Supplemental Materials

BINDER achieves accurate identification of hierarchical TADs by comprehensively characterizing consensus TAD boundaries

Yangyang Liu[1], Bingqiang Liu[2,*], Juntao Liu[1,*]

[1]School of Mathematics and Statistics, Shandong University (Weihai), Weihai, 264209, China

[2]School of Mathematics, Shandong University, Jinan 250100, China

[*]To whom correspondence should be addressed

Email addresses:
BL: bingqiang@sdu.edu.cn
JL: juntaosdu@126.com

Table of contents

Supplemental Discussions

Performance of BINDER on Dip-C data

We tested the performance of BINDER and the other eight compared TAD callers on Dip-C data of 16 GM12878 single cells downloaded from GEO: GSE117876. The results in Supplemental Fig. S11 showed that the number of TADs identified by BINDER was basically second only to the number of TADs identified by SpectralTAD and deDocM, both of which has excellent robustness on ultra-sparse Hi-C data (Li et al. 2018a; Cresswell et al. 2020). Besides, we tested the performance of BINDER and the other compared TAD callers in terms of the precision and the number of CTCF-matched boundaries, which were shown in Supplemental Fig. S12-13. From the results, we found that the three TAD callers including OnTAD, deDocE, and CATAD demonstrate higher precision, while the others including BINDER show similar precision across the 16 cells. However, in terms of the number of CTCF-matched boundaries, all of them (OnTAD, deDocE, and CATAD) perform much worse than BINDER, and only deDocM, IS, and SpectralTAD show better performance than BINDER while the precision of IS and deDocM is lower than BINDER in most cases. Based on the above results, it suggested that BINDER demonstrates a relatively good overall performance in single cell Dip-C data compared to other TAD callers. Here, we would like to emphasize that although BINDER shows excellent robustness in sparse randomly down-sampled datasets (with a minimum down-sampling rate of 1/100), it is specially designed according to the characteristics of bulk Hi-C datasets - using community discovery algorithm and a neural network model based on three TAD boundary features. However, single-cell datasets, such as Flyamer's dataset (Flyamer et al. 2017), exhibit a distinctive characteristic from bulk Hi-C data: much more severe sparsity. Li et al. mentioned that the number of contacts for bulk Hi-C data is comparable to that of Flyamer's dataset with a down-sampling rate of 1/800, which far exceeds the rate of 1/100 in our robustness benchmarking test (Li et al. 2021). The performance of BINDER on the 16 single cell datasets may suggest that the severe sparsity of the Hi-C data has limited the performance of BINDER in its current version. However, this test leads us to considering an updated version of BINDER specifically for the characteristics of scHi-C or DipC data, e.g., retraining the neural network by using extremely sparse Hi-C matrices and adjusting the Infomap algorithm to make it suitable for single-cell

Hi-C data.

Convergence comparison

In this study, we say that a TAD conforms to the principle of convergence if its left and right boundaries are matched by a CTCF labeled “Forward” and “Reverse”, respectively (Rao et al. 2014). And for convenience, we call a TAD a convergence TAD if it conforms to the principle of convergence. To compare convergence of TADs identified by BINDER and the other compared TAD callers, we tested their outputs on 50kb, 25kb, and 10kb GM12878 data in terms of the number and proportion of convergence TADs. As shown in Supplemental Fig. S10, we can see that BINDER shows the best performance in terms of both the number and proportion of convergence TADs, indicating its more accurate TAD identification ability.

An example showing breaking a down-coregulated TAD relates to a cell-type transition

Although Zhan et al. analyzed the coregulation of TADs in ESCs and NPCs (Zhan et al. 2017), to the best of our knowledge, studies on coregulated TADs in hematopoietic cell lines are relatively scarce. Below we provide a possible example, the breakdown of a down-coregulated TAD in the MEP cell type in the MK differentiation path (Supplemental Fig. S24A). The down-coregulated TAD contains two genes *Xpo7* and *Dok2*, and we can see that the expression values of both *Xpo7* and *Dok2* undergo a sharp decline before MEP and a sharp increase in MK (Supplemental Fig. S24B), which clearly demonstrates that the breaking of the down-coregulated TAD may closely relates to the transition of cell types from MEP to MK. About the *Xpo7* gene, although Hattangadi et al. indicated that it appears to play an important role in erythropoiesis, its specific mechanism in MK lineage remains to be investigated (Hattangadi et al. 2014). Moreover, during the terminal differentiation of erythropoiesis, this gene is highly induced, based on which we infer that *Xpo7* may also play important roles during the differentiation of MK. *Dok2* are closely related rasGAP-associated docking proteins expressed preferentially in hematopoietic cells, and several studies reports have shown that *Dok2* are involved in myeloid homeostasis (Yasuda et al. 2004; Coppin et al. 2016). This gene regulates the expression of *Klf1* gene by directly binding to its promoter region and Klf1 is an important transcription factor during the differentiation of erythropoiesis.

Though related studies mainly focus on the roles of *Dok2* during the differentiation of erythropoiesis, we can infer that *Dok2* may also plays an important regulatory role during the the differentiation of MK since MEP can differentiate into both RBC and MK.

Co-active (co-inactive) TADs

We first define a TAD as a co-active (co-inactive) TAD if 80% of the genes within that TAD have an expression value greater than 0 (for 0), and then we calculated *Z*-scores in terms of the number of co-active and co-inactive TADs as we defined above. The results are shown in Supplemental Fig. S20.

We found the features of co-active and co-regulation of genes within TADs were different in the following ways: (i) the number of co-active (co-inactive) TADs was significantly higher compared to the number of up- (down-) coregulated TADs; (ii) The number of up- (down-) coregulated TADs in the GR path showed a clear upward trend, whereas the number of co-active (co-inactive) TADs was low in GR cell types relative to other cell types; (iii) *Z*-scores calculated in co-active analysis were generally high compared to those calculated in the coregulation analysis, leading to *P*-values close to 0; (iv) *Z*-score was increasing in co-active TADs from GMP to GR while decreasing from MEP to MK, but decreasing in both in co-inactive TAD, which may suggest a difference between the co-active and co-inactive patterns within TADs in different differentiation pathways.

Selection of Hi-C normalization method

To test the robustness of BINDER in terms of the Hi-C data normalization methods, we evaluated the performance of BINDER using different normalization methods including iterative correction and eigenvector decomposition (ICE) (Imakaev et al. 2012a), Knight-Ruiz (KR) (Knight and Ruiz 2012), square root Vanilla Coverage (sqrtVC) (Rao et al. 2014), and Sequential Component Normalization (SCN) (Cournac et al. 2012), and the results are shown in Supplemental Table S10. The results show that BINDER exhibits a robust performance with different normalization methods, with SCN-BINDER reaching the highest precision and ICE-BINDER achieving the highest number of CTCF-matched TAD boundaries. For convenience, we added an option in our tool that allows the users to select different

normalization methods based on their needs, and the default set is SCN.

Selection of community discovery algorithms

About the selection of community discovery algorithms, our consideration is mainly based on the characteristics of the Hi-C matrices. For example, the community discovery is performed on sub-matrices extracted from the diagonal of a Hi-C contact matrix. The values of these sub-matrices generally conform to a power law that has a scaling exponent close to −1 in many species (Szabo et al. 2019), which means that the closer to the diagonal the larger the value, a feature that distinguishes this network from other networks. The Infomap algorithm based on information flow compression focuses on the time of information flow of the entire network structure rather than overly focusing on edges of high weights and theoretically satisfies our requirements. Other methods, such as modularity-based Louvain algorithm, usually tend to place more edges with high weights inside a community (Rosvall and Bergstrom 2008). Therefore, these algorithms overly consider those edges that are located near the diagonal when processing sub-matrices with power-law distributions, thus losing the consideration of the overall network structure. As for methods based on spectral clustering, they usually require the users to set the number of clusters in advance. However, different parameter choices may lead to quite different clustering results. For example, SBTD needs users to set a parameter k to set the number of clusters. For SpectralTAD, although it does not require a pre-given parameter, it introduces an additional approach named Silhouettes for automatically choosing the number of clusters. Taken together, in order to avoid artificial parameterization or to avoid possible risks due to the additional approach introduction for choosing parameters, we chose the Infomap algorithm for community discovery that is suitable for Hi-C data and requires no additional parameter.

Mouse hematopoietic differentiation tree

During mouse hematopoietic cell differentiation, long-term hematopoietic stem cell (LT-HSC), positioned at the root of the differentiation tree, can differentiate into short-term hematopoietic stem cell (ST-HSC). Subsequently, ST-HSC differentiates into multipotent progenitor (MPP), marking the initial divergence into common myeloid progenitor (CMP)

and common lymphoid progenitor (CLP). This branching point initiates the distinct pathways leading to myeloid and lymphoid cell lineages. Further down the myeloid lineage, CMP differentiates into granulocyte-macrophage progenitor (GMP) and megakaryocyte-erythroid progenitor (MEP). MEP retains the capacity to differentiate into the megakaryocyte progenitor (MKP), ultimately leading to the terminal differentiation of myeloid cells, including megakaryocytes (MK) and granulocytes (GR) (Seita and Weissman 2010; Zhang et al. 2018).

Supplemental Methods

Normalization of Hi-C matrix

To test the effect of different normalization methods on identifying TAD, we used iterative correction and eigenvector decomposition (ICE) (Imakaev et al. 2012b), Knight-Ruiz (KR) (Knight and Ruiz 2012), square root Vanilla Coverage (sqrtVC) (Rao et al. 2014) and Sequential Component Normalization (SCN) (Cournac et al. 2012) to normalize Hi-C contact matrix. ICE normalization was performed with a Python package "iced" (v0.5.13) (https://github.com/hiclib/iced), and sqrtVC and SCN normalization was performed with a Python package "prody" (v2.4.0) (Zhang et al. 2021). As for KR normalization, the data we used from Rao and colleagues (GSE63525) provide vectors for performing KR normalization, e.g., chr1_10kb.KRnorm, which is denoted as X. The dimensions of this vector are the same as the dimensions of the rows or columns of the corresponding Hi-C contact matrices, and so we can KR-normalize a raw contact matrix M by using the following equation:

$$M^{'} = diag(X) \cdot M \cdot diag(X)$$

where $diag(X)$ represents the diagonal matrix with X as the diagonal and M is the raw Hi-C contact matrix.

And the built-in KR normalization in BINDER is from Kumar's code (Kumar et al. 2017).

Parameter setup for TAD callers

We performed all of the other eight TAD callers compared to BINDER using their default parameters on the same data (even though different software require different matrix formats, we have transformed the input data in the format they require) to ensure as much fairness as possible in our benchmarking test.

For methods that identify non-nested TADs, we ran them as described below. For TopDom, we used the default window.size=5 for all data tested. For MSTD, we used the MSTD_v1 method from "MSTDlib", a package for Python provided by the author of MSTD, with its default parameters. For SBTD, we used SBTD.py provided by the author to run it and set k=3 by default. For IS, we set the following parameters: -bmoe 3 -nt 0.1 -v -im mean. In addition, since IS does not directly output TADs but rather TAD boundaries, it is not involved in comparisons about TADs. For deDoc, we provide Hi-C input directly and leave all the other

parameters as default. For TAD, where CS_threshold we set by default to 0.8. For methods that identify nested TADs including SpectralTAD and OnTAD, we ran them as follows. For SpectralTAD, we downloaded its package and ran it with its default settings. For OnTAD, we provided the input Hi-C data to OnTAD directly and make no changes to all its default parameters.

Sequential Component Normalization (SCN)

The SCN method is a global normalization method for contact maps independent of the protocol that generates it, and it can be applied to any genomic contact map. In addition, this method has the advantage of eliminating potential experimental bias, thus providing a clearer indication of the frequency of interactions between any pair of restriction fragments in the genome. And the detailed procedures are as follows (Cournac et al. 2012).

First, each column vector of the raw contact matrix is normalized using the Euclidean norm, and then each row vector of the resulting matrix is normalized. The whole process is repeated in turn until every row and column of the matrix is normalized to 1 and the matrix is once again a symmetric matrix. This normalization can be thought of as a sequence of expansions and contractions of the interaction vectors so that they tend to a sphere of radius 1 in the interaction space. Usually, a small number of iterations is sufficient to ensure convergence.

Infomap community discovery algorithm

The Infomap algorithm equates searching for the optimal community partition to minimizing the global information encoding, where the map equation serves as the objective function of this optimization process. And the detailed procedures are as follows (Rosvall and Bergstrom 2008).

The community partition M is defined as the partition of the set of n vertices into m communities such that each vertex is assigned to only one community. The map equation $L(M)$ gives the average number of bits per step required to describe an infinite random walk on a network partitioned according to M:

$$L(M) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^{m} p_{\circlearrowright}^{i} H(\mathcal{P}^{i}) \quad (1)$$

The first term of this equation gives the average number of bits needed to describe the movement between modules and the second term gives the average number of bits needed to describe the movement within modules.

First, consider the first part. The probability that the random walker switches communities at each step is:

$$q_{\curvearrowright} = \sum_{i=1}^{m}\left(\tau\frac{n-n_i}{n-1}\sum_{\alpha\in i}p_\alpha + (1-\tau)\sum_{\alpha\in i}\sum_{\beta\notin i}p_\alpha w_{\alpha\beta}\right) \tag{2}$$

where τ (0.15 by default) is a teleportation probability that connects any two vertices with positive probability. n_i is the number of nodes in community i. p_α is the probability that a random walker visits vertex α. $w_{\alpha\beta}$ is the probability that vertex α transfers to vertex β.

The entropy of movements between communities is

$$H(\mathcal{Q}) = \sum_{i=1}^{m}\frac{q_{i\curvearrowright}}{\sum_{j=1}^{m}q_{j\curvearrowright}}\log\left(\frac{q_{i\curvearrowright}}{\sum_{j=1}^{m}q_{j\curvearrowright}}\right) \tag{3}$$

where $q_{i\curvearrowright}$ is the probability of existing community *i*.

Then, consider the second part. To weight the entropy of movements within community *i*, we compute

$$p_{\circlearrowright}^{i} = q_{i\curvearrowright} + \sum_{\alpha\in i}p_\alpha \tag{4}$$

The entropy of movements within community *i* is

$$H(\mathcal{P}^i) = \frac{q_{i\curvearrowright}}{q_{i\curvearrowright}+\sum_{\beta\in i}p_\beta}\log\left(\frac{q_{i\curvearrowright}}{q_{i\curvearrowright}+\sum_{\beta\in i}p_\beta}\right) + \sum_{\alpha\in i}\frac{p_\alpha}{p_\alpha+\sum_{\beta\in i}p_\beta}\log\left(\frac{p_\alpha}{p_\alpha+\sum_{\beta\in i}p_\beta}\right) \tag{5}$$

Finally, a deterministic greedy search algorithm is used, followed by a heat-bath algorithm to improve the results by a simulated annealing approach.

Evaluation of validation set during MLP model training

At each epoch, we evaluate the performance of the trained model on the validation set by calculating the accuracy (ACC). Specifically, we iterate through each threshold from 0-1 in step of 0.01, calculate the ACC value corresponding to each threshold (labels of samples for which the model's predicted value is greater than the threshold are set to 1, and the labels of

samples less than or equal to the threshold are set to 0), and select the largest ACC value as the ACC value for that epoch. ACC is defined as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

where TP, FP, TN, and FN denote the number of samples for true positives, false positives, true negatives, and false negatives, respectively.

Description of dual boundary

The purpose of introducing the dual boundary concept in BINDER is to show that we consider the "seam" between two bin i and bin i+1, or more delicately the middle of the two nucleotides adjacent to bin i and bin i+1, to be the boundary of a TAD, and this definition is intended to make the boundary of a TAD more precise. For example, if a TAD is located from the leftmost nucleic acid of bin i towards the rightmost nucleic acid of bin j, then it is identified by two dual boundaries (i-1, i) and (j, j+1) based on the definition of dual boundary. However, the TAD is recorded by two bins i and j according to traditional definition, which is hard to precisely describe the exact boundaries of the TAD. In the whole pipeline of BINDER, each module is designed based on this precisely defined boundaries. We added the above description of dual boundary in the Methods section of the revised version. However, in the final output, for convenience, BINDER still gives the common definition of TAD. For example, we find two neighboring TADs which are defined by three dual boundaries: (10, 11), (22, 23), and (54, 55), then it outputs the two TADs like:

TAD_rank	left_bin	right_bin
1	11	22
2	23	54

Generation of test sets for MLP model

We generated 4 test sets using data of 23 chromosomes (22 autosomes+X sex chromosomes) from GM12878 cell line normalized by 4 normalization methods (SCN, ICE, KR, sqrtVC) at 50kb, 25kb and 10kb. Given the data for a certain normalization method, we extract 430 positive samples (215) and negative samples (215) in each chromosome at each resolution to

ensure a balance of positive and negative samples in the test set. If the positive (negative) samples are less than 215, they are all added to the test set. Finally, we generated 4 test sets with sample sizes 29590, 29588, 29588 and 29588 from SCN, ICE, KR and sqrtVC normalized data, respectively.

Early stopping to avoid overffiting

In the MLP model training process, we use early stopping method to avoid overfitting, i.e., if the model does not exceed the ACC value of the i-th epoch in all epochs after the *i*-th epoch (*n*=10 by default in BINDER), the model with the highest ACC value in the epochs before the *i*-th epoch (including the *i*-th epoch) is selected as the final model and the training is terminated. This mechanism can not only effectively avoid overfitting, but also save training resources.

As Supplemental Fig. S22 shows (we present the results of model training for 100 epochs), we will choose the model trained in the 29th epoch as our final model because in this epoch the model produces the highest ACC value of 72.65 in validation set and none of the models in the 30th-39th epochs have ACC values above 72.65. Furthermore, we can see that after 29 epochs the model quickly shows signs of overfitting, so the model after 100 epochs would be overfitting if there was no early stopping.

Ablation experiments

Since BINDER is a multistep pipeline, in order to better test the contribution of each algorithmic component, we did ablation experiments for BINDER to improve the interpretability of its performance and results are shown in Table S10.

First, given a certain normalization method for Hi-C contact matrix, we tested the performance of BINDER with or without the MLP model and with different features combinations, benchmarked against the whole chromosome (22 autosomes and X sex chromosome) of 50kb GM12878 cell line in terms of precision we have defined before, the number of CTCF-matched TAD boundaries, and the number of TADs (we trained a corresponding MLP model for each different combination of features). The results show that the performance of BINDER framework after integrating MLP model far exceeds that of

BINDER framework using only Infomap algorithm. And besides, although BINDER framework using only MLP model achieves higher precision compared to BINDER framework using only Infomap in the four normalization scenarios, the number of TADs and CTCF-matched TAD boundaries by this framework are lower (Here, we limit the BINDER framework with only MLP model to output the same number of boundaries used to generate TAD as the BINDER framework with only Infomap - we set the $count$ value of all dual boundaries to 0, and subsequently select them according to their corresponding reliability scores given by MLP model from high to low, until we reach the specified number of boundaries). Therefore, the strategy of integrating community discovery in network and neural network model is reasonable, and of course the results demonstrate that BINDER's "consensus boundary" strategy has both high precision and is capable of outputting a good enough number of TADs.

Second, we performed feature ablation experiments on the 220-dimensional features of BINDER's MLP model—110-dimensional local interaction density, 100-dimensional directionality index (DI), and 10-dimensional *P*-value of the Wilcoxon rank sum test. In Table S10, we found that BINDER framework with only 100-dimensional DI feature performs best with the highest precision and large TADs output when only one feature is used, suggesting that DI is able to better feature boundaries of TADs. Furthermore, the similarity of the results of Info+D+DI+P and Info+D+DI frameworks and the mostly poor results of the Info+P framework suggests that the 10-dimensional *P*-value of the Wilcoxon rank sum test feature plays a minimal role in the BINDER framework.

Finally, the difference in performance using different normalizations of BINDER (Info+D+DI+P) is not too large, with SCN-BINDER having the highest precision and ICE-BINDER outputting the highest number of TAD and CTCF-matched TAD boundaries. Thus, in order to provide users with a variety of normalization options, we set the normalization parameter in our BINDER software.

Jaccard index

Jaccard index is a classical measure of similarity between two sets, here we use it to measure the similarity between two TAD sets. Considering that an element in a TAD set is a TAD, i.e.,

an integer binary, rather than a single integer, and considering the tolerance of small bias near TAD boundaries, we give a definition of whether two TADs are equal in the following.

Given two sets of TADs $M = \{t_i^M = (a_i^M, b_i^M) | 1 \leq i \leq m\}$ and $N = \{t_i^N = (a_i^N, b_i^N) | 1 \leq i \leq n\}$, where a_i^M, b_i^M, a_i^N and b_i^N are base pair positions. For two TADs $t_i^M = (a_i^M, b_i^M) \in M$ and $t_j^N = (a_j^N, b_j^N) \in N$, we define $t_i^M = t_j^N$ if and only if $|a_i^M - a_j^N| \leq r$ and $|b_i^M - b_j^N| \leq r$, where r is resolution of Hi-C contact matrix. Thus, we can define the Jaccard index of M and N:

$$J(M,N) = \frac{|M \cap N|}{|M \cup N|}$$

where the intersection and union operations of two TAD sets are defined based on the definition of equality between TADs defined above.

Weighted similarity

The definition of Weighted similarity (WS) is given by Li et al (Li et al. 2018b), but in the final part we made a modification. The following describes how to compute the WS of two TAD sets.

Given two sets of TADs $M = \{t_1^M, t_2^M, \cdots, t_m^M\}$ and $N = \{t_1^N, t_2^N, \cdots, t_n^N\}$. First, the similarity score $s_{i,j}$ of any two TADs t_i^M and t_j^N in M and N is defined as:

$$s_{i,j} = \frac{L(t_i^M \cap t_j^N)}{\sqrt{L(t_i^M) \cdot L(t_j^N)}}$$

where $L(t_i^M \cap t_j^N)$ denotes the number of intersecting bins between t_i^M and t_j^N, and $L(t_i^M)$ and $L(t_j^N)$ denote bin numbers of t_i^M and t_j^N, respectively.

Then, the similarity score of t_i^M on N is defined as:

$$s_i(N) = \max_{j=1,2,\cdots,n} s_{i,j}$$

So, the weighted similarity M relative to N is defined as follows:

$$s(M,N) = \frac{\sum_{i=1}^{m} s_i(N) \cdot L(t_i^M)}{\sum_{i=1}^{m} L(t_i^M)}$$

Since $s(M,N)$ is not equal to $s(N,M)$, the mutually weighted similarity $S(N,M)$ of M and N is defined to be:

$$S(N,M) = \frac{s(M,N) + s(N,M)}{2}$$

Hierarchical clustering for the enrichment of epigenetic factors near TAD boundaries

For a given boundary B_i (nucleotide coordinate), we create a 10-dimensional count vector for which all elements are 0, denoted as V_i. Ten different epigenetic markers CTCF, SMC3, RAD21, H3K4me3, H3K4me1, H3K9me3, H3k27ac, H3K27me3, H3K36me3, and DNase are represented in order. We say that B_i has an E-match if B_i is contained by the positional interval (Chip-seq) of a certain peek of some epigenetic marker E. For each occurrence of an E-match, we add 1 count at the corresponding coordinate of V_i belonging to that epigenetic marker. Then, we normalize V_i as follows:

$$V_i^{'}(k) = \frac{V_i(k)}{\sum_{j=1}^{10} V_i(j)}, k = 1,2,\cdots,10$$

We obtain the epigenetic marker enrichment vector $V_i^{'}$ corresponding to B_i, which satisfies:

$$\sum_{k=1}^{10} V_i^{'}(k) = 1$$

Thus, each boundary B_i has an epigenetic marker enrichment vector corresponding to it. Finally, we perform hierarchical clustering on the matrix composed of these vectors.

Hierarchical clustering of 8 mouse hematopoietic cells

First, a gene is defined as active if it is expressed (TPM > 0) in all 8 mouse hematopoietic cell types. Here, we identified a total of 18,152 active genes from 20 chromosomes (19 autosomes and 1 X sex chromosome). Then, each cell type is matched with an expression vector of active genes of dimension 18152, where the element value is the $\log_2$-normalized expression value of the gene ($\log_2$(1+TPM)). Finally, we performed hierarchical clustering of the eight cell types in terms of the Pearson correlation coefficients of the active gene expression vectors between any two cell types.

CTCF-based evaluation criterion

In mammals, TAD boundaries are frequently enriched in chromatin architectural proteins CCCTC-Bindering factor (CTCF) and cohesin, which is considered to be a mechanism by

which they work together to promote "loop extrusion" to construct TADs (Zheng and Xie 2019). In addition, a recent model suggests that the formation of TADs in the genome relies on obligatory alternation of CTCF site clusters (Nanni et al. 2020). Thus, we believe it is logical to use CTCF-based information to assess the quality of TAD boundaries in the absence of quantitative assessment approaches.

Here, we propose a method called CTCF-based evaluation criterion (CEC) for evaluating the quality of predicted TAD boundaries based on CTCF annotations. Assuming that a boundary x locates at the intersection of two bins, B_i and B_{i+1}, we define the neighborhood $U(x,\delta)$ centered at x with a radius of the half-resolution length δ to be the docking region of boundary x. Then boundary x is called CTCF-matched if there is an intersection of a CTCF binding interval with $U(x,\delta)$. Thus, the number of CTCF-matched boundaries for a TAD caller is defined as the number of all CTCF-matched boundaries in its predicted boundaries, and its precision is defined as the ratio of CTCF-matched boundaries out of all its predicted boundaries. Subsequently, the formula for the precision of a given TAD caller C is as follows:

$$precision(C) = \frac{\#\ CTCF - matched\ boundaries}{\#\ boundaries\ identified\ by\ C}$$

Peak enrichment score

In order to assess the enrichment degree of a certain factor f in the vicinity of TAD boundaries, the peak enrichment score $E(f)$ of factor f is defined as follows. Suppose a TAD caller predicts a set of genome-wide TAD boundaries $B = \{b_1, b_2, \ldots, b_n\}$ from a particular cell line. For any b_i, we consider an area of radius 200kb centered on it and partition it into 40 intervals of length 10kb. Then, the factor f that intersects one of these intervals causes the value of that region to add the peak score of f. After performing this operation for all boundaries, the normalization is performed by first summing the corresponding elements of the n 40-dimensional vectors to be $V_{sum} = \{v_1, v_2, \cdots, v_{40}\}$, and then each element being divided by the sum of the vector V_{sum}, and the normalized 40-dimensional vector is denoted as $V_{norm} = \{v_1^{'}, v_2^{'}, \cdots, v_{40}^{'}\}$. So, V_{norm} satisfies the following equation:

$$\sum_{i=1}^{40} v_1^{'} = 1$$

Finally, the peak enrichment score $E(f)$ is defined as follows:

$$E(f) = \frac{\frac{1}{2}(v_{20}^{'} + v_{21}^{'})}{\frac{1}{10}(\sum_{i=1}^{5} v_i^{'} + \sum_{i=36}^{40} v_i^{'})}$$

Size of TADs associated with boundaries

In Supplemental Fig. S16, we explored the enrichment of transcription factors near the boundaries of TADs associated with different sizes, because a TAD boundary may be recruited by several TADs of different sizes, so we need to define for it the exact size of the TADs associated with it. Here, given a boundary, we take the size of the TAD (bin size) associated with it based on the maximum size of those TADs that are bounded by it.

TAD hierarchy change of genes (THC)

Suppose that there are two cell lines C_1 and C_2 and a particular gene g, where the genomic location of g is [a, b] and b-a>100,000. The sets of TADs identified by BINDER on C_1 and C_2 are denoted as T_1 and T_2, respectively, and neither T_1 nor T_2 is an empty set. For any TAD t within genomic location [c, d] belonging to T_1, t is called a matching TAD for gene g in T_1 if $a>c$ and $b<d$. The set consisting of all matching TADs for gene g in T_1 is denoted by T_m^1. Similarly, the set consisting of matching TADs for gene g in T_2 is denoted by T_m^2. Thus, the set of co-preserved TADs for g on C_1 and C_2 is defined as $T_m = T_m^1 \cap T_m^2$. Then, the *THC* of gene g on C_1 and C_2 is defined as:

$$THC(g) = \max_{t \in T_m} \left(H_2(g|t) - H_1(g|t)\right)$$

where $H_k(g|t)$ denote the TAD hierarchy of t in C_k (k=1, 2). If T_m^1 (T_m^2) is an empty set, then $H_1(g|t) = 0$ ($H_2(g|t) = 0$).

Selection of eligible genes

We selected genes in mm10 (GRCm38) that are greater than 100 kilo base pairs in length and do not have a zero expression value in either of the two compared cell lines, here are ST and GR cell lines, and screened a total of 1013 genes, which we refer to as eligible genes.

Random permutation of genome

In Figure 7, we have done random permutation of the genome to calculate *P*-value for each co-regulation TAD, and we describe how we did it below. We actually randomized all the positions of genes on the 20 chromosomes (19 autosomes and 1 X chromosome) of mouse. Suppose that there are n genes in the whole genome, and each gene g corresponds to a position p. Then we can obtain the (gene, gene position) set of the genome:

$$G = \{(g_1, p_1), (g_2, p_2), \cdots, (g_n, p_n)\}$$

Subsequently, we randomly order all the genes to get a new gene order $\{g_{i_1}, g_{i_2}, \cdots, g_{i_n}\}$. Keeping the gene positions unchanged, we thus get a random permutation genome G':

$$G' = \{(g_{i_1}, p_1), (g_{i_2}, p_2), \cdots, (g_{i_n}, p_n)\}$$

***Z*-score and *P*-values of up- (down-) coregulated TADs**

The *Z*-score is calculated with reference to the definition by Zhan et al (Zhan et al. 2017). In a cell line *C*, the number of up- (down-) coregulated TADs in its whole genome is denoted as N_{up} (N_{down}). Then, we counted the number of up- (down-) coregulated TADs in each randomly permutated genome for a total of *M* (*M*=100) times, with the result of each count noted as N_{up}^{i} (N_{down}^{i}), where *i*=1, 2, ..., *M*. The *Z*-score of the up- (down-) coregulated TADs of cell line *C* is then defined as follows:

$$z - score_{up(down)} = \frac{N_{up(down)} - N_{up(down)}^{mean}}{\sqrt{\frac{1}{M-1}\sum_{i=1}^{M}\left(N_{up(down)}^{i} - N_{up(down)}^{mean}\right)^2}}$$

where $N_{up(down)}^{mean} = \frac{1}{M}\sum_{i=1}^{M} N_{up(down)}^{i}$.

P-values are calculated by *Z*-scores as follows:

$$p = 1 - \Phi(|z|)$$

where Φ is the cumulative distribution function of the standard normal distribution.

Definition of upregulated and downregulated genes

Here, we used limma (v3.50.3) to find upregulated and downregulated genes. limma is an R package that can be used for differential expression analysis of data from microarray experiments (Ritchie et al. 2015). Meanwhile, limma's analysis results are stable even when a

small number of experimental arrays are used.

In Figure 7, we used RNA-Seq data from 8 types hematopoetic cells of mouse (2 replicates for each type of cell), and we used the data of LT-HSC as “controls” to analyze the differential expression genes in the other 7 cell species. We defined those genes with logFC greater than 1 (logFC less than -1) and *P*-value less than 0.05 as upregulated (downregulated) genes.

Supplemental Figures

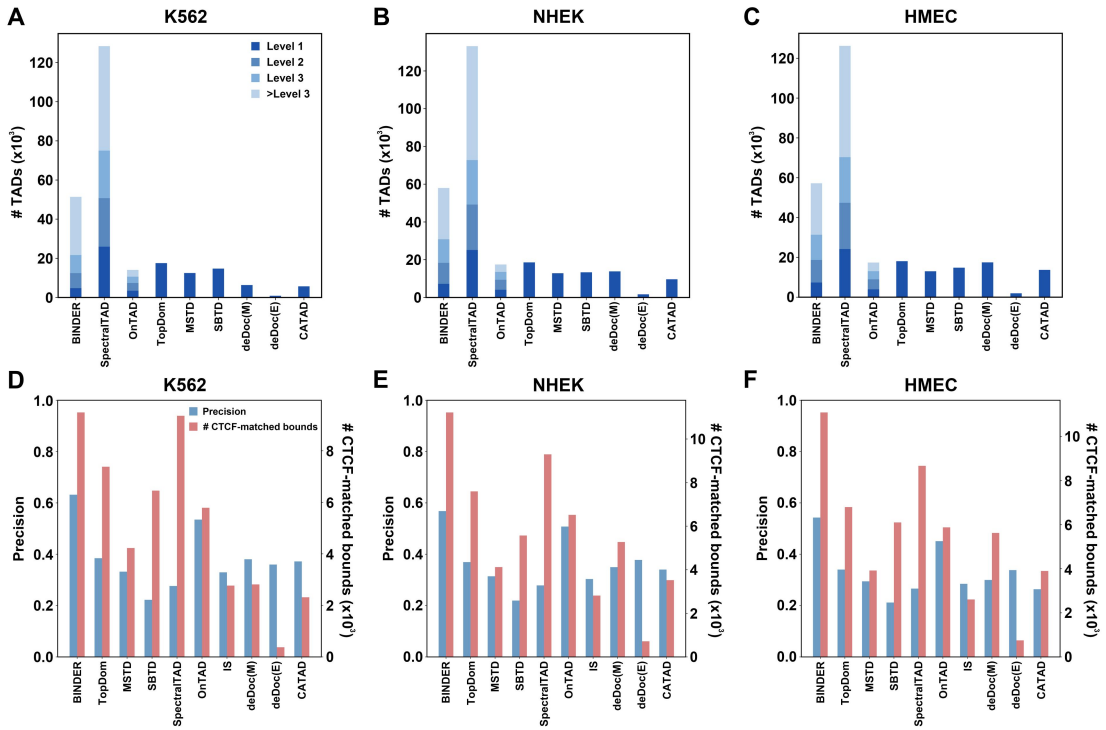

Fig. S1 Number of nested or non-nested TADs identified by BINDER and the other compared TAD callers on 10kb (**A**) K562, (**B**) NHEK, and (**C**) HMEC data. Precision and the number of CTCF-matched TAD boundaries of BINDER and the other compared TAD callers on 10kb (**D**) K562, (**E**) NHEK and (**F**) HMEC data.

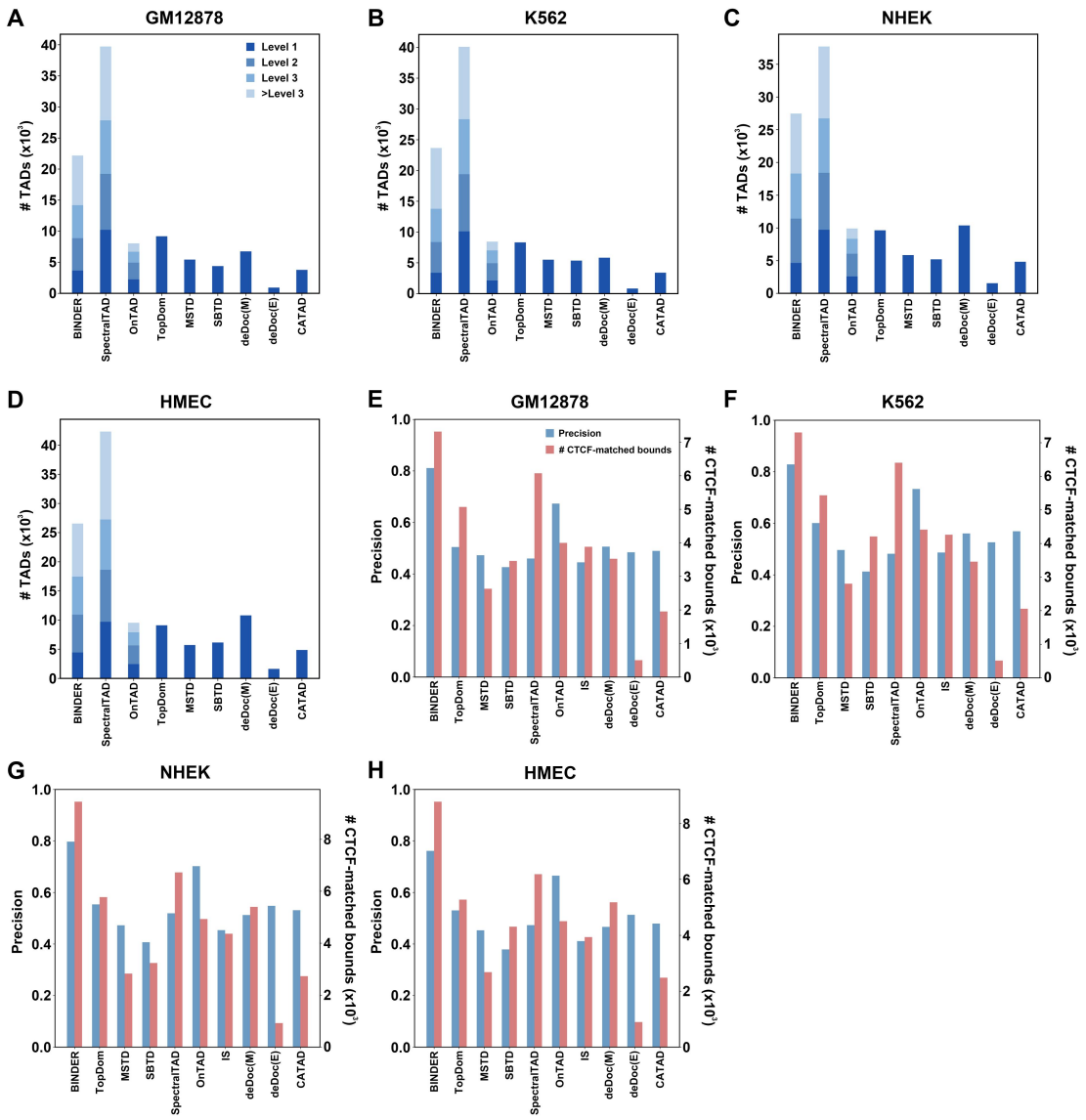

Fig. S2 Number of nested or non-nested TADs identified by BINDER and the other compared TAD callers on 25kb (**A**) GM12878, (**B**) K562, (**C**) NHEK and (**D**) HMEC data. Precision and the number of CTCF-matched TAD boundaries of BINDER and the other compared TAD callers on 25kb (**E**) GM12878, (**F**) K562, (**G**) NHEK and (**H**) HMEC data.

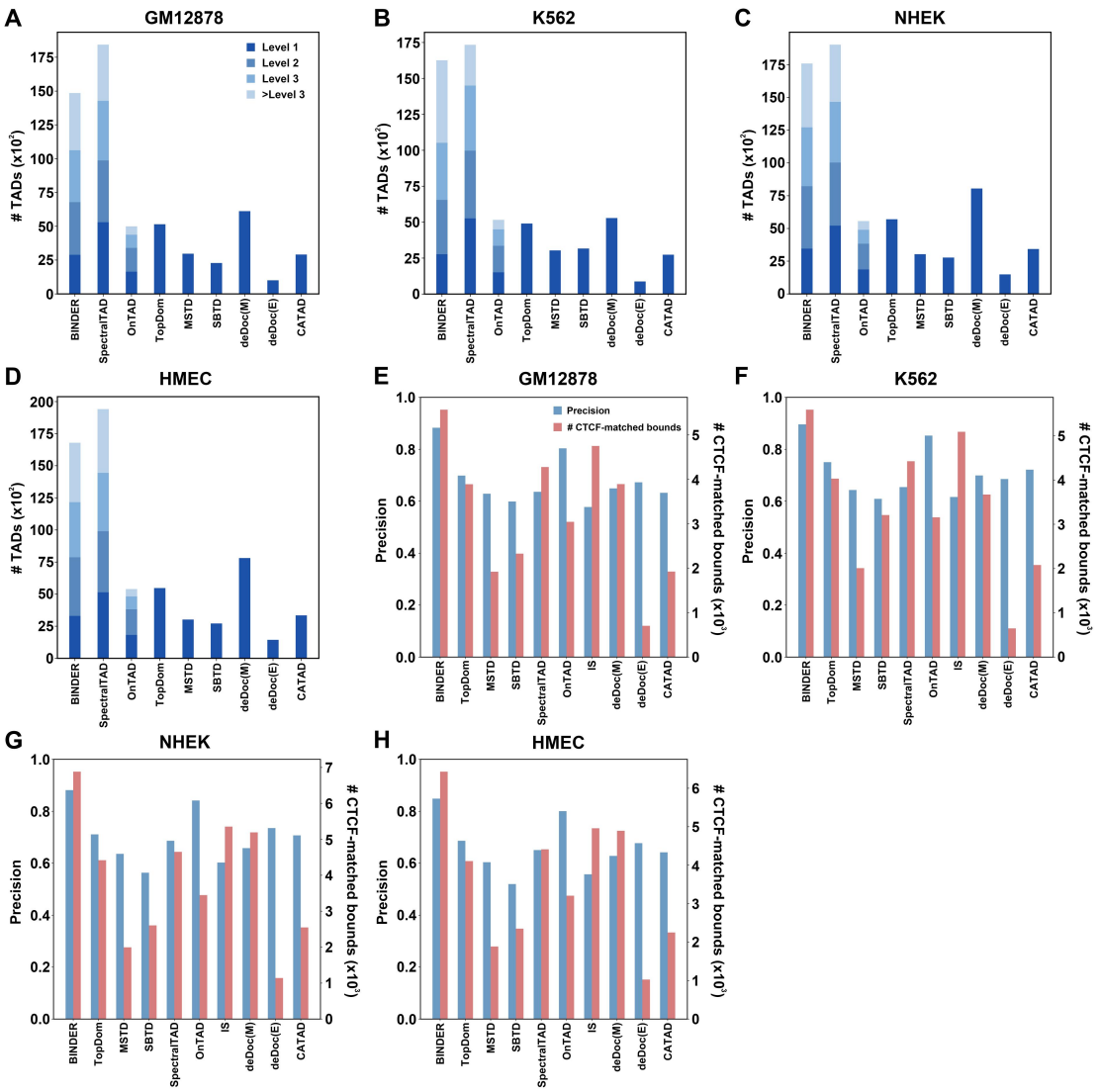

Fig. S3 Number of nested or non-nested TADs identified by BINDER and the other compared TAD callers on 50kb (**A**) GM12878, (**B**) K562, (**C**) NHEK and (**D**) HMEC data. Precision and the number of CTCF-matched TAD boundaries of BINDER and the other compared TAD callers on (**E**) GM12878, (**F**) K562, (**G**) NHEK and (**H**) HMEC data.

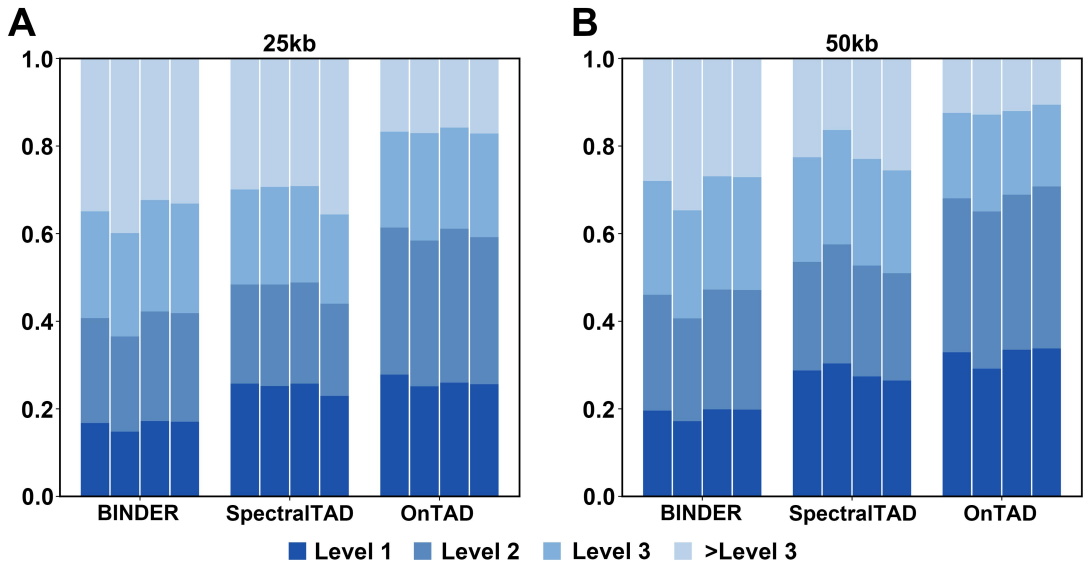

Fig. S4 Proportion of TADs at different levels identified by BINDER, SpectralTAD and OnTAD (four bars for each algorithm from left to right for the results of GM12878, K562, NHEK and HMEC respectively) at resolutions of (**A**) 25kb and (**B**) 50kb.

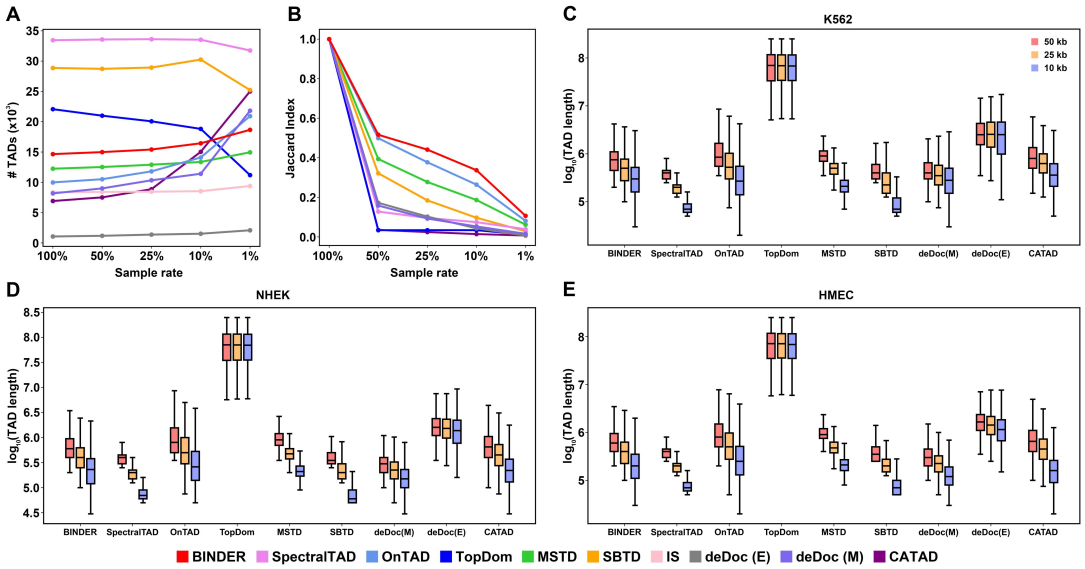

Fig. S5 (**A**) Number of TAD boundaries and (**B**) Jaccard index of BINDER and the compared eight TAD callers at 4 down-sampling rates. Length of TADs ($\log_{10}$ of bp length) identified by BINDER and the other compared TAD callers on (**C**) K562, (**D**) NHEK and (**E**) HMEC data at resolutions of 50kb, 25kb, and 10kb.

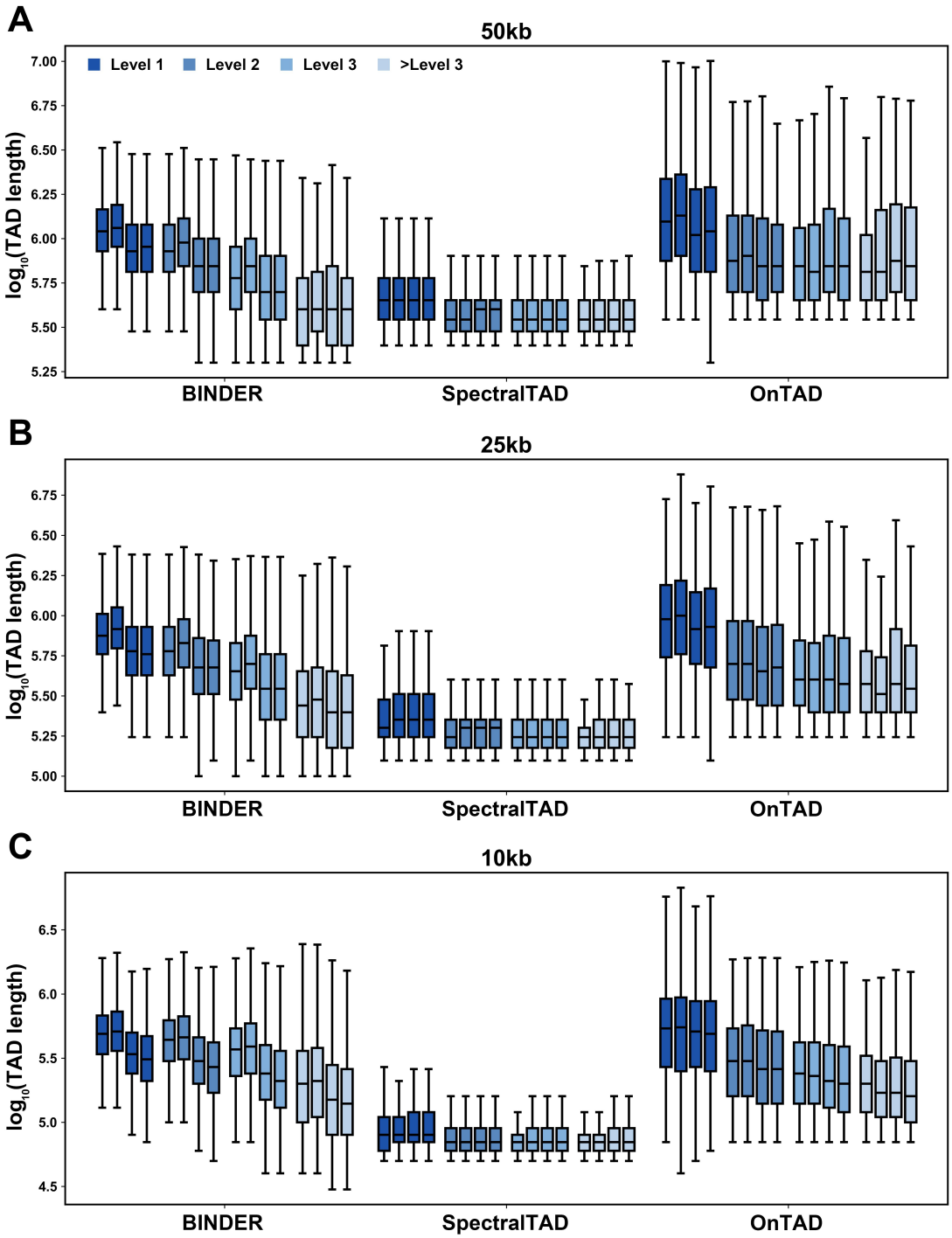

Fig. S6 Lengths ($\log_{10}$ bp length) of TADs identified by BINDER, SpectralTAD and OnTAD at different levels on GM12878, K562, NHEK and HMEC data at resolutions of 50kb, 25kb, and 10kb (boxplots of the same color from left to right for the results of GM12878, K562, NHEK and HMEC, respectively).

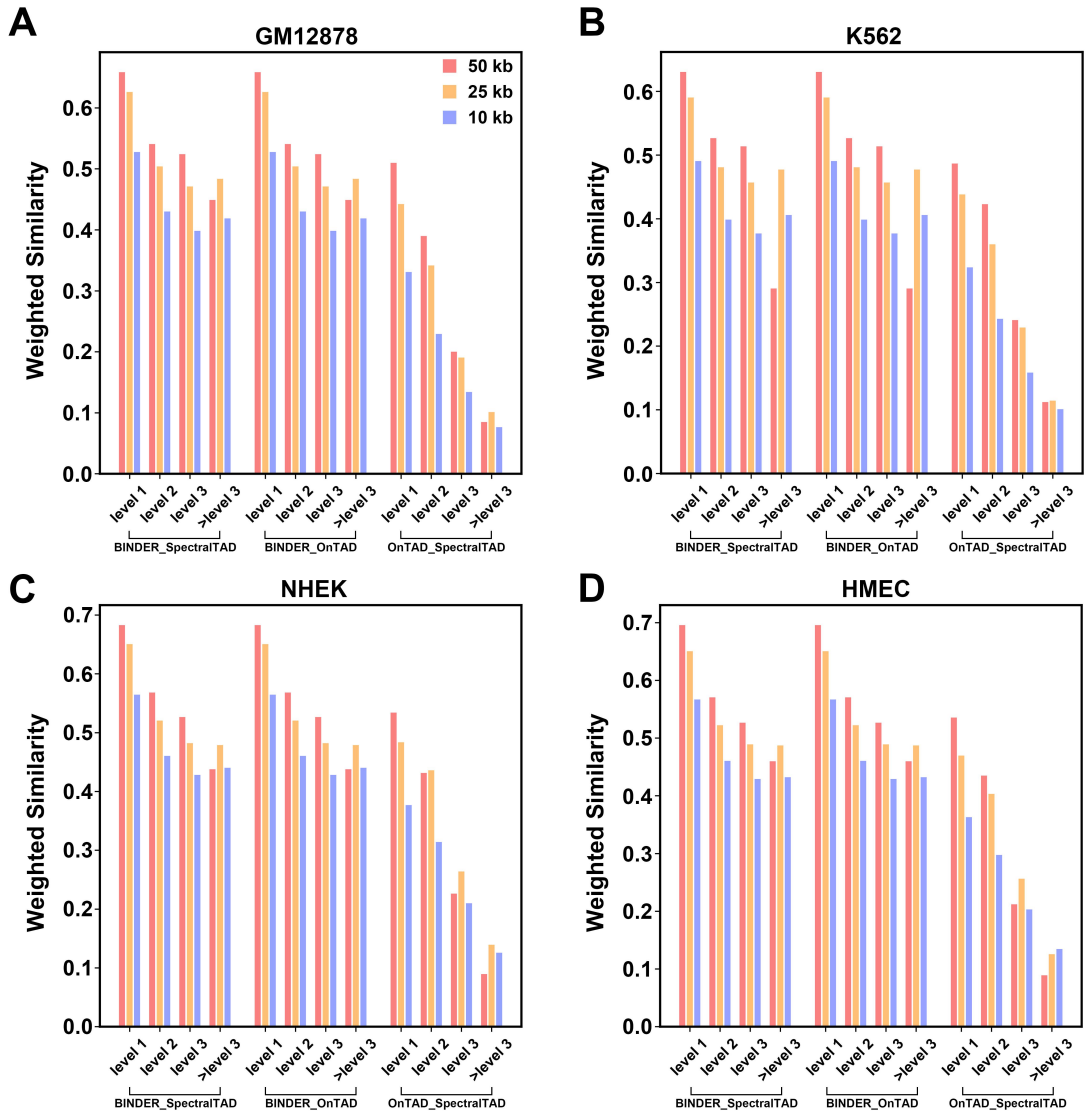

Fig. S7 Comparison of weighted similarity between TADs of level 1, level 2, level 3, and levels greater than 3 identified by BINDER, SpectralTAD, and OnTAD on (**A**) GM12878, (**B**) K562, (**C**) NHEK, and (**D**) HMEC data at resolutions of 50kb, 25kb, and 10kb.

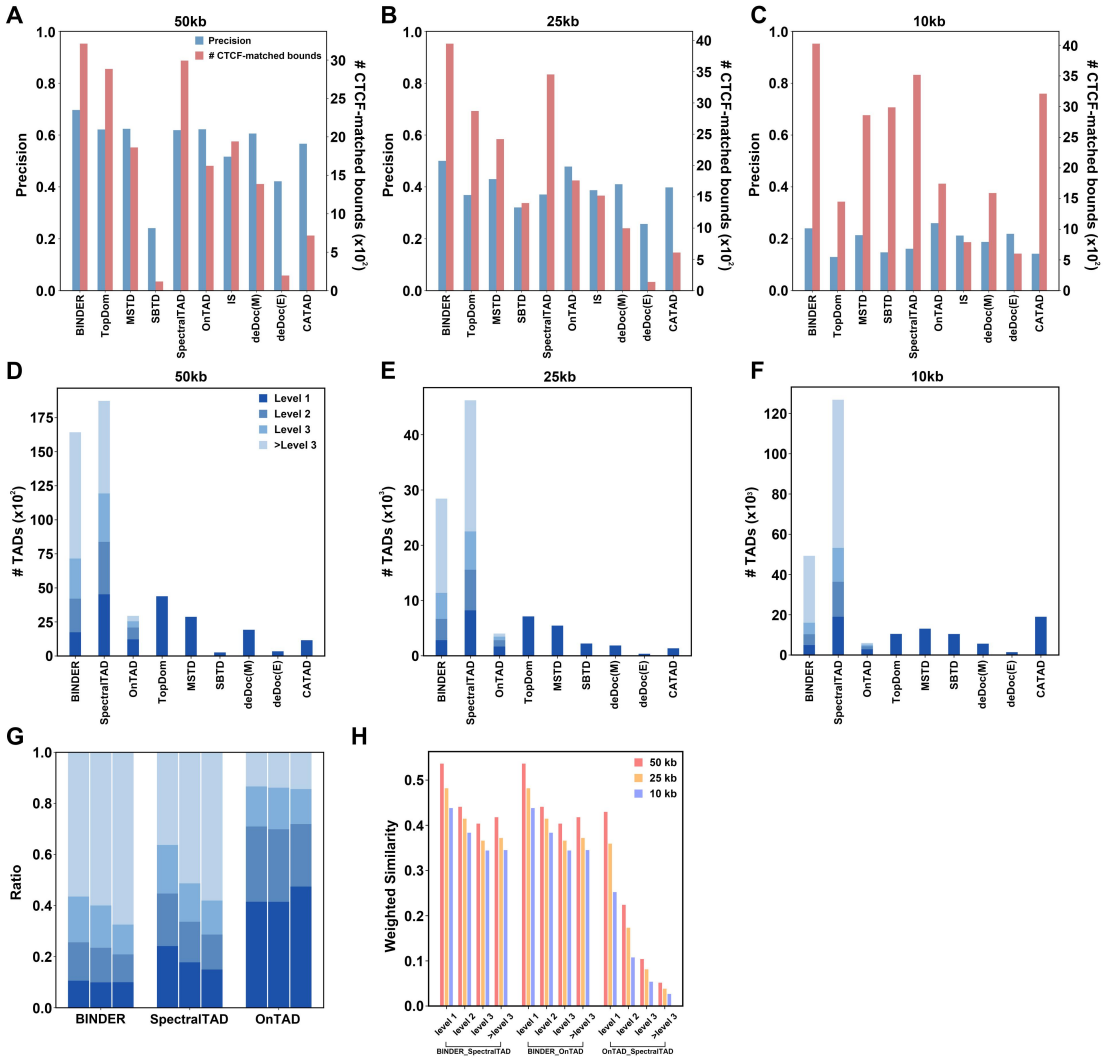

Fig. S8 Precision and the number of CTCF-matched TAD boundaries of BINDER and the other compared TAD callers on hg38 GM12878 data at resolutions of (**A**) 50kb, (**B**) 25kb, and (**C**) 10kb. Number of nested or non-nested TADs identified by BINDER and the other compared callers on hg38 GM12878 data at resolutions of (**D**) 50kb, (**E**) 25kb, and (**F**) 10kb. (**G**) Proportion of TADs at different levels identified by BINDER, SpectralTAD and OnTAD (three bars for each algorithm from left to right for the results at resolutions of 50kb, 25kb, and 10kb, respectively). (**H**) Weighted similarity between TADs of level 1, level 2, level 3, and level greater than 3 identified by BINDER, SpectralTAD, and OnTAD on the hg38 GM12878 data.

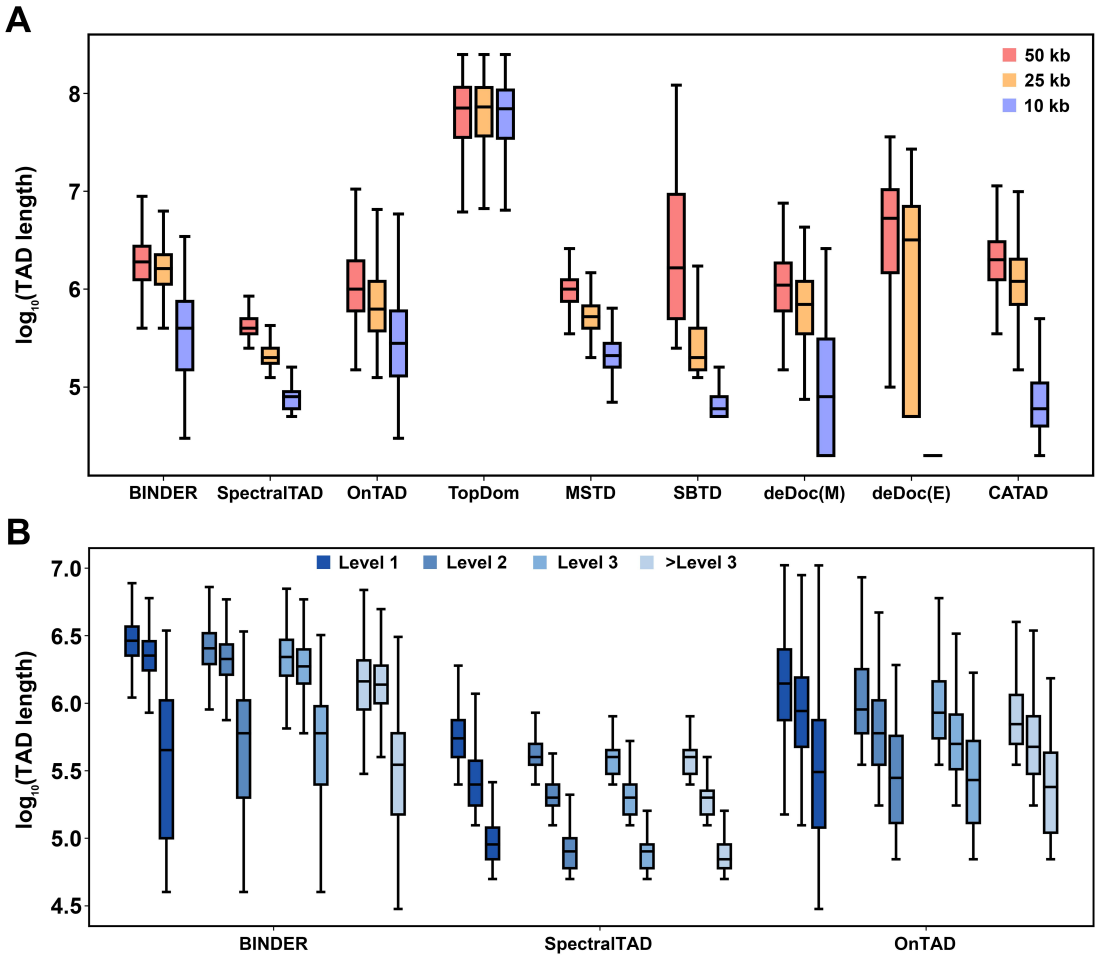

Fig. S9 (**A**) Length of TADs ($\log_{10}$ of bp length) identified by BINDER and the other compared TAD callers on hg38 GM12878 at resolutions of 50kb, 25kb, and 10kb. (**B**) Length of TADs ($\log_{10}$ bp length) identified by BINDER, SpectralTAD and OnTAD at different levels on hg38 GM12878 data (boxplots of the same color from left to right for the result of BINDER at resolutions of 50kb, 25kb, and 10kb respectively).

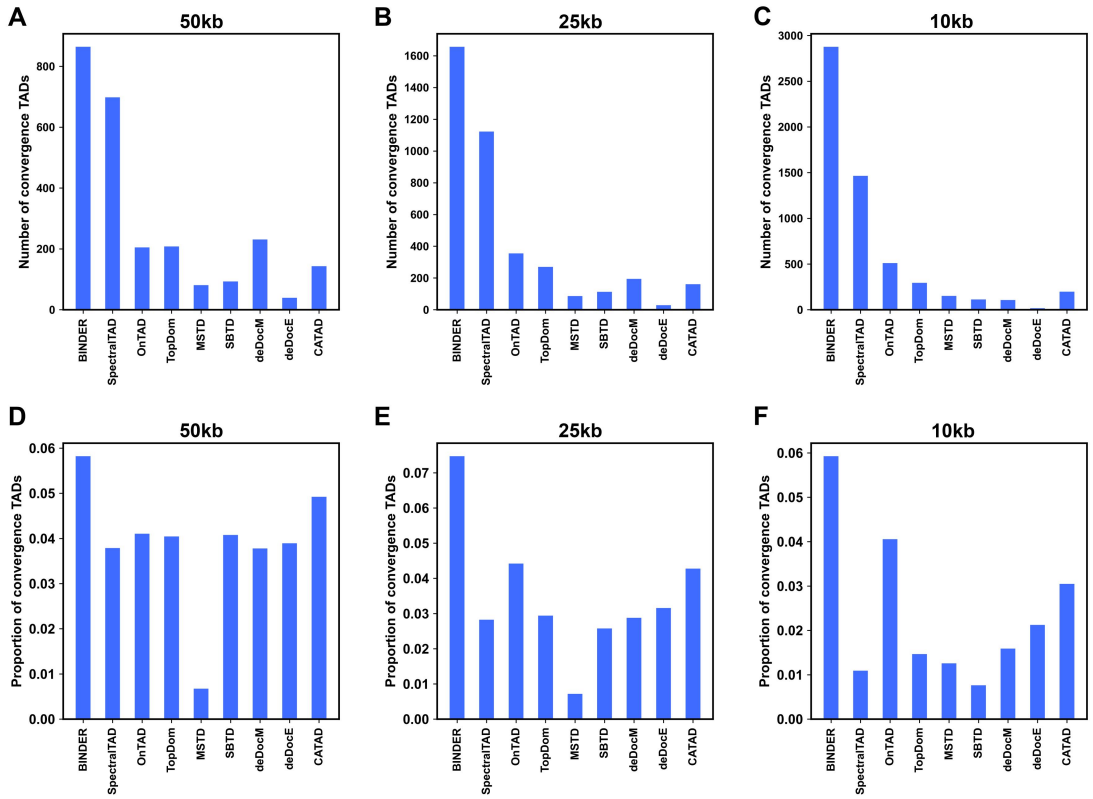

Fig. S10 Comparison of convergence TADs identified by BINDER and the other compared TAD callers on 50kb, 25kb, and 10kb GM12878 data in terms of (**A-C**) the number of convergence TADs and (**D-F**) the proportion of convergence TADs. We say that a TAD conforms to the principle of convergence if its left and right boundaries are matched by a CTCF peak from ChIP-seq data labeled "Forward" and "Reverse", respectively. And for convenience, we call a TAD a convergence TAD if it conforms to the principle of convergence. The proportion convergence TADs is defined by the number of convergence TADs divided by the number of all TADs.

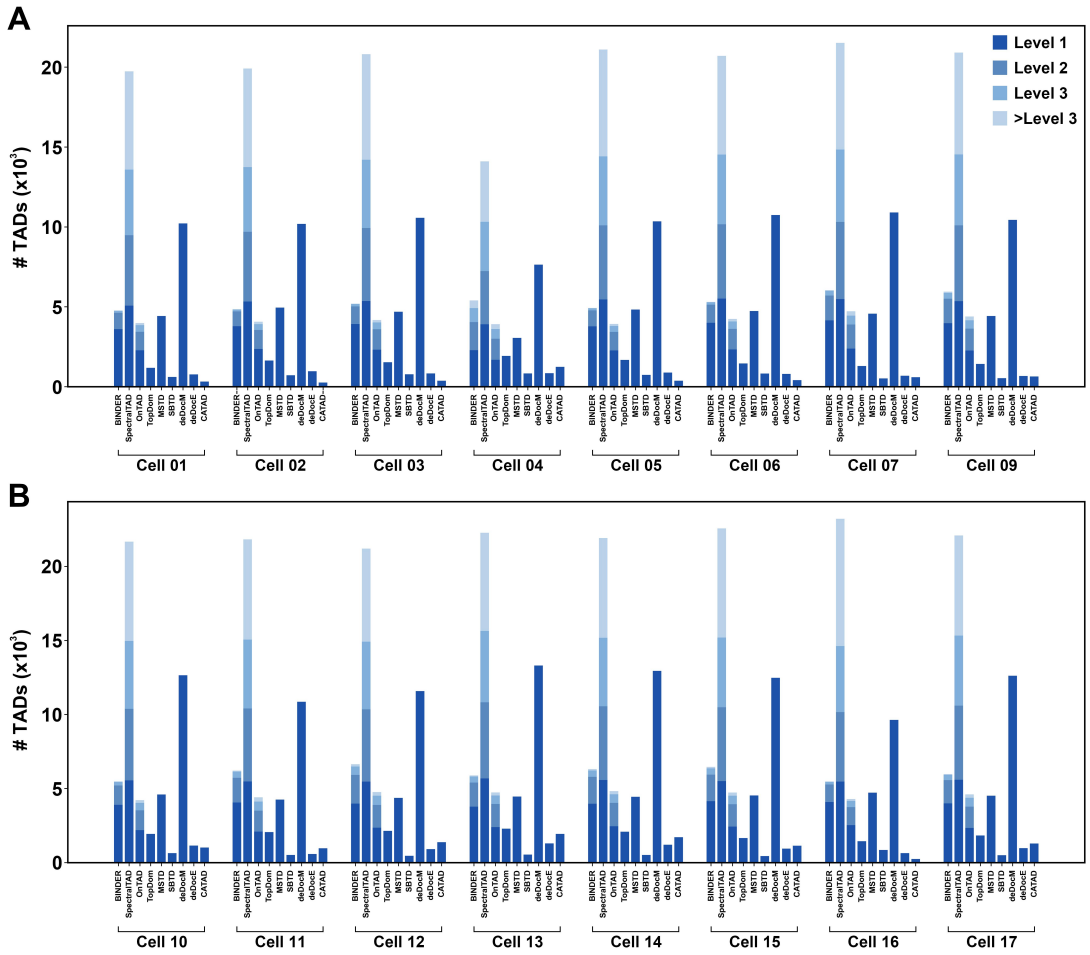

Fig. S11 Number of nested or non-nested TADs identified by BINDER and the other compared TAD callers on Dip-C data of 16 GM12878 single cells.

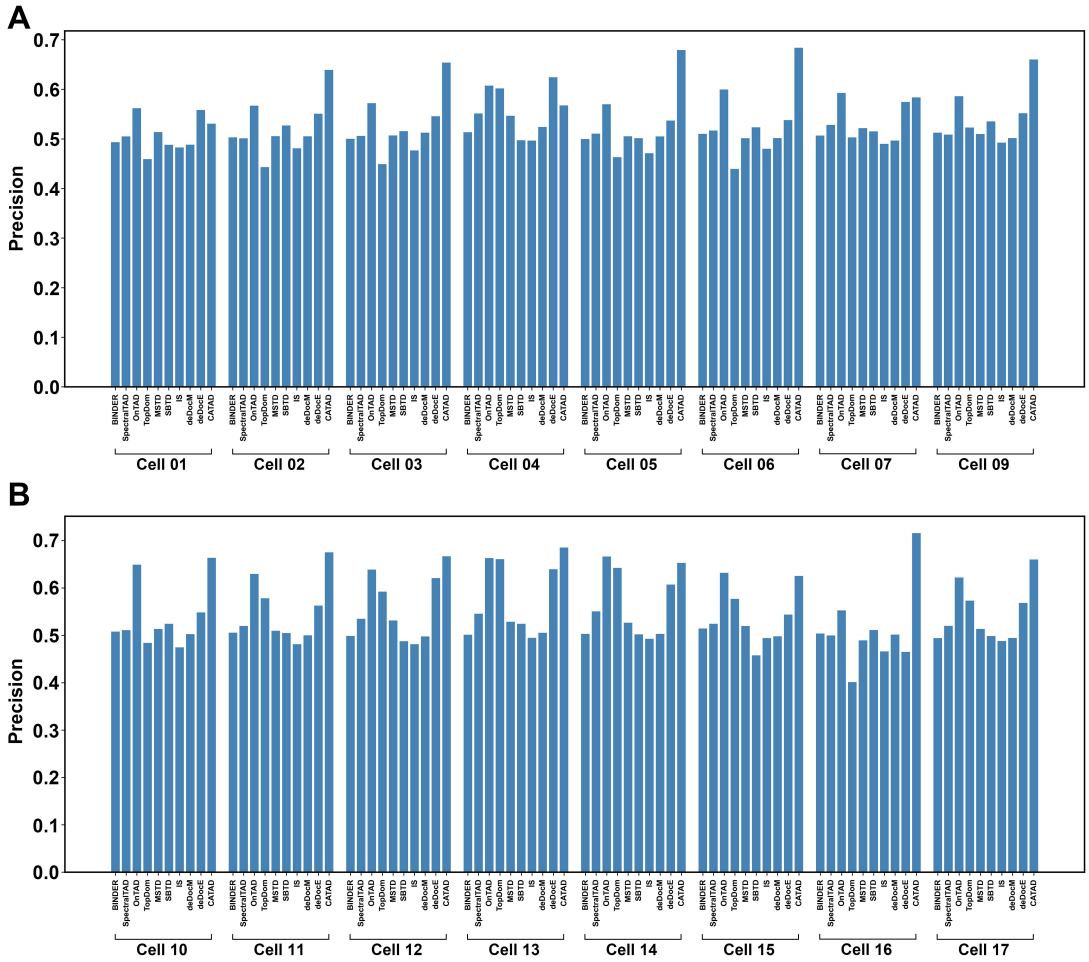

Fig. S12 Precision of TAD boundaries identified by BINDER and the other TAD callers on Dip-C data of 16 GM12878 single cells.

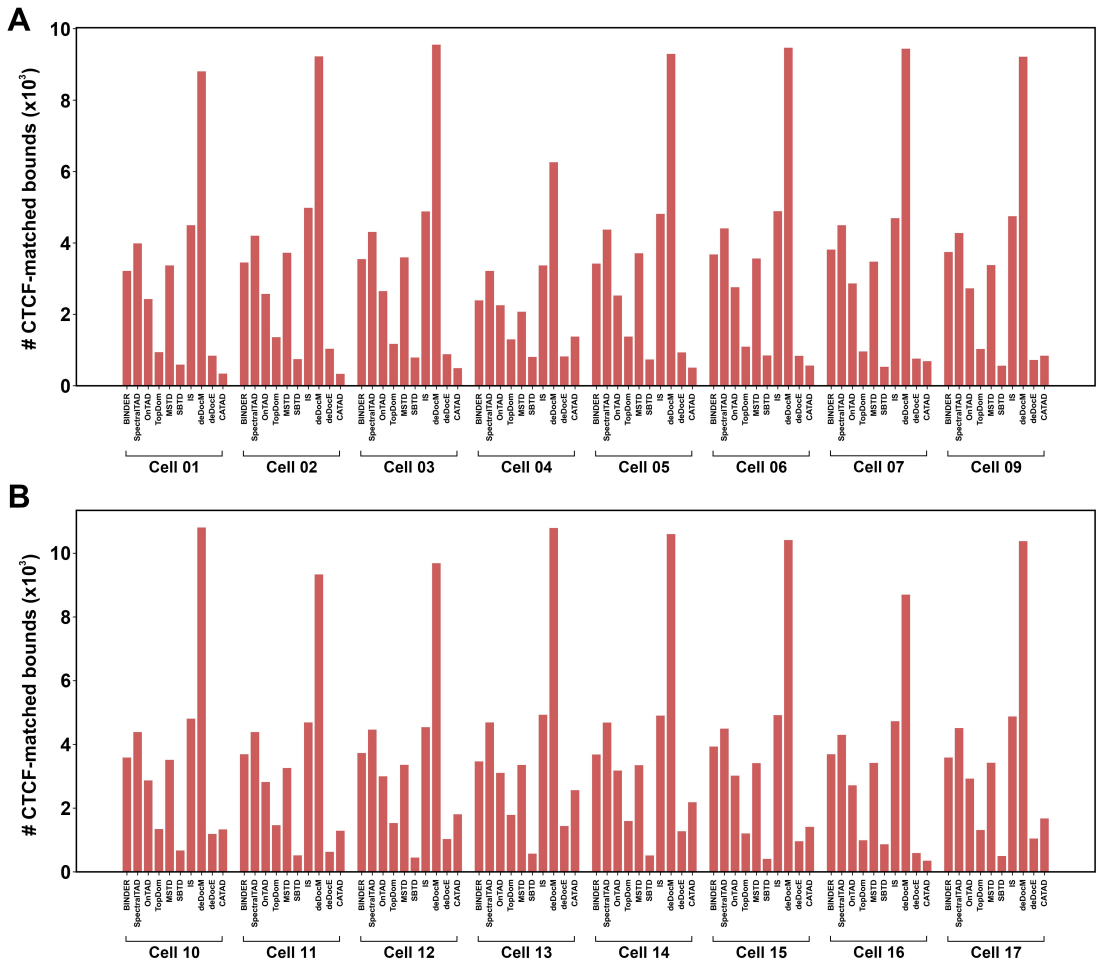

Fig. S13 Number of CTCF-matched boundaries identified by BINDER and the other TAD callers on Dip-C data of 16 GM12878 single cells.

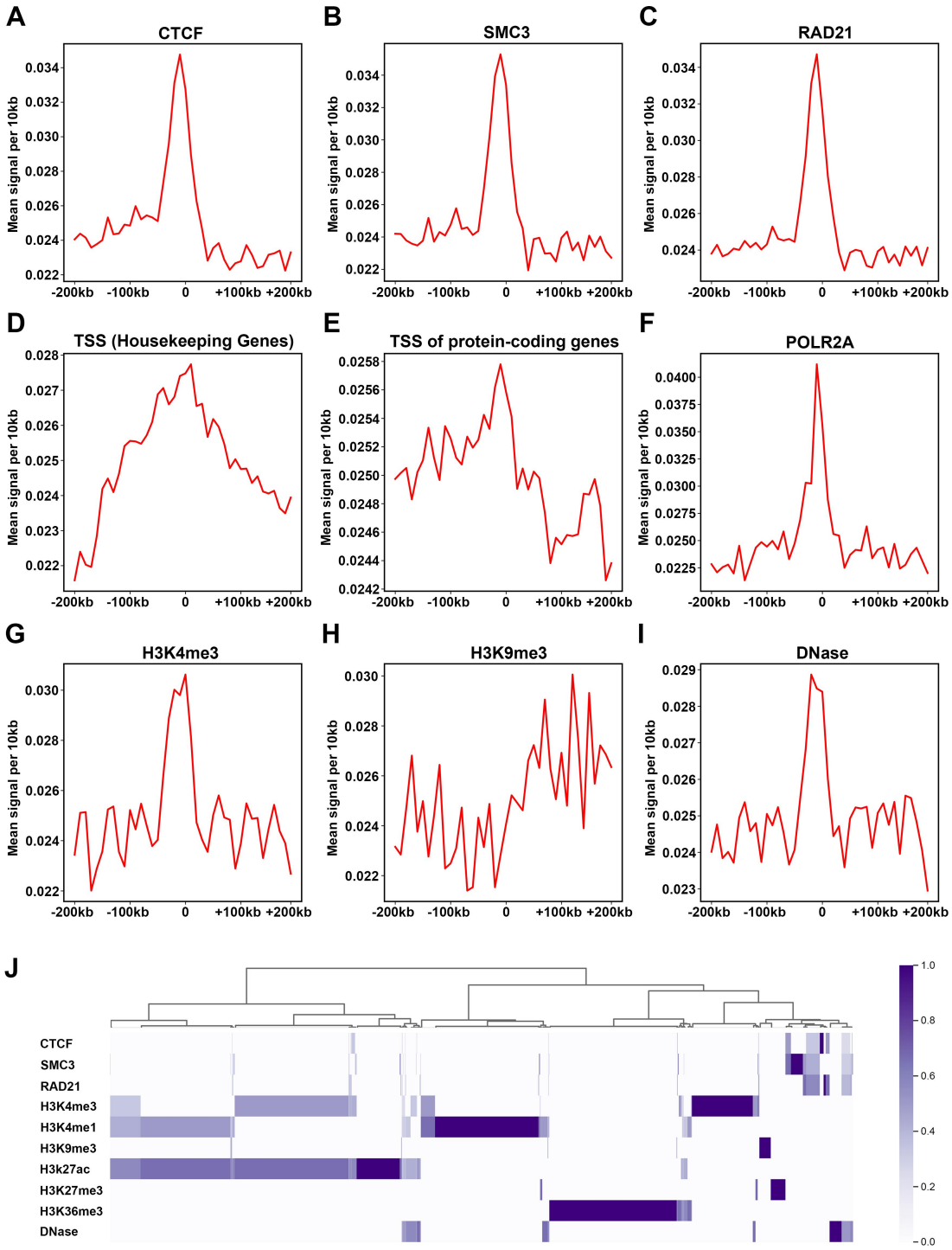

Fig. S14 Enrichment of epigenetic factors near TAD boundaries identified by BINDER on hg38 GM12878 data - (**A**) CTCF, (**B**) SMC3, (**C**) RAD21, (**D**) TSS of housekeeping genes, (**E**) TSS of protein-coding genes, (**F**) POLR2A, (**G**) H3K4me3, (**H**) H3K9me3, and (**I**) DNase. (**J**) Hierarchical clustering of hierarchical TAD boundaries based on nine epigenetic factors and DNase-seq, where each feature of the boundary is normalized to 0-1 (see details in "Hierarchical clustering for the enrichment of epigenetic factors near TAD boundaries" of the Supplemental Methods section). All of the results above were plotted based on the output of BINDER on hg38 GM12878 data at resolution of 10kb.

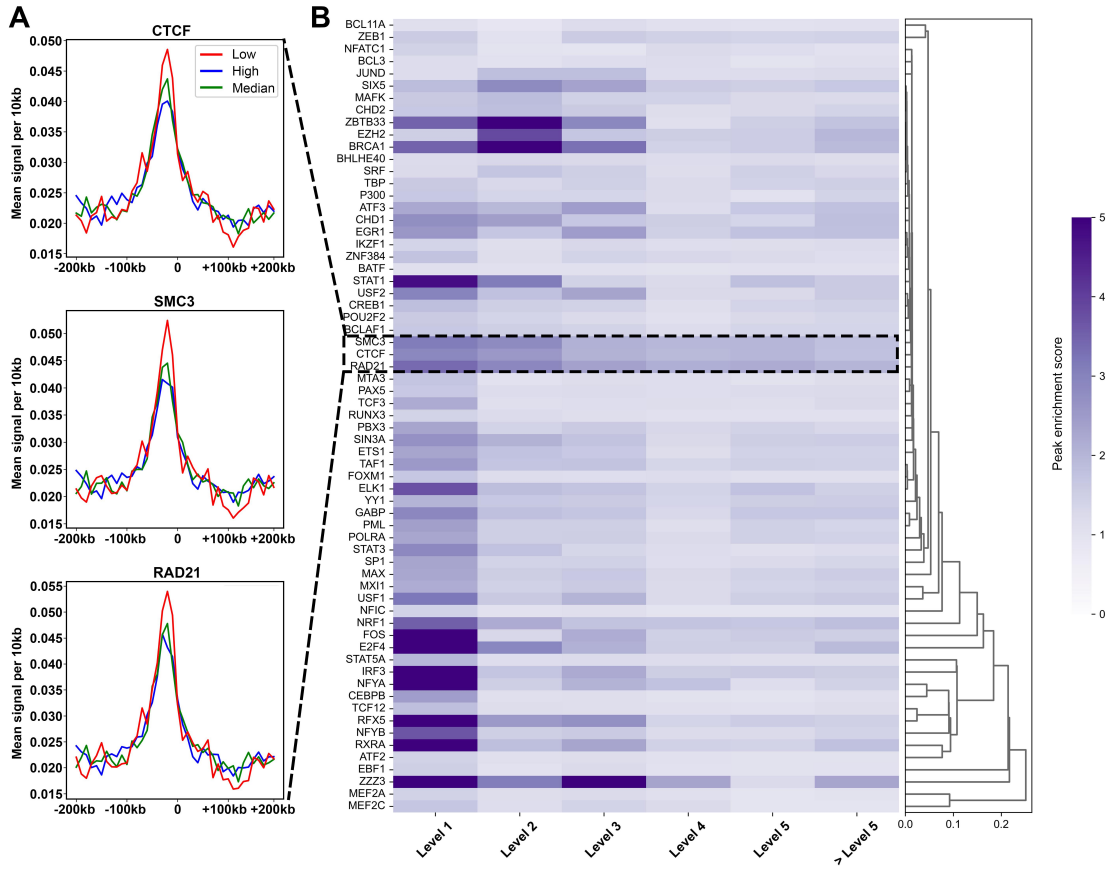

Fig. S15 (**A**) Enrichment of CTCF, SMC3, and RAD21 near TAD boundaries of low, medium and high levels, which is indicated by the red, green and blue lines, respectively (low: level<4, median: 3<level<6, high: level>5). (**B**) Hierarchical clustering of 66 TFs by their peak enrichment scores near hierarchical boundaries identified by BINDER, with darker colors indicating higher peak enrichment scores. All of the results above were plotted based on the output of BINDER on hg38 GM12878 data at resolution of 10kb.

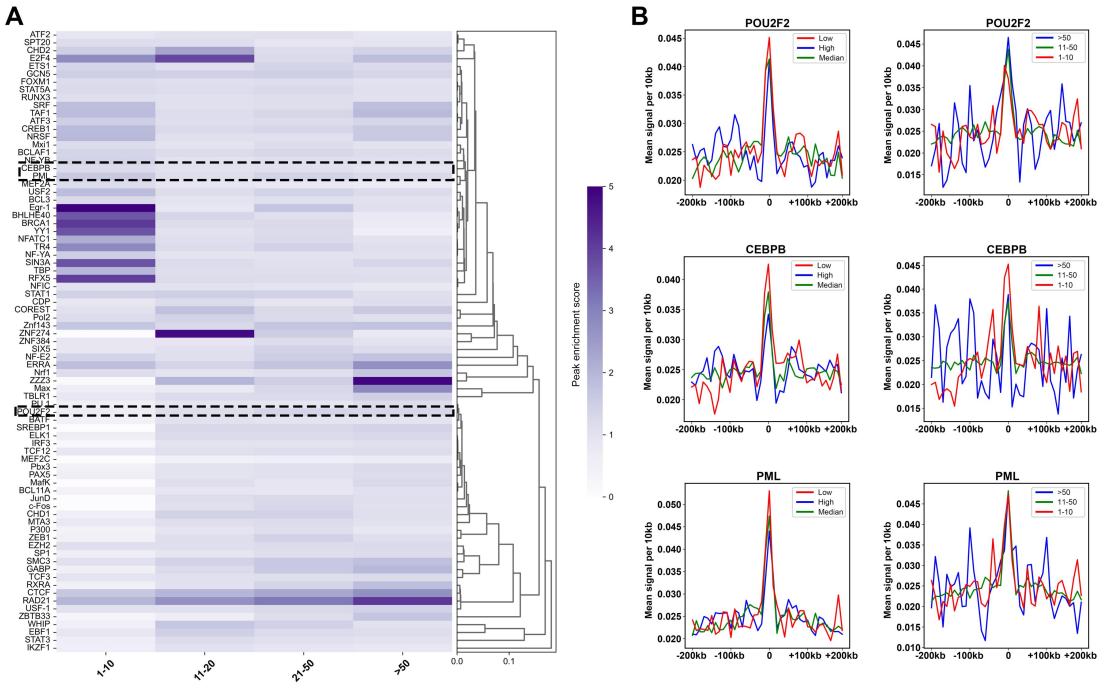

Fig. S16 Relationship between the enrichment of transcription factors (TFs) and the size of TAD. (**A**) Heatmap of the enrichment of TFs at boundaries of TADs of different sizes. (**B**) Enrichment of POU2F2, CEBPB, and PML (all highly enriched near level 2 TAD boundaries, see Fig. 4B) near boundaries of different levels (left) and near TADs of different sizes (right) (details about the definition of "size of TADs associated with boundaries" were described in Supplemental Methods).

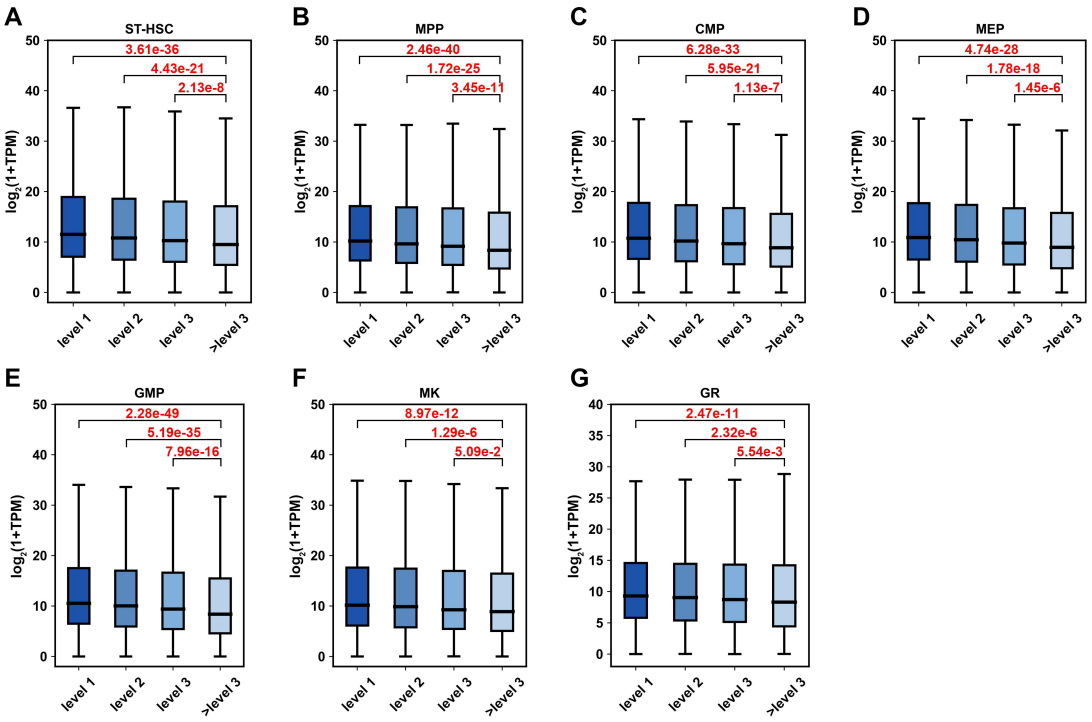

Fig. S17 Boxplots of gene expression ($\log_2(1 + \text{TPM})$) within TADs of different levels (level 1, level 2, level 3, and > level 3) of (**A**) ST-HSC, (**B**) MPP, (**C**) CMP, (**D**) MEP, (**E**) GMP, (**F**) MK, and (**G**) GR (*P*-values: Wilcoxon rank sum test).

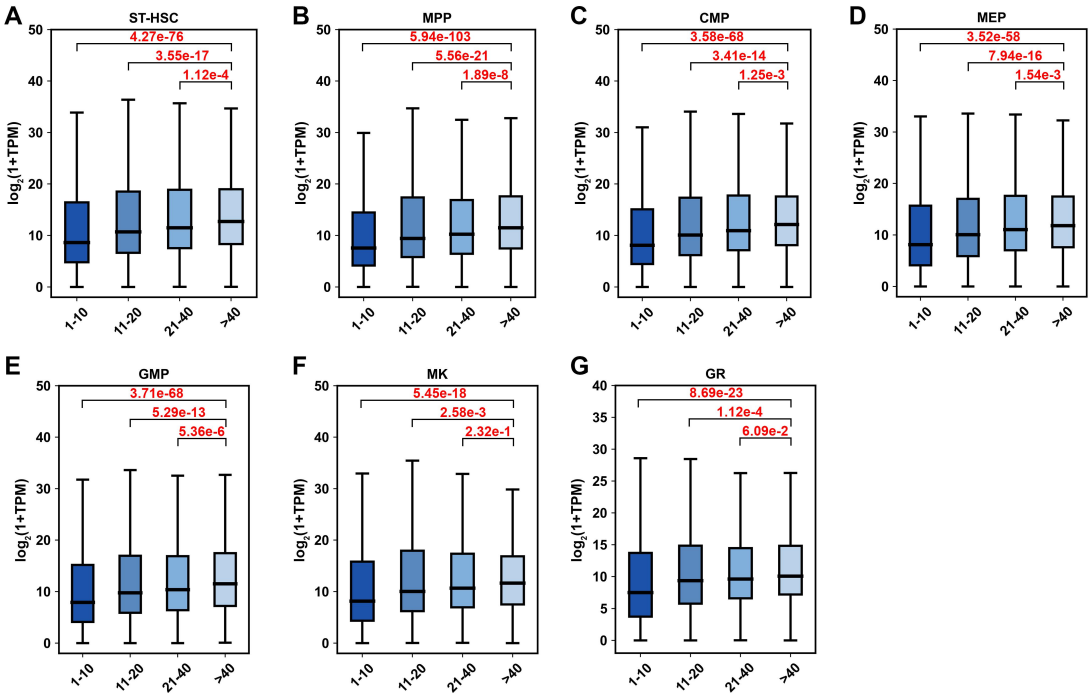

Fig. S18 Boxplots of gene expression ($\log_2(1 + \text{TPM})$) within TADs of different lengths (1-10, 11-20, 21-40, >40, bin length) of (**A**) ST-HSC, (**B**) MPP, (**C**) CMP, (**D**) MEP, (**E**) GMP, (**F**) MK, and (**G**) GR (*P*-values: Wilcoxon rank sum test).

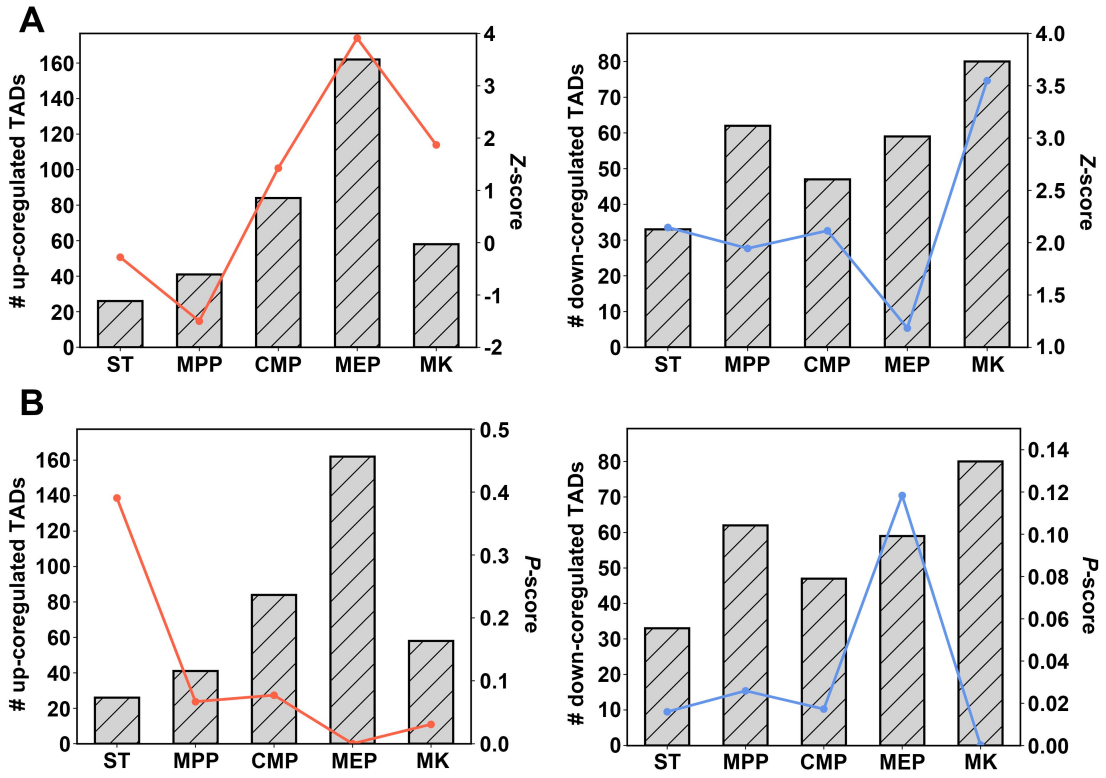

Fig. S19 (**A**) Number of up-coregulated and down-coregulated TADs in MK path and corresponding *Z*-scores for each cell type. (**B**) Number of up-coregulated and down-coregulated TADs in MK path and corresponding *P*-values for each cell type.

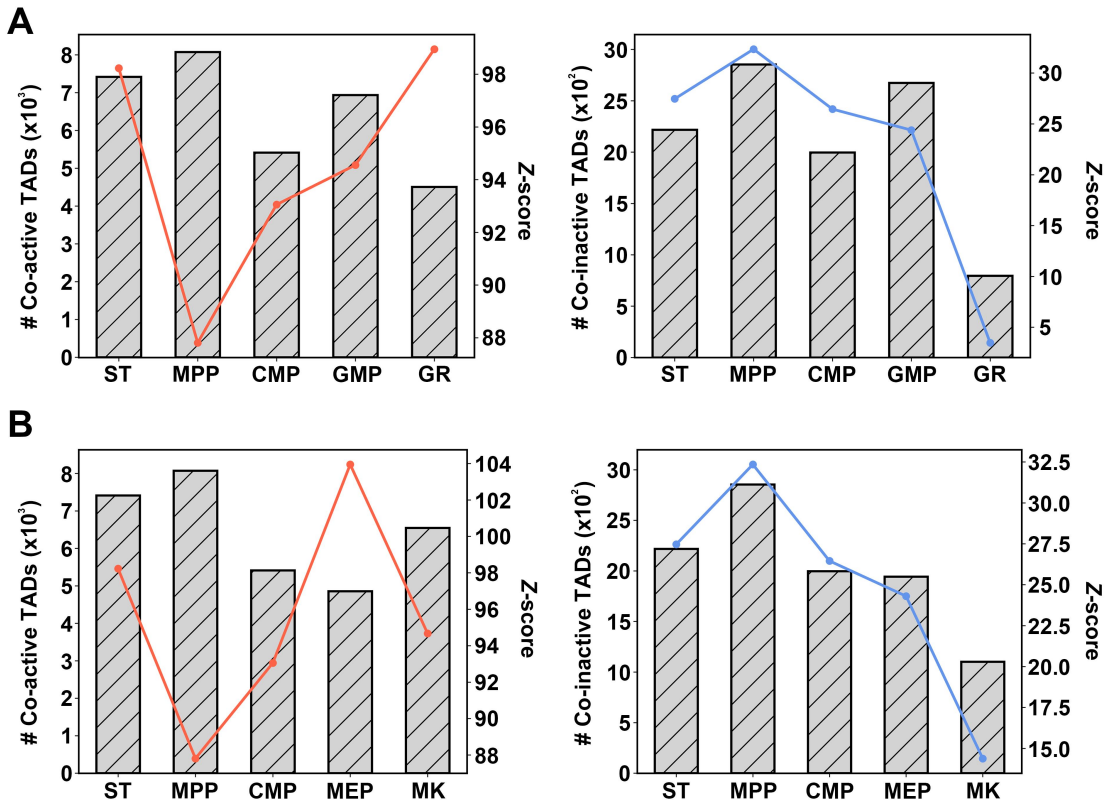

Fig. S20 (**A**) Number of co-active and co-inactive TADs in GR path and corresponding *Z*-scores for each cell type. (**B**) Number of co-active and co-inactive TADs in MK path and corresponding *Z*-scores for each cell type.

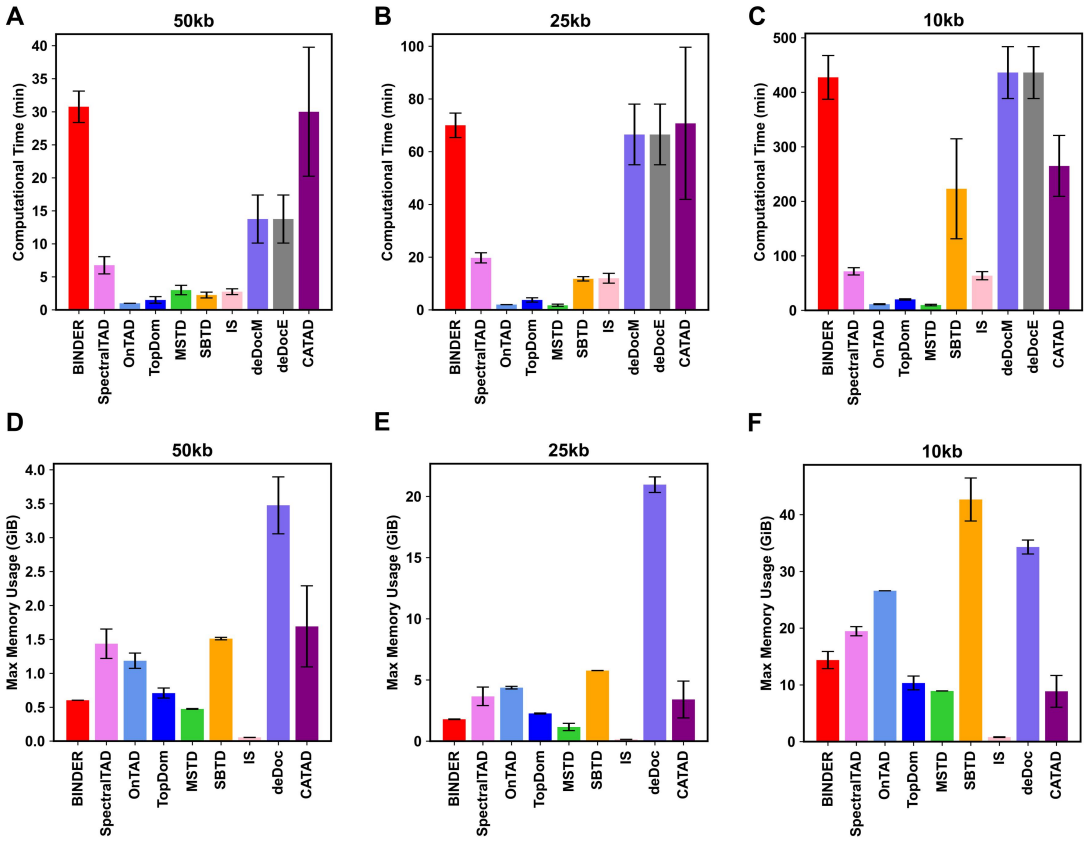

Fig. S21 Computational time of running BINDER and the other compared TAD callers in four cell lines at resolutions of (**A**) 50kb, (**B**) 25kb, and (**C**) 10kb of the four cell lines. Maximum memory usage of running BINDER and the other compared TAD callers in four cell lines at resolutions of (**D**) 50kb, (**E**) 25kb, and (**F**) 10kb of the four cell lines.

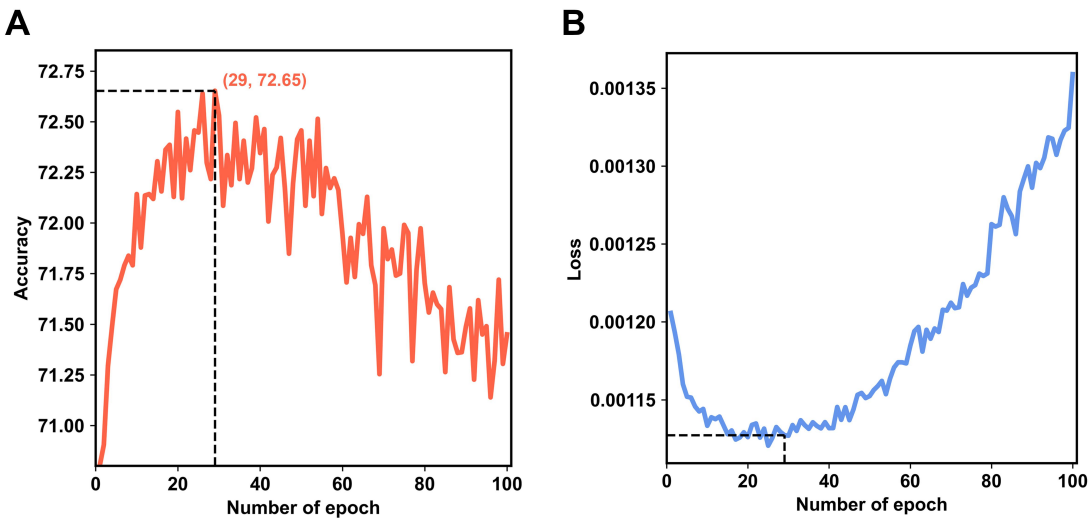

Fig. S22 Performance of the neural network during the training of it. (**A**) Accuracy and (**B**) loss curves in the validation set are drawn.

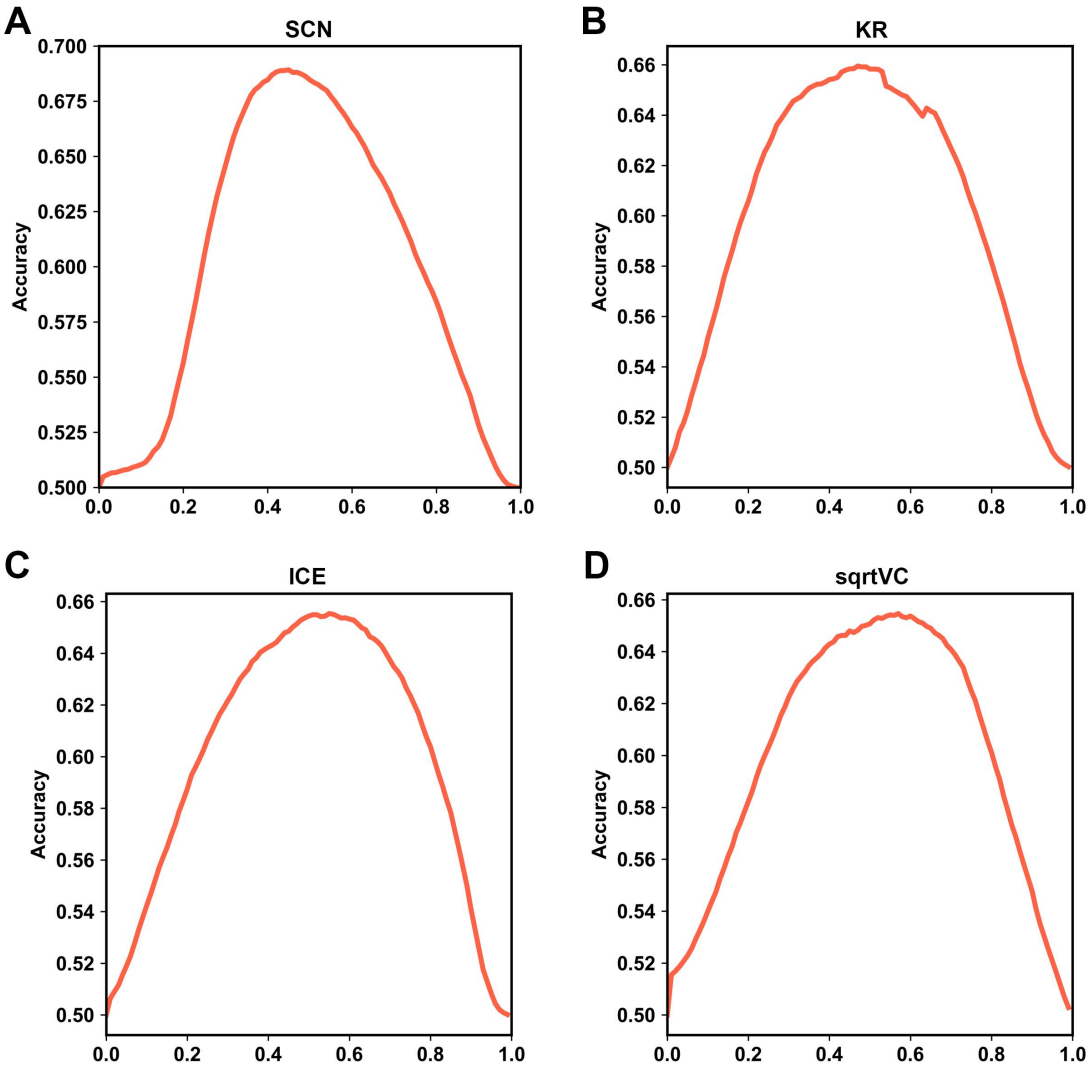

Fig. S23 Accuracy of 4 trained neural network models based on SCN, KR, ICE, and sqrtVC on four individual test sets generated from Hi-C data normalized by (**A**) SCN, (**B**) KR, (**C**) ICE, and (**D**) sqrtVC methods, respectively, where x-axis represents threshold for determining the label (0 or 1) of a dual boundary.

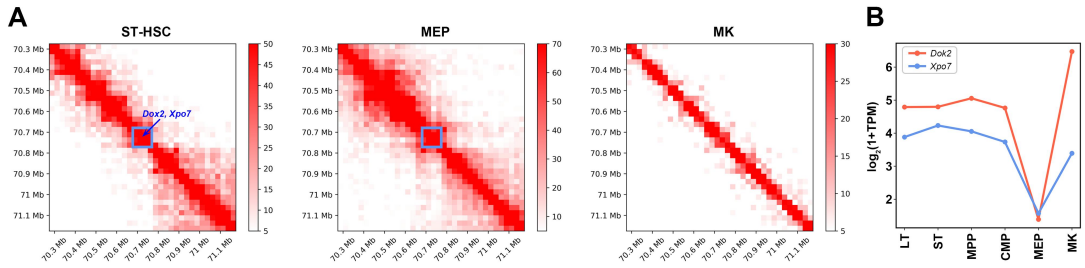

Fig. S24 An example of how breaking a down-coregulated TAD relates to a cell-type transition. (**A**) Hi-C sub-heatmaps of ST, MEP, and MK cell types containing a down-coregulated TAD example in MEP containing *Xpo7* and *Dok2* genes. This TAD is indicated by a blue box. (**B**) Gene expression of *Xpo7* and *Dok2* genes in MK path.

Supplemental References

Coppin E, De Grandis M, Pandolfi PP, Arcangeli ML, Aurrand-Lions M, Nunes JA. 2016. Dok1 and Dok2 Proteins Regulate Cell Cycle in Hematopoietic Stem and Progenitor Cells. *J Immunol* **196**: 4110-4121.

Cournac A, Marie-Nelly H, Marbouty M, Koszul R, Mozziconacci J. 2012. Normalization of a chromosomal contact map. *BMC Genomics* **13**: 436.

Cresswell KG, Stansfield JC, Dozmorov MG. 2020. SpectralTAD: an R package for defining a hierarchy of topologically associated domains using spectral clustering. *BMC Bioinformatics* **21**: 319.

Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. *Trends Genet* **29**: 569-574.

Flyamer IM, Gassler J, Imakaev M, Brandao HB, Ulianov SV, Abdennur N, Razin SV, Mirny LA, Tachibana-Konwalski K. 2017. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544**: 110-114.

Frankish A, Carbonell-Sala S, Diekhans M, Jungreis I, Loveland JE, Mudge JM, Sisu C, Wright JC, Arnan C, Barnes I et al. 2023. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res* **51**: D942-D949.

Hattangadi SM, Martinez-Morilla S, Patterson HC, Shi J, Burke K, Avila-Figueroa A, Venkatesan S, Wang J, Paulsen K, Gorlich D et al. 2014. Histones to the cytosol: exportin 7 is essential for normal terminal erythroid nuclear maturation. *Blood* **124**: 1931-1940.

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012a. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods* **9**: 999-1003.

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012b. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9**: 999-1003.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**: D493-496.

Knight PA, Ruiz D. 2012. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis* **33**: 1029-1047.

Kumar R, Sobhy H, Stenberg P, Lizana L. 2017. Genome contact map explorer: a platform for the comparison, interactive visualization and analysis of genome contact maps. *Nucleic Acids Research* **45**: e152-e152.

Li A, Yin X, Xu B, Wang D, Han J, Wei Y, Deng Y, Xiong Y, Zhang Z. 2018a. Decoding topologically associating domains with ultra-low resolution Hi-C data by graph structural entropy. *Nature Communications* **9**.

Li A, Yin X, Xu B, Wang D, Han J, Wei Y, Deng Y, Xiong Y, Zhang Z. 2018b. Decoding topologically associating domains with ultra-low resolution Hi-C data by graph structural entropy. *Nat Commun* **9**: 3265.

Li X, Zeng G, Li A, Zhang Z. 2021. DeTOKI identifies and characterizes the dynamics of chromatin TAD-like domains in a single cell. *Genome Biol* **22**: 217.

Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, Myers Z, Sud P, Jou J, Lin K et al. 2020. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* **48**: D882-D889.

Nanni L, Ceri S, Logie C. 2020. Spatial patterns of CTCF sites define the anatomy of TADs and their boundaries. *Genome Biol* **21**: 197.

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**: 1665-1680.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**: e47-e47.

Rosvall M, Bergstrom CT. 2008. Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci U S A* **105**: 1118-1123.

Seita J, Weissman IL. 2010. Hematopoietic stem cell: self-renewal versus differentiation. *Wiley Interdiscip Rev Syst Biol Med* **2**: 640-653.

Szabo Q, Bantignies F, Cavalli G. 2019. Principles of genome folding into topologically associating domains. *Science Advances* **5**.

Yasuda T, Shirakata M, Iwama A, Ishii A, Ebihara Y, Osawa M, Honda K, Shinohara H, Sudo K, Tsuji K et al. 2004. Role of Dok-1 and Dok-2 in myeloid homeostasis and suppression of leukemia. *J Exp Med* **200**: 1681-1687.

Zhan Y, Mariani L, Barozzi I, Schulz EG, Bluthgen N, Stadler M, Tiana G, 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.

Zhang S, Krieger JM, Zhang Y, Kaya C, Kaynak B, Mikulska-Ruminska K, Doruker P, Li H, Bahar I. 2021. ProDy 2.0: increased scale and scope after 10 years of protein dynamics modelling with Python. *Bioinformatics* **37**: 3657-3659.

Zhang Y, Gao S, Xia J, Liu F. 2018. Hematopoietic Hierarchy - An Updated Roadmap. *Trends Cell Biol* **28**: 976-986.

Zheng H, Xie W. 2019. The role of 3D genome organization in development and cell differentiation. *Nat Rev Mol Cell Biol* **20**: 535-550.