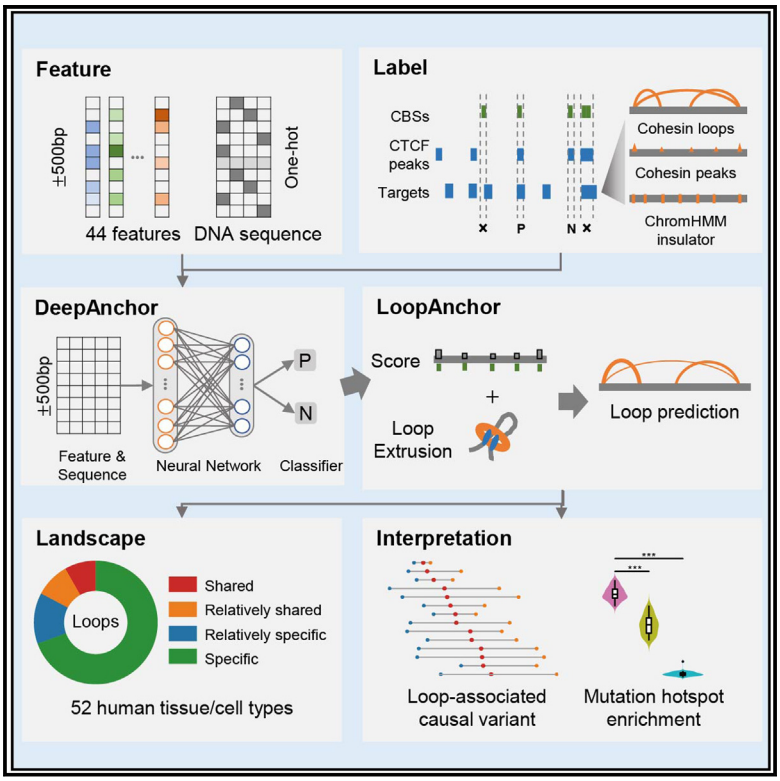
Patterns

Inferring CTCF-binding patterns and anchored loops across human tissues and cell types

Graphical abstract



Highlights

- A deep-learning model, DeepAnchor, characterizes CTCF-binding consequences
- Fine-grained, base-wise genomic/epigenomic features shape CTCF-binding patterns
- An enhanced implementation, LoopAnchor, outperforms existing loop prediction methods
- A landscape is established for CTCF-anchored loops across 52 human tissue/cell types

Authors

Hang Xu, Xianfu Yi, Xutong Fan, ..., Kexin Chen, Dandan Huang, Mulin Jun Li

Correspondence

chenkexin@tjmuch.com (K.C.), mikey.huang2011@gmail.com (D.H.), mulinli@connect.hku.hk (M.J.L.)

In brief

Xu, Yi, Fan, et al. introduce a deep-learning-based model, DeepAnchor, which uses fine-grained features to predict CTCF-binding sites. Prediction scores from this model are integrated into an existing loop extrusion model to predict CTCF-anchored loops, resulting in improved performance compared to current methods. Using this model, called LoopAnchor, they create a comprehensive map of CTCF-anchored loops for 52 different human tissue and cell types, advancing our understanding of how CTCF-controlled regulatory elements influence gene regulation in a context-specific manner.

Xu et al., 2023, Patterns *4*, 100798
August 11, 2023
https://doi.org/10.1016/j.patter.2023.100798

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Inferring CTCF-binding patterns and anchored loops across human tissues and cell types

Hang Xu,[1,2,12] Xianfu Yi,[3,12] Xutong Fan,[3,12] Chengyue Wu,[4] Wei Wang,[1] Xinlei Chu,[1] Shijie Zhang,[5] Xiaobao Dong,[6] Zhao Wang,[5] Jianhua Wang,[3] Yao Zhou,[3] Ke Zhao,[5] Hongcheng Yao,[7] Nan Zheng,[8] Junwen Wang,[9] Yupeng Chen,[4] Dariusz Plewczynski,[10] Pak Chung Sham,[7] Kexin Chen,[1,*] Dandan Huang,[11,*] and Mulin Jun Li[1,3,13,*]

[1]Department of Epidemiology and Biostatistics, Key Laboratory of Prevention and Control of Human Major Diseases (Ministry of Education), National Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin 300070, China

[2]Singapore Immunology Network (SIgN), Agency for Science, Technology and Research (A*STAR), Singapore 138648, Singapore

[3]Department of Bioinformatics, The Province and Ministry Co-sponsored Collaborative Innovation Center for Medical Epigenetics, School of Basic Medical Sciences, Tianjin Medical University, Tianjin 300070, China

[4]Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Tianjin Medical University, Tianjin 300070, China

[5]Department of Pharmacology, Tianjin Key Laboratory of Inflammation Biology, School of Basic Medical Sciences, Tianjin Medical University, Tianjin 300070, China

[6]Department of Genetics, School of Basic Medical Sciences, Tianjin Medical University, Tianjin 300070, China

[7]Centre for PanorOmic Sciences-Genomics and Bioinformatics Cores, The University of Hong Kong, Hong Kong 999077, China

[8]Department of Network Security and Informatization, Tianjin Medical University, Tianjin 300070, China

[9]Department of Health Sciences Research and Center for Individualized Medicine, Mayo Clinic, Scottsdale, AZ 85259, USA

[10]Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland

[11]Wuxi School of Medicine, Jiangnan University, Wuxi 214122, China

[12]These authors contributed equally

[13]Lead contact

*Correspondence: chenkexin@tjmuch.com (K.C.), mikey.huang2011@gmail.com (D.H.), mulinli@connect.hku.hk (M.J.L.)

https://doi.org/10.1016/j.patter.2023.100798

THE BIGGER PICTURE In complex organisms, highly compact chromatin organizes long DNA molecules in the cell nucleus. Surprisingly, these DNA molecules can remain untangled, a property that has recently been attributed to a process called chromatin loop extrusion. Cohesin proteins, which serve as the basic units for loop extrusion, work together with specific transcription factors to create distinct DNA loops. These loops, which are controlled by an elegant architecture protein called the CCCTC-binding factor (CTCF), play a crucial role in regulating genes, cell development, and disease progression. Current computational methods for predicting CTCF-mediated chromatin loops take into account both DNA sequence and chromatin features, but they struggle to capture fine-grained patterns, possibly due to limitations in machine-learning algorithms. Deep-learning algorithms could reveal detailed patterns and enhance our understanding of CTCF binding's functional consequences.

Proof-of-Concept: Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

CCCTC-binding factor (CTCF) is a transcription regulator with a complex role in gene regulation. The recognition and effects of CTCF on DNA sequences, chromosome barriers, and enhancer blocking are not well understood. Existing computational tools struggle to assess the regulatory potential of CTCF-binding sites and their impact on chromatin loop formation. Here we have developed a deep-learning model, DeepAnchor, to accurately characterize CTCF binding using high-resolution genomic/epigenomic features. This has revealed distinct chromatin and sequence patterns for CTCF-mediated insulation and looping. An optimized implementation of a previous loop model based on DeepAnchor score excels in predicting CTCF-anchored loops. We have established a compendium of CTCF-anchored loops across 52 human tissue/cell types, and this suggests that genomic disruption of these loops could be a general mechanism of disease pathogenesis. These computational models and resources can help investigate how CTCF-mediated *cis*-regulatory elements shape context-specific gene regulation in cell development and disease progression.

INTRODUCTION

The diverse interactions between enhancers and target genes have been systematically profiled to clarify tissue/cell-type-specific transcriptional regulation.[1,2] However, the mechanisms underlying the precise relationships between genes and enhancers at varied distances remain unclear. As the main insulator-related transcription factor (TF) discovered in vertebrates, CCCTC-binding factor (CTCF) assists cohesin in chromatin loop formation, which is believed to constitute the fundamental biophysical basis for distal gene regulation.[3–9] The loop extrusion model[10,11] states that cohesin dimers load onto DNA at NIPBL-binding sites and slide along the DNA until they reach the anchors, which are typically bound by specific cofactors, such as CTCF, YY1, and ERα.[12–14] CTCF-anchored loops enclose the chromatin loops mediated by YY1 and ERα, which directly mediate enhancer-promoter interactions.[14] The nested genome units are termed insulated neighborhoods (INs). Substantial evidence suggested that these structures constitute the mechanistic underpinnings of higher-order chromosome organizations,[15,16] such as topologically associating domains (TADs).[17,18] Enhancers tend to regulate genes within the same IN/TAD, while the regulation across different INs/TADs is more likely to be prevented.[19,20] Therefore, characterizing CTCF-binding patterns and their regulatory consequences would greatly facilitate analyses of in-depth 3D genome regulation.

CTCF and its associated *cis*-regulatory elements (CREs), or CTCF-mediated CREs, are critical in regulating tissue/cell-type-specific gene expression by maintaining chromatin domain boundaries or blocking enhancer activities. Furthermore, CTCF is a versatile TF with several roles in different scenarios, which include gene activation, transcriptional repression, and pausing and alternative splicing.[21–26] Therefore, it is necessary to distinguish insulator or looping-related CTCF-binding sites (CBSs) from those with other functions. Numerous analyses of CTCF functional heterogeneity in the nucleus determined that the CTCF-binding motif could be the key to understanding CTCF behavior at specific sites.[27–29] For example, insulation potency relies greatly on both the number of CBSs in tandem and an upstream sequence flanking the core CTCF motif.[28] Additionally, the CBSs located at loop anchors tend to be arranged in convergent orientation in CTCF/cohesin chromatin loop formations.[30] However, these aforementioned studies either lacked the ability to enumerate the feature patterns of individual CBSs or were limited to specific genomic loci.

Numerous computational tools have been developed to qualitatively or quantitatively predict CTCF-mediated loops,[31–40] but few could specifically evaluate the regulatory potential of the DNA sequence at CBSs and how it affects loop formation. Moreover, the feature factors or combinations that determine the specificity among different types of functional CTCF-binding events remain elusive. Furthermore, it is uncertain whether this knowledge can be utilized to enhance genome-wide CTCF-anchored loop prediction. In this study, we developed an interpretable deep-learning model, termed DeepAnchor, to query the genomic/epigenomic feature types that determine whether a CBS is insulator/cohesin/loop associated. Large-scale base-wise genomic and epigenomic features the high-resolution critical patterns for CTCF-mediated insulation and looping. Incorporation of the predicted DeepAnchor score into a previous loop competition and extrusion model (LEM)[40] revealed that the score aided the combined model (LoopAnchor) in outperforming the existing CTCF-anchored loop prediction methods. Furthermore, a novel landscape of tissue/cell-type-specific CTCF-anchored loops across 52 human tissue/cell types was used to interpret disease-causal variants. Together with the compiled resources, this method will facilitate mechanistic research on 3D chromatin dynamic regulation during cell development and disease progression.

RESULTS

DeepAnchor enables high-confidence CTCF-binding pattern characterization

The key feature of DeepAnchor is the implementation of a deep-learning model that uses base-wise features to detect CTCF-mediated CREs. Typically, the sequence and chromatin status of CBSs in specific tissue/cell types are analyzed using a 1D feature vector, in which regional feature values are averaged for each CBS, regardless of heterogeneous signals across the locus.[32,35] However, a true CTCF-mediated CRE is differentiated from other CBSs by its in-depth architectures,[4,28,41] which presents new requirements for pre-modeling feature characterization. One straightforward solution is to profile a CTCF-binding event using the high-resolution features that surround the CBS. In our implementation, 44 quantitative base-wise genomic/epigenomic features within the ±500 bp region for each CBS were obtained from CADD annotation,[42] and the DNA sequence within ±500 bp of the CBS was extracted and represented by one-hot encoding (Table S1). Consequently, concatenating large-scale annotations with converted DNA features at the base-wise level generated a 1,000 × 48 feature matrix for each CBS (Figure 1A, top left).

We modeled three major forms of CBSs across the whole human genome: insulator-associated CBS (insulator CBS), cohesin-associated CBS (cohesin CBS), and loop-associated CBS (loop CBS). These CTCF-mediated CREs share a common core CTCF-binding motif but can display varied regulatory functions. To prepare high-quality training data for different types of CBS learning, each selected positive site was required to contain (1) putative CBS identified by motif scanning and (2) observed CBS detected by CTCF chromatin immunoprecipitation sequencing (ChIP-seq) at a given tissue/cell type. For insulator CBS, the selected sites were intersected with genomic segments exhibiting the ChromHMM insulator state (15-state model). For cohesin CBS, the selected sites were overlapped with peaks detected by cohesin ChIP-seq in matched cell types. For loop CBS, the selected sites were required to be in cohesin loop anchors, which are identified by chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) data from the same cell type (Figure 1A, top right). Furthermore, we randomly selected an equal number of negative samples from the whole CBS pool. To avoid ambiguous results, multiple CBSs sharing the same CTCF/cohesin ChIP-seq peak(s) or the same cohesin loop anchor(s) were excluded (see experimental procedures).

Incorporating base-wise features of the region spanning the selected CBSs significantly expands the feature volume and scale, which can yield an overwhelming amount of information

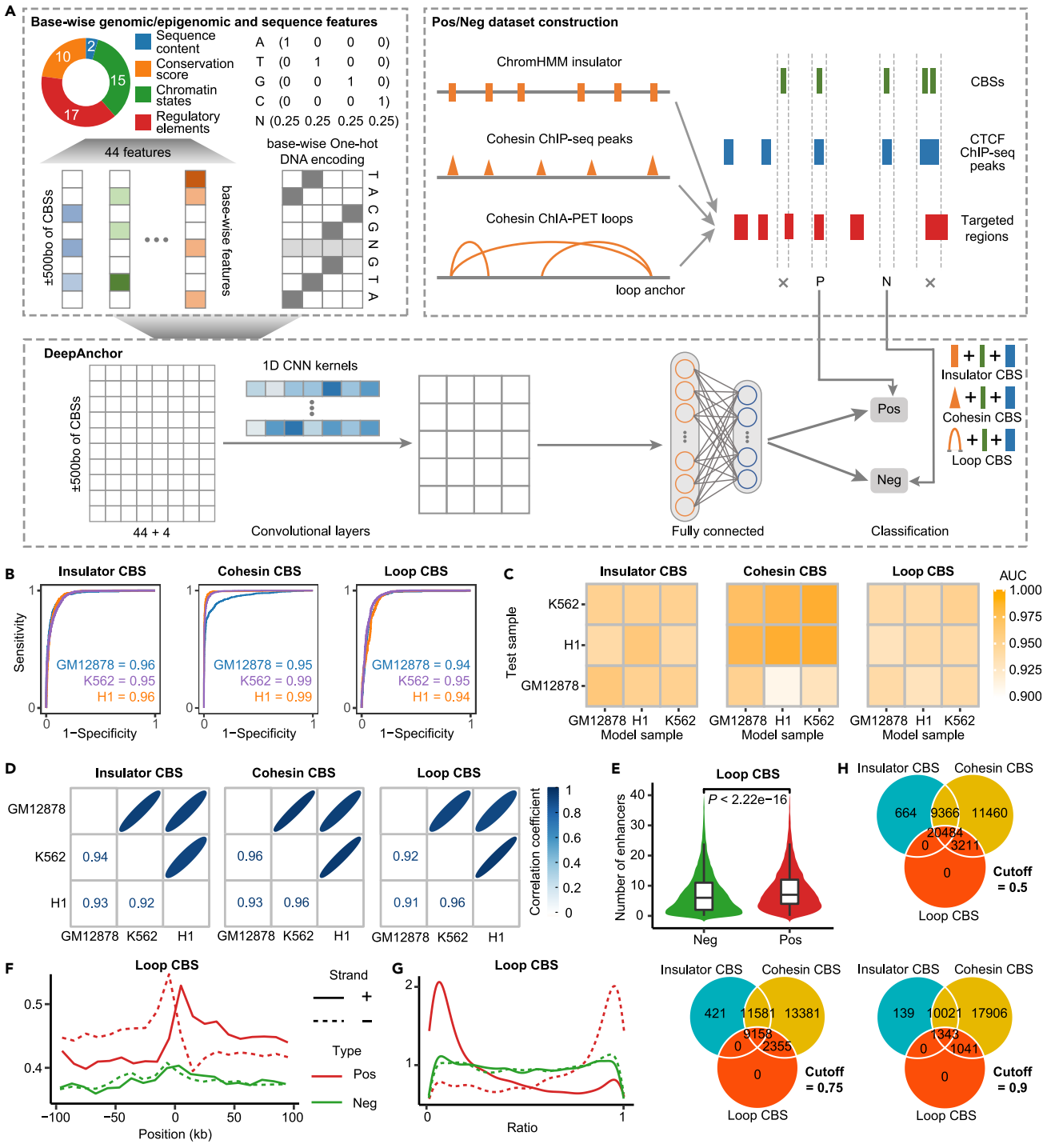


Figure 1. Structure and performance of DeepAnchor model

(A) Schematic view of DeepAnchor model. Base-wise chromatin features and DNA sequences are extracted for all candidate CBSs identified by motif scanning. Positive (Pos) and negative (Neg) datasets are constructed by considering both CTCF ChIP-seq peaks and targeted chromosome intervals, including ChromHMM insulator-associated CBS (insulator CBS), cohesin ChIP-seq signal-associated CBS (cohesin CBS), and cohesin ChIA-PET loop-associated CBS (loop CBS). A 1D-CNN model is then used to train a classifier for distinguishing positive CBSs (CTCF-mediated CREs) from other ones. The probability of CBSs being insulator/cohesin/loop-associated can be calculated and used as the DeepAnchor score for downstream analyses. Related terminologies are as follows. Insulator: an enhancer blocker or a barrier between heterochromatin and euchromatin. Chromatin loop: during the interphase of a cell, the condensed chromatin forms a 3D structure within the cell nucleus. The basic loop-like structure is called a chromatin loop. Loop anchor: given a chromatin loop detected by ChIA-PET, we call the endpoints of the loop on the chromosome a loop anchor. Insulator/cohesin/loop CBS: by using different targeted regions, we obtain different P/N datasets and train different DeepAnchor models. CBS predicted by different models will be named by a particular targeted region.

(legend continued on next page)

that often exceeds the processing capacity of classical machine-learning algorithms. For example, 1,000-bp regions can incorporate 1,000 × 48 = 48,000 features if the base-wise value is used. Important patterns were extracted from the feature matrices by a deep convolutional neural network (CNN) to overcome the challenge of addressing numerous features, which greatly reduced the computational burden and increased prediction performance. Training a classifier with positive and negative datasets enabled the DeepAnchor framework to generate a probability score that indicated the likelihood that a CBS region was a true CTCF-mediated CRE. This score is referred to as the DeepAnchor score in downstream analyses (Figure 1A, bottom; see experimental procedures).

The DeepAnchor model used high-quality CTCF/RAD21 ChIP-seq and RAD21 ChIA-PET data on the GM12878, K562, and H1-hESC cell lines, and DeepAnchor repeatability was tested by cross-validation (see experimental procedures). To minimize the potential influence of chromosome-specific confounding factors, the train-test-validation set was partitioned by dividing the CBSs into subsets according to chromosome. The training, validation, and test sets involved chr1–16, chr17–18, and the remaining chromosomes, respectively. Five-fold cross-validation across the three cell lines determined that the DeepAnchor models trained on different CBS sample types all had high areas under the receiver-operating characteristic (ROC) curves (AUC = 0.94–0.99) (Figure 1B). The comparability of DeepAnchor scores from different samples was demonstrated by cross-cell-type validation across the three cell lines. Remarkably, the DeepAnchor models demonstrated consistently high prediction performance on all testing pairs across different CBS types (all AUCs >0.9, Figure 1C), and the DeepAnchor scores were highly correlated among the three cell-type-specific models (Pearson's r = 0.91–0.96) (Figure 1D). These results suggested that CTCF-mediated CREs share conservative features in different contexts and highlighted the feasibility of predicting genome-wide functional CBS at the organism level.

We established a threshold to distinguish positive and negative CTCF-mediated CREs by identifying the optimal points on the ROC curve with the maximum Youden's J statistic. The optimal threshold was obtained when the DeepAnchor score was close to 0.5 for different models, where 86%–95% positive and 85%–88% negative samples in the training datasets were correctly assigned. The GM12878 DeepAnchor models and the optimal cutoff enabled the generation of genome-wide positive and negative CTCF-mediated CREs. The FANTOM5 enhancer dataset revealed that the positive CBSs colocalized with more enhancers within ±100 kbp regions than the negative CBSs on average ($p < 2.22e{-}16$, Mann-Whitney U test) (Figures 1E and S1A). Moreover, counting the number of enhancers in every 10 kbp region within ±100 kbp of CTCF-mediated CREs revealed a strand-oriented asymmetric pattern only for positive CBSs, particularly in loop CBSs and cohesin CBSs (Figures 1F and S1B). Consistent with previous findings,[43] the CTCF-mediated CRE density around TAD boundaries exhibited strand bias, whereby positive CBSs at the plus-strand were typically enriched at the left TAD boundary while those at the minus-strand were aggregated at the right TAD boundary. Contrastingly, such patterns were not observed for the negative CBSs (Figures 1G and S1C). Furthermore, we examined the intersection of the predicted CTCF-mediated CREs (at different thresholds) among the three CBS types. We determined that all loop CBSs and most insulator CBSs were cohesin related, which indicated that cohesin-bound CBSs might have additional functions beyond forming chromatin loops and insulation. Increasing the DeepAnchor cutoff score significantly reduced the overlap between the three CBS types, with only slight alterations in the overlap between the insulator CBSs and cohesin CBSs (Figure 1H). Together, these results demonstrated that the DeepAnchor model can accurately capture high-confidence CTCF-mediated CREs across the human genome.

DeepAnchor reveals distinct base-wise chromatin and sequence features for CTCF-mediated CREs

Although DeepAnchor uses the deep-learning structure to extract high-level configurations from large-scale features, it also presents challenges in interpreting feature importance. Base-wise representation of various genomic/epigenomic features in the DeepAnchor model enabled the comprehensive depiction of the CTCF-mediated CRE patterns. To visualize the underlying relationships among these features, we first clustered the feature scores of the training dataset and determined that these features could be generally partitioned into three major subsets: transcription-associated, conservation-associated, and chromatin-state-associated (Figure S2). The DeepAnchor output was clarified using SHapley Additive exPlanations (SHAP),[44] which is a game-theoretic approach. SHAP returns the Shapley value matrix with the same alignment as the input feature matrix, which depicts the contribution of each feature at each position. To evaluate the overall contribution of each feature, the mean Shapley values at all positions were calculated and summarized with the mean values (Figure 2A). Overall, the feature of CTCF-binding evidence at open chromatin (EncOCctcfPval) greatly contributed to the model. Furthermore, G/T/A/C Shapley values are top-ranked, which implies the importance of DNA sequence. As expected, open chromatin features (e.g., EncOCFairePVal, EncOCC) displayed higher

(B) Cross-validation ROC curves based on GM12878, K562, and H1-hESC datasets, respectively, among three types of CBSs.
(C) Cross-sample ROC curves for DeepAnchor models on different cell types among three types of CBSs.
(D) Correlation between DeepAnchor scores for three cell-type-specific models among three types of CBSs.
(E) Comparison of the number of enhancers around CBSs between predicted Pos and Neg CTCF-mediated CREs in loop CBS model. The Mann-Whitney U test was used to test the significance.
(F) Strand-oriented asymmetric pattern of enhancer enrichment at predicted Pos and Neg CTCF-mediated CREs in loop CBS model.
(G) Position and strand preference of TAD boundary enrichment by measuring the distance between each CBS and the 5′ end of TAD it belongs at predicted Pos and Neg CTCF-mediated CREs in loop CBS model.
(H) The intersection of the predicted CTCF-mediated CREs among three CBS types at different thresholds.
See also Figure S1 and Table S1.

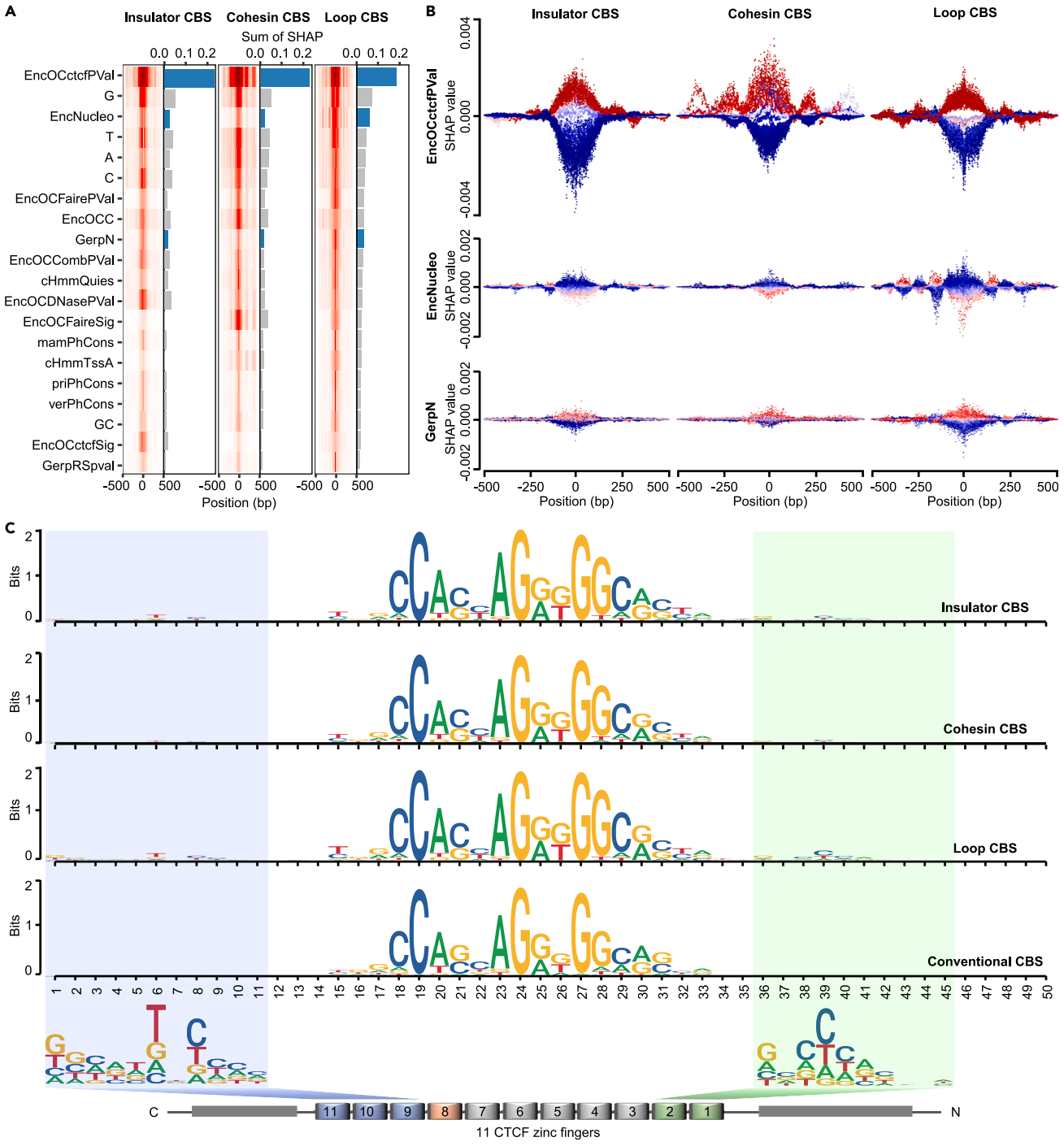


Figure 2. Base-wise analysis of sequence and chromatin patterns across different types of CTCF binding

(A) Feature importance analysis of top 20 features at a base-wise level among three types of CBSs. Heatmap: average absolute feature Shapley values of each position at ±500 bp of CBS. Bar plot: summation of absolute feature Shapley values across ±500 bp of CBS.

(B) Base-wise Shapley value distribution for three representative features across different types of CBSs. EncOCctcfPval, p value (PHRED-scale) of CTCF evidence for open chromatin; EncNucleo, maximum of ENCODE nucleosome position track score; GerpN, neutral evolution score defined by GERP++. Please refer to more feature descriptions from Table S1.

(C) Comparison of DNA motifs associated with positive CBSs among three types of CBSs in this study and a commonly used conventional CBS.

See also Figures S2–S4.

importance than other features (Figures 2A and S3). Interestingly, nucleosome positioning evidence (EncNucleo) displayed more contribution in the loop CBS model, which suggested that the CTCF footprinting pattern among nucleosomes could determine the CTCF-mediated chromatin loop formulation. A conservation feature, the neutral evolution score defined by GERP++ (GerpN), also demonstrated a weak preference in the loop CBS model (Figure 2A). Taken together, the results suggested that certain features might capture distinct and crucial patterns that can differentiate CTCF-mediated CREs from other types of CTCF-binding events.

To explore the base-wise schema of CTCF-mediated CREs, we analyzed the distribution of Shapley values derived from the DeepAnchor model for several important features at every position of the CBS and its surrounding region (±500 bp) using the testing dataset with randomly sampled CBSs. For example, CTCF-binding intensity and its occupancy (EncOCctcfPval) displayed a periodic pattern at cohesin/loop CBSs instead of insulator CBSs, in which higher binding intensity of the central CBS heavily contributed to positive CBS discrimination but the Shapley values inverted approximately 150 bp away from the central CBS (Figure 2B, top). Interestingly, the Shapley values of the nucleosome positioning signature (EncNucleo) were reciprocally distributed only at loop CBSs (Figure 2B, middle), and the nucleosome positioning level was inversely correlated with its Shapley value (Figure S4A). These results suggested that CTCF at loop anchor sites might be involved in expanding and protecting the linked DNA between adjacent nucleosomes, thereby facilitating loop formation. The base-wise visualization of feature importance also strongly supported a previous report which stated that CTCF binding anchors nucleosome positioning and leads to the presence of well-positioned nucleosome flanking sites at specific CBSs.[45,46] Furthermore, we determined that the Shapley values of the conservation feature (GerpN) exhibited a distribution pattern similar to those of CTCF-binding features, particularly in loop CBSs (Figure 2B, bottom). Consistent with previous results,[47] investigation of base-wise Shapley value distribution in the positive or negative training datasets revealed that CTCF-mediated CREs typically demonstrated stronger binding intensity than other CBS types (Figure S4B).

Previous CTCF multivalency studies revealed that CTCF binds on diverse sequences through combinatorial clustering of its 11 zinc fingers (ZFs),[27,48] yet the sequence determinants of CTCF-mediated insulation remain elusive. By applying MEME motif discovery to all predicted positive and negative CBSs, two additional weak motifs flanking the core CTCF motif were revealed at only the positive loop and insulator CBSs (Figure 2C). Agreeing with a sensitive insulator reporter assay,[28] ZFs 9–11 recognized the upstream motif 5 or 6 bp away from the core sequence and might contribute to CTCF CBS binding directionality.[49] Interestingly, the distance and sequence context of the downstream motif beside the core sequence differed from that of previous findings.[27] Nonetheless, this downstream sequence was associated with the CTCF N terminus and ZFs 1–2 and stabilizes cohesin engagement.[50,51] Therefore, the distinct sequence features and multiple lines of evidence for CTCF-binding patterns demonstrated the fidelity of the DeepAnchor model.

Incorporating the DeepAnchor score into LEM improves the accuracy of CTCF-anchored loop prediction

Although DeepAnchor models the insulative and looping potential of genome-wide CBSs at the organism level, it cannot be used to explain chromatin loop formation, as the loop CBS model only measures the single end of loops independently. We established the relationship between loop formation and the features of both CTCF-mediated anchors to apply the DeepAnchor score to CTCF-anchored loop prediction. A simple but effective model, LEM, has been described previously, in which the chromatin loops profiled in CTCF ChIA-PET experiments were clarified as cohesin blocking and localized at CBSs.[40] This quantitative model evaluates CTCF-anchored loop formation with only four features: CTCF-binding intensity, CTCF motif orientation, distance between CTCF-binding events, and loop competition. Notwithstanding the acquired specificity, we reasoned that incorporating the loop-associated DeepAnchor score into the LEM could improve CTCF-anchored loop prediction. To this end, we weighted the probability of CTCF binding in the original LEM by considering the looping potential of CTCF binding (the anchor score predicted by the loop-associated DeepAnchor model) and achieved a new implementation, LoopAnchor (see experimental procedures).

The performance of LoopAnchor was compared with that of five prevalent algorithms for CTCF-mediated interaction prediction: the Naive and Oti methods,[37] CTCF-MP,[35] Lollipop,[32] and LEM.[40] To ensure fair benchmarking, all supervised models were trained on GM12878 RAD21 ChIA-PET data and tested on K562 RAD21 ChIA-PET data. The LoopAnchor ROC-AUC (0.936) and PR-AUC (0.921) both suggested that it outperformed the other state-of-the-art methods (Figure 3A), which indicated the effectiveness of our model optimization. The correlation analysis between the predicted loop intensities and the observed ChIA-PET signals on K562 demonstrated that LoopAnchor achieved the highest correlation (Pearson's r = 0.653) among the six tools (Figure 3B), which also demonstrated the superiority of LoopAnchor for the quantitative measurement of CTCF-anchored loops.

We investigated whether our LoopAnchor model could capture the dynamic changes during cell development by applying it to human monocyte activation data with the paired *in situ* Hi-C and CTCF ChIP-seq data before and after exposure to phorbol 12-myristate 13-acetate.[52] The LoopAnchor model accurately predicted 25 of 34 gained loops, and all six lost CTCF-anchored loops (Figure 3C). For example, Hi-C detected eight loops on the *JAG1* locus at chromosome 20:9547732-11575097 (GRCh37/hg19), among which one gained loop was identified only at differentiated macrophages (Figure 3D). Consistent with this, LoopAnchor not only detected all unchanged loops but also accurately predicted the gained loop (Figure 3D; see experimental procedures). Notably, the loop intensity differences in this genomic region predicted by LoopAnchor between monocytes and macrophages were highly correlated with the loop score changes measured from the Hi-C data (Pearson's r = 0.796). These results indicate that LoopAnchor could quantitatively capture local loop structure changes during the dynamic cellular process.

A landscape of CTCF-anchored loops across 52 human tissue/cell types

Given the simplicity and improved performance of the LoopAnchor model, we used existing large-scale CTCF ChIP-seq data to derive

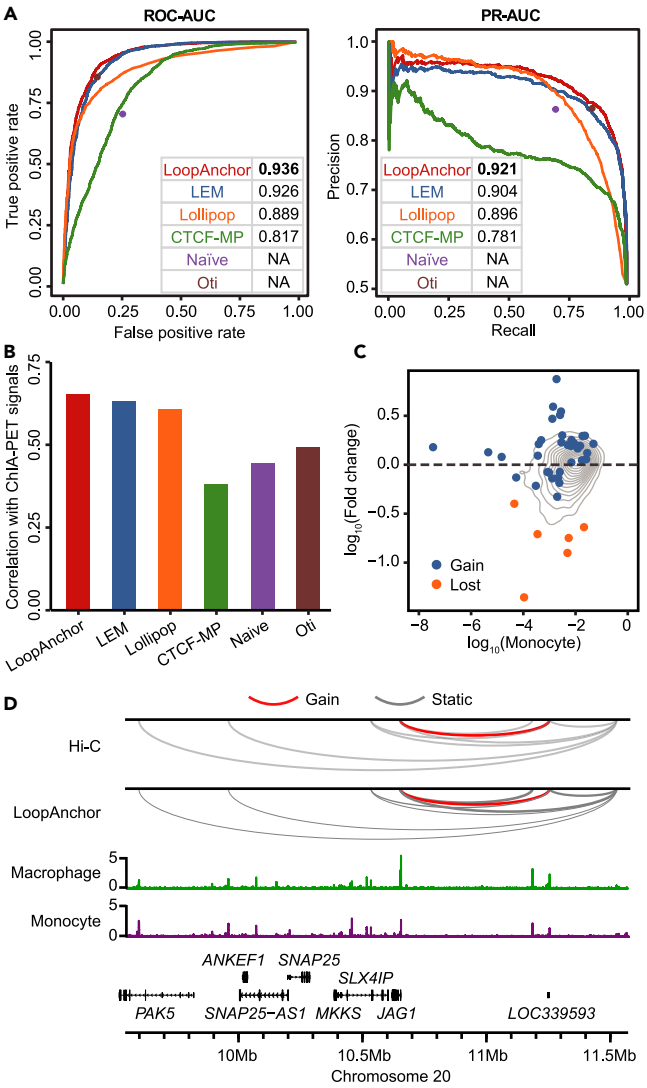


Figure 3. Performance evaluation of LoopAnchor for CTCF-anchored loop prediction

(A) ROC curves, precision-recall curves, and associated AUCs among LoopAnchor and five state-of-the-art methods. All supervised models, such as LoopAnchor, LEM, Lollipop, and CTCF-MP, were trained on GM12878 RAD21 ChIA-PET data and independently tested on K562 RAD21 ChIA-PET data.

(B) Correlation between predicted scores and real RAD21 ChIA-PET loop intensity on K562.

(C) Gained and lost loops from monocyte to macrophage differentiation by comparing LoopAnchor prediction with Hi-C observation. $\log_{10}$(Fold change) is the transformed fold change of predicted loop intensity for a specific loop between macrophage and monocyte; $\log_{10}$(Monocyte) is the transformed Hi-C loop score observed in monocyte; blue dot is lost Hi-C loop; orange dot is gained Hi-C loop.

(D) Example of loops predicted by LoopAnchor at *JAG1* locus. For Hi-C loops, line color is used to distinguish the gained and static loops. For LoopAnchor, line width represents the predicted loop intensity, and the loop with a fold change of intensity >3 is marked in red.

a global picture of CTCF-mediated chromatin interactions across various human tissue/cell types. Initially, we uniformly processed 740 CTCF ChIP-seq datasets collected from ENCODE,[2] CistromDB,[53] and ChIP-Atlas[54] utilizing the ENCODE TF ChIP-seq processing pipeline (Table S2; see experimental procedures). Our data quality control, peak calling filtering, and redundancy removal process retrieved a dataset containing 168 high-quality CTCF ChIP-seq data related to 113 normal and 55 cancer samples spanning 32 tissues and 20 anatomical cell types (Figure 4A and Table S3; see experimental procedures). Each biosample contained approximately 100,000 peaks (mean = 104,691, median = 96,296, Figure S5A). Genome-wide CTCF-mediated loops for each biosample were detected with LoopAnchor and LEM, which yielded approximately 20,000 loops per biosample (LoopAnchor mean = 16,255, median = 16,637; LEM mean = 17,826, median = 18,252; Figures S5B and S5C).

Based on the shared pattern of detected loops across 168 biosamples, the LoopAnchor loops were classified into four distinct types: (1) 8.2% (n = 6,179) tissue/cell-type-shared loops in >75% of the biosamples; (2) 9.0% (n = 6,802) tissue/cell-type-relatively-shared loops in >50% but <75% of the biosamples; (3) 13.3% (n = 10,016) tissue/cell-type-relatively-specific loops in >25% but <50% of the biosamples; (4) 69.4% (n = 52,216) tissue/cell-type-specific loops in <25% of the biosamples (Figure 4B; see experimental procedures). The LEM loops were also classified into these four types based on the same strategy (Figure S6A). Compared to LEM, LoopAnchor detected more shared loops but fewer specific loops globally (tissue/cell-type-shared loops: 8.2% vs. 3.9%; tissue/cell-type-specific loops: 69.4% vs. 80.3%; Figures 4B and S6A). This suggested that LoopAnchor was more effective for capturing conservative CTCF-anchored loops across different tissue/cell types. Investigation of the LoopAnchor-predicted loop intensity score distribution revealed a significant difference among the four loop categories ($p < 2.22e{-}16$, Kruskal-Wallis test, Figure 4C), where the tissue/cell-type-shared loops received the highest scores while the tissue/cell-type-specific loops received the lowest scores. A similar pattern was observed for the LEM-derived loop intensity score ($p < 2.22e{-}16$, Kruskal-Wallis test, Figure S6B). To evaluate the tissue/cell-type specificity of the classified CTCF-mediated loops, we used the predicted chromatin loops based on Peakachu for 42 human tissue/cell types[55] and compared the number of associated tissue/cell types among different categories, whereby the tissue/cell-type distribution of the LoopAnchor-predicted loops also demonstrated significant differences among the four loop categories ($p < 2.22e{-}16$, Kruskal-Wallis test, Figure 4D). As expected, the number of associated tissue/cell types gradually decreased from the tissue/cell-type-shared group (mean = 4, median = 4.743) to the tissue/cell-type-specific group (mean = 2, median = 2.314), which indicated the effectiveness of our loop classification. A similar trend was identified for the LEM-detected loops ($p < 2.22e{-}16$, Kruskal-Wallis test, Figure S6C).

In summary, we established a comprehensive collection of CTCF-anchored loops that exhibit both commonalities and specificities across a wide range of human tissue/cell types. All predicted loops for the 168 selected biosamples and all 764 biosamples can be visualized as separated tracks at UCSC Track Data Hubs (https://genome.ucsc.edu/cgi-bin/hgHubConnect) by entering the customized hub URLs (https://raw.githubusercontent.com/mulinlab/LoopAnchor/master/loopanchor/data/hubs/hubs_landscape.txt and https://raw.githubusercontent.com/mulinlab/LoopAnchor/master/loopanchor/data/hubs/hubs_all.txt, respectively).

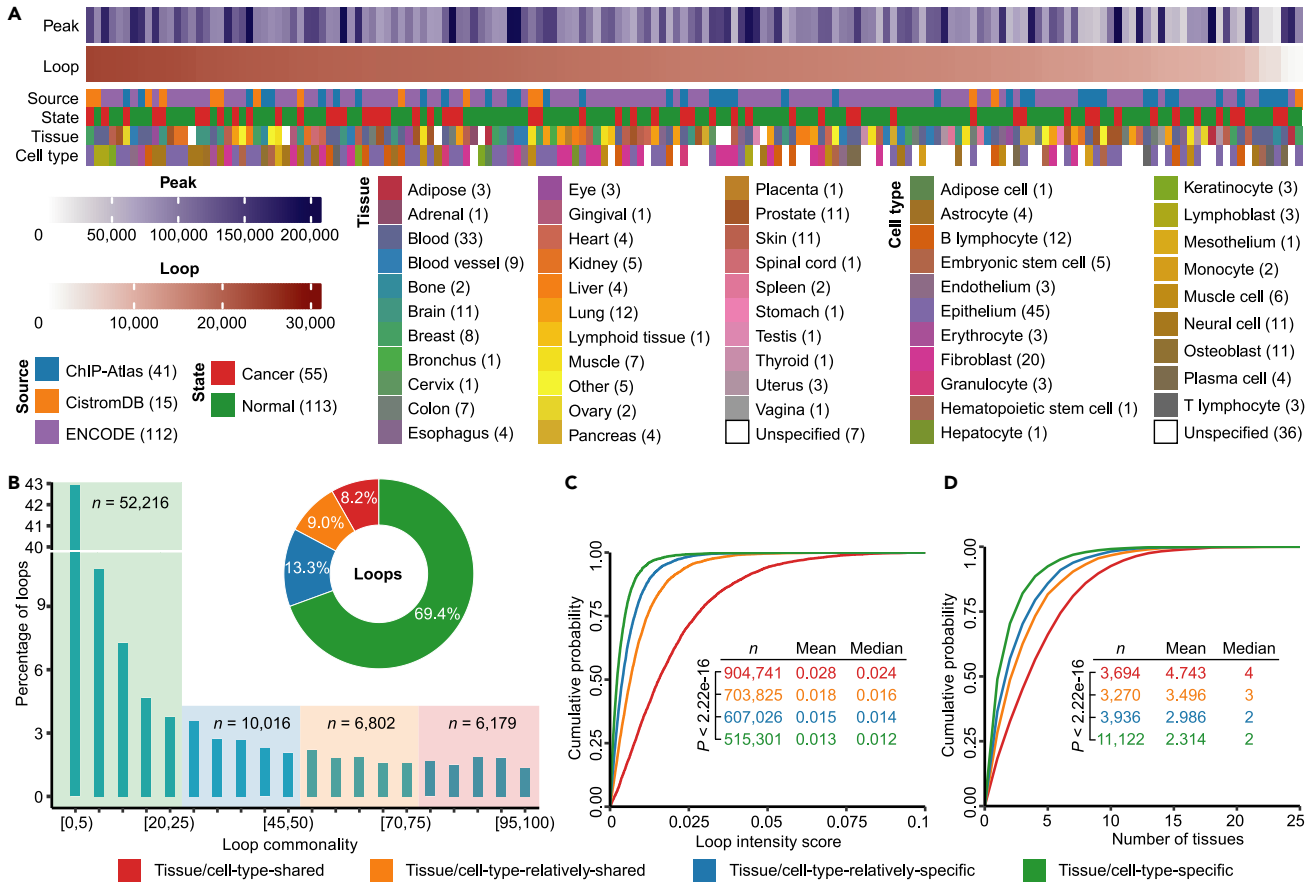


Figure 4. Landscape of predicted CTCF-anchored loops across 32 human tissues and 20 cell types

(A) Overview of predicted CTCF-anchored loops across 168 biosamples (columns) and biological conditions (rows).

(B) Classification of CTCF-anchored loops according to their shared patterns. All predicted loops were classified into four categories, namely tissue/cell-type-specific, tissue/cell-type-relatively-specific, tissue/cell-type-relatively-shared, and tissue/cell-type-shared.

(C) Comparison of loop intensity score for loops in four categories. The cumulative probability was calculated, and the Kruskal-Wallis test was used to test the significance.

(D) Validation of tissue distribution for loops across four categories using predicted chromatin loops based on Peakachu for 42 human tissue/cell types. The cumulative probability was calculated, and the Kruskal-Wallis test was used to test the significance.

See also Figures S5 and S6; Tables S2 and S3.

Tissue/cell-type-specific loop anchors are highly enriched at disease-causal loci

Numerous lines of evidence supported the premise that CTCF/cohesin-binding sites are highly mutated in cancer[56–58] or are constantly shaped via evolutionary selection.[59,60] However, whether context-specific CTCF-mediated looping and associated loop anchors are more likely linked to disease-causal genomic loci has not been systematically tested. We collected 12,738 causal variants for 54 blood-related autoimmune diseases from genome-wide association studies (GWASs)[61] (Table S4) and sampled-matched control variants to evaluate the genome-wide enrichment of autoimmune disease-causal variants on the LoopAnchor-predicted CTCF-anchored loops in normal tissues with at least three biosamples (see experimental procedures). Notably, based on the scores measured by the permutation tests' p values, blood tissue (mean p = 0.0036) was ranked the top tissue for autoimmune disease variant enrichment in the loop anchors (Figure 5A). Additionally, the score distribution comparison revealed that blood tissue was significantly more enriched than 50% of the other tissues (6 of 13, one-tailed Mann-Whitney U test, false discovery rate [FDR] < 0.1, Figure 5A). As most disease-causal regulatory variants exhibit tissue/cell type specificity in phenotypically relevant contexts, these results supported the idea that context-dependent loop information can better interpret GWAS disease-causal variants for complex diseases.

The relevance of CTCF-mediated loop anchors in cancer mutation was assessed by obtaining somatic mutations from the International Cancer Genome Consortium (ICGC) whole-genome mutation aggregation[62] and extracting recurrent mutations in the non-coding regions (see experimental procedures). Comparison with permuted non-recurrent mutations revealed that higher-recurrence mutations obtained greater enrichment in the CTCF-mediated loop anchors in cancer biosamples (two-tailed Mann-Whitney U test, Figure 5B).

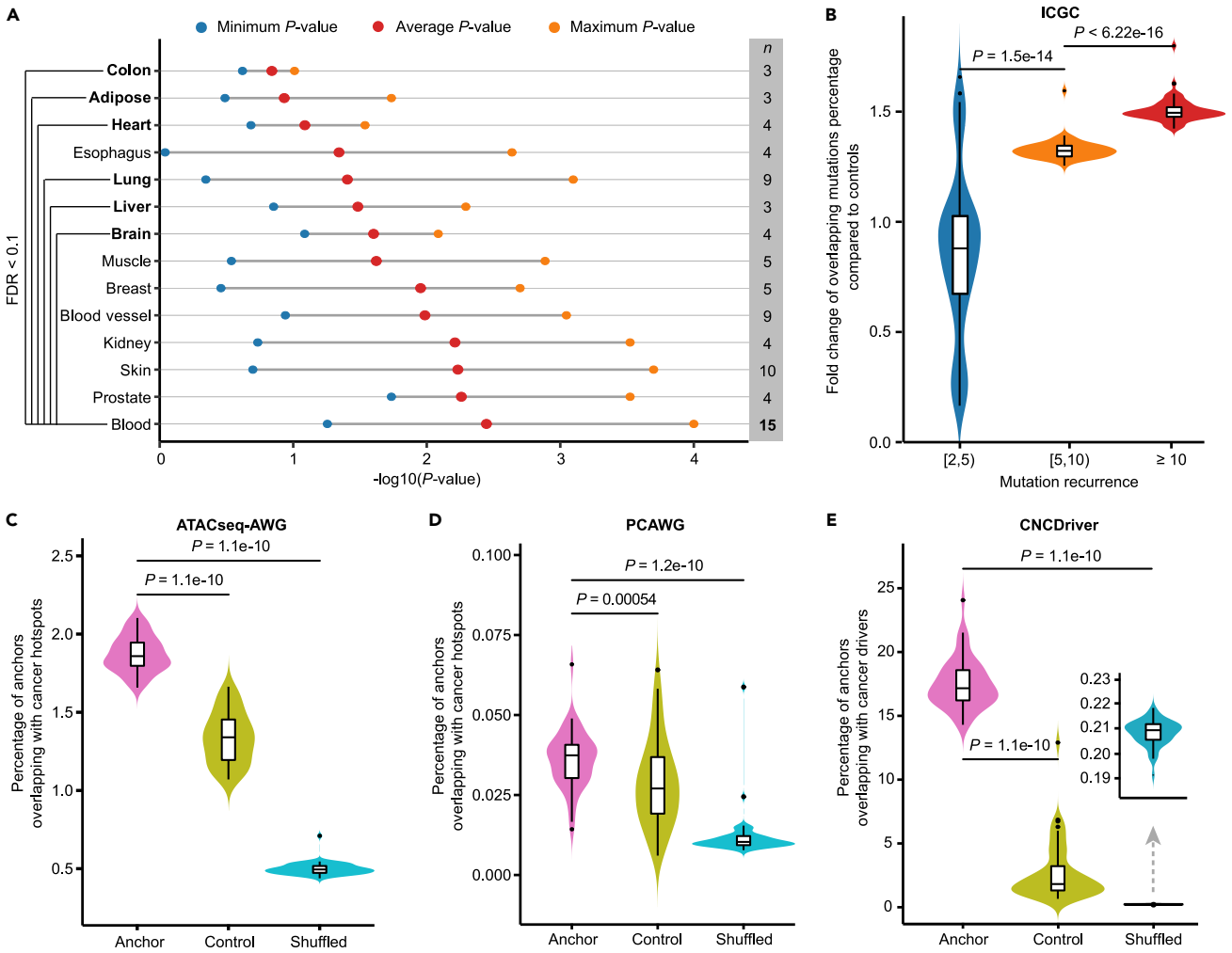


Figure 5. Disease-causal variants and somatic hotspots enrichment

(A) Enrichment significance distribution for autoimmune disease-causal variants among different normal tissues. The tissues were ordered by their average p values. The one-tailed Mann-Whitney U test (FDR < 0.1) revealed that the blood tissue was significantly more enriched than the tissues highlighted in bold.

(B) Comparison of fold change distribution among different somatic mutation recurrence categories. The two-tailed Mann-Whitney U test was used to test the significance.

(C–E) Comparison of overlapped percentage with somatic hotspots using different datasets from ATACseq-AWG (C), PCAWG (D), and CNCDriver (E). The paired two-tailed Mann-Whitney U test was used to test the significance.

See also Table S4.

Whole-genome cancer mutation hotspots collected from three pan-cancer whole-genome analyses—assay for transposase-accessible chromatin using sequencing analysis working group (ATAC-seq-AWG),[63] Pan-Cancer Analysis of Whole Genomes (PCAWG),[64] and CNCDriver[65]—were used to examine whether the cancer mutation hotspots occurred more frequently in CTCF-mediated loop anchors than in other CTCF-binding loci or shuffled non-coding genomic regions (see experimental procedures). We obtained consistent results from the different datasets, where the loop anchors were highly enriched with cancer mutation hotspots (two-tailed Mann-Whitney U test, Figures 5C–5F). These findings suggested that CTCF-anchored loop disruption in the human genome might be a common causal mechanism underlying disease pathogenesis.

DISCUSSION

The multifaceted roles of CTCF in the nucleus motivated ongoing investigations into its phenotypic and mechanistic functions. However, the precise mechanism by which CTCF recognizes and interacts with insulators to exert its effects on chromosome barriers and enhancer blocking require further examination. In this study, we developed a novel computational model to accurately predict genome-wide CTCF-mediated CREs. The incorporation of large-scale base-wise genomic and epigenomic features within a deep-learning model revealed several high-resolution distinct chromatin and sequence features of CTCF-mediated insulation. Importantly, two additional sequence motifs flanking the core CTCF motif at the positive loop and insulator-associated CBSs were identified. Subsequently, the predicted

insulator score was used to optimize the previous LEM and achieved better performance in CTCF-anchored loop prediction. Based on the model, we established a novel compendium of tissue/cell-type-specific and -shared CTCF-anchored loops across 52 human tissue/cell types. Finally, we demonstrated that tissue/cell-type-specific loop anchors are highly enriched at disease-causal loci. Therefore, our results enhance understanding of CTCF-mediated insulation and loop formation. Together with the compiled resource, the new method provides useful approaches for studying the dynamic regulation of 3D chromatin during cell differentiation and disease progression.

Although many computational models can predict CTCF-mediated loops with varied sensitivity and specificity,[31–40,66] they typically learn from CTCF ChIA-PET data and rarely evaluate the regulatory potential and loop attributes among different CBS types. Here, we applied cohesin ChIA-PET/ChIP-seq data, ChromHMM-predicted insulator, and strict CTCF-binding evidence to specifically analyze the CTCF-binding patterns at three types of CTCF-mediated CREs. This yielded a unique tool for characterizing CTCF-binding consequence and loop formation through the loop extrusion mechanism. Notably, the limited availability and varied library construction quality of CTCF/cohesin ChIA-PET restricted the broad training of context-specific models. However, our cross-cell-type comparisons demonstrated high agreement among the DeepAnchor models trained with different ChIA-PET data, which indicated that the discrimination of true CTCF-mediated CREs might rely on several conservative features revealed by our interpretable deep-learning model, such as higher binding intensity, well-positioned nucleosomes, and two unique motifs flanking the central CBS.

The introduction of the DeepAnchor score into LEM[40] improved CTCF-anchored loop prediction performance and demonstrated that the new LoopAnchor method could achieve better quantitative estimation. However, the quantitative predictions in both LEM and LoopAnchor only measured the contribution of each component independently. For example, the insulation potential of CBS and loop competition are dependent processes based on the dynamic chromatin context,[67] which motivates our future optimization direction. Given that all existing methods do not yield CTCF-associated anchor prediction results, we could not evaluate whether incorporating other anchor scores would achieve better performance. Instead, we performed systematic comparisons with five prevalent algorithms for CTCF-mediated loop prediction. While our method was limited to inferring specific CTCF-bound CREs and loops, recent studies indicated that some TFs could function as novel architectural proteins to regulate genome organization,[68] such as YY1,[69] ZNF143,[70] MAZ,[71] BHLHE40,[72] CTCFL,[73] MyoD,[74] and ZBTB3.,[75] some of which are independent of CTCF binding. Based on the targeted regions and negative controls, our model can easily be extended to train classifiers to extract the base-wise features of the binding patterns of these architectural proteins. Therefore, predicting genome-wide full-spectrum anchor sites and their associated looping events warrants further in-depth investigations.

The CTCF-anchored loop landscape established by applying LoopAnchor to 168 uniformly processed human CTCF ChIP-seq biosamples is a valuable resource for 3D genome and regulatory genomics studies. As the anchor scores were estimated at the organism level, LoopAnchor only requires CTCF-binding profiles to accurately predict CTCF-anchored loops. This would greatly simplify the application, particularly for studying the dynamic 3D CTCF code during cell development and disease progression. Additionally, the compiled tissue/cell-type-shared and -specific loops can facilitate the interpretation of disease-causal variants identified by GWASs and somatic non-coding driver mutations in cancers.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Mulin Jun Li (mulinli@connect.hku.hk).

Materials availability

This study did not generate new unique materials.

Data and code availability

The source codes of LoopAnchor are freely available under MIT License at https://github.com/mulinlab/LoopAnchor[76] or https://bitbucket.org/xuhang01/loopanchor. The source code has also been released with v.1.0.0 as the version of the repository at publication time.

All predicted loops for 168 selected biosamples and all 764 biosamples can be visualized and compared as separated tracks at UCSC Track Data Hubs (https://genome.ucsc.edu/cgi-bin/hgHubConnect) by entering customized hub URLs https://raw.githubusercontent.com/mulinlab/LoopAnchor/master/loopanchor/data/hubs/hubs_landscape.txt or https://raw.githubusercontent.com/mulinlab/LoopAnchor/master/loopanchor/data/hubs/hubs_all.txt, respectively.

Training dataset preparation

Three types of target regions were prepared from high-quality functional genomics data of three Tier 1 ENCODE cell lines (GM12878, K562, and H1-hESC), including insulators estimated by ChromHMM[77] (https://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeBroadHmm), cohesin (RAD21) ChIP-seq narrow peaks (CohesinDB: https://cohesindb.iqb.u-tokyo.ac.jp/download/downAllObject/allcohesin.final.tsv.gz), and cohesin (RAD21) ChIA-PET loops (ENCODE: ENCSR752QCX, ENCSR000CAC, ENCSR543YTV).[78] For ChromHMM insulator segments and RAD21 ChIP-seq narrow peaks, the intervals were merged to avoid overlap. RAD21 ChIA-PET data were preprocessed with ChIA-PET2 using default parameters.[79] To increase specificity, inter-chromosomal interactions and intra-chromosomal interactions with less than two pair-end tags were excluded. The anchors of loops were extracted and merged to obtain non-redundant anchors. Three types of CBSs were then prepared by intersecting the target intervals with candidate CBSs identified by motif scanning as well as CTCF ChIP-seq narrow peaks (ENCODE: ENCSR000AKB, ENCSR000BPJ, ENCSR000BNH), including insulator CBS, cohesin CBS, and loop CBS. Thus, for each cell type, the candidate CBSs co-localized with both CTCF ChIP-seq peaks and targeted regions were selected as positive sets while all others constituted the negative pool. To avoid biases, the CBSs that shared the same target region from the positive sets were removed because it was difficult to distinguish which one was the true positive.

Feature preparation

The DNA-binding motif of CTCF (MA0139.1) was downloaded from JASPAR2020.[80] We used FIMO, a tool from the MEME Suite (v5.1.1),[81] to scan the CTCF-binding motif across the whole human genome (GRCh37/hg19), which generated 110,059 CBSs with p value $<1e-5$. According to the topic relevance, we selected 44 base-wise genomic/epigenomic features from the CADD (v1.4) annotation database (https://cadd.gs.washington.edu/static/ReleaseNotes_CADD_v1.4.pdf).[42,82] For each CBS, the feature values within ±500 bp region were extracted and stored together to form a feature matrix. All features were then scaled using the min-max scaling algorithm. The 0.1 and 99.9 percentiles were used as the minimum and maximum values to reduce the influence of extreme high/low values. In addition, the DNA sequence of

the ±500 bp region surrounding the CBS center was obtained using BEDTools.[83] A, T, C, and G were converted into numerical values using one-hot encoding: A (1,0,0,0), T (0,1,0,0), C (0,0,0,1), and G (0,0,1,0). Finally, by concatenating the large-scale annotations with converted DNA features at the base-wise level, a 1,000 × 48 feature matrix was generated for each CBS.

DeepAnchor model and fitting

For each CBS $n = 1, \ldots, N$, genomic/epigenomic features and DNA sequence features at base-wise level for the ±500 bp regions surrounding the CBS center were extracted. By concatenating features and one-hot encoding sequence, a two-dimensional 1,000 × 48 signal matrix x_n is generated for each CBS n. To extract feature patterns from the signal matrix, DeepAnchor uses 1D convolutional kernels to process the signal matrix. For each layer l, the forward propagation from layer $l-1$ to l is expressed as follows:

$$x_k^l = b_k^l + \sum_{i=1}^{N_{l-1}} conv1D\left(w_{ik}^l, x_i^{l-1}\right), \quad \text{(Equation 1)}$$

where N_l is the number of neurons at layer l, x_i^{l-1} and x_k^l are the i^{th} neuron at layer $l-1$ and k^{th} neuron at layer l, and w_{ik}^l and b_k^l are the kernel and the scalar bias from layer $l-1$ to layer l. DeepAnchor implements two 1D-CNN layers, each followed by a max-pooling layer and a drop-out layer. After 1D-CNN layers, all the outputs are fully connected. Three fully connected layers convert the outputs of 1D-CNN layers to lower dimensions and finally train a classifier with sigmoid activation,

$$p_1 = \frac{e^{x_1^f}}{e^{x_1^f} + e^{x_0^f}},$$

$$p_0 = 1 - p_1,$$

where x^f is the result of fully connected layers which contain only two elements for positive training and negative target, respectively. p_1 is the probability that CBS n can be a positive CBS. The DeepAnchor model was implemented using the TensorFlow framework. To avoid overfitting, a balanced negative set was extracted by randomly selecting equal number of samples from the negative set pool. Finally, the positive and negative datasets were divided into train, test, and valid sets according to chromosomes. Samples on chr1–16 were selected as the training set, chr17–18 as the validation set, and the others as the test set. We considered p_1 and p_2 as the possibility that a CBS is a true target-associated CBS or not, and the binary cross entropy was calculated as the loss of model. The input data were split into small batches with a batch size of 50, and the model was trained for 20 epochs. Early stopping was implemented to stop training if the validation loss no longer decreased.

Evaluation of the DeepAnchor model

Five-fold cross-validation was used to evaluate the DeepAnchor model to generate ROC curves and corresponding AUCs. To define positive and negative CBSs, a cutoff was set by finding the points on the ROC curve with the largest Youden's J statistic. As with the training procedure on GM12878, DeepAnchor models were trained for each cell type, and the models were also validated by all test sets belonging to three cell types. After training the model, DeepAnchor computed a score within [0, 1] for all potential CBSs, where larger value indicates a higher possibility for a CBS to be CTCF-mediated CREs. According to Youden's J statistic, CBSs with anchor scores >0.5 were selected as positive CTCF-mediated CREs, while negative CTCF-mediated CREs have anchor scores <0.5. Human enhancer annotation was downloaded from the FANTOM5 database[84] (https://fantom.gsc.riken.jp/5/datafiles/latest/extra/Enhancers/). The number of enhancers for positive/negative insulators was counted if the distance between them was smaller than 100 kbp. We evaluated the differences of involved enhancers between positive and negative CBSs using the Mann-Whitney U test. The estimated TAD data were derived from the GM12878 Hi-C assay,[17] in which there are 4,386 TADs and the median length of TADs is 520 kbp. To evaluate whether CBSs have a positional preference or not, we counted the distance between each CBS and the 5′ end of TAD that contains the CBS. Because of the variable length of TAD, the distance of CBS was scaled by the length of TAD it belongs to.

Base-wise feature contribution analyses

SHAP[44] was used to interpret the output of DeepAnchor. Since the input of the DeepAnchor model is a 1,000 × 48 feature matrix, SHAP returns the Shapley value matrix with the same alignment as the input feature matrix. The shap package[85] was used to analyze the well-trained DeepAnchor model and use the training dataset as background examples to receive an expectation. An explainer was then generated based on the model and background dataset. The explainer calculated the Shapley values for every position via testing datasets. The training and test dataset was the same as the ones used for training the DeepAnchor model. The feature contribution was interpreted by the mean Shapley values for all positions.

Algorithm of LoopAnchor

The DeepAnchor score reflects the phenomenon that CBSs do not have equal functionality in biological processes such as the formation of chromatin loops. However, in previously described algorithms for predicting CTCF-mediated chromatin loops none of them addressed such inequality, although they have used specific features. They also failed to make use of base-wise genomic or epigenomic features. Therefore, it is possible to improve the performance by incorporating the DeepAnchor score in CTCF-anchored loop predictions. Enlightened by a recent published model, LEM,[40] we integrated the loop-associated DeepAnchor score into the model to predict CTCF-anchored loops. The probability of insulator-associated CTCF binding at CBS i was estimated as

$$p_i = \frac{S_i \cdot t_i}{S_i \cdot t_i + a \cdot \overline{F_{re}}},$$

where S_i is the CTCF ChIP-seq signal and a is a constant that has been estimated in the original paper. In this new implementation, t_i is the loop-associated DeepAnchor score at CBS i; thus, revised mean CTCF signal $\overline{F_{re}}$ is given by

$$\overline{F_{re}} = \frac{\sum_i^n S_i \cdot t_i}{n}.$$

Comparison with state-of-the-art methods for CTCF-mediated loop prediction

Oti code was obtained from Oti et al.,[37] and the Naive method code was derived from Oti code by setting the recursive round to 1. Default parameters were used to run Oti and Naive methods, and loops with a score >0 were considered positive loops while others were treated as negative. CTCF-MP,[35] Lollipop,[32] and LEM[40] were installed and used on relevant cell types according to their GitHub repositories. To facilitate comparison among different methods, all methods were trained on the GM12878 cell line and tested on the K562 cell line. Balanced (Pos/Neg = 1) and unbalanced (Pos/Neg = 1/5) gold datasets were prepared, respectively. ROC curve, PR curve, and corresponding AUC were then plotted and calculated. The Pearson correlation between predicted loop intensity with real data was also calculated to show the performance of all methods.

Application of LoopAnchor for dynamic loop detection during macrophage development

CTCF ChIP-seq data were downloaded for both monocyte and macrophage cells[52] (GEO: GSE96800). CTCF ChIP-seq data with replication were processed with ENCODE standard pipeline to retrieve bigWig and peak files. LoopAnchor was used to predict CTCF-anchored loops with peak files for both cell states. The cell-state-specific loops based on Hi-C loops detected by HiCCUPS[86] (GEO: GSE63525) was also downloaded. Because Hi-C loops do not need to be mediated by CTCF and cohesin, loops without CTCF peaks in anchor regions on both sides were filtered out. We mainly focused on the "gained" and "lost" loops from monocyte to macrophage cells by comparing the predicted loop intensity L of the same loop between the two cell states. The fold change of loop intensity for a specific loop between macrophage and

monocyte is defined as L_{macro}/L_{mono} and thresholding at $|L_{macro}/L_{mono}| > 3$ for "gained" and "lost" loops.

Landscape construction of CTCF-anchored loops across human tissue/cell types

We performed a systematic collection of CTCF ChIP-seq datasets from ENCODE,[2] CistromDB,[53] and ChIP-Atlas,[54] and uniformly processed them using the ENCODE TF ChIP-seq processing pipeline (https://github.com/ENCODE-DCC/chip-seq-pipeline2). In brief, the clean reads were mapped to the human genome (GRCh37/hg19) with BWA (v0.7.17)[87] followed by the post-alignment filtering. Finally, the peaks were called using SPP (v1.15).[88] To select the high-quality ChIP-seq biosamples, we first filtered out any outliers with excessively low or high peaks. All B lymphocyte biosamples prefixed with "GM" were then excluded except for GM12878. For each mapped tissue or anatomical cell type, redundant biosamples were removed according to the Jaccard similarity clustering of called peaks, selecting only one biosample in each cluster with the maximal number of peaks. Finally, LoopAnchor and LEM were applied to detect genome-wide CTCF-mediated loops for each biosample.

Definition of tissue/cell-type-shared and -specific loops

All predicted CTCF-anchored loops were classified into four categories based on their shared patterns across different tissue/cell types. Specifically, loops that were shared by no less than 75% of tissue/cell types were classified as tissue/cell-type-shared loops. Loops that were shared by no less than 50% but less than 75% of tissue/cell types were classified as tissue/cell-type-relatively-shared loops. Loops that were shared by no less than 25% but less than 50% of tissue/cell types were classified as tissue/cell-type-relatively-specific loops. The remaining loops which were shared by less than 25% of tissue/cell types were classified as tissue/cell-type-specific loops. The public Hi-C loops detected by Peakachu[55] for 42 tissue/cell types were obtained from the 3D genome browser[89] (http://3dgenome.fsm.northwestern.edu/publications.html). The "pairtopair" command from BEDTools[83] with the parameter "-type both" was then used to compare them with loops predicted by LoopAnchor. The Kruskal-Wallis test, the non-parametric substitute for ANOVA, was employed to compare the distribution of loop intensity scores and the number of tissues across the four categories.

GWAS disease-causal variant enrichment

We collected 12,738 causal variants for 54 blood-related autoimmune diseases from CAUSALdb.[61] To evaluate the genome-wide enrichment of autoimmune disease-causal variants on the CTCF-anchored loops predicted by LoopAnchor, we first randomly sampled the same number of control variants (10,000 times) with matched allele frequency using vSampler.[90] For each normal biosample, the predicted CTCF-anchored loops were flattened into unique anchors, and the "intersect" command from BEDTools[83] was used to examine the colocalization between variants and anchors. The enrichment p value was determined by computing the number of permutations in which the percentage of overlapping variants was higher than that of the real dataset for each biosample. This analysis was only conducted on normal tissues that had more than three biosamples. Here we treated the permutation test p values as "scores" and compared them across tissue groups. Pairwise comparisons between the blood tissue group and non-blood tissue groups were conducted using the one-tailed Mann-Whitney U test. Multiple test correction was applied to calculate the FDR using the Benjamini-Hochberg method.

Non-coding somatic mutation and hotspot enrichment

The genome-wide somatic mutations were downloaded from the ICGC Data Portal (release 28).[62] Candidate non-coding somatic mutations were retained by removing those overlapping with exons and splicing sites and were classified into three categories based on their mutation recurrence ($\in [2,5)$, $\in [5,10)$, ≥ 10). The somatic mutations with only one recurrence were used as background for generating control datasets. We created 10,000 control datasets by randomly sampling the equivalent mutations for each category. For each cancer biosample, the predicted CTCF-anchored loops were flattened into unique anchors, and the number of intersecting hits between somatic mutations in each category and anchors in a particular biosample was calculated. By comparing it with the average count value from the control datasets, the fold change of overlapping mutation percentage was derived. The Mann-Whitney U test was used to compare the difference in fold changes among the three categories for each pair. In addition, candidate somatic mutation hotspots or driver regions were curated from three previous catalogs, including ATACseq-AWG,[63] PCAWG,[64] and CNCDriver.[65] For each cancer biosample, the loops predicted by LoopAnchor with a score ≥ 0.01 were retained and flattened to unique anchors as the "Anchor" datasets, while the same number of non-anchor regions were sampled from remaining loops with the smaller predicted scores as "Control" datasets. By excluding real hotspots and restricting them to non-coding genomic regions, the "shuffle" command from BEDTools[83] was also used to retrieve shuffled hotspots with 100 permutations as "Shuffled" datasets. The Mann-Whitney U test was used to compare the difference in the overlapped percentage of somatic hotspots among these generated datasets in pairwise manner.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2023.100798.

ACKNOWLEDGMENTS

This work is supported by the National Key R&D Program of China (2021YFC2500400 and 2021YFC2500403), the National Natural Science Foundation of China (32070675 to M.J.L. and 32270717 to M.J.L.), and the Natural Science Foundation of Tianjin (19JCJQJC63600 to M.J.L. and 19JCQNJC09000 to X.Y.).

AUTHOR CONTRIBUTIONS

Conceptualization, M.J.L. and H.X.; methodology, H.X. and X.Y.; investigation, H.X., X.Y., X.F., and D.H.; resources, C.W., W.W., X.C., D.H., S.Z., X.D., Z.W., Jianhua Wang, Y.Z., K.Z., H.Y., and N.Z.; writing—original draft, H.X., X.Y., and M.J.L.; writing—review and editing, Y.C., Junwen Wang, K.C., D.P., and P.C.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 12, 2022
Revised: January 25, 2023
Accepted: June 20, 2023
Published: July 12, 2023

REFERENCES

1. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. https://doi.org/10.1038/nature11247.
2. ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710. https://doi.org/10.1038/s41586-020-2493-4.
3. Braccioli, L., and de Wit, E. (2019). CTCF: a Swiss-army knife for genome organization and transcription regulation. Essays Biochem. *63*, 157–165. https://doi.org/10.1042/EBC20180069.
4. Ong, C.T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nat. Rev. Genet. *15*, 234–246. https://doi.org/10.1038/nrg3663.
5. Phillips, J.E., and Corces, V.G. (2009). CTCF: master weaver of the genome. Cell *137*, 1194–1211. https://doi.org/10.1016/j.cell.2009.06.001.
6. Ali, T., Renkawitz, R., and Bartkuhn, M. (2016). Insulators and domains of gene expression. Curr. Opin. Genet. Dev. *37*, 17–26. https://doi.org/10.1016/j.gde.2015.11.009.

7. Ghirlando, R., and Felsenfeld, G. (2016). CTCF: making the right connections. Genes Dev. *30*, 881–891. https://doi.org/10.1101/gad.277863.116.

8. Nichols, M.H., and Corces, V.G. (2015). A CTCF code for 3D genome architecture. Cell *162*, 703–705. https://doi.org/10.1016/j.cell.2015.07.053.

9. Gabriele, M., Brandão, H.B., Grosse-Holz, S., Jha, A., Dailey, G.M., Cattoglio, C., Hsieh, T.H.S., Mirny, L., Zechner, C., and Hansen, A.S. (2022). Dynamics of CTCF- and cohesin-mediated chromatin looping revealed by live-cell imaging. Science *376*, 496–501. eabn6583. https://doi.org/10.1126/science.abn6583.

10. Davidson, I.F., Bauer, B., Goetz, D., Tang, W., Wutz, G., and Peters, J.M. (2019). DNA loop extrusion by human cohesin. Science *366*, 1338–1345. https://doi.org/10.1126/science.aaz3418.

11. Kim, Y., Shi, Z., Zhang, H., Finkelstein, I.J., and Yu, H. (2019). Human cohesin compacts DNA by loop extrusion. Science *366*, 1345–1349. https://doi.org/10.1126/science.aaz4475.

12. Sanborn, A.L., Rao, S.S.P., Huang, S.C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. USA *112*, E6456–E6465. https://doi.org/10.1073/pnas.1518552112.

13. He, C., Wang, X., and Zhang, M.Q. (2014). Nucleosome eviction and multiple co-factor binding predict estrogen-receptor-alpha-associated long-range interactions. Nucleic Acids Res. *42*, 6935–6944. https://doi.org/10.1093/nar/gku327.

14. Beagan, J.A., Duong, M.T., Titus, K.R., Zhou, L., Cao, Z., Ma, J., Lachanski, C.V., Gillis, D.R., and Phillips-Cremins, J.E. (2017). YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. Genome Res. *27*, 1139–1152. https://doi.org/10.1101/gr.215160.116.

15. Hnisz, D., Day, D.S., and Young, R.A. (2016). Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. Cell *167*, 1188–1200. https://doi.org/10.1016/j.cell.2016.10.024.

16. Beagan, J.A., and Phillips-Cremins, J.E. (2020). On the existence and functionality of topologically associating domains. Nat. Genet. *52*, 8–16. https://doi.org/10.1038/s41588-019-0561-1.

17. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380. https://doi.org/10.1038/nature11082.

18. Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385. https://doi.org/10.1038/nature11049.

19. Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schujiers, J., Lee, T.I., Zhao, K., and Young, R.A. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell *159*, 374–387. https://doi.org/10.1016/j.cell.2014.09.030.

20. Islam, Z., Saravanan, B., Walavalkar, K., Farooq, U., Singh, A.K., Radhakrishnan, S., Thakur, J., Pandit, A., Henikoff, S., and Notani, D. (2023). Active enhancers strengthen insulation by RNA-mediated CTCF binding at chromatin domain boundaries. Genome Res. *33*, 1–17. https://doi.org/10.1101/gr.276643.122.

21. Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature *479*, 74–79. https://doi.org/10.1038/nature10442.

22. Vostrov, A.A., and Quitschke, W.W. (1997). The zinc finger protein CTCF binds to the APBbeta domain of the amyloid beta-protein precursor promoter. Evidence for a role in transcriptional activation. J. Biol. Chem. *272*, 33353–33359. https://doi.org/10.1074/jbc.272.52.33353.

23. Filippova, G.N., Fagerlie, S., Klenova, E.M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P.E., Collins, S.J., and Lobanenkov, V.V. (1996). An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. Mol. Cell Biol. *16*, 2802–2813. https://doi.org/10.1128/mcb.16.6.2802.

24. Oh, S., Shao, J., Mitra, J., Xiong, F., D'Antonio, M., Wang, R., Garcia-Bassets, I., Ma, Q., Zhu, X., Lee, J.H., et al. (2021). Enhancer release and retargeting activates disease-susceptibility genes. Nature *595*, 735–740. https://doi.org/10.1038/s41586-021-03577-1.

25. Guo, Y., Monahan, K., Wu, H., Gertz, J., Varley, K.E., Li, W., Myers, R.M., Maniatis, T., and Wu, Q. (2012). CTCF/cohesin-mediated DNA looping is required for protocadherin alpha promoter choice. Proc. Natl. Acad. Sci. USA *109*, 21081–21086. https://doi.org/10.1073/pnas.1219280110.

26. Lobanenkov, V.V., Nicolas, R.H., Adler, V.V., Paterson, H., Klenova, E.M., Polotskaja, A.V., and Goodwin, G.H. (1990). A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. Oncogene *5*, 1743–1753.

27. Nakahashi, H., Kieffer Kwon, K.R., Resch, W., Vian, L., Dose, M., Stavreva, D., Hakim, O., Pruett, N., Nelson, S., Yamane, A., et al. (2013). A genome-wide map of CTCF multivalency redefines the CTCF code. Cell Rep. *3*, 1678–1689. https://doi.org/10.1016/j.celrep.2013.04.024.

28. Huang, H., Zhu, Q., Jussila, A., Han, Y., Bintu, B., Kern, C., Conte, M., Zhang, Y., Bianco, S., Chiariello, A.M., et al. (2021). CTCF mediates dosage- and sequence-context-dependent transcriptional insulation by forming local chromatin domains. Nat. Genet. *53*, 1064–1074. https://doi.org/10.1038/s41588-021-00863-6.

29. Ribeiro-Dos-Santos, A.M., Hogan, M.S., Luther, R.D., Brosh, R., and Maurano, M.T. (2022). Genomic context sensitivity of insulator function. Genome Res. *32*, 425–436. https://doi.org/10.1101/gr.276449.121.

30. Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. Cell *162*, 900–910. https://doi.org/10.1016/j.cell.2015.07.038.

31. Lv, H., Dao, F.Y., Zulfiqar, H., Su, W., Ding, H., Liu, L., and Lin, H. (2021). A sequence-based deep learning approach to predict CTCF-mediated chromatin loop. Briefings Bioinf. *22*, bbab031. https://doi.org/10.1093/bib/bbab031.

32. Kai, Y., Andricovich, J., Zeng, Z., Zhu, J., Tzatsos, A., and Peng, W. (2018). Predicting CTCF-mediated chromatin interactions by integrating genomic and epigenomic features. Nat. Commun. *9*, 4221. https://doi.org/10.1038/s41467-018-06664-6.

33. Wang, W., Gao, L., Ye, Y., and Gao, Y. (2021). CCIP: Predicting CTCF-mediated chromatin loops with transitivity. Bioinformatics *37*, 4635–4642. https://doi.org/10.1093/bioinformatics/btab534.

34. Kuang, S., and Wang, L. (2021). Deep learning of sequence patterns for CCCTC-binding factor-mediated chromatin loop formation. J. Comput. Biol. *28*, 133–145. https://doi.org/10.1089/cmb.2020.0225.

35. Zhang, R., Wang, Y., Yang, Y., Zhang, Y., and Ma, J. (2018). Predicting CTCF-mediated chromatin loops using CTCF-MP. Bioinformatics *34*, i133–i141. https://doi.org/10.1093/bioinformatics/bty248.

36. Matthews, B.J., and Waxman, D.J. (2018). Computational prediction of CTCF/cohesin-based intra-TAD loops that insulate chromatin contacts and gene expression in mouse liver. Elife *7*, e34077. https://doi.org/10.7554/eLife.34077.

37. Oti, M., Falck, J., Huynen, M.A., and Zhou, H. (2016). CTCF-mediated chromatin loops enclose inducible gene regulatory domains. BMC Genom. *17*, 252. https://doi.org/10.1186/s12864-016-2516-6.

38. Ibn-Salem, J., and Andrade-Navarro, M.A. (2019). 7C: Computational chromosome conformation capture by correlation of ChIP-seq at CTCF motifs. BMC Genom. *20*, 777. https://doi.org/10.1186/s12864-019-6088-0.

39. Cao, F., Zhang, Y., Cai, Y., Animesh, S., Zhang, Y., Akincilar, S.C., Loh, Y.P., Li, X., Chng, W.J., Tergaonkar, V., et al. (2021). Chromatin interaction neural network (ChINN): a machine learning-based method for predicting

chromatin interactions from DNA sequences. Genome Biol. *22*, 226. https://doi.org/10.1186/s13059-021-02453-5.

40. Xi, W., and Beer, M.A. (2021). Loop competition and extrusion model predicts CTCF interaction specificity. Nat. Commun. *12*, 1046. https://doi.org/10.1038/s41467-021-21368-0.
41. Lee, R., Kang, M.K., Kim, Y.J., Yang, B., Shim, H., Kim, S., Kim, K., Yang, C.M., Min, B.G., Jung, W.J., et al. (2022). CTCF-mediated chromatin looping provides a topological framework for the formation of phase-separated transcriptional condensates. Nucleic Acids Res. *50*, 207–226. https://doi.org/10.1093/nar/gkab1242.
42. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. *47*, D886–D894. https://doi.org/10.1093/nar/gky1016.
43. Clarkson, C.T., Deeks, E.A., Samarista, R., Mamayusupova, H., Zhurkin, V.B., and Teif, V.B. (2019). CTCF-dependent chromatin boundaries formed by asymmetric nucleosome arrays with decreased linker length. Nucleic Acids Res. *47*, 11181–11196. https://doi.org/10.1093/nar/gkz908.
44. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. *2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9.
45. Fu, Y., Sinha, M., Peterson, C.L., and Weng, Z. (2008). The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet. *4*, e1000138. https://doi.org/10.1371/journal.pgen.1000138.
46. Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.Y., Cui, K., and Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res. *19*, 24–32. https://doi.org/10.1101/gr.082800.108.
47. Luan, J., Xiang, G., Gómez-García, P.A., Tome, J.M., Zhang, Z., Vermunt, M.W., Zhang, H., Huang, A., Keller, C.A., Giardine, B.M., et al. (2021). Distinct properties and functions of CTCF revealed by a rapidly inducible degron system. Cell Rep. *34*, 108783. https://doi.org/10.1016/j.celrep.2021.108783.
48. Rhee, H.S., and Pugh, B.F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell *147*, 1408–1419. https://doi.org/10.1016/j.cell.2011.11.013.
49. Yin, M., Wang, J., Wang, M., Li, X., Zhang, M., Wu, Q., and Wang, Y. (2017). Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites. Cell Res. *27*, 1365–1377. https://doi.org/10.1038/cr.2017.131.
50. Li, Y., Haarhuis, J.H.I., Sedeño Cacciatore, Á., Oldenkamp, R., van Ruiten, M.S., Willems, L., Teunissen, H., Muir, K.W., de Wit, E., Rowland, B.D., and Panne, D. (2020). The structural basis for cohesin-CTCF-anchored loops. Nature *578*, 472–476. https://doi.org/10.1038/s41586-019-1910-z.
51. Pugacheva, E.M., Kubo, N., Loukinov, D., Tajmul, M., Kang, S., Kovalchuk, A.L., Strunnikov, A.V., Zentner, G.E., Ren, B., and Lobanenkov, V.V. (2020). CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. Proc. Natl. Acad. Sci. USA *117*, 2020–2031. https://doi.org/10.1073/pnas.1911708117.
52. Phanstiel, D.H., Van Bortle, K., Spacek, D., Hess, G.T., Shamim, M.S., Machol, I., Love, M.I., Aiden, E.L., Bassik, M.C., and Snyder, M.P. (2017). Static and dynamic DNA loops form AP-1-Bound activation hubs during macrophage development. Mol. Cell *67*, 1037–1048.e6. https://doi.org/10.1016/j.molcel.2017.08.006.
53. Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., Chen, C.H., Brown, M., Zhang, X., Meyer, C.A., and Liu, X.S. (2019). Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Res. *47*, D729–D735. https://doi.org/10.1093/nar/gky1094.
54. Oki, S., Ohta, T., Shioi, G., Hatanaka, H., Ogasawara, O., Okuda, Y., Kawaji, H., Nakaki, R., Sese, J., and Meno, C. (2018). ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. EMBO Rep. *19*, e46255. https://doi.org/10.15252/embr.201846255.
55. Salameh, T.J., Wang, X., Song, F., Zhang, B., Wright, S.M., Khunsriraksakul, C., Ruan, Y., and Yue, F. (2020). A supervised learning framework for chromatin loop detection in genome-wide contact maps. Nat. Commun. *11*, 3428. https://doi.org/10.1038/s41467-020-17239-9.
56. Katainen, R., Dave, K., Pitkänen, E., Palin, K., Kivioja, T., Välimäki, N., Gylfe, A.E., Ristolainen, H., Hänninen, U.A., Cajuso, T., et al. (2015). CTCF/cohesin-binding sites are frequently mutated in cancer. Nat. Genet. *47*, 818–821. https://doi.org/10.1038/ng.3335.
57. Fang, C., Wang, Z., Han, C., Safgren, S.L., Helmin, K.A., Adelman, E.R., Serafin, V., Basso, G., Eagen, K.P., Gaspar-Maia, A., et al. (2020). Cancer-specific CTCF binding facilitates oncogenic transcriptional dysregulation. Genome Biol. *21*, 247. https://doi.org/10.1186/s13059-020-02152-7.
58. Guo, Y.A., Chang, M.M., Huang, W., Ooi, W.F., Xing, M., Tan, P., and Skanderup, A.J. (2018). Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. Nat. Commun. *9*, 1520. https://doi.org/10.1038/s41467-018-03828-2.
59. Kentepozidou, E., Aitken, S.J., Feig, C., Stefflova, K., Ibarra-Soria, X., Odom, D.T., Roller, M., and Flicek, P. (2020). Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. Genome Biol. *21*, 5. https://doi.org/10.1186/s13059-019-1894-x.
60. Sadowski, M., Kraft, A., Szalaj, P., Wlasnowolski, M., Tang, Z., Ruan, Y., and Plewczynski, D. (2019). Spatial chromatin architecture alteration by structural variations in human genomes at the population scale. Genome Biol. *20*, 148. https://doi.org/10.1186/s13059-019-1728-x.
61. Wang, J., Huang, D., Zhou, Y., Yao, H., Liu, H., Zhai, S., Wu, C., Zheng, Z., Zhao, K., Wang, Z., et al. (2020). CAUSALdb: a database for disease/trait causal variants identified using summary statistics of genome-wide association studies. Nucleic Acids Res. *48*, D807–D816. https://doi.org/10.1093/nar/gkz1026.
62. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93. https://doi.org/10.1038/s41586-020-1969-6.
63. Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, S.W., et al. (2018). The chromatin accessibility landscape of primary human cancers. Science *362*, eaav1898. https://doi.org/10.1126/science.aav1898.
64. Rheinbay, E., Nielsen, M.M., Abascal, F., Wala, J.A., Shapira, O., Tiao, G., Hornshøj, H., Hess, J.M., Juul, R.I., Lin, Z., et al. (2020). Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature *578*, 102–111. https://doi.org/10.1038/s41586-020-1965-x.
65. Liu, E.M., Martinez-Fundichely, A., Diaz, B.J., Aronson, B., Cuykendall, T., MacKay, M., Dhingra, P., Wong, E.W.P., Chi, P., Apostolou, E., et al. (2019). Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole Genomes. Cell Syst. *8*, 446–455.e8. https://doi.org/10.1016/j.cels.2019.04.001.
66. Deng, Y., Tang, L., Zhou, X., Wang, W., and Li, M. (2022). Integrating extrusion complex-associated pattern to predict cell type-specific long-range chromatin loops. iScience *25*, 105687. https://doi.org/10.1016/j.isci.2022.105687.
67. Davidson, I.F., and Peters, J.M. (2021). Genome folding through loop extrusion by SMC complexes. Nat. Rev. Mol. Cell Biol. *22*, 445–464. https://doi.org/10.1038/s41580-021-00349-7.
68. Yi, X., Zheng, Z., Xu, H., Zhou, Y., Huang, D., Wang, J., Feng, X., Zhao, K., Fan, X., Zhang, S., et al. (2021). Interrogating cell type-specific cooperation of transcriptional regulators in 3D chromatin. iScience *24*, 103468. https://doi.org/10.1016/j.isci.2021.103468.
69. Weintraub, A.S., Li, C.H., Zamudio, A.V., Sigova, A.A., Hannett, N.M., Day, D.S., Abraham, B.J., Cohen, M.A., Nabet, B., Buckley, D.L., et al. (2017). YY1 Is a structural regulator of enhancer-promoter Loops. Cell *171*, 1573–1588.e28. https://doi.org/10.1016/j.cell.2017.11.008.
70. Bailey, S.D., Zhang, X., Desai, K., Aid, M., Corradin, O., Cowper-Sal Lari, R., Akhtar-Zaidi, B., Scacheri, P.C., Haibe-Kains, B., and Lupien, M. (2015). ZNF143 provides sequence specificity to secure chromatin

interactions at gene promoters. Nat. Commun. *2*, 6186. https://doi.org/10.1038/ncomms7186.

71. Ortabozkoyun, H., Huang, P.Y., Cho, H., Narendra, V., LeRoy, G., Gonzalez-Buendia, E., Skok, J.A., Tsirigos, A., Mazzoni, E.O., and Reinberg, D. (2022). CRISPR and biochemical screens identify MAZ as a cofactor in CTCF-mediated insulation at Hox clusters. Nat. Genet. *54*, 202–212. https://doi.org/10.1038/s41588-021-01008-5.
72. Hu, G., Dong, X., Gong, S., Song, Y., Hutchins, A.P., and Yao, H. (2020). Systematic screening of CTCF binding partners identifies that BHLHE40 regulates CTCF genome-wide distribution and long-range chromatin interactions. Nucleic Acids Res. *48*, 9606–9620. https://doi.org/10.1093/nar/gkaa705.
73. Debruyne, D.N., Dries, R., Sengupta, S., Seruggia, D., Gao, Y., Sharma, B., Huang, H., Moreau, L., McLane, M., Day, D.S., et al. (2019). BORIS promotes chromatin regulatory interactions in treatment-resistant cancer cells. Nature *572*, 676–680. https://doi.org/10.1038/s41586-019-1472-0.
74. Wang, R., Chen, F., Chen, Q., Wan, X., Shi, M., Chen, A.K., Ma, Z., Li, G., Wang, M., Ying, Y., et al. (2022). MyoD is a 3D genome structure organizer for muscle cell identity. Nat. Commun. *13*, 205. https://doi.org/10.1038/s41467-021-27865-6.
75. Wang, Z., Liang, Q., Qian, X., Hu, B., Zheng, Z., Wang, J., Hu, Y., Bao, Z., Zhao, K., Zhou, Y., et al. (2023). An autoimmune pleiotropic SNP modulates IRF5 alternative promoter usage through ZBTB3-mediated chromatin looping. Nat. Commun. *14*, 1208. https://doi.org/10.1038/s41467-023-36897-z.
76. Xuhang01. (2023). Xuhang01/LoopAnchor:v.1.0.0(V.1.0.0). Zenodo. https://doi.org/10.5281/zenodo.8008481.
77. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49. https://doi.org/10.1038/nature09906.
78. Grubert, F., Srivas, R., Spacek, D.V., Kasowski, M., Ruiz-Velasco, M., Sinnott-Armstrong, N., Greenside, P., Narasimha, A., Liu, Q., Geller, B., et al. (2020). Landscape of cohesin-mediated chromatin loops in the human genome. Nature *583*, 737–743. https://doi.org/10.1038/s41586-020-2151-x.
79. Li, G., Chen, Y., Snyder, M.P., and Zhang, M.Q. (2017). ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis. Nucleic Acids Res. *45*, e4. https://doi.org/10.1093/nar/gkw809.
80. Fornes, O., Castro-Mondragon, J.A., Khan, A., van der Lee, R., Zhang, X., Richmond, P.A., Modi, B.P., Correard, S., Gheorghe, M., Baranašić, D., et al. (2020). JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *48*, D87–D92. https://doi.org/10.1093/nar/gkz1001.
81. Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018. https://doi.org/10.1093/bioinformatics/btr064.
82. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315. https://doi.org/10.1038/ng.2892.
83. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.
84. Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., et al. (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol. *16*, 22. https://doi.org/10.1186/s13059-014-0560-6.
85. Lundberg, S.M., Nair, B., Vavilala, M.S., Horibe, M., Eisses, M.J., Adams, T., Liston, D.E., Low, D.K.W., Newman, S.F., Kim, J., and Lee, S.I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat. Biomed. Eng. *2*, 749–760. https://doi.org/10.1038/s41551-018-0304-0.
86. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021.
87. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
88. Kharchenko, P.V., Tolstorukov, M.Y., and Park, P.J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat. Biotechnol. *26*, 1351–1359. https://doi.org/10.1038/nbt.1508.
89. Wang, Y., Song, F., Zhang, B., Zhang, L., Xu, J., Kuang, D., Li, D., Choudhary, M.N.K., Li, Y., Hu, M., et al. (2018). The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol. *19*, 151. https://doi.org/10.1186/s13059-018-1519-9.
90. Huang, D., Wang, Z., Zhou, Y., Liang, Q., Sham, P.C., Yao, H., and Li, M.J. (2021). vSampler: fast and annotation-based matched variant sampling tool. Bioinformatics *37*, 1915–1917. https://doi.org/10.1093/bioinformatics/btaa883.

Patterns, Volume *4*

Supplemental information

Inferring CTCF-binding patterns and anchored loops across human tissues and cell types

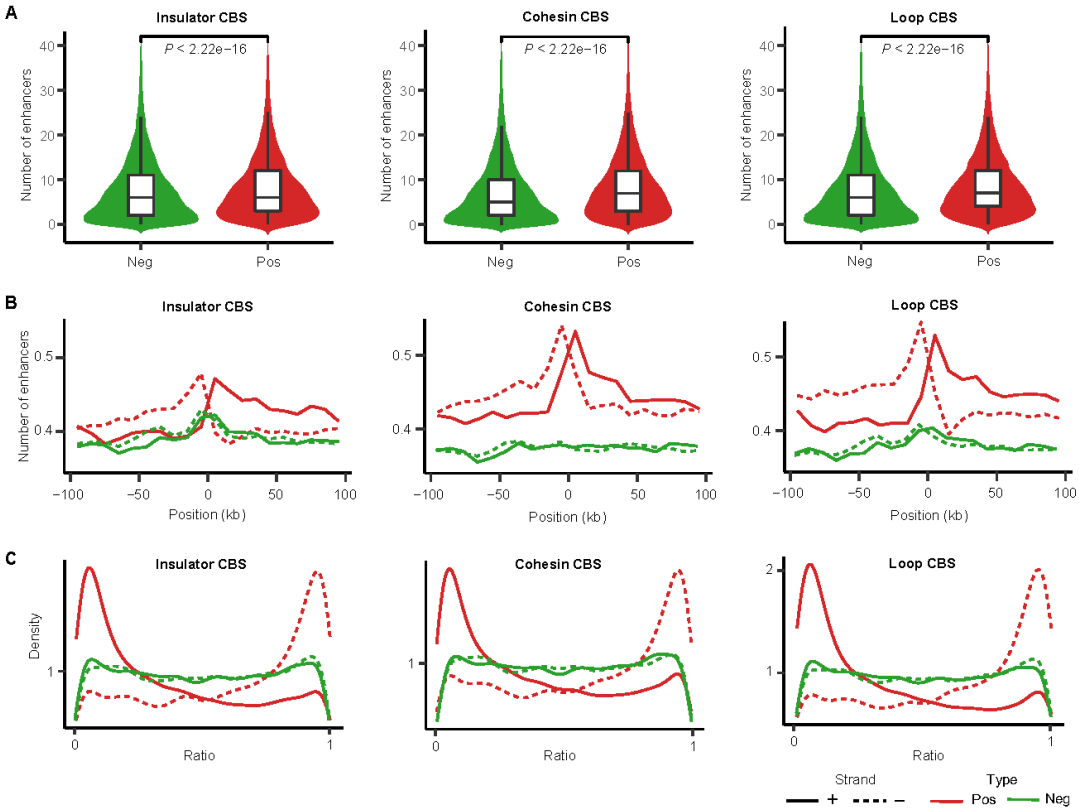
Hang Xu, Xianfu Yi, Xutong Fan, Chengyue Wu, Wei Wang, Xinlei Chu, Shijie Zhang, Xiaobao Dong, Zhao Wang, Jianhua Wang, Yao Zhou, Ke Zhao, Hongcheng Yao, Nan Zheng, Junwen Wang, Yupeng Chen, Dariusz Plewczynski, Pak Chung Sham, Kexin Chen, Dandan Huang, and Mulin Jun Li

Supplementary Information

This file includes:

Supplementary Figures S1 to S6

Supplementary Figures

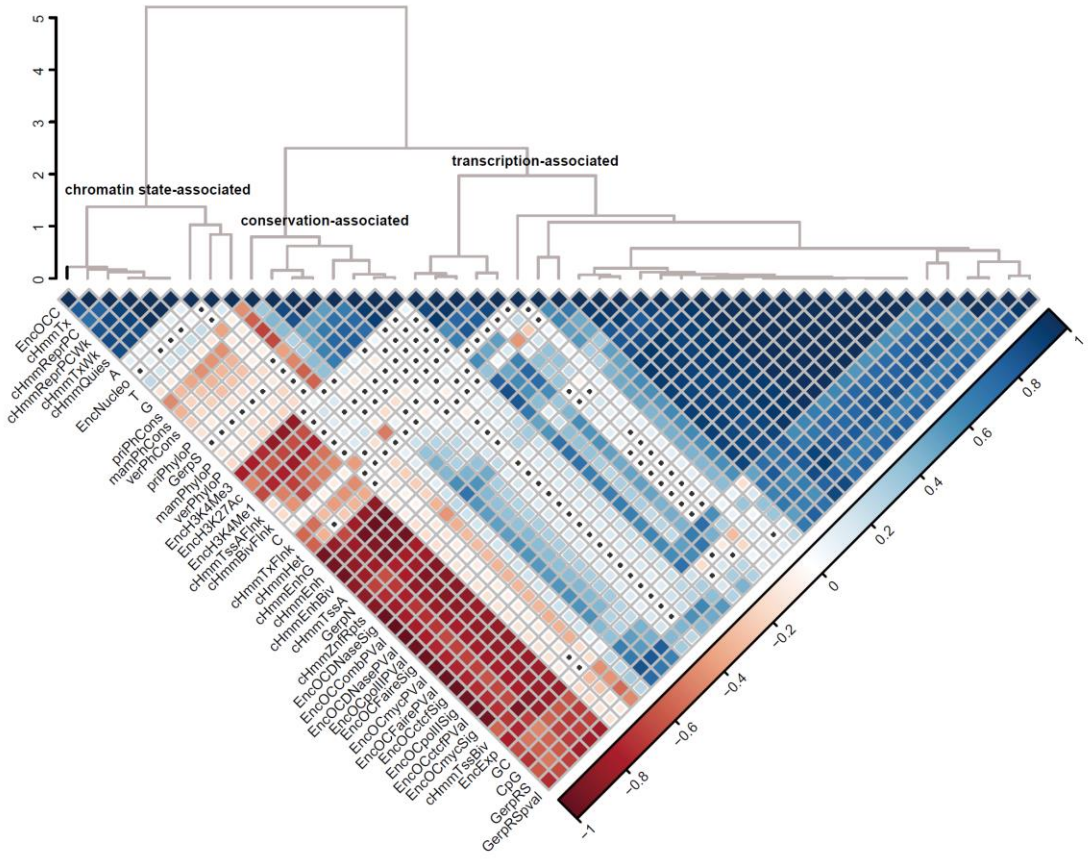


Supplementary Figure S1. Evaluations of positive and negative CTCF-mediated CREs predicted by different types of DeepAnchor CBS models on enhancer and TAD relevance, related to Figure 1.

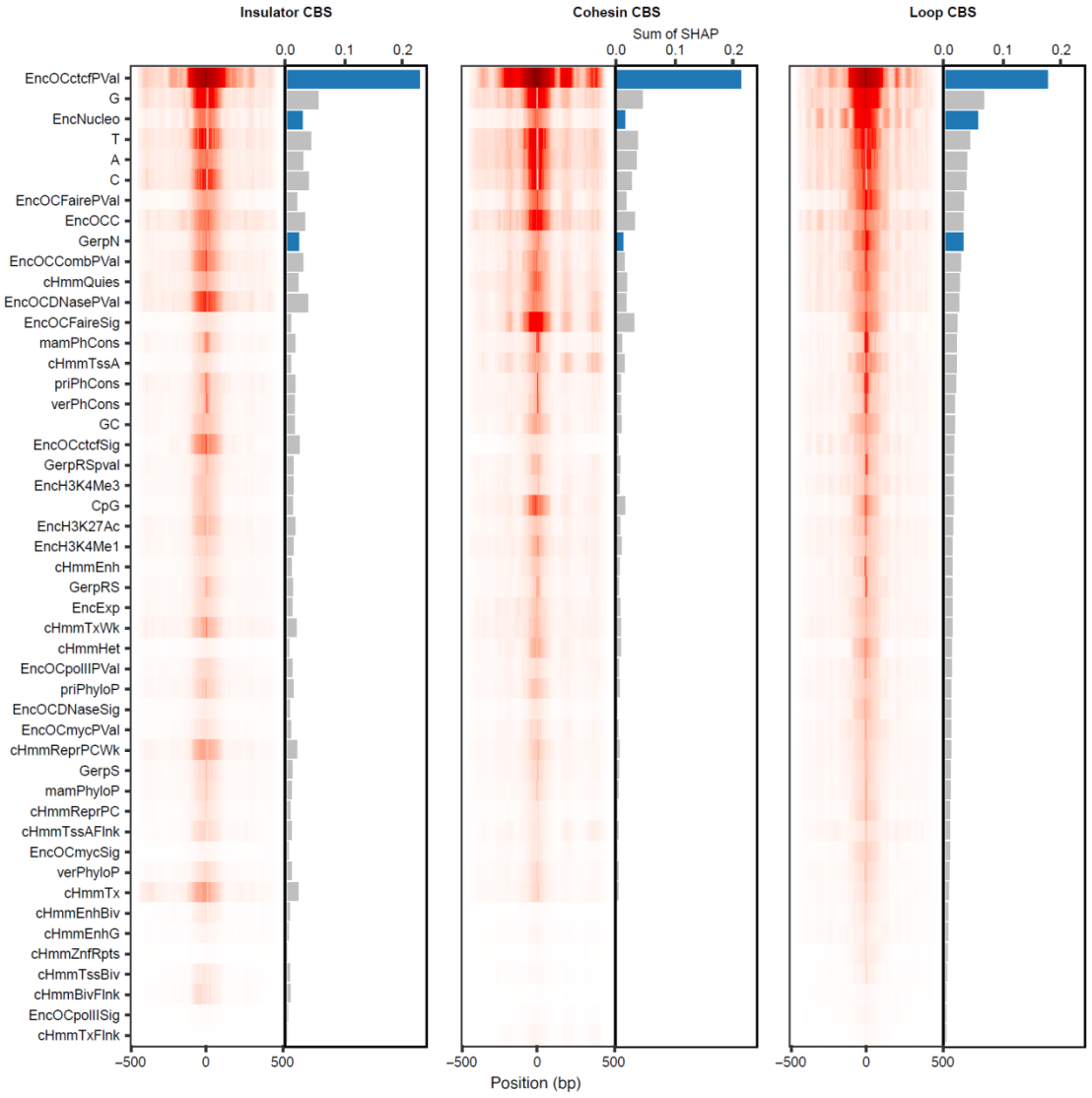
A. Comparison of number of enhancers around CBSs between predicted Pos and Neg CTCF-mediated CREs among three types of DeepAnchor CBS models. The Mann-Whitney U test was used to test the significance.

B. Strand-oriented asymmetric pattern of enhancer enrichment at predicted Pos and Neg CTCF-mediated CREs among three types of DeepAnchor CBS models.

C. Position and strand preference of TAD boundary enrichment by measuring the distance between each CBS and the 5’ end of TAD it belongs at predicted Pos and Neg CTCF-mediated CREs among three types of DeepAnchor CBS models.



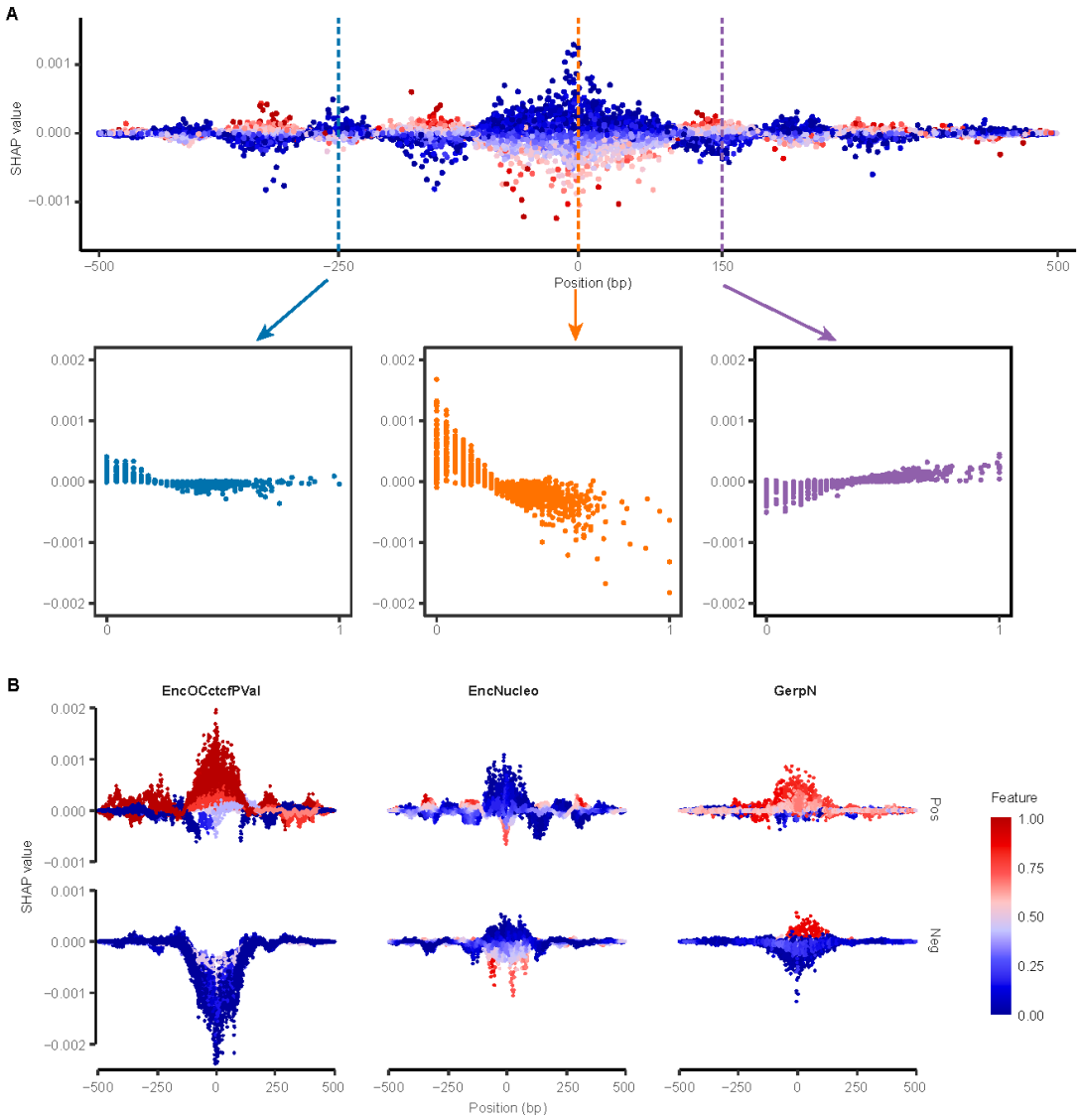
Supplementary Figure S2. Hierarchical clustering of 44 genomic/epigenomic features on training dataset, related to Figure 2.



Supplementary Figure S3. Feature importance analysis of all used features at base-wise level among three types of CBSs, related to Figure 2.

Heatmap: Average absolute feature Shapley values of each position at ±500 bp of CBS.

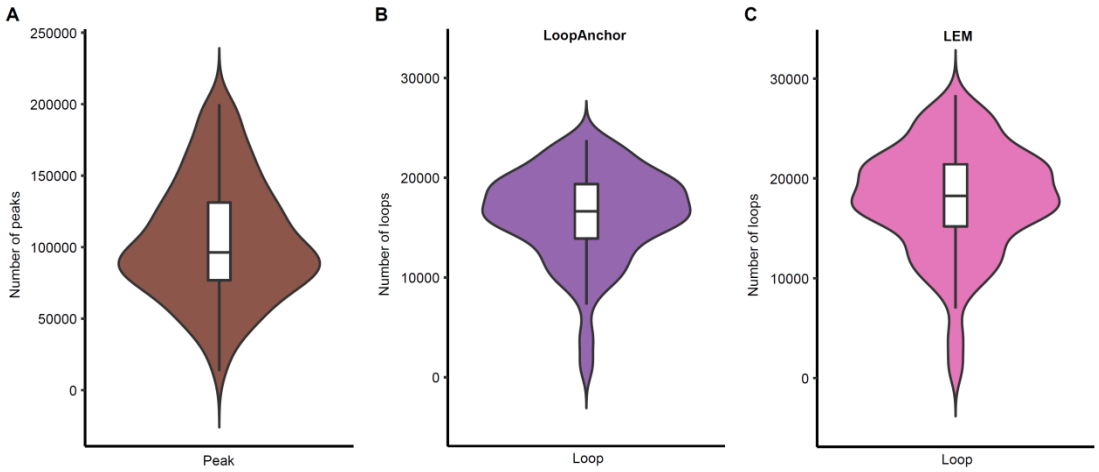
Bar plot: Summation of absolute feature Shapley values across ±500 bp of CBS.



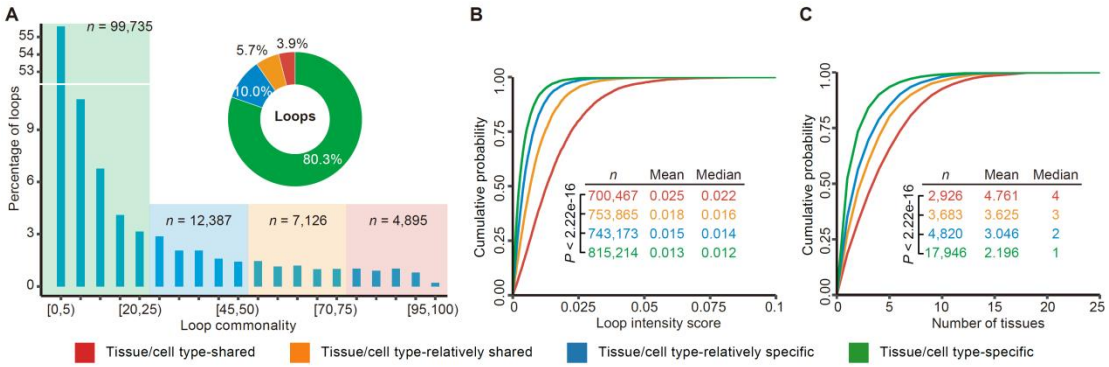
Supplementary Figure S4. Base-wise analysis of feature contributions for loop CBS model in different angles, related to Figure 2.

A. Base-wise anti-correlated pattern between nucleosome positioning level and Shapley value.

B. Base-wise Shapley value distribution for three representative features, Up: positive loop CBSs, Down: Negative loop CBSs.



Supplementary Figure S5. The number distribution of CTCF ChIP-seq peaks from different biosamples (A), and loops predicted by LoopAnchor (B) and LEM (C) methods, related to Figure 4.



Supplementary Figure S6. Predicted CTCF-anchored loops across 32 human tissues and 20 cell types using LEM, related to Figure 4.

A. Classification of CTCF-anchored loops according to their shared patterns. All predicted loops were classified into four categories, including tissue/cell type-specific, tissue/cell type-relatively specific, tissue/cell type-relatively shared, and tissue/cell type-shared.

B. Comparison of loop intensity score for loops in four categories. The cumulative probability was calculated and the Kruskal-Wallis test was used to test the significance.

C. Validation of tissue distribution for loops across four categories using predicted chromatin loops by Peakachu for 42 human tissue/cell types. The cumulative probability was calculated and the Kruskal-Wallis test was used to test the significance.