Supplemental Material for

Defining the separation landscape of topological domains for decoding consensus domain organization of 3D genome

Dachang Dang[1], Shao-Wu Zhang[1*], Ran Duan[2], Shihua Zhang[3,4,5,6*]

[1]Key Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xi'an 710072, China;

[2]Department of Software Engineering, Yunnan University, Kunming 650500, China;

[3]NCMIS, CEMS, RCSDS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China;

[4]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China;

[5]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China;

[6]Key Laboratory of Systems Health Science of Zhejiang Province, School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Hangzhou 310024, China.

*To whom correspondence should be addressed. Tel/Fax: +86 029 88431308; +86 01 82541360; Emails: zhangsw@nwpu.edu.cn and zsh@amss.ac.cn.

Supplementary Figures

Supplementary Tables

Additional notes on the method section

Comparison of different TAD-calling methods

We compared TADs identified by 16 TAD-calling methods on seven human cell lines by calculating some metrics between domain sets or boundary sets, such as the absolute differences of TAD number and average size, Jaccard index and the Measure of Concordance.

Jaccard index between boundary sets. Here we introduced a modified Jaccard index to evaluate the similarity between two sets of TAD boundaries A and B:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

where $0 \leq J(A,B) \leq 1$. Here, we adopt a tolerance radius of 1 bin when defining the intersection between two sets of TAD boundaries. If the distance between two boundaries doesn't exceed 1 bin, they will be determined as the shared ones between boundary sets.

Measure of Concordance (MoC) between domain sets. MoC was introduced by Zufferey et al. (Zufferey et al. 2018) to compare TAD partitions and it is defined as follows:

$$MoC(P,Q) = \begin{cases} 1, & if\ N_p = N_Q = 1; \\ \frac{1}{\sqrt{N_p N_Q} - 1}\left(\sum_{i=1}^{N_p}\sum_{j=1}^{N_Q}\frac{\|F_{i,j}\|^2}{\|P_i\|\|Q_j\|} - 1\right), & \text{otherwise} \end{cases}$$

where P and Q are two domain sets, N_p and N_Q represent the number of TADs contained in these two sets. P_i and Q_j are two TADs with domain lengths of $\|P_i\|$ and $\|Q_j\|$, where $\|F_{i,j}\|$ denotes the overlap length between P_i and Q_j. Here the length indicates the number of base pairs of DNA sequence.

We calculated these metrics mainly for comparison of TADs across methods or across datasets. For comparison across methods, we kept the dataset the same and calculated the metrics between TADs identified by different methods. For comparison across datasets, we kept the method the same and calculated the metrics between TADs found by this method on different Hi-C datasets.

Enrichment of structural proteins at TAD boundaries. For the structural proteins (e.g., CTCF, RAD21, and SMC), we counted the binding peaks in each 10-kb bin and built up the profile of the median values around the boundaries identified by a certain method. Then we computed the fold change between their binding peaks at TAD boundaries versus adjacent flanking regions.

Analyses of TAD separation landscapes of multiple cell lines

For each bin along the genome, we calculated the average boundary scores among seven cell lines based on their TAD separation landscapes. We sorted these bins by the average scores in ascending order, and selected some quartiles, including 0%, 30%, 44%, 58%, 72%, 86%, and 100%, to divide all bins into six levels. The bins contained in the first level (0~30%) have a boundary score of 0 and bins with higher levels would have larger boundary scores. For bins with different levels, we computed the number of cell lines in which they have non-zero boundary scores and counted the number of housekeeping genes in each bin, as well as the average number of CTCF binding peaks across cell lines, and we also calculated the average phastCons scores for DNA sequences of these CTCF binding peaks (**Supplemental Fig. S6A**). Besides, we clustered the Hi-C samples from different cell lines based on the Pearson correlations between their TAD separation landscapes (**Fig. 4B**) or 1D indicators like DI, IS, and CI (**Supplemental Fig. S6B**).

Two metrics used for the identification of three types of boundary regions

The within-cluster sum of squared error (WCSSE) is the sum of the squared differences between each sample and its cluster center across all clusters:

$$WCSSE(X, C) = \sum_{i=1}^{K} \sum_{x_j \in C_i} \left(x_j - \mu_i\right)^2$$

where K denotes the cluster number, $X = \{x_1, x_2, \dots, x_n\}$ denotes the set of boundary regions, C_i is the i-th cluster of *k*-means and μ_i is the cluster center. The number of clusters that minimize the sum of squared error can be viewed as optimal.

The silhouette coefficient (SC) is a measure of how similar a sample is to its own cluster compared to other clusters, i.e.,

$$SC = \frac{1}{n} \sum_{i=1}^{n} \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

where $a(i) = \frac{1}{|C_i|-1} \sum_{x_j \in C_i, j \neq i} \left(x_i - x_j\right)^2$ reflects the average distance for sample i within a cluster and $b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{x_j \in C_k} \left(x_i - x_j\right)^2$ reflects the minimum average distance between the sample in other clusters for sample i. The number of clusters that maximize the silhouette coefficient can be viewed as the optimal one.

Overlap between three types of boundary regions and boundaries reported by individual TAD-calling method

While other single methods only report the position of a TAD boundary (usually a bin with a fixed length). Therefore, we can't divide their boundaries into three types as we defined. But for each boundary region we got, we examined whether it overlapped with the boundary reported by a single method (**Supplemental Fig. S14E**). We found that all three types of boundary regions showed different overlap ratios among different methods. All methods had a low ratio for the Narrow-weak type, indicating that they might fail to detect such boundary regions.

Comparison of ConsTADs and other TAD-calling methods

After obtaining the ConsTADs, we compared them with the results of other TAD-calling methods in two aspects and we found that ConsTADs could reveal more domains with H3K36me3/H3K27me3 differential signal and had the highest agreement with other methods in the classification of DNA replication domains.

Identification of topological domains with significant H3K36me3/H3K27me3 differential signal

We calculated the average signal of H3K36me3 and H3K27me3 in each 50-kb bin along Chromosome 2 in GM12878 and then got the fold change of signal for each bin by dividing the mean signal across the whole chromosome. For each bin, we computed the ratio between the H3K36me3 and H3K27me3 fold change, similarly to Zufferey et al. (Zufferey et al. 2018), we termed them as LR values or LR intervals and a positive value indicates a bias to H3K36me3, while a negative one indicates a bias to H3K27me3, and these biases usually stay the same in some consecutive intervals and the positions with altered bias are recorded. We collected the topological domains identified by all 16 TAD-calling methods as well as the ConsTADs we defined and calculated the average LR values within each domain and shuffled the LR intervals 1000 times to derive a null distribution of LR values within domains. For each domain, an empirical *p*-value can be calculated by comparing its observed average LR values with the null

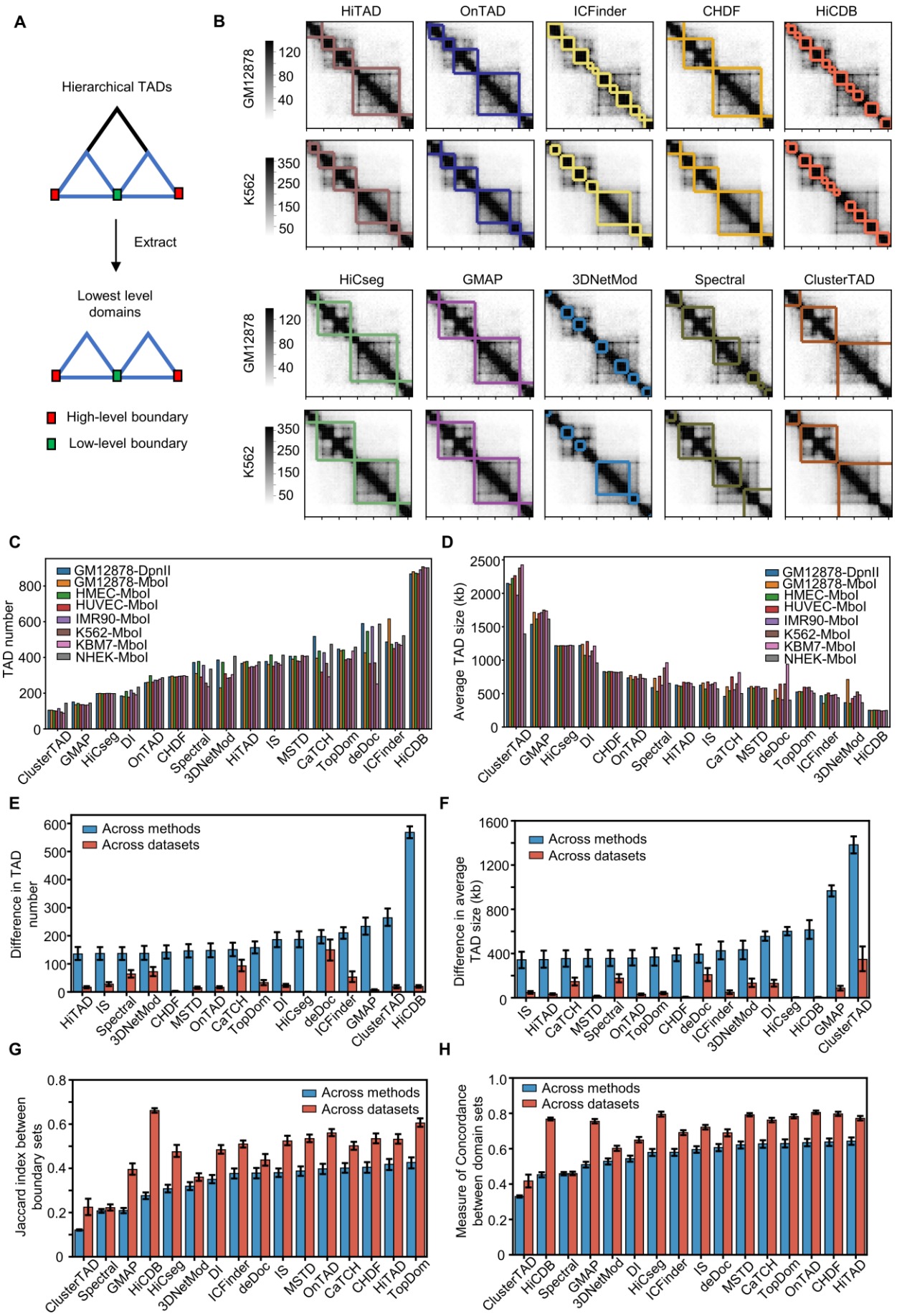
distribution. The *p*-values for all the domains identified by each TAD-calling method were corrected using the Benjamini-Hochberg procedure and domains with a corrected *p*-value smaller than 0.1 were considered with significant H3K36me3/H3K27me3 differential signal. Domains under 150 kb (three bins at 50 kb resolution) in length were excluded from this analysis because they were too short (**Supplemental Fig. S15C**). We also calculated the distances from all bias-changing points to the nearest domain boundary for each method (**Supplemental Fig. S15D**).

Consistency score of the domain replication cluster

Just like the ConsTADs, for each method, the topological domains were divided into five clusters based on their DNA replication timing, and bins within these domains were also assigned to a certain cluster. Then for each bin along the chromosome, a consistency score of domain replication cluster was calculated as the proportion of methods in which the bin has the same cluster assignment as the current method (**Supplemental Fig. S15F**).

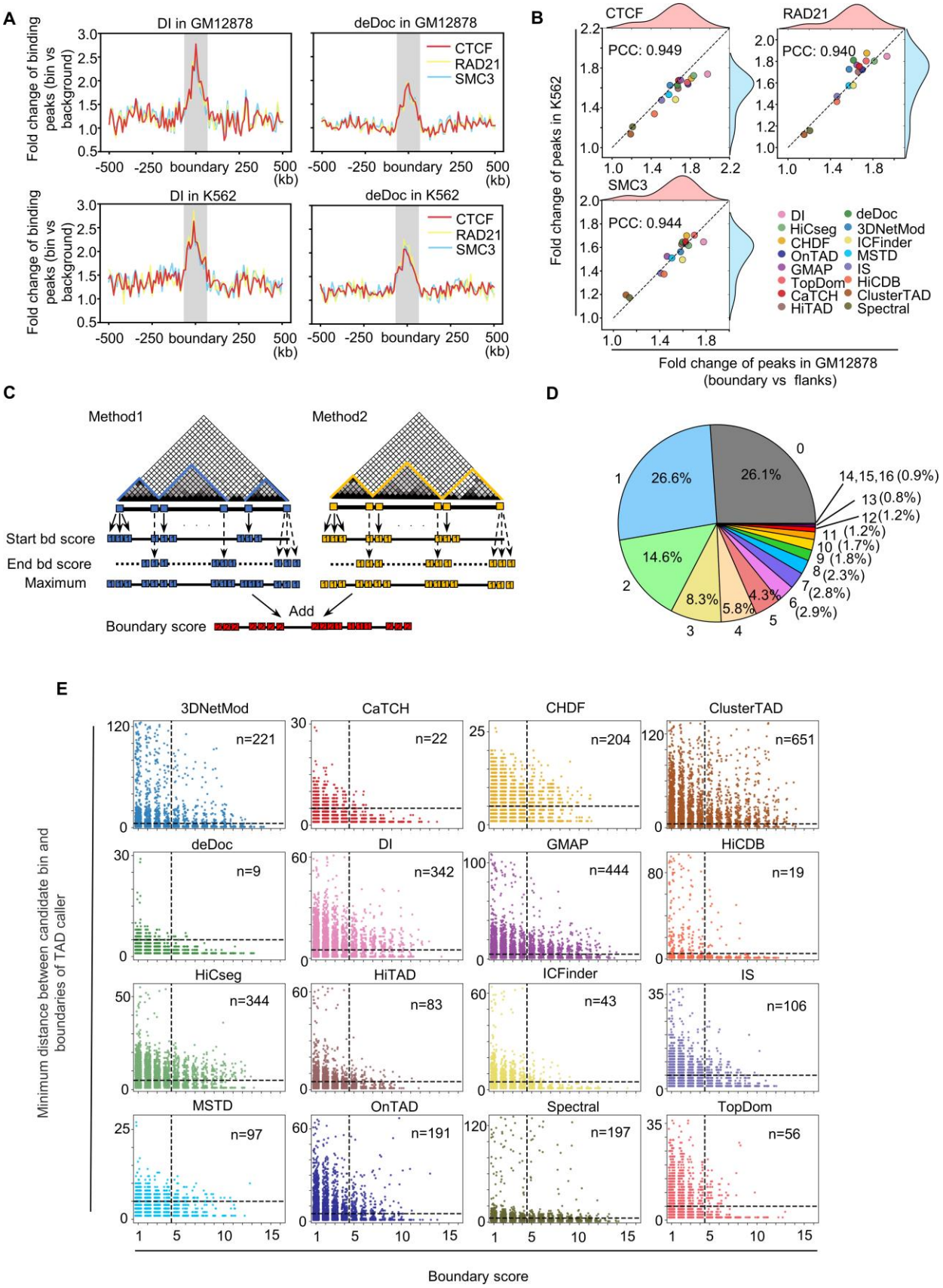
Exploring boundary probabilities of boundary regions at the single-cell level

We obtained the 3D coordinates for probes representing 250-kb loci on Chromosome 2 of IMR90 single cells and these coordinates are separated for each imaged chromosomal copy. We first calculated the 3D spatial distances between pairs of imaged chromatin loci and constructed the spatial distance matrices for every single chromosome. We then identified chromatin domains in single chromosomes following the procedures proposed by (Su et al. 2020). In this way, we can assign a boundary probability to each probe, indicating the number of single chromosomes that consider as the domain boundary over the total number of chromosomes. We then identified the boundary regions on Chromosome 2 of IMR90 with a Hi-C contact map with 50 kb resolution and used the LiftOver (Hinrichs et al. 2006) to convert all 50-kb bins from human genome assembly hg19 to hg38. For each probe representing a 250-kb locus, we assigned it a label corresponding to the type of boundary region that covered the largest proportion of it, and if a probe did not overlap with any boundary regions, it would be defined as non-boundary (**Supplemental Fig. S22E**). Besides, for every single chromosome, if the distance between two loci is below 500nm, they will be considered as contacting each other. Thus, we got the overall proximity frequency matrix of all these single chromosomes by dividing the contact frequency of each pair of loci by the total number of single chromosomes (**Supplemental Fig. S22F**).



Supplemental Fig. S1. Comparison of TADs identified by the 16 TAD-calling methods on Hi-C data from different cell lines.

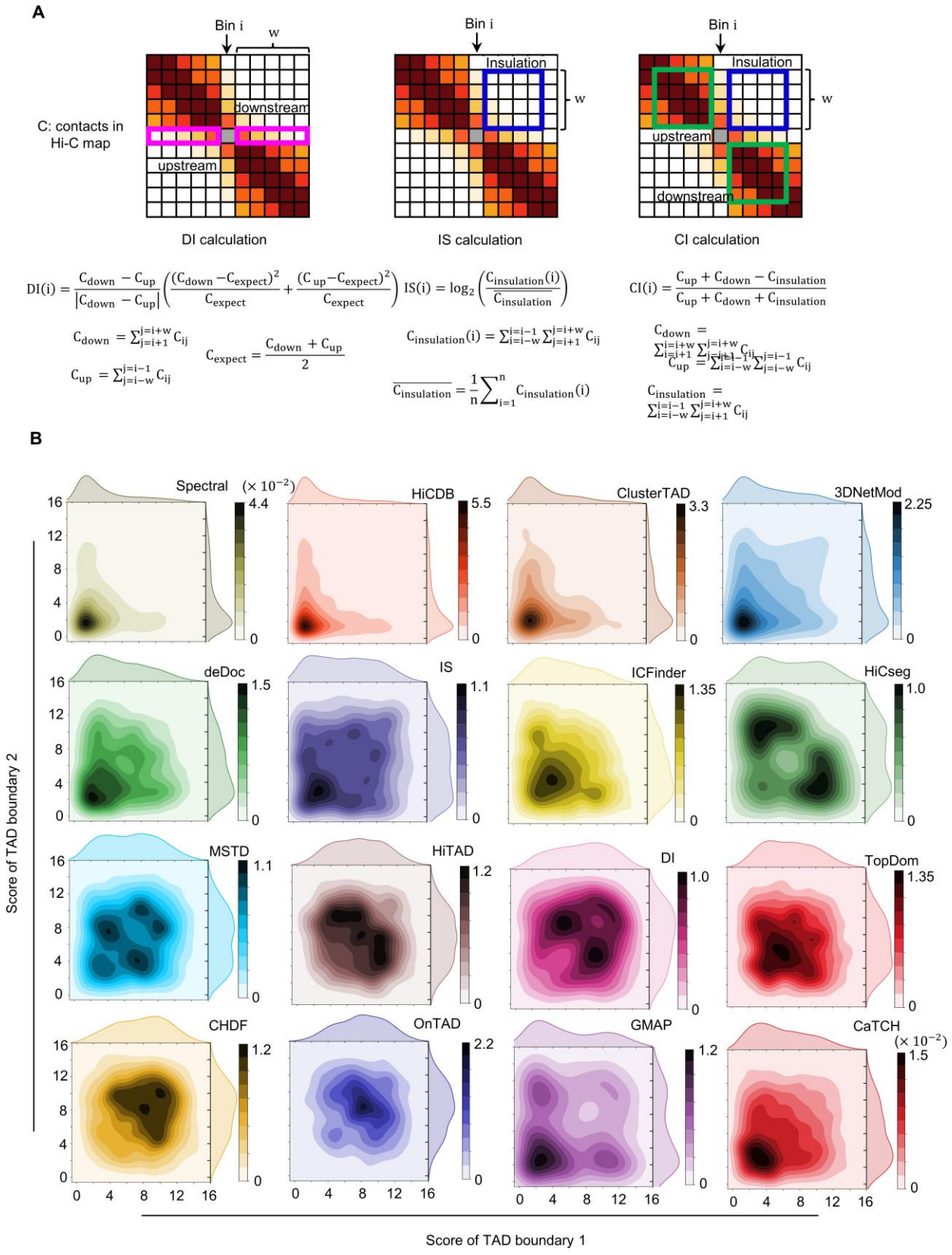
(A) Illustration of extracting lowest level domains for hierarchical TADs reported by methods like OnTAD, HiTAD and 3DNetMod. (B) TADs identified by different methods (10 methods here) on the same chromatin region (Chr 2: 10.45 – 13.55 Mb) of GM12878 and K562. (C, D) Comparison of the number (C) and average size (D) of TADs identified by 16 methods on Chromosome 2 of eight Hi-C datasets. Methods are sorted by the average values among eight datasets. (E-H) Comparison of the difference in the number (E) and size (F) of TADs, as well as the Jaccard index between boundary sets (G) and the MoC between domain sets (H) across TADs identified by each method and other methods on the same dataset or across TADs identified by each method on the different datasets. Methods are sorted by the difference or consistency of TADs across methods. The error bars represent the 95% confidence intervals.



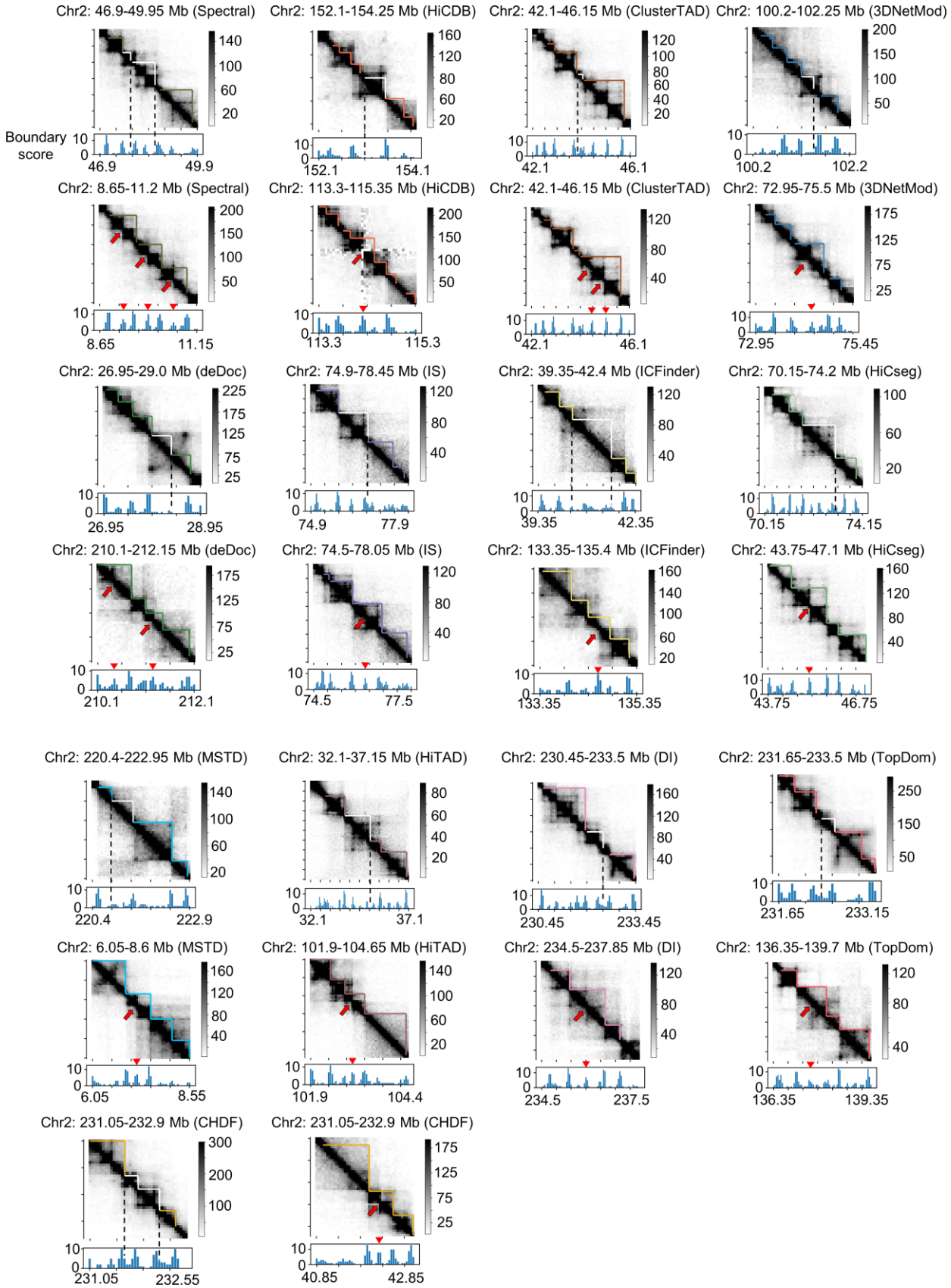
Supplemental Fig. S2. Enrichment of TFs in TAD boundaries and illustration of the boundary voting strategy.

(A) Average profiles for fold changes of CTCF, RAD21, and SMC3 binding peaks around boundaries identified by DI and deDoc in GM12878 and K562. The fold changes are calculated for each 10-kb bin by dividing the average peak numbers across the chromosome. (B) Comparison of fold changes of CTCF,

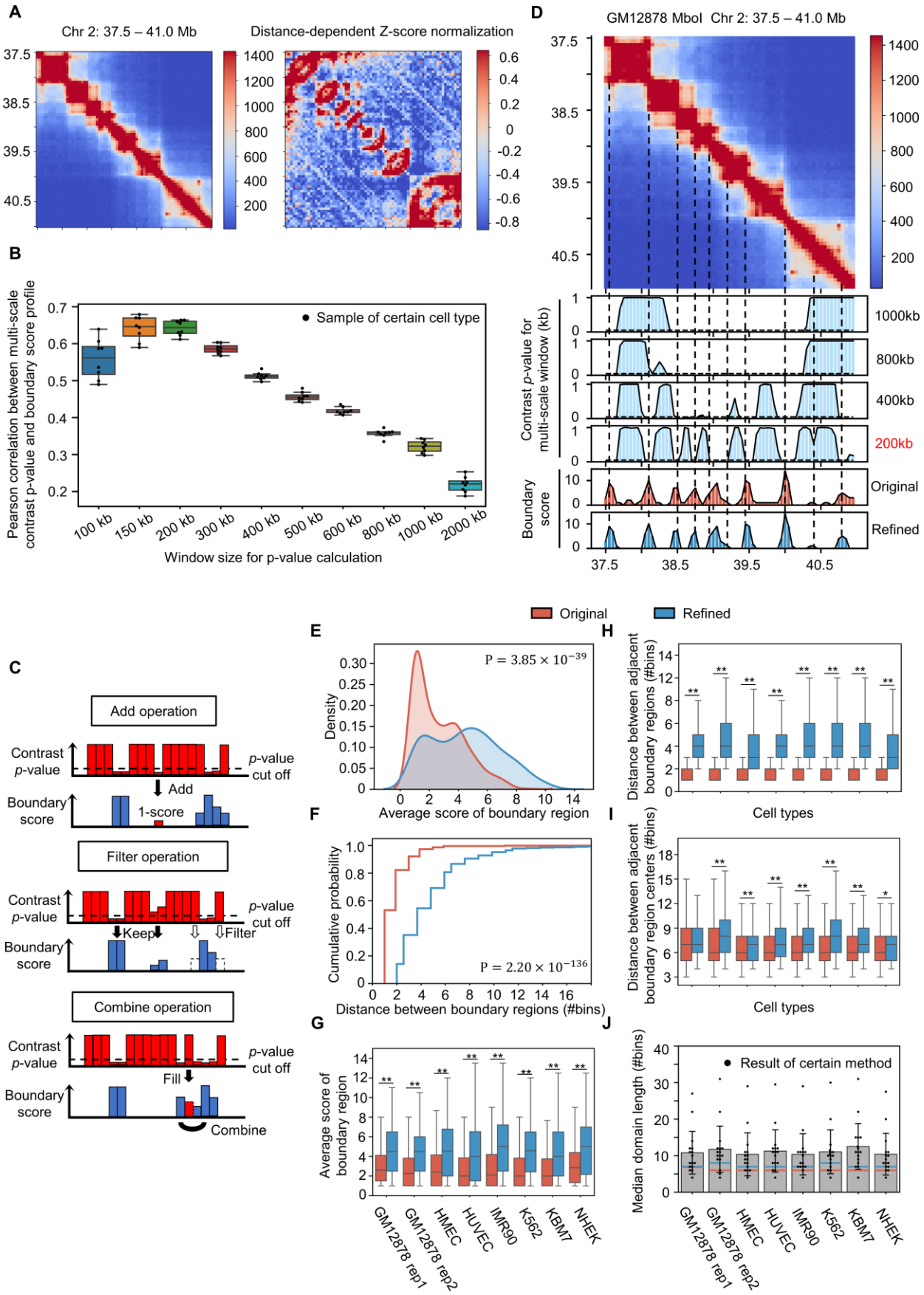
RAD21, and SMC3 binding peaks for boundaries identified by the 16 methods between GM12878 and K562. The fold change for each method is calculated by dividing the signal of the boundary by the average signal of flanking regions in the profiles shown in (A). The distribution of fold changes for different methods in GM12878 or K562 are shown and their Pearson correlation coefficients (PCC) are calculated. (C) Illustration of the boundary voting strategy. Each boundary can contribute one score to the surrounding bins according to a radius (e.g., a radius of one here). (D) Proportion of 50-kb bins with different scores on Chromosome 2 of GM12878. (E) The minimum distance between bins with different scores and the boundaries identified by each method. Two thresholds including five for the score and five bins for the minimum distance are indicated by vertical and horizontal dotted lines respectively. Dots in the upper right areas represent high-scoring bins away from the boundaries identified by each method and the numbers of such dots are shown. A horizontal tiny perturbation is added to each dot to avoid overlap.



Supplemental Fig. S3. Illustration of 1D indicators and boundary score distribution for TADs.
(A) Illustration of how to calculate three kinds of 1D topological indicators including DI, IS, and CI. (B) Density profiles of two boundary scores for TADs identified by 16 methods.



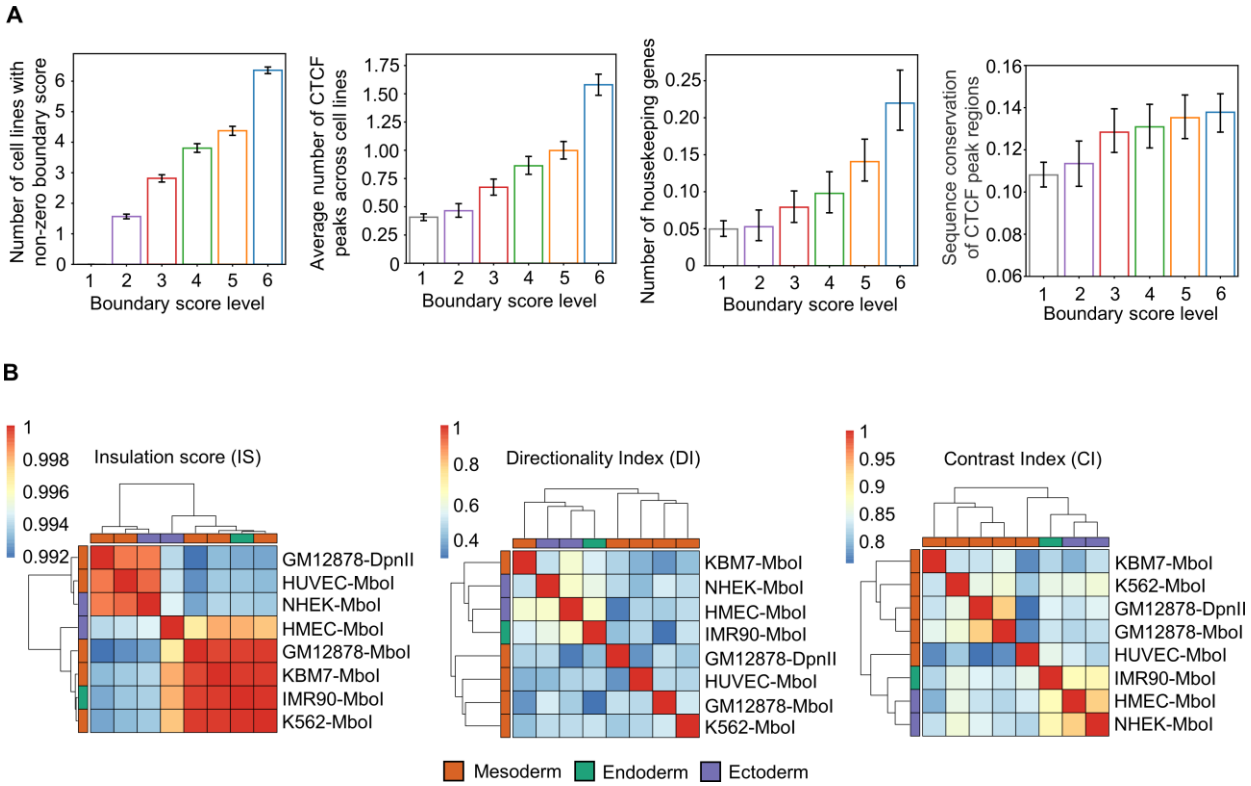
Supplemental Fig. S4. Hi-C contact maps around unreliable TADs and missed TADs for different methods (see Figs. 2D, 2E).



Supplemental Fig. S5. Construction of the TAD separation landscape.

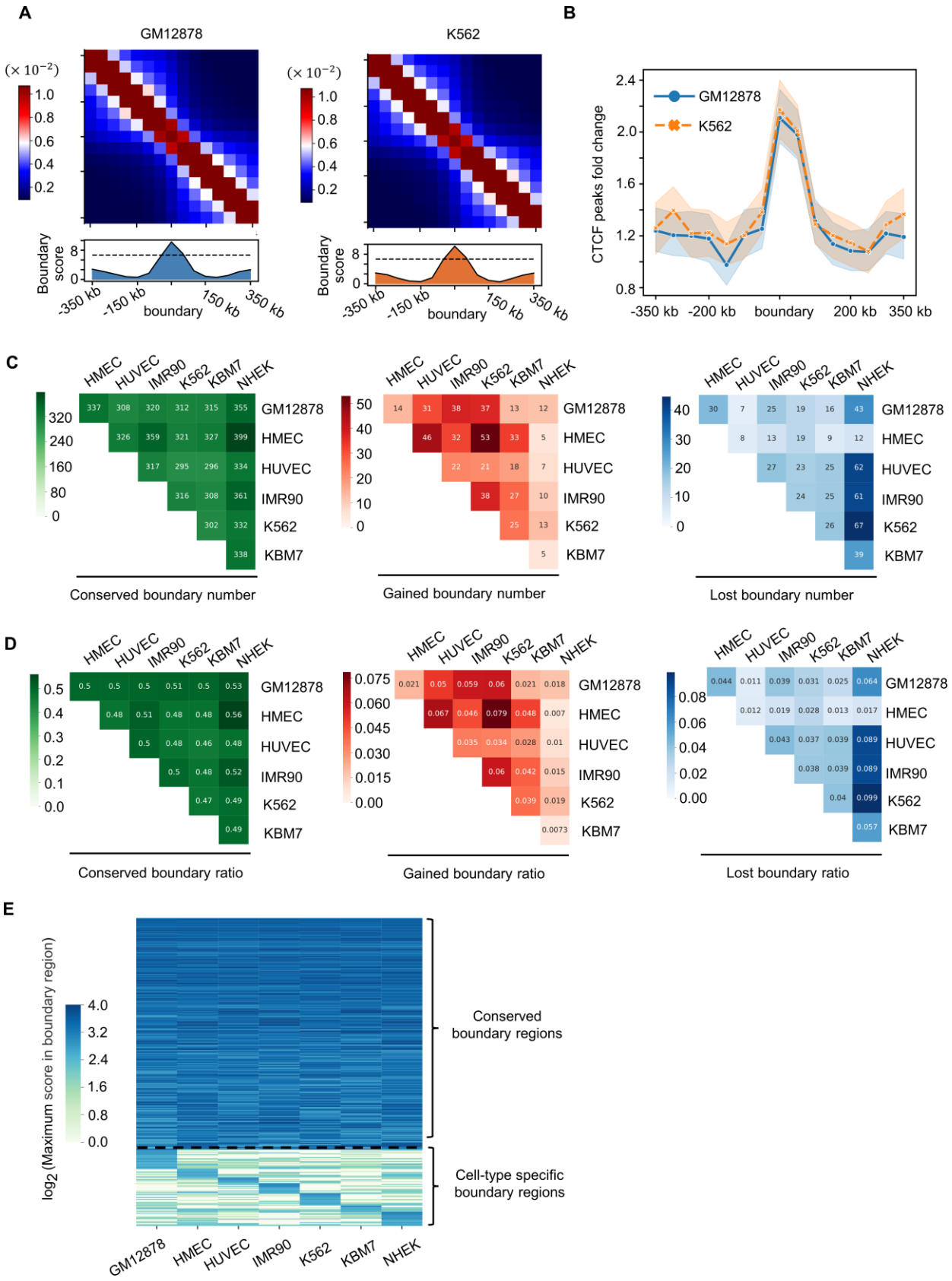
(A) An example of the Hi-C contact map before (left) and after (right) the distance-dependent z-score normalization. (B) Pearson correlation between the boundary score profile and the contrast *p*-value profiles calculated using multi-scale window size. Each dot represents the result for a sample of a certain

cell type. At each scale, eight samples are involved, including seven cell types GM12878 (with two replicates), HVEC, HUVEC, IMR90, K562, KBM7, and NHEK. For each sample, the scale with the highest correlation is chosen for the construction of the TAD separation landscape. (C) Three operations (Add, Filter, Combine) are used for constructing the TAD separation landscape. The Add operation adds one score to bins with zero boundary scores but with *p*-values below the cut-off. The Filter operation turns the boundary scores to zero for bins with *p*-values greater than the cut-off, but the boundary scores for bins in the valleys of the *p*-value profiles are kept. The Combine operation combines two adjacent boundary regions separated by one bin gap and the gap will be filled with the average boundary score of the upper and lower bins. The *p*-value cut-off is set as 0.05 in this study. (D) The Hi-C contact map of a region on Chromosome 2 of GM12878 accompanied by the corresponding multi-scale contrast *p*-value profiles. The original boundary score profile as well as the refined profile are also shown. The selected scale of the contrast *p*-value for constructing the TAD separation landscape is marked red. For each scale, the *p*-value cut-off is set as 0.05 (denoted by the horizontal dashed lines). The vertical dashed lines indicate the positions of refined boundary regions. (E, F) Comparison of the average score distribution for the boundary regions (E) and the cumulative distributions of the distance between the adjacent boundary regions (F) before and after refining for the GM12878 MboI sample. The Kolmogorov-Smirnov test is used to get the *p*-value. (G-I) Comparison of the average score for the boundary regions (G), the distance between the adjacent boundary regions (H), and the distance between centers of the adjacent boundary regions before and after refining for eight samples from the different cell types. The Mann-Whitney *U* tests are performed, * represents *p*-value < 3×10^{-4} and ** represents *p*-value < 10^{-6}. (J) The median length of TAD domains for the 16 TAD-calling methods on eight samples from different cell types. Each black dot represents the result of a certain method. The median distances between centers of the adjacent boundary regions before and after refining are shown as red and blue lines on each bar.



Supplemental Fig. S6. Analysis of bins with different boundary scores and the clustering of Hi-C sample.

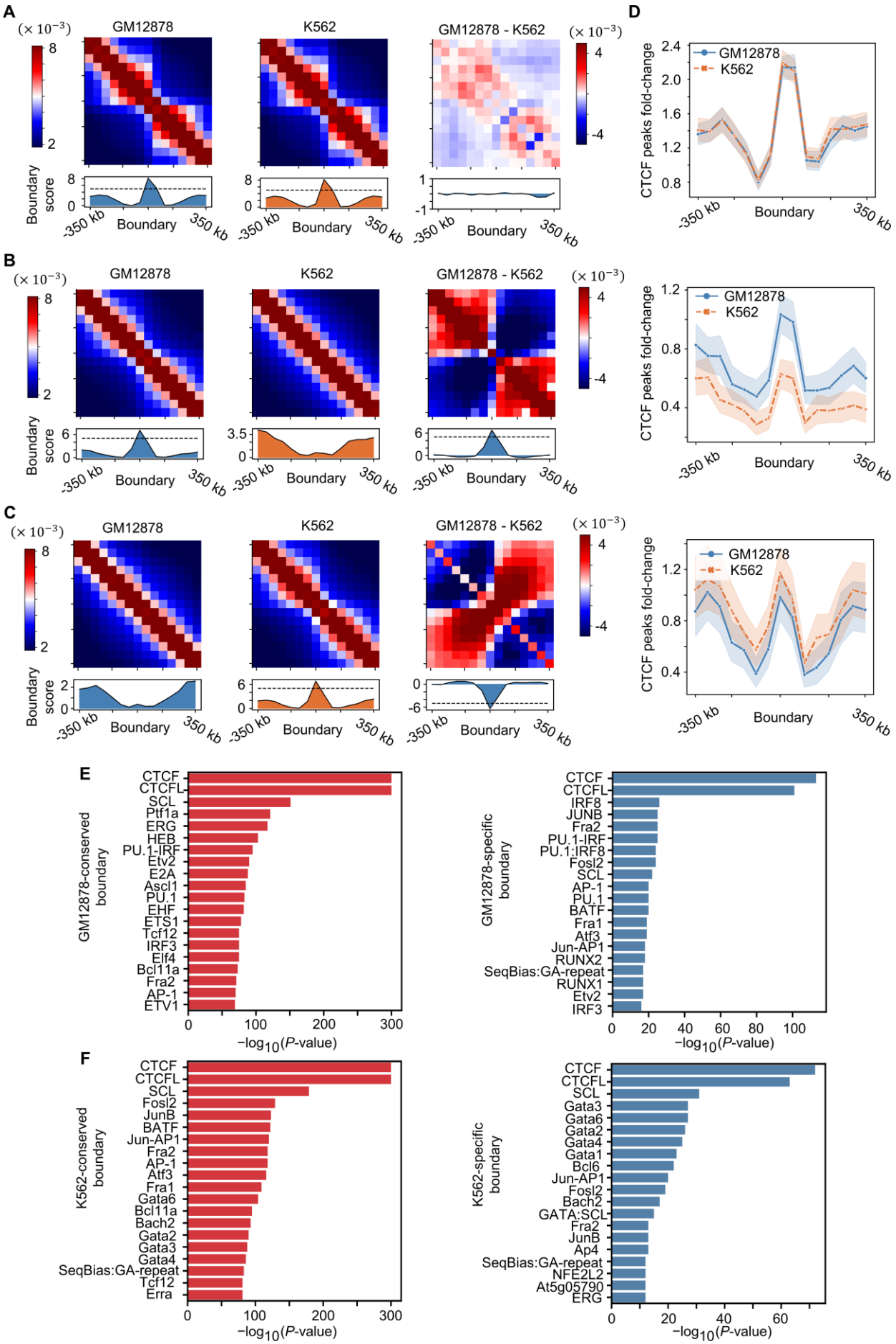
(A) The relationship between the boundary score level and the conservation of boundary across cell lines, the average number of CTCF peaks across cell lines, the number of housekeeping genes, and the conservation of DNA sequence for CTCF binding regions. (B) Clustering of multiple Hi-C samples from different cell lines based on the Pearson correlation of 1D indicators including IS, DI, and CI.



Supplemental Fig. S7. Comparison of boundary regions.

(A) The aggregated Hi-C contact maps around the conserved boundary regions between GM12878 and K562, combined with the average boundary score profiles of the corresponding regions. The dotted lines indicate a boundary score of five. (B) Profiles of CTCF peak fold change (bin vs background) around the conserved boundary regions between GM12878 and K562. The shaded areas represent the 95%

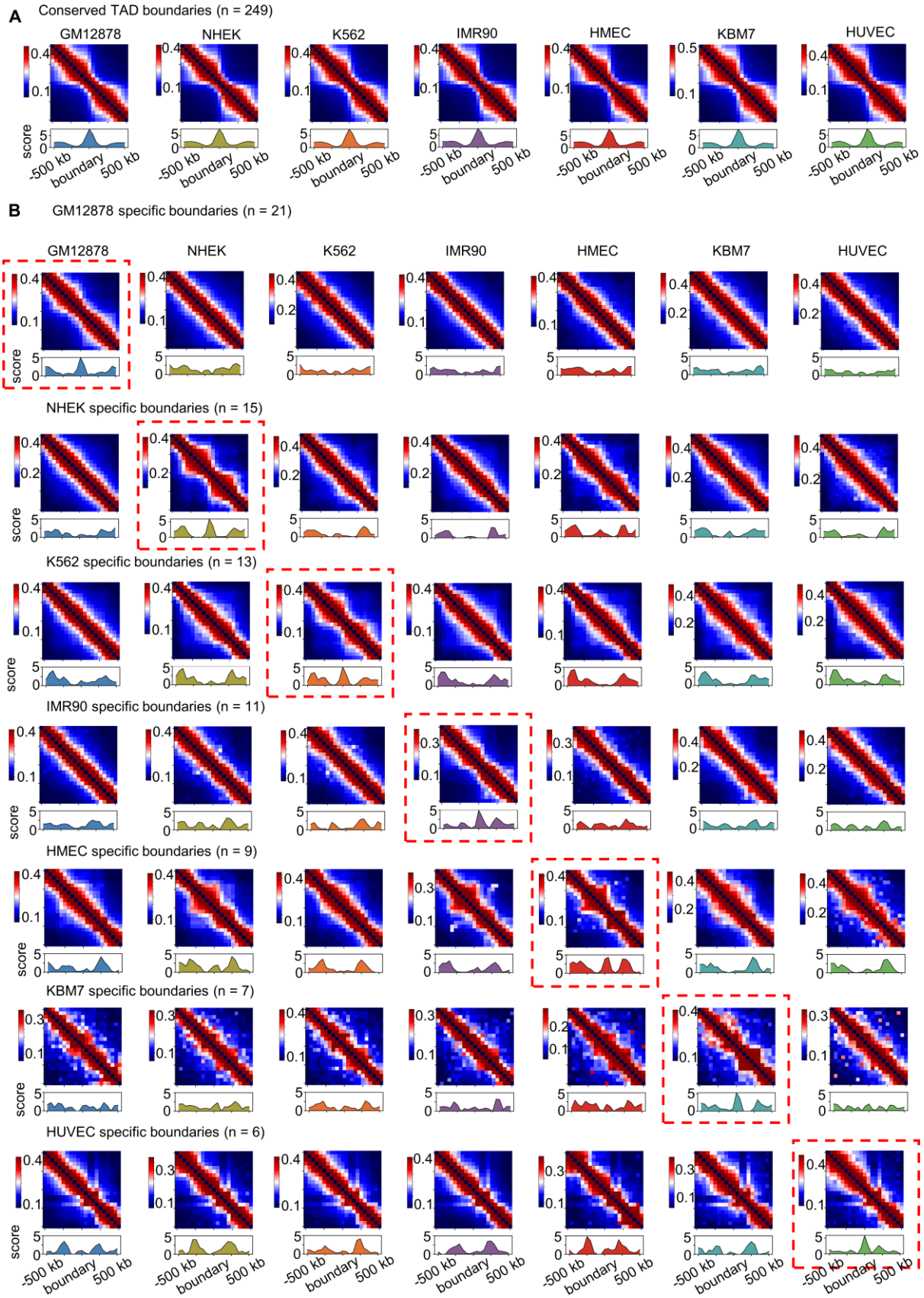
confidence intervals in 1000 bootstraps. (C, D) The number (C) and ratio (D) of conserved, cell-type gained and cell-type lost boundary regions for pairwise comparisons between seven cell lines. The ratios are calculated based on the unions of the boundary regions between two cell types. The gained and lost boundary regions are defined in terms of the cell lines in each row. (E) Heatmap of the maximum boundary scores in the conserved and cell-type specific boundary regions among seven cell lines.



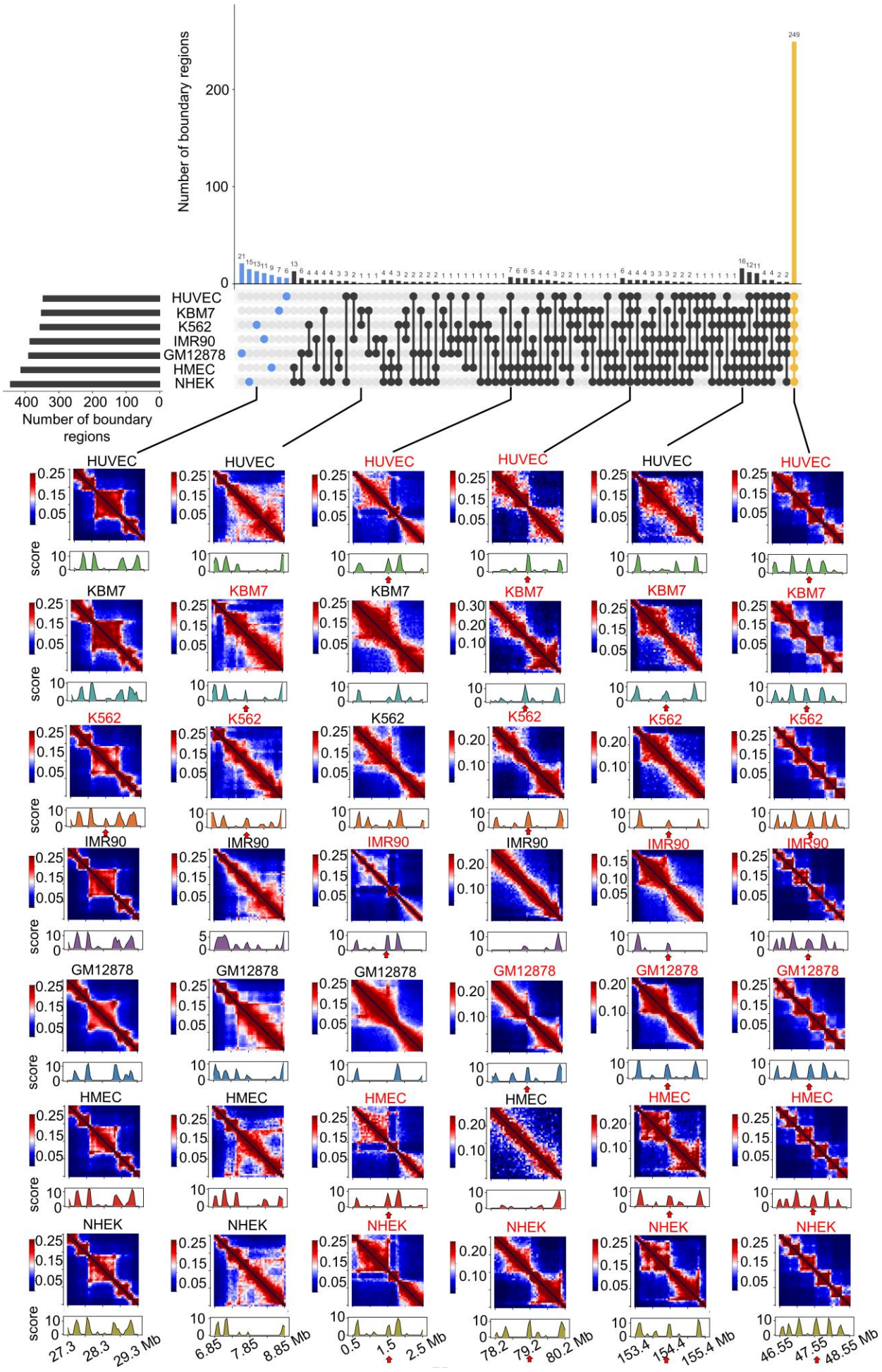
Supplemental Fig. S8. Genome-wide comparison of the boundary region between GM12878 and

K562.

(A-C) The aggregated Hi-C contact maps and boundary score profiles around the conserved (A), GM12878-specific (B), and K562-specific boundaries (C) between GM12878 and K562, as well as the difference between these aggregated maps. (D) The CTCF peaks profiles around the conserved, GM12878-specific, K562-specific boundaries. (E) The top 20 TFs enriched in conserved and GM12878-specific boundaries. (F) The top 20 TFs enriched in conserved and K562-specific boundaries.

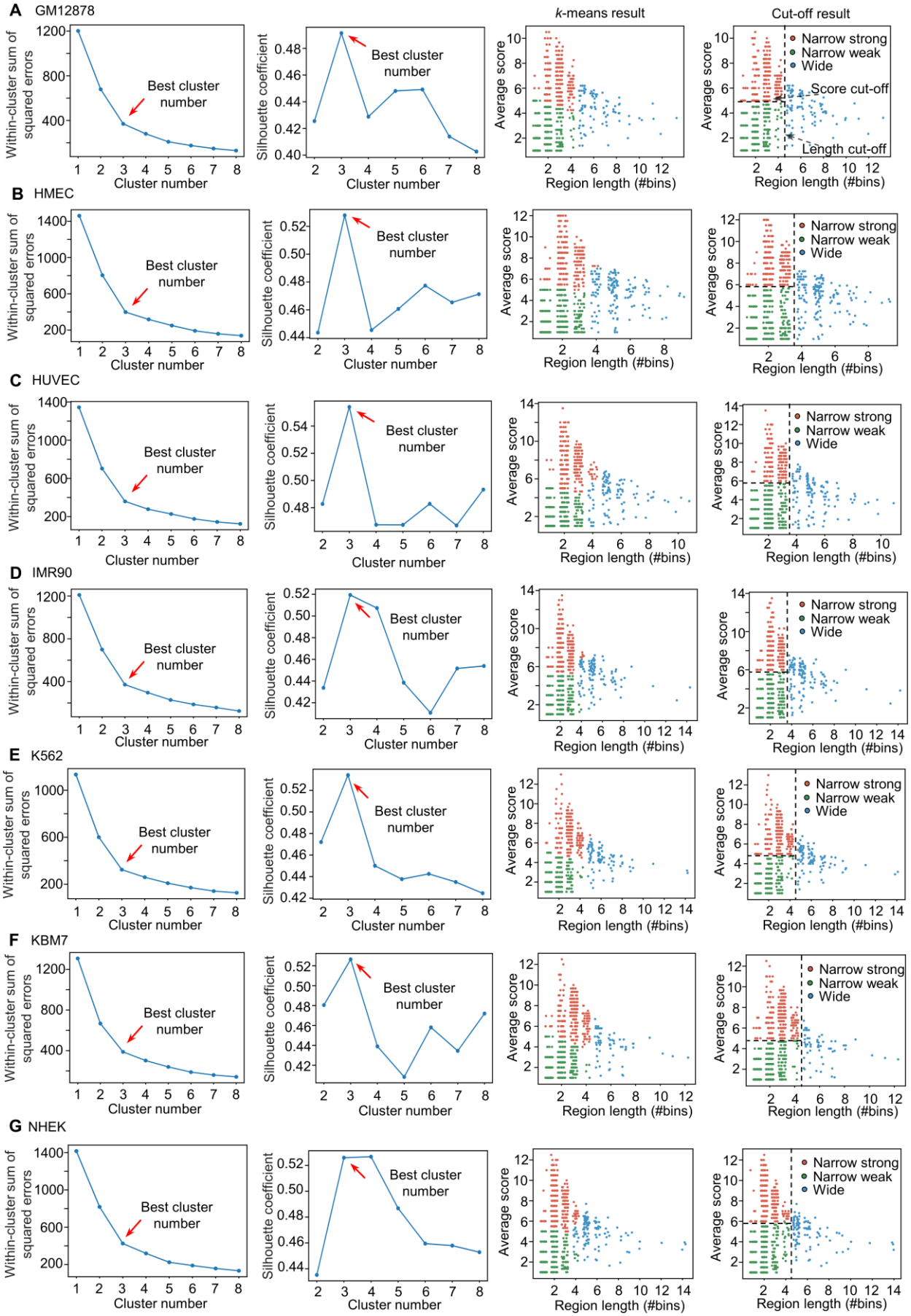


Supplemental Fig. S9. Illustration of the conserved and cell-type specific boundary regions.
(A, B) The aggregated Hi-C contact maps around the conserved boundary regions (A) and each kind of cell-type specific boundary regions (B) combined with the average boundary score profiles. The numbers of the boundary regions are shown above each and the corresponding cell type is marked with a red frame.


Number of boundary regions
HUVEC
KBM7
K562
IMR90
GM12878
HMEC
NHEK
Number of boundary regions
score
27.3
28.3
29.3 Mb
6.85
7.85
8.85 Mb
0.5
1.5
2.5 Mb
78.2
79.2
80.2 Mb
153.4
154.4
155.4 Mb
46.55
47.55
48.55 Mb

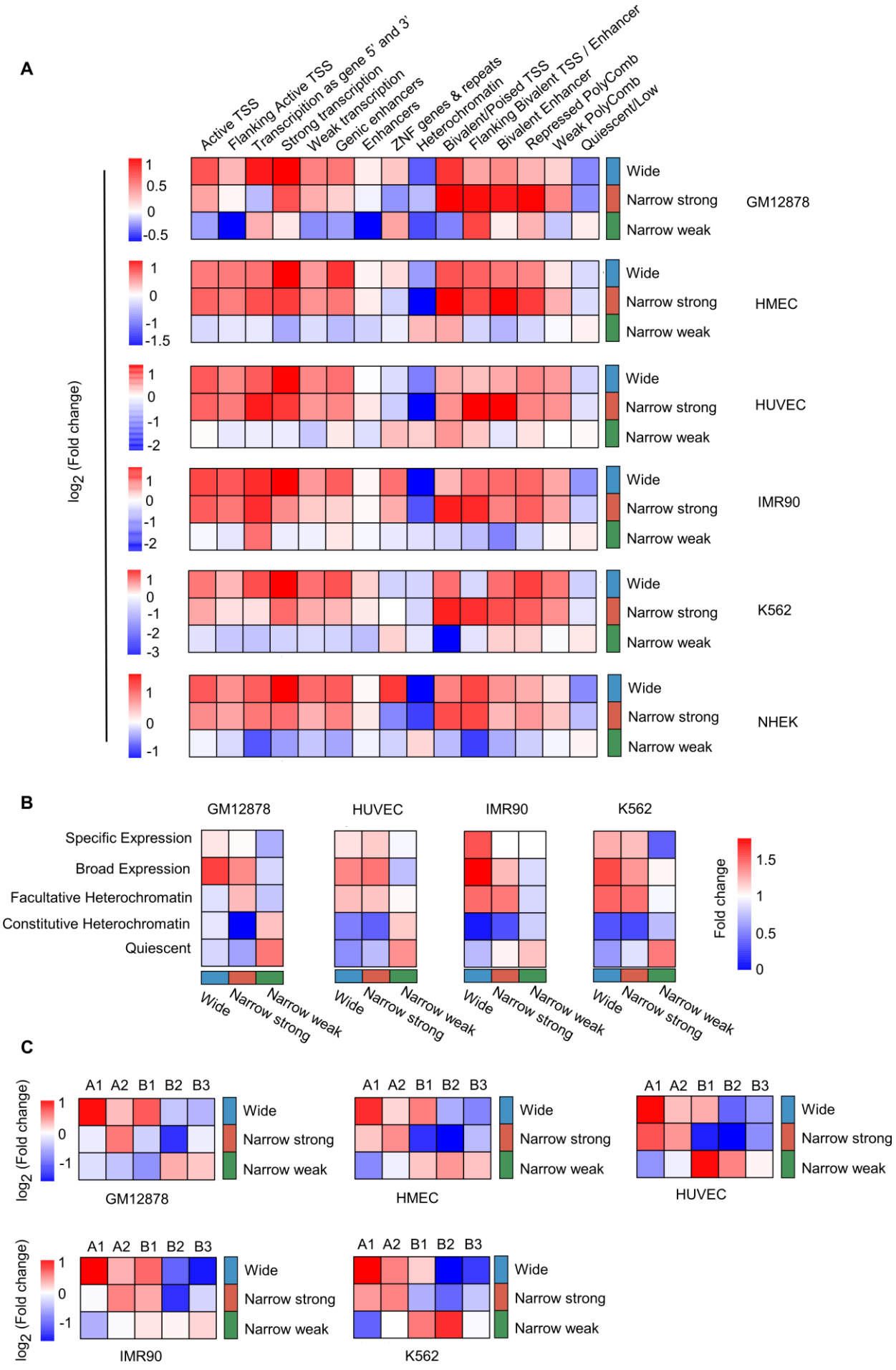
Supplemental Fig. S10. Boundary regions shared by a part of the seven cell lines.

Numbers of the cell-type specific boundaries or boundaries shared by part of the seven cell lines are shown above. The examples of Hi-C contact maps centered on boundary regions belonging to certain types are shown below. These boundary regions are marked by red arrows and cell lines sharing these boundary regions are marked in red.



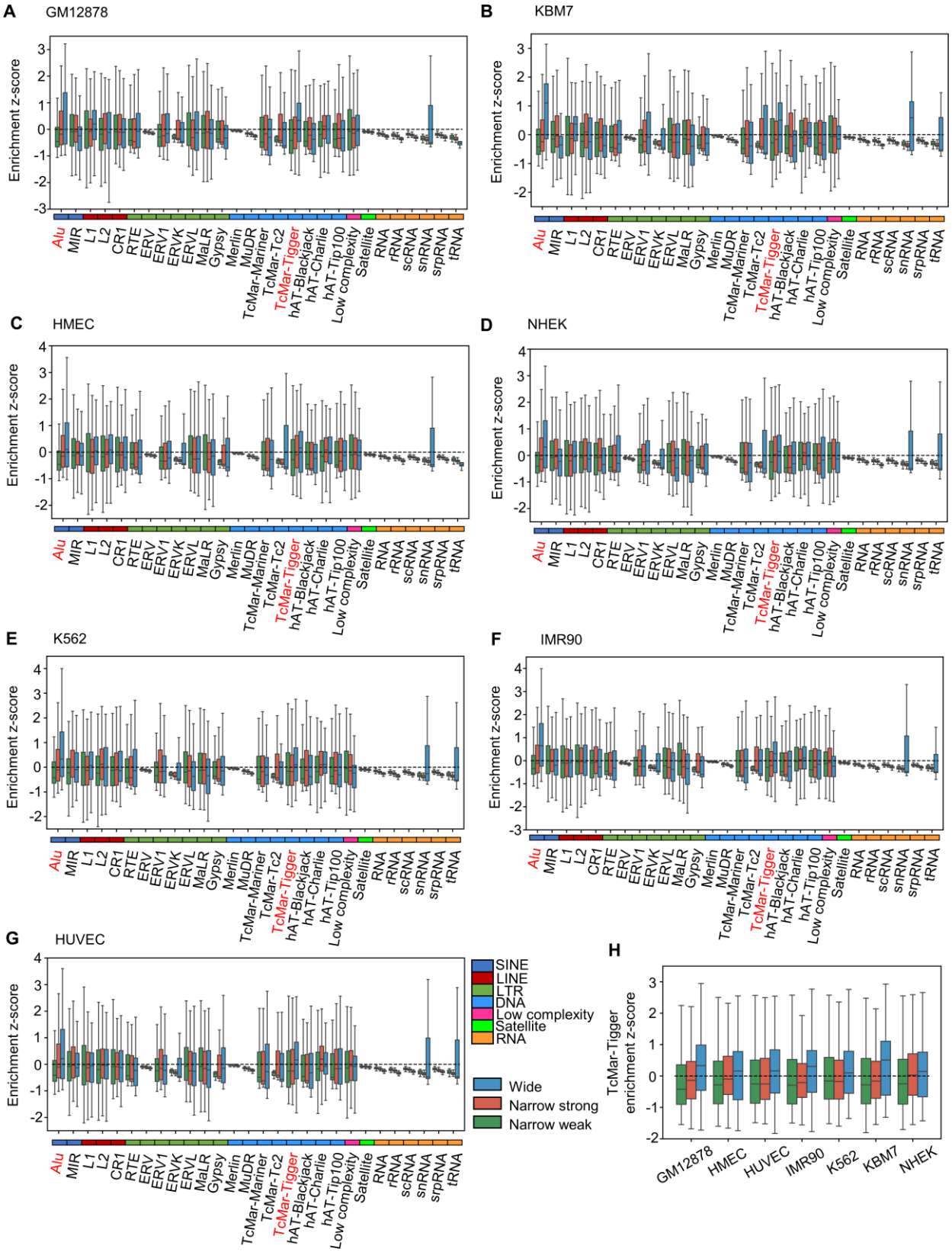
Supplemental Fig. S11. Identification of three types of boundary regions based on the TAD separation landscapes in seven cell lines.

(A-G) For each cell line, four images are shown from left to right. Two images on the left depict the relationships between the within-cluster sum of squared errors or the silhouette coefficient and the number of clusters for *k*-means, respectively. The right two show the clustering results of *k*-means with the optimal number of clusters and the clustering results according to the selected thresholds for the length and the average score of boundary region respectively. The best number of clusters is marked by red arrows and the thresholds for region length and average score are indicated by dashed lines.


A
Active TSS
Flanking Active TSS
Transcription as gene 5' and 3'
Strong transcription
Weak transcription
Genic enhancers
Enhancers
ZNF genes & repeats
Heterochromatin
Bivalent/Poised TSS
Flanking Bivalent TSS / Enhancer
Bivalent Enhancer
Repressed PolyComb
Weak PolyComb
Quiescent/Low
log2 (Fold change)
Wide
Narrow strong
Narrow weak
GM12878
HMEC
HUVEC
IMR90
K562
NHEK
B
GM12878
HUVEC
IMR90
K562
Specific Expression
Broad Expression
Facultative Heterochromatin
Constitutive Heterochromatin
Quiescent
Fold change
C
A1 A2 B1 B2 B3
GM12878
HMEC
HUVEC
IMR90
K562

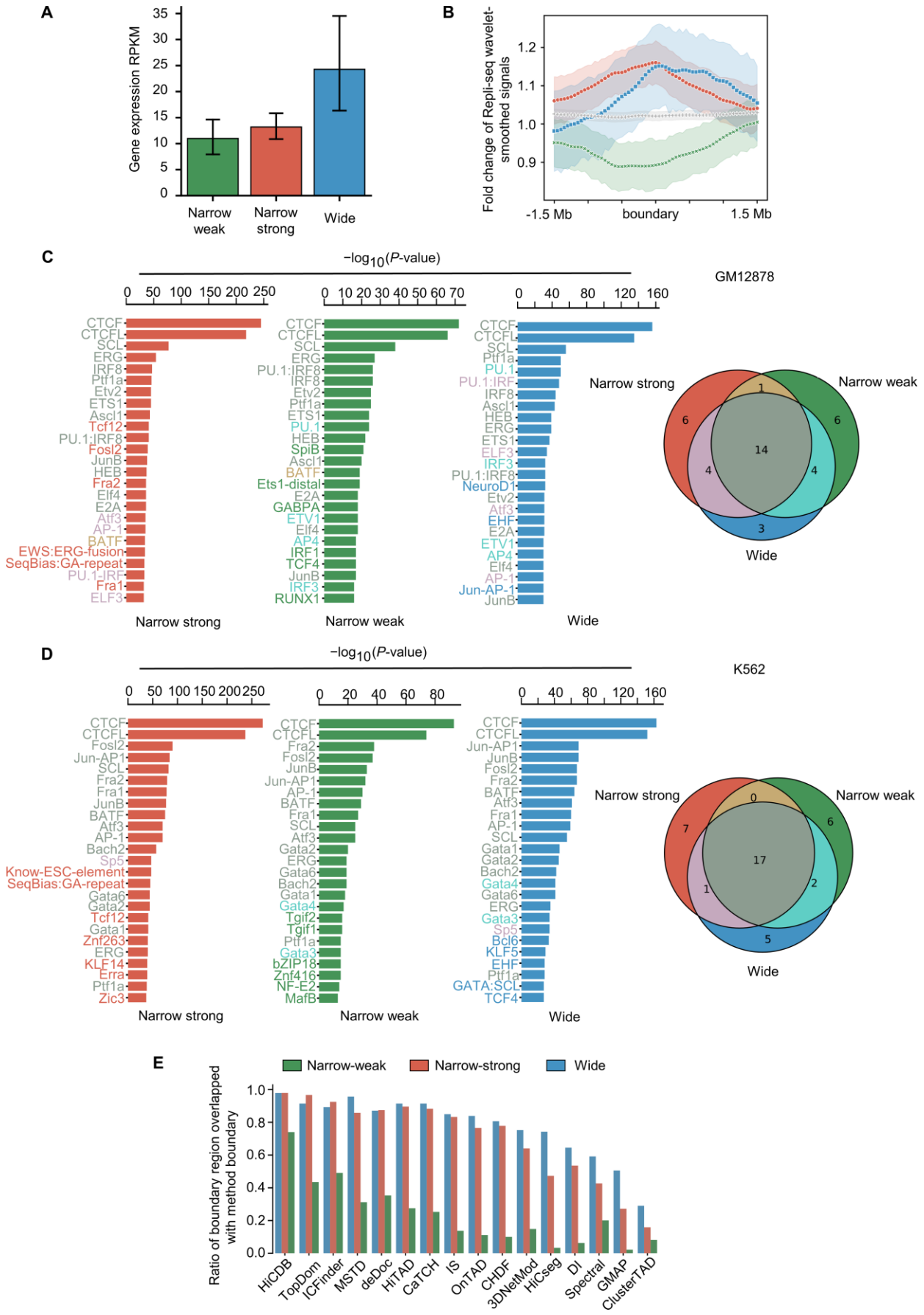
Supplemental Fig. S12. Enrichment analysis of chromatin states and subcompartments for three types of boundary regions.

(A-C) Fold change profiles of ChromHMM states (A), Segway states (B), and five kinds of subcompartments (C) calculated for the three types of boundary regions in multiple cell lines. Fold change is defined as the total length of the state or subcompartment in boundary regions divided by the expected length of the state across the whole chromosome.



Supplemental Fig. S13. Enrichment analysis of repeat elements in three types of boundary regions.

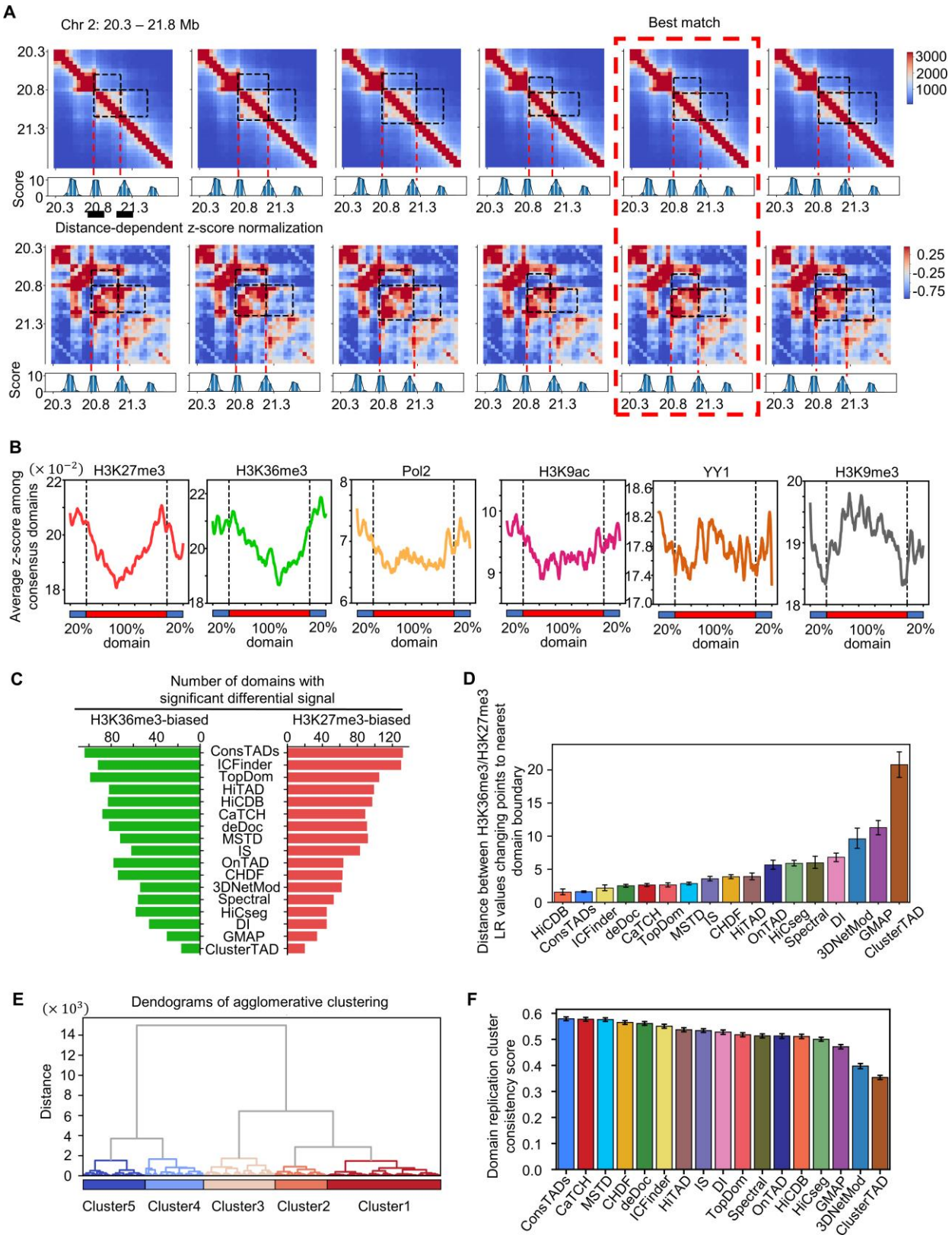
(A-G) Enrichment z-score of repeat elements in three types of boundary regions identified from seven cell lines. These repeat elements can be divided into seven types, labeled with different colors. (H) The enrichment of TcMar-Tigger elements is shown separately. The dashed lines indicate the z-score of zero.



Supplemental Fig. S14. Biological characterization of the three types of boundary regions.

(A) FPKM of genes in the three types of boundary regions in GM12878. (B) Fold change profiles of Repli-seq wavelet-smoothed signal around the three types of boundaries in GM12878. (C, D) Top 25 TFs from

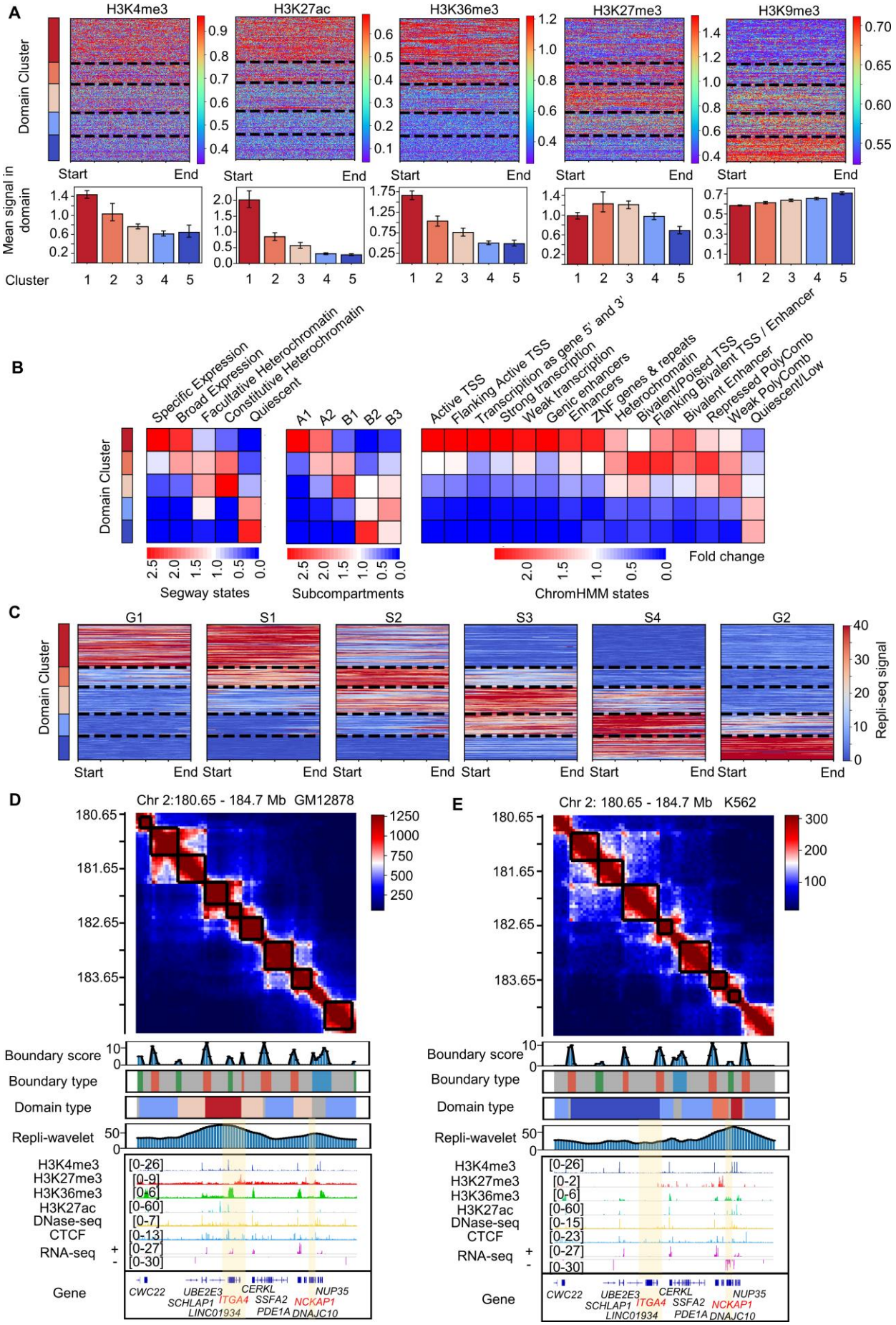
HOMER TF datasets ranked based on their motif enrichment in the open chromatin of the three types of boundary regions for GM12878 (C) and K562 (D), respectively. The significance of TF enrichment is calculated with the hypergeometric test by HOMER. The intersections between TFs enriched in the three types of boundary regions are also shown and the names of the TFs are colored according to the intersections. (E) The ratio of three types of boundary regions that overlapped with boundaries reported by 16 TAD-calling methods.



Supplemental Fig. S15. Illustration of boundary matching and additional analysis of biological features within domains.

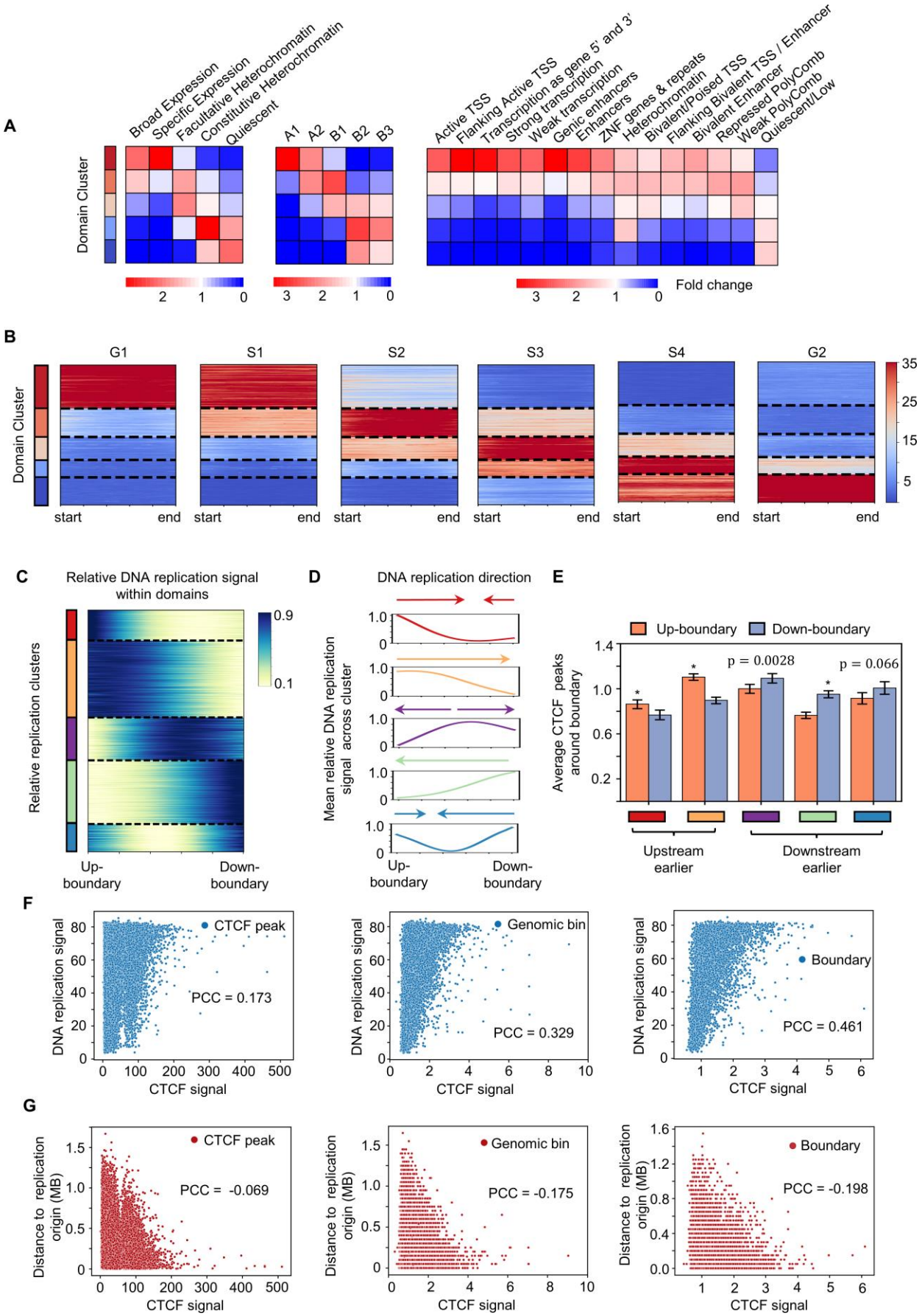
(A) Schematic diagram of the boundary matching process to define ConsTADs. For the two selected adjacent boundary regions (marked by black bars), one bin is selected from each of them, all six combinations are shown and the selected bins are indicated by the red dashed line. For each combination, the corresponding domain region, upstream and downstream region is indicated by the black dashed lines. The combination of bins with the best average rank in terms of the boundary score

and domain signal enrichment would be selected to form the ConsTADs. (B) Profiles of the remaining six kinds of biological features within ConsTADs and adjacent regions (see Fig. 6A). (C) Number of domains with a significant differential signal of H3K36me3 or H3K27me3 for ConsTADs and the 16 TAD-calling methods. (D) Distribution of the distance between H3K36me3/H3K27me3 LR values changing points to the nearest domain boundary from the 16 TAD-calling methods and ConsTADs. (E) Dendogram of agglomerative clustering for ConsTADs based on the DNA replication signal and all these domains are divided into five clusters (see Fig. 6C). (F) Consistency scores of domain replication clusters for all bins along Chromosome 2 in GM12878 under each TAD-calling method.



Supplemental Fig. S16. Biological properties of the five kinds of replication domain clusters.

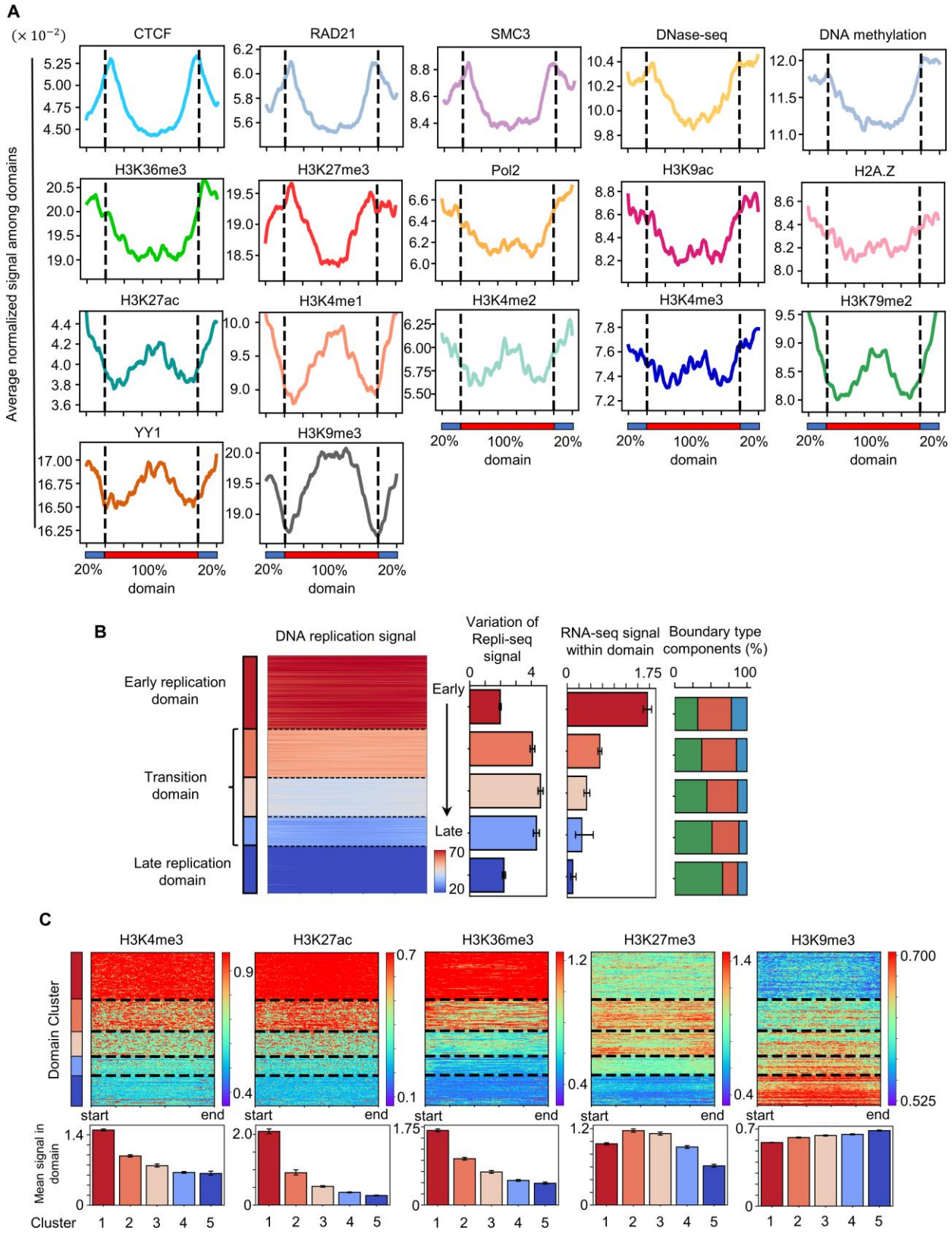
(A) Profiles of several biological features within the five domain clusters accompanied by the mean signal in each domain. (B) Fold change profiles of Segway states (left), five kinds of subcompartments (middle), and ChromHMM states (right) calculated for the five domain clusters over the background. (C) Profiles of Repli-seq signal for the six phases of the cell cycle within the five domain clusters. (D and E) The Hi-C contact maps and the ConsTADs for a region on Chromosome 2 in GM12878 (C) and K562 (D). The TAD separation landscape, domain type annotation, as well as the profiles of some biological features are also shown below. Regions corresponding to two genes *ITGA4* and *NCKAP1* are marked with yellow shades.



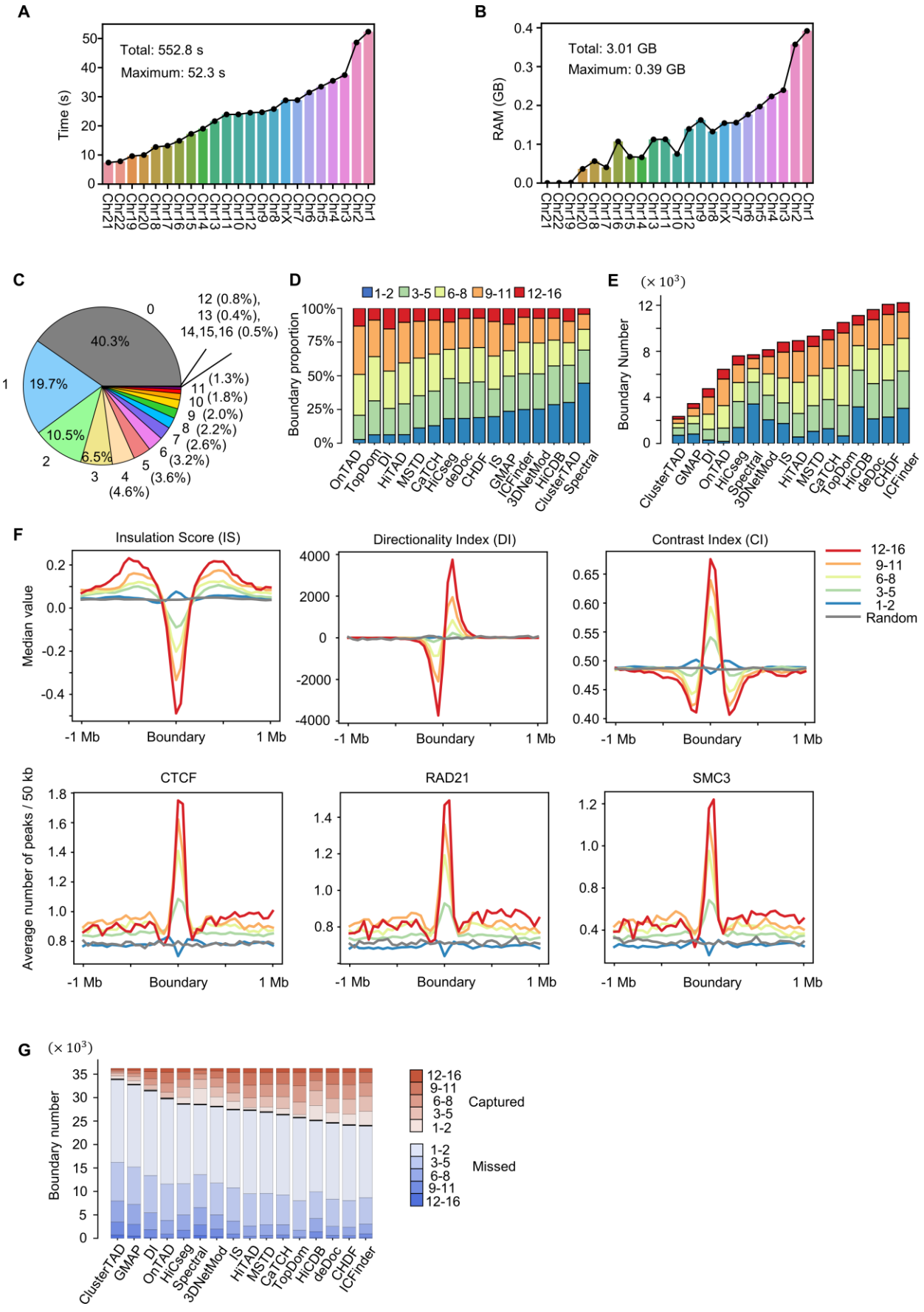
Supplemental Fig. S17. Genome-wide analysis of five types of replication domains or relative replication domains.

(A) Fold change profiles of Segway states (left), five kinds of subcompartments (middle), and

ChromHMM states (right) calculated for the five domain clusters over the background. (B) Profiles of the Repli-seq signal for the six phases of the cell cycle within the five domain clusters. (C) Five clusters of ConsTADs with distinct relative DNA replication models, indicating the relative early or late replication times of chromatin regions within each domain. (D) Mean profiles of relative DNA replication signals for five clusters of ConsTADs in (C). Arrows indicate the direction of DNA replication from early to late. (E) Average number of CTCF binding peaks around the upstream and downstream boundary for each domain in five clusters. Mann-Whitney *U* tests were performed to get the *p*-values, * represents *p*-value < 0.0001. (F) The relationship between the replication timing signal and the CTCF signal for CTCF binding peaks, all genomic bins, and the ConsTADs boundaries. (G) The relationship between the CTCF signal and the distance to the origin of replication for CTCF binding peaks, all genomic bins, and the ConsTADs boundaries. PCC means Pearson correlation coefficient. The origins of replication are defined as the peaks in the Repli-seq profile.



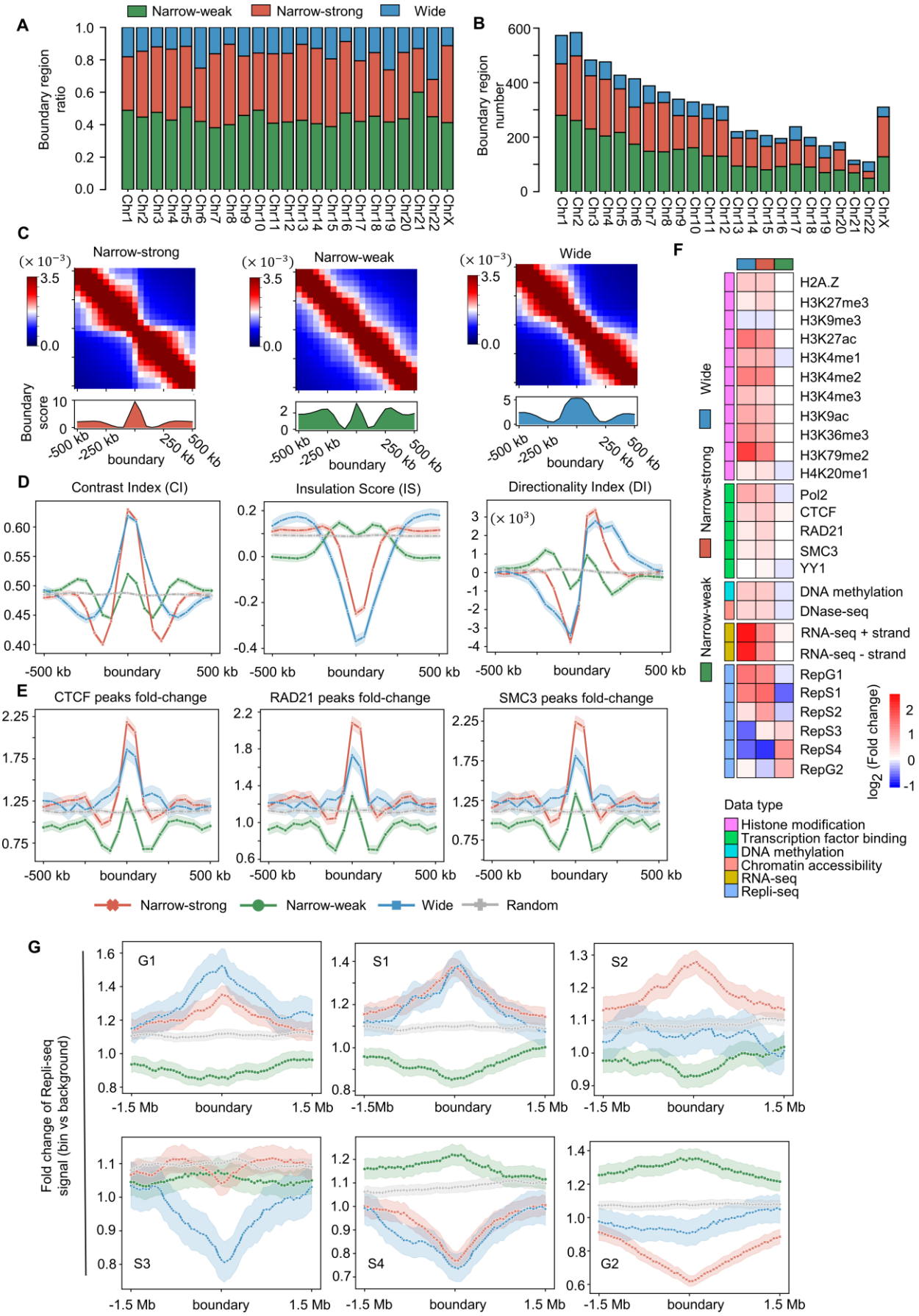
Supplemental Fig. S18. Genome-wide analysis of ConsTADs and five types of replication domains. (A) Relative profiles of several biological features within ConsTADs and adjacent regions in GM12878. (B) Five types of ConsTADs with distinct DNA replication signals accompanied by the variances of Repli-seq signal, the average RNA-seq signal, and the boundary type components within each type of domains. (C) Profiles of several biological features within the five domain clusters accompanied by the mean signal in each domain.



Supplemental Fig. S19. Evaluation of 16 TAD-calling methods based on the boundary voting strategy for genome-wide results and the Computation time and RAM used by ConsTADs.

(A and B) Computation time and RAM used by ConsTADs for all chromosomes of GM12878. The Hi-C

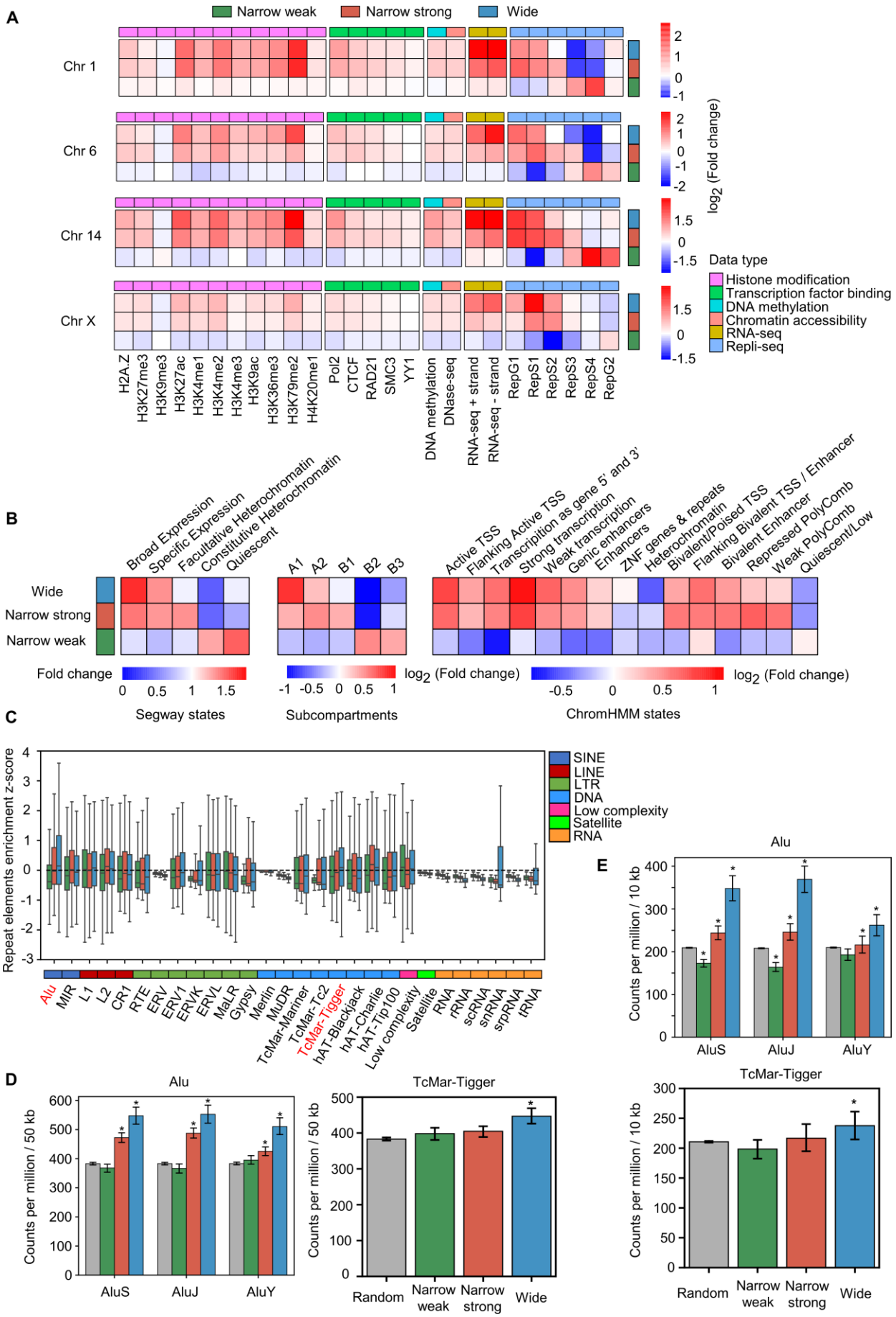
contact maps are generated at 50 kb resolution. (C) Proportion of genome-wide bins with different boundary scores for GM12878. (D) Proportion of boundaries with different levels of boundary score for 16 methods. Methods are sorted in ascending order by the proportion of boundaries belonging to the first level (ranging from 1 to 2). (E) Number of boundaries with different levels of boundary score for 16 methods. Methods are sorted in ascending order by the total number of boundaries. (F) Profiles of three topological indicators (Insulation Score [IS], Directionality Index [DI], Contrast Index [CI]) and profiles of three structural proteins (CTCF, RAD21, SMC3) within 2-Mb regions centered on boundaries with different boundary score levels or randomly selected regions. (G) Number of boundaries with different boundary scores captured or missed by each method. Methods are sorted in ascending order by the number of captured boundaries.



Supplemental Fig. S20. Genome-wide identification and analysis of three types of boundary regions in GM12878.

(A and B) The ratio and number of three types of boundary regions found on all chromosomes for

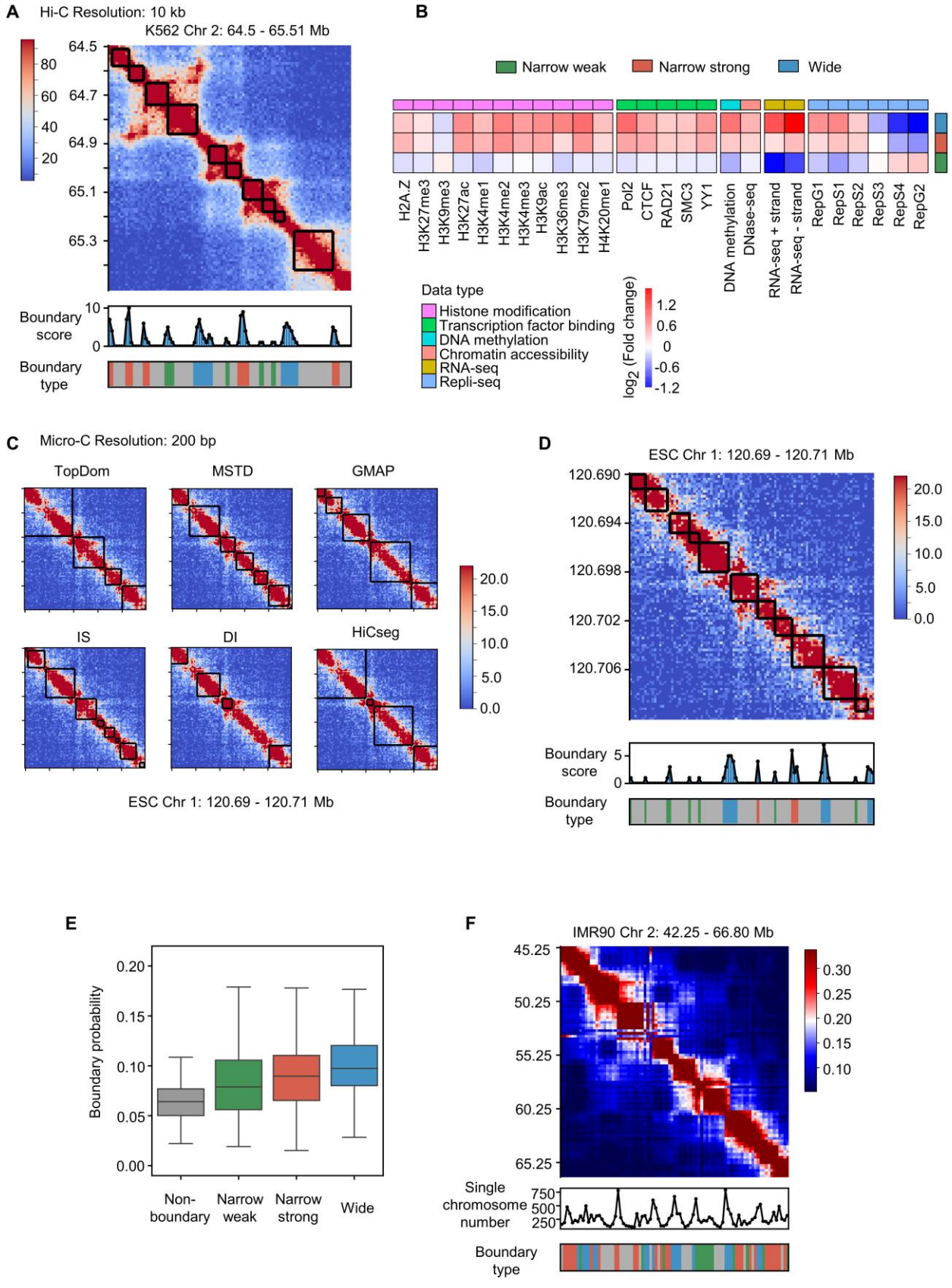
GM12878. (C) Aggregated Hi-C contact maps around NSBs (left panel), NWBs (middle panel), and WBs (right panel), combined with the average boundary score profiles in GM12878. (D) Profiles of three kinds of numerical indices including CI (left panel), IS (middle panel), and DI (right panel) around different types of boundaries and randomly selected regions in GM12878. The shaded areas represent the 95% confidence intervals in 1000 boostraps. (E) Profiles of three kinds of biological signals including CTCF (left panel), RAD21 (middle panel), and SMC3 (right panel) around different types of boundaries and randomly selected regions in GM12878. (F) Fold change profiles of multiple types of biological data constructed for three types of boundary regions in GM12878. (G) Fold change profiles of the Repli-seq signal around three types of boundaries and randomly selected regions in six phases of the cell cycle: G1, S1, S2, S3, S4, and G2.



Supplemental Fig. S21. Genome-wide analysis of three types of boundary regions in GM12878.

(A) Fold change profiles of multiple types of biological data constructed for three types of boundary

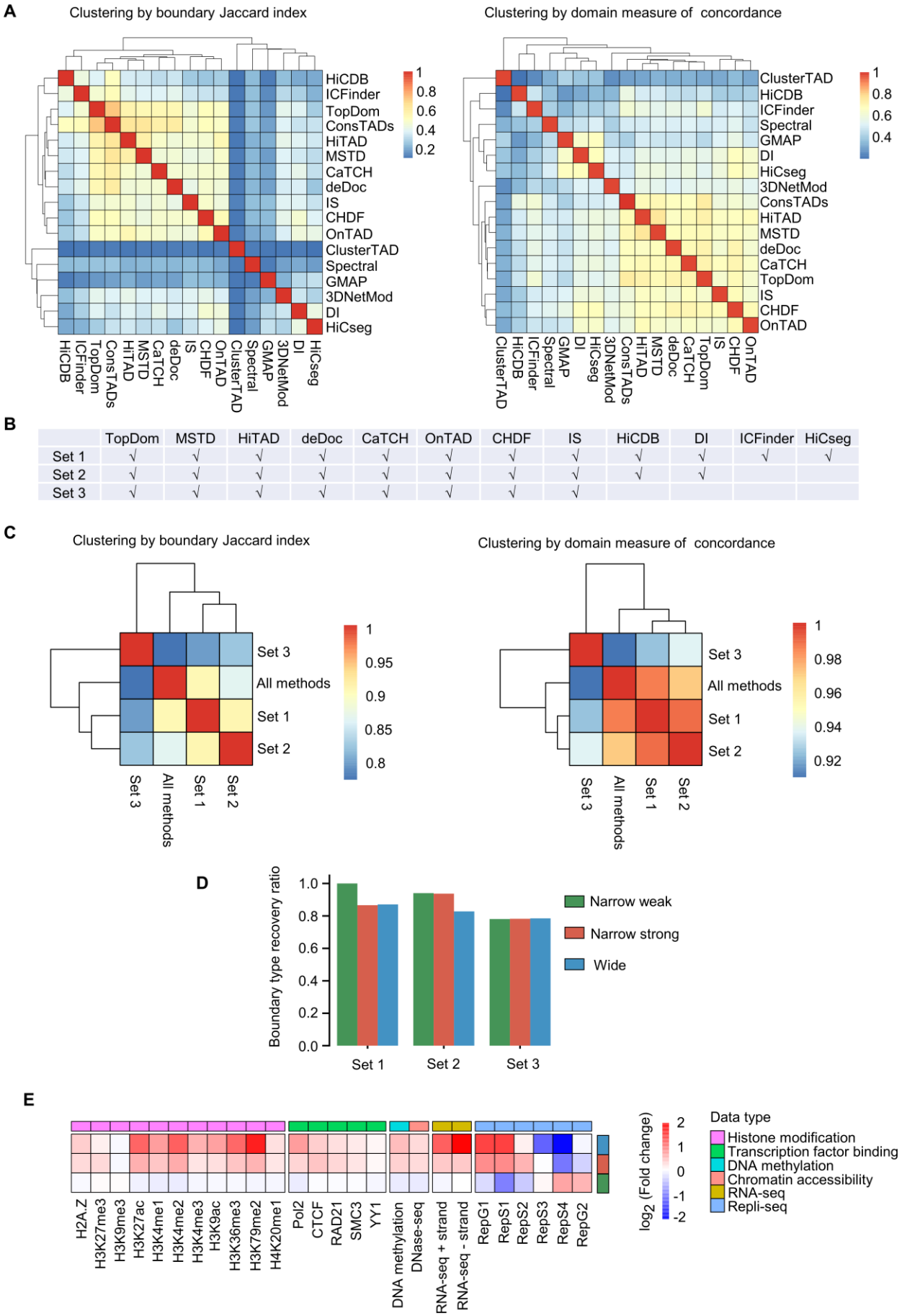
regions in chromosomes 1, 6, 14 and X in GM12878. (B) Fold change profiles of Segway states, subcompartments, and chromHMM states calculated for the three types of boundary regions in GM12878. Fold change is defined as the total length of a state or subcompartment in boundary regions divided by its expected length across the whole chromosome. (C) Enrichment z-score of repeat elements in three types of boundary regions identified from GM12878. These repeat elements can be divided into seven types, labeled with different colors. (D) Density of the *Alu* subfamilies and TcMar-Tigger in three types of boundaries identified with 50-kb Hi-C data under hg19 reference genome in GM12878. (E) Density of the *Alu* subfamilies and TcMar-Tigger in three types of boundaries identified with 10-kb Hi-C data under hg38 reference genome in K562. Mann-Whitney *U* tests were performed between boundary regions and randomly selected regions, respectively, * represents p-value < 0.01.



Supplemental Fig. S22. ConsTADs in Hi-C and Micro-C contact maps with 10 kb and 200 bp resolution and the relationship between boundary type and boundary probability in single cells.

(A) An example of ConsTADs and boundary regions found on the Hi-C contact map of Chromosome 2 in K562 at 10 kb resolution. (B) Fold change profiles of multiple types of biological data constructed for three types of boundary regions found on the Hi-C contact map with 10 kb resolution. (C) Contact

domains found by six representative TAD-calling methods on 200-bp Micro-C contact maps. (D) ConsTADs and boundary regions found in the same region in (C). (E) Probability of genomic regions labeled as the three types of boundary regions (or non-boundary ones) acting as TAD boundaries on Chromosome 2 of IMR90 single cells. (F) A representative proximity frequency matrix of genomic regions on the 3029 single chromosomes of IMR90 cells and the number of chromosomes in which each genomic region is defined as a TAD boundary. The label for each region indicating its boundary type is also shown.



Supplemental Fig. S23. Comparison of ConsTADs and 16 TAD-calling methods as well as a light version of ConsTADs with fewer methods.

(A) Clustering results of ConsTADs and 16 TAD-calling methods based on boundary Jaccard index (left) and domain measure of concordance (right). (B) Three sets containing part of TAD-calling methods. (C) The similarity of ConsTADs integrated by all 16 methods and three method sets, clustering results are based on the Jaccard index (left) and domain measure of concordance (right). (D) The ratio of three types of boundary regions recovered by results of three method sets. (E) Fold change profiles of multiple types of biological data constructed for three types of boundary regions based on the method set 2.

Supplemental Table S1. The details of data used in this study.

See additional file Supplemental_Table_S1.xlsx

Supplemental Table S2. The parameters used in this study for the 16 TAD-calling methods.

See additional file Supplemental_Table_S2.xlsx

Supplemental Table S3. Computation time and RAM used by 16 TAD-calling methods and ConsTADs.

Computation time and RAM used by 16 TAD-calling methods on GM12878 Chr1-ChrX (50 kb)					
Index	Method	All time (s)	Max time (s)	All RAM (GB)	Max RAM (GB)
1	OnTAD	35.473	3.72	15.64	0.79
2	Spectral	85	11	71.65	4.08
3	TopDom	85	12	13.984	0.745
4	MSTD	160.460489	18.3561101	1.058	0.046
5	CaTCH	219	26	64.01	3.32
6	IS	226	24	0.998	0.057
7	GMAP	272.44	26.02	34.43	1.87
8	HiCDB	308	308	0.83	0.83
9	deDoc	479	90	253.43	14.33
10	ICFinder	745	79	151.8	6.96
11	HiTAD	883	883	4.12	4.12
12	DI	906	78	16.805	0.743
13	HiCseg	1474	355	67.57	4.12
14	CHDF	4043	899	11.02	1.59
15	ClusterTAD	77491	13961	128.68	6.22
16	3DNetMod	132661	8834	5.28	0.27
*****	ConsTADs	552.8	52.3	3.01	0.39

Reference

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34**: D590-598.

Su JH, Zheng P, Kinrot SS, Bintu B, Zhuang X. 2020. Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. *Cell* **182**: 1641-1659 e1626.

Zufferey M, Tavernari D, Oricchio E, Ciriello G. 2018. Comparison of computational methods for the identification of topologically associating domains. *Genome Biol* **19**: 217.