Published online 25 May 2022

Nucleic Acids Research, 2022, Vol. 50, No. 15 ***e88***
https://doi.org/10.1093/nar/gkac422

Detecting TAD-like domains from RNA-associated interactions

Yu Wei Zhang [1], Lingxi Chen [1] and Shuai Cheng Li [1,2,*]

[1]Department of Computer Science, City University of Hong Kong, Hong Kong and [2]Department of Biomedical Engineering, City University of Hong Kong, Hong Kong

Received March 14, 2022; Revised May 02, 2022; Editorial Decision May 03, 2022; Accepted May 09, 2022

ABSTRACT

Topologically associated domains (TADs) are crucial chromatin structural units. Evidence has illustrated that RNA–chromatin and RNA–RNA spatial interactions, so-called RNA-associated interactions (RAIs), may be associated with TAD-like domains (TLDs). To decode hierarchical TLDs from RAIs, we proposed SuperTLD, a domain detection algorithm incorporating imputation. We applied SuperTLD on four RAI data sets and compared TLDs with the TADs identified from the corresponding Hi-C datasets. The TLDs and TADs share a moderate similarity of hierarchies ≥ 0.5312 and the finest structures ≥ 0.8295. Comparison between boundaries and domains further demonstrated the novelty of TLDs. Enrichment analysis of epigenetic characteristics illustrated that the novel TLDs exhibit an enriched CTCF by 0.6245 fold change and H3 histone marks enriched within domains. GO analysis on the TLD novel boundaries exhibited enriched diverse terms, revealing TLDs' formation mechanism related closely to gene regulation.

INTRODUCTION

RNAs physically associate with chromatin or other RNAs via two main mechanisms (1–6). The newly transcribed RNAs, nascent RNAs, remain at their template DNA sequence sites and interact with chromatin in *cis-interactions*. After being released from the site of transcription, RNAs interact with specific genomic loci, or RNAs, through RNA: DNA hybrid formation or protein-mediated mechanism in *trans-interactions*. The RNA-associated interactions (RAIs) are involved in multiple molecular mechanisms, such as DNA methylation (7), long non-coding RNA regulation (3,5,6,8) and fusion transcript formation (9).

Recent experimental technologies capture the genome-wide interactions between RNA and DNA in the nucleus, such as MARGI (10), GRID-seq (8), ChAR-seq (11), iMARGI (12) and RADICL-seq (13). Through various strategies, these technologies capture the chimeric fragments between chromatin-associated RNAs and their spatially adjacent genomic sequences. Compared to technologies that merely capture a specific RNA's chromatin binding regions (14), these methods can handle both coding and non-coding RNAs, thus contributing to a global view of RNA–chromatin interactions. In addition, experimental technologies are developed to infer the spatial associations between different RNAs (4,15) through chromatin association or RNA-binding protein mediation. This work refers to RNA–chromatin interactions and RNA–RNA interactions collectively as RAIs.

Several works have demonstrated that RAIs are related to 3D chromatin structures, especially topologically associated domains (TADs) (8,9,11). Chromosomes are organized into TADs in which DNA sequences within the same domain interact more frequently than sequences from adjacent TADs (16–19). TADs are hierarchical; adjacent TADs can form a new TAD. Li *et al.* demonstrated that GRID-seq has high global concordance with Hi-C data in mESCs and S2 cells, and chromatin-associated RNAs are predominantly confined within TADs (8). Guh *et al.* showed that the lncRNA Xist can prevent the formation of TADs through repelling positive chromatin factors and cohesin (6). Bonetti *et al.* found that the DNA tags of RNA–chromatin interactions are significantly enriched at TAD boundaries (13). Based on the fact that TAD boundaries constrain the spread of transcriptional activities, Bonetti *et al.* further demonstrated its barrier effect as preventing the free diffusion of RNA migration. Altogether, robust evidence has shown that RAIs could extrapolate genomic interactions related to RNA production and regulation. However, the systematic survey of RAI inferred TAD-like domains (TLDs), including the computational identification method and comprehensive evaluation, remains challenging.

To this end, we introduced SuperTLD, an imputation-based domain detection method, to infer hierarchical TLDs from RAIs. SuperTLD first imputes the missing interaction frequencies through a negative binomial model with a mean-variance linear dependency for genes. Then a Bayesian correction is incorporated into the structural information theory to detect the hierarchical domains

[*]To whom correspondence should be addressed. Tel: +852 34429412; Fax: +852 34420503; Email: shuaicli@cityu.edu.hk

from the imputed RAIs. In this work, we collected three RNA–chromatin interaction data sets, iMARGI, GRID-seq, RADICL-seq and one RNA–RNA spatial interaction data set RIC-seq. We detected their hierarchical TLDs via SuperTLD and explored the inferred TLDs' structural and functional properties.

MATERIALS AND METHODS

RAI data sets

In this study, we considered four RAI data sets, among which three RNA–chromatin interaction data sets are from iMARGI(12), GRID-seq(8), RADICL-seq(13) and one RNA–RNA spatial interaction data set is from RIC-seq (4). For each dataset, we partition the genome into bins, with each bin along the genome containing the same number of consecutive nucleic acids. The *interaction map* is constructed for each chromosome using the intra-chromosome interactions. The RIC-seq data can be organized as a symmetric bin-bin interaction map $A \in \mathbb{R}^{n\times n}$, $a_{i,j} \geq 0$. The interaction maps of iMARGI, RADICL-seq, and GRID-seq can be represented as $A \in \mathbb{R}^{m\times n}$, $a_{i,j} \geq 0$, where m and n are the number of genes and bins, respectively.

(1) *The human embryonic kidney (HEK293T) cell line RNA–chromatin interactions from iMARGI* (12) with accession number GSE122690. We downloaded the RNA–DNA interaction read pairs (in BEDPE format) and constructed the RNA–chromatin interaction map through binning at resolution 100 kb, with the reference genome hg38.

(2) *The mouse embryonic stem cell line RNA–chromatin interactions from RADICL-seq*(13) with accession number GSE132190. We downloaded the processed interactions (in tab-delimited text format) with crosslinking using 1% formaldehyde. We derived the RNA–chromatin interaction map at bin resolution 100 kb, with mm10 reference genome.

(3) *The Drosophila S2 cell line RNA–chromatin interactions from GRID-seq* (8) with accession number GSE82312. We downloaded the processed data where the GRID-seq value is normalized interaction density between RNA and DNA (in tab-delimited text format), with Drosophila genome dm3. Given the short chromosome size of Drosophila melanogaster, We used a bin resolution of 40kb to derive the RNA–chromatin interaction map.

(4) *The HeLa RNA–RNA spatial interactions from RIC-seq* (4) with accession number GSE127188. We downloaded the processed RNA–RNA interactions (in hic format) with hg19 reference genome and derived the symmetric interaction map at bin resolution 100kb.

Hi-C data and preprocessing

We downloaded the corresponding Hi-C datasets of the four RAI data sets. These are the human embryonic kidney cell line GSE44267 (20), mouse embryonic stem cell line GSE96107 (21), Drosophila melanogaster embryonic cell line GSE34453 (22), and HeLa cell line (requested from dbGap). We performed Hi-C read alignment using the same reference genome as each RAI data set and built the Hi-C contact map with the same bin resolution (23). Knight-Ruiz normalization (24) is applied to the raw contact map to equalize the contact numbers of each bin. The normalized Hi-C contact map can be represented as $H \in \mathbb{R}^{n\times n}$, where each cell $h_{i,j} \geq 0$ is the normalized interaction frequency between bin i and j.

Overview of SuperTLD

SuperTLD, which includes data imputation and hierarchical domain detection, is proposed for inferring TLD structures from the RAI dataset. The RAI interaction map is first normalized to achieve a uniform read coverage across bins for all the genes, as shown in Figure 1A. Then we assume the normalized interaction frequency follows a negative binomial distribution, whose parameters are estimated with maximum log-likelihood by matrix factorization-based stochastic gradient descent. Then the posterior mean is derived as the imputed RAI. Next, the hierarchical domain detection designed for the incomplete graph is applied to the imputed RAI interaction maps for TLD hierarchy inference (Figure 1B).

Impute the missing interaction frequencies via negative binomial model. A drawback of the RAI capture technology is that low-abundance RNAs may escape from the detection (1), resulting in the sparseness and uneven coverage of the RAI interaction maps. Therefore, we adopted the negative binomial random variation similar to SAVER (25) and integrated matrix factorizations to impute RAI data for mitigating the missing interaction frequencies.

Genomic sequencing technologies are often affected by biological and technical biases, producing different read coverage for each genomic region (26), which makes it challenging to identify the true positive signals. Therefore, the read coverage is divided by the mean read coverage to achieve uniform read coverage across bins. As for symmetric interaction maps of the RIC-seq data set, Knight-Ruiz normalization algorithm (23,24) is adopted to equalize the row and column sum while maintaining symmetry.

Let $m_{i,j}$ be the normalized interaction frequency between gene i (or bin i for RIC-seq interaction maps) and bin j in the RAI interaction map. We model $m_{i,j}$ as a negative binomial random variable via the following Poisson–Gamma mixture distribution

$$\begin{aligned} m_{i,j} &\sim \mathrm{Poisson}(\lambda_{\mathrm{i,j}}) \\ \lambda_{i,j} &\sim \mathrm{Gamma}(\alpha_{\mathrm{i,j}}, \beta_{\mathrm{i,j}}), \end{aligned} \qquad (1)$$

where $\lambda_{i,j}$ represents the true interaction frequency, with a gamma prior added to account for its uncertainty (25), $\alpha_{i,j}$ is the shape parameter and $\beta_{i,j}$ is the rate parameter of the gamma prior. Under the shape-rate parametrization, the mean $\mu_{i,j}$ and variance $\sigma^2_{i,j}$ of the gamma-distributed $\lambda_{i,j}$ is

$$\mu_{i,j} = \frac{\alpha_{i,j}}{\beta_{i,j}},\ \sigma^2_{i,j} = \frac{\alpha_{i,j}}{\beta^2_{i,j}}$$

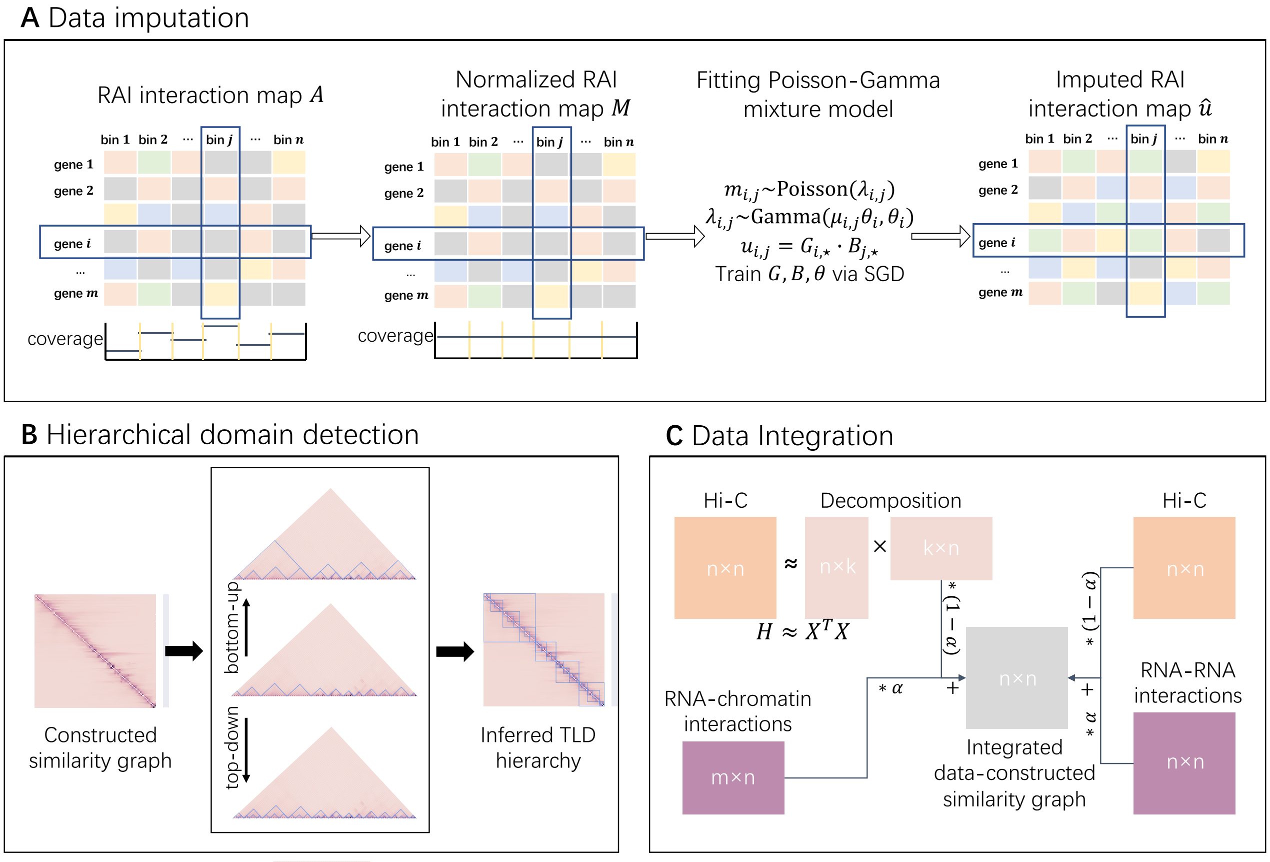
A Data imputation
RAI interaction map A
bin 1
bin 2
···
bin j
···
bin n
gene 1
gene 2
gene i
...
gene m
coverage
Normalized RAI interaction map M
Fitting Poisson-Gamma mixture model
$m_{i,j} \sim \text{Poisson}(\lambda_{i,j})$
$\lambda_{i,j} \sim \text{Gamma}(\mu_{i,j}\theta_i, \theta_i)$
$u_{i,j} = G_{i,\star} \cdot B_{j,\star}$
Train G, B, θ via SGD
Imputed RAI interaction map $\hat{u}$
B Hierarchical domain detection
Constructed similarity graph
bottom-up
top-down
Inferred TLD hierarchy
C Data Integration
Hi-C
Decomposition
n×n
≈
n×k
×
k×n
$H \approx X^T X$
$* (1-\alpha)$
$* \alpha$
+
RNA-chromatin interactions
m×n
Integrated data-constructed similarity graph
RNA-RNA interactions

Figure 1. SuperTLD method. (**A**) Data imputation. Given the RNA–chromatin interaction map, we first perform normalization to equalize the read coverage across bins. Then we fit a Poisson-Gamma mixture model with a mean-variance linear dependency of genes. To derive the posterior gamma distribution, we propose two latent factor matrices G, B for genes and bins that decompose μ and then alternately train them to maximize log-likelihood via stochastic gradient descent. Then the posterior mean $\hat{\mu} = \hat{G}\hat{B}^T$ is derived as the imputed value using the learned $\hat{G}$, $\hat{B}$. (**B**) To infer the TLD hierarchy from the imputed RAI interaction map, we construct the similarity graph of bins and employ dynamic programming to find a partition with minimum structural entropy. The algorithm can be iteratively applied to increase the hierarchy layers with bottom-up and top-down approaches. While the top-down approach is by independently applying the dynamic programming to each partition, the bottom-up approach models each partition as a vertex and form a new similarity graph where the dynamic programming is applied. (**C**) This flow chart shows the data integration for RNA–chromatin interactions and RNA–RNA interactions with the corresponding Hi-C. For the RNA–chromatin interaction map (the left part), the Hi-C contact map is first decomposed to the k-feature representation of bins X. Next, X is concatenated with the imputed RNA–chromatin interactions $\hat{\mu}$ along with the bins through a scaling factor α. Then the similarity graph is constructed from this concatenated matrix. The integrated data is the weighted sum between the RIC-seq interaction map and the Hi-C contact map as they share the same bin resolution (the right part).

To address the mean-variance dependency for genes in RAI data, we assume a Poisson-like distribution of a gene such that the variance scales linearly with the mean (constant Fano factor) (25). The constant Fano factor F can be expressed as $F_i = \frac{\sigma^2_{i,j}}{\mu_{i,j}} = \frac{1}{\theta_i}$ for RNA–chromatin interaction map, and $F_{i,j} = \frac{\sigma^2_{i,j}}{\mu_{i,j}} = \frac{1}{\sqrt{\theta_i\theta_j}}$ for RNA–RNA interaction map. The gamma prior changes to $\lambda_{i,j} \sim \text{Gamma}(\mu_{\text{i,j}}\theta_\text{i}, \theta_\text{i})$ and $\lambda_{i,j} \sim \text{Gamma}(\mu_{\text{i,j}}\sqrt{\theta_\text{i}\theta_\text{j}}, \sqrt{\theta_\text{i}\theta_\text{j}})$, respectively.

Our goal is to derive the posterior gamma distribution of $\lambda_{i,j}$ given the observed interaction frequency $m_{i,j}$, so as to use posterior mean $\hat{\mu}_{i,j}$ as the imputed value. Inspired by SMURF(27), we maximize the log-likelihood by matrix factorization-based stochastic gradient descent (SGD) algorithm. First we propose two latent factor matrices, $G \in \mathbb{R}^{m\times k}$ for genes and $B \in \mathbb{R}^{n\times k}$ for bins, where k is the number of latent factors. We assume $\mu_{i,j} = G_{i,\star} \cdot B_{j,\star}$ for RNA–chromatin interaction map, and $\mu_{i,j} = B_{i,\star} \cdot B_{j,\star}$ for RNA–RNA interaction map, respectively. Then, we initialize the latent factor matrix via applying singular value decomposition on M. Next we update the latent factor matrices to maximize the log-likelihood via SGD. The objective function for RNA–chromatin interaction map and RNA–RNA interaction map are defined as Equations (2) and (3), respectively.

$$\begin{aligned}\min_{G,B,\theta} \sum_{(i,j)\in L} & -\theta_i g_i \cdot b_j \log\theta_i + \log\Gamma(\theta_i g_i \cdot b_j) \\ & - \log\Gamma(m_{i,j} + \theta_i g_i \cdot b_j) + (m_{i,j} + \theta_i g_i \cdot b_j) \\ & \times \log(1+\theta_i) + \Omega(||g_i||_2^2 + ||b_j||_2^2) \end{aligned} \tag{2}$$

$$\begin{aligned}\min_{B,\theta} \sum_{(i,j)\in L, i\leq j} & -\sqrt{\theta_i\theta_j} b_i \cdot b_j \log\sqrt{\theta_i\theta_j} \\ & + \log\Gamma(\sqrt{\theta_i\theta_j} b_i \cdot b_j) - \log\Gamma(m_{i,j} + \sqrt{\theta_i\theta_j} b_i \cdot b_j) \\ & + (m_{i,j} + \sqrt{\theta_i\theta_j} b_i \cdot b_j)\log(1 + \sqrt{\theta_i\theta_j}) \\ & + \Omega(||b_i||_2^2 + ||b_j||_2^2) \end{aligned} \quad (3)$$

where L is the set of $m_{i,j} > 0$, $g_i = G_{i,\star} \in \mathbb{R}^k$ and $b_i = B_{i,\star} \in \mathbb{R}^k$. Ω is the hyper-parameter controlling the L2 penalty to avoid over-fitting. G, B, θ are alternately updated via SGD until the loss converges (see Supplementary Note 1). Finally, we derive the imputed interaction frequency $\hat{\mu} = \hat{G}\hat{B}^T$ or $\hat{\mu} = \hat{B}\hat{B}^T$ with the learned latent factor matrices $\hat{G}$, $\hat{B}$.

Infer TLD hierarchy from RAI interaction map. Given an imputed RAI interaction map of a chromosome, we next construct the similarity graph. We model each bin as a vertex and the interaction frequency as edge weights. For the imputed RNA–chromatin interaction map, we model the edge weight as the inner product between two vectors of the corresponding bins. The adjacency matrix of the similarity graph is further normalized by shrinking the magnitude of its singular values.

Finding the TLDs is equivalent to forming a hierarchy of the vertices. We adapt the tool SuperTAD (28), which is designed for Hi-C contact maps to construct the TAD hierarchy. SuperTAD employs dynamic programming to find an optimal coding tree, the hierarchy, with the objective of minimizing graph structural entropy. The graph structural entropy measures the uncertainty embedded in the partitioned graph. The similarity graph may underestimate the actual structural entropy (29). To address this, besides the negative binomial model-based interaction recovery aforementioned, we adapt SuperTAD to the similarity graph by incorporating a Bayesian correction during dynamic programming and iteratively inferring hierarchy layer by layer.

We briefly review the definition of structural information (entropy) as follows. Denote the similarity graph as $G = (V, E)$. A *coding treeT* of G forms a hierarchical partitioning of vertices. Each node of T contains a subset of vertices while the root λ_T represents all vertices. We term the vertex subset coded by a node $u \in T$ as $S_T(u)$. The parent of a node u is termed as $p_T(u)$. For each tree node $u \in T$, $u \neq \lambda_T$, the structural entropy of u is defined as

$$H_T(G; u) = -\frac{g(u)}{\mathcal{V}(G)} \log_2 \frac{\mathcal{V}(u)}{\mathcal{V}(p_T(u))}$$

where $g(u)$ is the weight sum of edges between vertices in and not in $S_T(u)$. $\mathcal{V}(u)$ is the volume of vertices in $S_T(u)$, i.e., the sum of vertices' degree. $\mathcal{V}(G)$ is the the volume of the graph G. If a node u is leaf that encodes a vertex, then $g(u) = \mathcal{V}(u)$.

Denote the vertices as $(v_1, v_2, ..., v_n)$, which are ordered bins according to the chromosome. Let $S(n, k)$ be the structural entropy of the optimal coding tree that partitions n vertices into k disjoint subsets. For each state in S, we traverse every vertex $v_i \in \{v_{k-1}, v_k, ..., v_{n-1}\}$ listed before v_n to minimize the sum of structural entropy $S(i, k-1) + H(G; u)$ where u is the node encoding $\{v_{i+1}, v_{i+2}, ..., v_n\}$ and $H(G; u)$ is the structural entropy of the subtree with root u. The recurrent relation of dynamic programming is as follows:

$$\begin{aligned} S(n,k) &= \min_{k-1\leq i<n} \{S(i, k-1) + H(G; u)\} \\ &= \min_{k-1\leq i<n} \{S(i, k-1) + H_T(G; u) \\ &\quad + \sum_{i+1\leq j\leq n} H_T(G; v_j)\} \end{aligned} \quad (4)$$

The similarity graphs can be sparse, to avoid the underestimation of structural entropy, we assume a uniform connected graph prior and incorporate the Bayesian approach to estimate the actual values of $g(u)$, $\mathcal{V}(u)$ (see Supplementary Note 2). We use the Bayesian estimated parameters to compute each state in $S(n, k)$. The optimal partitioning can be found in time $O(n^3)$ through dynamic programming with the recurrence in Equation (4). The algorithm determines the optimal k with minimal structural entropy $k_{opt} = \arg\min_{k\in[1,n]} S(n, k)$ by enumerating all the possible k values.

The dynamic programming is iteratively applied to form the hierarchy, as shown in Figure 1B. we adopt a bottom-up approach, which models each partition as a vertex and form a new similarity graph for the k vertices. Then the dynamic programming approach is applied to the new similarity graph. In this work, a two-layer TLD hierarchy is adopted.

Assessment of the structural similarity between TLDs and TADs

To assess the structural property of TLDs, we compare TLDs with TADs that calculate the correlation of distance decay patterns between two similarity graphs and measure the similarity of domains and boundaries.

One feature in the Hi-C contact map is distance decay; that is, the contact frequency of two bins decreases as their genomic distance increases (30,31). As RAIs are captured through spatial adjacency as Hi-C, the RAI-constructed similarity graph of bins also shows a distance decay pattern. The farther two bins are, the less they interact. We calculate the average distribution of contact frequency over distance for each similarity graph as its distance decay distribution. We compute a Pearson correlation coefficient between two distance decay distributions to depict the similarity between the RAI-constructed similarity matrix and Hi-C contact map.

To assess the similarity between TLD and TAD partitions, we adopt normalized mutual information (NMI) and overlapping ratio (OR) (28). NMI aims at measuring the similarity of two disjoint partitions, while the OR measures the similarity between two hierarchical structures. We apply NMI to nodes with height 1 (nodes that are the parent of leaves) in the inferred coding tree to evaluate the agreement of TLD and TAD at the finest level. The value of NMI ranges from 0 to 1 as zero indicates no mutual information while one indicates perfect correlation. OR is a symmetric metric, ranging from 0 to 1. One indicates TAD hierarchy and TLD hierarchy are the same, while zero indicates the two hierarchies contain no intersection between any pair of domains.

To explore whether the TLDs employ merely a subset of TAD boundaries or have novel boundaries, we roughly define four types of relationships that are termed as matched, merged, split, and shifted (32,33). A matched domain is defined as both TLD boundaries lying within one bin (align) of two boundaries of a TAD. A merged or split domain represents that one boundary aligns with a TAD and the other aligns inside another TAD or identical TAD. A shifted domain is defined as both boundaries lying within five bins but not aligning with a TAD's two boundaries. Split and shifted domains generate novel TLD boundaries among the four relationships.

Assessment of epigenetic characteristics enrichment of TLDs

Evidence has demonstrated that TADs play a vital role in gene regulation (16,18,19). The boundary of TADs is enriched in insulating features that interrupt the interactions across two TADs (20,22). To assess the enrichment of transcriptional repressors at TLD boundaries, we downloaded the CTCF TF ChIP-seq peaks from ENCODE (https://www.encodeproject.org/) for four cell lines. The ChIP-seq peaks are summed into 5-kb intervals around boundaries. Then we calculate intervals' average peak number from two regions. One is the region at the TAD boundaries (the bin detected as a boundary, referred to as *peak*), the other is the 50kb region located 500kb away from the boundaries at both sides (referred to as *background*). We compute the fold change of average peak number between *peak* and *background* per boundary. We take the average across boundaries minus one. The zero value of the average fold change stands for no enrichment around boundaries.

Published works have validated that TADs are frequently enriched in repressing/activating histone marks H3K27me3/H3K36me3 (34,35). To assess the enrichment of two histone modifications within TLDs, H3K27me3 (repressing) and H3K36me3 (activating), we downloaded the Histone ChIP-seq data from ENCODE (https://www.encodeproject.org/). Unlike CTCF binding that produced sharp and narrow peaks, histone marks spread over more extended regions. The histone marks are better modeled by increasing or decreasing their level rather than by single discrete peaks. Therefore, the fold change over control (in bigWig format) for histone marks are downloaded for three cell lines, HEK293, HeLa, and mESC.

The ChIP-seq signals are summed into intervals with a fixed 10% of the resolution. Next, we calculate the $\log_{10}$-ratio between H3K27me3 and H3K36me3 for each interval (LR value) and take the average LR values of intervals within each domain as the observed LR values. Then the empirical p-value for each domain is calculated via shuffling intervals 1000 times and corrected through false discovery rate with Benjamini–Hochberg method. With the constraint that corrected P-value <0.1, we report the fraction of TLDs that significantly enriched in either H3K27me3 or H3K36me3. A higher faction reflects that the inferred hierarchy has more Histone H3 modifications enriched.

Integration of RAI data with Hi-C data

As the TLDs inferred from RAI data exhibit TADs' characteristics, we integrate each RAI data set with the corresponding Hi-C to enhance the inferred domains' performance in the evaluation analysis.

To integrate asymmetric RNA–chromatin interactions with Hi-C (as shown in Figure 1C, the left part), we first apply the eigen-decomposition on the Hi-C contact map H to derive bin's k-feature vector X such that $H = X^T X$. Then X is concatenated with imputed RNA–chromatin interaction map $\hat{\mu}$ along with the bins, weighted by a factor $\alpha \in [0, 1]$.

To integrate the RNA–RNA interactions with Hi-C (as shown in Figure 1 C, the right part), the imputed interaction map and Hi-C contact map are weighted summed through a scaling factor α, as they share the same bin resolution.

RESULTS

SuperTLD infers hierarchical TLDs from RAI data sets

We developed SuperTLD, a hierarchical domain detection method incorporating imputation, using RAIs to infer hierarchical TAD-like domains. Given any RAI interaction map, SuperTLD applies matrix factorization-based stochastic gradient descent to impute the missing interaction frequency via a negative binomial model. In TLD hierarchy detection, SuperTLD adopts dynamic programming to find a partition with minimum structural entropy, where Bayesian correction is incorporated to address the underestimation of structural information.

This work collected four RAI data sets and performed SuperTLD on each chromosome's interaction map from each data set with the default parameter value. SuperTLD takes 18.5 minutes on average to run on all chromosomes of a human cell line (see Supplementary Note 3: Supplementary Table S1) and supports genome-wide multiprocess computing.

To explore the structural and functional properties of the inferred TLDs, we assessed the structural similarity between TLDs and TADs and evaluated their epigenetic characteristics enrichment using CTCF TF ChIP-seq and Histone H3 modification ChIP-seq data.

TLDs exhibit similar structures but high boundary variation compared to TADs

To explore the structural properties of TLDs, we compared TLDs with TADs and assessed the similarity between the two hierarchies. TAD hierarchy is detected from the normalized Hi-C contact maps via our domain detection method. We adopted five metrics to quantify the similarity between TLDs and TADs: Pearson correlation coefficient (PCC) of distance decay, overlapping ratio (OR), normalized mutual information (NMI), common ratio, and the percentage of various domain relationships. To evaluate the hierarchy inference potential of RAIs, we calculated the PCC between the distance decay distribution of the RAI-constructed similarity graph and the Hi-C contact map. Then, we applied OR and NMI to measure the similarity of TLDs and TADs on the aspect of hierarchical structure and the finest structure, respectively. Next, we identified the common boundaries and calculated the common ratio as the percentage of common boundaries in TLD boundaries. Last, we defined four types of domain relationships, namely, match, merge, split, and shift. We quantified the percentage of each type of domain in TLDs. We selected one chromosome from each RAI dataset to illustrate the result.

First, the RAI-constructed similarity matrices exhibited a high correlation of distance decay with the Hi-C contact map. The distance decay pattern characterizes the chromosome polymer as distribution, inherently accounting for randomness and structural variability (30). A high correlation of distance decay with the Hi-C contact map reveals the RAIs' potential for TLD inference. We drew the distribution of distance decay for the RAI-constructed similarity matrix and Hi-C contact map and normalized the curve by dividing the maximum of the distribution to make the comparison more intuitive. As shown in Figure 2A, the normalized curves rapidly decay within the closest 20 bins and then flatten out. The curves of GRID-seq and RIC-seq decay much faster and converge within 10 bins. The similarity matrices of all data sets achieve a significant PCC over 0.95 with the Hi-C contact map.

Second, the inferred TLDs show a moderate structural similarity with TADs. To quantify the agreement of structures between TLDs and TADs, we adopted OR on two hierarchies and NMI on the disjoint partitions at the finest level. As shown in Table 1, the RADICL-seq inferred TLDs rank the top under both OR and NMI, while the GRID-seq inferred TLDs have the smallest OR. Note that iMARGI is excluded in the similarity comparison across data sets, as different isolates of HEK293T exhibit inconsistent DNA copy numbers leading to possible intrinsic structure differences between TLDs and TADs (see Supplementary Note 4: Supplementary Figures S1 and S2). We drew the heatmap of each RAI-constructed similarity matrix and Hi-C contact map and annotated the TLDs and TADs through blocks. As shown in Figure 2 B, RADICL-seq and RIC-seq inferred TLDs exhibit a similar structure and share more common boundaries (the red circles placed along the diagonal) with TADs. iMARGI infers few TLDs on the short arm of chr10, leading to a relatively low similarity at the finest level (NMI: 0.8252) with TADs. GRID-seq fails to infer the higher-level TLDs of chr2R and ranks the last under OR measurement among all RAI data sets.

Third, TLDs possess a large percentage of novel boundaries. Considering that regions within TLD are enriched for RNA-related DNA-DNA interactions, the boundaries of TLDs may act as a subset of TAD boundaries. To test the novelty of RAI inferred TLDs, we calculated the common ratio as the percentage of common boundaries in TLD boundaries, where the common boundary is the bin that both TLD and TAD use as a boundary. As shown in Table 1, RADICL-seq inferred TLDs have the highest common ratio of 0.2467 among RAI data sets, showing a high percentage of novel boundaries. To explore the relationship between TLDs and TADs, we defined the matched domain, merged domain, split domain, and shifted domain. Note that merge or match do not generate novel TLD boundaries while split and shift do. A TLD that cannot be classified into any type is a novel domain. We counted the domains in each relationship and calculated their percentage in TLDs. As shown in Table 1, more TLDs are merged or split from TADs than matching or shifting. Additionally, a large percentage of RAI inferred TLDs are novel domains.

Altogether, our analysis indicates that RAI inferred TLDs show a moderate similarity of structure but a high variation of boundaries and domains compared to TADs.

Novel TLDs show enriched epigenetic characteristics

The TAD boundaries act as insulators to obstruct the spread of transcriptional activities, such that the promoter preferentially interacts with enhancers within the same TAD (20,36). We evaluated the enrichment of epigenetic characteristics for TLDs and compared the results with TADs.

To assess the enrichment of transcriptional repressor CTCF at inferred TLD boundaries, we computed the fold change of CTCF peaks for every boundary and took the average as the result of the TLD hierarchy. We also calculated the *P*-value of CTCF fold change to measure the significance. As shown in Table 2, iMARGI inferred TLD boundaries exhibit the largest CTCF fold change of 0.7386 (*P*-value of 0.0198). GRID-seq inferred TLD boundaries achieve a larger CTCF fold change than TADs among the four RAI data sets.

Next, we evaluated the enrichment of histone marks by calculating the fraction of TLDs that significantly enriched in H3K27me3/H3K36me3. A higher fraction indicates that more inferred TLDs in the hierarchical coding tree are biologically meaningful. We also compared the result with TAD hierarchies. As shown in Table 2, iMARGI inferred TLDs achieve the highest ratio as 80.64% among data sets, also higher than TADs (75.51%). RIC-seq inferred TLDs show a low CTCF fold change (0.1022) at the boundaries, but significant histone marks enrichment within most domains (77.10%). This finding drives us to explore the heterogeneity within TLD boundaries further.

Except for the common boundaries shared with TADs, we call the rest TLDs as unique boundaries or novel boundaries. The CTCF enrichment analysis is applied to the three boundary groups for each data set. These are Hi-C unique boundaries, RAI unique boundaries, and common boundaries. As shown in Figure 3, the common boundaries between TADs and TLDs for each cell line exhibit significant CTCF enrichment, revealing their strong insulation strength. The iMARGI and GRID-seq inferred TLD unique boundaries also show a relatively high CTCF enrichment, demonstrating that RAIs can infer novel biologically relevant boundaries.

Last, GO analysis is performed on TLDs' common boundaries and novel boundaries. As shown in Figure 4, most of the genes on the common boundaries are enriched for terms related to cytoplasm or cytosol, while genes on the novel boundaries of different TLD hierarchies are enriched for different terms. This result further proved the novel TLDs' functional validity.

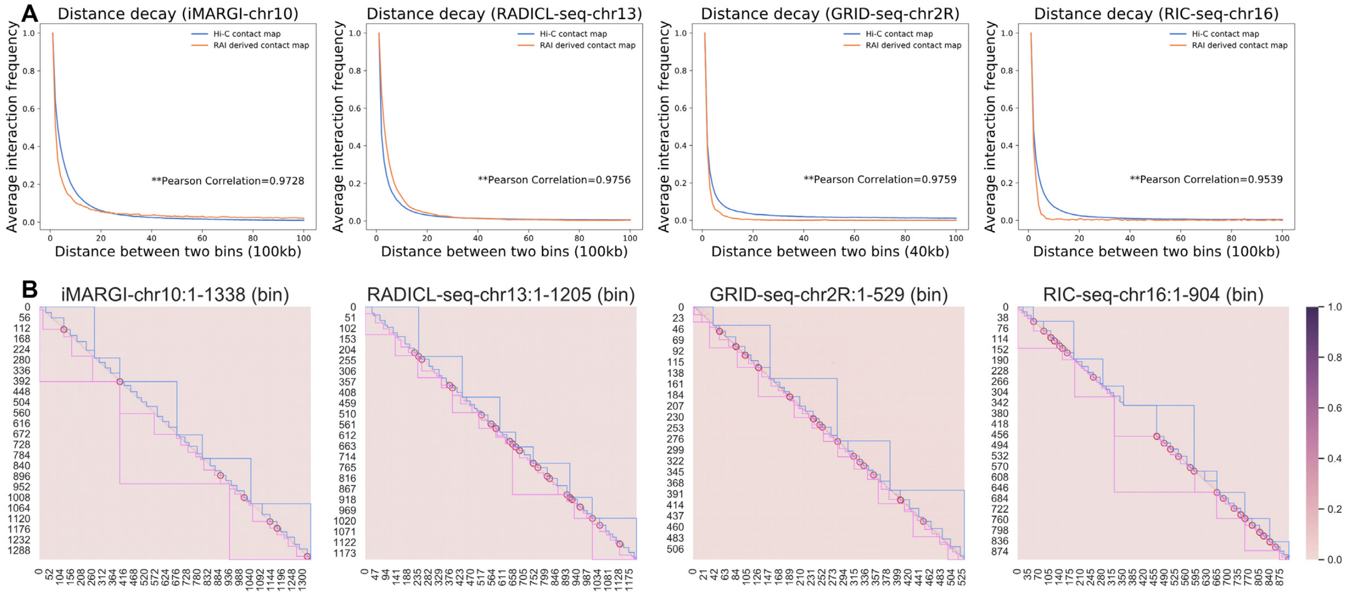

Figure 2. Analysis of structural similarity between TLDs and TADs. (**A**) The normalized distance decay curve from each RAI-constructed similarity graph as well as the Hi-C contact map. We calculate the distance decay as the average interaction frequency (vertical axis) for a pair of bins separated by a given distance (horizontal axis). The text on the bottom right shows the PCC between two distributions, and '**' stands for statistically extreme significant *p*-value<0.01. The plot shows the distribution in the closet 100 bins (10 Mb for iMARGI, RADICL-seq, RIC-seq; 4Mb for GRID-seq). (**B**) The heatmap of the Hi-C contact map (upper triangle) and the RAI-constructed similarity matrix (lower triangle). The boundaries of the inferred hierarchies are drawn using different colors in the upper triangle (TADs in blue; TLDs in violet). The red circles along the diagonal annotated the common boundaries shared by TADs and TLDs. The four heatmaps are normalized between 0-1 and share the rightmost color bar.

Table 1. The structural similarity between TLDs and TADs

TLDs	OR	NMI	Common ratio	Matched[a]	Merged	Split	Shifted
iMARGI-chr10	0.6252	0.8252	0.2321	2(6.45%)	12(38.71%)	3(9.68%)	0
RADICL-seq-chr13	0.6905	0.8818	0.2467	7(7.95%)	22(25%)	23(26.14%)	3(3.41%)
GRID-seq-chr2R	0.5312	0.8507	0.1678	2(2.22%)	9(10%)	27(30%)	6(6.67%)
RIC-seq-chr16	0.6687	0.8542	0.1548	3(2.29%)	8(6.11%)	47(35.88%)	7(5.34%)

[a]The integer in front of the brackets indicates the number of matched domains, and the percentage indicates its proportion in TLDs. The columns 'merged', 'split', 'shifted' are the same.

Table 2. Epigenetic characteristics enrichment of TLDs and TADs

TLDs	CTCF fold change	CTCF fold change (TADs)[a]	CTCF *P*-value	CTCF *P*-value (TADs)	H3K* ratio[b]	H3K* ratio (TADs)
iMARGI-chr10	0.7386	1.1696	0.0198	0	80.64%	75.51%
RADICL-seq-chr13	0.2487	0.9017	0.0278	0	60.23%	72.46%
GRID-seq-chr2R	0.2008	0.1269	0.1826	0.3652	N/A[c]	N/A
RIC-seq-chr16	0.1022	0.2133	0.1783	0.2398	77.10%	81.82%

[a]The column name suffixed with '(TAD)' represents the analysis result of TADs. The rest columns are the same.
[b]'H3K* ratio' denotes the fraction of domains that significantly enriched in H3K27me3/H3K36me3 histone marks.
[c]The result is not applicable due to a lack of data.

Integration of the RAIs with Hi-C improves the TAD hierarchy inference

As the TLDs exhibit similar structural and functional properties as TADs, we integrated RAIs with Hi-C to see if data integration can enhance the performance of inferring TAD hierarchy, i.e., inferring more and functionally enriched domains. As shown in Figure 1C, the integration is conducted after the imputation of RAIs. We introduced a scaling factor $\alpha \in [0, 1]$ for data integration, where $\alpha = 1$ denotes no Hi-C information and $\alpha = 0$ denotes no RAI information. Then the integrated interaction map is used to conduct the hierarchical domain detection.

We tested the α value from 0 to 1 by 0.05 on each RAI data set (see Supplementary Note 5: Supplementary Tables S2–S5). With the increasing proportion of Hi-C data, the integrated data of iMARGI and RADICL-seq infers a first increasing and then decreasing number of domains, while GRID-seq and RIC-seq show a decreasing pattern. For the RIC-seq data set, $\alpha = 0.95$ vastly reduces the number of inferred domains from 131 ($\alpha = 1.0$) to 57 and enhances the epigenetic characteristics enrichment (CTCF

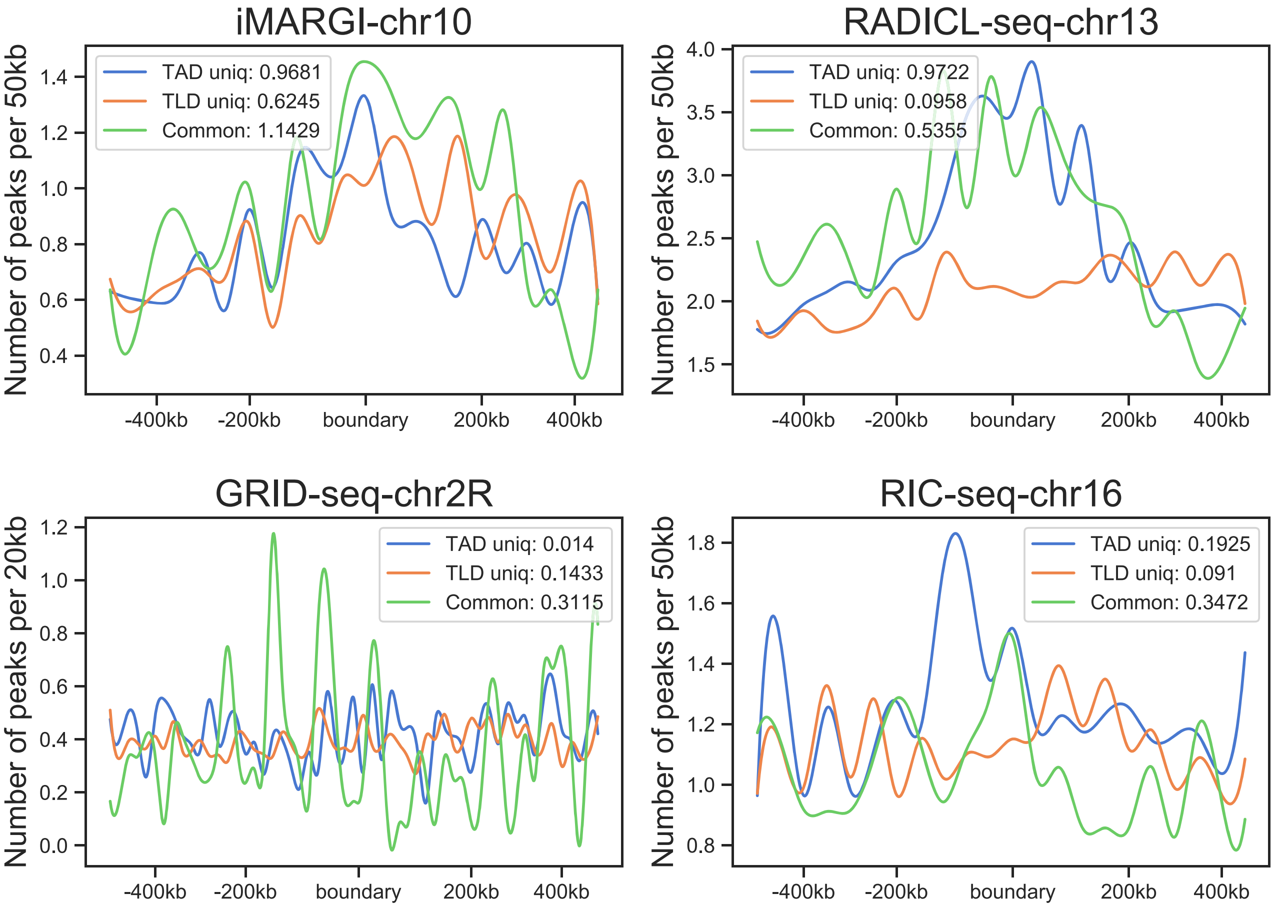

Figure 3. Analysis of CTCF enrichment at TAD unique boundaries, TLD unique boundaries, and common boundaries. The plot shows the distribution of the average number of CTCF ChIP-seq peaks around the boundaries (±500 kb). We identified the common boundaries and divided the rest into unique boundaries of TADs and TLDs. The legend indicates the assigned color and the average CTCF fold change for each group of boundaries. We use half of the bin size as the profile resolution to smooth the curve.

fold change: from 1.0225 to 1.8535; H3K* ratio: from 77.10% to 83.93%), revealing that data integration helps to eliminate false-positive domains inferred from RIC-seq data set. Note that the structural similarity of inferred domains with TADs may not linearly improve with the increase of Hi-C proportion.

The epigenetic characteristics enrichment analysis supports the integrated data's effectiveness in TAD inference. Domains inferred from the integrated data show enhancement in the boundaries' CTCF fold change and the H3K27me3/H3K36m3 enrichment within domains. Among the four data sets, GRID-seq integrated data achieves the largest increase in CTCF fold change by 0.1919, and iMARGI integrated data achieves the largest increase in the percentage of histone H3 mark-enriched domains by 3.13%. The results of iMARGI, GRID-seq, and RIC-seq demonstrate that the integrated data performs better than the sole data set, except iRADICL-seq.

DISCUSSION

In this work, we proposed a novel method SuperTLD that uses RAIs to infer the TLD hierarchy structure, which comprises the data imputation and hierarchical domain detection. SuperTLD supports asymmetric or symmetric interaction maps of RAIs as input. By assessing the four public RAI data sets from human, mouse, and Drosophila melanogaster, we found that SuperTLD works well in the analysis of RAIs, and the inferred TLDs exhibit significant epigenetic characteristics enrichment. To explore the structural properties of TLDs, we adopted five evaluation metrics to assess their similarity with TADs. We adopted CF ChIP-seq and histone ChIP-seq for TLDs' epigenetic characteristics enrichment analysis.

The gene amount and RNAs follow a non-uniform distribution along the genome. Unlike Hi-C, RAI data sets require a proper decision on the bin size when performing SuperTLD. A small bin size renders the interaction map too sparse to infer TADs. A large bin size results in a rough interaction map where only megabase-size domains can be observed.

We roughly defined four domain relationships between TLDs and TADs and classify each TLD. A large fraction of TLDs are merged or split from TADs, however, the reasons for this discrepancy remain unknown.

We have shown that functional evaluation of TLDs inferred from RAI data sets reveals the enriched epigenetic characteristics at TLD boundaries or within domains,

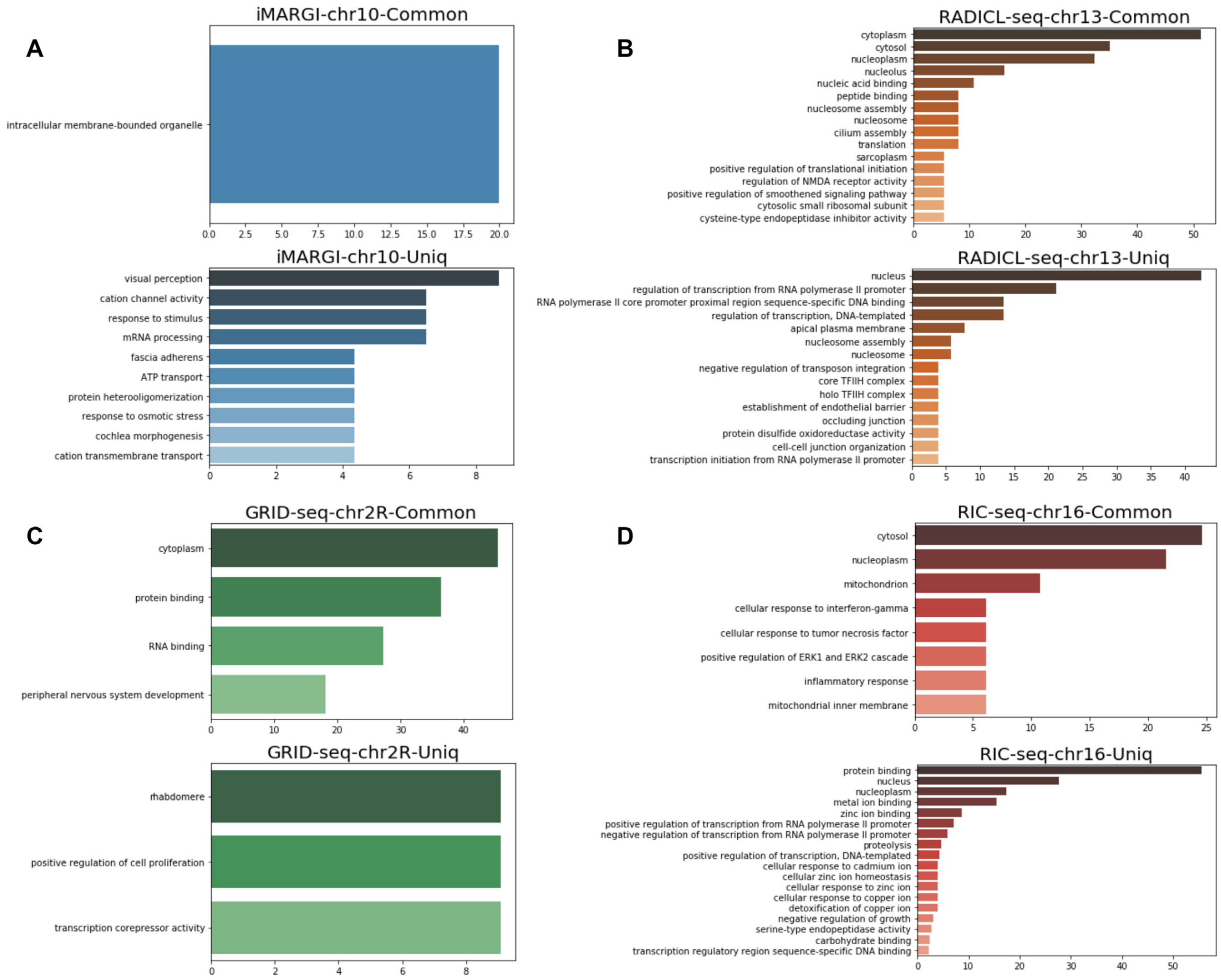

Figure 4. GO analysis of genes on the common boundaries between TADs and TLDs and TLD novel boundaries for (**A**) iMARGI, (**B**) RADICL-seq, (**C**) GRID-seq and (**D**) RIC-seq. The x-axis of each barplot represents the percentage of enriched genes for a given term. The terms are sorted by the descending order of percentage for each barplot. When drawing for RIC-seq inferred TLDs, we set a threshold of 5% for common boundaries and 2% for novel boundaries; the terms with a lower percentage are removed from the plot.

though the enrichment varies across RAI data sets. For example, the TLD boundaries inferred from iMARGI exhibit the most considerable CTCF fold change of 0.7386, while those from RIC-seq show the most minor CTCF fold change of 0.1022. Removing the common boundaries with TADs, iMARGI inferred novel TLD boundaries exhibit the largest CTCF fold change of 0.6245, while those from RIC-seq show the smallest CTCF fold change of 0.091 (only a slight enrichment of CTCF at boundaries), revealing the advantages of iMARGI in TLD inference. With more data available in the future, we can perform more detailed comparisons between RAI data sets spanning the difference in species, cell lines, and chromosomes.

In the last section, we proposed integrating RAIs and Hi-C to infer more domains with functional enrichment. The integrated data infers new domains exhibiting strong CTCF enrichment at the boundaries. Our evaluation of the data integration proved the enhanced performance of TAD hierarchy inference. We suggest integrating more interaction sources to dissect the TAD structures and functions, such as protein–RNA interaction, protein–DNA interaction, etc.

CONCLUSIONS

We proposed a novel method SuperTLD, comprising interaction imputation and hierarchical domain detection, to infer the hierarchical TLDs from RAIs. We collected four RAI data sets and applied SuperTLD on each chromosome. The comparison experiments demonstrate that TLDs share a moderate structural similarity with TADs but vary on boundaries. The novel TLD boundaries are enriched for epigenetic characteristics. GO analysis reveals the clear difference between the shared boundaries with Hi-C and novel boundaries. Moreover, we integrated RAIs data with Hi-C and found the superior of the inferred domain in the epigenetic characteristics enrichment, revealing the effectiveness of multiple data sources integration in TAD inference.

DATA AVAILABILITY

All the data used in this paper can be retrieved from public databases. All the experiments are reproducible with the dedicated version of the software with default arguments. SuperTLD is an open-source Python package with source code freely available at https://github.com/deepomicslab/SuperTLD.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Wei Wu for constructive criticism of the manuscript.

FUNDING

Strategy Research Grant [7005215]. Funding for open access charge: Strategic Interdisciplinary Research Grant [7005215].

Conflict of interest statement. None declared.

REFERENCES

1. Li,X. and Fu,X.-D. (2019) Chromatin-associated RNAs as facilitators of functional genomic interactions. *Nat. Rev. Genet.*, **20**, 503–519.
2. Chen,W., Yan,Z., Li,S., Huang,N., Huang,X., Zhang,J. and Zhong,S. (2018) RNAs as proximity-labeling media for identifying nuclear speckle positions relative to the genome. *Iscience*, **4**, 204–215.
3. Kuo,C.-C., Hänzelmann,S., Sentürk Cetin,N., Frank,S., Zajzon,B., Derks,J.-P., Akhade,V.S., Ahuja,G., Kanduri,C., Grummt,I. *et al.* (2019) Detection of RNA–DNA binding sites in long noncoding RNAs. *Nucleic Acids Res.*, **47**, e32.
4. Cai,Z., Cao,C., Ji,L., Ye,R., Wang,D., Xia,C., Wang,S., Du,Z., Hu,N., Yu,X. *et al.* (2020) RIC-seq for global in situ profiling of RNA–RNA spatial interactions. *Nature*, **582**, 432–437.
5. Sentürk Cetin,N., Kuo,C.-C., Ribarska,T., Li,R., Costa,I.G. and Grummt,I. (2019) Isolation and genome-wide characterization of cellular DNA: RNA triplex structures. *Nucleic Acids Res.*, **47**, 2306–2321.
6. Guh,C.-Y., Hsieh,Y.-H. and Chu,H.-P. (2020) Functions and properties of nuclear lncRNAsâfrom systematically mapping the interactomes of lncRNAs. *J. Biom. Sci.*, **27**, 44.
7. Jones,L., Hamilton,A.J., Voinnet,O., Thomas,C.L., Maule,A.J. and Baulcombe,D.C. (1999) RNA–DNA interactions and DNA methylation in post-transcriptional gene silencing. *Plant Cell*, **11**, 2291–2301.
8. Li,X., Zhou,B., Chen,L., Gou,L.-T., Li,H. and Fu,X.-D. (2017) GRID-seq reveals the global RNA–chromatin interactome. *Nat. Biotechnol.*, **35**, 940.
9. Yan,Z., Huang,N., Wu,W., Chen,W., Jiang,Y., Chen,J., Huang,X., Wen,X., Xu,J., Jin,Q. *et al.* (2019) Genome-wide colocalization of RNA–DNA interactions and fusion RNA pairs. *Proc. Nat. Acad. Sci. U.S.A.*, **116**, 3328–3337.
10. Sridhar,B., Rivas-Astroza,M., Nguyen,T.C., Chen,W., Yan,Z., Cao,X., Hebert,L. and Zhong,S. (2017) Systematic mapping of RNA–chromatin interactions in vivo. *Curr. Biol.*, **27**, 602–609.
11. Bell,J.C., Jukam,D., Teran,N.A., Risca,V.I., Smith,O.K., Johnson,W.L., Skotheim,J.M., Greenleaf,W.J. and Straight,A.F. (2018) Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts. *Elife*, **7**, e27024.
12. Wu,W., Yan,Z., Nguyen,T.C., Chen,Z.B., Chien,S. and Zhong,S. (2019) Mapping RNA–chromatin interactions by sequencing with iMARGI. *Nat. Protoc.*, **14**, 3243–3272.
13. Bonetti,A., Agostini,F., Suzuki,A.M., Hashimoto,K., Pascarella,G., Gimenez,J., Roos,L., Nash,A.J., Ghilotti,M., Cameron,C.J. *et al.* (2020) RADICL-seq identifies general and cell type–specific principles of genome-wide RNA–chromatin interactions. *Nat. Commun.*, **11**, 1018.
14. Quinn,J.J., Ilik,I.A., Qu,K., Georgiev,P., Chu,C., Akhtar,A. and Chang,H.Y. (2014) Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification. *Nat. Biotechnol.*, **32**, 933–940.
15. Morf,J., Wingett,S.W., Farabella,I., Cairns,J., Furlan-Magaril,M., Jimenez-Garcia,L.F., Liu,X., Craig,F.F., Walker,S., Segonds-Pichon,A. *et al.* (2019) RNA proximity sequencing reveals the spatial organization of the transcriptome in the nucleus. *Nat. Biotechnol.*, **37**, 793–802.
16. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
17. Eagen,K.P. (2018) Principles of chromosome architecture revealed by Hi-C. *Trends Biochem. Sci.*, **43**, 469–478.
18. Dixon,J.R., Gorkin,D.U. and Ren,B. (2016) Chromatin domains: the unit of chromosome organization. *Mol. Cell*, **62**, 668–680.
19. Lupiáñez,D.G., Spielmann,M. and Mundlos,S. (2016) Breaking TADs: how alterations of chromatin domains result in disease. *Trends Genet.*, **32**, 225–237.
20. Zuin,J., Dixon,J.R., van der Reijden,M.I., Ye,Z., Kolovos,P., Brouwer,R.W., van de Corput,M.P., van de Werken,H.J., Knoch,T.A., van IJcken,W.F. *et al.* (2014) Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Nat. Acad. Sci. U.S.A.*, **111**, 996–1001.
21. Bonev,B., Cohen,N.M., Szabo,Q., Fritsch,L., Papadopoulos,G.L., Lubling,Y., Xu,X., Lv,X., Hugnot,J.-P., Tanay,A. *et al.* (2017) Multiscale 3D genome rewiring during mouse neural development. *Cell*, **171**, 557–572.
22. Sexton,T., Yaffe,E., Kenigsberg,E., Bantignies,F., Leblanc,B., Hoichman,M., Parrinello,H., Tanay,A. and Cavalli,G. (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, **148**, 458–472.
23. Durand,N.C., Shamim,M.S., Machol,I., Rao,S.S., Huntley,M.H., Lander,E.S. and Aiden,E.L. (2016) Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.*, **3**, 95–98.
24. Knight,P.A. and Ruiz,D. (2013) A fast algorithm for matrix balancing. *IMA J. Numer. Anal.*, **33**, 1029–1047.
25. Huang,M., Wang,J., Torre,E., Dueck,H., Shaffer,S., Bonasio,R., Murray,J.I., Raj,A., Li,M. and Zhang,N.R. (2018) SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods*, **15**, 539–542.
26. Imakaev,M., Fudenberg,G., McCord,R.P., Naumova,N., Goloborodko,A., Lajoie,B.R., Dekker,J. and Mirny,L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.
27. Wang,B., Pu,J., Chen,L. and Li,S. (2022) SMURF: embedding single-cell RNA-seq data with matrix factorization preserving selfconsistency. bioRxiv doi: https://doi.org/10.1101/2022.04.22.489140, 22 April 2022, preprint: not peer reviewed.
28. Zhang,Y.W., Wang,M.B. and Li,S.C. (2021) SuperTAD: robust detection of hierarchical topologically associated domains with optimized structural information. *Genome Biol.*, **22**, 45.
29. Smiljanić,J., Edler,D. and Rosvall,M. (2020) Mapping flows on sparse networks with missing links. *Phys. Rev. E*, **102**, 012302.
30. Lajoie,B.R., Dekker,J. and Kaplan,N. (2015) The Hitchhikerâs guide to Hi-C analysis: practical guidelines. *Methods*, **72**, 65–75.
31. Zhang,Y., An,L., Xu,J., Zhang,B., Zheng,W.J., Hu,M., Tang,J. and Yue,F. (2018) Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nat. Commun.*, **9**, 750.
32. Cresswell,K.G. and Dozmorov,M.G. (2020) TADCompare: an R package for differential and temporal analysis of topologically associated domains. *Front. Genet.*, **11**, 158.

33. Li,X., Zeng,G., Li,A. and Zhang,Z. (2021) DeTOKI identifies and characterizes the dynamics of chromatin TAD-like domains in a single cell. *Genome Biol.*, **22**, 217.
34. Rao,S.S., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
35. Zufferey,M., Tavernari,D., Oricchio,E. and Ciriello,G. (2018) Comparison of computational methods for the identification of topologically associating domains. *Genome Biol.*, **19**, 217.
36. van Arensbergen,J., van Steensel,B. and Bussemaker,H.J. (2014) In search of the determinants of enhancer–promoter interaction specificity. *Trends cell Biol.*, **24**, 695–702.