Hierarchical Domain Structure Reveals the Divergence of Activity among TADs and Boundaries

Lin An[1], Tao Yang[1], Jiahao Yang[2], Johannes Nuebler[3], Qunhua Li[4*], Yu Zhang[4*]

[1]Bioinformatics and Genomics program, Pennsylvania State University, University Park, PA, [2]Tsinghua University, Beijing, China, [3]Massachusetts Institute of Technology, Cambridge, Massachusetts, [4] Department of Statistics, Pennsylvania State University, University Park, PA
*To whom correspondence should be addressed.

Email addresses:
LA: lua137@psu.edu; TY: txy146@psu.edu; JY: yangjh14@mails.tsinghua.edu.cn; JN: jnuebler@mit.edu; QL: qunhua.li@psu.edu; YZ: yzz2@psu.edu.

Abstract

Mammalian genomes are organized into different levels. As a fundamental structural unit, Topologically Associating Domains (TADs) play a key role in gene regulation. Recent studies showed that hierarchical structures are present in TADs. Precise identification of hierarchical TAD structures however remains a challenging task. We present OnTAD, an Optimal Nested TAD caller from Hi-C data, to identify hierarchical TADs. Systematical comparison with existing methods shows that OnTAD has significantly improved accuracy, reproducibility and running speed. OnTAD reveals new biological insights on the role of different TAD levels and boundary usage in gene regulation and the loop extrusion model. The OnTAD is available at: https://github.com/anlin00007/OnTAD

Background

Previous studies have shown that human genome is highly compacted and organized at different levels in nucleus, and the spatial organization of chromatin is essential for gene regulation [1]. Advanced sequencing technologies, such as 3C (Chromatin Conformation Capture), 4C, 5C, ChIA-PET, Hi-C and Hi-ChIP, have emerged to measure 3D chromatin structure at different resolutions [2]. Among them, Hi-C [3] obtains measurement of chromatin interaction frequency across the entire genome. It has been shown that some local regions in the human genome tend to interact most frequently, which are termed as ‘Topologically

Associating Domains' (TADs) [4]. It is shown that CTCF and cohesin proteins are usually enriched at TAD boundary regions to form isolated local environment [5]. It is further shown that TADs are relatively conserved across different cell-types and even species [5,6]. As a result, TADs are widely deemed as a basic architectural unit to study gene regulatory activities. To date, several computational methods have been developed to locate TADs in the genome. For example, Dixon et al. [5] uses 'Directionality Index' to estimate the shift of interaction direction towards upstream and downstream of each locus and identify boundaries of TADs. Other methods, such as TOPDOM [7] and Insulation Score [8], convert the TAD boundary finding problem to a local minimum identification problem by calculating average interaction frequency of surrounding regions at each locus.

While many earlier TAD calling methods treat TADs as a single structure, recent high-resolution studies have shown that TADs can form hierarchies, with sub-TADs nested within larger TADs [9–13]. Several recent TAD calling methods therefore aim to identify hierarchical TAD structures. For example, TADtree [9] uses a linear model to interpolate contact enrichment to distinguish inner TADs and outer TADs from local background. rGMAP [10] assumes that the interaction frequency in sub-TADs is different from larger TADs and applies a Gaussian Mixture model to identify both types of TADs. Arrowhead [11] uses heuristics approach to uncover corners of TADs at multiple sizes, thereby revealing hierarchical TAD structures. 3D-Net [12] utilizes a maximization of network modularity method to identify TADs at different levels. And finally, IC-Finder [13] uses a clustering method to reconstruct the TAD hierarchy.

The aforementioned methods brought new biological insights towards chromatin formation and their potential roles in gene regulation. Yet, there are still several limitations to be tackled. First, many methods use *ad hoc* thresholds to call hierarchical TADs at different levels, where the choice of thresholds is empirical and thus may not be applicable to other data. Secondly, most existing methods are sensitive to sequencing depth and mapping resolution of the Hi-C contact matrix. Thirdly, the long running time and large memory usage limit the utility of many TAD callers in high-resolution data. Finally, the performance of these TAD callers has not been sufficiently evaluated in real data, which adds to the difficulty in picking the right method for TAD calling [14,15]. There is a pressing need for a TAD calling method that can efficiently, accurately and robustly uncover hierarchical TAD structures in high-resolution Hi-C data, and the method needs to be justified for its performance in various realistic settings.

We present OnTAD, an Optimal Nested TAD caller to uncover hierarchical TAD structures from Hi-C data. Our approach first scans through the genome with a sliding window to identify candidate TAD boundaries, and then uses a recursive dynamic programming

algorithm to optimally assemble them into hierarchical TADs based on some scoring function. Systematic evaluation shows that OnTAD substantially outperforms existing TAD callers in terms of both TAD boundary identification and hierarchical TAD assembly. OnTAD also produces more reproducible results across biological replicates, at different resolutions, and in different sequencing depths than the existing methods. Using OnTAD, we uncovered some novel insights regarding the potential biological functions of TAD structures. For examples, we observed that active epigenetic states are substantially more enriched in inner TADs than in outer TADs. The boundaries of nested TADs have higher CTCF enrichment, active epigenetic states and gene expression than the boundaries of TADs without hierarchical structures. In addition, we observed significant asymmetry of TAD boundary sharing, which supports the asymmetric loop extrusion model [16]. Taken together, OnTAD enables creation of new hypotheses on the role of chromatin structures in gene regulation.

Result

The OnTAD algorithm

OnTAD takes a Hi-C contact matrix as the input and calls TADs in two steps. In the first step, the method finds candidate TAD boundaries using an adaptive local minimum search algorithm inspired by TOPDOM [7]. Specifically, it scans through the diagonal of the Hi-C matrix using a *W* by *W* diamond-shaped window (Figure 1a) and calculates the average contact frequency within each window. The locations at which the average contact frequency reaches a significant local minimum are identified as candidate TAD boundaries (see Methods). As the sizes of TADs are unknown, the method repeats the above steps using a series window-sizes for W= 1,2.,…,K, to uncover all possible boundaries for TADs in different sizes. The union of the candidate boundaries of all window sizes is used to assemble TADs in the next step (Figure 1b). Here, K depends on the resolution of the Hi-C matrix and the maximum TAD size that the user wants to call. For instance, for a 10kb resolution Hi-C matrix and a maximum TAD size of 2Mb, K=2000/10=200.

In the second step, OnTAD assembles TADs by selectively connecting pairs of candidate boundaries using a nested dynamic programming algorithm (see Methods). To form a TAD, OnTAD requires the mean contact frequency between the two boundaries to be greater than that of the surrounding area outside of the TAD by a user-defined margin. The nested dynamic programming algorithm first identifies the optimal partition of the genome that yields the largest possible TADs between candidate boundaries (Supplementary Figure 1), and then subTADs will

be called recursively within each identified TAD. This dynamic programming framework allows us to obtain the optimal solutions with respect to maximizing some score function (defined in Methods), producing a TAD organization that best fits to the observed Hi-C contact matrix.

Comparison with existing TAD calling methods

Many TAD calling methods have been developed to date, yet a comprehensive evaluation of their performance have not been reported partly due to lack of proper measures of accuracy and a gold standard data set [14,15]. Here, we compared OnTAD with four widely used TAD calling methods (DomainCaller [5], rGMAP [10], Arrowhead [11] and TADtree [9]) using the Hi-C data in GM12878 from Rao et al. [11]. We ran each method in the settings recommended in their manuals. All the evaluations below are based on genome-wide 10Kb Hi-C data, unless explicitly mentioned.

Accuracy of TAD boundary detection

We first evaluated the accuracy of TAD boundary detection. CTCF is known to be an essential architectural protein to form TAD structures [4]. Thus, a high concentration of CTCF signal is expected at the TAD boundaries. We computed the average CTCF ChIP-seq signal in the boundaries identified by each TAD calling method as well as the neighborhood regions. As shown in Figure 2a left panel, all methods showed enriched CTCF signal in the identified TAD boundaries over the surrounding regions (fold change > 1.2). Among them, OnTAD had the highest CTCF enrichment (mean signal 1.26X greater than that of the second highest method). We also computed the enrichment of cohesin proteins (rad21 and smc3), which are also key components in the formation of TADs [16]. Again, the boundaries identified by OnTAD showed a substantially higher enrichment than those identified by other methods (mean signal 1.19X and 1.10X greater than that of the second highest method, respectively) (Figure 2a middle and right panel). Taken together, the stronger enrichment of CTCF and cohesin suggest that OnTAD better predicts TAD boundaries than other methods.

Accuracy of TAD assembly

We next evaluated the accuracy of TAD calling. If TADs are accurately called, one would expect a high proportion of the variation in the contact frequencies over the Hi-C matrix, after accounting for distance dependence, is explained by TAD calls. We developed a metric called TAD-adjR^2, which is a modified version of the adjusted R^2 (see methods), to measure the proportion of Hi-C signal variation explained by TAD calls. Because the contact frequencies

decay over the genomic distance between the pair of interacting loci, we stratified the contacts by the genomic distance and calculated TAD-adjR^2 within each stratum. As shown in Figure 2b, OnTAD has a higher TAD-adjR^2 at most genomic distances (0-1.5Mb) than all the other methods (AUC: OnTAD: 0.70, Arrowhead: 0.67, DomainCaller: 0.64, rGMAP: 0.66 and TADtree: 0.57), indicating that OnTAD produces the best classification between TADs and non-TAD regions.

Reproducibility of TADs and boundaries

Another important criterion for TAD calling is the reproducibility of both TADs and their boundaries. To measure reproducibility of TAD boundaries, we used Jaccard index to calculate the agreement of boundaries (Figure 2c-e) between two TAD calling results. To measure reproducibility of TADs, we treated each TAD covered region as a cluster of bins in the genome, and then measured the agreement of cluster assignments between two TAD calling results using the adjusted rand index (Supplementary Figure 3a-c). We evaluated the reproducibility in three scenarios: 1) between biological replicates (GM12878, 10Kb) (Figure 2c, Supplementary Figure 3a); 2) across different resolutions (5Kb, 10Kb, 25Kb) (Figure 2d, Supplementary Figure 3b); and 3) at different sequencing depths (original sequencing depth versus 1/4, 1/8, 1/16 and 1/32 of the total number of reads) (Figure 2e, Supplementary Figure 3c). As shown, OnTAD had the highest TAD boundary reproducibility between biological replicates, at different sequencing depths and between 10Kb resolution and 25Kb resolution (t-test p-value < 0.01). Also, OnTAD had the highest TAD calling reproducibility across different Hi-C resolutions and at different sequencing depths (t-test p-value < 0.01).

Run time comparison

We recorded the run time of different methods. In all of our analysis, OnTAD ran notably faster than the other methods (Supplementary Table 1). For example, it took OnTAD 655 seconds to analyze 10Kb resolution data for the whole genome on the high performance computing cluster with Xeon E5-2680CPU and 72Gb RAM. It was 3X faster than Arrowhead, 24X faster than DomainCaller, 28X faster than rGMAP, and 263X faster than TADtree.

Hierarchical TADs are more active than singleton TADs

Among all TADs identified by OnTAD, a majority (94%) of them belong to hierarchical structures, and a small group of TADs do not contain or belong to any other TADs. We refer to the former

as ‘Hierarchical TADs’ and the latter as ‘singletons’ (Figure 3a). We hypothesized that these two types of TAD may have different regulatory potentials and examined their association with various epigenetic marks.

CTCF enrichment

We first compared the CTCF enrichment at the boundaries of the two types of TADs. We observed that the boundaries of hierarchical TADs are substantially more enriched with CTCF signal than singletons (Figure 3b) (mean CTCF signals are 3.61 and 2.45, respectively). To investigate what leads to the difference in CTCF enrichment, we compared the average number of CTCF sites at the boundaries of the two types of TADs. Indeed, the boundaries of hierarchical TADs had more CTCF sites than those of singletons (mean numbers of CTCF sites are 0.545 and 0.305, respectively).

Epigenetic profiles

It has been reported that chromatin interactions are strongly associated with local epigenetic profiles [5] [11] [17]. We thus expect to observe enrichment of active epigenetic states within TADs. To perform this analysis, we obtained the active epigenetic states from the 36-state IDEAS segmentation in 6 ENCODE cell types [18]. We assigned hierarchical TADs into five levels, with level one being the outermost TADs, level two being the subTADs nested under one layer of outer TAD, and level five being the subTADs nested under four or more layers of outer TADs, respectively. We observed that the active epigenetic states are increasingly enriched along the levels of TADs (p-value of correlation < 2.2e-16) (Figure 3c, d). In comparison, singletons are noticeably less active compared with hierarchical TADs (especially level >2). Similar pattern was also observed in other cell types (K562 and Huvec) (Supplementary Figure 4). Taken together, our results showed that hierarchical TADs are on average more active than singletons, and within hierarchical TADs, inner TADs (e.g., sub TADs) are more active than outer TADs.

Active genes in hierarchical TADs

We further investigated how gene expression is associated with TAD hierarchies. Using the RNA-seq data of GM12878 from ENCODE consortium (www.encodeproject.org), we calculated the density of the number of expressed genes (FPKM > 5) within TADs at each level, which is defined as the number of expressed genes per bin (i.e., 10Kb region). For genes covered by multiple TADs, we associated them with the innermost TADs. We found that as the TAD level

increases, the density of the number of expressed genes also increases, i.e., genes are more frequently activated within inner TADs (p-value of correlation < 2.2e-16) (Supplementary Figure 5a). The same trend of positive association between the number of expressed genes and the TAD level (p-value < 2.2e-16) was also observed in the K562 cell (Supplementary Figure 5b)

Shared TAD boundaries are asymmetric and more active than other boundaries

It has been reported that TAD boundaries are interaction hotspots and are highly active [19]. We also observed that, for TADs at all levels, the number of expressed genes and the enrichment of active epigenetic states are significantly higher at the TAD boundaries than at the internal regions of TADs (all t-test p-values < 0.001) (Figure 3e, Supplementary Figure 6). It thus warrants a separate characterization of TAD boundaries. While associated enhancer functions and cross cell-type conservation of TAD boundaries stratified by insulation strengths have been reported [20], functions of TAD boundaries stratified by hierarchical levels have not been done.

For hierarchical TADs, we observed that their boundaries are frequently shared by multiple TADs in a common hierarchical branch. Interestingly, for boundaries shared by more than three TADs on either side, we observed that the numbers of TADs on the two sides of the boundaries are significantly different, showing a preference of asymmetric hierarchical structures over symmetric structures (Supplementary Table 2, Chi-Square p-value < 1e-10). We hypothesized that the boundary usage may play an essential role in maintaining the hierarchical structures and regulating gene activities. To investigate this hypothesis, we classified boundaries into five categories, according to the maximum number of TADs sharing the boundary on either side (Figure 4a). A boundary is classified as level one if it is used by only one TAD on either side, or as level five if it is shared by five or more TADs on either side. The number of boundaries assigned to each category is shown in Supplementary Figure 7.

Epigenetic and genomic profiles

We checked the enrichment of active epigenetic states at different boundary levels. We noticed a significant positive linear correlation between the enrichment of active epigenetic states and the number of times each boundary is shared (e.g., p-value = 7.43e-05 for Tss, 0.00204 for TssCtcf and 0.001215 for Enh) (Figure 4b). We further studied the relationship between gene expression level and the boundary sharing. Again, we observed a significant positive linear correlation (p-value = 0.007) between the number of times a boundary is shared and the gene expression level. In particular, the gene expression level at the boundaries that are shared by 5 or more TADs is higher than other boundaries (Figure 4c). Therefore, we call the boundaries

shared by 5 or more TADs as super boundaries. We posit that super boundaries are more active than boundaries shared by one or two TADs.

The asymmetric loop extrusion model

We next asked if the observed asymmetry in boundary sharing is associated with the mechanism of loop formation. A recent study in yeast suggests that loops are formed in an asymmetric process with the cohesin anchors on one side and reels from another side [21]. One possibility of asymmetric boundary usage is that loop extrusion may be stopped by some mechanism related to CTCF or other architectural protein's binding orientation at the boundaries. In such case, stopping of loop extrusion will depend on genomic location, and when loop extrusion stops, it will result in boundary usage asymmetry on the two sides. This may also explain asymmetric loop extrusion observed in experiments. It is known that the anchor sites in yeast are supported by Ycg1 HEAT-repeat and Brn1 kleisin subunits [22]. In human, however, little is known about the proteins supporting the anchor sites. We therefore performed transcription factor (TF) enrichment analysis of 161 TF ChIP-seq data from ENCODE consortium at different levels of TAD boundaries. We found that many structural-related TFs, such as SIX5, CTCFL, HMGN3 and CHD1, are significantly and increasingly enriched across different levels of boundaries, especially in super boundaries (Figure 4d). These enriched TFs may play an important role in forming and maintaining hierarchical TADs.

Discussion

While hierarchical structures in TAD formation have been reported [9,10,12,13], it remains unclear how the TAD hierarchies are involved in gene regulation mechanisms. This is partly due to the lack of a TAD calling method that can systematically identify all TAD hierarchies from Hi-C data. Here we introduce OnTAD, a new method to uncover the hierarchical TAD structures from Hi-C data with substantially improved sensitivity and specificity than existing methods. OnTAD yields optimal solutions with respect to its scoring function, subject to the accuracy of the pre-selected sets of candidate TAD boundaries identified by its local minimum-searching algorithm. Our comprehensive evaluation shows that OnTAD substantially outperforms the existing tested TAD calling methods in accuracy, reproducibility and running speed. Importantly, OnTAD classifies TADs and TAD boundaries into different levels based on the identified hierarchies, thereby enabling systematic investigation of the interplay between hierarchical TADs and gene regulation.

Using OnTAD, we observed several novel insights associated with TAD structures and boundaries. In particular, hierarchical TADs are on average significantly more active than singleton TADs. The boundaries that are shared by multiple TADs, e.g., ‘super boundaries’, are also significantly more enriched in active epigenetic states and active genes than the boundaries used exclusively by one TAD. Intriguingly, the super boundaries showed a significant orientation imbalance of sharing, which is in concordance with the asymmetric loop extrusion model.

In addition, recent studies show a possible explanation that TADs are formed by the corporatingon of cis-acting factors (e.g., cohesin complex) and stalling factors (e.g., CTCF). We calculated linear correlation between cohesion density and interaction strength signal within TADs and showed a highly significant positive correlation (p-value < 1e-200). Similar analysis done on CTCF density also showed a significant positive correlation (p-value = 2.93e-197) (Supplementary Figure 8). It thus confirms the importance of cohesin proteins and CTCF in the formation of TADs and contacts reported in recent studies.

A current limitation of OnTAD is that it relies on the hierarchical TAD assumption, i.e., no two TADs can be partially overlapping with each other beyond the shared boundaries. This assumption is required for the dynamic programming to find an optimal solution in polynomial time. To investigate the hierarchical structure assumption, we ran OnTAD on high resolution (10Kb) in-situ Hi-C data in GM12878. We segregated regions around the corner of each TAD into four 5*5 quadrants and calculated the average contact frequency of each quadrant (Supplementary figure 9). If the hierarchical TAD assumption holds, we expect to observe high interaction frequency in the quadrant within TAD (quadrant 1). And at the same time, the two quadrants (2 & 3) on the sides of each TAD corner should not have high interaction frequency simultaneously. As shown in Supplementary Figure 9, the mean frequency patterns of the four quadrants for most of the TAD corners are consistent with our expectations, suggesting that the hierarchical TAD assumption holds for a majority of the genome. On the other hand, a remedial solution to the violation of such assumption is to remove the signals from the called TADs and then rerun OnTADs on the de-clumped HiC data to identify additional TADs.

Given the superior power, robustness and efficiency of OnTAD over the existing methods, our algorithm will be also useful for calling TADs across different cell types to uncover both shared and cell-type specific TADs. It will be particularly interesting to investigate if certain levels of hierarchical TAD structures may change across cell types, and how such changes are

associated with differential gene regulation. With systematical identification of TAD hierarchies by OnTAD, new biological insights can be generated towards understanding the chromatin role in gene regulation.

Method

Notations and data preprocessing

Let X denote a symmetric Hi-C matrix, where each entry (i,j) in the matrix is a value quantifying the chromatin interaction strength between bins i and j. Let $X[a{:}b, c{:}d] = \{(i,j): a \le i \le b, c \le j \le d\}$ denote a sub-matrix of X. A candidate TAD between bins a and b corresponds to a diagonal block matrix $X_{[a,b]}=X[a{:}b, a{:}b]$, where the mean of the entries in $X_{[a,b]}$ is expected to be higher than that in its neighboring matrices. Because of the distance dependency in Hi-C data, i.e., the dependence of contact frequency on the proximity of the interaction loci, we normalize the Hi-C matrix before TAD calling by subtracting the mean counts at each distance from the original contact matrix.

Recursive TAD calling algorithm

We develop a TAD calling algorithm to assemble TADs from the candidate boundaries. Several issues need to be considered in the design of the algorithm in order to produce biologically meaningful TADs. First, because a region may be shared by multiple candidate TADs, the scores of these TADs can be strongly correlated. Second, in the TADs with nested structures, the scores of the TADs and their nested sub-TADs are convoluted. Third, some boundaries may be shared between TADs due to the biological mechanisms of loop formation. Last, the algorithm needs to be computationally efficient to call TADs in the genome scale.

To address these issues, we developed a recursive algorithm to identify the set of TADs that gives the optimal partition of the genome according to a scoring function g(X) (see the next section). Our algorithm assumes that any given two TADs are either completely non-overlapping (but can share boundaries) or completely overlapping (i.e. one TAD is nested within the other). While this assumption sometimes may not be true, it greatly reduces the complexity of the problem while still enabling us to 1) de-convolute nested TAD structures, 2) impose shared boundaries, and 3) obtain an efficient algorithmic solution. When it is violated, i.e., the boundaries of the TADs cross with each other, our method can still produce a reasonable approximation (Supplementary Fig.1C).

Briefly, the algorithm works as follows. Given a matrix $X_{[a,b]}$, the algorithm starts at the root level to first find the best bin i ($a \le i < b$) to partition the matrix into two submatrices, $X_{[a,i]}$ and $X_{[i,b]}$, such that $X_{[i,b]}$ is the largest right-most TAD in $X_{[a,b]}$. Since $X_{[a,i]}$ and $X_{[i,b]}$ are disjointed, the TADs within each submatrix can be called separately in a recursive manner. At each recursive step, the parent matrix is partitioned into two sub-matrices, and TADs are called within each sub-matrix using the same recursive formula (Supplementary Fig.2A). The recursion stops when $i=a$, i.e., the sub-matrix $X_{[a,i]}$ contains no TAD. After a recursive step is completed, it identifies the best TADs in the current branch according to the scoring function, de-convolutes the TAD signals in the parent matrix by removing signals of inner TADs, and evaluates if the parent matrix itself is a TAD. This process is repeated until the recursion returns to the root level (Supplementary Fig.2B). Note that, because every TAD is the largest right-most TAD of some parent matrix in a recursive branch, this recursive procedure guarantees to traverse all TADs, even though only the largest right-most TAD is called at each step.

The scoring function

Our scoring function $g(X_{[a,b]})$ for matrix $X_{[a,b]}$ is defined as

$$g(X_{[a,b]}) = max_i \begin{cases} 0 & i = a \\ \max\left(0, g(X_{[a,i]}) + h(X_{[i,b]})\right) & i = a+1, \dots, b-1 \end{cases}$$

where $h(X_{[i,b]}) = g(X_{[i,b]}) + \Delta(X_{[i,b]}|\text{sub TADs})$

Here, $g(X_{[a,b]})$ is the score of TADs within $X_{[a,b]}$, not including the score for $X_{[a,b]}$ itself being a TAD. It is calculated by finding the best left boundary of the largest right-most TAD in $X_{[a,b]}$. $h(X_{[i,b]})$ is the score of the largest right-most TAD in $X_{[a,b]}$. It is the sum of the score of TADs within $X_{[i,b]}$ and the score of $X_{[i,b]}$ itself being a TAD, namely $\Delta(X_{[i,b]}|\text{sub TADs})$. For any diagonal block matrix to be called a TAD, its mean signal is required to be greater than the means of its neighboring regions on both sides. We therefore define

$$\Delta(X_{[i,b]}|sub\ \text{TADs}) = m(X_{[i,b]}|\text{sub TADS}) - \max\left(\overline{X[i{:}b\text{-}(b\text{-}i\text{+}1),i{:}b]}, \overline{X[i{:}b,i{:}b\text{+}(b\text{-}i\text{+}1)]}\right) - \lambda$$

where $m(X_{[i,b]}|\text{sub TADs})$ denotes the mean of $X_{[i,b]}$, excluding the TADs within $X_{[i,b]}$, as returned by the recursion; λ denotes a positive penalty parameter; $X[i{:}b\text{-}(b\text{-}i\text{+}1),i{:}b]$ and $X[i{:}b\text{-}(b\text{-}i\text{+}1),i{:}b]$ are equal-sized off-diagonal matrices in the adjacent flanking regions of $X_{[i,b]}$; and finally, $\bar{X}$ denotes the mean of X. We note that $\Delta(X_{[i,b]}|sub\ \text{TADs})$ is calculated based on the TADs returned from $g(X_{[i,b]})$. That is, we do not directly optimize $g(X_{[i,b]}) + \Delta(X_{[i,b]}|sub\ \text{TADs})$.

When the score of a candidate TAD is <0, it is likely not a real TAD. We therefore set a lower bound on the score at 0 and do not output the "TAD" with a score 0.

Computation complexity of the TAD calling algorithm

We performed an analysis on the computational complexity for our recursive algorithm. For a lxl Hi-C matrix, if all bins are potential boundaries, then the recursion needs to visit $l(l+1)/2$ diagonal block sub-matrices. As there are l size 1 diagonal block matrices, the computation complexity for computing the scores of all size 1 matrices is $O(l)$. Given the scores of size 1 matrices, we can calculate the scores of size 2 matrices. There are (l-1) of them, each enumerating through (2-1) partitions. Hence the time complexity is O((2-1)(l-1)). Following the same calculation, the scores of one sub-matrix of size k will be computed by enumerating (k-1) partitions. As there are (l-k+1) of them, the time complexity is O((k-1)(l-k+1)). Similar calculation can be done for the mean of sub-matrices. As a result, the total complexity to obtain the scores of all sub-matrices from size 1 to l is $O(l^3)$.

Empirically, the computational complexity is much lower than the above due to some further reductions. First, because potential TAD boundaries are limited to the TOPDOM local minimums, this substantially reduces the number of partitions from $O(l^3)$ to $O(m^3)$, where m is the number of candidate boundaries. Second, because TADs usually are smaller than 2Mb, the maximum TAD size to be called (d) typically is much smaller than l. This constraint effectively reduces the time complexity of our algorithm from $O(m^3)$ to $O(md^2)$. Furthermore, because TADs usually are formed between neighboring boundaries, we set a constraint in the recursive procedure to limit the TADs to be formed only between candidate boundaries that are no more than five neighbors apart.

Identification of candidate TAD boundaries

We identify candidate TAD boundaries by finding the bins with the local minimum TOPDOM statistics [7]. The TOPDOM statistics is the mean Hi-C signal of a square submatrix with one corner of the sub matrix touching the diagonal of the Hi-C matrix. The bin touched by the submatrix is a likely TAD boundary when the mean of the submatrix is at local minimum, as the latter indicates that the submatrix does not overlap with any TADs. The original TOPDOM paper only computed the statistics at a fixed window size. To identify all candidate TAD boundaries for TADs in different sizes, we calculated the TOPDOM statistics at all window sizes ranging from 1 to a maximum TAD size (d) specified by users. Here, we used d=2Mb.

For each window size, we obtained a vector of the local minimum of the TOPDOM statistic on the genome. A bin i is identified as a local minimum if its TOPDOM statistics is the smallest value in the neighborhood of [i-5, i+5] and is smaller than the largest statistics in the neighborhood by at least 1.96S, where S is the standard deviation of the TOPDOM statistic in the entire matrix. Figure 1b shows examples of the local minimums on the genome. Since local minimums at different window sizes capture the information of TADs in different sizes, we merged local minimums across different window sizes and used the corresponding bins as candidate TAD boundaries.

TAD-adjR2 for assessing accuracy of TAD calling

Because TADs are regions with frequent local interactions, a reasonable TAD caller is expected to classify the regions with high contact frequencies as TADs and the regions with low contact frequencies as non-TADs, i.e. gaps between TADs. At any given genomic distance, the variation between Hi-C signals should be largely explained by the classification of TADs. How well the variation can be explained by the classification of TADs can reflect the accuracy of TAD calling. Based on this intuition, we developed a metric similar to R-square in regression models to evaluate the accuracy of TAD calling. Let *Y* denote the contact frequency at a given genomic distance, *n* denote the number of bins at this distance, *p* denote the number of called TADs whose sizes are greater or equal to the genomic distance, $\hat{Y}$ denote the average contact frequency within each TAD and in the background region, and $\acute{Y}$ denote the overall mean contact frequency. For each genomic distance, the TAD-adjR^2 is defined as

$$\hat{R}^2_{tad} = 1 - \frac{\frac{1}{n-p-1}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \acute{Y}\right)^2}$$

This quantity essentially measures the proportion of variance in Hi-C signal that is explained by the classification of TADs, adjusting for the number of TADs and genomic distance.

Enrichment of expressed genes

We merged biological replicates and computed the average FPKM for each gene. Genes with FPKM > 5 are deemed as expressed genes. For each TAD level, we compute the density of expressed gene as the number of expressed genes per 10Kb. For TADs with nested structures, genes covered by the inner level TADs are excluded in the calculation of gene density for outer TADs.

Epigenetic state enrichment

We used the 36 epigenetic states segmentation identified by IDEAS [18]. Let n_i denote the total number of 200bp windows that have IDEAS-assigned epigenetic states, and $n_{s,i}$ denote the number of 200bp windows annotated as state *s* at a TAD boundary *i*. For a given state *s,* its enrichment in a set of *M* boundaries is computed as

$$E(s) = \frac{\sum_{i=1}^{M} n_{s,i} + 1}{B_s \sum_{i=1}^{M} n_i + 1}$$

where B_s is the proportion of state *s* in the whole genome. The 1's in the formula of E(s) are added to avoid dividing by 0.

Data

Hi-C data: The Hi-C data is obtained from Rao et al. 2014 (GEO accession number: GSE63525). Among them, three human cell types (B-lymphoblastoid cells (GM12878), umbilical vein endothelial cells (HUVEC) and erythrocytic leukaemia cells (K562)) were included in this study. The Hi-C matrices mapped at 5Kb, 10Kb and 25Kb resolutions were used in this study.

Epigenomic data: The histone modification and gene expression data were downloaded from the NIH Roadmap Epigenomics project (http://www.roadmapepigenomics.org/), including H2A.Z, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me3 and H4K20me1. The ChIP-seq data of CTCF and cohesin protein (Rad21 and Smc3) were downloaded from ENCODE project (https://www.encodeproject.org/). The downloaded data were in BigWig format. The 'bigWigAverageOverBed' was used to segment signal into windows according to the resolution of Hi-C data.

Epigenetic states: The IDEAS segmentation of the 6 ENCODE cell type/tissues (GM12878, H1h-ESC, Hela-S3, HepG2, HUVEC, K562) was downloaded from (http://main.genome-browser.bx.psu.edu/). The 36-state IDEAS model trained on 10 marks (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3, H3K36me3, H3K20me1, PolII and CTCF), as well as DNase-seq and Faire-seq, was applied to this study.

Figure Legends

Figure 1 | Overview of the OnTAD pipeline. a, OnTAD uses a sliding diamond-shaped window to calculate the average contact frequency within the window at each locus on the genome. The five loci marked by letters 'a'-'e' are examples being evaluated as potential TAD boundaries, with 'd' being a clear false positive. **b,** The calculated average contact frequency from diamond-shaped windows, at different window sizes (W), show local minimums (red arrow) that are indicating potential TAD boundaries. **c,** OnTAD optimally assembles possible boundary pairs with Dynamic Programming (see methods) **d,** Visualization of final output from OnTAD.

Figure 2 | Evaluation of TAD calling methods. a, Average ChIP-Seq signal at TAD boundaries and surrounding regions (+/- 10 bins) (from left to right, CTCF, Smc3 and Rad21). **b,** Proportions of Hi-C signal variability explained by the called TADs (measured by TAD-adjR2) as a function of genomic distance between two interacting loci. (area under the curve: OnTAD: 0.37, Arrowhead: 0.3, DomainCaller: 0.25, rGMAP:0.26 and TADtree: 0.11) **c,d&e,** Reproducibility of TAD boundaries (Jaccarrd index): **c,** between two biological replicates (GM12878, 10Kb) **d,** across multiple resolutions (5Kb vs 10Kb) and **e,** across different down sampled sequencing depths (GM12878, original vs 1/4, 1/8, 1/16 and 1/32 of original sequencing depth).

Figure 3 | Hierarchical TADs are more active than singletons. a, an illustration of hierarchical levels of TADs. The levels are assigned from external to internal. The TADs covered by cyan dash line are assigned to level 1, by blue dash line are assigned to level 2, by orange dash line are assigned to level 3, and singletons are also assigned to level 1 (cyan). **b,** mean CTCF signal at the boundaries specific to hierarchical TADs (yellow), specific to singletons (cyan), and shared by both (green). The boundaries of hierarchical TADs have the highest enrichment of CTCF signal. **c & d,** Enrichment of epigenetic states at the boundaries of different levels of TADs. The enrichment (y-axis, log2) of active states increases as the TAD level increases (x-axis). The gap genomic regions not covered by any TADs are used as the background for calculating enrichments. **e,** Distribution of RNA-seq signal (FPKM) at the boundaries (blue) and within TADs (red).

Figure 4 | Super Boundaries are highly active. a, an illustration of the TAD boundary levels. The boundary levels are defined as the maximum number of TADs on either side that used this boundary. The yellow dots refer to level 1 boundaries as they are shared by a maximum of one

TAD on either side. The purple dots refer to level 2 boundaries as they are shared by two TADs on either side. The red dot refers to a level 3 boundary by this logic. **b,** Enrichment of epigenetic states at different levels of TAD boundaries. Super boundaries (shared by >5 TADs) are significantly enriched with Tss related states than others. **c,** Distribution of expression levels of genes whose transcription start sites are overlapping with different levels of TAD boundaries. **d,** potential TFs recruited at super boundaries. The enrichment of ChIP-seq TF peaks at boundaries against genome-wide background is shown in log2 scale. The super boundaries are significantly enriched with TFs that have chromatin structure related functions (marked with red boxes).

Supplementary figure 1 | Illustration of convoluted TAD structures. a, Candidate TADs (a,c) and (b,d) are both suboptimal, as their scores may be driven by a real TAD (b,c). **b,** Two real TADs (a,c) and (b,c) are nested, which makes the score of (a,c) convoluted with the score of (b,c). **c,** Real TADs (a,c) and (b,d) are partially overlapping, which may be recaptured as nested TADs (b,c), (a,c) and (a,d).

Supplementary figure 2 | Illustration of the recursive TAD calling algorithm. a, starting from the entire matrix, we partition the matrix into two matrices: the one that potentially forms the largest right-most TAD (triangles marked in black), and the remaining part. We call the same function on each sub-matrix to recursively identify nested TAD structures, and the best partition is determined by a scoring function. **b,** each recursion step identifies the best set of TADs in its matrix under consideration, and return the TAD calls back to its parent until the root.

Supplementary figure 3 | TAD reproducibility under different measurements. a, Adjusted rand index between TADs from two biological replicates (GM12878, 10Kb). **b**, Adjusted rand index across TADs from Hi-C data in multiple resolutions (GM12878, 5Kb, 10Kb and 25Kb). We excluded TADtree as it took too much computing resources on high resolution data. **c**, Adjusted rand index between TADs from Hi-C data in original sequencing depth and in different down sampled sequencing depth (GM12878, 1/4, 1/8, 1/16 and 1/32 of original sequencing depth). All comparisons are calculated on autochromosomes individually (excluding chr1 and chr9 due to no results produced by some methods).

Supplementary figure 4 | **Enrichment of epigenetic states at the boundaries of different levels of TADs.** Same method was used with figure 4c. **a,** K562 **b,** Huvec

Supplementary figure 5 | Density of the number of expressed genes in different levels of TADs a, GM12878 **b,** K562

Supplementary figure 6 | **Enrichment of active epigenetic states at the TAD boundaries (solid line) versus inside TADs (dashed line).** Y-axis denotes fold enrichment of three active epigenetic states (Tss, Enh and PromCtcf). X-axis denotes the boundaries and TADs at different levels.

Supplementary figure 7 | **Number of TAD boundaries in each level.** The level is defined as the maximum number of TADs on either side that share this boundary.

Supplementary figure 8 | Cohesin signal strongly correlates with mean contact frequency in TADs. a, scatter plot for TAD mean signal (x-axis) versus Cohesin signal (y-axis) in all TADs. **b,** scatter plot for TAD mean signal versus CTCF signal in all TADs.

Supplementary figure 9 | Contact frequency is unbalanced on the two sides of hierarchical TAD corners. The regions around TAD corners are segregated into four quadrants (1-4 on the top right figure). We then averaged contact frequency of each TAD corner by quadrants. As shown in the heatmap, quadrant 1 has highest average contact frequency as it is within TADs. Meanwhile, majority quadrant 2 and 3 shows unequal average contact frequencies, suggesting the outer TADs tend to form on one side of inner TADs rather than both sides.

Supplementary Table 1 | Comparison of running time of different methods on high resolution Hi-C data (GM12878 10Kb).

Supplementary Table 2 | Number of TADs on each side of a boundary that share this boundary (GM12878 10Kb).

Declarations

Acknowledgments

This work was supported by the NIH R01 GM121613 and NIH R24 DK106766, NIH training grant T32 GM102057 (CBIOS training program to The Pennsylvania State University), a Huck Graduate Research Innovation Grant, and by NIH grants R01GM109453.

Availability of data and materials

OnTAD is available at https://github.com/anlin00007/OnTAD.

Authors' contributions

YZ and LA implemented OnTAD. YZ and QL conceived the method. LA, TY, and JY conducted the analysis. LA, YZ, QL and TY wrote the manuscript with assistance from the other authors. JN assisted the interpretation of the results. All authors read and approved the final manuscript.

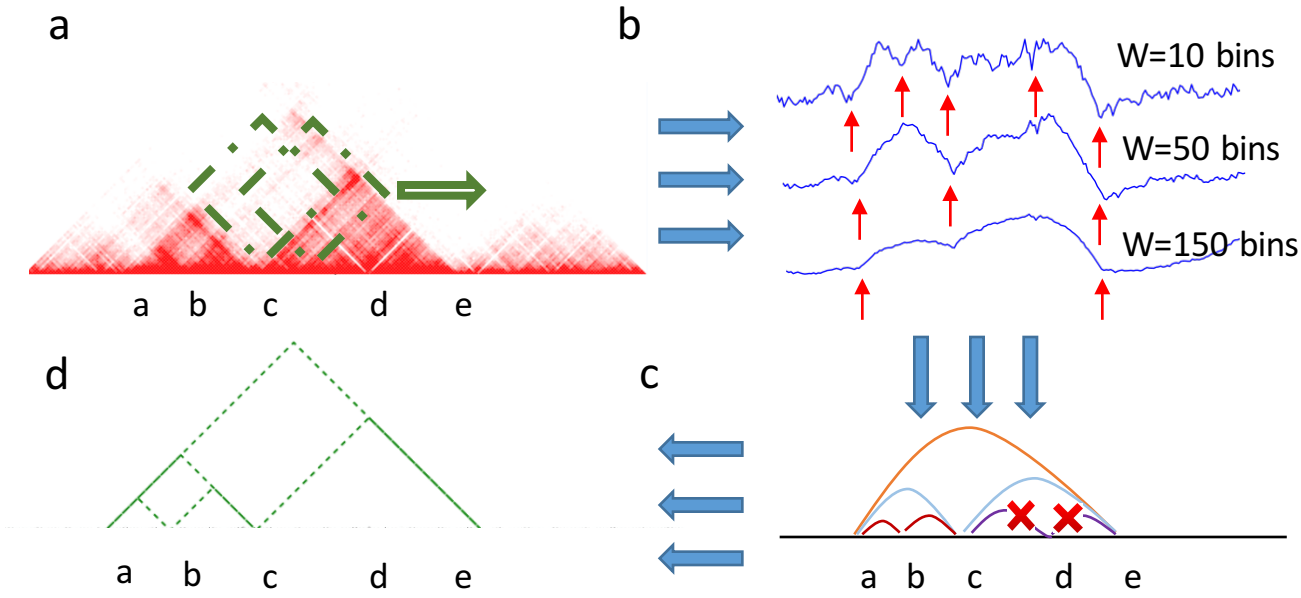
Competing interests

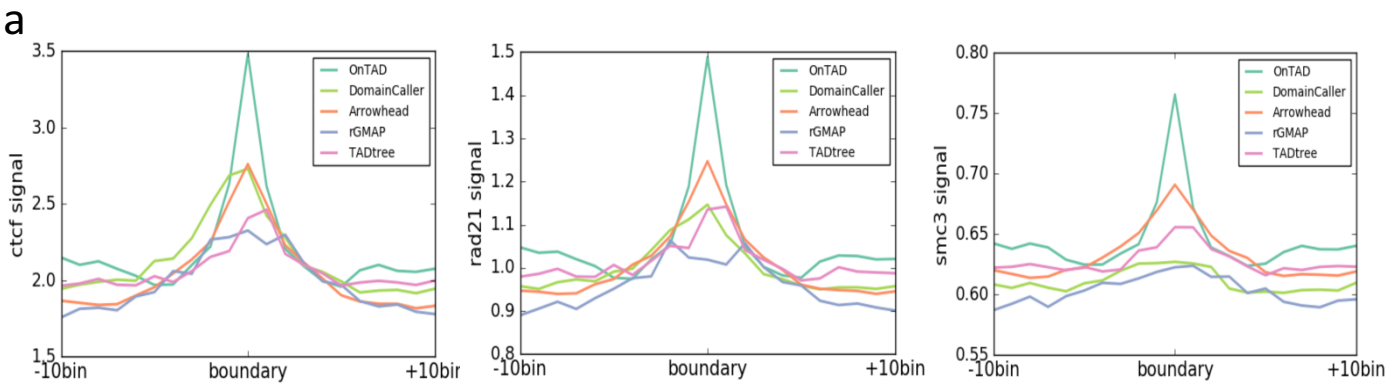
The authors declare that they have no competing interests.

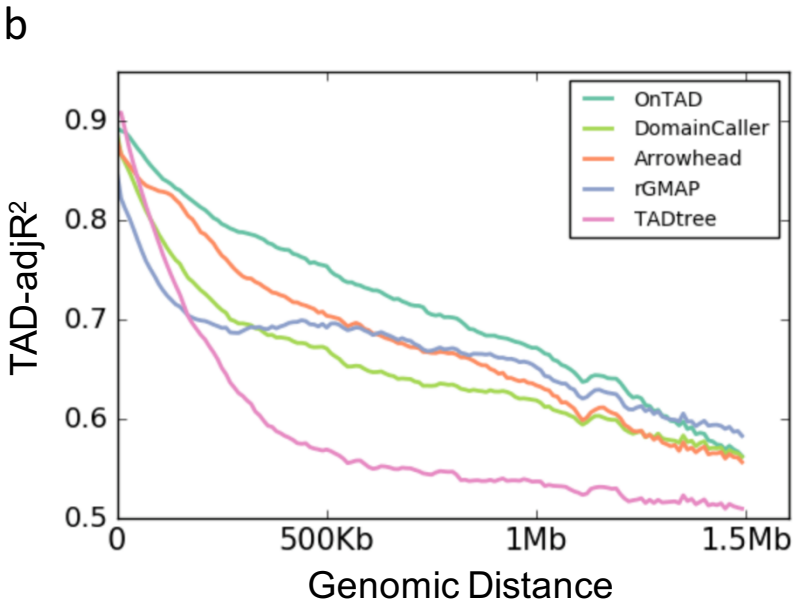
Reference

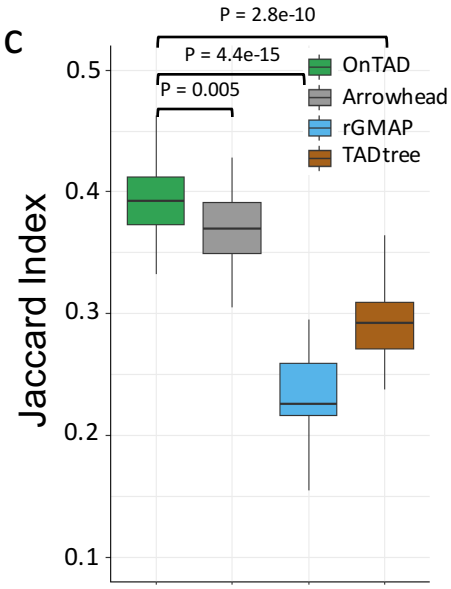
1. Won H, de la Torre-Ubieta L, Stein JL, Parikshak NN, Huang J, Opland CK, et al. Chromosome conformation elucidates regulatory relationships in developing human brain. Nature [Internet]. Nature Publishing Group; 2016;538:523–7. Available from: http://dx.doi.org/10.1038/nature19847%5Cnhttp://10.1038/nature19847%5Cnhttp://www.nature.com/nature/journal/v538/n7626/abs/nature19847.html#supplementary-information
2. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet [Internet]. 2013;14:390–403. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3874835/
3. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science (80-) [Internet]. 2009;326:289–93. Available from: http://science.sciencemag.org/content/326/5950/289.abstract
4. Dixon JR, Gorkin DU, Ren B. Chromatin Domains: The Unit of Chromosome Organization. Mol Cell [Internet]. Elsevier Inc.; 2016;62:668–80. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1097276516301812
5. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. Nature [Internet]. 2012;485:376–80. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3356448/
6. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, et al. Chromatin Architecture Reorganization during Stem Cell Differentiation. Nature [Internet]. 2015;518:331–6. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4515363/
7. Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, et al. TopDom: An efficient and deterministic method for identifying topological domains in genomes. Nucleic Acids Res. 2015;44:1–13.
8. Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, et al. Condensin-Driven Remodeling of X-Chromosome Topology during Dosage Compensation. Nature [Internet]. 2015;523:240–4. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4498965/

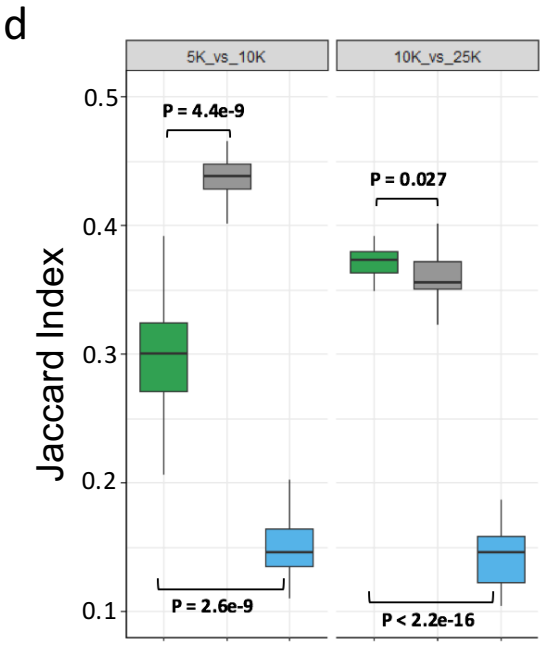
9. Weinreb C, Raphael BJ. Identification of hierarchical chromatin domains. Bioinformatics. 2016;32:1601–9.
10. Yu W, He B, Tan K. Identifying topologically associating domains and subdomains by Gaussian Mixture model And Proportion test. Nat Commun [Internet]. Springer US; 2017;8:535. Available from: http://www.nature.com/articles/s41467-017-00478-8
11. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. Elsevier; 2015;159:1665–80.
12. Norton HK, Emerson DJ, Huang H, Kim J, Titus KR, Gu S, et al. Detecting hierarchical genome folding with network modularity. Nat Methods. 2018;15:119–22.
13. Haddad N, Vaillant C, Jost D. IC-Finder: Inferring robustly the hierarchical organization of chromatin folding. Nucleic Acids Res. 2017;45.
14. Dali R, Blanchette M. A critical assessment of topologically associating domain prediction tools. Nucleic Acids Res [Internet]. Oxford University Press; 2017;45:2994–3005. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5389712/
15. Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. Comparison of computational methods for Hi-C data analysis. Nat Publ Gr [Internet]. Nature Publishing Group; 2017;14:14–9. Available from: http://dx.doi.org/10.1038/nmeth.4325
16. Sanborn AL, Rao SSP, Huang S-C, Durand NC, Huntley MH, Jewett AI, et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc Natl Acad Sci [Internet]. 2015;112:201518552. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4664323&tool=pmcentrez&rendertype=abstract
17. Ho JWK, Jung YL, Liu T, Alver BH, Lee S, Ikegami K, et al. Comparative analysis of metazoan chromatin organization. Nature [Internet]. The Author(s); 2014;512:449. Available from: http://dx.doi.org/10.1038/nature13415
18. Zhang Y, An L, Yue F, Hardison RC. Jointly characterizing epigenetic dynamics across multiple human cell types. Nucleic Acids Res [Internet]. 2016;gkw278-. Available from: http://nar.oxfordjournals.org/content/early/2016/04/29/nar.gkw278.full
19. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Rep [Internet]. The Authors; 2016;17:2042–59. Available from: http://dx.doi.org/10.1016/j.celrep.2016.10.061
20. Gong Y, Lazaris C, Sakellaropoulos T, Lozano A, Kambadur P, Ntziachristos P, et al. Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. Nat Commun. Nature Publishing Group; 2018;9:542.
21. Ganji M, Shaltiel IA, Bisht S, Kim E, Kalichava A, Haering CH, et al. Real-time imaging of DNA loop extrusion by condensin. Science (80-) [Internet]. 2018; Available from: http://science.sciencemag.org/content/early/2018/02/21/science.aar7831.abstract
22. Kschonsak M, Merkel F, Bisht S, Metz J, Rybin V, Hassler M, et al. Structural Basis for a Safety-Belt Mechanism That Anchors Condensin to Chromosomes. Cell [Internet]. Elsevier; 2018;171:588–600.e24. Available from: http://dx.doi.org/10.1016/j.cell.2017.09.008

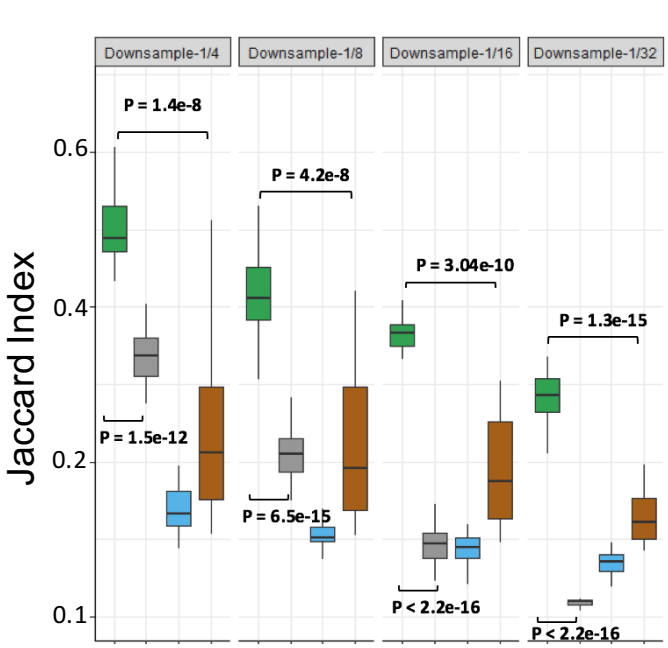

a
b
W=10 bins
W=50 bins
W=150 bins
a b c d e
d
c
a b c d e

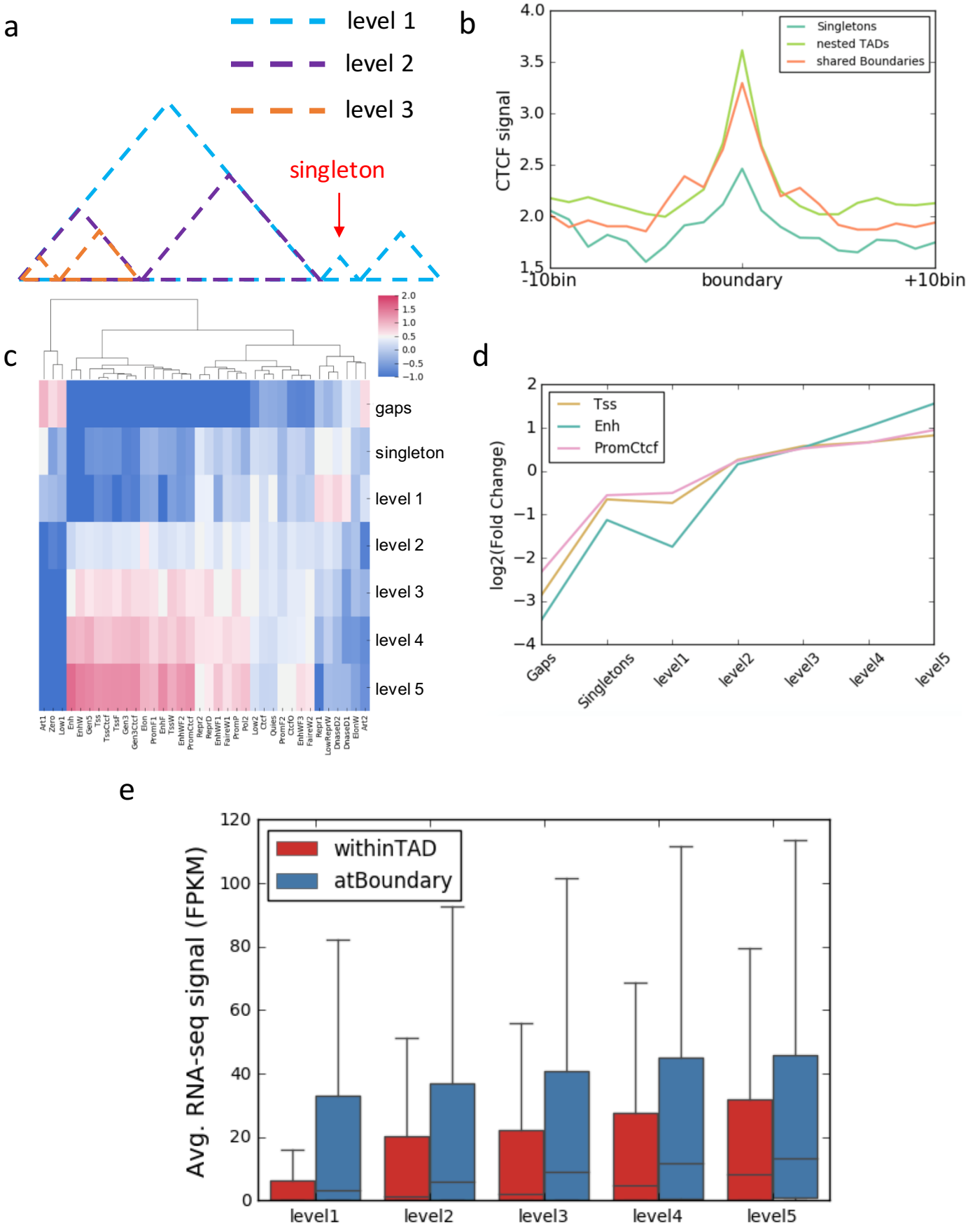

a
ctcf signal
rad21 signal
smc3 signal
-10bin
boundary
+10bin
OnTAD
DomainCaller
Arrowhead
rGMAP
TADtree

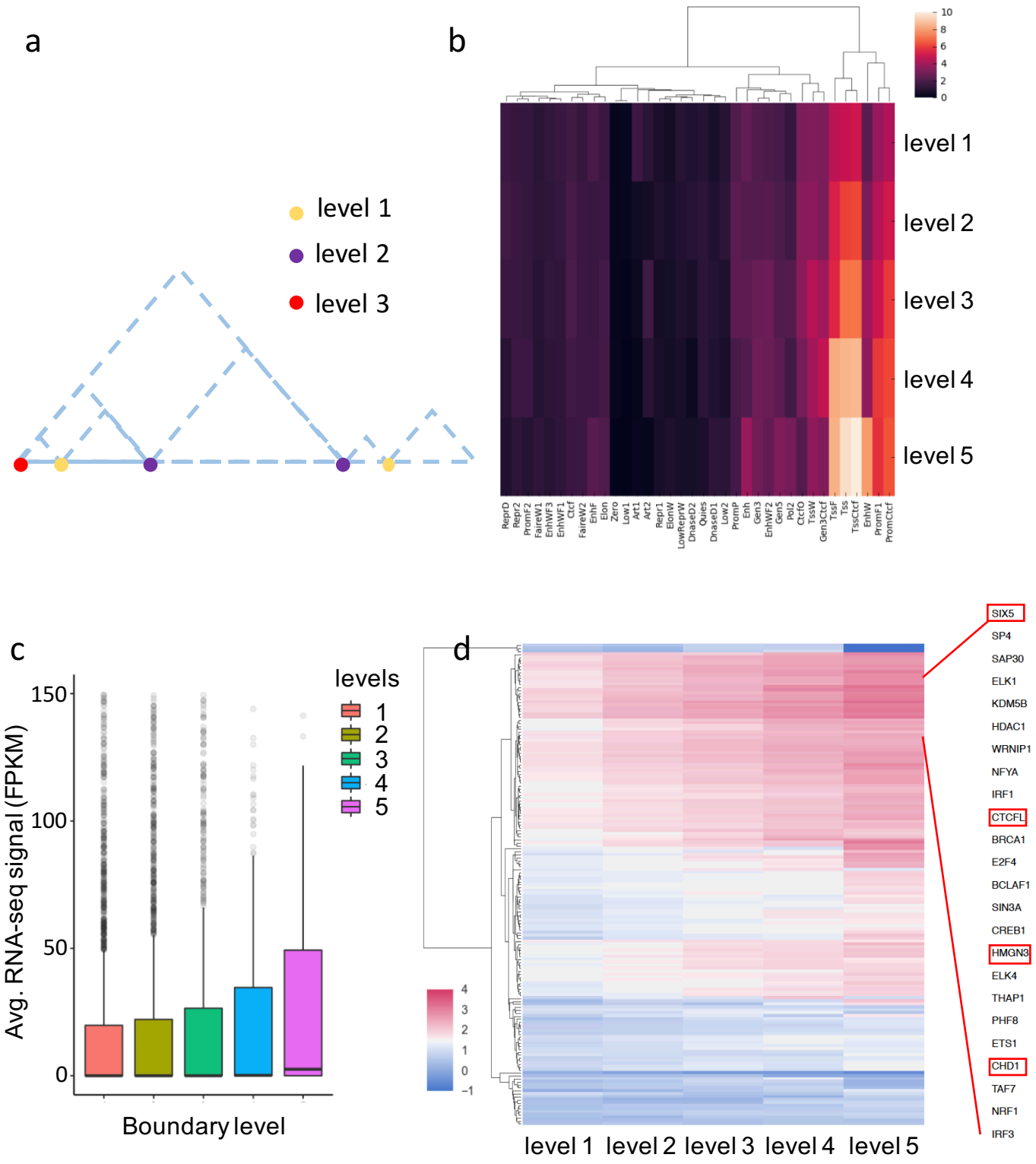

b
TAD-adjR²
Genomic Distance
0
500Kb
1Mb
1.5Mb
OnTAD
DomainCaller
Arrowhead
rGMAP
TADtree

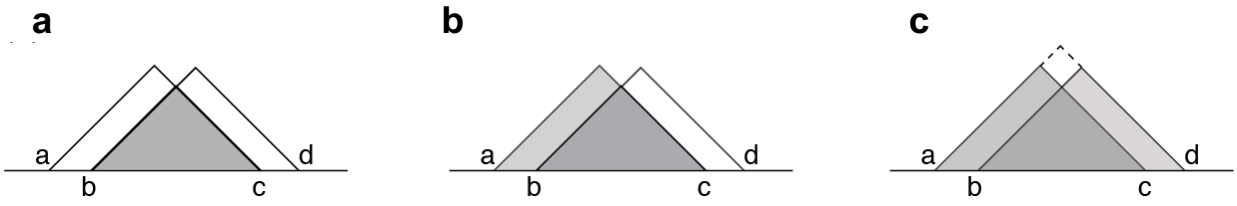

c
Jaccard Index
P = 2.8e-10
P = 4.4e-15
P = 0.005
OnTAD
Arrowhead
rGMAP
TADtree


d
5K_vs_10K
10K_vs_25K
Jaccard Index
P = 4.4e-9
P = 2.6e-9
P = 0.027
P < 2.2e-16

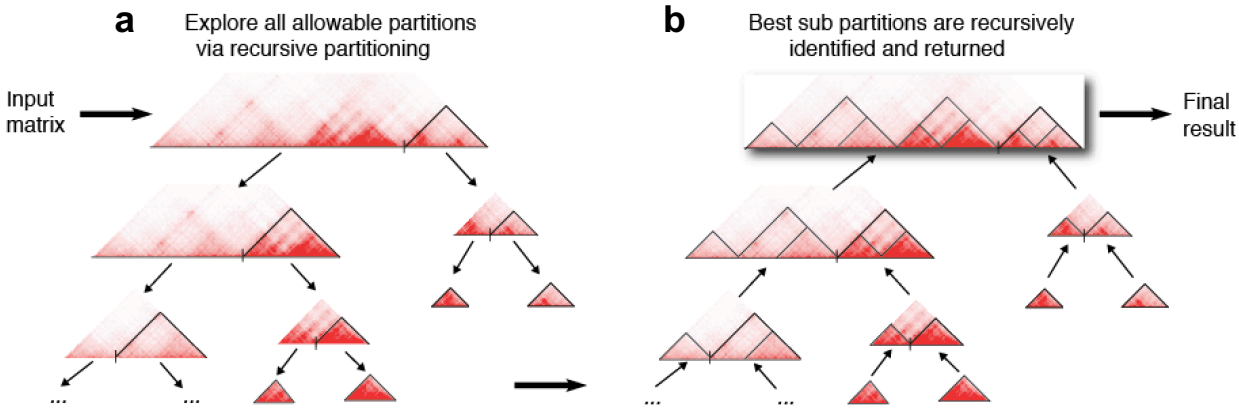

e
Downsample-1/4
Downsample-1/8
Downsample-1/16
Downsample-1/32
Jaccard Index
P = 1.4e-8
P = 1.5e-12
P = 4.2e-8
P = 6.5e-15
P = 3.04e-10
P < 2.2e-16
P = 1.3e-15
P < 2.2e-16


a
level 1
level 2
level 3
singleton
b
Singletons
nested TADs
shared Boundaries
CTCF signal
4.0
3.5
3.0
2.5
2.0
1.5
-10bin
boundary
+10bin
c
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
gaps
singleton
level 1
level 2
level 3
level 4
level 5
d
Tss
Enh
PromCtcf
log2(Fold Change)
2
1
0
−1
−2
−3
−4
Gaps
Singletons
level1
level2
level3
level4
level5
e
withinTAD
atBoundary
Avg. RNA-seq signal (FPKM)
120
100
80
60
40
20
0
level1
level2
level3
level4
level5

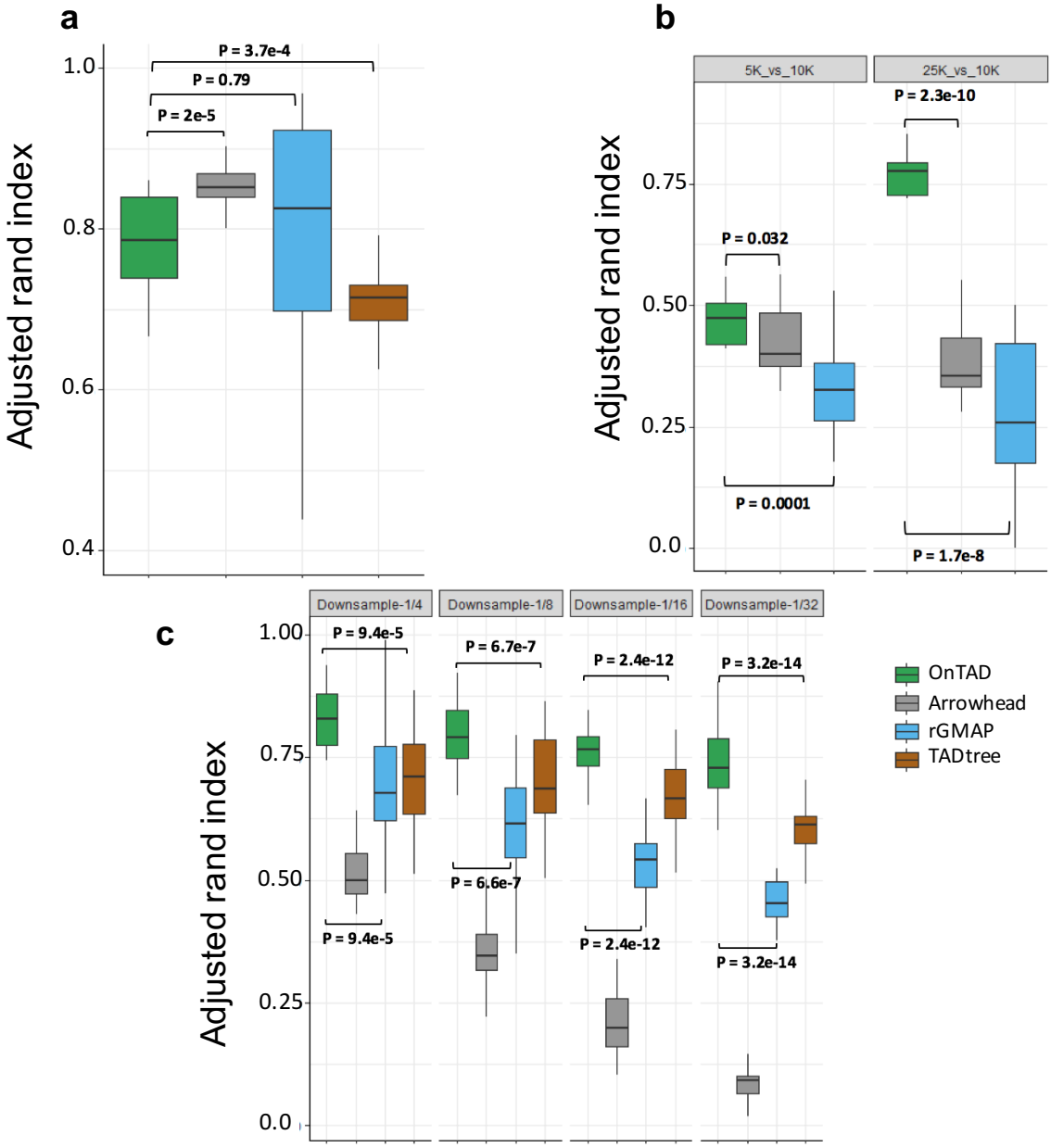

a
level 1
level 2
level 3
b
level 1
level 2
level 3
level 4
level 5
ReprD
ReprI
PromF2
FaireW1
EnhWF3
EnhWF1
Ctcf
FaireW2
EnhF
Elon
Zero
Low1
Art1
Art2
Repr1
ElonW
LowReprW
DnaseD2
Quies
DnaseD1
Low2
PromP
Enh
Gen3
EnhWF2
Gen5
Pol2
CtcfO
TssW
Gen3Ctcf
TssF
Tss
TssCtcf
EnhW
PromF1
PromCtcf
c
levels
1
2
3
4
5
Avg. RNA-seq signal (FPKM)
150
100
50
0
Boundary level
d
SIX5
SP4
SAP30
ELK1
KDM5B
HDAC1
WRNIP1
NFYA
IRF1
CTCFL
BRCA1
E2F4
BCLAF1
SIN3A
CREB1
HMGN3
ELK4
THAP1
PHF8
ETS1
CHD1
TAF7
NRF1
IRF3
level 1
level 2
level 3
level 4
level 5



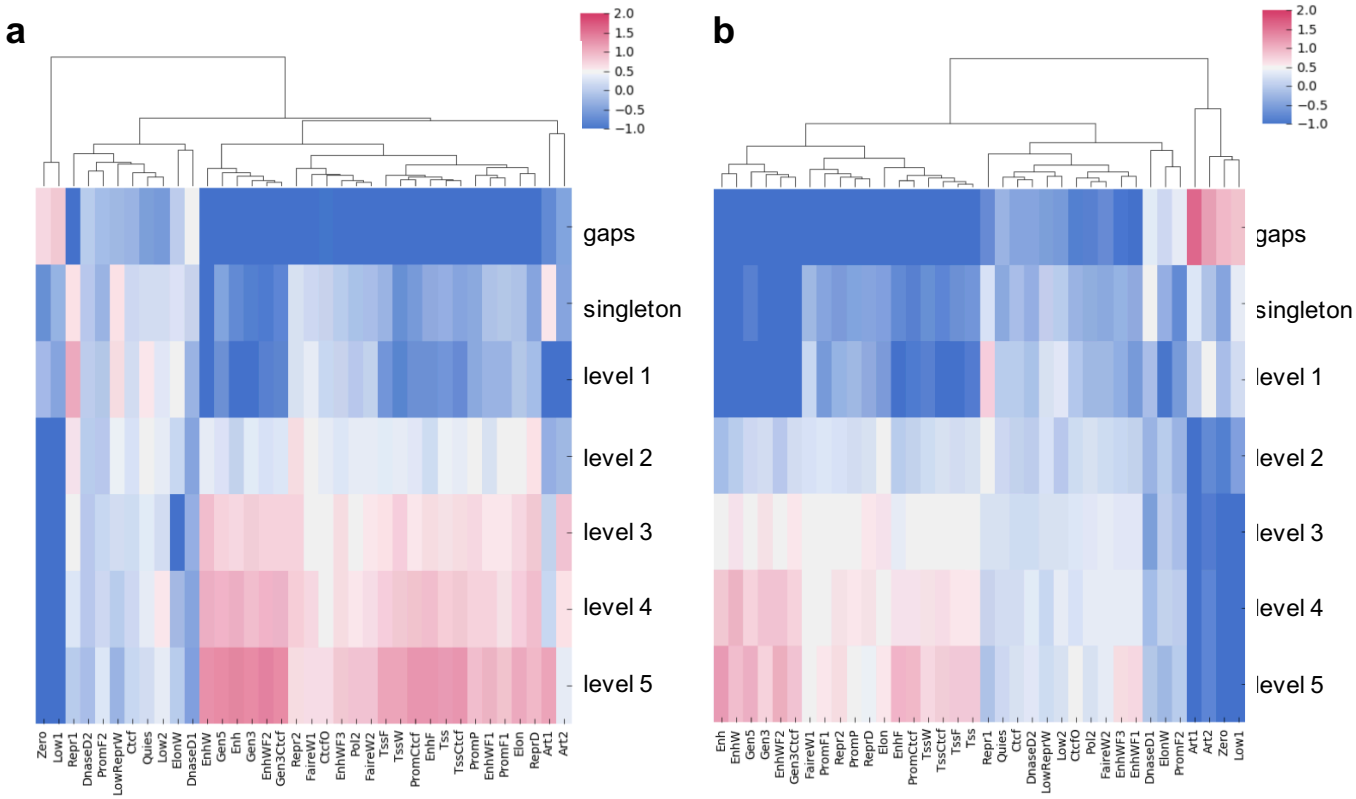
Supplementary figure 1 | **Illustration of convoluted TAD structures.** a, Candidate TADs (a,c) and (b,d) are both suboptimal, as their scores may be driven by a real TAD (b,c). b, Two real TADs (a,c) and (b,c) are nested, which makes the score of (a,c) convoluted with the score of (b,c). c, Real TADs (a,c) and (b,d) are partially overlapping, which may be recaptured as nested TADs (b,c), (a,c) and (a,d).



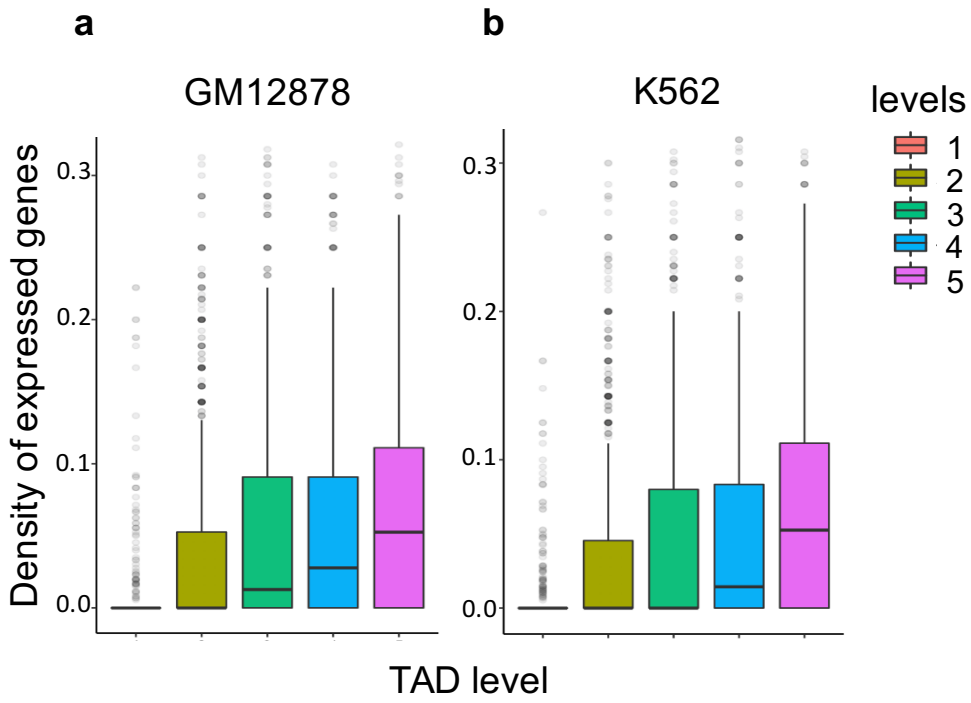
Supplementary figure 2 | Illustration of the recursive TAD calling algorithm. a, Starting from the entire matrix, we partition the matrix into two matrices: the one that potentially forms the largest right-most TAD (triangles marked in black), and the remaining part. We call the same function on each sub-matrix to recursively identify nested TAD structures, and the best partition is determined by a scoring function. b, Each recursion step identifies the best set of TADs in its matrix under consideration, and return the TAD calls back to its parent until the root.



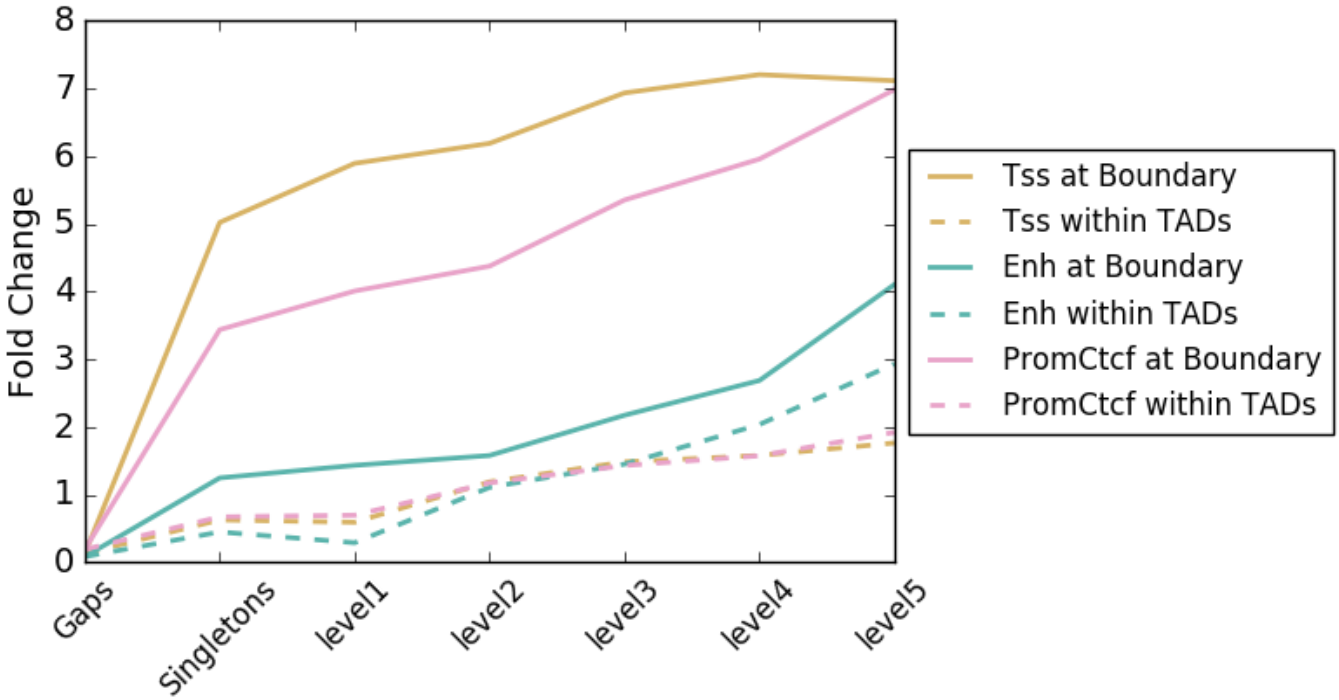
Supplementary figure 3 | TAD reproducibility under different measurements. a, Adjusted rand index between TADs from two biological replicates (GM12878, 10Kb). b, Adjusted rand index across TADs from Hi-C data in multiple resolutions (GM12878, 5Kb, 10Kb and 25Kb). We excluded TADtree as it took too much computing resources on high resolution data. **c,** Adjusted rand index between TADs from Hi-C data in original sequencing depth and in different down sampled sequencing depth (GM12878, 1/4, 1/8, 1/16 and 1/32 of original sequencing depth). All comparisons are calculated on autochromosomes individually (excluding chr1 and chr9 due to no results produced by some methods).



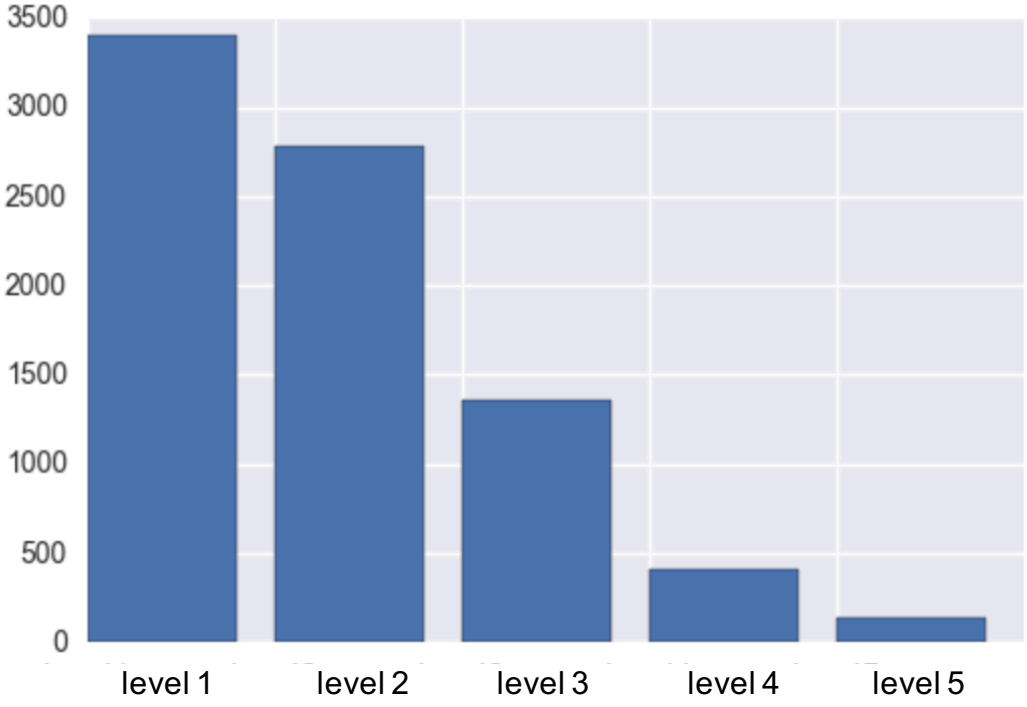
Supplementary figure 4 | Enrichment of epigenetic states at the boundaries of different levels of TADs. Same method was used with figure 4c. a, K562 b, Huvec



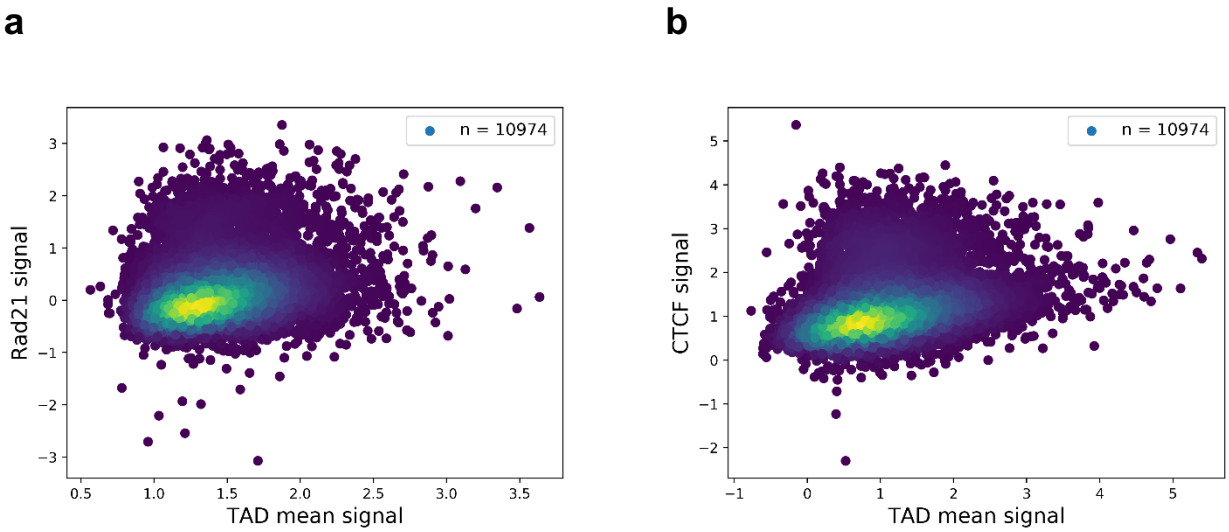
Supplementary figure 5 | Density of the number of expressed genes in different levels of TADs a, GM12878 b, K562



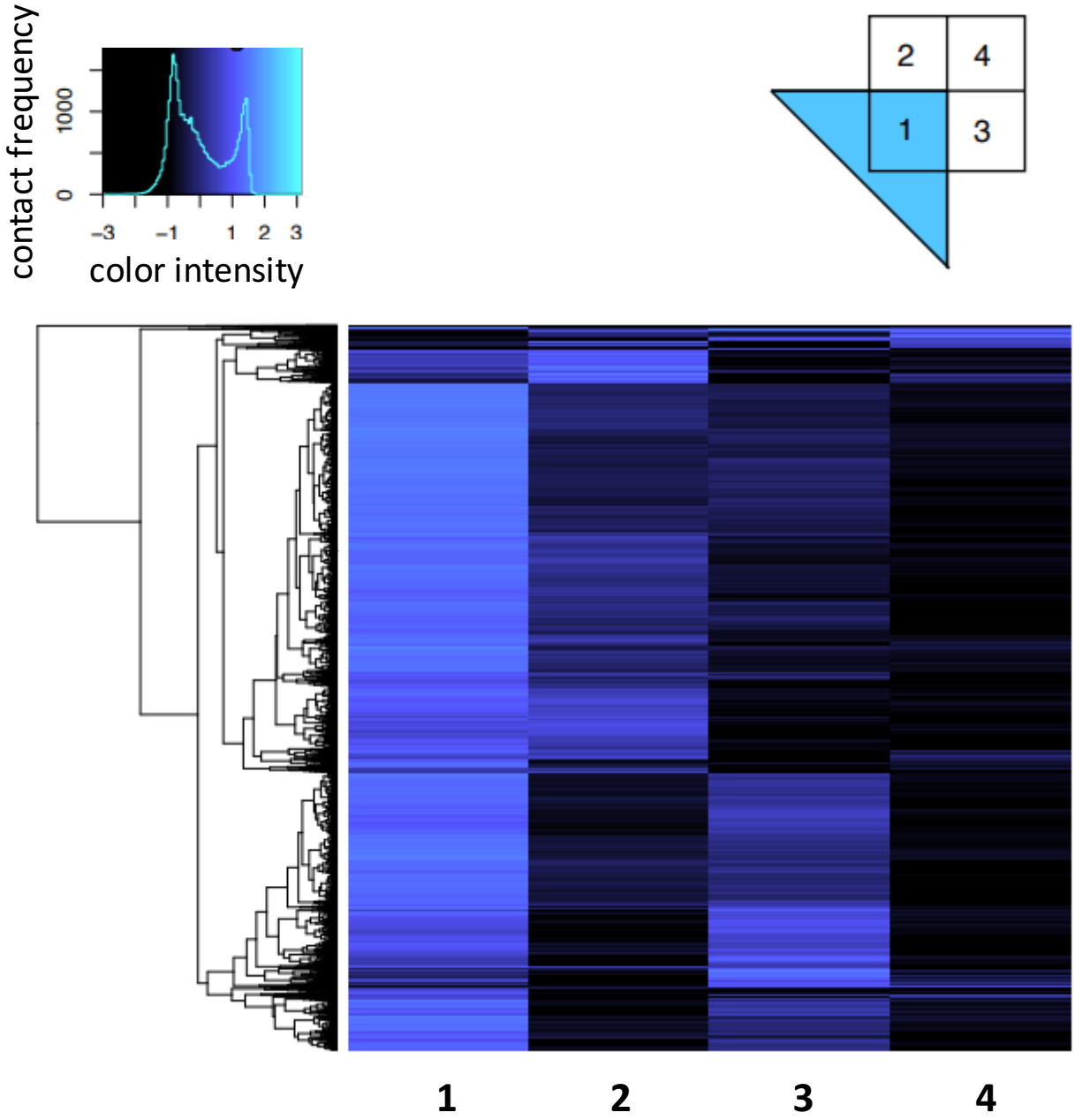
Supplementary figure 6 | Enrichment of active epigenetic states at the boundaries (solid line) versus inside TADs (dashed line). Y-axis denotes fold enrichment of three active epigenetic states (Tss, Enh and PromCtcf). X-axis denotes the boundaries and TADs at different levels.



Supplementary figure 7 | Number of TAD boundaries in each level. The level is defined as the maximum number of TADs on either side that share this boundary.



Supplementary figure 8 | Cohesin signal strongly correlates with mean contact frequency in TADs. a, scatter plot for TAD mean signal (x-axis) versus Cohesin signal (y-axis) in all TADs. b, scatter plot for TAD mean signal versus CTCF signal in all TADs.



Supplementary figure 9 | Contact frequency is unbalanced on the two sides of hierarchical TAD corners. The regions around TAD corners are segregated into four 5*5 quadrants (1-4 on the top right figure). We then averaged contact frequency of each TAD corner by quadrants. As shown in the heatmap, quadrant 1 has highest average contact frequency as it is within TADs. Meanwhile, majority quadrant 2 and 3 shows unequal average contact frequencies, suggesting the outer TADs tend to form on one side of inner TADs rather than both sides.