary annotation by identifying locally matched chromosome conformations from the reference panel

Putative TAD boundaries predicted by the single-sample non-parametric test described above are often off by one or more bins, due to the noisy nature of the data. For a given left or right putative TAD boundary predicted at b_i , we define LMCCs as the subset of the Hi-C samples from our reference panel that have a predicted boundary with 25 kb (5 bins) of b_i . We then update the study sample's boundary scores for the 10-bin region centered at b_i as the mean boundary scores of the study sample and all selected reference samples (Additional file 1: Note S2). We also provide users the option to compute refined boundary scores as a weighted sum of each sample's boundary scores in RobusTAD, with the weights set according to each sample's read depth. This approach slightly improved accuracy in our experiments. Last, we update the boundary call as the peak position among the refined boundary scores.

Assembly of nested TADs from predicted boundaries

Given an intra-chromosomal contact map $M = \{m_{ij}\}$ and sets B^L and B^R of previously identified left and right boundaries, we sought to pair left and right boundaries to form hierarchical domains. Similar to OnTAD [15], we do not allow partial overlaps between domains in TAD predictions. This fully nested TAD hierarchy assumption allows us to use a dynamic programming algorithm to find a globally optimal solution. Our dynamic programming algorithm is inspired by the Nussinov algorithm [35] for RNA secondary structure prediction.

We denote the ordered multi-set of TAD boundaries as $(b_1, b_2, \dots, b_{n-1}, b_n)$, where $b_i \in B^L \cup B^R$. We define the globally optimal solution as the nested TAD hierarchy that maximizes the sum of scores of all TADs in the hierarchy, subject to the TADs' left and right boundaries being selected (potentially with repetition) from B^L and B^R . We create the dynamic programming table T of size $n \times n$, where T_{ij} stores the maximum sum of domain scores for all nested domains within region $[b_i, b_j]$. The forward pass of the dynamic programming fills the upper triangular portion of T , using the following recursion:

$$T_{ij} = \max_{i < k < j} T_{ik} + T_{kj} + \delta(i, j)$$

$$\delta(i, j) = \begin{cases} S_{[b_i, b_j]} & \text{if } b_i \in B^L \text{ and } b_j \in B^R \text{ and } S_{[b_i, b_j]} \geq \lambda \\ 0 & \text{otherwise} \end{cases}$$

λ defines the minimum score for pairing b_i and b_j as a domain. Since the TAD score is defined similarly to the TAD boundary score, we set $\lambda = 0.2$, which approximates the minimum TAD boundary scores observed in most of our experiments. In addition, we found that setting λ too high may result in fewer TAD predictions, but values in the range of $[0, 0.3]$ perform well in practice (Additional file 1: Fig. S15). The evaluation of $\delta(i, j)$ depends on k in the recursion function of T_{ij} as $S_{[b_i, b_j]}$ requires excluding nested TADs within $[b_i, b_j]$ which are identified with the dynamic programming algorithm in previous steps. We sequentially fill entries in T from the first to the furthest diagonals. To start, we initialize the first upper diagonal as

$$T_{i,i+1} = \begin{cases} S_{[b_i, b_{i+1}]} & b_i \in B^L, b_{i+1} \in B^R, \text{ and } S_{[b_i, b_{i+1}]} \geq \lambda \\ 0 & \text{otherwise} \end{cases}$$

Last, we select the optimal set of domains that maximize the sum of TAD scores for genomic region $[b_1, b_n]$ by backtracking from T_{1n} . The time complexity of this dynamic programming algorithm is $O(n^3)$ (where n is the number of boundaries) for filling the scoring table if $\delta(i, j)$ can be evaluated in constant time. However, since the evaluation of $\delta(i, j)$ involves comparing two sets of values that depend on i and j , this comparison could take up to $O(n)$ time in the worst case. As a result, the overall time complexity of the algorithm could increase to $O(n^4)$. RobusTAD can be executed on a typical human data set at 5 kb resolution in about 1 day. Some domain boundaries in b may be absent from the TAD hierarchy. They are treated as false positive domain boundaries or involved in partial overlap domains, which does not satisfy our assumption.

Curating Hi-C reference panel

We downloaded 177 published human Hi-C datasets (Additional file 1: Table S1) from the GEO database and uniformly processed them with distiller [46]. Reads were mapped against hg38 and we discarded reads with a mapping quality < 10 . This produced Hi-C contact maps at fixed resolutions and stored processed contact maps in multi-resolution cooler format (.mcool). Lastly, read count matrices were normalized using Cooler's iterative correction algorithm [38, 39]. We applied the single-sample version of RobusTAD with default parameters to calculate boundary scores for all of these Hi-C samples at 5 kb resolution and saved boundary scores and boundary calls as a reference database, to be used for the reference-panel based version of RobusTAD. This reference panel is specific to the data resolution. Applying RobusTAD to annotate Hi-C data at a different resolution requires users to create a corresponding reference panel, using the tools provided in the package.

Enrichment analysis and Measure of Concordance

We followed Zufferey et al. [24] to analyze enrichment of H3 K36 me3 and H3 K27 me3 histone marks and CTCF, SMC3, and RAD21 structural proteins within TADs or at their boundaries. For structural protein enrichment, we calculated the fold-change by comparing peak counts in a narrow interval around a boundary to those in distant flanks (Fold change = $\frac{\text{peak}}{\text{background}} - 1$). For histone marks, we calculated the average log10-ratio in small intervals within TADs and obtained empirical p -values using ten shuffles.

To compare TAD partitions, we used the Measure of Concordance (MoC) [24], which ranges from 0 (absence of concordance) to 1 (full concordance) and is defined as follows,

$$\text{MoC}(\mathbf{P}, \mathbf{Q}) = \begin{cases} 1 & \text{if } N_P = N_Q = 1 \\ \frac{1}{\sqrt{N_P N_Q - 1}} \left(\sum_{i=1}^{N_P} \sum_{j=1}^{N_Q} \frac{|\mathbf{F}_{i,j}|^2}{|\mathbf{P}_i||\mathbf{Q}_j|} - 1 \right) & \text{otherwise} \end{cases}$$

where $\mathbf{P} = \{\mathbf{P}_i\}$, and $\mathbf{Q} = \{\mathbf{Q}_i\}$ are sets of TADs including N_P and N_Q TADs, $\mathbf{F}_{i,j}$ is the overlap region between \mathbf{P}_i and \mathbf{Q}_j , and $|\cdot|$ represents cardinality. We only included TADs without any smaller TAD in this analysis.

Alternative approaches

This study compared RobustTAD to 14 other TAD callers. We ran TopDom, ArmatuS, Arrowhead, EAST, CaTCH, Domaincall (DI), GMAP, ICFinder, and HiCseg as suggested in [24]. As we performed the analysis at 5kb resolution, we have updated parameters related to resolutions accordingly. We ran HiTAD, RefHiC, and deDoc with their default settings. OnTAD: We set maxsz = 600 to allow OnTAD to detect TADs as large as 3 Mb. Grinch: following Lee and Roy [17], we detected TADs by setting the expected TAD length as 2 Mb, 1 Mb, and 500 Kb in three runs and combined all results. We observed tools such as Grinch reported invalid TAD annotations of length less than three bins and excluded these invalid annotations from further investigation. In addition, the convention of TAD definition often varies from tool to tool (with ± 1 bin shift). We converted them to the convention used in RobustTAD (i.e., boundary refers to the start of the left/right furthestmost bin inside the TAD).

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-025-03568-9>.

Additional file 1: PDF document containing supplementary notes, figures, and table.

Additional file 2: Peer review history.

Acknowledgements

The authors thank Dr. Yue Li and Dr. Jacek Majewski for useful discussions in this project.

Review history

The review history is available as Additional file 2.

Peer review information

Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

Y.Z. and M.B. conceived the study. R.D. co-conceived and implemented domain boundary annotation, and wrote portions of the manuscript. Y.Z. implemented RobustTAD, performed data analyses, and wrote the manuscript. M.B. supervised the project and wrote the manuscript. All authors read and approved the final paper.

Funding

This work was funded by a Genome Quebec/Canada grant to M.B.. Y.Z. is supported by FRQNT Doctoral (B2X) Research Scholarships.

Data availability

RobustTAD is available at <https://github.com/zhyanlin/RobustTAD>, and <https://zenodo.org/records/15011037> [47], under MIT license. All scripts and data required to reproduce figures and analyses are available at <https://doi.org/10.5281/zenodo.8306238> [48].

Hi-C data: All Hi-C data used in this project are from previous publications [3, 42, 49–88] and were downloaded from public repositories. Accession numbers are provided in Additional file 1: Table S1.

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 29 September 2023 Accepted: 4 April 2025

Published online: 19 May 2025

References

- Bonev B, Cavalli G. Organization and function of the 3d genome. *Nat Rev Genet*. 2016;17(11):661–78.
- Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009;326(5950):289–93.
- Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. 2014;159(7):1665–80.
- Krietenstein N, Abraham S, Venev SV, Abdennur N, Gibcus J, Hsieh T-HS, et al. Ultrastructural details of mammalian chromosome architecture. *Mol Cell*. 2020;78(3):554–65.
- Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. 2012;485(7398):376–80.
- Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. Comparison of computational methods for Hi-C data analysis. *Nat Methods*. 2017;14(7):679–85.
- Dali R, Blanchette M. A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Res*. 2017;45(6):2994–3005.
- Beagan JA, Phillips-Cremens JE. On the existence and functionality of topologically associating domains. *Nat Genet*. 2020;52(1):8–16.
- Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically associating domains. *Sci Adv*. 2019;5(4):eaaw1668.
- Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, Zhou XJ. Topdom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res*. 2016;44(7):e70.
- Filippova D, Patro R, Duggal G, Kingsford C. Identification of alternative topological domains in chromatin. *Algorithm Mol Biol*. 2014;9(1):1–11.
- Li A, Yin X, Xu B, Wang D, Han J, Wei Y, et al. Decoding topologically associating domains with ultra-low resolution hi-c data by graph structural entropy. *Nat Commun*. 2018;9(1):1–12.
- Wang X-T, Cui W, Peng C. HiTAD: detecting the structural and functional hierarchies of topologically associating domains from chromatin interactions. *Nucleic Acids Res*. 2017;45(19):e163.
- Ardakany AR, Lonardi S. Efficient and accurate detection of topologically associating domains from contact maps. In: 17th International Workshop on Algorithms in Bioinformatics (WABI 2017). Leibniz International Proceedings in Informatics (LIPIcs), vol. 88. Wadern: Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik; 2017. pp. 22:1–22:11.
- An L, Yang T, Yang J, Nuebler J, Xiang G, Hardison RC, et al. OnTAD: hierarchical domain structure reveals the divergence of activity among tads and boundaries. *Genome Biol*. 2019;20(1):1–16.
- Zhan Y, Mariani L, Barozzi I, Schulz EG, Blüthgen N, Stadler M, et al. Reciprocal insulation analysis of hi-c data shows that tads represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome Res*. 2017;27(3):479–90.
- Lee D-I, Roy S. Grinch: simultaneous smoothing and detection of topological units of genome organization from sparse chromatin contact count matrices with matrix factorization. *Genome Biol*. 2021;22(1):1–31.
- Yu W, He B, Tan K. Identifying topologically associating domains and subdomains by gaussian mixture model and proportion test. *Nat Commun*. 2017;8(1):1–9.
- Lévy-Leduc C, Delattre M, Mary-Huard T, Robin S. Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics*. 2014;30(17):i386–92.
- Haddad N, Vaillant C, Jost D. Ic-finder: inferring robustly the hierarchical organization of chromatin folding. *Nucleic Acids Res*. 2017;45(10):e81.
- Chen F, Li G, Zhang MQ, Chen Y. Hicdb: a sensitive and robust method for detecting contact domain boundaries. *Nucleic Acids Res*. 2018;46(21):11239–50.
- Crane E, Bian Q, Patton McCord R, Lajoie BR, Wheeler BS, Ralston EJ, et al. Condensin-driven remodelling of x chromosome topology during dosage compensation. *Nature*. 2015;523(7559):240–4.
- Chen J, Hero AO III, Rajapakse I. Spectral identification of topological domains. *Bioinformatics*. 2016;32(14):2151–8.
- Zufferey M, Tavernari D, Oricchio E, Ciriello G. Comparison of computational methods for the identification of topologically associating domains. *Genome Biol*. 2018;19(1):1–18.
- Soler-Vila P, Cusco P, Farabella I, Di Stefano M, Marti-Renom MA. Hierarchical chromatin organization detected by tadpole. *Nucleic Acids Res*. 2020;48(7):e39.
- Liu K, Li H-D, Li Y, Wang J, Wang J. A comparison of topologically associating domain callers based on Hi-C data. *IEEE/ACM Trans Comput Biol Bioinforma*. 2022;20(1):15–29.
- Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Šali A. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*. 2000;29(1):291–325.
- Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. *Annu Rev Genomics Hum Genet*. 2009;10:387–406.
- Howe BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*. 2009;5(6):e1000529.

30. Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. *Am J Hum Genet.* 2018;103(3):338–48.
31. Sauerwald N, Singhal A, Kingsford C. Analysis of the structural variability of topologically associated domains as revealed by Hi-C. *NAR Genom Bioinforma.* 2020;2(1):lqz008.
32. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.* 2016;17(8):2042–59.
33. Dekker J, Belmont AS, Guttman M, Leshyk VO, Lis JT, Lomvardas S, et al. The 4d nucleome project. *Nature.* 2017;549(7671):219–26.
34. Zhang Y, Blanchette M. Reference panel guided topological structure annotation of Hi-C data. *Nat Commun.* 2022;13(1):7426.
35. Nussinov R, Jacobson AB. Fast algorithm for predicting the secondary structure of single-stranded rna. *Proc Natl Acad Sci.* 1980;77(11):6309–13.
36. Weinreb C, Raphael BJ. Identification of hierarchical chromatin domains. *Bioinformatics.* 2016;32(11):1601–9.
37. Zhang YW, Wang MB, Li SC. Supertad: robust detection of hierarchical topologically associated domains with optimized structural information. *Genome Biol.* 2021;22(1):1–20.
38. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nat Methods.* 2012;9(10):999.
39. Abdennur N, Mirny LA. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics.* 2020;36(1):311–6.
40. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. 2018. arXiv preprint arXiv:1802.03426.
41. Flyamer IM, Illingworth RS, Bickmore WA. Coolpup.py: versatile pile-up analysis of hi-c data. *Bioinformatics.* 2020;36(10):2980–5.
42. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489(7414):57–74.
43. Libbrecht MW, Chan RCW, Hoffman MM. Segmentation and genome annotation algorithms for identifying chromatin state and other genomic patterns. *PLoS Comput Biol.* 2021;17(10):e1009423.
44. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res.* 2011;12:2825–30.
45. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat Methods.* 2020;17:261–72.
46. Open2C, Abdennur N, Fudenberg G, Flyamer IM, Galitsyna AA, Goloborodko A, et al. Pairtools: from sequencing data to chromosome contacts. *PLoS Comput Biol.* 2024;20(5):e1012164.
47. Zhang Y. Robustad: Reference panel based annotation of nested topologically associating domains. 2025. <https://zenodo.org/records/1501103>.
48. Zhang Y. Robustad experiment. 2023. <https://doi.org/10.5281/zenodo.8306238>.
49. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, et al. Chromatin architecture reorganization during stem cell differentiation. *Nature.* 2015;518(7539):331–6.
50. Bunting KL, Soong TD, Singh R, Jiang Y, Béguelin W, Poloway DW, et al. Multi-tiered Reorganization of the Genome during B Cell Affinity Maturation Anchored by a Germinal Center-Specific Locus Control Region. *Immunity.* 2016;45(3):497–512.
51. Haarhuis JHI, van der Weide RH, Blomen VA, Yáñez-Cuna JO, Amendola M, van Ruiten MS, et al. The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell.* 2017;169(4):693–707.
52. Fritz AJ, Ghule PN, Boyd JR, Tye CE, Page NA, Hong D, et al. Intracellular and higher-order chromatin organization of the major histone gene cluster in breast cancer. *J Cell Physiol.* 2018;233(2):1278–90.
53. Rubin AJ, Barajas BC, Furlan-Magaril M, Lopez-Pajares V, Mumbach MR, Howard I, et al. Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation. *Nat Genet.* 2017;49(10):1522–8.
54. Rao SSP, Huang S-C, St Hilaire BG, Engreitz JM, Perez EM, Kieffer-Kwon KR, et al. Cohesin Loss Eliminates All Loop Domains. *Cell.* 2017;171(2):305–20.
55. Luo Z, Rhie SK, Lay FD, Farnham PJ. A Prostate Cancer Risk Element Functions as a Repressive Loop that Regulates HOXA13. *Cell Rep.* 2017;21(6):1411–7.
56. Wutz G, Várnai C, Nagasaka K, Cisneros DA, Stocsits RR, Tang W, et al. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.* 2017;36(24):3573–99.
57. Zirkel A, Nikolic M, Sofiadis K, Mallm JP, Brackley CA, Gothe H, et al. HMGB2 Loss upon Senescence Entry Disrupts Genomic Organization and Induces CTCF Clustering across Cell Types. *Mol Cell.* 2018;70(4):730–44.
58. Kojic A, Cuadrado A, De Koninck M, Giménez-Llorente D, Rodríguez-Corsino M, Gómez-López G, et al. Distinct roles of cohesin-SA1 and cohesin-SA2 in 3D chromosome organization. *Nat Struct Mol Biol.* 2018;25(6):496–504.
59. Rodrigues P, Patel SA, Harewood L, Olan I, Vojtasova E, Syafruddin SE, et al. B-Dependent Lymphoid Enhancer Co-option Promotes Renal Carcinoma Metastasis. *Cancer Discov.* 2018;8(7):850–65.
60. Lyu X, Rowley MJ, Corces VG. Architectural Proteins and Pluripotency Factors Cooperate to Orchestrate the Transcriptional Response of hESCs to Temperature Stress. *Mol Cell.* 2018;71(6):940–55.
61. Heinz S, Texari L, Hayes MGB, Urbanowski M, Chang MW, Givarkes N, et al. Transcription Elongation Can Affect Genome 3D Structure. *Cell.* 2018;174(6):1522–36.
62. Guo Y, Perez AA, Hazelett DJ, Coetzee GA, Rhie SK, Farnham PJ. CRISPR-mediated deletion of prostate cancer risk-associated CTCF loop anchors identifies repressive chromatin loops. *Genome Biol.* 2018;19(1):160.
63. Amat R, Böttcher R, Le Dily F, Vidal E, Quilez J, Cuartero Y, et al. Rapid reversible changes in compartments and local chromatin organization revealed by hyperosmotic shock. *Genome Res.* 2019;29(1):18–28.
64. Raviram R, Rocha PP, Luo VM, Swanzey E, Miraldi ER, Chuong EB, et al. Analysis of 3D genomic interactions identifies candidate host genes that transposable elements potentially regulate. *Genome Biol.* 2018;19(1):216.

65. Le Dily F, Vidal E, Cuartero Y, Quilez J, Nacht AS, Vicent GP, Carbonell-Caballero J, Sharma P, Villanueva-Cañas JL, Ferrari R, De Llobet LI, Verde G, Wright RHG, Beato M. Hormone-control regions mediate steroid receptor-dependent genome organization. *Genome Res.* 2019;29(1):29–39.
66. Nir G, Farabella I, Pérez Estrada C, Ebeling CG, Beliveau BJ, Sasaki HM, et al. Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLoS Genet.* 2018;14(12):e1007872.
67. Kantidze OL, Luzhin AV, Nizovtseva EV, Safina A, Valieva ME, Golov AK, et al. The anti-cancer drugs curaxins target spatial genome organization. *Nat Commun.* 2019;10(1):1441.
68. Bertero A, Fields PA, Ramani V, Bonora G, Yardimci GG, Reinecke H, et al. Dynamics of genome reorganization during human cardiogenesis reveal an RBM20-dependent splicing factory. *Nat Commun.* 2019;10(1):1538.
69. Paulsen J, Lijakat Ali TM, Nekrasov M, Delbarre E, Baudement MO, Kurscheid S, et al. Long-range interactions between topologically associating domains shape the four-dimensional genome during differentiation. *Nat Genet.* 2019;51(5):835–43.
70. Battle SL, Doni Jayavelu N, Azad RN, Hesson J, Ahmed FN, Overbey EG, et al. Enhancer Chromatin and 3D Genome Architecture Changes from Naïve to Primed Human Embryonic Stem Cell States. *Stem Cell Rep.* 2019;12(5):1129–44.
71. Wu S, Fatkhutdinov N, Rosin L, Luppino JM, Iwasaki O, Tanizawa H, et al. ARID1A spatially partitions interphase chromosomes. *Sci Adv.* 2019;5(5):eaaw5294.
72. Tian L, Shao Y, Nance S, Dang J, Xu B, Ma X, et al. Long-read sequencing unveils IGH-DUX4 translocation into the silenced IGH allele in B-cell acute lymphoblastic leukemia. *Nat Commun.* 2019;10(1):2789.
73. Lalonde S, Codina-Fauteux VA, de Bellefon SM, Leblanc F, Beaudoin M, Simon MM, et al. Integrative analysis of vascular endothelial cell genomic features identifies AIDA as a coronary artery disease candidate gene. *Genome Biol.* 2019;20(1):133.
74. Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, et al. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat Genet.* 2019;51(9):1380–8.
75. Jung I, Schmitt A, Diao Y, Lee AJ, Liu T, Yang D, et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat Genet.* 2019;51(10):1442–9.
76. Dall'Agnese A, Caputo L, Nicoletti C, di Iulio J, Schmitt A, Gatto S, et al. Transcription Factor-Directed Re-wiring of Chromatin Architecture for Somatic Cell Nuclear Reprogramming toward trans-Differentiation. *Mol Cell.* 2019;76(3):453–72.
77. Ooi WF, Nargund AM, Lim KJ, Zhang S, Xing M, Mandoli A, et al. Integrated paired-end enhancer profiling and whole-genome sequencing reveals recurrent *CCNE1* and *IGF2* enhancer hijacking in primary gastric adenocarcinoma. *Gut.* 2020;69(6):1039–52.
78. Achinger-Kawecka J, Valdes-Mora F, Luu PL, Giles KA, Caldon CE, Qu W, et al. Epigenetic reprogramming at estrogen-receptor binding sites alters 3D chromatin landscape in endocrine-resistant breast cancer. *Nat Commun.* 2020;11(1):320.
79. Akdemir KC, Le VT, Chandran S, Li Y, Verhaak RG, Beroukhi R, et al. Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat Genet.* 2020;52(3):294–305.
80. Wutz G, Ladurner R, St Hilaire BG, Stocsits RR, Nagasaka K, Pignard B, et al. ESCO1 and CTCF enable formation of long chromatin loops by protecting cohesinstag1 from WAPL. *Elife.* 2020;9:e52091.
81. Higashijima Y, Matsui Y, Shimamura T, Nakaki R, Nagai N, Tsutsumi S, et al. Coordinated demethylation of H3K9 and H3K27 is required for rapid inflammatory responses of endothelial cells. *EMBO J.* 2020;39(7):e103949.
82. Kloetgen A, Thandapani P, Ntziachristos P, Ghebrechristos Y, Nomikou S, Lazaris C, et al. Three-dimensional chromatin landscapes in T cell acute lymphoblastic leukemia. *Nat Genet.* 2020;52(4):388–400.
83. Sati S, Bonev B, Szabo Q, Jost D, Bensadoun P, Serra F, et al. 4D Genome Rewiring during Oncogene-Induced and Replicative Senescence. *Mol Cell.* 2020;78(3):522–38.
84. Rosencrance CD, Ammouri HN, Yu Q, Ge T, Rendleman EJ, Marshall SA, Eagen KP. Chromatin Hyperacetylation Impacts Chromosome Folding by Forming a Nuclear Subcompartment. *Mol Cell.* 2020;78(1):112–26.
85. Casa V, Moronta Gines M, Gade Gusmao E, Slotman JA, Zirkel A, Josipovic N, et al. Redundant and specific roles of cohesin STAG subunits in chromatin looping and transcriptional control. *Genome Res.* 2020;30(4):515–27.
86. Stik G, Vidal E, Barrero M, Cuartero Y, Vila-Casadesús M, Mendieta-Esteban J, et al. CTCF is dispensable for immune cell transdifferentiation but facilitates an acute inflammatory response. *Nat Genet.* 2020;52(7):655–61.
87. Yang J, McGovern A, Martin P, Duffus K, Ge X, Zarrineh P, et al. Analysis of chromatin organization and gene expression in T cells identifies functional genes for rheumatoid arthritis. *Nat Commun.* 2020;11(1):4402.
88. Brown MA, Dotson GA, Ronquist S, Emons G, Rajapakse I, Ried T. TCF7L2 silencing results in altered gene expression patterns accompanied by local genomic reorganization. *Neoplasia.* 2021;23(2):257–69.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.