doi: 10.1093/nar/gkx735

HiTAD: detecting the structural and functional hierarchies of topologically associating domains from chromatin interactions

Xiao-Tao Wang, Wang Cui and Cheng Peng*

Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China

Received June 07, 2017; Revised July 19, 2017; Editorial Decision August 09, 2017; Accepted August 11, 2017

ABSTRACT

A current question in the high-order organization of chromatin is whether topologically associating domains (TADs) are distinct from other hierarchical chromatin domains. However, due to the unclear TAD definition in tradition, the structural and functional uniqueness of TAD is not well studied. In this work, we refined TAD definition by further constraining TADs to the optimal separation on global intra-chromosomal interactions. Inspired by this constraint, we developed a novel method, called HiTAD, to detect hierarchical TADs from Hi-C chromatin interactions. HiTAD performs well in domain sensitivity, replicate reproducibility and inter cell-type conservation. With a novel domain-based alignment proposed by us, we defined several types of hierarchical TAD changes which were not systematically studied previously, and subsequently used them to reveal that TADs and sub-TADs differed statistically in correlating chromosomal compartment, replication timing and gene transcription. Finally, our work also has the implication that the refinement of TAD definition could be achieved by only utilizing chromatin interactions, at least in part. HiTAD is freely available online.

INTRODUCTION

Recent years have seen rapid development in exploring high-order organization of chromatin due to the chromosome conformation capture (3C) technique (1) and its derivatives, such as 4C (2,3), 5C (4), ChIA-PET (5), Hi-C (6), TCC (7), Capture Hi-C (8) and *in situ* Hi-C (9), etc. It is now known that chromatin is neatly packed in nucleus, in which topologically associating domain (TAD) is a kind of structural unit in linking chromatin organization and biological functions, at least in drosophila (10) and mammalian genomes (11,12). It was reported that TAD could constrain enhancer-promoter targeting in gene regulation (13,14), shape replication timing (15) and determine pathogenicity of genomic duplications (16). The switch of TAD boundary was observed in mouse limb development (17), and the boundary knockout on mouse model directly proved that the disruption of TAD boundary led to development disease (18). The studies on cancer genomes also revealed that mutations occurred on TAD boundaries could contribute to oncogene activation (19,20), implying the association of TAD disruption with tumorigenesis.

TAD itself is a hierarchical organization which needs to be further clarified. TAD is traditionally defined as a continuous chromatin region in which the loci interact with each other more frequently than the loci outside the region (11,12,21). However, different levels of chromatin domains satisfy this criterion more or less, especially with the improvement of data quality and sequencing depth. By thoroughly investigating specific chromatin regions with 5C, Phillips-Cremins *et al*. found that there existed smaller chromatin domains (called sub-TADs) inside the traditional TADs (22). Further comparisons revealed that TADs were stable across cell types, whereas sub-TADs could vary greatly to facilitate gene regulation. By improving Hi-C experimental pipeline and sequencing depth, Rao *et al*. observed hierarchical overlapping among different chromatin domains in the genome-wide scale (9). A recent work also revealed that TADs exhibited structural heterogeneity and functional diversity in mammalian genomes (23). These phenomena suggest the existence of hierarchical domains in chromatin, which cannot be explained by traditional TADs.

Several methods have been proposed to identify hierarchical chromatin domains from Hi-C chromatin interactions. Rao *et al*. proposed an Arrowhead transformation on bias-corrected chromatin interaction matrix and then used dynamic programming to identify chromatin domains at multiple scales simultaneously (9). TADtree proposed by Weinreb and Rahpael used a weighted interval scheduling with multiplicities to find TAD forest (24). However, the high computational complexity in this algorithm limits its available resolutions. Matryoshka (bioRxiv https://doi.

*To whom correspondence should be addressed. Tel: +86 27 87280877; Email: pengcheng@mail.hzau.edu.cn org/10.1101/032953) proposed by Malik and Patro identified various chromatin domains at different resolutions, and then the consensus hierarchy through domain clustering was used to generate hierarchical chromatin domains. Recently, a network modularity based method was proposed to identify hierarchical chromatin domains by utilizing different resolution parameter values (bioRxiv https://doi.org/10.1101/089011). The method CaTCH identified large levels of hierarchical chromatin domains by using only a single parameter, reciprocal insulation (25). However, it did not point out the TAD positions in the hierarchical domains. Instead, additional data, such as CCCTC-binding factor (CTCF) enrichment, were needed as references to identify the TADs. Similarly, HBM identified hierarchical domains until the given chromatin was merged into a single cluster, without pointing out the TAD positions (26). Finally, some methods, such as Armatus (27), spectral method (28), Mr-TADFinder (bioRxiv https://doi.org/10.1101/097345) and IC-Finder (29), could identify overlapped or hierarchical domains across different parameter values, but they did not automatically reconcile the hierarchy and consensus among these domains.

The aforementioned methods mainly treat TADs as local insulations but neglect their global properties, making it hard to judge where the TADs stand in the hierarchical domains. In this work, except the local insulations, we further constrain TADs to the optimal separation on intra-chromosomal interactions. To facilitate representation, TAD and its smaller chromatin domains are together called hierarchical TAD in this work. Inspired by our TAD constraint, we developed an iterative optimization procedure, called HiTAD, to detect hierarchical TADs from Hi-C chromatin interactions, and then applied HiTAD to analyzing Hi-C and *in situ* Hi-C datasets with different sequencing depths involving several human and mouse cell types (Supplementary Table S1). Compared to the selected two methods (Arrowhead and TADtree), HiTAD can detect more hierarchical TADs with higher replicate reproducibility and inter cell-type conservation. With a novel domain-based alignment strategy, we defined several change types of hierarchical TADs which were not systematically studied. Our analyses on these hierarchical TADs show that TADs and sub-TADs differ in correlating chromosomal compartment, replication timing domain and transcriptional regulation.

MATERIALS AND METHODS

Data sources, processing and representation

The *in situ* Hi-C datasets of human cell types GM12878, IMR90 and K562 were downloaded from NCBI with accession number GSE63525 (9). For traditional Hi-C, two independently generated datasets of human cell type GM12878 were downloaded from NCBI with accession numbers GSE48592 (30) and GSE63525 (9) respectively. The datasets of human cell type IMR90 and mouse cell types mESC and Cortex were downloaded from NCBI with accession numbers GSE43070 (14) and GSE35156 (11) respectively. The dataset of human cell type Panc1 was downloaded from ENCODE (31). Raw Hi-C data were processed and corrected by using software hiclib (32). The bins located in gap regions were removed from calculation but included in visualization. The summary of Hi-C datasets is listed in Supplementary Table S1. With respect to ChIP-Seq and RNA-Seq datasets, the processed data were downloaded from ENCODE (31), including epigenomic and binding peaks from ChIP-Seq and expression of long RNA contigs from RNA-Seq. The called domain boundaries of replication timing were downloaded from public website (http://mouseencode.org/publications/mcp05/) (33). Human CTCF motif was downloaded from a database for ENCODE transcription factors (http://compbio.mit.edu/encode-motifs/) (34), and mouse CTCF motif was scanned in Fimo (35) by using the same human PWM as input. The human and mouse reference genomes hg19 and mm10 were used in sequence alignments.

HiTAD overview

The idea behind HiTAD is that TADs are optimal domains to separate intra-chromosomal interactions in global level. Combining the fact that TADs can also be divided into smaller domains in a hierarchical way, the detection of hierarchical TADs can be transformed into an iterative optimization procedure by defining appropriate objective functions from interaction frequencies (Figure 1). In this work, the objective function is defined as the enrichment between intra-domain interaction frequencies and inter-domain interaction frequencies in a way to reduce the impact of genomic distance. To speed up the calculation, an adaptive directionality index (DI)-based Hidden Markov Model (HMM) is proposed to sensitively generate a genome-wide pool of bottom domains by using only local insulation. Then these bottom domains are used as basic elements to detect TADs by using global intra-chromosomal interactions under given objective function. To better perform TAD detection, a recursive formula is used to solve the optimization problem. These detected TADs are next used to generate corresponding sub-TADs in a similar way, but with the bottom domains localized within the TAD as initial domain pools. Similar procedure is applied to subsequent-level domain detection until bottom domains are met. Finally, under a domain-based alignment proposed by us, the reproducible domains from two replicates are maintained to guarantee the accuracy in the highly variable Hi-C chromatin interactions. HiTAD is a fast and memory-saving method that can be implemented in PC (Supplementary Table S2). Next, we will introduce the detailed implementation of HiTAD.

Bottom boundaries detected by adaptive directionality index

To sensitively detect the boundaries with various domain sizes, we proposed adaptive DI by modifying traditional DI (11,22). Let $(f_{ij})_{N\times N}$ represent the matrix of chromatin interaction frequencies after bias correction, where N is the number of bins at the given resolution. For every selected bin i, the adaptive DI is defined as:

$$DI_i = \frac{\frac{1}{W_i}\sum_{k=1}^{W_i} U_k - \frac{1}{W_i}\sum_{k=1}^{W_i} D_k}{\sqrt{\frac{\sum_{k=1}^{W_i}\left(U_k - \bar{U}_i\right)^2 + \sum_{k=1}^{W_i}\left(D_k - \bar{D}_i\right)^2}{W_i(W_i-1)}}} \qquad (1)$$

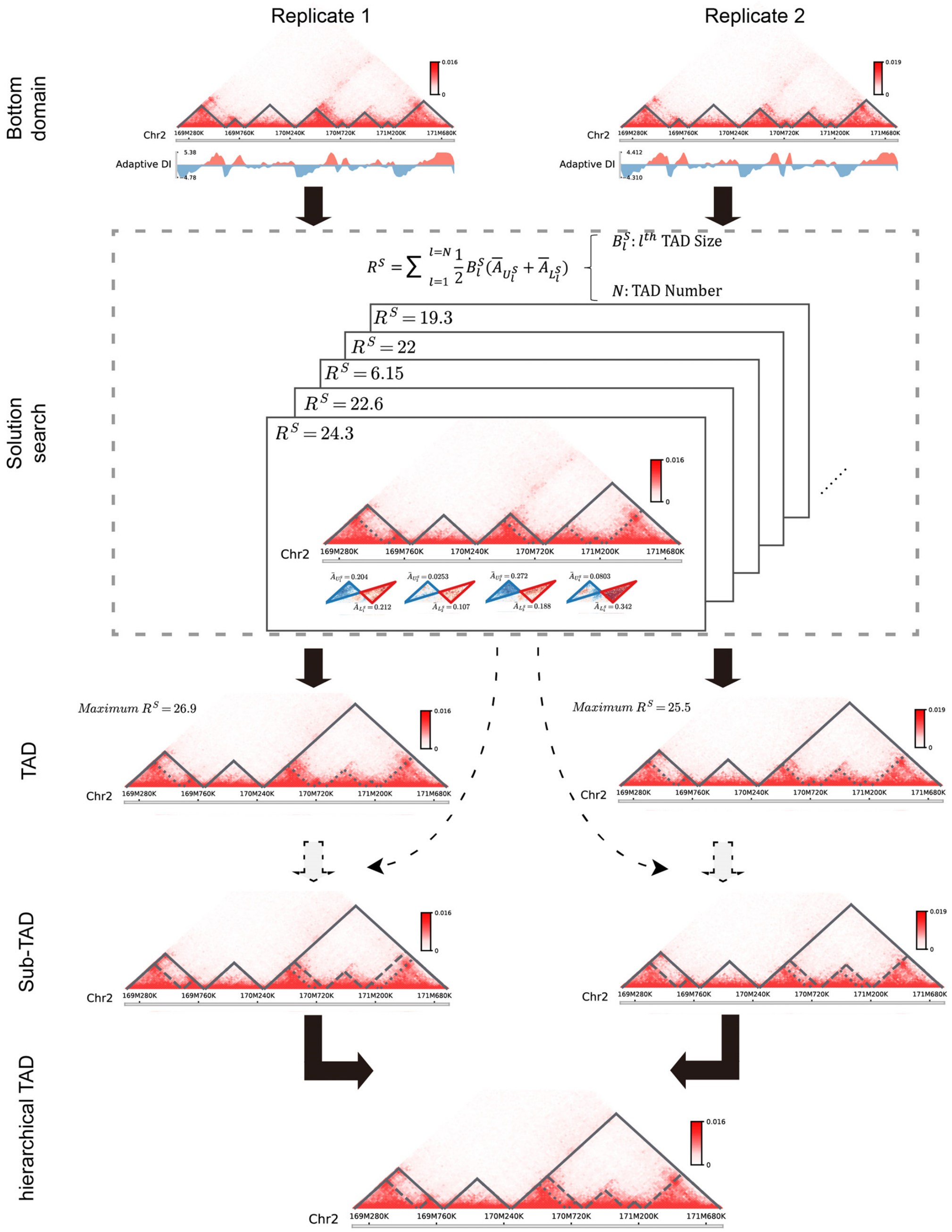
Replicate 1
Replicate 2
Bottom domain
Chr2
169M280K
169M760K
170M240K
170M720K
171M200K
171M680K
0.016
0.019
0
Adaptive DI
5.38
-4.78
4.412
-4.310
Solution search
$R^S = \sum_{l=1}^{l=N} \frac{1}{2} B_l^S (\bar{A}_{U_l^S} + \bar{A}_{L_l^S})$
B_l^S: l^{th} TAD Size
N: TAD Number
$R^S = 19.3$
$R^S = 22$
$R^S = 6.15$
$R^S = 22.6$
$R^S = 24.3$
$\bar{A}_{U_1^S} = 0.204$
$\bar{A}_{L_1^S} = 0.212$
$\bar{A}_{U_2^S} = 0.0253$
$\bar{A}_{L_2^S} = 0.107$
$\bar{A}_{U_3^S} = 0.272$
$\bar{A}_{L_3^S} = 0.188$
$\bar{A}_{U_4^S} = 0.0803$
$\bar{A}_{L_4^S} = 0.342$
TAD
Maximum $R^S = 26.9$
Maximum $R^S = 25.5$
Sub-TAD
hierarchical TAD

Figure 1. HiTAD workflow. First, the bottom domains are detected by using adaptive DI-based HMM. Second, under the objective function derived from chromatin interactions, a recursive formula is applied to searching all possible TADs by using bottom domains as input. Third, TADs are generated by maximizing the objective function from the searching space. Fourth, sub-TADs and other level domains are generated in a similar way but with localized bottom domains as input to optimization problem. Finally, the hierarchical domains reproducible from two replicates are maintained. The calculations were performed at 20 kb resolution on *in situ* Hi-C dataset IMR90.

where U_k denotes the upstream interaction frequency between bin i and bin $i-k$, D_k denotes the corresponding downstream interaction frequency and W_i is the window size on bin i. Since domain sizes vary from tens of kilobases to several megabases at currently available resolutions, the window size W_i is determined adaptively based on local interaction environment (Supplementary Figure S1). Specifically, let $S_i\ (k) = \{\begin{matrix} 1, & f_{i-k,i} - f_{i,i+k} \geq 0 \\ 0, & f_{i-k,i} - f_{i,i+k} < 0 \end{matrix}$ denote the interaction bias when comparing upstream bin $i-k$ to downstream bin $i+k$, in which 1 represents upstream bias and 0 represents downstream bias. Then four kinds of state transitions from $S_i(k)$ to $S_i(k+1)$ can be observed ($0 \to 0,\ 0 \to 1,\ 1 \to 0, 1 \to 1$). Generally, the state transitions should be statistically same from $S_i(k)$ to $S_i(k+m)$ if bin $i-k-m$ or bin $i+k+m$ is located in the same domain with bin i. Let $T_i\ (k) = \frac{1}{k+1} \sum_{j\ =\ 1}^{j\ =\ k} S_i(j)$, and then the maximum or minimum values of $T_i(k)$ ($k\ =\ 1, 2, 3, \cdots$) are selected as candidate window sizes. Since $T_i(k)$ is discrete, the maximum or minimum is determined locally by comparing its 10 nearest neighbor values, i.e. five neighbors in each side. To further guarantee the best selection on window size, a chi-square statistics is constructed as $\chi^2\ (i, k) = \sum_{j=1}^{4} \frac{(O_j(i,k) - E_j(i,k))^2}{E_j(i,k)}$, where $O_j(i, k)$ is the observed frequency in one of the four kinds of transitions, $E_j(i, k)$ is the expected frequency which is set to be $(k - 1)/4$ and k represents the candidate window sizes selected above. Then the minimum value of k satisfying $\chi^2(i, k) \geq \chi^2_{0.05(3)}$ is selected as the final window size W_i. Next, the calculated adaptive DIs from Equation 1 are used as input in HMM to detect bottom boundaries. Five states (start, upstream bias, no bias, downstream bias and end) and corresponding state transitions are set in HMM (Supplementary Table S3). Three-distribution Gaussian mixture is used to emit state and Baum–Welch algorithm is used to perform training on data. The detected boundaries are reused to further improve sensitivity through following procedure. Let $B\ = \{b_1^0, b_2^0, \cdots, b_n^0\}$ denote the initial boundaries from the adaptive DI-based HMM, where b_l^0 is the genomic position of the lth boundary and n is the total boundary number. Let $D_j^0 =\ [b_j^0, b_{j+1}^0]$ denote the domain where bin i localizes. The new window size of bin i is set to $W_i =\ \max\{i - b_j^0, b_{j+1}^0 - i\}$, and the corresponding adaptive DI is recalculated by using Equation 1. All recalculated adaptive DIs are used as input in HMM to detect new bottom boundaries and bottom domains. This recalculation is repeated until more than 95% boundaries detected from two neighbor steps can be aligned to each other (see 'Domain and boundary alignment' section). Generally, the convergence is achieved by only one to three iterations. We also performed comparison between adaptive DI-based HMM and traditional DI-based HMM. Since traditional DI depends on the fixed window sizes, different values were selected for thorough comparisons. The results show that most boundaries from traditional DI-based HMM can be detected by adaptive DI-based HMM. By contrast, many boundaries from adaptive DI-based HMM cannot be detected by traditional one (Supplementary Figure S2 and Table S4). The major reason underlying this difference is that traditional DI cannot reconcile the various domain sizes on boundary detection by using only one fixed window size.

Hierarchical TAD detection

The finally generated boundary set above is denoted as $B = \{b_1, b_2, \cdots, b_n\}$. The TAD identification is transformed into selecting a best subset $B^s = \{b_1^s, b_2^s, \cdots, b_m^s | m \leq n\}$ from set B to generate corresponding TAD set $D^s = \{D_1^s = [b_1^s, b_2^s],\ D_2^s = [b_2^s, b_3^s], \cdots, D_{m-1}^s = [b_{m-1}^s, b_m^s]\}$. To eliminate the impact of genomic distance on interaction frequency, the arrowhead transformation proposed by Rao *et al.* (9) is used to measure the interaction-frequency difference between intra-domain interactions and inter-domain interactions:

$$A_i\ (k) = \frac{f_{i-k,i} - f_{i,i+k}}{f_{i-k,i} + f_{i,i+k}},$$

where bin i is the genomic position and k is the genomic distance from bin i. Under this calculation, the square region of TAD $D_l^s = [b_l^s, b_{l+1}^s]$ should be separated to form an upper-triangle region $U_l^s = \{(i, k) | i \in (b_l^s,\ \frac{b_l^s + b_{l+1}^s}{2}),\ k \in (i - b_l^s,\ b_{l+1}^s - i)\}$ and a lower-triangle region $L_l^s = \{(i, k) | i \in (\ \frac{b_l^s + b_{l+1}^s}{2},\ b_{l+1}^s),\ k \in (b_{l+1}^s - i,\ i - b_l^s)\}$, where k is the genomic distance dependent on bin i (Supplementary Figure S3). The objective function is defined as:

$$\begin{cases} R^s = \sum_{l=1}^{l=m-1} \frac{1}{2} \left(b_{l+1}^s - b_l^s\right) \left(\bar{A}_{U_l^s} + \bar{A}_{L_l^s}\right) \\ \bar{A}_{U_l^s} = \mathrm{mean}\left(-w_{i,i+k} \cdot A_i\left(k\right)\right), (i, k) \in U_l^s \\ \bar{A}_{L_l^s} = \mathrm{mean}\left(w_{i-k,i} \cdot A_i\left(k\right)\right), (i, k) \in L_l^s \end{cases}, \quad (2)$$

where $w_{i,i+k}$ and $w_{i-k,i}$ are the weights on the interaction-frequency differences and $mean(A_i(k))$ represents the average of $A_i(k)$ in the corresponding region. Since $A_i(k)$ in the upper-triangle region represents the difference between upstream inter-domain interaction and intra-domain interaction, it tends to be negative in the TAD region. By contrast, $A_i(k)$ in the lower-triangle region represents the difference between intra-domain interaction and downstream inter-domain interaction, and tends to be positive. Then the values in the upper-triangle region are set to $-A_i(k)$.

The weight in Equation 2 is set to fold change between the observed and expected interaction frequencies, in which a previous procedure is used to calculate the expected interaction frequency by considering both genomic distance and local interaction background (23). Briefly, let $f\ (k) = \frac{\sum_{|i-j|=k} f_{ij}}{b_{l+1}^s - b_l^s - k}$ denote the average interaction frequency at genomic distance k in TAD $D_l^s = [b_l^s, b_{l+1}^s]$, and then the smoothed interaction frequency $F(k)$ is obtained by using B-Spline approximation. Since the number of chromatin interactions decreases with the genomic distance k, $F(k)$ may fluctuate when k increases gradually. To alleviate the impact of these fluctuations, the expected interaction frequency is defined as $E_{ij} = \{ \begin{matrix} F(|i-j|),\ |i-j| \leq k_r \\ F(k_r),\ |i-j| > k_r \end{matrix}$, where k_r is the first turning point. The final expected interaction frequency at position (i, j) is calculated by following a previous window strategy to take local interaction background into consideration (9):

$$E_{ij}^* = \frac{\sum_{m=i-w}^{m=i+w} \sum_{n=j-w}^{n=j+w} f_{mn} - \sum_{m=i-p}^{m=i+p} \sum_{n=j-p}^{n=j+p} f_{mn} - \sum_{m=i-w}^{m=i-p-1} f_{mj} - \sum_{m=i+p+1}^{m=i+w} f_{mj} - \sum_{n=j-w}^{n=j-p-1} f_{in} - \sum_{n=j+p+1}^{n=j+w} f_{in}}{\sum_{m=i-w}^{m=i+w} \sum_{n=j-w}^{n=j+w} E_{mn} - \sum_{m=i-p}^{m=i+p} \sum_{n=j-p}^{n=j+p} E_{mn} - \sum_{m=i-w}^{m=i-p-1} E_{mj} - \sum_{m=i+p+1}^{m=i+w} E_{mj} - \sum_{n=j-w}^{n=j-p-1} E_{in} - \sum_{n=j+p+1}^{n=j+w} E_{in}} \cdot E_{ij}$$

where $2p + 1$ and $2w + 1$ are the two square-window sizes centered at (i, j) with $p\ =\ 1$ and $w\ =\ 3$. Then the corresponding weight is defined as $w_{ij}^* = {f_{ij}}/{E_{ij}^*}$. Finally, all weights in the TAD $D_l^s = [b_l^s, b_{l+1}^s]$ are normalized by using a piecewise function:

$$W_{ij} = \begin{cases} 1,\ w_{ij}^* \geq w_u \\ 0.5 \cdot \frac{w_{ij}^* - 1}{w_u - 1} + 0.5,\ 1 \leq w_{ij}^* < w_u \\ 0.5 \cdot \frac{w_{ij}^* - w_l}{1 - w_l},\ w_l < w_{ij}^* < 1 \\ 0,\ 0 < w_{ij}^* \leq w_l \end{cases}, \quad (3)$$

where, w_u and w_l are the 99 percentile and 1 percentile respectively to reduce the impact of outliers. If $w_{ij}^* =\ 1$ in Equation 3, the corresponding weight is set to 0.5. The same value (0.5) is set to the chromatin interactions within bottom domains to avoid their over impacts on objective function since $A_i(k)$ in bottom domains are generally large due to very high interaction frequencies. The chromatin interactions with $f_{ij} = 0$ are excluded from weight calculation and objective function.

Under the objective function defined in Equation 2, TAD detection is transformed into searching the solution with maximum value. Then a recursive formula is used to solve this optimization problem. Specifically, for the first bottom domain d_1, the objective function is calculated by:

$$R_1 = \frac{1}{2}\,(b_2 - b_1)\left(\bar{A}_{U_{D1}} + \bar{A}_{L_{D1}}\right),$$

where $D1$ represents the set of all kinds of TAD selections on domain d_1 in the first step, and U_{D1} and L_{D1} are the corresponding upper-triangle and lower-triangle regions defined previously. Actually, the start domain d_1 can only be an independent TAD or merged with consecutive downstream bottom domains (Supplementary Figure S4a). The set of TAD selections on domain d_1 in the initial step is used in next step. In step l on domain d_l, the objective function is calculated:

$$R_l = \begin{cases} \frac{1}{2}\,(b_{l+1} - b_l)\left(\bar{A}_{U_{Dl}} + \bar{A}_{L_{Dl}}\right) + \max\left(R_{l-1}\right), & Dl \in S(I) \\ \frac{1}{2}\,(b_{l+1} - b_l)\left(\bar{A}_{U_{Dl}} + \bar{A}_{L_{Dl}}\right) + R_{l-1}, & Dl \in S(II) \end{cases},$$

where S(I) and S(II) represent two subsets for domain d_l. In subset S(I), domain d_l is independent on the candidate TADs generated in previous steps, whereas in subset S(II), d_l has already been contained in previous steps and should be re-evaluated (Supplementary Figure S4b). The ultimate goal is to find the boundary subset B^s to maximize objective function R_{n-1}, which can be written as $\max_{B^s} R_{n-1}$. To fasten the calculation, the maximum TAD size is limited to 4 Mb in the aforementioned search procedure in this work. However, different selections on TAD-size limitation only slightly influence the accuracy of TAD detection (Supplementary Table S5).

The sub-TADs in each TAD D_l are detected by using the same procedure as TAD detection except that the input boundary set is replaced by the boundary subset $B^{D_l} = \{b_k | b_{l-1} \leq b_k \leq b_l,\ b_k \in B\}$, where b_{l-1} and b_l are the upstream and downstream boundaries of TAD D_l respectively. The subsequent level domains in each sub-TAD are detected in the similar way until bottom domains are met. Finally, the TADs and other domains reproducible from two replicates are maintained to form hierarchical TADs. The reproducibility is calculated by using following alignment strategy.

Domain and boundary alignment

Traditionally, domain and boundary are aligned by matching boundaries with nearest genomic positions between two Hi-C datasets. This strategy generally assigns a threshold in advance, but neglects the usage of domain themselves. To facilitate the domain and boundary matching for hierarchical TADs, a domain-based alignment strategy is proposed in this work by considering all domains in the same time.

Specifically, let $B^1 = \{b_1^1, b_2^1, \cdots, b_m^1\}$ and $B^2 = \{b_1^2, b_2^2, \cdots, b_n^2\}$ denote two boundary sets. For chromatin region $d_i^1 = [b_p^1, b_q^1]$ $(q > p)$ in set B^1 and chromatin region $d_j^2 = [b_u^2, b_v^2]$ $(v > u)$ in set B^2, let (b_1, b_2, b_3, b_4) represent the ascending order of corresponding genomic positions $(b_p^1, b_q^1, b_u^2, b_v^2)$. The overlap ratio between these two regions is defined as:

$$OR\left(d_i^1, d_j^2\right) = \begin{cases} 0, & b_q^1 = b_2 \ or \ b_v^2 = b_2 \\ \frac{b_3 - b_2}{b_4 - b_1}, & b_q^1 \neq b_2 \ and \ b_v^2 \neq b_2 \end{cases}.$$

Correspondingly, for region d_i^1 in set B^1, the best mapping in set B^2 is defined as $d_{i'}^2$ satisfying $OR\,(d_i^1,\ d_{i'}^2) = \max_{d_k^2}\{OR(d_i^1,\ d_k^2)\}$, where d_k^2 represents any chromatin region in set B^2. This directional mapping is denoted as $\{d_i^1\} \to \{d_{i'}^2\}$. The reverse mapping $\{d_j^2\} \to \{d_{j'}^1\}$ from set B^2 to set B^1 can be defined in the same way. If $d_j^2 = d_{i'}^2$ and $d_i^1 = d_{j'}^1$, then chromatin regions d_i^1 and d_j^2 are defined as bidirectional mapping, i.e. $\{d_i^1\} \leftrightarrow \{d_j^2\}$. When performing alignment, the selected bottom domain in one boundary set can be mapped to a chromatin region composed of several consecutive bottom domains in another boundary set. By integrating the two directional mappings, the consecutive bottom domains in the same boundary set are combined to form the combinatorial bidirectional mapping. As shown in Supplementary Figure S5, there are three bottom domains $\{d_i^1, d_{i+1}^1, d_{i+2}^1\}$ and $\{d_j^2, d_{j+1}^2, d_{j+2}^2\}$ in boundary set B^1 and set B^2 respectively. According to the mapping definition, the domain mapping from set B^1 to set B^2 is $\{d_i^1\} \to \{d_j^2, d_{j+1}^2\}$, $\{d_{i+1}^1\} \to \{d_{j+1}^2\}$ and $\{d_{i+2}^1\} \to \{d_{j+2}^2\}$, and the reverse mapping is $\{d_j^2\} \to \{d_i^1\}$, $\{d_{j+1}^2\} \to \{d_i^1, d_{i+1}^1\}$ and $\{d_{j+2}^1\} \to \{d_{i+2}^2\}$. Combing these two directional mappings yields the combinatorial bidirectional mapping: $\{d_i^1, d_{i+1}^1\} \leftrightarrow \{d_j^2, d_{j+1}^2\}$ and $\{d_{i+2}^1\} \leftrightarrow \{d_{j+2}^2\}$.

The hierarchical TAD alignment is based on the mapping strategy defined above. First, bidirectional mapping is performed for all bottom domains in two boundary sets B^1 and B^2, including the combinatorial bidirectional mapping. Second, for each chromosome, starting from the first TAD in the domain set D^1 generated from boundary set B^1, the bottom domains of current TAD are extracted and bidirectional mapping is searched in the other domain set D^2 generated from boundary set B^2. If not all bottom domains are bidirectionally mapped, the TAD and its downstream TAD are merged to search bidirectional mapping again. This TAD merging and mapping iteration is repeated until bidirectional mapping is achieved. Then the hierarchical levels of mapped bottom domains in domain set D^2 are extracted and recorded for corresponding TADs in domain set D^1. All level sub-domains in current TAD or merged TADs are extracted and the same mapping procedure is performed. Third, the next TAD without performing mapping is selected as a new start to repeat above procedure until all TADs in the selected chromosome undergo hierarchical mapping, including all level sub-domains. Fourth, the next chromosome is selected to perform the same hierarchical mapping until all chromosomes are done.

Other calculations in HiTAD analysis

In method comparison, the hierarchical TAD detection in TADtree is performed by using recommended parameters.

Reproducibility is defined as the number of reproducible domains dividing the maximum domain number in two replicates. Conservation ratio between two cell types is defined in the similar way by using the domain numbers in two cell types to replace those in two replicates. In method evaluation, the random hierarchical TADs are generated by using the following iterative shuffling procedure. To preserve original hierarchies, the TAD sizes and corresponding subdomain sizes are recorded for each chromosome. The TAD positions are shuffled according to the recorded TAD sizes in each chromosome, and then the sub-TAD positions are shuffled in each TAD according to the recorded sub-TAD sizes. The above shuffling is repeated until no hierarchical levels are recorded in the original hierarchical TADs. As for boundary analysis, the enrichments of epigenomic peaks, CTCF peaks and CTCF motifs are calculated by following previous works (11,36) and the compartments are calculated at 200 Kb resolution by using package hiclib (32).

RESULTS

Method comparison

To investigate the performance of HiTAD, we first compared HiTAD to two other methods with available software, Arrowhead implemented in Juicer (37) and TADtree (24). Matryoshka and the network modularity based method were excluded from comparisons due to software-installation difficulty and code unavailability respectively. The methods CaTCH (25) and HBM (26) were excluded since they do not explicitly point out TAD positions. The methods for traditional domains, such as DI based HMM (11), HiCseg (38), TopDom (39), TADbit (bioRxiv https://doi.org/10.1101/036764), HiCExplorer (bioRxiv https://doi.org/10.1101/115063), Armatus (27), spectral method (28), MrTADFinder and IC-Finder (29), were also excluded since they don not explicitly detect or automatically output hierarchical TADs. The selected methods were applied to both traditional and *in situ* Hi-C datasets under 40 kb resolution since it is pretty hard for TADtree to be run at higher resolutions. To simplify the statement, traditional and *in situ* Hi-C datasets are denoted by different suffixes, such as GM12878-T and GM12878-I respectively. To clarify hierarchical levels, TAD is denoted as level 0, sub-TAD is denoted as level 1 and subsequent domain level is denoted as level 2, etc.

HiTAD outperforms the other two methods in domain sensitivity, replicate reproducibility and inter cell-type conservation (Figure 2). HiTAD detects more domains in all levels for both *in situ* and traditional Hi-C datasets, especially in level 0, level 1 and level 2 domains (Figure 2A). This sensitivity improvement mainly arises from the fact that HiTAD successfully detects domains in more chromatin regions compared to the other two methods (Supplementary Figure S6a). Besides, the different domain size distributions can also contribute to differences in domain numbers. Generally, the domain sizes from HiTAD are smaller than those from Arrowhead but larger than or comparable to those from TADtree (Supplementary Figure S6b). Next GM12878-I and GM12878-T with four biological replicates available (Supplementary Table S1) were selected to evaluate reproducibility since HiTAD generally needs two replicates to detect hierarchical TADs. The four replicates were divided into two groups to independently generate two sets of hierarchical TADs for HiTAD. As for the other two methods, the two replicates in the same group were merged in hierarchical TAD detection. Our calculation shows that HiTAD outperforms the other two methods in the replicate reproducibility in all hierarchical levels on both GM12878-I and GM12878-T (Figure 2B). Similarly, the hierarchical TADs detected by HiTAD are more conserved across cell types than those detected by the other two methods in level 0, level 1 and level 2 domains (Figure 2C).

Finally, it should be noted that the metrics used in this work just reflect parts of algorithmic performance. Generally, higher domain number and chromatin coverage could only indicate better sensitivity, reproducibility is an important aspect of algorithms, and conservation ratio reflects the biological aspect of hierarchical TADs. We think these metrics together can reflect the algorithmic performance in hierarchical TAD detection, but they are not necessarily equal to algorithmic accuracy. However, there is currently no golden standard to evaluate the accuracy of hierarchical TADs. Compared to traditional TADs, it is more difficult to evaluate hierarchical TADs since this evaluation contains both chromatin domains and the hierarchies (Supplementary Figure S7). Better metrics or standards may be developed in the future.

Method evaluation

To obtain better details, hierarchical TADs were detected at 20 kb resolution for *in situ* Hi-C datasets in the following analyses. Since the numbers of level 2 and level 3 domains were quite limited, these domains were combined in next calculations. The shared boundary among different level domains was classified as lower-level boundary.

We first evaluated HiTAD by measuring the insulation effects of detected hierarchical boundaries. Two histone modifications, H3K36me3 and H3K27me3, were selected to represent active and inactive signals. The boundary insulation was calculated by following a previous procedure (9). Briefly, in each level of hierarchical TADs, every domain was divided into 10 bins, and the strength of histone modification was recorded for every bin. Then an $N \times 20$ matrix was generated, where 20 columns represent the signal strengths of two consecutive domains and rows represent all possible consecutive domains. The correlations of the columns of this matrix reflect how the epigenomic signals in any two bins are correlated. Our calculations show that the signals in the same domain are highly correlated to each other, but the signal correlations between two consecutive domains are sharply separated in the boundaries (Figure 3A and Supplementary Figure S8). Furthermore, the lower level boundaries exhibit stronger insulation effects. As for the controls generated from random shuffling, there are no such sharp separations in the boundaries. These results together indicate that the detected hierarchical boundaries, especially the level 0 and level 1 boundaries, exhibit insulation effects.

We next performed enrichment analysis on different level boundaries. Two histone modifications, H3K4me3 and H3K4me1, were selected to represent promoters and en-

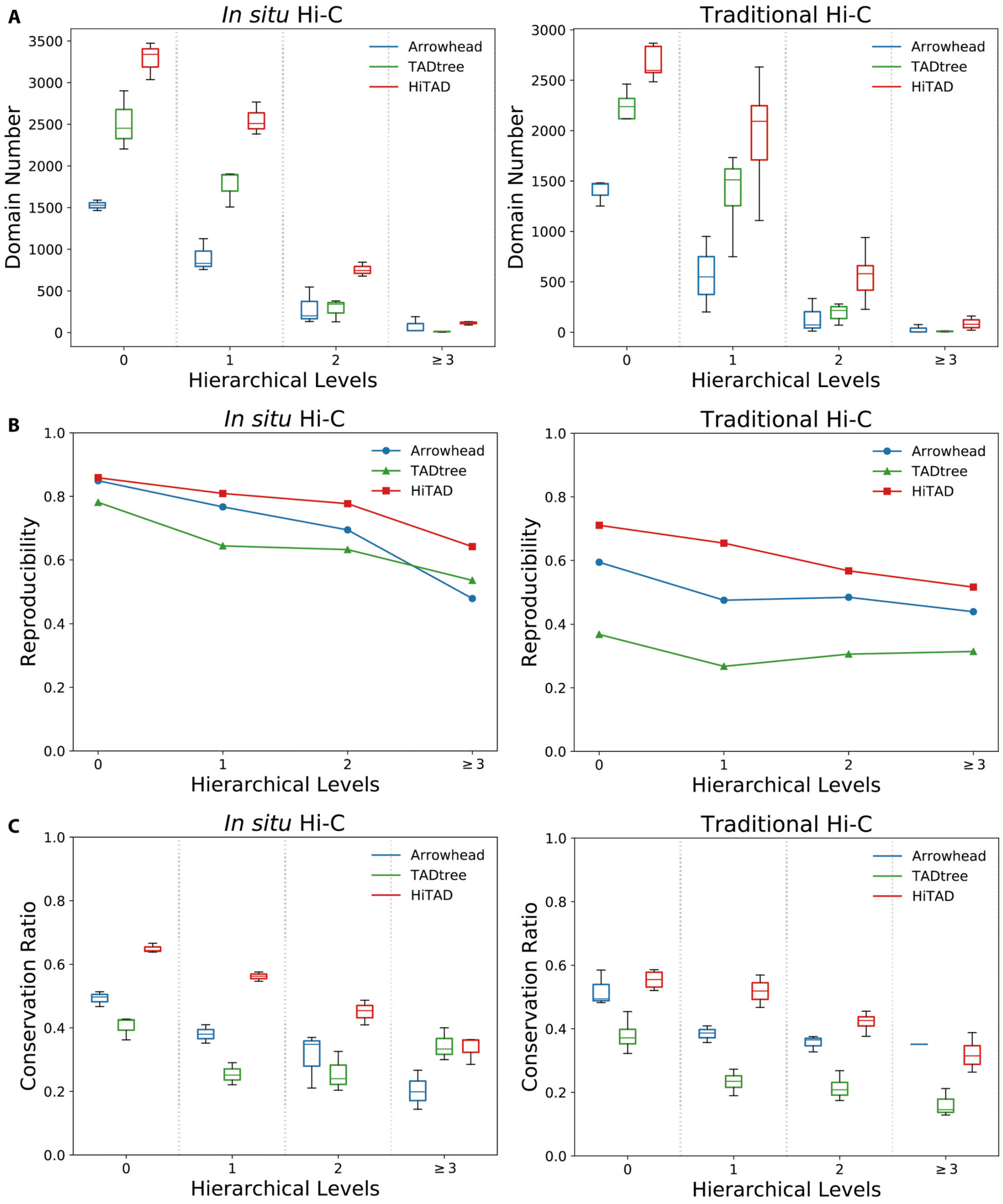

Figure 2. Performance comparison. Three methods, Arrowhead, TADtree and HiTAD, are included to perform comparisons. The left and right figures illustrate the results calculated from *in situ* and traditional Hi-C datasets respectively. (**A**) Domain number. (**B**) Replicate reproducibility. (**C**) Inter cell-type conservation.

hancers respectively. The key architectural protein CTCF was also included since this protein plays an important role in shaping chromatin domains (40). The calculated results show that different level boundaries are enriched in promoter signal H3K4me3 and CTCF binding sites but are a little depleted in enhancer signal H3K4me1 (Figure 3B and Supplementary Figure S9a), consistent with traditional TAD analysis (9,11). However, higher level boundaries generally show lower signals in both boundaries and nearby background regions. It was reported that the divergent CTCF motifs shaped the domain boundaries (9,36,41–43), so we further analyzed the composition of CTCF motif directions in these binding sites. The result shows that different level boundaries are enriched in divergent CTCF motifs in overall (Figure 3C and Supplementary Figure S9b), but the higher-level boundaries show lower densities on both total and divergent CTCF motifs. The results from histone modifications, CTCF binding sites and motif direction together indicate that almost all level boundaries show similarities with traditional TAD boundaries, but different level boundaries exhibit different signal strengths. These signal differences may partially explain the insulation differences among hierarchical boundaries. However, the limited domain numbers and relatively low reproducibility of higher

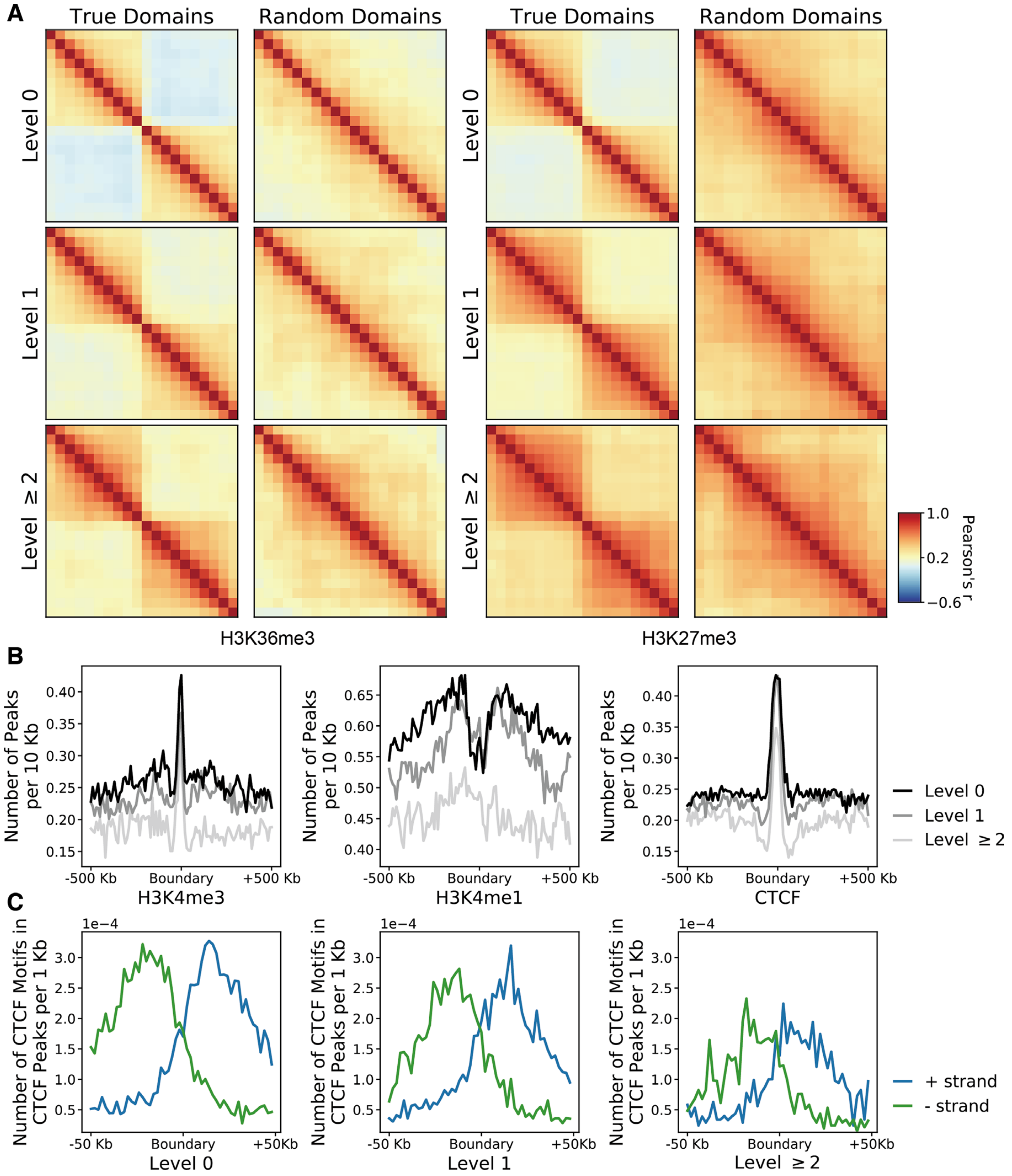

Figure 3. HiTAD evaluation by insulation effects and signal enrichments. (**A**) Boundary insulation measured by signal correlations. There are 20 bins in each heatmap, in which the first and second 10 bins represent two consecutive domains respectively. The left two heatmaps are calculated from active signal H3K36me3, while the right two heatmaps are calculated from inactive signal H3K27me3. (**B**) Signal enrichment in different level boundaries. Three representative signals are shown, including H3K4me3, H3K4me1 and CTCF peaks. (**C**) Enrichment of directional CTCF motifs. The divergent CTCF motifs on boundaries are composed of minus-strand CTCF motifs in upstream regions and plus-strand CTCF motifs in downstream regions.

level domains can influence the reliable analysis on boundary enrichments, and thorough works are needed to further investigate their functional characteristics in the future.

Boundary-level analysis on structural and functional properties of hierarchical TADs

Due to limited domain numbers, level 2 and level 3 domains were excluded from calculations in this and following sections to simplify analysis. The shared boundaries between TADs (level 0 domains) and sub-TADs (level 1 domains) were classified as TAD boundaries. When comparing two cell types, TAD boundaries and sub-TAD boundaries undergo different structural changes, including conserved TAD boundary, conserved sub-TAD boundary, disappeared TAD boundary, disappeared sub-TAD boundary and TAD-to-sub-TAD boundary switch (Supplementary Figure S10). Statistically, TAD boundaries exhibit quite high inter cell-type conservation, whereas sub-TAD boundaries are quite dynamic across cell types (Figure 4A and Supplementary Figure S11). Next, we extended traditional analyses on compartment (44) and replication timing (15) to investigate the similarities and differences between TAD

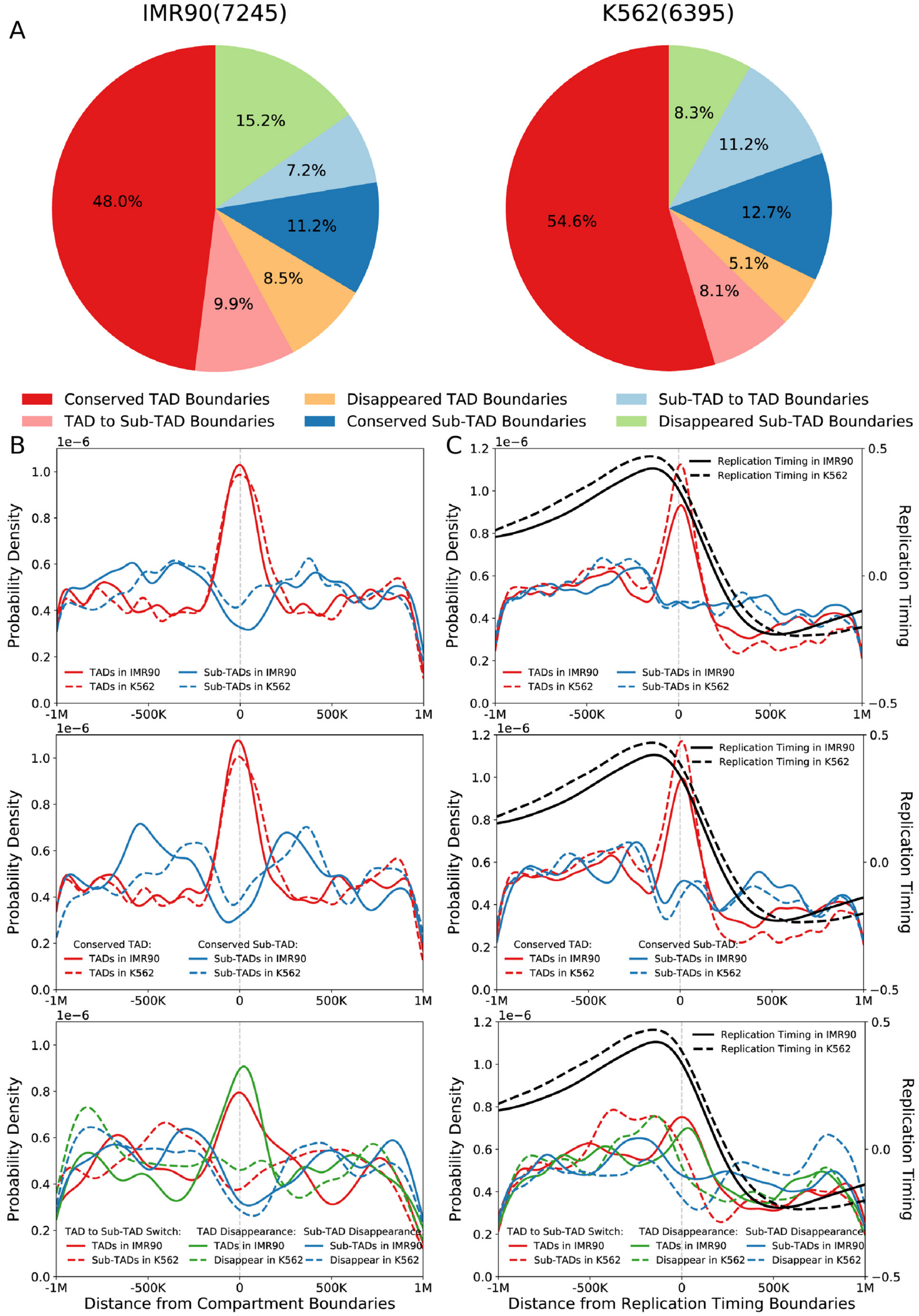

Figure 4. Structural and functional analysis on boundary-level changes. (**A**) The left and right pie charts illustrate the proportion of hierarchical boundary changes by using IMR90-I and K562-I as references respectively. The total boundary number is presented in the bracket for each cell type. (**B**) The top figure illustrates the enrichment difference between TAD boundaries and sub-TAD boundaries in the same cell type (intra cell-type comparison). The medium and bottom figures illustrate the relationship between hierarchical boundary changes and compartment boundary changes across cell types (inter cell-type comparison). The x-axis denotes the genomic distance to compartment boundaries, and the y-axis denotes the probability density. (**C**) The enrichments on replication timing boundaries are presented in the same way as compartment boundaries, except that the replication timing transitions are included in each sub-figure. The x-axis denotes the genomic distance to replication timing boundaries. The left y-axis denotes the probability density of domain boundary enrichment, and the right y-axis denotes the replicating timing.

boundaries and sub-TAD boundaries by using both intra cell-type and inter cell-type comparisons. The presented results are mainly calculated from IMR90-I and K562-I since there are replication timing data in human cell types IMR90 and K562.

TAD boundaries but not sub-TAD boundaries mainly separate higher-order chromosomal compartments. Lieberman-Aiden *et al.* proposed A-B (active-inactive) compartments in original Hi-C work (6), so we performed enrichment analysis on these compartment boundaries. Figure 4B illustrates that the A-B compartment boundaries are enriched in TAD boundaries but depleted in sub-TAD boundaries when performing intra cell-type analyses on IMR90-I and K562-I independently. As for inter cell-type comparisons, different change types were analyzed separately. The trends in conserved TAD and sub-TAD boundaries are the same as those from intra cell-type analysis. If TAD boundaries switch to sub-TAD boundaries from one cell type to the other, the enrichment-to-depletion switch is also observed simultaneously. Similarly, if TAD boundaries disappear in the other cell type, the enrichment on compartment boundaries also disappears. In the case of sub-TAD disappearance, the stronger depletion is observed in the corresponding cell type (Figure 4B and Supplementary Figure S12a). The same phenomena are observed in other inter cell-type comparisons (Supplementary Figure S13). These results together suggest that the changes of compartment boundaries are mainly accompanied with the changes of TAD boundaries across cell types. Combining intra cell-type and inter cell-type analyses, we can conclude that the TAD but not sub-TAD boundaries are mainly involved in correlating higher-order compartment.

TAD boundaries but not sub-TAD boundaries mainly separate replication timing domains. Similar to compartment boundaries, replication timing boundaries are enriched in TAD boundaries but depleted a little in sub-TAD boundaries in both IMR90 and K562 independently (Figure 4C). As for inter cell-type comparisons, the conserved TAD boundaries are enriched in replication timing boundaries, and the conserved sub-TAD boundaries in overall are depleted in replication timing boundaries. If TAD boundaries in IMR90-I switch to sub-TAD boundaries or even disappear in K562-I, the enrichments disappear simultaneously. No enrichment is observed in the case of sub-TAD boundary disappearance (Figure 4C and Supplementary Figure S12b). In summary, the changes of replication timing boundaries are also mainly accompanied with the changes of TAD boundaries across cell types, indicating the difference between TAD and sub-TAD boundaries in correlating the replication timing boundaries.

Domain-level analysis on structural and functional properties of hierarchical TADs

We next investigated domain-level changes and corresponding transcriptional associations for hierarchical TADs. Compared to boundary-level change, the domain-level change is more complicated since it covers wider chromatin region and involves simultaneous changes of several boundaries. Through different inter cell-type comparisons, we observed several types of domain-level changes for hierarchical TADs, which are defined as conserved TAD, semi-conserved TAD, merged TAD, split TAD and undefined TAD in this work. Conserved TAD is conserved completely in all hierarchical levels, while semi-conserved TAD is conserved in TAD level but dynamic in sub-TAD level. These two cases together represent traditional TAD conservation across cell types. Merged TAD represents the situation that two or more TADs are merged to form a new TAD from one cell type to another cell type, while split TAD represents the reverse situation. The undefined TAD denotes the other TAD change with no clear definition in this work. Similar to TAD change, the semi-conserved TAD can be further classified as merged sub-TAD, split sub-TAD and undefined sub-TAD (Supplementary Figure S14).

By using IMR90-I as reference, 33.5% TADs are totally conserved and 11.8% TADs are semi-conserved when comparing IMR90-I to K562-I (Figure 5A). Interestingly, the undefined sub-TAD change is not observed in this comparison, and only very few cases are observed in other inter cell-type comparisons in this work (Supplementary Figure S15). These consistent results indicate that most sub-TAD changes in semi-conserved TADs only involve sub-TAD mergence or sub-TAD split without complicated combinations. As for TAD-level change, around 31.7% and 8.6% TADs show TAD mergence and split respectively, and the rest 14.4% TADs undergo complicated domain-level changes in IMR90-I (Figure 5A). Similar trends are observed in K562-I (Figure 5A) and other inter cell-type comparisons (Supplementary Figure S15).

We next investigated the relationship between domain changes and gene expressions. The undefined sub-TADs and undefined TADs were excluded from calculation since there are no consistent domain changes in these two cases. Compared to conserved TADs, the sub-TAD mergence and split in semi-conserved TADs exhibit transcriptional upregulation and downregulation respectively. The similar trends are observed in TAD mergence and split (Figure 5B and Supplementary Figure S16). To further reveal the functional roles of TAD mergence and split, we analyzed the gene expression changes on two consecutive TADs jointly by using following procedure. First, each TAD was separated into 100 bins and the total RNA expression in each bin was normalized by corresponding bin size. Second, the RNA expressions on 100 bins were summed and averaged in each TAD of two consecutive TADs. Then the one with lower average RNA expression was classified as inactive TAD, and the other one was classified as active TAD. Third, the average RNA expression in each bin was calculated by using all inactive TADs or active TADs respectively. Figure 5C illustrates the calculated results for TAD mergence. Intuitively, the transcriptional gap between inactive TADs and active TADs is quite large in the original cell type IMR90, but this gap is shrunk after the TAD mergence in cell type K562. To better show the differences, the 100 average RNA expression values in the two kinds of TADs were separately presented by using box plots. Figure 5D clearly illustrates the significant upregulation and downregulation in the originally inactive and active TADs respectively. The reverse effect is observed in TAD split (Figure 5E and F). The same regulatory effects are also observed in other inter cell-type comparisons (Supplementary Figure S17). These converged

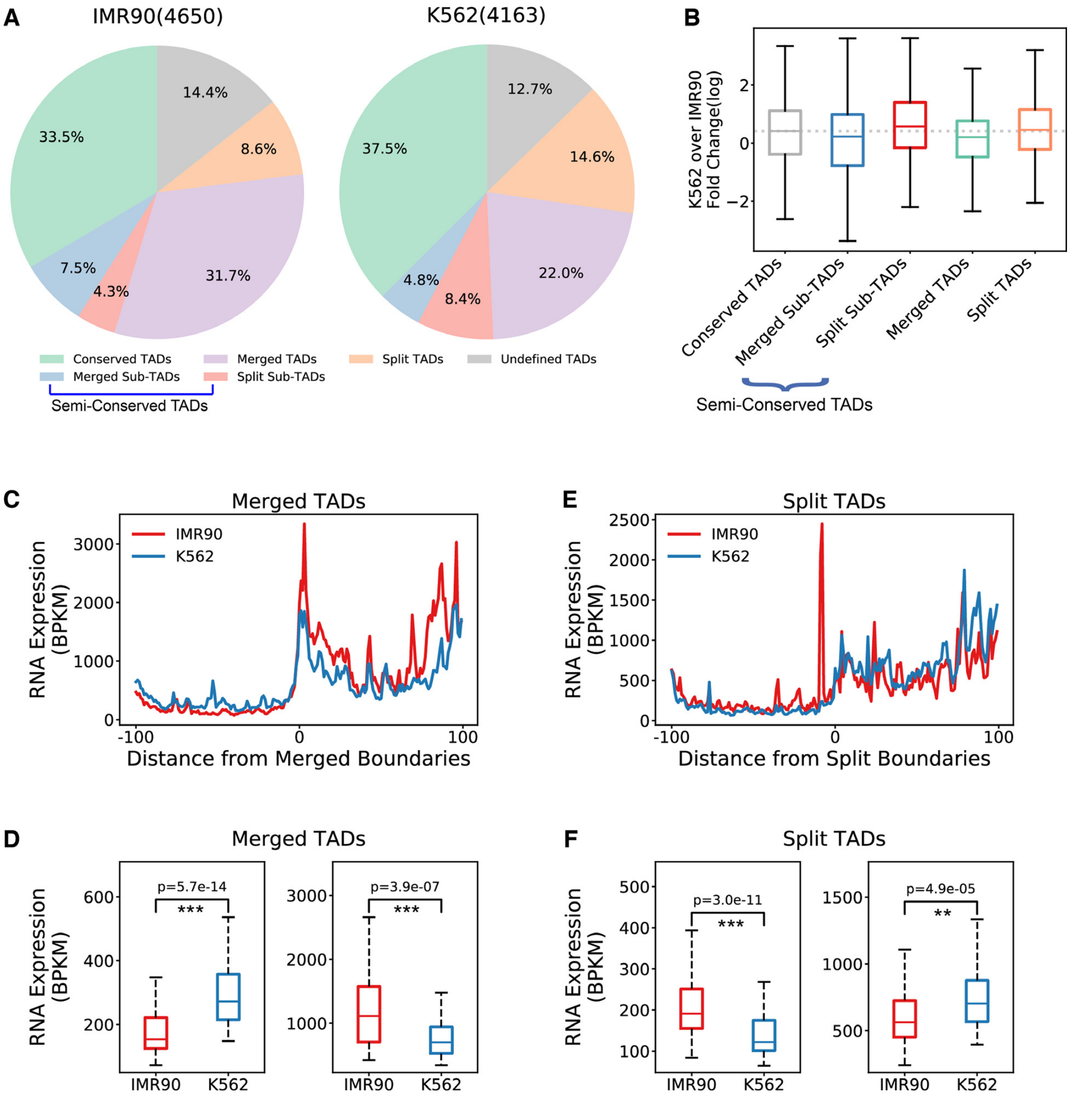

Figure 5. Structural and functional analysis on domain-level changes. (**A**) The left and right pie charts illustrate the proportion of hierarchical domain changes by using IMR90-I and K562-I as references respectively. The total TAD number is presented in the bracket for each cell type. (**B**) The transcriptional fold changes are presented separately for different domain changes when comparing K562 to IMR90. (**C**) The averaged RNA expressions on merged TADs are presented in the order of inactive to active TADs. Merged TAD represents that the two consecutive TADs in IMR90 are merged to be a new TAD in K562. BPKM: bases per kilobase per million mapped bases. (**D**) Box plots of average RNA expressions on merged TADs illustrate the transcriptional changes between cell types. The box plots in the left and right figures are calculated from the 100 bins in inactive and active TADs respectively. The Mann–Whitney U test was used to calculate *P*-values ($^{*}P < 0.05$, $^{**}P <$ 1e-3 and $^{***}P <$ 1e-5). (**E**) The averaged RNA expressions on split TADs are presented in the same order as merged TADs. Split TAD represents that a TAD in IMR90 are split into two consecutive TADs in K562. (**F**) Box plots of average RNA expressions on split TADs illustrate the transcriptional changes between cell types, in the same way as merged TADs.

results suggest the role of TAD mergence in shrinking the transcriptional gap between two consecutive TADs. With respect to sub-TAD mergence and split, we did not observe completely consistent patterns in all inter cell-type comparisons (Supplementary Figure S18). Thorough works are needed to further clarify the transcriptional difference between TAD mergence and sub-TAD mergence in the future.

DISCUSSION

A current question in the field of hierarchical chromatin domain is whether TADs are structurally and functionally distinct from sub-TADs and other domains. Since traditional TAD definition is a little ambiguous, Dixon *et al.* recently attempted to define TAD to be stable through cell divisions and conserved through cell lineages (45). As discussed in a recent paper (46), this definition is hard to implement computationally. However, by utilizing only chromatin interactions in one cell type, HiTAD detects TADs and corresponding hierarchical domains with pretty high replicate reproducibility and inter cell-type conservation, partially satisfying the TAD definition through biological processes. The detected hierarchical TADs are also evaluated by insulation effects as well as signal enrichments. In summary, our work suggests that the refinement of TAD definition, including hierarchical structure and biological function, can be achieved by only analyzing high-quality chromatin interactions, at least in part.

HiTAD adopts some strategies different from traditional methods. First, except the local insulations used by traditional methods, HiTAD also utilizes global intra-chromosomal interactions by constraining TADs to the best separation on individual chromosomes. In other words, the TAD detection in HiTAD attempts to make use of the advantages of local and global intra-chromosomal interactions in a concerted way. Second, biological replicates are used to generate final hierarchical TADs in our method. The advantage of this strategy is that flexible methods can be used to generate sufficient bottom domains and corresponding hierarchies for every replicate, instead of considering the high variations of chromatin interactions in the initial start. Though the usage of replicate reproducibility can improve the accuracy and reliability from highly variant data, it does not mean that the obtained boundaries and domains are reproducible in all replicates. This is because two replicates cannot represent all cases in the population. In addition, the replicate reproducibility requires the comparable data quality from two replicates. Otherwise, the low quality data will dominate the final results over the high quality data. In this case, we suggest that the results from high quality data or merged data should be used. However, it is not difficult currently to generate Hi-C data with enough quality for HiTAD due to the improvement of sequencing technology and the reduction of sequencing cost. Third, compared to traditional boundary-based alignment, domain-based alignment in our method adaptively matches the domains and boundaries by eliminating the choice of distance threshold between the two aligned boundaries. The domain-based alignment also allows unmatched domains and boundaries automatically.

The change types of hierarchical TADs defined in this work can pave the way for further studies on chromosomal structures and functions. In this work, we defined several inter cell-type changes on boundaries and domains respectively for hierarchical TADs and explored their structural and functional roles for the first time. Compared to boundary-level change, the domain-level change generally covers wider chromatin region and involves several boundaries. For transcriptional regulation, we used a simple method to investigate the domain-level relationship between hierarchical TADs and gene expressions. But for the chromosomal compartment and replication timing domain, their inter cell-type changes also involve several boundaries. In this way, the domain-level analysis will meet the combinatorial problem when simultaneously comparing several boundaries, so the boundary-level enrichments were performed on chromosomal compartment and replication timing by following previous works (15,44) to simplify analysis. In addition, both boundary-level and domain-level analyses were simplified by only considering two levels, TAD and sub-TAD. More complicated structural changes can be observed if taking more hierarchical levels into consideration, but it will be more difficult to depict their biological functions, especially with the limited number of higher level boundaries and domains. In spite of simplified analyses, we still revealed the structural and functional differences between TADs and sub-TADs by using these hierarchical TAD changes. Since these phenomena are quite common in different inter cell-type comparisons, it is quite possible that these analyses can be applied to other mammalian cell types not covered in this work.

TAD and sub-TAD differ in correlating chromosomal compartment and replication timing. In the chromosomal structure, the TAD boundaries but not sub-TAD boundaries mainly correlate the chromosomal compartments. As for the replication timing domain, the same trend is observed. These together argue that TADs are the main structural units in linking higher-order chromosomal organization and replication timing. However, this does not necessarily mean that sub-TADs have no effect. The identification resolution of compartment is relatively low due to sequencing depth, making subtle comparisons difficult. Sub-compartment was recently proposed through extremely deep sequencing, but it is hard to perform reliable analysis on sub-compartment due to the limited data involving only one cell type GM12878 (9). The analysis on replication timing domains also meets the resolution problem. The sub-domains of replication timing, like sub-TADs, may be observed in the future with technology development. The structural and functional differences between TAD and sub-TADs need further investigation with better technologies and algorithms.

TAD and sub-TAD can separate regulatory activity but with different insulation effect. TAD boundaries generally show stronger insulations than sub-TAD boundaries. And the transcriptional association of TAD mergence/split is a little different from that of sub-TAD mergence/split when performing domain-level comparisons. Combining the fact that TAD is more stable than sub-TAD across cell types, these results argue that TAD may insulate the global gene regulation in a relatively stable way and sub-TAD further facilitates local gene regulation in a dynamic way, consistent with previous work (22). With regard to the potential mechanism underlying hierarchical TADs, TAD and sub-TAD boundaries share similar trends in CTCF binding sites and divergent motifs, but with different density. This may partially explain the differences in insulation and stability between TAD boundaries and sub-TAD boundaries. However, recent research showed that transcription could contribute to the formation of chromatin domains by helping position another key architectural protein complex, cohesion (47). Further studies are needed to clarify the detailed relationship among transcription, CTCF binding and hierarchical TAD.

Finally, our method performs well in detecting TADs and sub-TADs, but it is less sensitive to detect higher level domains, especially level ≥3 domains. Figure 2 illustrates that it is currently difficult to sensitively and reproducibly detect higher level domains. This is because the chromatin interactions are highly variable around the boundary regions of these small domains. In HiTAD, the locally high variations and the small window sizes together can make the adaptive DIs fluctuate around candidate boundary regions, which leads to the failure in domain reproducibility. Though this strategy excludes irreproducible domains from two replicates, the percentage of reproducible domains still decreases in higher level domains when more than two replicates are used to measure the reproducibility (Figure 2B). In the future, better methods can be utilized or developed to balance the sensitivity and reproducibility in detecting higher level domains.

CONCLUSION

In this work, we developed a novel method HiTAD to detect hierarchical TADs from Hi-C chromatin interactions by further constraining TAD to optimal chromatin interaction separation in chromosomal level. HiTAD performs well in domain sensitivity, replicate reproducibility and inter-cell-type conservation. We evaluated the detected hierarchical TADs by calculating insulation effects and signal enrichments on different level boundaries. By defining boundary-level and domain-level changes for hierarchical TADs, we systematically investigated the structural and functional differences between TADs and sub-TADs. The intra cell-type and inter cell-type analyses together revealed that TADs and sub-TADs differed in correlating higher-order compartment, replication timing and transcriptional regulation. With better technology and algorithm, the structural and functional characteristics of hierarchical TADs can be further explored in the near future.

AVAILABILITY

HiTAD is integrated to a Python package called TADLib, which is freely available online at https://pypi.python.org/pypi/TADLib.

SUPPLEMENTARY DATA

Supplementary data are available at NAR Online.

FUNDING

Fundamental Research Funds for the Central Universities [2662015PY197, 2662016PY034]; National Natural Science Foundation of China [31200951]. Funding for open access charge: Fundamental Research Funds for the Central Universities [2662015PY197, 2662016PY034]; National Natural Science Foundation of China [31200951].

Conflict of interest statement. None declared.

REFERENCES

1. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.
2. Simonis,M., Klous,P., Splinter,E., Moshkin,Y., Willemsen,R., de Wit,E., van Steensel,B. and de Laat,W. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.*, **38**, 1348–1354.
3. Zhao,Z., Tavoosidana,G., Sjolinder,M., Gondor,A., Mariano,P., Wang,S., Kanduri,C., Lezcano,M., Sandhu,K.S., Singh,U. *et al.* (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.*, **38**, 1341–1347.
4. Dostie,J., Richmond,T.A., Arnaout,R.A., Selzer,R.R., Lee,W.L., Honan,T.A., Rubio,E.D., Krumm,A., Lamb,J., Nusbaum,C. *et al.* (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, **16**, 1299–1309.
5. Fullwood,M.J., Liu,M.H., Pan,Y.F., Liu,J., Xu,H., Mohamed,Y.B., Orlov,Y.L., Velkov,S., Ho,A., Mei,P.H. *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**, 58–64.
6. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
7. Kalhor,R., Tjong,H., Jayathilaka,N., Alber,F. and Chen,L. (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.*, **30**, 90–98.
8. Dryden,N.H., Broome,L.R., Dudbridge,F., Johnson,N., Orr,N., Schoenfelder,S., Nagano,T., Andrews,S., Wingett,S., Kozarewa,I. *et al.* (2014) Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res.*, **24**, 1854–1868.
9. 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.
10. 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.
11. 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.
12. Nora,E.P., Lajoie,B.R., Schulz,E.G., Giorgetti,L., Okamoto,I., Servant,N., Piolot,T., van Berkum,N.L., Meisig,J., Sedat,J. *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381–385.
13. Sanyal,A., Lajoie,B.R., Jain,G. and Dekker,J. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109–113.
14. Jin,F., Li,Y., Dixon,J.R., Selvaraj,S., Ye,Z., Lee,A.Y., Yen,C.A., Schmitt,A.D., Espinoza,C.A. and Ren,B. (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**, 290–294.
15. Pope,B.D., Ryba,T., Dileep,V., Yue,F., Wu,W., Denas,O., Vera,D.L., Wang,Y., Hansen,R.S., Canfield,T.K. *et al.* (2014) Topologically associating domains are stable units of replication-timing regulation. *Nature*, **515**, 402–405.
16. Franke,M., Ibrahim,D.M., Andrey,G., Schwarzer,W., Heinrich,V., Schopflin,R., Kraft,K., Kempfer,R., Jerkovic,I., Chan,W.L. *et al.* (2016) Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, **538**, 265–269.
17. Andrey,G., Montavon,T., Mascrez,B., Gonzalez,F., Noordermeer,D., Leleu,M., Trono,D., Spitz,F. and Duboule,D. (2013) A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science*, **340**, 1234167.
18. Lupianez,D.G., Kraft,K., Heinrich,V., Krawitz,P., Brancati,F., Klopocki,E., Horn,D., Kayserili,H., Opitz,J.M., Laxova,R. *et al.* (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, **161**, 1012–1025.
19. Flavahan,W.A., Drier,Y., Liau,B.B., Gillespie,S.M., Venteicher,A.S., Stemmer-Rachamimov,A.O., Suva,M.L. and Bernstein,B.E. (2016) Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, **529**, 110–114.
20. Hnisz,D., Weintraub,A.S., Day,D.S., Valton,A.L., Bak,R.O., Li,C.H., Goldmann,J., Lajoie,B.R., Fan,Z.P., Sigova,A.A. *et al.* (2016) Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, **351**, 1454–1458.
21. Bouwman,B.A. and de Laat,W. (2015) Getting the genome in shape: the formation of loops, domains and compartments. *Genome Biol.*, **16**, 154.
22. Phillips-Cremins,J.E., Sauria,M.E., Sanyal,A., Gerasimova,T.I., Lajoie,B.R., Bell,J.S., Ong,C.T., Hookway,T.A., Guo,C., Sun,Y. *et al.* (2013) Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, **153**, 1281–1295.
23. Wang,X.T., Dong,P.F., Zhang,H.Y. and Peng,C. (2015) Structural heterogeneity and functional diversity of topologically associating domains in mammalian genomes. *Nucleic Acids Res.*, **43**, 7237–7246.
24. Weinreb,C. and Raphael,B.J. (2016) Identification of hierarchical chromatin domains. *Bioinformatics*, **32**, 1601–1609.

25. Zhan,Y., Mariani,L., Barozzi,I., Schulz,E.G., Bluthgen,N., Stadler,M., Tiana,G. and Giorgetti,L. (2017) 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.*, **27**, 479–490.
26. Shavit,Y., Walker,B.J. and Lio,P. (2016) Hierarchical block matrices as efficient representations of chromosome topologies and their application for 3C data integration. *Bioinformatics*, **32**, 1121–1129.
27. Filippova,D., Patro,R., Duggal,G. and Kingsford,C. (2014) Identification of alternative topological domains in chromatin. *Algorithms Mol. Biol.*, **9**, 14.
28. Chen,J., Hero,A.O. 3rd and Rajapakse,I. (2016) Spectral identification of topological domains. *Bioinformatics*, **32**, 2151–2158.
29. Haddad,N., Vaillant,C. and Jost,D. (2017) IC-Finder: inferring robustly the hierarchical organization of chromatin folding. *Nucleic Acids Res.*, **45**, e81.
30. Selvaraj,S., R Dixon,J., Bansal,V. and Ren,B. (2013) Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.*, **31**, 1111–1118.
31. Bernstein,B.E., Birney,E., Dunham,I., Green,E.D., Gunter,C. and Snyder,M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
32. 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.
33. Yue,F., Cheng,Y., Breschi,A., Vierstra,J., Wu,W., Ryba,T., Sandstrom,R., Ma,Z., Davis,C., Pope,B.D. *et al.* (2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, **515**, 355–364.
34. Kheradpour,P. and Kellis,M. (2014) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.*, **42**, 2976–2987.
35. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
36. Gomez-Marin,C., Tena,J.J., Acemel,R.D., Lopez-Mayorga,M., Naranjo,S., de la Calle-Mustienes,E., Maeso,I., Beccari,L., Aneas,I., Vielmas,E. *et al.* (2015) Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 7542–7547.
37. 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.
38. Levy-Leduc,C., Delattre,M., Mary-Huard,T. and Robin,S. (2014) Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics*, **30**, i386–i392.
39. Shin,H., Shi,Y., Dai,C., Tjong,H., Gong,K., Alber,F. and Zhou,X.J. (2016) TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.*, **44**, e70.
40. Ong,C.T. and Corces,V.G. (2014) CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.*, **15**, 234–246.
41. Rudan,M.V., Barrington,C., Henderson,S., Ernst,C., Odom,D.T., Tanay,A. and Hadjur,S. (2015) Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.*, **10**, 1297–1309.
42. Tang,Z., Luo,O.J., Li,X., Zheng,M., Zhu,J.J., Szalaj,P., Trzaskoma,P., Magalska,A., Wlodarczyk,J., Ruszczycki,B. *et al.* (2015) CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell*, **163**, 1611–1627.
43. Guo,Y., Xu,Q., Canzio,D., Shou,J., Li,J., Gorkin,D.U., Jung,I., Wu,H., Zhai,Y., Tang,Y. *et al.* (2015) CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell*, **162**, 900–910.
44. Dixon,J.R., Jung,I., Selvaraj,S., Shen,Y., Antosiewicz-Bourget,J.E., Lee,A.Y., Ye,Z., Kim,A., Rajagopal,N., Xie,W. *et al.* (2015) Chromatin architecture reorganization during stem cell differentiation. *Nature*, **518**, 331–336.
45. Dixon,J.R., Gorkin,D.U. and Ren,B. (2016) Chromatin domains: the unit of chromosome organization. *Mol. Cell*, **62**, 668–680.
46. Dali,R. and Blanchette,M. (2017) A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Res.*, **45**, 2994–3005.
47. Busslinger,G.A., Stocsits,R.R., van der Lelij,P., Axelsson,E., Tedeschi,A., Galjart,N. and Peters,J.M. (2017) Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. *Nature*, **544**, 503–507.