Inferring CTCF insulators and anchored loops across human tissues and cell types

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

[1]Department of Epidemiology and Biostatistics, Tianjin Key Laboratory of Molecular Cancer Epidemiology, The Province and Ministry Co-sponsored Collaborative Innovation Center for Medical Epigenetics, National Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China.

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

[3]School of Biomedical Engineering, Tianjin Medical University, Tianjin, China.

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

[5]Department of Genetics, School of Basic Medical Sciences, Tianjin Medical University, Tianjin, China.

[6]Institute of Aging, Key Laboratory of Alzheimer's Disease of Zhejiang Province, School of Mental Health and The Affiliated Kangning Hospital, Wenzhou Medical University, Wenzhou, China.

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

[8]Department of Health Sciences Research and Center for Individualized Medicine, Mayo Clinic, Scottsdale, Arizona 85259, USA.

[9]Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland.

[10]Department of Bioinformatics, School of Basic Medical Sciences, Tianjin Medical University, Tianjin, China.

[#]The authors contributed equally to this work

*Correspondence: mulinli@connect.hku.hk (M.J.L.)

Abstract

How CTCF recognizes insulators to exert chromosome barrier or enhancer blocking effects remains to be interrogated. Despite many computational tools were developed to predict CTCF-mediated loops qualitatively or quantitatively, few could specially evaluate the insulative potential of DNA sequence at CTCF binding sites (CBSs) and how it affects chromatin loop formation. Here, we developed a deep learning model, DeepAnchor, to precisely predict the insulative potential of CBS. By incorporating base-wise genomic/epigenomic features, we revealed distinct chromatin and sequence features for CTCF-mediated insulation at a high resolution, such as two sequence motifs flanking the core CTCF motif only at positive CTCF insulators. Besides, we leveraged the predicted insulator score to optimize the loop extrusion model and achieved the best performance in

predicting CTCF-anchored loops. We established a compendium of context-specific CTCF-anchored loops across 52 human tissue/cell types and found that genomic disruption of CTCF-anchored loops may represent a general causal mechanism of disease pathogenesis. These computational models, together with the established resource, could facilitate the mechanistic research on how the CTCF-mediated insulation shapes context-specific gene regulation in cell development and disease progression.

Introduction

Variable interactions between enhancer and target gene have been systematically profiled in explaining tissue/cell type-specific transcriptional regulation (Consortium 2012; Consortium et al. 2020), but the mechanisms underlying the precise wiring between genes and the enhancers from varied distances remain unclear. As the main insulator-related transcription factor (TF) discovered in vertebrates, CTCF assists cohesin to form chromatin loops, which is believed to be the fundamental biophysical basis for distal gene regulation (Phillips and Corces 2009; Ong and Corces 2014; Nichols and Corces 2015; Ali et al. 2016; Ghirlando and Felsenfeld 2016; Braccioli and de Wit 2019; Gabriele et al. 2022). According to the loop extrusion model (Davidson et al. 2019; Kim et al. 2019), cohesin dimers load on DNA at NIPBL binding sites and slide along with DNA until reaching the anchors, which are usually bound by specific cofactors, such as CTCF, YY1, and ERα (He et al. 2014; Sanborn et al. 2015; Beagan et al. 2017). Chromatin loops mediated by YY1 and ERα are found to form enhancer-promoter interactions, while CTCF-anchored loops generally enclose the formers (Beagan et al. 2017), assembling genome units that are named insulated neighborhoods (INs). Enhancers tend to regulate genes within the same IN while the regulation across different INs is more likely to be prevented (Dowen et al. 2014). Therefore, characterizing CTCF binding at insulators (CTCF insulators) will greatly facilitate the analyses of fine-grained 3D genome regulation.

CTCF insulators play a critical role in controlling tissue/cell type-specific gene expression by maintaining chromatin domain boundaries or blocking enhancer activities, however, CTCF has been revealed as a versatile TF that plays several other roles in different scenarios, including gene activation, transcriptional repression, pausing, and alternative splicing (Lobanenkov et al. 1990; Filippova et al. 1996; Vostrov and Quitschke 1997; Shukla et al. 2011; Guo et al. 2012; Oh et al. 2021). Therefore, it is necessary to distinguish insulator-related CTCF binding sites (CBSs) from those with other functions. There are a lot of works that analyze the functional heterogeneity of CTCF in the nucleus, which revealed that the binding motif of CTCF, formed by an asymmetric core motif and several subunits, is the key to understanding the manner of CTCF at insulator (Nakahashi et al. 2013; Huang et al. 2021b). Studies have uncovered that insulation potency heavily depends on the number of CBSs in tandem and an upstream flanking sequence at the core CTCF motif (Huang et al. 2021b). Besides, in the formation of CTCF/cohesin chromatin loops, the CBSs at loop anchors tend to be arranged in convergent orientation (Guo et al. 2015). However, these studies are either unable to count the feature patterns of individual CBSs or limited to given genomic loci. Besides, a large number of computational tools were

developed to qualitatively or quantitatively predict CTCF-mediated loops (Oti et al. 2016; Kai et al. 2018; Matthews and Waxman 2018; Zhang et al. 2018; Ibn-Salem and Andrade-Navarro 2019; Cao et al. 2021; Kuang and Wang 2021; Lv et al. 2021; Wang et al. 2021; Xi and Beer 2021), but few could specially evaluate the insulative potential of DNA sequence at CBS and how it affects the loop formation. Moreover, it remains elusive to see which factors or combinations of features determine the specificity of CTCF insulators and CTCF-anchored loops across the whole genome.

Here, we develop an interpretable deep learning model, DeepAnchor, to ask what types of genomic/epigenomic features determine a CBS to be insulator-associated. By leveraging large-scale base-wise genomic and epigenomic features, we uncovered several critical patterns for CTCF-mediated insulation at a high resolution. We embedded the predicted insulator score to the previous loop extrusion model (Xi and Beer 2021), and found the optimized model LoopAnchor outperforms the existing methods for predicting CTCF-anchored loops. Furthermore, for the first time, we established a landscape of tissue/cell type-specific CTCF-anchored loops across 52 human tissue/cell types and applied them to interpret disease-causal variants. This method together with the compiled resources will facilitate the mechanistic research on the dynamic regulation of 3D chromatin during cell development and disease progression.

Results

DeepAnchor enables high-confidence CTCF insulator detection

The key feature of DeepAnchor is the implementation of a deep learning model that leverages base-wise features for CTCF insulator detection. Previous studies usually analysed the sequence and chromatin status of CBS in particular tissue/cell types with a one-dimensional feature vector, in which regional feature values are averaged for each CBS regardless of heterogeneous signals across the locus (Kai et al. 2018; Zhang et al. 2018). However, a true CTCF insulator differs from other CBSs in fine-grained architecture (Ong and Corces 2014; Huang et al. 2021b), which raises new requirements for feature characterization before modelling. One straightforward solution is to profile an insulator locus using high-resolution features surrounding the insulator locus. In our implementation, for each CBS, 44 quantitative base-wise genomic/epigenomic features within the ±500bp region were obtained from CADD annotation (Rentzsch et al. 2019), and DNA sequence within ±500bp of the CBS was also extracted and represented by one-hot encoding (Table S1). Thus, by concatenating large-scale annotations with converted DNA features at the base-wise level, a feature matrix with a size of 1000×48 was generated for each CBS (Figure 1A top left).

To prepare high-quality ground truth for CTCF insulator learning, we required that each selected site should contain 1) putative CBS by motif scanning; 2) observed CBS by CTCF ChIP-seq at given tissue/cell type; 3) cohesin loop anchor by ChIA-PET at given tissue/cell type (Figure 1A top right). Positive CTCF insulators are derived from the overlapped

cohesin loop anchors, while negative controls are not. To avoid ambiguous results, multiple CBSs shared the same CTCF ChIP-seq peak(s) or the same cohesin loop anchor(s) are excluded (see Methods). Including base-wise features of the region that spans insulators will largely expand the feature volume and scale, which will introduce a flood of information that is usually beyond the ability of classical machine learning algorithms. For example, regions with 1000bp can incorporate 1000×48=48,000 features if a base-wise value is applied. To address this challenge, we applied a deep convolutional neural network (CNN) to extract important patterns from the feature matrixes, which greatly reduces the calculation burden and could increase prediction performance. After training a classifier on a positive and negative dataset, the DeepAnchor framework returns the probability of a CBS region belonging to a true CTCF insulator, which is called the DeepAnchor score in the downstream analyses (Figure 1A bottom, see Methods).

We first trained a DeepAnchor model using a high-quality RAD21 ChIA-PET data on the GM12878 cell line and tested the repeatability of DeepAnchor by cross-validation (see Methods). To avoid the potential influence of chromosome-specific confounding factors, we divided the train-test-valid set by splitting CBSs into subsets each on different chromosomes. That is, we selected chr1-16 as the training set, chr17-18 as the validation set, and others as the test set. This DeepAnchor model received a mean area under the ROC curve (AUC) of 0.95 via five-fold cross-validation (Figure 1B). We set a cut-off to define positive and negative CTCF insulators by finding the points on the ROC curve with the largest Youden's J statistic. The optimal cut-point was obtained when the DeepAnchor score is close to 0.5, where 94% positive and 87% negative samples in training datasets are correctly assigned (Figure 1C). To demonstrate that the DeepAnchor scores derived from different samples are comparable, we trained the DeepAnchor model on three cell lines (including GM12878, K562, and H1-hESC) and performed cross-cell type-validation. Notably, the DeepAnchor models exhibited consistently high cross-sample prediction performance on all testing pairs (Figure 1D) and the DeepAnchor scores are highly correlated among the three models (Pearson's r = 0.921-0.952) (Figure 1E), suggesting CTCF insulators share conservative features in different contexts and the feasibility to predict genome-wide CTCF insulators at the organismal level.

With the GM12878 DeepAnchor model and the optimal cut-off, we generated 23,695 positive and 86,364 negative CTCF insulators genome-widely (Figure 1F). Using the FANTOM5 enhancer dataset, we found that positive CTCF insulators colocalize with more enhancers within ±100kb regions than negative CTCF insulators on average (P-value<2.22e-16, Mann Whitney U Test) (Figure 1G). Besides, by counting the number of enhancers in every 10kb bin within ±100kb of CTCF insulators, we noted that there is a strand-oriented asymmetric pattern for positive CTCF insulators, but not negative ones (Figure 1H). Consistent with previous findings (Clarkson et al. 2019), the density of CTCF insulators around the topologically associated domain (TAD) boundaries also shows strand bias, in which positive CTCF insulators at plus-strand are usually enriched at the left TAD boundary, while those at minus-strand are aggregated at right TAD boundary. However, no such patterns were observed for negative CTCF insulators (Figure 1I). These results

demonstrated that our DeepAnchor model can accurately capture high-confidence CTCF insulators across the whole genome.

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

Although DeepAnchor leverages a deep learning structure to extract high-level configurations from large-scale features, it also raises challenges in analysing feature importance and its biological insight. Base-wise representation of various genomic/epigenomic features in the DeepAnchor model makes it possible to depict a comprehensive picture of the patterns of CTCF insulation. To visualize underlying relationships among these features, we first clustered the feature scores upon the training dataset and found these features could be generally partitioned into three major subsets, including transcription-associated, conservation-associated, and chromatin state-associated feature sets (Figure S1). We used SHapley Additive exPlanations (SHAP) (Lundberg et al. 2020), a game-theoretic approach to explaining the output of DeepAnchor. SHAP returns the Shapley value matrix with the same alignment as the input feature matrix, which shows the contribution of each feature at each position. To inspect the overall contribution of each feature, we calculated the mean Shapley values at all positions and then summarized the mean values (Figure 2A). Overall, features like CTCF binding evidence at open chromatin (EncOCctcfPval) and nucleosome positioning evidence (EncNucleo) greatly contribute to the model, suggesting the CTCF footprinting pattern among nucleosomes could determine the formulation of CTCF insulation. Shapley values of G/T/A/C are also top-ranked, which implies the importance of DNA sequence. As expected, open chromatin features (e.g., EncOCFairePVal and EncOCC) display higher importance than other features. Besides, a conservation feature, the neutral evolution score defined by GERP++ (GerpN), is at a high rank (Figure 2A). Taken together, these results indicate that some informative features may capture the unique and important patterns that can distinguish CTCF insulators from other types of CTCF binding events.

To explore the base-wise schema of CTCF insulation, for several important features, we investigated the distribution of Shapley values derived from the DeepAnchor model at every position of CBS and its surrounding region (±500bp) upon the testing dataset with 1036 CBSs. For example, CTCF binding intensity and its occupancy (EncOCctcfPval) display a periodical pattern at CTCF insulators, in which higher binding intensity of central CBS heavily contributes to positive CTCF insulator discrimination but the Shapley values flip at around 150bp far away from central CBS (Figure 2B left). Interestingly, we observed that the Shapley values of nucleosome positioning signature (EncNucleo) are reciprocally distributed with those of CTCF binding intensity (Figure 2B middle), and the nucleosome positioning level is anti-correlated with its Shapley value (Figure S2A), indicating that CTCF could enlarge and protect the linked DNA between its two neighboring nucleosomes to facilitate the insulation formation. Such base-wise visualization of feature contribution also greatly supports the previous finding that the binding of CTCF provides an anchor point for positioning nucleosomes thus resulting in an array of well-positioned nucleosomes flanking sites at CTCF insulator (Fu et al. 2008; Cuddapah et al. 2009). We also found that the

Shapley values of the conservation feature (GerpN) showed similar distribution as those of CTCF binding features (Figure 2B right). In addition, consistent with previous results (Luan et al. 2021), an investigation of base-wise Shapley value distribution upon positive or negative training datasets demonstrated that CTCF insulators usually demonstrated stronger binding intensity than other types of CBSs (Figure S2B).

Previous CTCF multivalency studies revealed that CTCF reads sequence diversity through combinatorial clustering of its 11 zinc fingers (ZFs) (Rhee and Pugh 2011; Nakahashi et al. 2013), yet the sequence determinants of CTCF-mediated insulation remain elusive. By applying MEME motif discovery to all predicted positive and negative CTCF insulators, we revealed two additional weak motifs flanking the core CTCF motif only at positive CTCF insulators (Figure 2C). In agreement with a sensitive insulator reporter assay (Huang et al. 2021b), the upstream motif separated from the core sequence by 5 or 6 bp is recognized by ZFs 9-11 and may contribute to the directionality of CTCF binding to CBSs (Yin et al. 2017). Interestingly, the downstream motif next to the core sequence is different from previous findings in terms of distance and sequence context (Nakahashi et al. 2013). Nonetheless, this downstream sequence associated with CTCF N terminus and ZFs 1-2 has been documented to stabilize cohesin engagement (Dong et al. 2020). Thus, the distinct sequence features together with multiple lines of existing evidence for CTCF insulation further demonstrate the fidelity of the DeepAnchor model.

Incorporating DeepAnchor score into loop extrusion model improves the performance of CTCF-anchored loop prediction

Although DeepAnchor models the insulative potential of genome-wide CBSs at the organismal level, accurate prediction of context-specific insulation and CTCF-anchored loop will be essential in understanding 3D chromatin interactions and fine-grained gene regulation. By assuming that looping interactions found in the CTCF ChIA-PET experiments are due to the blocking and localization of cohesin at CBSs, a simple but effective loop extrusion model (LEM) was previously developed (Xi and Beer 2021). This quantitative model evaluates CTCF-anchored loop formation with only four features, including CTCF-binding intensity, CTCF motif orientation, the distance between CTCF-binding events, and loop competition. Albeit the acquired specificity, we reasoned that incorporating the DeepAnchor score into LEM could further improve the performance of CTCF-anchored loop prediction. To this end, by considering the insulative potential of CTCF binding, we weighed the probability of CTCF binding in the original LEM model and reached a new implementation, call LoopAnchor (see Methods).

We compared the performance of LoopAnchor with five prevalent algorithms for CTCF-mediated interaction prediction, including Naïve and Oti methods (Oti et al. 2016), CTCF-MP (Zhang et al. 2018), Lollipop (Kai et al. 2018), and LEM (Xi and Beer 2021). To ensure a fair benchmark, all supervised models were trained on GM12878 RAD21 ChIA-PET data and tested on K562 RAD21 ChIA-PET data. We found that LoopAnchor outperforms other state-of-art methods in terms of both ROC-AUC (0.936) and PR-AUC (0.921) (Figure 3A), indicating the effectiveness of our model optimization. Besides, the correlation analysis

between the predicted loop intensities and the observed ChIA-PET signals on K562 showed that LoopAnchor receives the highest correlation (Pearson's r = 0.653) among the six tools (Figure 3B), which also demonstrates the superiority of LoopAnchor for quantitative measurement of CTCF-anchored loop.

To investigate whether our LoopAnchor mode could capture the dynamic loop changes during cell development, we applied our LoopAnchor model to human monocytes activation data with paired *in situ* Hi-C and CTCF ChIP-seq before and after exposure to PMA (Phanstiel et al. 2017). We found our LoopAnchor model can accurately predict 25/34 gained loops and all six lost CTCF-anchored loops (Figure 3C). For example, on the *JAG1* locus at chromosome 20:9547732-11575097, there are eight loops detected by Hi-C, among which one gained loop is identified only at differentiated macrophages (Figure 3D). Consistently, LoopAnchor not only detected all unchanged loops but also accurately predicted the gained loop (Figure 3D, see Methods). Notably, in this genomic region, the difference of loop intensity between monocyte and macrophage cells predicted by LoopAnchor is highly correlated with the loop score changes measured from Hi-C data (Pearson's r = 0.796), indicating that LoopAnchor can quantitatively capture the local loop structure changes at the dynamic cellular process.

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

Given the simplicity and gained performance of the LoopAnchor model, we sought to leverage existing large-scale CTCF ChIP-seq data to learn a global picture of CTCF-mediated chromatin interactions across different human tissue/cell types. We first uniformly processed 740 CTCF ChIP-seq datasets collected from ENCODE (Consortium et al. 2020), CistromDB (Zheng et al. 2019), and ChIP-Atlas (Oki et al. 2018) based on the ENCODE TF ChIP-seq processing pipeline (Table S2, see Methods). According to data quality control, peak calling filtering, and redundancy removing, we totally retrieved 168 high-quality CTCT ChIP-seq datasets, relating to 113 normal and 55 cancer samples across 32 tissues and 20 anatomical cell types (Figure 4A and Table S3, see Methods). Each biosample contains around 100,000 peaks (mean = 104,691, median = 96,296, Figure S3A). We applied LoopAnchor and LEM to detect genome-wide CTCF-mediated loops for each biosample respectively, and we found that nearly 20,000 loops were detected per biosample (mean = 16,255 and median = 16,637 for LoopAnchor, mean = 17,826 and median = 18,252 for LEM, Figure S3B and S3C).

Based on the shared pattern of detected loops across 168 biosamples, we first classified all LoopAnchor loops into four types: (1) 8.2% (*n* = 6,179) tissue/cell type-shared loops in more than 75% biosamples; (2) 9.0% (*n* = 6,802) tissue/cell type-relatively shared loops in more than 50% but less than 75% of biosamples; (3) 13.3% (*n* = 10,016) tissue/cell type-relatively specific loops in more than 25% but less than 50% of biosamples; (4) 69.4% (*n* = 52,216) tissue/cell type-specific loops in less than 25% of biosamples (Figure 4B, see Methods). Based on the same strategy, the LEM loops were also classified into these four types (Figure S4A). Compared to LEM, LoopAnchor can detect more shared loops but less specific loops globally (8.2% vs. 3.9% for tissue/cell type-shared loops, and 69.4% vs. 80.3%

for tissue/cell type-specific loops, Figure 4B and Figure S4A), suggesting LoopAnchor could capture more conservative CTCF-anchored loops across different tissue/cell types. By investigating the distribution of LoopAnchor-predicted loop intensity scores, we found a significant difference among the four loop categories (*P*-value < 2.22e-16, two-tailed Mann-Whitney U test, Figure 4C), in which the tissue/cell type-shared loops received the highest scores and the tissue/cell type-specific loops got the lowest scores. A similar pattern was also observed for the loop intensity score derived from the LEM method (Figure S4B). To evaluate the tissue/cell type specificity of classified CTCF-mediated loops, we leveraged the predicted chromatin loops by Peakachu for 42 human tissue/cell types (Salameh et al. 2020) and compared the number of associated tissue/cell types among different categories. We found that the tissue/cell type distribution of the LoopAnchor-predicted loops also showed significant differences among the four loop categories (*P*-value < 2.22e-16, two-tailed Mann-Whitney U 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), indicating the effectiveness of our loop classification. A similar trend was also identified for the detected loops from the LEM method (Figure S4C). Together, we established a compendium of tissue/cell type commonality and specificity for CTCF-anchored loops across large-scale human tissue/cell types. 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/hubs_landscape.txt or https://raw.githubusercontent.com/mulinlab/LoopAnchor/master/hubs_all.txt, respectively.

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

Multiple lines of evidence have supported that CTCF/cohesion binding sites are highly mutated in cancers (Katainen et al. 2015; Guo et al. 2018; Fang et al. 2020) or constantly shaped via evolutionary selection (Sadowski et al. 2019; Kentepozidou et al. 2020). However, whether context-specific CTCF insulation and associated loop anchors are more likely linked to disease-causal genomic loci has not been systematically tested. By collecting 12,738 causal variants for 54 blood-related autoimmune diseases from genome-wide association studies (GWASs) (Wang et al. 2020) (Table S4) and sampling matched control variants, we evaluated the genome-wide enrichment of autoimmune disease-causal variants on the CTCF-anchored loops predicted by LoopAnchor in normal tissues with more than three biosamples (see Methods). Notably, the autoimmune disease-causal variants were significantly enriched in the loop anchors in blood tissue (mean *P*-value = 0.00357) compared to other tissues (mean *P*-value ranges from 0.14465 (colon) to 0.0055 (prostate), Figure 5A). As the majority of disease-causal regulatory variants show tissue/cell type specificity in phenotypically relevant contexts, these results further support that context-dependent loop information could better interpret GWAS disease-causal variants for complex diseases.

To evaluate the relevance of CTCF-mediated loop anchors in the cancer mutation process,

we collected somatic mutations from ICGC whole genome mutation aggregation (Consortium 2020) and extracted recurrent mutations in the non-coding regions (see Methods). By comparing with permuted non-recurrent mutations, we found that mutations with higher recurrence obtain greater enrichment in CTCF-mediated loop anchors for cancer biosamples (two-tailed Mann-Whitney U test, Figure 5B). Besides, based on the whole-genome cancer mutation hotspots collected from three pan-cancer whole-genome analyses, including ATACseq-AWG (Corces et al. 2018), PCAWG (Rheinbay et al. 2020), and CNCDriver (Liu et al. 2019), we tested whether the mutation hotspots are more frequently occurred in CTCF-mediated loop anchors than other CTCF binding loci or shuffled non-coding genomic regions (see Methods). We uncovered consistent results among different datasets, in which the loop anchors were highly enriched with cancer mutation hotspots (two-tailed Mann-Whitney U test, Figure 5C-F). These results suggest genomic disruption of CTCF-anchored loops may thus represents a general causal mechanism of disease pathogenesis.

Discussion

The multiple critical functions of CTCF in the nucleus stimulate constant phenotypic and mechanistic research in the field, yet how CTCF recognizes insulators to exert chromosome barrier or enhancer blocking effects remains to be interrogated. In this study, we developed a novel computational model to accurately predict CTCF insulators and CTCF-anchored loops genome-widely. By incorporating large-scale base-wise genomic and epigenomic features within a deep learning model, we revealed several distinct chromatin and sequence features for CTCF-mediated insulation at a high resolution. Importantly, we found two additional sequence motifs flanking the core CTCF motif only at positive CTCF insulators. Besides, we leveraged the predicted insulator score to optimize the previous loop extrusion model and achieved better performance in predicting CTCF-anchored loops. Based on the model, for the first time, we established a compendium of tissue/cell type-specific and -shared CTCF-anchored loops across 52 human tissue/cell types. Finally, we showed that tissue/cell type-specific loop anchors are highly enriched at disease-causal loci. Therefore, our results deepen the understanding of CTCF-mediated insulation and loop formation. The new method together with the compiled resource also provides useful approaches for studying the dynamic regulation of 3D chromatin during cell differentiation and disease progression.

Although many computational models that can predict CTCF-mediated loops with varied sensitivity and specificity (Oti et al. 2016; Kai et al. 2018; Matthews and Waxman 2018; Zhang et al. 2018; Ibn-Salem and Andrade-Navarro 2019; Cao et al. 2021; Kuang and Wang 2021; Lv et al. 2021; Wang et al. 2021; Xi and Beer 2021), they usually learn from CTCF ChIA-PET data and barely evaluate the loop attribute among different types of CBS. Here, we applied cohesin ChIA-PET data and strict CTCF binding evidence to specifically analyze CTCF binding at the insulator and its anchored loop, which provides a unique tool for characterizing loop formation through the loop extrusion mechanism. Notably, the

limited availability and varied library construction quality of CTCF/cohesin ChIA-PET restrict the training of context-specific models broadly, however, our cross-cell type comparisons showed a high agreement among DeepAnchor models trained from different ChIA-PET data, indicating the discrimination of true CTCF insulators may rely on several conservative features revealed by our interpretable deep learning model, such as higher binding intensity, well-positioned nucleosomes as well as two unique motifs flanking the central CBS.

By introducing the DeepAnchor CTCF insulator score into a loop competition and extrusion model (Xi and Beer 2021), we observed improved performance in predicting CTCF-anchored loops. However, in both LEM and LoopAnchor, the quantitative predictions only measure the contribution of each component independently. For example, the insulation potential of CBS and loop competition are dependent processes based on dynamic chromatin context (Davidson and Peters 2021), which motivates a future optimization direction. Besides, our method is limited to inferring insulators and loops bound by CTCF, but recent studies indicated that some TFs could function as novel architectural proteins in regulating genome organization (Yi et al. 2021), such as YY1 (Weintraub et al. 2017), ZNF143 (Bailey et al. 2015), MAZ (Ortabozkoyun et al. 2022), BHLHE40 (Hu et al. 2020), CTCFL (Debruyne et al. 2019), etc., and some of which are independent of CTCF binding. Thus, predicting full spectrum insulators and their associated looping events across the genome deserve further in-depth investigation.

The landscape of CTCF-anchored loops established by applying LoopAnchor to 168 uniformly processed human CTCF ChIP-seq biosamples could be a useful resource in 3D genome and regulatory genomics studies. Since CTCF insulator scores were estimated at the organism level, LoopAnchor still only requires CTCF binding profiles to accurately predict CTCF-anchored loops, which greatly simplifies the application, particularly for studying dynamic 3D CTCF code during cell development and disease progression. In addition, the compiled tissue/cell type-shared and -specific loops could facilitate the interpretation of disease-causal variants identified by GWAS and somatic non-coding driver mutations in cancers.

Methods

DeepAnchor model

Here we describe the statistical model of DeepAnchor to identify insulator-associated CTCF binding sites (CBSs). For each CBS $n = 1, \ldots, N$, we extracted genomic/epigenomic features and DNA sequence features at a base-wise level for the ± 500bp region of the CBS center. By concatenating features and one-hot encoding sequence, a signal matrix x_n is generated for 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,

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

Where N_l is the number of neurons at layer l, x_i^{l-1} and x_k^l are the ith neuron at layer $l-1$ and k^{th} neuron at layer l, 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 belongs to a positive CTCF insulator.

Feature preparation

DNA binding motif of CTCF (MA0139.1) was downloaded from JASPAR2020 (Fornes et al. 2019). We used FIMO of MEME Suite (v5.1.1) (Grant et al. 2011) to scan the CTCF binding motif across the whole human genome (hg19), resulting in 110,059 CTCF binding sites with *P*-value < 1e-5. According to the topic relevance, we selected 44 base-wise genomic/epigenomic features from the CADD annotation database (Kircher et al. 2014; Rentzsch et al. 2019). For each CBS, feature values of ±500kb surrounding region were extracted and stored together to form a feature matrix. All features were scaled by a min-max scaling algorithm. To reduce the influence of extreme high/low values, we extracted 0.1 and 99.9 percentile as the min and max values. Besides, we got the DNA sequence of the ±500kb region of the CBS center using BEDTools (Quinlan and Hall 2010). A, T, C, G were then transferred into numerical value with one-hot encoding: A (1,0,0,0), T (0,1,0,0), G (0,0,1,0), C (0,0,0,1). Finally, by concatenating large-scale annotations with converted DNA features at the base-wise level, a feature matrix with a size of 1000×48 was generated for each CBS.

Training dataset and model fitting

RAD21 ChIA-PET data were re-processed using ChIA-PET2 with default parameters (Li et al. 2017). Inter-chromosomal interactions and interactions with less than 2 PETs were filtered out. High-quality RAD21 ChIA-PET and CTCF ChIP-seq data on GM12878 cell line (Grubert et al. 2020) were used to prepare the major training dataset. Firstly, the CBSs without overlapped CTCF ChIP-seq peaks were filtered out. Then we selected the remaining CBSs intersecting with the ChIA-PET anchor region as positive sets while considering the ones that do not overlap with ChIA-PET anchors as negative sets. CBSs were excluded from both positive sets and negative sets if they locate in the same ChIA-PET anchor or ChIP-seq peak since their identity cannot be fully determined. To avoid overfitting, we extracted a balanced negative set by randomly selecting samples from the total negative set. Finally, positive and negative datasets were divided into train, test, and valid set according to chromosomes. Samples on chr1-16 were selected as the training set, chr17-18 were validation set and the others were test set. We considered p_1 and p_2 as the possibility the CBS belongs to a true CTCF insulator or not, and we calculated binary cross entropy as the loss of model. Input data were divided into small batches with batch sizes of 50. The model was trained for 20 epochs, but early stopping was enabled when the validation loss does not decrease any more.

Evaluation of DeepAnchor model

We performed five-fold cross-validation to evaluate the LoopAnchor model and plotted ROC curves and corresponding AUC. To prove the reproducibility and the cross-sample performance, we trained separate models on several RAD21 ChIA-PET/CTCF ChIP-seq datasets from K562 and H1-hESC cell lines (Grubert et al. 2020). We set a cut-off to define positive and negative CTCF insulators by finding the points on the ROC curve with the largest Youden's J statistic. Same with the training procedure on GM12878, DeepAnchor models were trained for each cell type and the models were validated by all test sets belonging to three cell types. After training the DeepAnchor model, we applied it to all potential CBSs across the whole human genome. DeepAnchor will calculate a score ranging from [0, 1], with a larger value meaning more likely to be a positive CTCF insulator.

Human enhancer annotation was downloaded from the FANTOM5 database (Lizio et al. 2015). According to Youden's J statistic, CBSs with anchor scores> 0.5 are selected as positive CTCF insulators, while negative CTCF insulators are those with anchor scores < 0.5. We counted the number of enhancers for positive/negative insulators if the distance between them is smaller than 100kb, we tested the differences of involved enhancers between positive and negative insulators using the Mann Whitney U Test. The estimated TAD data was derived from the GM12878 Hi-C assay (Dixon et al. 2012), in which there are 4,386 TADs and the median length of TADs is 520kb. To test whether CBSs have preference for a position or not, we counted the distance between each CBS and the 5' end of TAD it belongs. Since TADs have variable lengths, we normalized the distance by counting the ratio concerning the total length of the TAD. Therefore, the normalized distance of CBSs will be the percentage of relative position within TAD.

Base-wise feature contribution analyses

We used SHapley Additive exPlanations (SHAP) (Lundberg et al. 2020) to interpret the output of DeepAnchor. Since the input of the DeepAnchor model is a 48x1000 feature matrix, SHAP returns the Shapley value matrix with the same alignment as the input feature matrix. We leveraged the shap package (Lundberg et al. 2018) to load the well-trained DeepAnchor model and used the training dataset as background examples to receive an expectation over. Then an explainer was generated based on the model and background dataset. The explainer was then applied to predict the Shapley values for every position via testing datasets. Note that the training dataset and testing dataset were the same as the ones in the training DeepAnchor model. We also interpreted the feature contribution based on the mean Shapley values for all positions.

Optimization of loop extrusion model in LoopAnchor

Enlightened by a recent loop competition and extrusion model (Xi and Beer 2021), we integrated the DeepAnchor score into the model to predict CTCF-anchored loops. We estimated the probability of insulator-associated CTCF binding at CBS *i*:

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

where S_i is the CTCF ChIP-seq signal and t_i is the DeepAnchor score at CBS *i*. a is a constant which have been estimated in the original paper. 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-art methods for CTCF-mediated loop prediction

We acquired Oti code from Oti et.al (Oti et al. 2016), and the naïve method was derived from Oti code by setting the recursive round of the Oti method to 1. We used the default parameters to run Oti and Naïve methods, and loops with a score > 0 were considered positive loops while others were treated as negative ones. CTCF-MP (Zhang et al. 2018), Lollipop (Kai et al. 2018), and LEM (Xi and Beer 2021) 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. We also calculated the Pearson correlation coefficient between predicted loop intensity with real data to show the performance of all methods.

Application of LoopAnchor for dynamic loop detection during macrophage development

We downloaded CTCF ChIP-seq data for both monocyte and macrophage cells (Phanstiel et al. 2017). 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. We also obtained cell state-specific based on Hi-C loops detected by HiCCUPS (Rao et al. 2014). Because Hi-C loops are not necessary to be mediated by CTCF and cohesin, we filtered out loops without CTCF peaks in anchor regions on both sides. 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 systematically collected CTCF ChIP-seq datasets from ENCODE (Consortium et al. 2020), CistromDB (Zheng et al. 2019), and ChIP-Atlas (Oki et al. 2018), and uniformly processed them using the ENCODE TF ChIP-seq processing pipeline (https://github.com/ENCODE-DCC/chip-seq-pipeline2). Briefly, the clean reads were mapped to the human genome (GRCh37/hg19) with BWA (v0.7.17) (Li and Durbin 2009) followed by the post-alignment filtering, and finally, the peaks were called using SPP (v1.15) (Kharchenko et al. 2008). To select the high-quality ChIP-seq biosamples, we firstly filtered out the outliers with excessively low or high peaks. Then, all B lymphocyte biosamples prefixed with "GM" were removed except GM12878. For each mapped tissue or anatomical cell type, we dropped redundant biosamples according to Jaccard similarity clustering of called peaks and selected one biosample in each cluster with a maximal peak number. Finally, we applied LoopAnchor and LEM to detect genome-wide CTCF-mediated loops for each biosample respectively.

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

All predicted CTCF-anchored loops were generally classified into four categories according to their shared patterns across different tissue/cell types. Specifically, loops that were shared in no less than 75% tissue/cell types were classified as tissue/cell type-shared loops. Loops that were shared by no less than 50% but less than 75% 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% tissue/cell types were classified as tissue/cell type-relatively specific loops. The remaining loops which were shared by less than 25% tissue/cell types were classified as tissue/cell type-specific loops. The public Hi-C loops detected by Peakachu (Salameh et al. 2020) for 42 tissue/cell types were downloaded from the 3D genome browser (Wang et al. 2018). Then, the "pairtopair" command from BEDTools (Quinlan and Hall 2010) with parameter "-type both" was used to compare them with loops predicted by LoopAnchor.

GWAS disease-causal variant enrichment

We collected 12,738 causal variants for 54 blood-related autoimmune diseases from CAUSALdb (Wang et al. 2020). To test the genome-wide enrichment of autoimmune disease-causal variants on the CTCF-anchored loops predicted by LoopAnchor, we first sampled the same amount of control variants 10,000 times with matched allele frequency

using vSampler (Huang et al. 2021a). For each normal biosample, the predicted CTCF-anchored loops were flattened into unique anchors, and the "intersect" command from BEDTools (Quinlan and Hall 2010) was used to perform the colocalization between variants and anchors. The enrichment *P*-value was calculated by counting the number of permutations that receive a higher percentage of overlapping variants than the real dataset in each biosample. This analysis was only performed on normal tissues with more than three biosamples.

Non-coding somatic mutation and hotspot enrichment

The genome-wide somatic mutations were downloaded from ICGC Data Portal (Release 28) (Consortium 2020). Candidate non-coding somatic mutations were retained by removing these overlapped with exons and splicing sites and were classified into three categories according to their mutation recurrence (∈[2,5), ∈[5,10), ≧10). The somatic mutations with only one recurrence were used as the background for drawing control datasets. We generated 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 count of intersected hits between somatic mutations in each category and anchors in a particular biosample was calculated. By comparing to 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 test the difference in fold changes among the three categories. In addition, we also curated candidate somatic mutation hotspots or driver regions from three previous catalogs, including ATACseq-AWG (Corces et al. 2018), PCAWG (Rheinbay et al. 2020), and CNCDriver (Liu et al. 2019). For each cancer biosample, the loops predicted by LoopAnchor with a score≥0.01 were kept 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 (Quinlan and Hall 2010) was also used to retrieve shuffled hotspots with 100 permutations as "Shuffled" datasets. The Mann-Whitney U test was used to test the difference in the overlapped percentage of somatic hotspots among these generated datasets.

Data availability

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/hubs_landscape.txt or https://raw.githubusercontent.com/mulinlab/LoopAnchor/master/hubs_all.txt, respectively.

Code availability

The source codes of LoopAnchor are freely available under MIT License at https://github.com/mulinlab/LoopAnchor.

Acknowledgements

Funding

National Natural Science Foundation of China [32070675, 31871327]; Natural Science Foundation of Tianjin [19JCJQJC63600].

Competing interests

The authors declare no conflict of interest.

Contributions

M.J.L. and H.X. conceived of the project. H.X. and X.Y. performed the analyses and implemented the software. W.W., X.C., D.H., S.Z., X.D., Z.W., J.W., Y.Z., K.Z. and H.Y. contributed the data collection and processing. J.W., K.C., D.P. and P.C.S. contributed to manuscript polishing and provided analysis suggestions. H.X., X.Y. and M.J.L. wrote the manuscript. The authors read and approved the final submission.

References

Ali T, Renkawitz R, Bartkuhn M. 2016. Insulators and domains of gene expression. *Curr Opin Genet Dev* **37**: 17-26.

Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R, Akhtar-Zaidi B, Scacheri PC, Haibe-Kains B, Lupien M. 2015. ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat Commun* **2**: 6186.

Beagan JA, Duong MT, Titus KR, Zhou L, Cao Z, Ma J, Lachanski CV, Gillis DR, Phillips-Cremins JE. 2017. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res* **27**: 1139-1152.

Braccioli L, de Wit E. 2019. CTCF: a Swiss-army knife for genome organization and transcription regulation. *Essays Biochem* **63**: 157-165.

Cao F, Zhang Y, Cai Y, Animesh S, Zhang Y, Akincilar SC, Loh YP, Li X, Chng WJ, 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.

Clarkson CT, Deeks EA, Samarista R, Mamayusupova H, Zhurkin VB, Teif VB. 2019. CTCF-dependent chromatin boundaries formed by asymmetric nucleosome arrays with decreased linker length. *Nucleic Acids Res* **47**: 11181-11196.

Consortium EP. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

Consortium EP, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A et al. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**: 699-710.

Consortium ITP-CAoWG. 2020. Pan-cancer analysis of whole genomes. *Nature* **578**: 82-93.

Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, Silva TC, Groeneveld C, Wong CK, Cho SW et al. 2018. The chromatin accessibility landscape of primary human cancers. *Science* **362**.

Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, 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.

Davidson IF, Bauer B, Goetz D, Tang W, Wutz G, Peters JM. 2019. DNA loop extrusion by human cohesin. *Science* **366**: 1338-1345.

Davidson IF, Peters JM. 2021. Genome folding through loop extrusion by SMC complexes. *Nat Rev Mol Cell Biol* **22**: 445-464.

Debruyne DN, Dries R, Sengupta S, Seruggia D, Gao Y, Sharma B, Huang H, Moreau L, McLane M, Day DS et al. 2019. BORIS promotes chromatin regulatory interactions in treatment-resistant cancer cells. *Nature* **572**: 676-680.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376-380.

Dong X, Huang D, Yi X, Zhang S, Wang Z, Yan B, Chung Sham P, Chen K, Jun Li M. 2020. Diversity spectrum analysis identifies mutation-specific effects of cancer driver genes. *Commun Biol* **3**: 6.

Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, Weintraub AS, Schujiers J, Lee TI, Zhao K et al. 2014. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**: 374-387.

Fang C, Wang Z, Han C, Safgren SL, Helmin KA, Adelman ER, Serafin V, Basso G, Eagen KP, Gaspar-Maia

A et al. 2020. Cancer-specific CTCF binding facilitates oncogenic transcriptional dysregulation. *Genome Biol* **21**: 247.

Filippova GN, Fagerlie S, Klenova EM, Myers C, Dehner Y, Goodwin G, Neiman PE, Collins SJ, Lobanenkov VV. 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.

Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D et al. 2019. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* doi:10.1093/nar/gkz1001.

Fu Y, Sinha M, Peterson CL, Weng Z. 2008. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet* **4**: e1000138.

Gabriele M, Brandao HB, Grosse-Holz S, Jha A, Dailey GM, Cattoglio C, Hsieh TS, Mirny L, Zechner C, Hansen AS. 2022. Dynamics of CTCF- and cohesin-mediated chromatin looping revealed by live-cell imaging. *Science* doi:10.1126/science.abn6583: eabn6583.

Ghirlando R, Felsenfeld G. 2016. CTCF: making the right connections. *Genes Dev* **30**: 881-891.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**: 1017-1018.

Grubert F, Srivas R, Spacek DV, 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.

Guo Y, Monahan K, Wu H, Gertz J, Varley KE, Li W, Myers RM, Maniatis T, Wu Q. 2012. CTCF/cohesin-mediated DNA looping is required for protocadherin alpha promoter choice. *Proc Natl Acad Sci U S A* **109**: 21081-21086.

Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, 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.

Guo YA, Chang MM, Huang W, Ooi WF, Xing M, Tan P, Skanderup AJ. 2018. Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nat Commun* **9**: 1520.

He C, Wang X, Zhang MQ. 2014. Nucleosome eviction and multiple co-factor binding predict estrogen-receptor-alpha-associated long-range interactions. *Nucleic Acids Res* **42**: 6935-6944.

Hu G, Dong X, Gong S, Song Y, Hutchins AP, 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.

Huang D, Wang Z, Zhou Y, Liang Q, Sham PC, Yao H, Li MJ. 2021a. vSampler: fast and annotation-based matched variant sampling tool. *Bioinformatics* **37**: 1915-1917.

Huang H, Zhu Q, Jussila A, Han Y, Bintu B, Kern C, Conte M, Zhang Y, Bianco S, Chiariello AM et al. 2021b. CTCF mediates dosage- and sequence-context-dependent transcriptional insulation by forming local chromatin domains. *Nat Genet* **53**: 1064-1074.

Ibn-Salem J, Andrade-Navarro MA. 2019. 7C: Computational chromosome conformation capture by correlation of ChIP-seq at CTCF motifs. *BMC Genomics* **20**: 777.

Kai Y, Andricovich J, Zeng Z, Zhu J, Tzatsos A, Peng W. 2018. Predicting CTCF-mediated chromatin interactions by integrating genomic and epigenomic features. *Nat Commun* **9**: 4221.

Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE, Ristolainen H, Hanninen UA, Cajuso T et al. 2015. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet*

47: 818-821.

Kentepozidou E, Aitken SJ, Feig C, Stefflova K, Ibarra-Soria X, Odom DT, Roller M, Flicek P. 2020. Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biol* **21**: 5.

Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* **26**: 1351-1359.

Kim Y, Shi Z, Zhang H, Finkelstein IJ, Yu H. 2019. Human cohesin compacts DNA by loop extrusion. *Science* **366**: 1345-1349.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**: 310-315.

Kuang S, Wang L. 2021. Deep learning of sequence patterns for CCCTC-binding factor-mediated chromatin loop formation. *J Comput Biol* **28**: 133-145.

Li G, Chen Y, Snyder MP, Zhang MQ. 2017. ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis. *Nucleic Acids Res* **45**: e4.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Liu EM, Martinez-Fundichely A, Diaz BJ, Aronson B, Cuykendall T, MacKay M, Dhingra P, Wong EWP, Chi P, Apostolou E et al. 2019. Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole Genomes. *Cell Syst* **8**: 446-455 e448.

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.

Lobanenkov VV, Nicolas RH, Adler VV, Paterson H, Klenova EM, Polotskaja AV, Goodwin GH. 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.

Luan J, Xiang G, Gomez-Garcia PA, Tome JM, Zhang Z, Vermunt MW, Zhang H, Huang A, Keller CA, Giardine BM et al. 2021. Distinct properties and functions of CTCF revealed by a rapidly inducible degron system. *Cell Rep* **34**: 108783.

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2**: 56-67.

Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, Low DK, Newman SF, Kim J et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* **2**: 749-760.

Lv H, Dao FY, Zulfiqar H, Su W, Ding H, Liu L, Lin H. 2021. A sequence-based deep learning approach to predict CTCF-mediated chromatin loop. *Brief Bioinform* **22**.

Matthews BJ, Waxman DJ. 2018. Computational prediction of CTCF/cohesin-based intra-TAD loops that insulate chromatin contacts and gene expression in mouse liver. *Elife* **7**.

Nakahashi H, Kieffer Kwon KR, 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.

Nichols MH, Corces VG. 2015. A CTCF code for 3D genome architecture. *Cell* **162**: 703-705.

Oh S, Shao J, Mitra J, Xiong F, D'Antonio M, Wang R, Garcia-Bassets I, Ma Q, Zhu X, Lee JH et al. 2021.

Enhancer release and retargeting activates disease-susceptibility genes. *Nature* **595**: 735-740.

Oki S, Ohta T, Shioi G, Hatanaka H, Ogasawara O, Okuda Y, Kawaji H, Nakaki R, Sese J, Meno C. 2018. ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep* **19**.

Ong CT, Corces VG. 2014. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet* **15**: 234-246.

Ortabozkoyun H, Huang PY, Cho H, Narendra V, LeRoy G, Gonzalez-Buendia E, Skok JA, Tsirigos A, Mazzoni EO, Reinberg D. 2022. CRISPR and biochemical screens identify MAZ as a cofactor in CTCF-mediated insulation at Hox clusters. *Nat Genet* **54**: 202-212.

Oti M, Falck J, Huynen MA, Zhou H. 2016. CTCF-mediated chromatin loops enclose inducible gene regulatory domains. *BMC Genomics* **17**: 252.

Phanstiel DH, Van Bortle K, Spacek D, Hess GT, Shamim MS, Machol I, Love MI, Aiden EL, Bassik MC, Snyder MP. 2017. Static and Dynamic DNA Loops form AP-1-Bound Activation Hubs during Macrophage Development. *Mol Cell* **67**: 1037-1048 e1036.

Phillips JE, Corces VG. 2009. CTCF: master weaver of the genome. *Cell* **137**: 1194-1211.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

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

Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. 2019. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**: D886-D894.

Rhee HS, Pugh BF. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**: 1408-1419.

Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, Tiao G, Hornshoj H, Hess JM, Juul RI, Lin Z et al. 2020. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**: 102-111.

Sadowski M, Kraft A, Szalaj P, Wlasnowolski M, Tang Z, Ruan Y, Plewczynski D. 2019. Spatial chromatin architecture alteration by structural variations in human genomes at the population scale. *Genome Biol* **20**: 148.

Salameh TJ, Wang X, Song F, Zhang B, Wright SM, Khunsriraksakul C, Ruan Y, Yue F. 2020. A supervised learning framework for chromatin loop detection in genome-wide contact maps. *Nat Commun* **11**: 3428.

Sanborn AL, Rao SS, Huang SC, Durand NC, Huntley MH, Jewett AI, Bochkov ID, 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 U S A* **112**: E6456-6465.

Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R, Oberdoerffer S. 2011. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**: 74-79.

Vostrov AA, Quitschke WW. 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.

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.

Wang W, Gao L, Ye Y, Gao Y. 2021. CCIP: Predicting CTCF-mediated chromatin loops with transitivity. *Bioinformatics* doi:10.1093/bioinformatics/btab534.

Wang Y, Song F, Zhang B, Zhang L, Xu J, Kuang D, Li D, Choudhary MNK, 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.

Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ, Cohen MA, Nabet B, Buckley DL et al. 2017. YY1 Is a structural regulator of enhancer-promoter Loops. *Cell* **171**: 1573-1588 e1528.

Xi W, Beer MA. 2021. Loop competition and extrusion model predicts CTCF interaction specificity. *Nat Commun* **12**: 1046.

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.

Yin M, Wang J, Wang M, Li X, Zhang M, Wu Q, Wang Y. 2017. Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites. *Cell Res* **27**: 1365-1377.

Zhang R, Wang Y, Yang Y, Zhang Y, Ma J. 2018. Predicting CTCF-mediated chromatin loops using CTCF-MP. *Bioinformatics* **34**: i133-i141.

Zheng R, Wan C, Mei S, Qin Q, Wu Q, Sun H, Chen CH, Brown M, Zhang X, Meyer CA et al. 2019. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res* **47**: D729-D735.

Figure Legends

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 CBSs. Positive (Pos) and negative (Neg) dataset is constructed by considering both CTCF ChIP-seq peaks and cohesin ChIA-PET loop anchors. A 1D-CNN model (DeepAnchor) is then used to train a classifier for distinguishing anchor-associated CBSs versus other ones. Probability of CBSs being anchor-associated can be calculated and used as DeepAnchor score for downstream analyses.

B. ROC curves of five-fold cross-validation based on GM12878 dataset.

C. DeepAnchor score thresholding (by finding the points on ROC curve with largest Youden's J statistic) and predicted results for training dataset.

D. Cross-sample ROC curves based on GM12878, K562 and H1-hESC datasets, respectively.

E. Correlation between DeepAnchor scores predicted by models trained on different datasets.

F. Distribution of DeepAnchor score across the whole genome.

G. Comparison of number of enhancers around CBSs between predicted Pos and Neg CTCF insulators.

H. Strand-oriented asymmetric pattern of enhancer enrichment at predicted Pos and Neg CTCF insulators.

I. 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 insulators.

Figure 2. Base-wise analysis of anchor-associated patterns for CTCF binding

A. Feature importance analysis at base-wise level. Up: Summation of absolute feature Shapley values across $\pm$500bp of CTCF insulator. Down: Average absolute feature Shapley values of each position at $\pm$500bp of CTCF insulator.

B. Base-wise Shapley value distribution for three representative features.

C. DNA motifs associated with positive CTCF insulators (up) and negative CTCF insulators (down).

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. Gain and lost loops from monocyte to macrophage differentiation by comparing LoopAnchor prediction with Hi-C observation, log10(Fold change) is the transformed fold change of predicted loop intensity for specific loop between macrophage and monocyte; log10(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 gained and static loops. For LoopAnchor, linewidth represents the predicted loop intensity and the loop with fold change of intensity >3 was marked with red colour.

Figure 4. The 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, including 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 two-tailed Mann-Whitney U test was used to test the significance.

D. 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 two-tailed Mann-Whitney U test was used to test the significance.

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.

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.

Figure 1

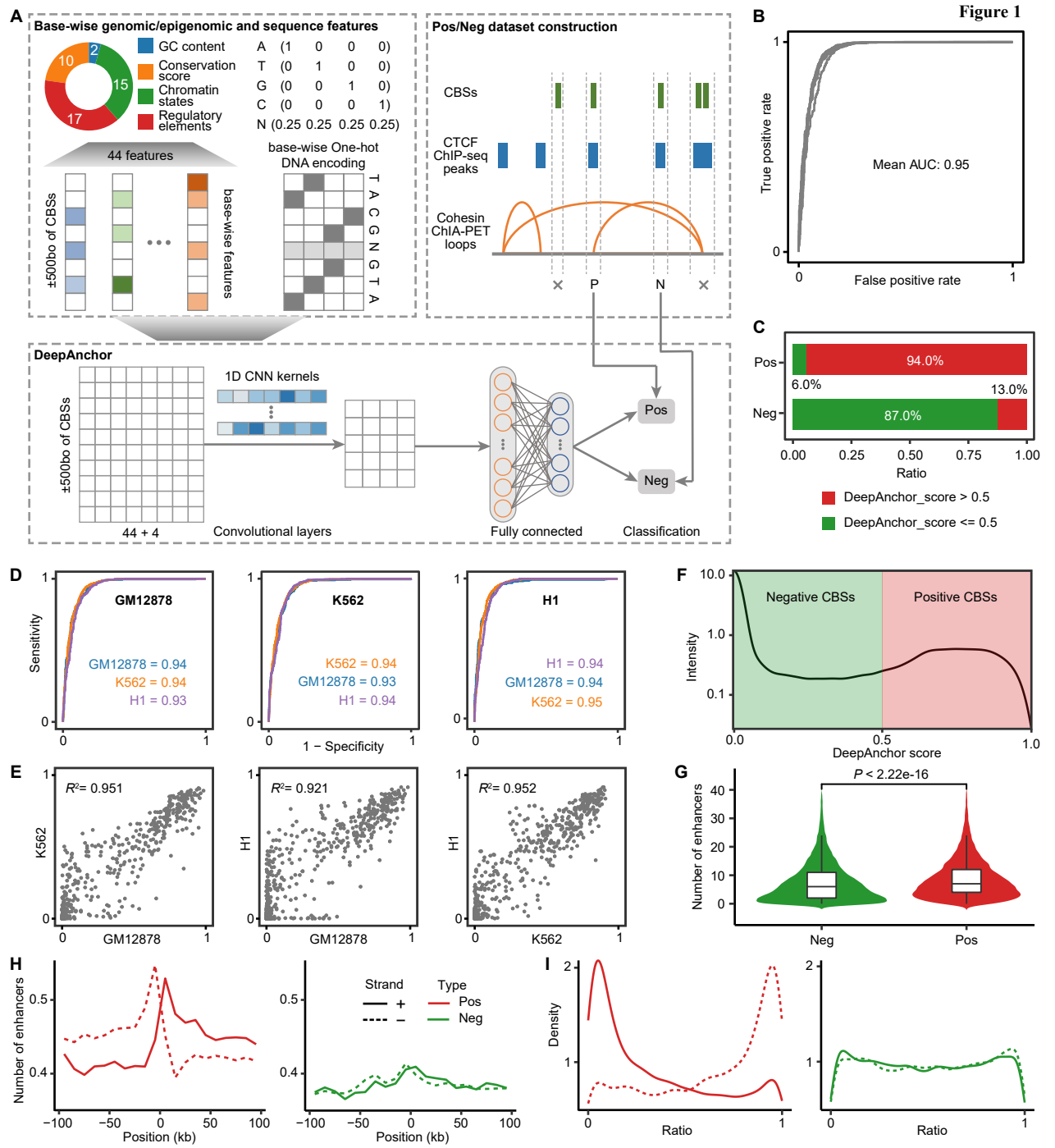

A
Base-wise genomic/epigenomic and sequence features
GC content
Conservation score
Chromatin states
Regulatory elements
2
15
17
10
44 features
±500bo of CBSs
base-wise features
A (1 0 0 0)
T (0 1 0 0)
G (0 0 1 0)
C (0 0 0 1)
N (0.25 0.25 0.25 0.25)
base-wise One-hot DNA encoding
T A C G N G T A
Pos/Neg dataset construction
CBSs
CTCF ChIP-seq peaks
Cohesin ChIA-PET loops
P
N
DeepAnchor
1D CNN kernels
±500bo of CBSs
44 + 4
Convolutional layers
Fully connected
Classification
Pos
Neg
B
True positive rate
False positive rate
Mean AUC: 0.95
C
Pos
Neg
94.0%
6.0%
13.0%
87.0%
Ratio
DeepAnchor_score > 0.5
DeepAnchor_score <= 0.5
D
GM12878
K562
H1
Sensitivity
1 − Specificity
GM12878 = 0.94
K562 = 0.94
H1 = 0.93
K562 = 0.94
GM12878 = 0.93
H1 = 0.94
H1 = 0.94
GM12878 = 0.94
K562 = 0.95
F
Negative CBSs
Positive CBSs
Intensity
DeepAnchor score
E
R²= 0.951
R²= 0.921
R²= 0.952
K562
H1
GM12878
G
P < 2.22e-16
Number of enhancers
Neg
Pos
H
Number of enhancers
Position (kb)
Strand
+
−
Type
Pos
Neg
I
Density
Ratio

Figure 2

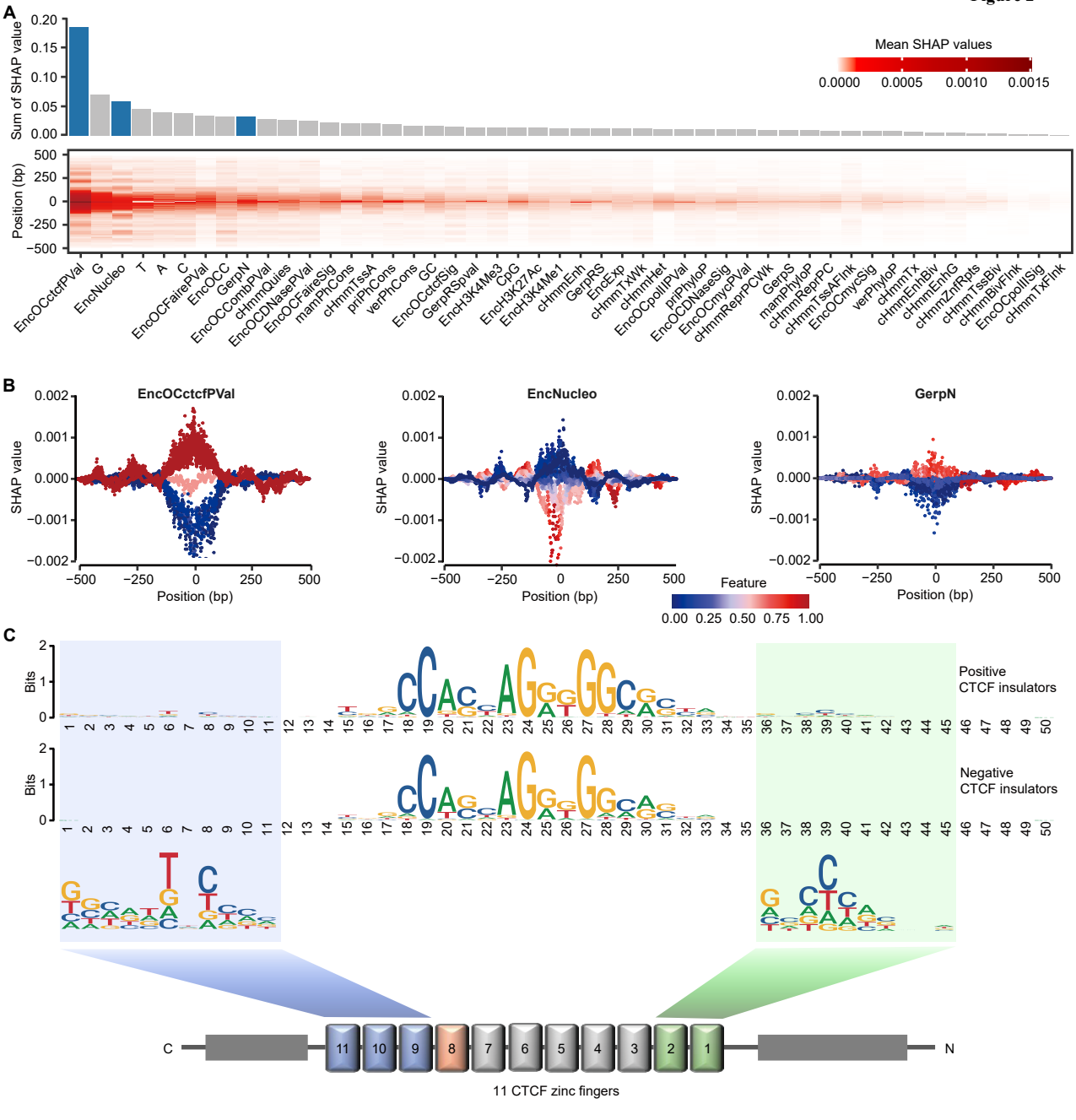

A
Sum of SHAP value
Mean SHAP values
0.0000 0.0005 0.0010 0.0015
Position (bp)
EncOCctcfPVal
G
EncNucleo
T
A
C
EncOCFairePVal
EncOCC
GerpN
EncOCCombPVal
cHmmQuies
EncOCDNasePVal
EncOCFaireSig
mamPhCons
cHmmTssA
priPhCons
verPhCons
GC
EncOCctcfSig
GerpRSpval
EncH3K4Me3
CpG
EncH3K27Ac
EncH3K4Me1
cHmmEnh
GerpRS
EncExp
cHmmTxWk
cHmmHet
EncOCpolIIPval
priPhyloP
EncOCDNaseSig
EncOCmycPval
cHmmReprPCWk
GerpS
mamPhyloP
cHmmReprPC
cHmmTssAFlnk
EncOCmycSig
verPhyloP
cHmmTx
cHmmEnhBiv
cHmmEnhG
cHmmZnfRpts
cHmmTssBiv
cHmmBivFlnk
EncOCpolIISig
cHmmTxFlnk
B
EncOCctcfPVal
EncNucleo
GerpN
SHAP value
Position (bp)
Feature
0.00 0.25 0.50 0.75 1.00
C
Bits
Positive CTCF insulators
Negative CTCF insulators
C
N
11 CTCF zinc fingers

Figure 3

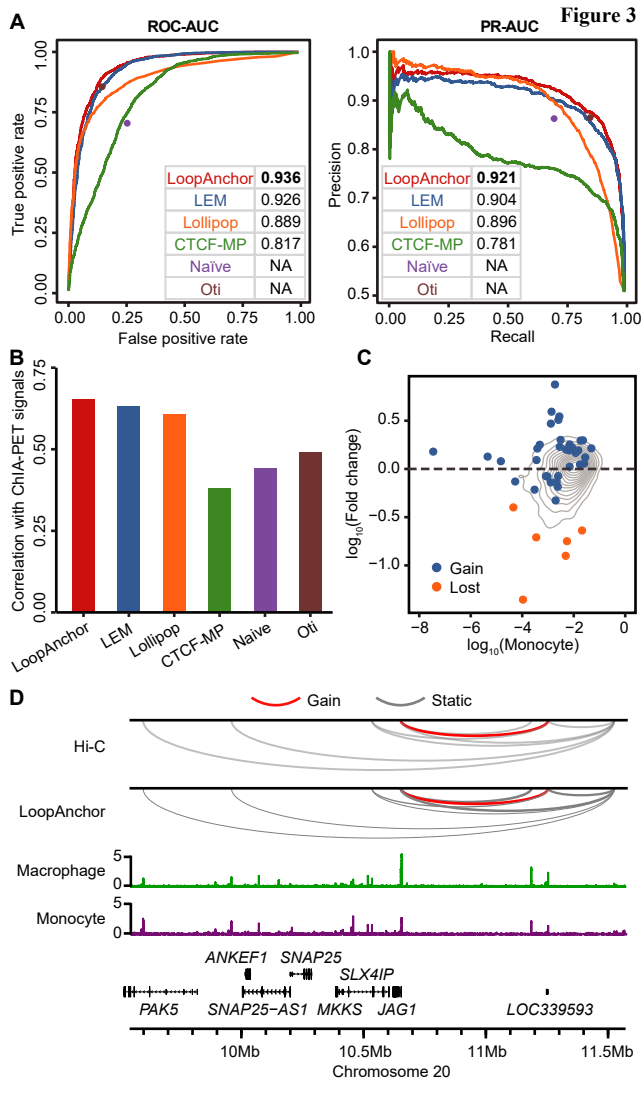

A
ROC-AUC
True positive rate
False positive rate
LoopAnchor 0.936
LEM 0.926
Lollipop 0.889
CTCF-MP 0.817
Naïve NA
Oti NA
PR-AUC
Precision
Recall
LoopAnchor 0.921
LEM 0.904
Lollipop 0.896
CTCF-MP 0.781
Naïve NA
Oti NA
B
Correlation with ChIA-PET signals
LoopAnchor
LEM
Lollipop
CTCF-MP
Naive
Oti
C
log10(Fold change)
log10(Monocyte)
Gain
Lost
D
Gain
Static
Hi-C
LoopAnchor
Macrophage
Monocyte
ANKEF1
SNAP25
SLX4IP
PAK5
SNAP25-AS1
MKKS
JAG1
LOC339593
10Mb
10.5Mb
11Mb
11.5Mb
Chromosome 20

Figure 4

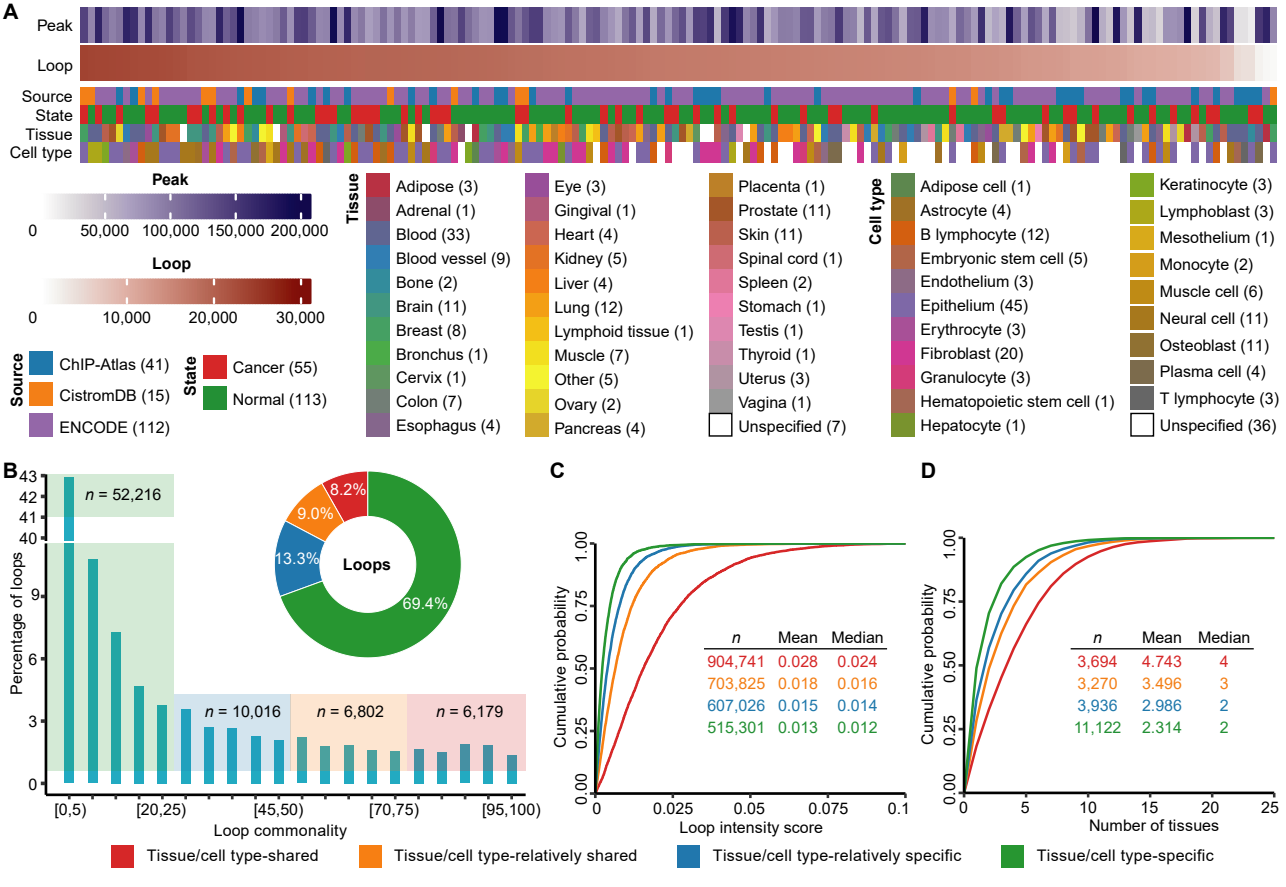
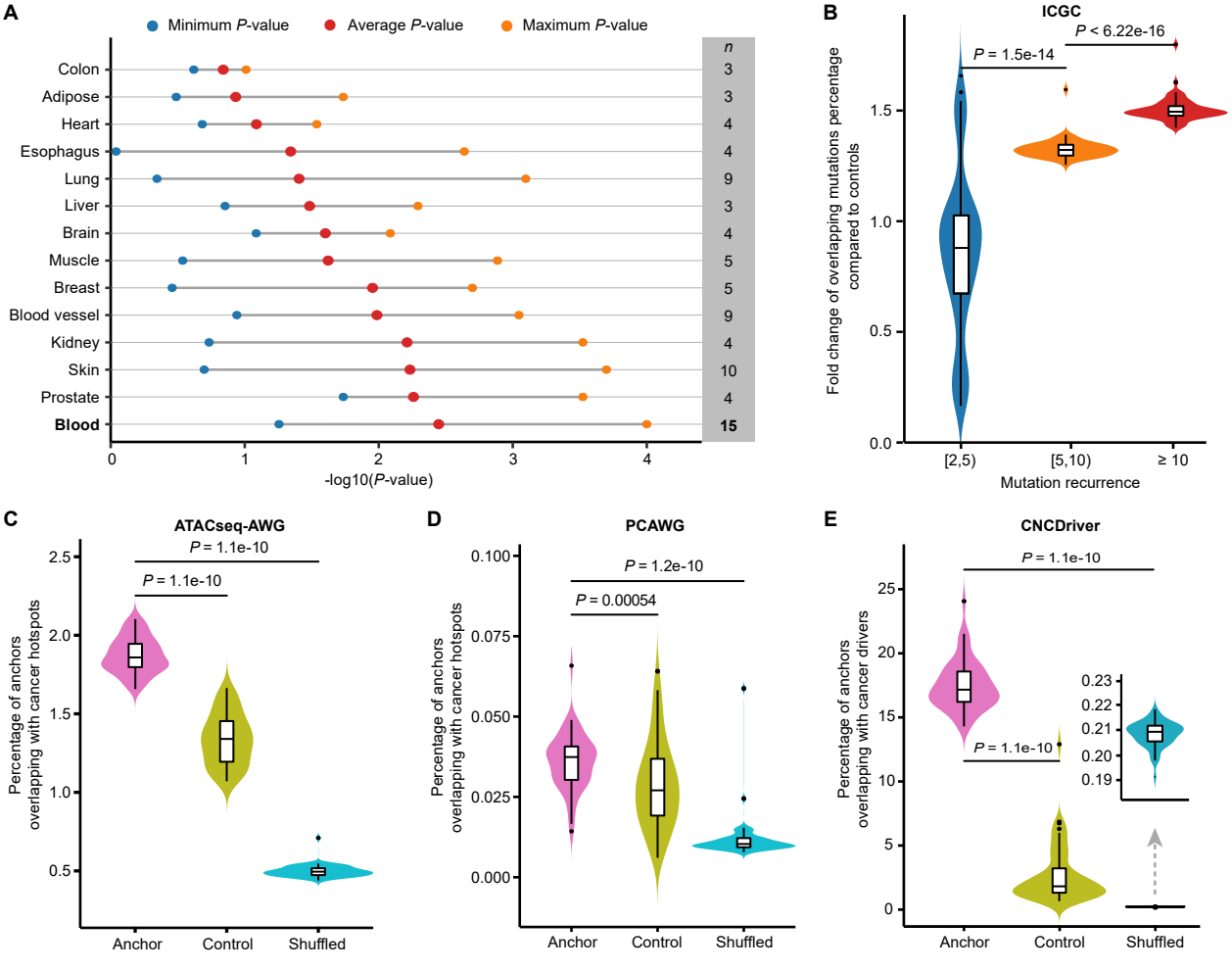

A
Peak
Loop
Source
State
Tissue
Cell type
Peak
0
50,000
100,000
150,000
200,000
Loop
0
10,000
20,000
30,000
Source
ChIP-Atlas (41)
CistromDB (15)
ENCODE (112)
State
Cancer (55)
Normal (113)
Tissue
Adipose (3)
Adrenal (1)
Blood (33)
Blood vessel (9)
Bone (2)
Brain (11)
Breast (8)
Bronchus (1)
Cervix (1)
Colon (7)
Esophagus (4)
Eye (3)
Gingival (1)
Heart (4)
Kidney (5)
Liver (4)
Lung (12)
Lymphoid tissue (1)
Muscle (7)
Other (5)
Ovary (2)
Pancreas (4)
Placenta (1)
Prostate (11)
Skin (11)
Spinal cord (1)
Spleen (2)
Stomach (1)
Testis (1)
Thyroid (1)
Uterus (3)
Vagina (1)
Unspecified (7)
Cell type
Adipose cell (1)
Astrocyte (4)
B lymphocyte (12)
Embryonic stem cell (5)
Endothelium (3)
Epithelium (45)
Erythrocyte (3)
Fibroblast (20)
Granulocyte (3)
Hematopoietic stem cell (1)
Hepatocyte (1)
Keratinocyte (3)
Lymphoblast (3)
Mesothelium (1)
Monocyte (2)
Muscle cell (6)
Neural cell (11)
Osteoblast (11)
Plasma cell (4)
T lymphocyte (3)
Unspecified (36)
B
n = 52,216
n = 10,016
n = 6,802
n = 6,179
Percentage of loops
Loop commonality
[0,5)
[20,25)
[45,50)
[70,75)
[95,100)
Loops
8.2%
9.0%
13.3%
69.4%
C
Cumulative probability
Loop intensity score
n Mean Median
904,741 0.028 0.024
703,825 0.018 0.016
607,026 0.015 0.014
515,301 0.013 0.012
D
Cumulative probability
Number of tissues
n Mean Median
3,694 4.743 4
3,270 3.496 3
3,936 2.986 2
11,122 2.314 2
Tissue/cell type-shared
Tissue/cell type-relatively shared
Tissue/cell type-relatively specific
Tissue/cell type-specific

Figure 5


A
Minimum P-value
Average P-value
Maximum P-value
n
Colon 3
Adipose 3
Heart 4
Esophagus 4
Lung 9
Liver 3
Brain 4
Muscle 5
Breast 5
Blood vessel 9
Kidney 4
Skin 10
Prostate 4
Blood 15
0 1 2 3 4
-log10(P-value)
B
ICGC
P = 1.5e-14
P < 6.22e-16
Fold change of overlapping mutations percentage compared to controls
0.0 0.5 1.0 1.5
[2,5) [5,10) ≥ 10
Mutation recurrence
C
ATACseq-AWG
P = 1.1e-10
P = 1.1e-10
Percentage of anchors overlapping with cancer hotspots
0.5 1.0 1.5 2.0 2.5
Anchor Control Shuffled
D
PCAWG
P = 1.2e-10
P = 0.00054
Percentage of anchors overlapping with cancer hotspots
0.000 0.025 0.050 0.075 0.100
Anchor Control Shuffled
E
CNCDriver
P = 1.1e-10
P = 1.1e-10
Percentage of anchors overlapping with cancer drivers
0 5 10 15 20 25
0.19 0.20 0.21 0.22 0.23
Anchor Control Shuffled