JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 28, Number 2, 2021

Pp. 133–145
DOI: 10.1089/cmb.2020.0225

Deep Learning of Sequence Patterns for CCCTC-Binding Factor-Mediated Chromatin Loop Formation

SHUZHEN KUANG[1,2] and LIANGJIANG WANG[1]

ABSTRACT

The three-dimensional (3D) organization of the human genome is of crucial importance for gene regulation, and the CCCTC-binding factor (CTCF) plays an important role in chromatin interactions. However, it is still unclear what sequence patterns in addition to CTCF motif pairs determine chromatin loop formation. To discover the underlying sequence patterns, we have developed a deep learning model, called DeepCTCFLoop, to predict whether a chromatin loop can be formed between a pair of convergent or tandem CTCF motifs using only the DNA sequences of the motifs and their flanking regions. Our results suggest that DeepCTCFLoop can accurately distinguish the CTCF motif pairs forming chromatin loops from the ones not forming loops. It significantly outperforms CTCF-MP, a machine learning model based on word2vec and boosted trees, when using DNA sequences only. Furthermore, we show that DNA motifs binding to several transcription factors, including ZNF384, ZNF263, ASCL1, SP1, and ZEB1, may constitute the complex sequence patterns for CTCF-mediated chromatin loop formation. DeepCTCFLoop has also been applied to disease-associated sequence variants to identify candidates that may disrupt chromatin loop formation. Therefore, our results provide useful information for understanding the mechanism of 3D genome organization and may also help annotate and prioritize the noncoding sequence variants associated with human diseases.

Keywords: 3D genome, chromatin loops, CTCF, deep learning, sequence motifs.

1. INTRODUCTION

THE HUMAN GENOME IS PACKAGED into highly complex structures in the nucleus with multiple levels of organization, such as chromatin loops and topologically associating domains (TADs) at the intermediate scale (Rao et al., 2014; Tang et al., 2015). The three-dimensional (3D) genome organization is critical for many cellular processes, including the transcriptional control of gene expression via enhancer–promoter interactions (Bonev and Cavalli, 2016). To characterize the 3D genome architecture, several high-throughput methods have been developed, including chromosome conformation capture (Hi-C) for detecting global chromatin interactions (Lieberman-Aiden et al., 2009) and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) for capturing genome-wide chromatin interactions mediated by specific protein

Departments of [1]Genetics and Biochemistry and [2]Biological Sciences, Clemson University, Clemson, South Carolina, USA.

factors (Fullwood et al., 2009). Results from recent studies indicate that 3D genome organization involves the architectural protein, CCCTC-binding factor (CTCF), and the structural maintenance of chromosomes complex, cohesin (Rao et al., 2014; Tang et al., 2015). The key roles of CTCF and cohesin are suggested by their colocalization on chromatin and enrichment at the anchors of chromatin loops and TAD boundaries (Wendt et al., 2008; Rao et al., 2014). Moreover, depletion of CTCF or cohesin results in the loss of loop structures (Nora et al., 2017; Rao et al., 2017).

Interestingly, CTCF-binding sites at loop anchors are mostly in the convergent orientation (Rao et al., 2014; Tang et al., 2015). The functional significance of the convergent orientation has been demonstrated by the topological change of chromatin loops and the alteration of gene expression if the CTCF-binding sites are inverted using CRISPR/Cas9 (Guo et al., 2015). The preferential orientation of CTCF motifs as well as the enrichment of CTCF and cohesin in the loop anchor regions may be explained by the loop extrusion model, which proposes that cohesin complex loads on the DNA and extrudes a progressively larger loop until reaching two convergent CTCF-binding sites (Sanborn et al., 2015; Fudenberg et al., 2016, 2017). However, the model cannot explain why many pairs of convergent CTCF-binding sites do not form chromatin loops. In addition, although the majority of chromatin loops have convergent CTCF-binding sites, some loops are formed between CTCF-binding sites in the tandem orientation (Rao et al., 2014). Recent studies suggest that, besides CTCF and cohesin, some other proteins such as BRD2, PDS5, and WAPL may also play important roles in establishing chromatin loop boundaries (Hsu et al., 2017; Wutz et al., 2017). It is thus interesting to examine what sequence patterns in addition to the presence of CTCF-binding sites are important for chromatin loop formation.

To unravel the sequence patterns for the formation of CTCF-mediated chromatin loops, CTCF-MP, a machine learning model based on word2vec and boosted trees, has been developed (Zhang et al., 2018). Word2vec is a computationally efficient method to learn word embedding using neural networks. It can encode each word in a text corpus as a vector in a continuous vector space where semantically similar words are located near each other (Mikolov et al., 2013). For CTCF-MP, words (*k*-mers) in DNA sequences are encoded as vectors using word2vec to learn sequence features and reduce input dimensionality. The relatively high performance of CTCF-MP with word2vec features suggests the existence of additional sequence patterns besides CTCF motifs in DNA sequences. However, the features learned by word2vec are hard to interpret. More recently, a deep learning model, called DeepMILO, has been developed to predict the impact of noncoding variants on insulator loops using DNA sequence as input (Trieu et al., 2020). Nevertheless, the sequence patterns learned by DeepMILO have not been fully examined.

Convolutional neural networks (CNNs) have attracted much attention in the field of biology because of the capability to discover informative motifs directly from the input sequences (Alipanahi et al., 2015; Kelley et al., 2016; Quang and Xie, 2016). Long short-term memory (LSTM) network can be used to learn long-range dependencies between sequence motifs (Hochreiter and Schmidhuber, 1997a; Quang and Xie, 2016; Angermueller et al., 2017). Moreover, attention mechanisms may be applied to biological problems to capture and emphasize the most important features within sequential input (Zhou et al., 2016; Chen et al., 2019; Li et al., 2019). However, these advanced deep learning techniques have not been fully utilized to model the sequence patterns for CTCF-mediated chromatin loop formation.

Genetic studies suggest that disruption of 3D genome organization can cause gene dysregulation and may be associated with the onset and progression of human disease (Norton and Phillips-Cremins, 2017; Anania and Lupiáñez, 2020). For instance, the deletion of CTCF anchor sites has been shown to cause the activation of genes outside the CTCF-CTCF loop (Ji et al., 2016). Mutations that can affect the boundaries of insulated neighborhoods are found in many types of cancer (Hnisz et al., 2016). Notably, most of the disease-associated single nucleotide polymorphisms (SNPs) from genome-wide association studies (GWAS) or high-throughput whole-genome sequencing are located in the noncoding regions, and how these sequence variants contribute to the pathogenesis of disease is still poorly understood. Thus, it is helpful to examine the effects of disease-associated variants on the formation of CTCF-mediated chromatin loops.

In this study, we have developed a new deep learning model, called DeepCTCFLoop, to predict whether a chromatin loop can be formed for a pair of CTCF motifs, either convergent or tandem, and learn the sequence patterns hidden in the adjacent sequences of CTCF motifs. DeepCTCFLoop utilizes a two-layer CNN and an attention-based bidirectional LSTM (BLSTM) network to learn and emphasize the relevant features, including the sequence motifs, the interactions between sequence motifs, and the long-range dependencies between high-level features. We show that DeepCTCFLoop can accurately predict CTCF-mediated chromatin loop formation. By examining the features learned by the first convolution

layer, we have identified several additional DNA-binding proteins, which may also be involved in the formation of CTCF-mediated chromatin loops. Furthermore, our results suggest that DeepCTCFLoop can be used to analyze disease-associated sequence variants for their effects on CTCF-mediated chromatin loop formation.

The source code and data sets used in this study for model construction are freely available at https://github.com/BioDataLearning/DeepCTCFLoop.

2. MATERIALS AND METHODS

2.1. Data collection and preprocessing

Data sets for model construction were downloaded for three different cell types, GM12878, HeLa, and K562. As previously described for CTCF-MP (Zhang et al., 2018), a positive instance was collected as a pair of convergent or tandem CTCF motifs in a chromatin loop region plus 250 nucleotides (nt) on each side of a motif, giving rise to a 1038-nt genomic sequence (Fig. 1). The locations of CTCF motifs were determined by scanning the human genome sequence (hg19, https://hgdownload.soe.ucsc.edu/goldenPath/hg19/chromosomes) using FIMO (Grant et al., 2011) with the known position weight matrix (PWM) of CTCF in JASPAR (MA0139.1) (Khan et al., 2018). The chromatin loop regions for the cell lines GM12878 and HeLa were downloaded from NCBIs Gene Expression Omnibus (Edgar, 2002) (GEO accession: GSE72816), and the regions of the cell line K562 were obtained from ENCODE (Feingold et al., 2004). The chromatin loop regions were detected by ChIA-PET for CTCF (Fullwood et al., 2009). The negative instances were compiled by randomly selecting the convergent or tandem CTCF motif pairs that were not in the chromatin loop regions, under the constraint that the distribution of the distances between CTCF motif pairs of negative instances was similar as for the positive instances (Fig. 1). The positive and negative instances for each cell line (30,627 positive and 30,624 negative instances for GM12878; 12,656 positive and 12,650 negative instances for HeLa; and 9900 positive and 9898 negative instances for K562) were randomly divided into training, validation, and test data with the ratio of 80%:10%:10%.

2.2. DeepCTCFLoop model construction

The architecture of DeepCTCFLoop is shown in Figure 2. The input of the model is the DNA sequence of convergent or tandem CTCF motif pairs and flanking regions. The DNA sequence is one-hot-encoded into a 4×1038 binary matrix with A = [1, 0, 0, 0], T = [0, 1, 0, 0], G = [0, 0, 1, 0], and C = [0, 0, 0, 1] and then used by DeepCTCFLoop to predict whether a chromatin loop can be formed between a pair of CTCF motifs. The predicted probability is expected to be close to 1 for a true chromatin loop and close to 0 for a negative instance.

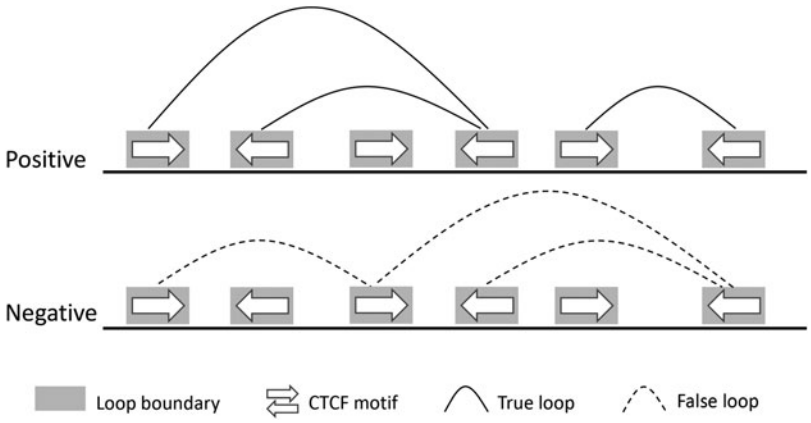

FIG. 1. Schematic diagram of convergent or tandem CTCF motif pairs used to compile the positive and negative instances. The positive instances are defined as the DNA sequences of CTCF motif pairs in the chromatin loop regions and their flanking regions. The negative instances are the DNA sequences of randomly selected CTCF motif pairs not in the chromatin loop regions but with the same orientations as the positive motif pairs, under the constraint that the distribution of the distances between the CTCF motif pairs of negative instances is similar as for the positive instances. CTCF, CCCTC-binding factor.

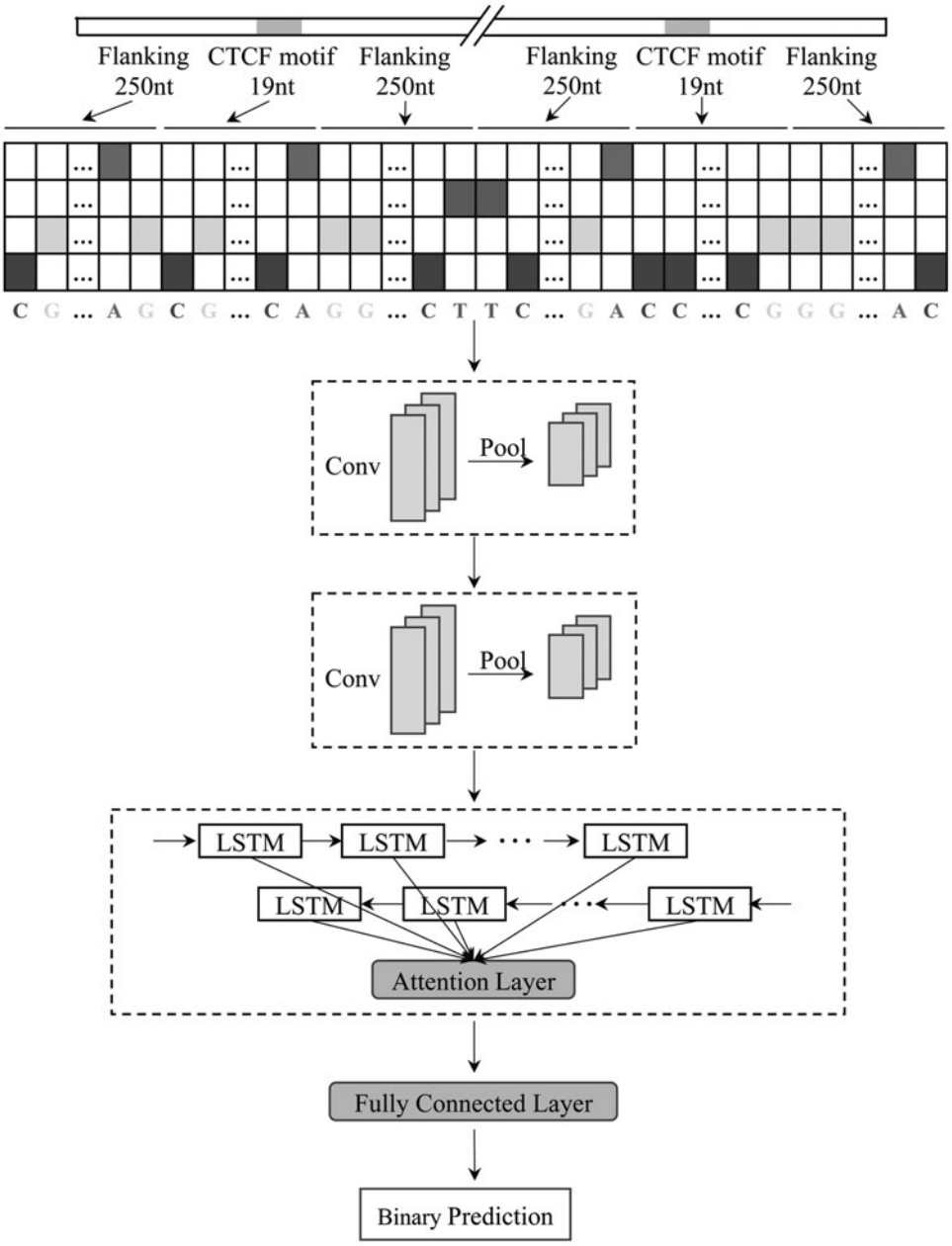

FIG. 2. Diagram of DeepCTCFLoop architecture. The DNA sequence of the CTCF motifs and their surrounding genomic sequences (250 nt) were taken as input by encoding into a binary matrix. Then, a two-layer convolutional neural network was used to learn the sequence motifs and high-level features. The bidirectional LSTM layer was used to learn the long-range dependencies between the high-level features. Next, an attention layer was used to capture the most important features for high model performance. Finally, two fully connected layers were used to combine the output from the attention layer and make the binary prediction. LSTM, long short-term memory.

The one-hot-encoded matrix of the DNA input is first fed into a 1D convolution layer. The N filters of the convolution layer with dimension $4 \times L$, where L is the length of a filter, convolve over the input matrix, resulting in N activation maps. The activation value a_{fi}^{s} for the filter f at the position i of an input sequence s is computed as:

$$a_{fi}^{s} = \max\left(0,\ \sum_{l=1}^{L}\sum_{d=1}^{4} w_{ld}^{f} s_{i+l,d}\right), \tag{1}$$

where w^f is the weight matrix for the filter f. The convolution filters function as motif detectors to discover the patterns within the input sequences. The parameters of the filters can be interpreted as PWMs. High activation values indicate the existence of a motif represented by a PWM at the corresponding positions in the input sequence.

After the first convolution layer, a max pooling layer is used to get the maximum activation value of spatially adjacent subregions. As a downsampling strategy, the max pooling layer can reduce input dimensionality and thus avoid model overfitting. Then, a second convolution layer followed by another max pooling layer is employed to learn the high-level interactions between sequence motifs. The model with the two-layer CNN was selected after comparison with one-layer CNN and three-layer CNN models.

Next, a layer of BLSTM is used to learn the long-range dependencies among the high-level features learned by the two-layer CNN. Compared with vanilla recurrent neural networks, LSTM is able to overcome the vanishing gradient problem (Hochreiter and Schmidhuber, 1997b). Each LSTM unit consists of an input gate, a forget gate and an output gate. These gates decide what information should be thrown away, be stored, or go to the output. LSTM is thus able to remember the information for a long period and learn the long-range dependencies. Here, BLSTM is used to scan the input both forward and backward.

Following the BLSTM layer, an attention layer is used to pay more attention on the most important features by assigning more weights to them (Zhou et al., 2016). The output is then fed into a fully connected layer, and the sigmoid function is used to calculate the probability of forming a chromatin loop.

In this study, the binary cross-entropy loss function was minimized using the Adam optimization algorithm with minibatches (Kingma and Ba, 2015). Dropout and L2 regularization were used to regularize the model. The early stopping procedure was also employed to avoid model overfitting. The model was implemented in Python using Keras 2.2.4 (https://github.com/fchollet/keras) with TensorFlow 1.5.0 as the backend. The hyperparameters for model training were tuned using Bayesian optimization via Hyperopt (Bergstra et al., 2013) with the data from GM12878, resulting in the number of CNN filters (N) as 208, the length of filters (L) as 13, the size of pooling layer as 4, the LSTM units as 64, the learning rate as 1e-4, the L2 regularization as 5e-5, the dropout rate after CNN as 0.43, and the dropout rate after the attention layer as 0.05. The average time used for model training and evaluation was about 2 hours with the data from the three cell lines.

2.3. Motif visualization and analysis

The filters of the first convolution layer were converted into PWMs as previously described (Kelley et al., 2016). Given a filter f with length L, it scanned all the positive sequences and calculated an activation value for each position i of a sequence s. If an activation value was greater than half of the maximum activation m of filter f over all positions of the positive sequences (Equation 2), the subsequence corresponding to that activation value was collected. The collected subsequences were aligned and converted into PWMs, which were then visualized using WebLogo (Crooks et al., 2004).

$$m = \max_{s,\,i} a_{fi}^{s}. \tag{2}$$

The PWMs were analyzed using hypergeometric and Kolmogorov–Smirnov (KS) tests to identify PWMs overrepresented in positive instances and the ones with different position distributions between positive and negative instances. The identified PWMs may represent the functional motifs in positive instances. The analyses were conducted separately for convergent or tandem CTCF motif pairs to explore the possible differences of sequence patterns.

The PWMs learned by DeepCTCFLoop were compared with the known motifs in the JASPAR database (Mathelier et al., 2014) using TomTom program from the MEME Suite (Bailey et al., 2009). Two motifs with E-value ≤0.05 were considered to be significantly matched.

2.4. Prioritization of candidate sequence variants

The candidate sequence variants associated with human diseases were obtained from the GWAS catalog database (ftp://ftp.ebi.ac.uk/pub/databases/gwas/releases/2020/05/04/gwas-catalog-associations.tsv) (Welter et al., 2014). The variants with p-values ≤1e-6 were collected for further analysis as suggested by GWAS to reduce the false-positive rate for disease association. The variants located in the chromatin loop regions were analyzed using DeepCTCFLoop to estimate their impacts on the formation of CTCF-mediated

chromatin loops. If a variant was predicted as negative and the predicted probability reduced more than 0.25 (compared with the wild type), it was selected as a candidate sequence variant that may disrupt CTCF-mediated chromatin loop formation.

2.5. Model performance evaluation

DeepCTCFLoop was evaluated using the test data set with the following performance metrics:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{3}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}, \tag{4}$$

$$\text{Specificity} = \frac{TN}{TN+FP}, \tag{5}$$

$$\text{MCC} = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \tag{6}$$

Here, *TP*, *TN*, *FP*, and *FN* represent the number of true positives, true negatives, false positives, and false negatives, respectively. Matthews correction coefficient (MCC) is often used as a robust metric of model performance. Moreover, the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) are used for model evaluation.

3. RESULTS AND DISCUSSION

3.1. Accurate prediction of CTCF-mediated chromatin loop formation

DeepCTCFLoop has been developed to predict the chromatin loop formation mediated by a pair of CTCF motifs in either convergent or tandem orientation and to discover the underlying sequence patterns (Fig. 1). It takes CTCF motif pairs and their surrounding genomic sequences as inputs (Fig. 2). The performance of DeepCTCFLoop was evaluated using the data sets derived from three different cell lines, including GM12878, HeLa, and K562. Since neural network architecture may affect performance, models with one to three layers of CNN were tested using the GM12878 data set. As shown in Figure 3,

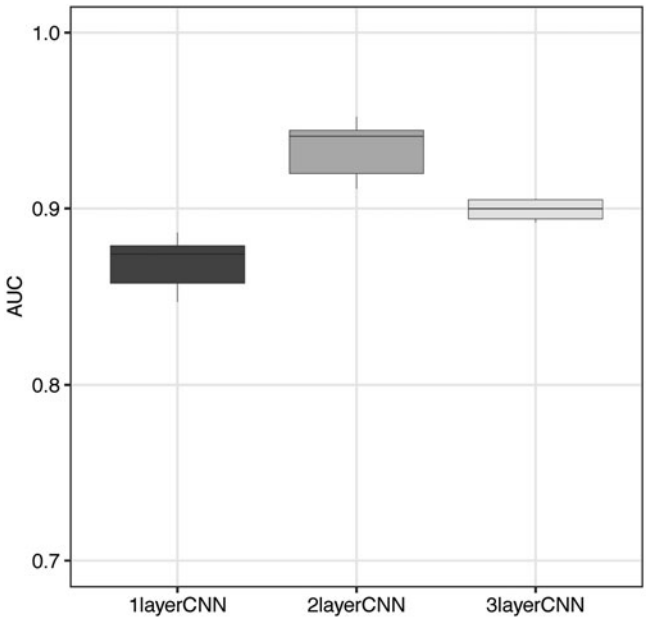

FIG. 3. Effect of neural network architecture on model performance. The AUC value was used to evaluate the model performance. The AUC values from 10 repetitions on the test data sets from GM12878 are shown. AUC, area under the ROC curve; CNN, convolutional neural networks; ROC, receiver operating characteristic.

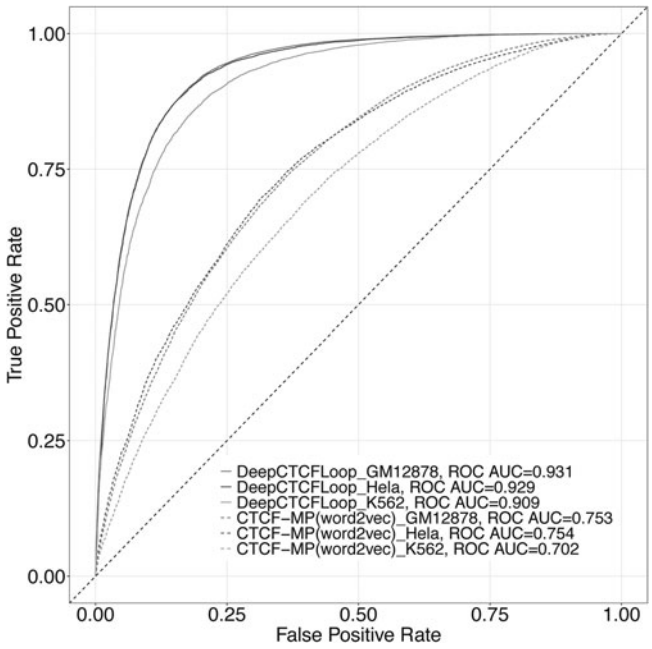

FIG. 4. ROC curves of DeepCTCFLoop and CTCF-MP (word2vec features only) on the test data sets of GM12878, Hela, and K562.

DeepCTCFLoop with the two-layer CNN achieved the best performance. The hyperparameters for model construction were also optimized using the GM12878 data set. Dropout, L2 regularization, and the early stopping procedure were used to avoid model overfitting.

As shown in Figure 4 and Table 1, DeepCTCFLoop achieved the mean AUC of 0.931 for GM12878, 0.929 for HeLa, and 0.909 for K562 on the test data sets for 10 repetitions. The high model performance suggests that DeepCTCFLoop has learned relevant features from the DNA sequences to distinguish the loop-forming CTCF motif pairs (positive instances) from noninteracting ones (negative instances). By comparison, CTCF-MP (Zhang et al., 2018) achieved relatively poor performance on the same data sets with the mean AUC of 0.753 for GM12878, 0.754 for HeLa, and 0.702 for K562, when only using the DNA sequence features from word2vec (Fig. 4 and Table 1). The superior performance of DeepCTCFLoop over CTCF-MP was also suggested by the significantly higher accuracy, sensitivity, specificity, and MCC values (Table 1). Although word2vec may capture the contextual information between *k*-mers (DNA words) by learning their semantical similarity, our results suggest that DeepCTCFLoop can capture more relevant information, such as sequence motifs and their relationships, from the input DNA sequences. This is consistent with the poor performance of word2vec on detecting informative motifs in a previous study (Trabelsi et al., 2019). Taken together, our results demonstrate the capability of DeepCTCFLoop to accurately predict the formation of chromatin loops mediated by convergent or tandem CTCF motif pairs.

Table 1. Superior Performance of DeepCTCFLoop Over CTCF-MP When Only Using DNA Sequences as the Input

	DeepCTCFLoop			*CTCF-MP (word2vec)*		
Metrics	GM12878	HeLa	K562	GM12878	HeLa	K562
Accuracy	0.861	0.860	0.830	0.685	0.678	0.619
Sensitivity	0.899	0.891	0.901	0.759	0.804	0.874
Specificity	0.824	0.829	0.759	0.612	0.552	0.365
MCC	0.725	0.721	0.668	0.377	0.373	0.275
AUC	0.931	0.929	0.909	0.753	0.754	0.702

For CTCF-MP, DNA sequences were encoded into vector features by word2vec. The average accuracy, sensitivity, specificity, MCC, and AUC for 10 repetitions on the test data sets from GM12878, HeLa, and K562 cells are shown.

AUC, area under the ROC curve; CTCF, CCCTC-binding factor; MCC, Matthews correction coefficient; ROC, receiver operating characteristic.

3.2. *Discovery of interesting sequence motifs for CTCF-mediated loop formation*

The superior performance of DeepCTCFLoop suggests that it may have learned the complex sequence patterns to distinguish loop-forming CTCF motif pairs from noninteracting ones. To understand the sequence patterns, the filters of the first convolutional layer were converted into PWMs (see Section 2). For the model trained with data from GM12878 cells (GM12878 model), 192 PWMs were derived and analyzed using hypergeometric and KS tests to identify the PWMs that might represent candidate motifs also involved in the chromatin loop formation mediated by convergent or tandem CTCF motif pairs. Among the 192 PWMs, 22 and 20 PWMs showed their significant enrichment (hypergeometric FDR ≤0.01) in chromatin loops with convergent and tandem motif pairs, respectively, and had significantly different sequence

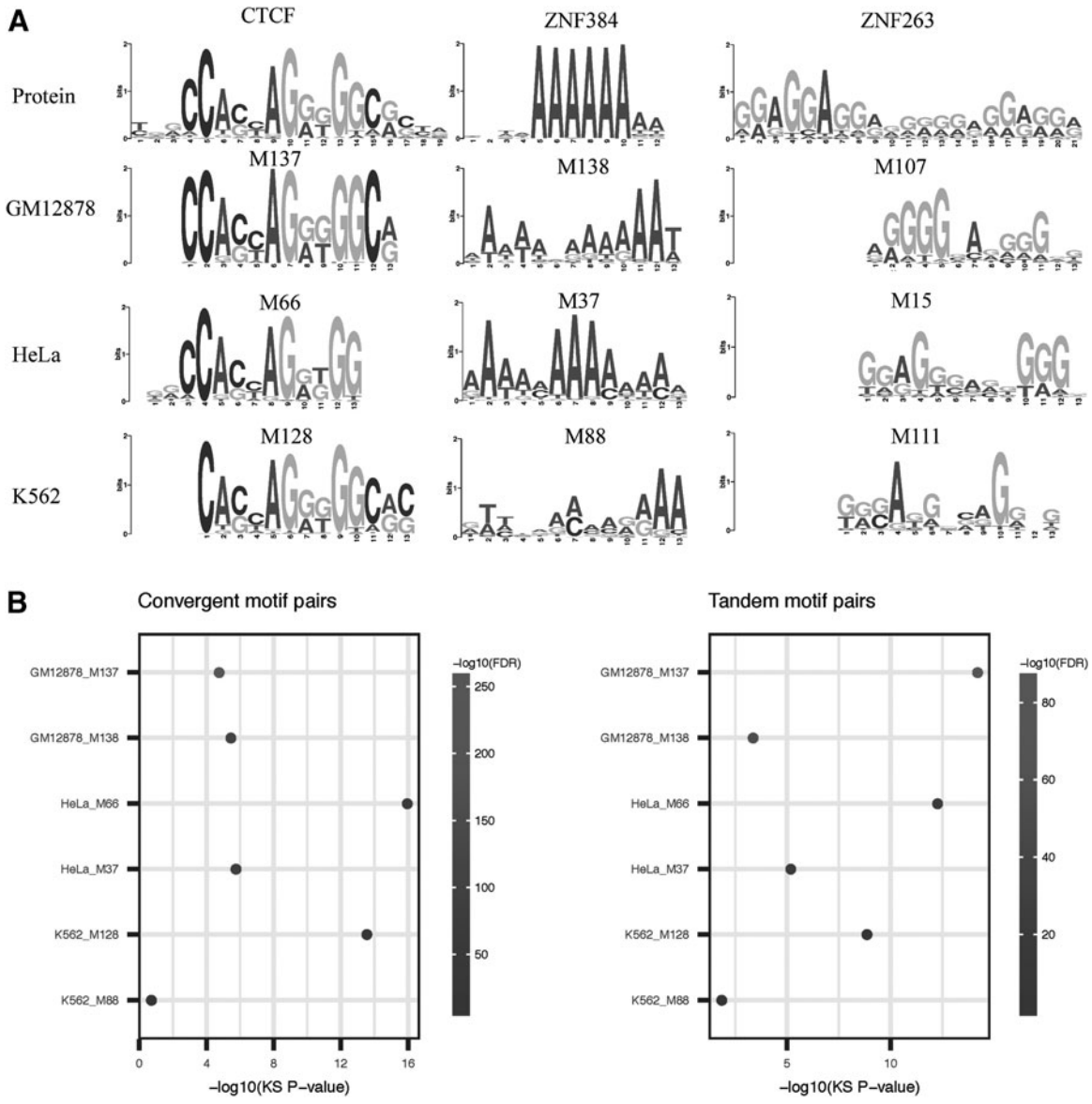

FIG. 5. Selected PWMs significantly matched with the DNA motifs of CTCF, ZNF384, and ZNF263. **(A)** Sequence logos of the selected PWMs. **(B)** Significant enrichment of some PWMs in the loop-forming convergent or tandem motif pairs and different sequence position distributions between positive and negative instances. The enrichment in positive instances was measured using the FDR value from hypergeometric test. The difference between position distributions was evaluated using the KS test. The PWMs learned from the GM12878, HeLa, and K562 data sets were compared with the known motifs in the JASPAR database using TomTom. FDR, false discovery rate; KS, Kolmogorov–Smirnov; PWM, position weight matrix.

position distributions between the positive and negative instances (KS p-value ≤0.01). Interestingly, 11 of these enriched PWMs were shared between the loop-forming convergent and tandem CTCF motif pairs. When compared with the known transcription factor (TF) motifs in the JASPAR database (Mathelier et al., 2014) using TomTom (Bailey et al., 2009), 7 of the 11 PWMs were matched with the CTCF motif, and the PWM M138 was significantly matched with the DNA-binding motif of TF ZNF384 (Fig. 5 and Table 2). Besides the shared PWMs, six and nine specific PWMs were enriched for the convergent and tandem CTCF motif pairs, respectively, suggesting possible differences in the underlying sequence patterns for chromatin loop formation. Moreover, besides M138, 10 additional PWMs were significantly matched with other TF motifs (Table 2). The results suggest that these DNA-binding TF proteins, especially ZNF384, are involved in the CTCF-mediated chromatin loop formation.

To identify the common motifs underlying chromatin loop formation in different cell types, we performed the same analysis for the HeLa and K562 models. For the HeLa model, 41 and 14 PWMs were found to show significant enrichment for the loop-forming convergent and tandem CTCF motif pairs, respectively, and different sequence position distributions between the positive and negative instances; for the K562 model, 29 and 8 such PWMs were identified. Notably, the DNA motif of ZNF384 was also matched by the PWM M37 of the HeLa model and M88 of the K562 model (Fig. 5 and Table 2). While M37 was found to be shared by the loop-forming convergent and tandem CTCF motif pairs in HeLa cells, M88 was only enriched for the convergent motif pairs in K562 cells. Besides ZNF384, the motif of ZNF263 was also matched by the PWMs from all three models (GM12878, HeLa, and K562), and several other TFs, including ASCL1, SP1, and ZEB1, were identified by two different models (Table 2).

Our results strongly suggest the involvement of ZNF384 in CTCF-mediated chromatin loop formation. ZNF384, a C2H2-type zinc finger protein, has been experimentally shown to be involved in chromatin looping and may contribute to the sequence specificity of loop formation by interacting with CTCF (Whalen et al., 2016). As for the other candidate TFs, ZNF263 is also a C2H2-type zinc finger protein, and its motif has been shown to be enriched in the conserved sequence of lncRNAs that are positioned at the loop end points and chromatin boundaries with significantly higher mutation levels in cancer (Amaral et al., 2018). ASCL1 is an evolutionarily conserved basic-helix-loop-helix TF, which has been shown to promote local chromatin accessibility at its target regions during neurogenesis and activate transcription by mostly

Table 2. List of Some Interesting Position Weight Matrices Learned by DeepCTCFLoop

DeepCTCFLoop PWM ID			*Consensus sequence of DeepCTCFLoop PWM*			*Consensus sequence of the known motif in JASPAR*	*Protein associated with the known motif*
GM12878	HeLa	K562	GM12878	HeLa	K562		
M138	M37	M88	AAAAAAA AAAAAT	AAAAAAAA AAAAA	GTTTAAAA AGAAA	TTTAAAAA AAAA	ZNF384
	M130		TTTAAGA GAAAAC				
	M151		TTTTTAAA AAAAA				
M107	M15	M111	AGGGGGA GGGGGG	GGAGGGA GGGGGG	GGGAGGG CAGGGG	GGAGGAGG AGGGGGA GGAGGA	ZNF263
—	M22	M77	—	TCCGCCG CTGGCG	CCACCAG GTGGCG	GCAGCAGC TGGCG	ASCL1
—	M152	M120	—	GCGCCCT GACCCC	TGCCCCT CCCCCC	GCCCCGC CCCC	SP1
		M57			CAGCCC TGCCTCC		
M14	M108	—	CCCCTCT GCCCAC	CGCGCCTG CGCCG	—	CCCACCT GCGC	ZEB1

The PWMs were learned from data of three cell types (GM12878, HeLa, and K562) and were significantly matched with the known DNA motifs of ZNF384, ZNF263, ASCL1, SP1, and ZEB1 using TomTom. The consensus sequences of the learned PWMs and the known DNA motifs, and the proteins associated with the known DNA motifs are shown.

PWM, position weight matrix.

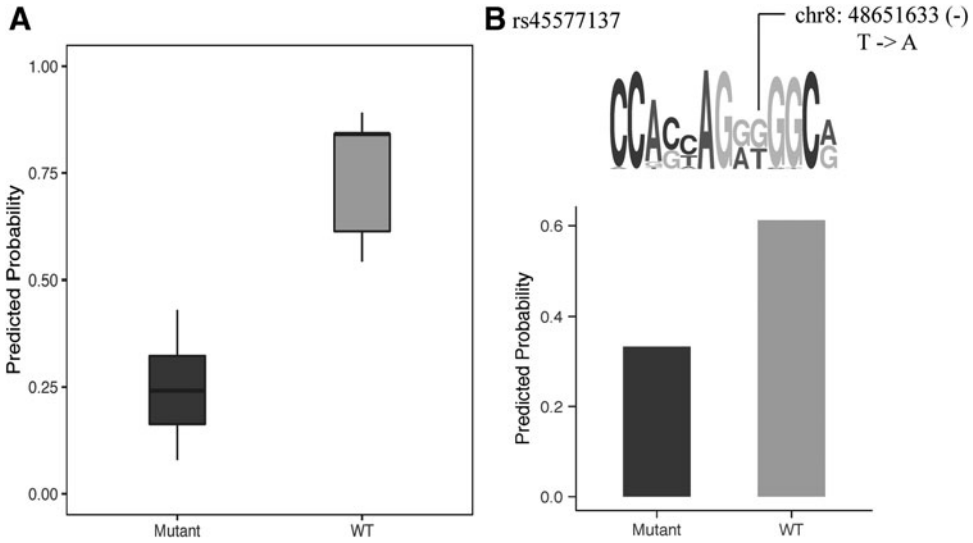

FIG. 6. Candidate sequence variants predicted by DeepCTCFLoop to affect chromatin loop formation. **(A)** Reduced probabilities of chromatin loop formation for nine disease-associated variants. **(B)** Effect of rs45577137 on CTCF-mediated chromatin loop formation. This variant is located within the second CTCF motif of a convergent motif pair for a chromatin loop on chromosome 8.

binding to distal enhancers (Raposo et al., 2015; Park et al., 2017; Aydin et al., 2019). The binding of ASCL1 to distal enhancers may facilitate chromatin loop formation via CTCF-mediated enhancer–promoter interactions (Ren et al., 2017). SP1 and ZEB1 have also been reported to be associated with chromatin regulation (Deshane et al., 2010; Aghdassi et al., 2012). Taken together, the results suggest that the binding of these TF proteins to specific DNA motifs may provide additional information for CTCF-mediated chromatin loop formation.

3.3. Application of DeepCTCFLoop to disease-associated sequence variants

Previous studies have shown that genetic mutations can cause human diseases by disrupting 3D genome organization (Zhang et al., 2012; Hnisz et al., 2016). As the majority of disease-associated SNPs are located in the noncoding regions with unknown mechanisms, we thus utilized DeepCTCFLoop to predict whether these sequence variants can affect CTCF-mediated chromatin loop formation. Particularly, when the GM12878 model was applied to the 524 disease-associated SNPs, 9 were predicted with high confidence to disrupt chromatin loop formation (Fig. 6A). One of the high-confident candidate variants was rs45577137, located within the second CTCF motif of a convergent motif pair for a chromatin loop on chromosome 8, and the nucleotide change T->A was predicted to significantly reduce the probability of chromatin loop formation (Fig. 6B). As rs45577137 resides in the regulatory region of CCCAT/enhancer-binding protein delta (CEBPD), which is activated in many inflammation-related diseases, such as Alzheimer's disease and cancer (Ko et al., 2015), we speculate that the disruption of insulated chromatin loops caused by rs45577137 may provide a possible mechanism for CEBPD activation in disease conditions. Taken together, DeepCTCFLoop provides a useful tool to analyze noncoding sequence variants for investigating potential chromatin loop disruption and pathogenic mechanisms.

4. CONCLUSIONS

In this study, we have developed a deep learning model, called DeepCTCFLoop, to predict whether a chromatin loop can be formed with a pair of convergent or tandem CTCF motifs and to discover the underlying sequence patterns in addition to the CTCF motif pair. The CTCF motifs and their flanking genomic sequences were taken as model input. When evaluated on three different cell types (GM12878, HeLa, and K562), DeepCTCFLoop showed superior performance and significantly outperformed a previous machine learning model, CTCF-MP, for sequence-based prediction of CTCF-mediated chromatin

loop formation. Interestingly, the DNA motifs of several TF proteins, including ZNF384, ZNF263, ASCL1, SP1, and ZEB1, were significantly matched with the PWMs learned by DeepCTCFLoop from data of GM12878, HeLa, and K562 cells, suggesting the potential roles of these DNA-binding proteins in CTCF-mediated chromatin loop formation. Notably, this role of ZNF384 was independently demonstrated by an experimental study (Whalen et al., 2016). We also utilized DeepCTCFLoop to analyze disease-associated sequence variants and identified some candidate variants predicted to disrupt CTCF-mediated chromatin loop formation. Therefore, the findings from this study not only provide useful information for understanding the mechanism of 3D genome organization but also help annotate and prioritize the noncoding sequence variants associated with human diseases.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

No external funding was received for this work.

REFERENCES

Aghdassi, A., Sendler, M., Guenther, A., et al. 2012. Recruitment of histone deacetylases HDAC1 and HDAC2 by the transcriptional repressor ZEB1 downregulates E-cadherin expression in pancreatic cancer. *Gut* 61, 439–448.

Alipanahi, B., Delong, A., Weirauch, M.T., et al. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33, 831–838.

Amaral, P.P., Leonardi, T., Han, N., et al. 2018. Genomic positional conservation identifies topological anchor point RNAs linked to developmental loci. *Genome Biol.* 19, 32.

Anania, C., and Lupiáñez, D.G. 2020. Order and disorder: Abnormal 3D chromatin organization in human disease. *Brief Funct. Genomics* 19, 128–138.

Angermueller, C., Lee, H.J., Reik, W., et al. 2017. DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* 18, 67.

Aydin, B., Kakumanu, A., Rossillo, M., et al. 2019. Proneural factors Ascl1 and Neurog2 contribute to neuronal subtype identities by establishing distinct chromatin landscapes. *Nat. Neurosci.* 22, 897–908.

Bailey, T.L., Boden, M., Buske, F.A., et al. 2009. MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208.

Bergstra, J., Yamins, D., and Cox, D.D. 2013. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. *12th Python Sci. Conf. (SCIPY 2013)* 13–20.

Bonev, B., and Cavalli, G. 2016. Organization and function of the 3D genome. *Nat. Rev. Genet.* 17, 661.

Chen, H., Gao, M., Zhang, Y., et al. 2019. Attention-based multi-NMF deep neural network with multimodality data for breast cancer prognosis model. *Biomed Res. Int.* 2019, 9523719.

Crooks, G.E., Hon, G., Chandonia, J.M., et al. 2004. WebLogo: A sequence logo generator. *Genome Res.* 14, 1188–1190.

Deshane, J., Kim, J., Bolisetty, S., et al. 2010. Sp1 regulates chromatin looping between an intronic enhancer and distal promoter of the human heme oxygenase-1 gene in renal cells. *J. Biol. Chem.* 285, 16476–16486.

Edgar, R. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210.

Feingold, E.A., Good, P.J., Guyer, M.S., et al. 2004. The ENCODE (ENCyclopedia of DNA Elements) project. *Science* 306, 636–640.

Fudenberg, G., Abdennur, N., Imakaev, M., et al. 2017. Emerging evidence of chromosome folding by loop extrusion. *Cold Spring Harb. Symp. Quant. Biol.* 82, 45–55.

Fudenberg, G., Imakaev, M., Lu, C., et al. 2016. Formation of chromosomal domains by loop extrusion. *Cell Rep.* 15, 2038–2049.

Fullwood, M.J., Liu, M.H., Pan, Y.F., et al. 2009. An oestrogen-receptor-α-bound human chromatin interactome. *Nature* 462, 58.

Grant, C.E., Bailey, T.L., and Noble, W.S. 2011. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* 27, 1017–1018.

Guo, Y., Xu, Q., Canzio, D., et al. 2015. CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* 162, 900–910.

Hnisz, D., Weintrau, A.S., Day, D.S., et al. 2016. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* 351, 1454–1458.

Hochreiter, S., and Schmidhuber, J. 1997a. Long short-term memory. *Neural Comput.* 9, 1735–1780.

Hochreiter, S., and Schmidhuber, J. 1997b. LSTM can solve hard long time lag problems, 473–479. In *Advances in Neural Information Processing Systems*. Eds: Mozer, M.C., Jordan, M.I., and Petsche, T. MIT Press. Cambridge, Massachusetts, USA.

Hsu, S.C., Gilgenast, T.G., Bartman, C.R., et al. 2017. The BET protein BRD2 cooperates with CTCF to enforce transcriptional and architectural boundaries. *Mol. Cell.* 66, 102–116.e7.

Ji, X., Dadon, D.B., Powell, B.E., et al. 2016. 3D chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell.* 18, 262–275.

Kelley, D.R., Snoek, J., and Rinn, J.L. 2016. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* 26, 990–999.

Khan, A., Fornes, O., Stigliani, A., et al. 2018. JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* 46, D260–D266.

Kingma, D.P., and Ba, J.L. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings*. San Diego, California, USA. arXiv: 1412.6980.

Ko, C.Y., Chang, W.C., and Wang, J.M. 2015. Biological roles of CCAAT/enhancer-binding protein delta during inflammation. *J. Biomed. Sci.* 22, 6.

Li, W., Wong, W.H., and Jiang, R. 2019. DeepTACT: Predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Res.* 47, e60.

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293.

Mathelier, A., Zhao, X., Zhang, A.W., et al. 2014. JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 42, D142–D147.

Mikolov, T., Chen, K., Corrado, G., et al. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013—Workshop Track Proceedings*. arXiv: 1301.3781. Scottsdale, Arizona, USA.

Nora, E.P., Goloborodko, A., Valton, A.L., et al. 2017. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* 169, 930–944.e22.

Norton, H.K., and Phillips-Cremins, J.E. 2017. Crossed wires: 3D genome misfolding in human disease. *J. Cell Biol.* 216, 3441–3452.

Park, N.I., Guilhamon, P., Desai, K., et al. 2017. ASCL1 reorganizes chromatin to direct neuronal fate and suppress tumorigenicity of glioblastoma stem cells. *Cell Stem Cell* 21, 209–224.

Quang, D., and Xie, X. 2016. DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* 11, e107.

Rao, S.S.P., Huang, S.C., Glenn St Hilaire, B., et al. 2017. Cohesin loss eliminates all loop domains. *Cell* 171, 305–320.e24.

Rao, S.S.P., Huntley, M.H., Durand, N.C., et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159, 1665–1680.

Raposo, A.A.S.F., Vasconcelos, F.F., Drechsel, D., et al. 2015. Ascl1 coordinately regulates gene expression and the chromatin landscape during neurogenesis. *Cell Rep.* 10, 1544–1556.

Ren, G., Jin, W., Cui, K., et al. 2017. CTCF-mediated enhancer-promoter interaction is a critical regulator of cell-to-cell variation of gene expression. *Mol. Cell* 67, 1049–1058.e1046.

Sanborn, A.L., Rao, S.S.P., Huang, S.C., 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–E6465.

Tang, Z., Luo, O.J., Li, X., et al. 2015. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* 163, 1611–1627.

Trabelsi, A., Chaabane, M., and Ben-Hur, A. 2019. Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities. *Bioinformatics* 14, i269–i277.

Trieu, T., Martinez-Fundichely, A., and Khurana, E. 2020. DeepMILO: A deep learning approach to predict the impact of non-coding sequence variants on 3D chromatin structure. *Genome Biol.* 21, 79.

Welter, D., MacArthur, J., Morales, J., et al. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42, D1001–D1006.

Wendt, K.S., Yoshida, K., Itoh, T., et al. 2008. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* 451, 796–801.

Whalen, S., Truty, R.M., and Pollard, K.S. 2016. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* 48, 488.

Wutz, G., Várnai, C., Nagasaka, K., et al. 2017. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.* 36, 3573–3599.

Zhang, R., Wang, Y., Yang, Y., et al. 2018. Predicting CTCF-mediated chromatin loops using CTCF-MP. *Bioinformatics* 34, i133–i141.

Zhang, X., Cowper-Sal-lari, R., Bailey, S.D., et al. 2012. Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. *Genome Res.* 22, 1437–1446.

Zhou, P., Shi, W., Tian, J., et al. 2016. Attention-based bidirectional long short-term memory networks for relation classification, 207–212. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016—Short Papers*. Berlin, Germany.

Address correspondence to:
Dr. Liangjiang Wang
Department of Genetics and Biochemistry
Clemson University
Clemson, SC 29634
USA

E-mail: liangjw@clemson.edu