Quantitative Biology 2023, 11(3): 275–286
https://doi.org/10.15302/J-QB-022-0316

RESEARCH ARTICLE

DeepRCI: predicting RNA-chromatin interactions via deep learning with multi-omics data

Yuanpeng Xiong[1,2], Xuan He[3], Dan Zhao[3], Tao Jiang[1,2,4,*], Jianyang Zeng[3,*]

[1] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2] BNRIST, Tsinghua University, Beijing 100084, China
[3] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China
[4] Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA
* Correspondence: jiang@cs.ucr.edu; zengjy321@tsinghua.edu.cn

Received November 23, 2022; Revised December 30, 2022; Accepted December 30, 2022

***Background*: Chromatin-associated RNA (caRNA) acts as a ubiquitous epigenetic layer in eukaryotes, and has been reported to be essential in various biological processes, including gene transcription, chromatin remodeling and cellular differentiation. Recently, numerous experimental techniques have been developed to characterize genome-wide RNA-chromatin interactions to understand their underlying biological functions. However, these experimental methods are generally expensive, time-consuming, and limited in identifying all potential sites, while most of the existing computational methods are restricted to detecting only specific types of RNAs interacting with chromatin.**

***Methods*: Here, we propose a highly interpretable computational framework, named DeepRCI, to identify the interactions between various types of RNAs and chromatin. In this framework, we introduce a novel deep learning component called variformer and integrate multi-omics data to capture intrinsic genomic features at both RNA and DNA levels.**

***Results*: Extensive experiments demonstrate that DeepRCI can detect RNA-chromatin interactions more accurately when compared to the state-of-the-art baseline prediction methods. Furthermore, the sequence features extracted by DeepRCI can be well matched to known critical gene regulatory components, indicating that our model can provide useful biological insights into understanding the underlying mechanisms of RNA-chromatin interactions. In addition, based on the prediction results, we further delineate the relationships between RNA-chromatin interactions and cellular functions, including gene expression and the modulation of cell states.**

***Conclusions*: In summary, DeepRCI can serve as a useful tool for characterizing RNA-chromatin interactions and studying the underlying gene regulatory code.**

Keywords: deep learning; multi-omics data; RNA-chromatin

Author summary: Chromatin-associated RNA (caRNA) acts as a ubiquitous epigenetic layer in eukaryotes, and has been reported to be essential in various biological processes. Here, we propose a highly interpretable computational framework, named DeepRCI, to identify the interactions between various types of RNAs and chromatin. DeepRCI can serve as a useful tool for characterizing RNA-chromatin interactions and studying the underlying gene regulatory code.

INTRODUCTION

The genomes of multicellular eukaryotes are mostly transcribed into protein-coding RNAs and noncoding RNAs [1,2]. Parts of these RNAs physically interact with chromatin and are thus termed chromatin-associated RNAs (caRNAs) [3,4]. The interactions between RNAs and chromatin can be generally

categorized into two groups, including cis-acting RNAs that interact with nearby genes and trans-acting RNAs that interact with distant genes on a genome-wide scale [5,6]. An increasing number of studies have shown that caRNAs can be involved in various essential biological processes [7–9]. For example, Kalwa *et al*. found that HOTAIR regulates adipogenic differentiation via triplex formation with PCDH7 and HOXB2 [10], Wang *et al*. reported that MIR100HG regulates p27 via triplex formation [11], and O'Leary *et al*. stated that PARTICLE binds to the MAT2A promoter CpG island as a triplex to contribute to the construction of gene-silencing machineries [12]. To identify the interaction profiles of various types of caRNAs at genomic loci, several experimental technologies have been developed, including capture hybridization analysis of RNA targets (CHART) [13], chromatin isolation by RNA purification (ChIRP) [14], and RNA antisense purification (RAP) [15]. These methods can be applied to detect the genome-wide chromatin interactions of a specific RNA. However, these "one-RNA-versus-the-genome" technologies cannot be applied to discover novel RNAs that interact with chromatin [16]. To address this limitation, MARGI (mapping RNA-genome interactions) [17] and its derivative, iMARGI [16], have been developed to characterize the interactions between chromatin and thousands of RNAs, including mRNAs and ncRNAs. Moreover, several high-throughput sequencing technologies, including GRID-seq (global RNA interactions with DNA by deep sequencing) [18] and RADICL-seq (RNA And DNA Interacting Complexes Ligated and sequenced) [4], have also been proposed to extract the RNA-chromatin interaction profiles of different cell types [18]. However, these experimental methods are generally expensive and time-consuming, and may miss a certain amount of interactions due to the bias caused by read mapping and employed probes.

In recent years, a number of rule-based statistical methods have been developed to predict the chromatin interacting potentials of the most widely distributed lncRNAs. For example, Kuo *et al*. proposed Triplex Domain Finder (TDF) to detect the DNA-binding domains of lncRNAs that mainly utilized statistical tests to evaluate the potentials of the triple-helix formations of multiple lncRNAs [19]. Buske *et al*. designed Triplex-Inspector to select the sequence-specific ligands and targets of lncRNAs with the corresponding gene locations and genomic architectures [20]. He *et al*. proposed LongTarget that utilized the non-canonical rules to detect motifs and binding sites in forming triplex [21]. In addition to these statistical methods, Zhang *et al*. proposed a deep learning based framework, named TriplexFPP, to detect the RNA-DNA triplex forming potentials [22], which applied a convolution neural network (CNN) to capture the spatial features of RNA/DNA sequences. This deep learning method outperformed the existing statistical methods but is still quite limited in accuracy, interpretability, and generalizability. Most of the aforementioned computational models were built only for specific types of RNAs, which means that they cannot be directly extended to other types of RNAs. In addition, although biological sequences such as RNA sequences generally have variable lengths, most of the existing deep learning models [23–25] are not designed to represent features naturally in such a scenario, and hence the input sequences fed into these models are often padded to a fixed length, which might introduce additional bias [26] and require extra calculation. Hence, processing sequences of variable lengths in a more flexible way may provide a more efficient model with better feature learning capability. Moreover, most of the existing prediction frameworks in this area are only designed for processing raw sequences and ignore the rich information available from multi-omics data. Previous studies have shown that chromatin conformation and chromatin accessibility are often associated with RNA-chromatin interactions [27–29]. Therefore, integrating multi-omics data may provide new useful insights into understanding the biological roles of RNA-chromatin interactions.

To overcome the limitations of the existing RNA-chromatin interaction prediction methods and decipher the biological functions of caRNAs, we propose a deep learning based framework, called DeepRCI (deep learning based RNA-chromatin interaction predictor), which consists of a novel deep learning layer, named variformer, to automatically handle the input sequences of variable lengths. To further improve the predictive capability and interpretability of the model, we integrate multi-omics data related to RNA-chromatin interactions using a Bayesian network, which modeled the underlying biological mechanisms of RNA-chromatin interactions [27–29] and further improved the generalizability of the proposed deep learning model (Fig. 1). We have evaluated our method on real data from the literature and showed that our model greatly outperformed the state-of-the-art baseline prediction methods, with at least 5% higher performance in terms of the area under the react operator curve (AUROC). Moreover, our model is highly interpretable and thus can help understand the underlying functions of the caRNAs. In particular, the sequence features captured by our model can be well matched to the known motifs in the CIS-BP motif database, and are also supported by

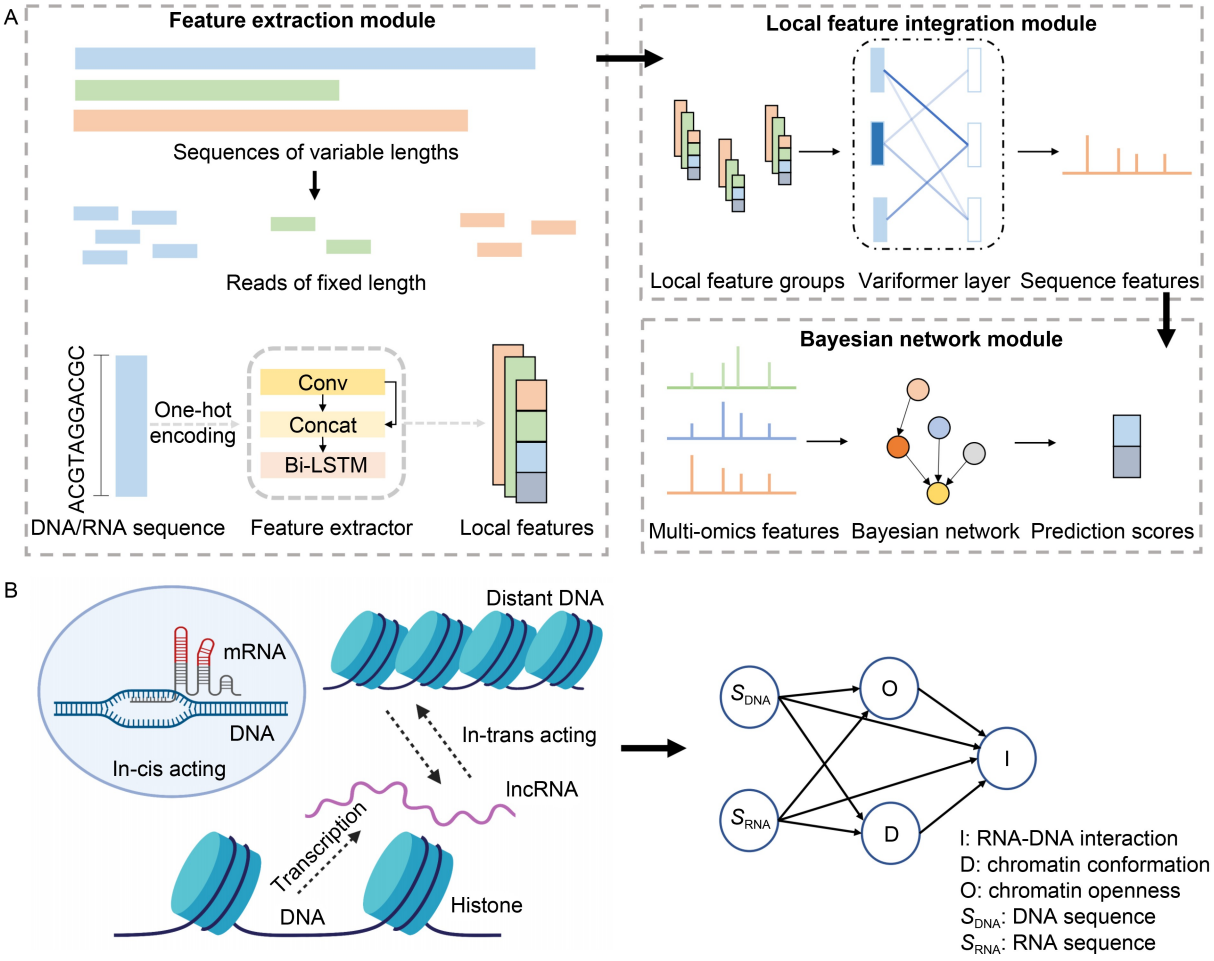

Figure 1. Overview of DeepRCI. (A) Schematic illustration of the pipeline. Sequences of RNAs and DNAs are first cut into short reads of a fixed length and then fed into a feature extractor to extract the local features. The feature extractor is composed of three basic operations. Conv: convolution layer, Concat: concatenation layer and Bi-LSTM: bi-directional long short term memory units. Then, the local features are fed into a local feature integration module, which mainly contains a so-called variformer layer, a novel deep learning layer used for extracting both local and global features from the sequences of variable lengths (see materials and methods for more details). Finally, the multi-omics features, including Hi-C data, chromatin open accessibility data, and the previously extracted sequence features are integrated into a Bayesian network module to predict the potential of RNA-chromatin interactions. (B) Illustration of the Bayesian network. A caRNA can not only form a triplex structure with nearby DNAs (in-cis acting) but also interact with distant DNAs (in-trans acting), which implies that RNA-chromatin interactions are closely related to chromatin conformation and chromatin openness. Based on this observation, we introduce a Bayesian network to model the underlying biological mechanism of RNA-chromatin interactions.

some experimental evidence in the literature. Downstream analyses revealed that DeepRCI can be applied to characterize some essential biological factors or processes associated with RNA-chromatin interactions, including gene expression and the modulation of cell states. To the best of our knowledge, DeepRCI is the first model that integrates multi-omics data to identify the interactions between various types of RNAs and chromatin. Our comprehensive tests demonstrated that DeepRCI can provide useful biological insights into the further understanding of the functional roles of RNA-chromatin interactions in gene regulation.

RESULTS

DeepRCI accurately predicted the interactions between RNA and chromatin

We first assessed the prediction performance of DeepRCI only on the sequence data (denoted as DeepRCI-seq) from the zhang2020 dataset [22], and compared DeepRCI with the state-of-the-art baseline prediction methods, including TriplexFPP [22] and modified version of DeepRCI-Seq where the variformer layer is replaced by traditional convolution layers for

sequence feature extraction (denoted as DeepConv). The test result shows that DeepRCI-seq achieved the best performance in terms of four metrics, including accuracy, area under the receiver operating characteristic (AUROC), area under the precision-recall curve (AUPRC), and F1-score (Fig. 2A).

In particular, our model achieved accuracy and F1-scores of 0.835 and 0.813, respectively, which were 7.5% and 14.2% higher than those of TriplexFPP, and 2.1%, and 2.25% higher than those of DeepConv, respectively. On average, the AUROC scores and AUPRC scores were also improved by nearly 5% compared to TriplexFPP or DeepConv. These test results on sequence data demonstrated that our model can predict RNA-chromatin interactions more accurately than the state-of-the-art baseline methods, and the proposed varifomer layer also greatly contributed to its performance and outperformed the traditional convolution layers. Next, we evaluated the prediction performance of DeepRCI on the stress20 dataset [2], which contained not only sequence data but also multi-omics data, including Hi-C and ATAC-seq data. Consequently, DeepRCI trained on multi-omics data achieved an AUROC score of 0.940 and an AUPRC score of 0.932, which were 12% and 10.3% higher than those of TripexFPP, respectively (Fig. 2B). We also assessed the performance of our model under different test settings, including the models trained using only sequence information or with the integration of multi-omics data excluding the Hi-C data (denoted as DeepRCI-atac). We also found that our model still outperformed the the-start-of-the-art methods significantly on radicl20 and grid17 datasets (Supplementary Notes and Fig. S1). These test results demonstrated that integrating multi-omics data can greatly improve the RNA-chromatin interaction prediction results (Fig. 2B).

To further explore the generalizability of our model on imbalanced data, we also compared DeepRCI which integrates multi-omics data through the Bayesian network with another model that simply concatenates multi-omics features (denoted as DeepRCI-simple) on two subsets of the stress20 dataset, where one is balanced (ratio of positive:negative samples = 1:1) while the other imbalanced (ratio of positive:negative samples = 1:10). DeepRCI tested on the balanced dataset achieved an AUROC of 0.943 and an AUPRC of 0.938, which were 1.5% and 1.4% higher than those of DeepRCI-simple, respectively (Fig. 3A). Unexpectedly, the superior performance of DeepRCI on the imbalanced data is much more evident than that on the balanced dataset (Fig. 3B). In particular, DeepRCI achieved an AUROC score of 0.925 and an AUPRC score of 0.576, which were 6% and 23.7% higher than those of DeepRCI-simple, respectively. These test results on

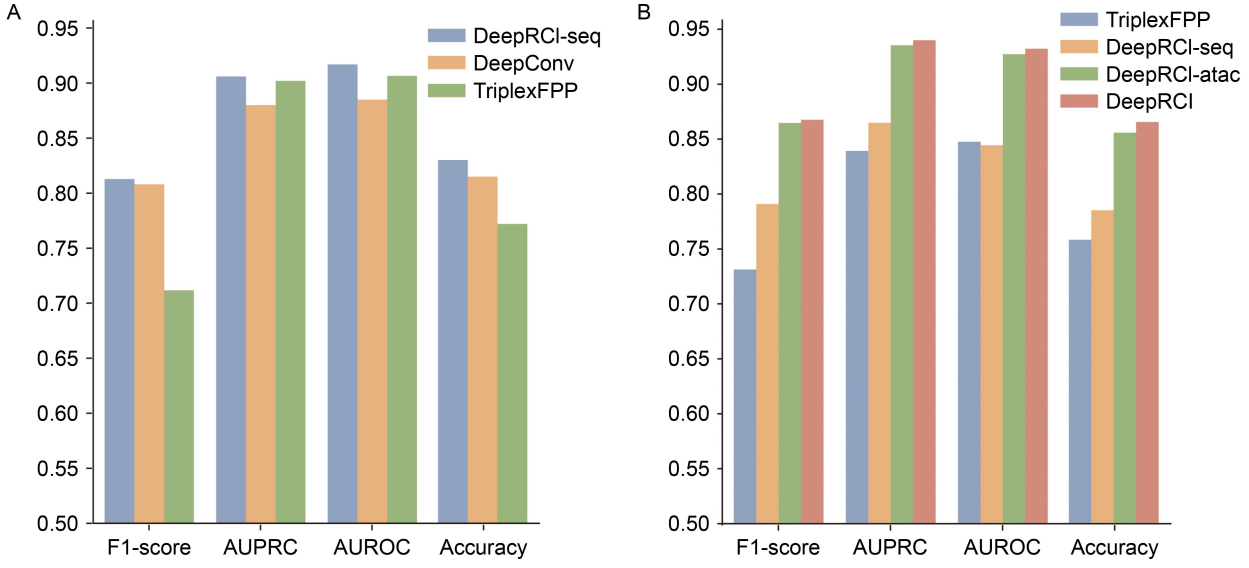

Figure 2. The performance of different models on lncRNAs from the zhang2020 dataset and multi-omics data from the stress20 dataset. (A) Performance comparison on the zhang2020 dataset among DeepRCI-seq, DeepConv and TriplexFPP, measured in terms of accuracy, F1-score, AUROC score, and AUPRC score. DeepRCI-seq represents the DeepRCI model trained with only sequence data. DeepConv represents the model using the traditional convolution layer rather than the proposed variformer layer for sequence feature extraction while keeping the other parts the same as in DeepRCI-seq. (B) Performance comparison on the stress20 dataset among DeepRCI, DeepRCI-seq, DeepRCI-atac, and TriplexFPP, measured in terms of accuracy, F1-score, AUROC score, and AUPRC score. DeepRCI-atac represents our model trained using multi-omics information except Hi-C data.

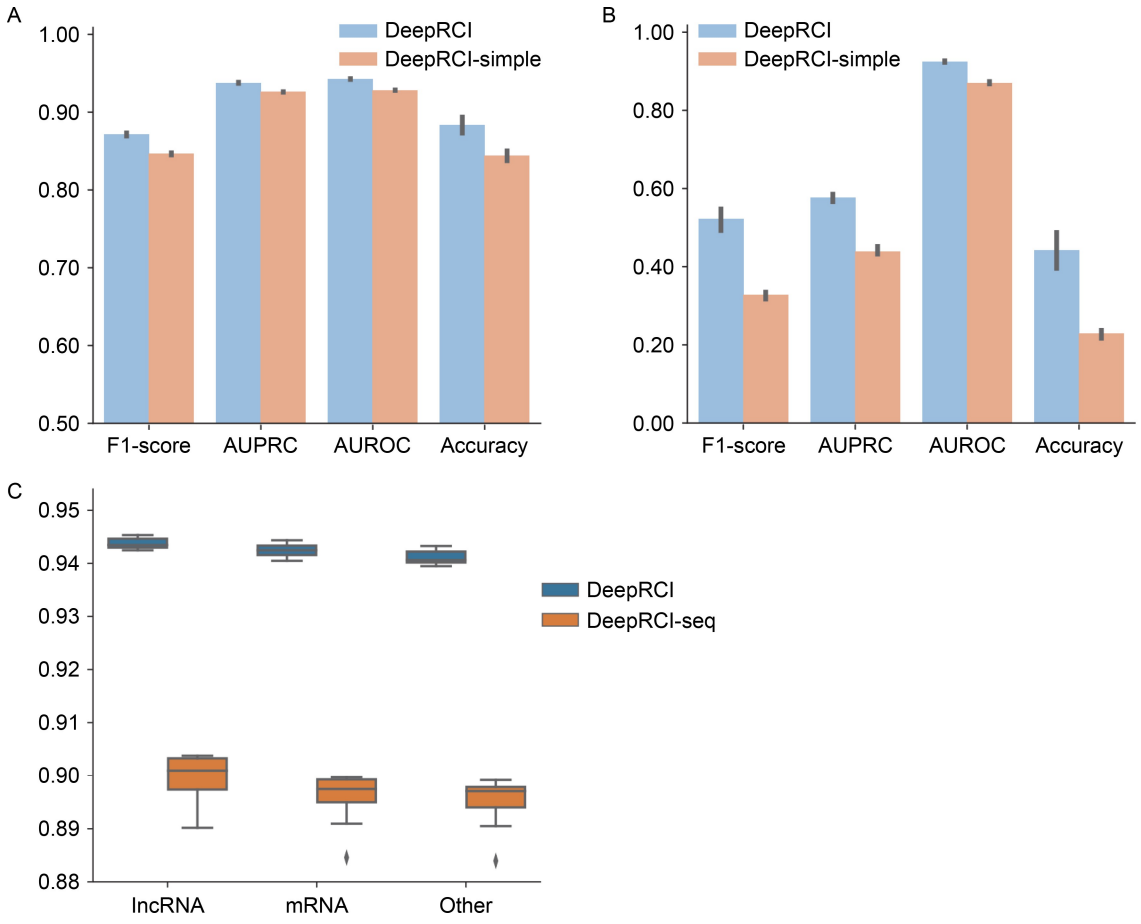

Figure 3. Ablation studies on the second dataset (*i.e.*, zhang2020). (A,B) Performance comparison on a balanced dataset (ratio of positive:negative samples=1:1) and an imbalanced dataset (ratio of positive:negative samples=1:10) between the model that integrated multi-omics data under the Bayesian framework (denoted as DeepRCI) and the model that incorporated multi-omics data through simple concatenation (denoted as DeepRCI-simple), measured in terms of F1-score, AUPRC, AUROC, and accuracy. (C) Performance comparison between the model trained with multi-omics data and the model trained only with sequence, measured in terms of AUROC score.

imbalanced data showed that our model can be applied potentially to a broad range of imbalanced settings to make high-quality predictions on RNA-chromatin interactions. After our model was validated, we further evaluated the prediction performance of our model on different types of RNAs in the stress20 dataset. More specifically, we applied DeepRCI and DeepRCI-seq to predict interactions of three types of RNAs contained in the dataset and found that DeepRCI achieved an average of 5% improvement in AUROC scores over DeepRCI-seq on all three types of RNAs (Fig. 3C). In addition, we observed that the prediction performance of DeepRCI on lncRNAs was significantly higher than that on other types of RNAs (P=8.8×10^{-4}, Wilcoxon rank-sum test). These test results showed that multi-omics data can largely contribute to the prediction performance on many types of RNAs.

DeepRCI provided a protein-binding prospective for understanding RNA-chromatin interactions

To interpret our deep learning model, we visualized and analyzed the sequence features captured by our model following the strategies described in [30,31]. More specifically, we first extracted the sequence features of 128 convolution kernels (each with a length of 18 bp) in the first convolution layer of the model and collected the sequences that activate the convolution filter to a value higher than the average level. These sequences were then aggregated together to generate positional weight matrices (PWMs) as the local sequence motifs in the original input sequences. Considering that DNAs and RNAs may contain different signals, we analyzed the sequence patterns of DNAs and RNAs separately. Next, these derived sequence features were mapped to the known binding motifs of DNA binding proteins and

RNA binding proteins obtained from the CIS-BP and CISBP-RNA databases [32] with TOMTOM [33], respectively. As expected, most of these sequences can be mapped to specific motifs in the databases (see Fig. 4 and Supplementary Tables S1 and S2). These analysis results indicated that our model can capture the meaningful biological signals of RNA-chromatin interactions to a certain extent. For example, the promoter region of the MYCN gene, whose motif was captured by our model, has been previously reported to be closely related to the natural lncRNA interacting sites [34].

DeepRCI can be applied to characterize cellular functions

We further analyzed the associations between RNA-chromatin interactions and cellular functions, including gene expression, the modulation of cell states and the formation of fusion genes (see Supplementary Notes S1 and Table S3), based on the prediction results of DeepRCI. More highly expressed genes tend to share a higher RNA-chromatin interaction ratio. It has been previously shown that noncoding RNAs, especially long noncoding RNAs [35,36], are important regulatory elements that regulate gene expression through various interactions, including RNA-chromatin, RNA-RNA and RNA-protein interactions [35,37,38]. To further investigate the relationships between RNA-chromatin interactions and gene expression, we first collected the gene expression data of HUEVCs and then calculated the intensities of the corresponding RNA-chromatin interactions predicted by DeepRCI for individual genes (see Supplementary Notes S2 for more details). As a result, genes with higher expression levels tend to have stronger RNA-chromatin interaction signals (Fig. 5A, Pearson correlation = 0.483, $P = 1.77 \times 10^{-302}$, two-tailed *t*-test). Considering that iMARGI-seq can generally exclude the influence of *in situ* mRNAs [16], this result effectively suggested a high correlation between noncoding RNAs and gene expression, which can also be supported by the previous studies [39–41]. In addition, we compared the RNA-chromatin interaction signals between highly expressed genes (top 25%) and lowly expressed genes (bottom 25%) on each euchromosome. The results showed that highly expressed genes tend to show high interaction intensities on individual euchromosomes (Fig. 5B and Supplementary Fig. S2). This observation is also consistent with a previous study, which showed that a set of 21 genes in

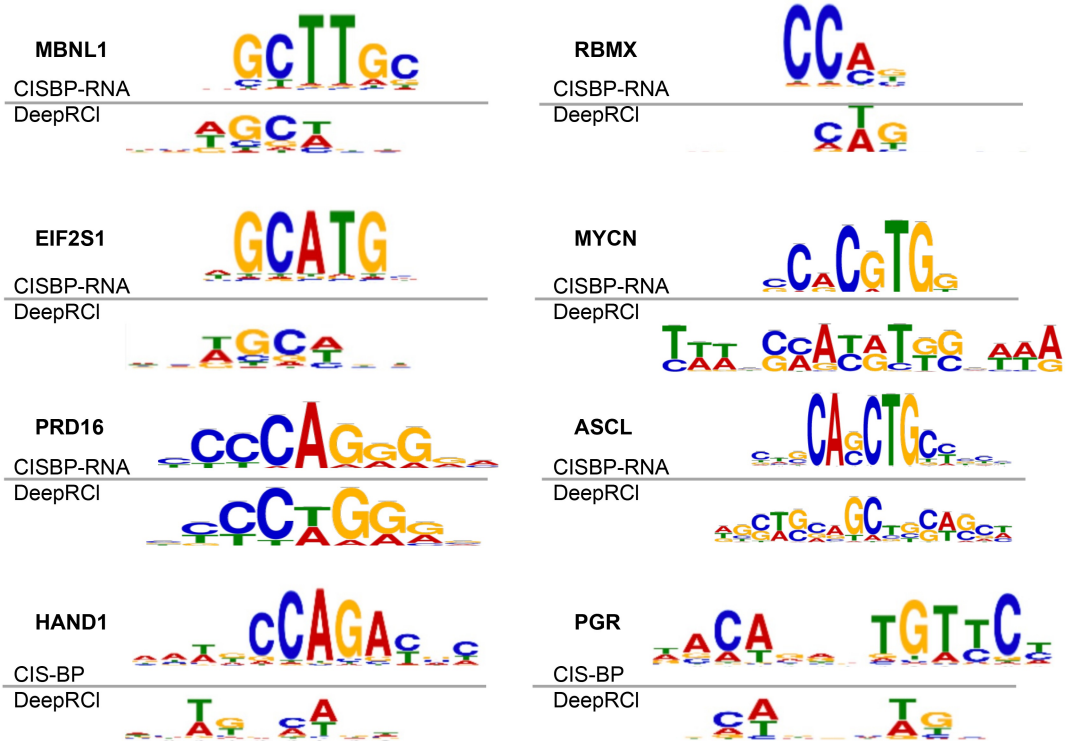

Figure 4. Local sequence motifs captured by DeepRCI for RNA-chromatin interactions. The motifs derived from the CIS-BP/CISBP-RNA database [28] and extracted by DeepRCI are shown above and below each line, respectively. The first three rows were generated from RNA sequences, while the fourth row was generated from DNA sequences. The promoter of gene MYCN has been reported to interact with a natural noncoding RNA [34].

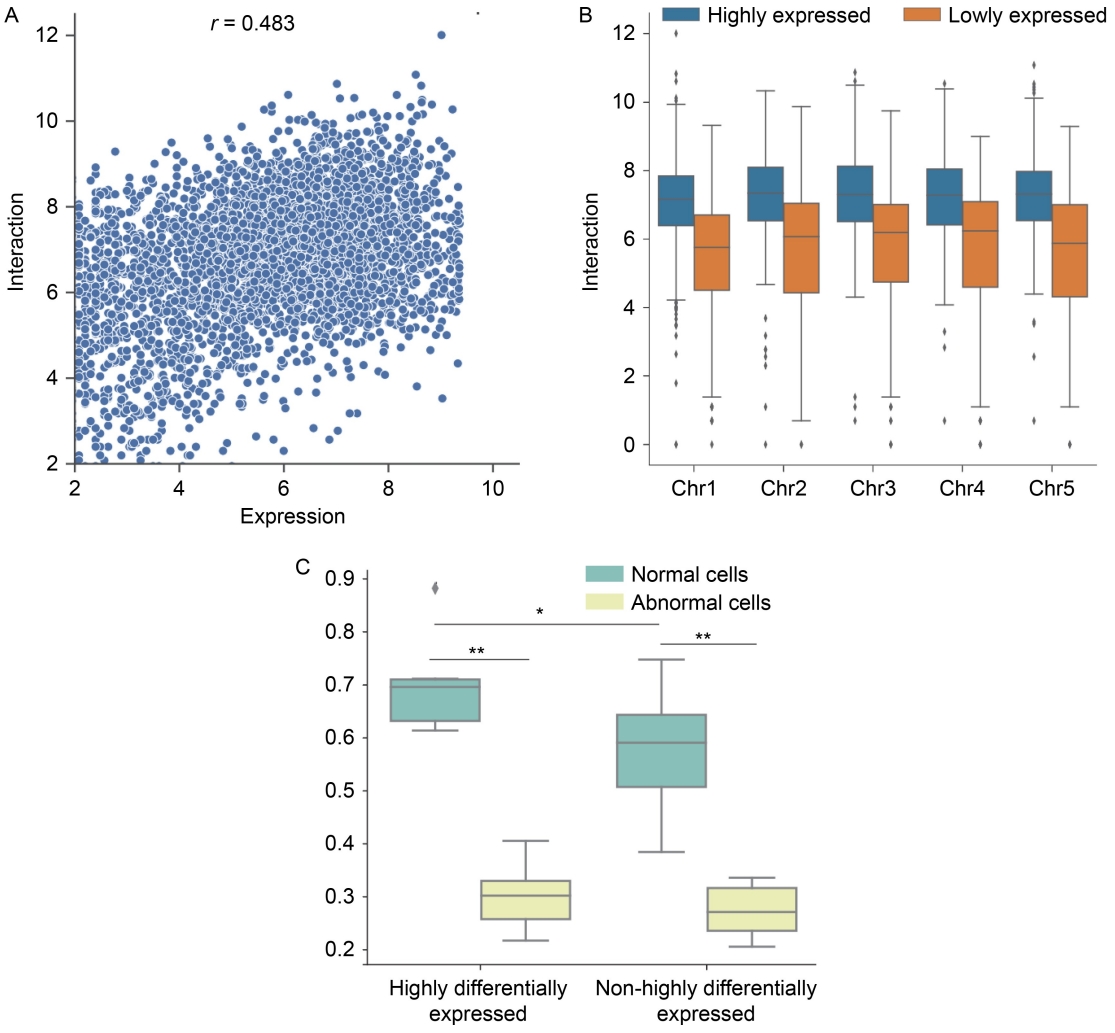

Figure 5. DeepRCI captured intrinsic correlations between RNA-chromatin interactions and cellular functions. (A) The scatter plot of gene expression levels and the corresponding interaction intensities (transformed by log2) for individual genes. Highly expressed genes tend to have high interaction intensities (Pearson correlation=0.483, $P = 1.77 \times 10^{-302}$, two-tailed *t*-test). (B) The box plot of RNA-chromatin interaction intensities (transformed by log2) between highly expressed genes (top 25%) and lowly expressed genes (bottom 25%) on euchromosomes. More highly expressed genes tend to have higher interaction intensities ($P = 1.23 \times 10^{-90}$, 1.82×10^{-50}, 8.06×10^{-39}, 4.89×10^{-22} , 1.88×10^{-37} from left to right, Wilcoxon rank-sum test). Results on more chromosomes can be found in Supplementary Fig. S2. (C) Comparison of the predicted interaction ratios (normalized interaction intensities) between highly differentially expressed genes (top 25 genes according to the log fold-changes) and non-highly differentially expressed genes (randomly sampled from the set of remaining genes) in normal and abnormal HUEVCs (*$P < 0.05$, **$P < 0.01$, Wilcoxon rank-sum test), respectively.

HUEVCs exhibited high co-expression with the LINC00607 RNA (a lncRNA interacting with chromatin) [2]. This further verified that DeepRCI can effectively capture the potential features of RNA-chromatin contacts. Global changes in RNA-chromatin interactions are associated with the modulation of cell state. To investigate whether there existed other types of cellular activities related to RNA-chromatin interactions, we compared and analyzed the gene expression data and predicted RNA-chromatin interaction profiles between normal and abnormal HUEVCs. More specifically, we first clustered the gene expression data (Supplementary Fig. S3A), and then collected the top 25 genes (marked as highly differentially expressed genes) according to their log fold-changes (Supplementary Fig. S3B). Next, we randomly selected the same number of genes from the set of the remaining genes (marked as non-highly differentially expressed genes) and then calculated the interaction ratios (normalized interaction intensities) corresponding to these two groups of genes. Ultimately, we found that the interaction ratios of highly differentially expressed genes in normal cells are

significantly higher than those in abnormal cells ($P = 3.9\times10^{-3}$, Wilcoxon rank-sum test), and such a phenomenon was also observed on non-highly differentially expressed genes (Fig. 5C, $P = 3.9 \times 10^{-3}$, Wilcoxon rank-sum test). Moreover, we found that the differences between highly differentially expressed genes and non-highly differentially expressed genes were significant in normal cells ($P < 0.05$, Wilcoxon rank-sum test) compared to abnormal cells ($P > 0.05$, Wilcoxon rank-sum test). These results indicated that the modulation of cell states might be accompanied by the change of the corresponding RNA-chromatin interactions, which can also be supported by the literature [2].

DISCUSSION

In this paper, we proposed a deep learning based framework for the prediction of RNA-chromatin interactions. To the best of our knowledge, our work is the first attempt to model the chromatin interactions of different types of RNAs. A novel layer, named variformer, was introduced in this framework to handle the sequences of variable lengths, which can capture long-range relations between distant sites along the input sequences through a multi-head attention mechanism. Though several computational methods have already been proposed for RNA-chromatin interaction prediction in recent years [19,22,42], most of them are restricted to predicting the interaction profiles of specific types of RNAs and insufficient in model interpretability. Here, our model provides a global view of sequence features, which can help interpret the deep learning model from a protein-binding perspective for understanding the RNA-chromatin interactions. Moreover, since our model integrates multi-omics data in a biologically reasonable scheme with a Bayesian network, it can capture more reliable features than models trained using only sequences and also alleviate the potential overfitting issue, especially on imbalanced datasets. In the downstream analyses, we further investigated the relationships between RNA-chromatin interactions and gene expression and gene activities. The analyses based on the prediction results of our model were highly consistent with the conclusions drawn from the previous studies, which thus suggested the potential applications of our model in studying RNA-chromatin interactions related biological problems. Note that the data used in this paper for calculating the associations between RNA-chromatin interactions and other variables were mainly limited to specific cell lines. Thus, further experiments to control relevant variables are still needed. In addition, the current version of our deep learning based model only considered single species, which may prevent one from uncovering the similarities and differences of RNA-chromatin interactions among different species. Therefore, integrating data from more species will be regarded as an important direction of our future work. As the investigation of RNA-chromatin interactions is attracting more and more research interests [13,14,43], we believe that our DeepRCI framework together with the improved experimental techniques [15,43] will provide more reliable insights into the studies of RNA-chromatin interactions.

MATERIALS AND METHODS

Datasets

To train and evaluate our model, we collected four benchmark datasets, including zhang2020 [22], stress20 [2], grid17 [18] and radicl20 [4]. The zhang2020 dataset was adopted from [22] and was mainly used for predicting the forming potentials of DNA:RNA triplexes. This dataset contains only long noncoding RNAs (lncRNAs) that can form triplexes with corresponding DNA sequences from chromatin. The stress20 dataset was derived from [2], and contains the interaction information of various types of RNAs, including cis- and trans-acting RNAs with their corresponding DNAs in human umbilical vein endothelial cells (HUVECs). The interaction read pairs of RNAs and DNAs in stress20 were downloaded from the GEO database (accession number GSM4006840). The Hi-C data and the ATAC-seq data were also integrated into the second dataset (accession numbers GSM4006837 and GSM4006837 from the GEO database, respectively), although we cannot find associated H-C and ATAC-seq data for the first dataset. More details for the remaining two datasets can be found in Supplementary Notes S3. In the downstream analyses, the gene expression data of HUVECs were collected from the GEO database (accession number GSM4006843). The gene expression data and RNA-chromatin interaction profiles of abnormal HUEVCs, induced with “high glucose and tumor necrosis factor α” (denoted as H + Tα), were curated from [2] (accession numbers GSM4006842 and GSM4006839 from the GEO database, respectively).

Data preprocessing

For the zhang2020 dataset, we removed lncRNAs with lengths greater than 5 kb due to the long-tailed distribution of sequence lengths (Supplementary Fig. S4). In the end, 35,327 negative samples and 517

positive samples were retained. For the remaining datasets, we first normalized the Hi-C data using HicNorm [44], and calculated the interaction frequencies of individual genomic loci at a resolution of 100 kb following the same protocol as in [2]. Then, we binned the ATAC-seq data at a resolution of 100 kb, and summed up the intensities of the detected peaks in the same bin. These two aforementioned features were collected as epigenomics features. Next, we mapped all short reads of the interaction pairs obtained by iMARGI [16] to the hg38 reference genome according to the corresponding genomic coordinates and strand types, and then collected these read pairs as sequence data of positive samples. Considering that there are a large number of RNAs that do not interact with chromatin in the whole genome, we adopted the following strategy to generate raw pairs of sequences for negative samples. More specifically, we converted the coordinates of interaction pairs obtained by iMARGI into the indexes of a sparse matrix M with the same resolution as for the Hi-C data. Here, each row represents a 100 kb fragment from all RNA sequences and each column represents a 100 kb fragment from all DNA sequences. Each zero elements M_{ij} of the matrix indicates that the i_{th} RNA fragment and the j_{th} DNA fragment has a high probability of being noninteracting. To avoid the model solely learning whether the DNA or RNA sequences have appeared in the training data or not, we adopted a strict sampling strategy. More specifically, for each RNA sequence R_i from positive samples, we randomly sampled a DNA sequence D_j from genomic loci that did not interact with the given R_i. Then, all the sequence pairs (R, D) were collected as negative samples. We also adopted an "edge swapping" strategy to generate negative samples (more details can be found in Supplementary Notes S4 and Fig. S5). To further reduce the redundancy in the sequence samples collected above, we employed the CD-HIT-EST [45] tool to remove samples that have a certain similarity (with a similarity score above 80%). In the end, we obtained 19,863 positive samples and 19,923 negative samples for the stress20 dataset.

Overview of DeepRCI

Our DeepRCI model is composed of three modules, including feature extraction, location feature integration, and a Bayesian network module (Fig. 1). The input sequences are first divided into short reads of a fixed length (see Suplementary Notes S5 and Fig. S6 for more details) and then encoded into binary vectors based on the one-hot encoding strategy [1,31]. The feature extractors shared between DNAs and RNAs are composed of convolution layers, concatenation layers, and bidirectional long short-term memory (LSTM) layers, which are designed for extracting local features from nucleotide sequences (Fig. 1A). The local features of short reads are first gathered into a group and then fed into a so-called variformer layer (see the next section for more details) to extract both local and global features of original input sequences. Unlike traditional transformers that handle sequences through padding and masking operations [46,47], variformer treats sequences as groups of local features, and then applies an attention mechanism [47] over each group to generate feature vectors of fixed sizes (Fig. 1B). Finally, the chromatin conformation and chromatin accessibility data are integrated with the previously extracted sequence features into a Bayesian network module to predict the potential of RNA-chromatin interactions. The hyperparamters of DeepRCI were determined by a line-search strategy (Suplementary Notes and Table S4).

The variformer layer

To deal with the input sequences of variable lengths, we propose a novel layer, called variformer. Specifically, all sequences of various lengths (denoted as $S \in \boldsymbol{R}^{L\times d}$, where L stands for the length of S and d stands for the feature dimension of each site) are first cut into short reads of a fixed length l (denoted as $s \in \boldsymbol{R}^{l\times d}$), and then a feature extraction layer shared among these reads is applied to each read to obtain the corresponding local features, which can be represented as a combination of several basic operations, that is,

$$f_{conv} = Pool(BN(Conv(s))), \tag{1}$$

$$f_{LSTM} = LSTM(f_{conv}), \tag{2}$$

$$c = Concat([f_{LSTM}, f_{conv}]), \tag{3}$$

where *Pool* stands for the max-pooling operation [36], *BN* stands for the batch normalization operation [48], *Concat* stands for the concatenation operation, c stands for the local features, and f_{conv} , f_{LSTM} stand for the features extracted by convolution and *LSTM* layers, respectively. These local features are then gathered into a local feature group g and fed into the variformer layer, which employs a multi-head attention mechanism [47] to extract the weights of individual reads in g and then combine them into a new feature vector of fixed length, that is,

$$r = \sum_j^{\frac{L}{l}} \lambda_{ij} W_j c_k, \tag{4}$$

where $\boldsymbol{r}_i \in \boldsymbol{R}^{d_k}$ stands for the output sequence features of

the i-th sample, d_k stands for the feature dimension of each site, $c_j \in \boldsymbol{R}^d$ stands for local features of the j-th read for the sequence $\boldsymbol{S}$, $W_j \in \boldsymbol{R}^{d_k \times d}$ stands for the learnable weight matrix, and λ_{ij} stands for a scalar value that assigns weights to individual reads along the sequence, that is,

$$\lambda_{ij} = \mathrm{softmax}\left[(W_{c_i} c_i)\left(W_{c_j} c_j\right)^T\right], \tag{5}$$

where $W_{c_i}, W_{c_j} \in R^{d_k \times d}$ stand for the learnable weight matrices for c_i and c_j, respectively.

In particular, the variformer layer in our model captures the dependencies between the distant sites along the input sequence through the multi-head attention mechanism, thus increasing the performance and interpretability of the model [47,49,50]. In addition, the variformer layer allows our model to process the input sequences of variable lengths automatically, thus improving the generalization of the model.

The Bayesian network module

To further improve the performance of our model, we introduce a Bayesian network module based on prior biological knowledge (more details can be found in Supplementary Notes S6) about RNA-chromatin interactions to integrate multi-omics data. Mathematically, the given training set $\boldsymbol{X}$ consisting of samples with multi-omics data and their corresponding label set $\boldsymbol{Y}$ can be represented as:

$$\boldsymbol{X} = \{D, O, S_{\mathrm{DNA}}, S_{\mathrm{RNA}}\}, \tag{6}$$

$$\boldsymbol{Y} = \{I\}, \tag{7}$$

where D stands for the chromatin conformation data, O stands for the chromatin open accessibility data, S_{DNA} stands for the input DNA sequences, S_{RNA} stands for the input RNA sequences, and I stand for the interaction states of RNA-DNA pairs. Then, the loss function $\boldsymbol{J}$ of our model can be defined as:

$$\boldsymbol{J} = \boldsymbol{E}\left(-I\log\left(\hat{I}(\Theta)\right)\right), \tag{8}$$

where $\boldsymbol{E}(\bullet)$ stands for the expectation of the cross entropy, Θ stands for the learnable weight parameters of the model, and $\hat{I}$ stands for the predicted interaction potential of RNA-DNA pairs, which is calculated by

$$\hat{I} = P(I, D, O, S_{\mathrm{DNA}}, S_{\mathrm{RNA}}), \tag{9}$$

where P stand for the joint probability of the random variables, including I, D, O, S_{DNA}, and S_{RNA}. As illustrated in Supplementary Notes S6, based on the Bayesian graph derived from known mechanisms of RNA-chromatin interactions (Fig. 1B), the joint probability $P(I, D, O, S_{\mathrm{DNA}}, S_{\mathrm{RNA}})$can be written as:

$$\begin{aligned} P(I, D, O, S_{\mathrm{DNA}}, S_{\mathrm{RNA}}) &= P(S_{\mathrm{DNA}})P(S_{\mathrm{RNA}}) \\ &P(D|S_{\mathrm{DNA}}, S_{\mathrm{RNA}}) \\ &P(I|S_{\mathrm{DNA}}, S_{\mathrm{RNA}}) \\ &P(I|D, O, S_{\mathrm{DNA}}, S_{\mathrm{RNA}}), \end{aligned} \tag{10}$$

where $P(\cdot|\cdot)$ stands for the conditional probability. Then, $\log\left(\hat{I}(\Theta)\right)$ be written as,

$$\begin{aligned} &m(I|D, O, S_{\mathrm{RNA}}, S_{\mathrm{DNA}}; \Theta_1) + f(D|S_{\mathrm{DNA}}, S_{\mathrm{RNA}}; \Theta_2) \\ &+ g(O|S_{\mathrm{DNA}}, S_{\mathrm{RNA}}; \Theta_3), \end{aligned} \tag{11}$$

where m, f and g stand for the neural networks (see Supplementary Notes and Fig. S6) that map the original features to the corresponding conditional probabilities, respectively, and Θ_1, Θ_2, and Θ_3 stand for the learnable weights of the model. More details about the derivation of Eq. (11) can be found in Supplementary Notes S6. In our framework, $m(I|D, O, S_{\mathrm{RNA}}, S_{\mathrm{DNA}}; \Theta_1)$ provides the final prediction with a posterior perspective of RNA-chromatin interactions given the observation D and O, while $f(D|S_{\mathrm{DNA}}, S_{\mathrm{RNA}}; \Theta_2)$ and $g(O|S_{\mathrm{DNA}}, S_{\mathrm{RNA}}; \Theta_3)$ serve as the prior constraints. In principle, our Bayesian network can alleviate the potential overfitting issue especially when the number of non-interacting sites is much larger than that of interacting sites [16, 18].

AVAILABILITY

github website (mlcb-thu/DeepRCI).

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at https://doi.org/10.15302/J-QB-022-0316.

ACKNOWLEDGEMENTS

The authors would like to thank BioRender for the creation of the Bayesian network shown in Fig. 1B. They thank Mr. Xingang Peng and Mr. Hantao Shu for the helpful discussions and suggestions on the manuscript. This work was supported in part by the National Natural Science Foundation of China (61872216, T2125007 to JZ, 31900862 to DZ), the National Key Research and Development Program of China (2018YFC0910404, 2021YFF1201300), the Turing AI Institute of Nanjing, the Tsinghua-Toyota Joint Research Fund and the US National Institute of Health grant (1R01NS125018).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Yuanpeng Xiong, Xuan He, Dan Zhao, Tao Jiang, and Jianyang Zeng declare that they have no conflict of interest or financial conflicts to disclose.

This article does not contain any studies with human or animal materials performed by any of the authors

REFERENCES

1. Bell, J. C., Jukam, D., Teran, N. A., Risca, V. I., Smith, O. K., Johnson, W. L., Skotheim, J. M., Greenleaf, W. J. and Straight, A. F. (2018) Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts. eLife, 7, e27024
2. Calandrelli, R., Xu, L., Luo, Y., Wu, W., Fan, X., Nguyen, T., Chen, C.-J., Sriram, K., Tang, X., Burns, A. B., *et al*. (2020) Stress-induced RNA-chromatin interactions promote endothelial dysfunction. Nat. Commun., 11, 5211
3. Li, X. and Fu, X.-D. (2019) Chromatin-associated RNAs as facilitators of functional genomic interactions. Nat. Rev. Genet., 20, 503−519
4. Bonetti, A., Agostini, F., Suzuki, A. M., Hashimoto, K., Pascarella, G., Gimenez, J., Roos, L., Nash, A. J., Ghilotti, M., Cameron, C. J. F., *et al*. (2020) RADICL-seq identifies general and cell type-specific principles of genome-wide RNA-chromatin interactions. Nat. Commun., 11, 1018
5. Antonov, I. and Medvedeva, Y. A. (2018) Purine-rich low complexity regions are potential RNA binding hubs in the human genome. F1000 Res., 7, 76
6. Kato, M. and Carninci, P. (2020) Genome-wide technologies to study RNA–chromatin interactions. Noncoding RNA, 6, 20
7. Watanabe, T., Tomizawa, S., Mitsuya, K., Totoki, Y., Yamamoto, Y., Kuramochi-Miyagawa, S., Iida, N., Hoki, Y., Murphy, P. J., Toyoda, A., *et al*. (2011) Role for piRNAs and noncoding RNA in *de novo* DNA methylation of the imprinted mouse *Rasgrf1* locus. Science, 332, 848−852
8. Miao, Y., Ajami, N. E., Huang, T.-S., Lin, F.-M., Lou, C.-H., Wang, Y.-T., Li, S., Kang, J., Munkacsi, H., Maurya, M. R., *et al*. (2018) Enhancer-associated long non-coding RNA LEENE regulates endothelial nitric oxide synthase and endothelial function. Nat. Commun., 9, 292
9. J. L. Rinn, M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu, S. A. Brugmann, L. H. Goodnough, J. A. Helms, P. J. Farnham, E. Segal, *et al*. (2007) Functional demarcation of active and silent chromatin domains in human *HOX* loci by noncoding RNAs. Cell, 129, 1311−1323
10. Kalwa, M., Hänzelmann, S., Otto, S., Kuo, C.-C., Franzen, J., Joussen, S., Fernandez-Rebollo, E., Rath, B., Koch, C., Hofmann, A., *et al*. (2016) The lncRNA HOTAIR impacts on mesenchymal stem cells via triple helix formation. Nucleic Acids Res., 44, 10631−10643
11. Wang, S., Ke, H., Zhang, H., Ma, Y., Ao, L., Zou, L., Yang, Q., Zhu, H., Nie, J., Wu, C., *et al*. (2018) LncRNA MIR100HG promotes cell proliferation in triple-negative breast cancer through triplex formation with p27 loci. Cell Death Dis., 9, 805
12. O'Leary, V. B., Ovsepian, S. V., Carrascosa, L. G., Buske, F. A., Radulovic, V., Niyazi, M., Moertl, S., Trau, M., Atkinson, M. J. and Anastasov, N. (2015) Particle, a triplex-forming long ncRNA, regulates locus-specific methylation in response to low-dose irradiation. Cell Rep., 11, 474−485
13. Simon, M. D. (2013) Capture hybridization analysis of RNA targets (CHART). Curr. Protoc. Mol. Biol., 101, 21−25
14. Chu, C. and Chang, H. Y. (2016) Understanding RNA-chromatin interactions using chromatin isolation by RNA purification (chirp). In: Polycomb Group Proteins, pp. 115–123. New York: Springer
15. Engreitz, J., Lander, E. S. and Guttman, M. (2015) RNA antisense purification (rap) for mapping RNA interactions with chromatin. In: Nuclear Bodies and Noncoding RNAs, pp. 183–197. New York: Springer
16. Wu, W., Yan, Z., Nguyen, T. C., Bouman Chen, Z., Chien, S. and Zhong, S. (2019) Mapping RNA-chromatin interactions by sequencing with iMARGI. Nat. Protoc., 14, 3243−3272
17. Sridhar, B., Rivas-Astroza, M., Nguyen, T. C., Chen, W., Yan, Z., Cao, X., Hebert, L. and Zhong, S. (2017) Systematic mapping of RNA-chromatin interactions *in vivo*. Curr. Biol., 27, 602−609
18. Zhou, B., Li, X., Luo, D., Lim, D.-H., Zhou, Y. and Fu, X.-D. (2019) GRID-seq for comprehensive analysis of global RNA-chromatin interactions. Nat. Protoc., 14, 2036−2068
19. Kuo, C.-C., Hänzelmann, S., Sentürk Cetin, N., Frank, S., Zajzon, B., Derks, J.-P., Akhade, V. S., Ahuja, G., Kanduri, C., Grummt, I., *et al*. (2019) Detection of RNA-DNA binding sites in long noncoding RNAs. Nucleic Acids Res., 47, e32
20. Buske, F. A., Bauer, D. C., Mattick, J. S. and Bailey, T. L. (2013) Triplex-inspector: an analysis tool for triplex-mediated targeting of genomic loci. Bioinformatics, 29, 1895−1897
21. He, S., Zhang, H., Liu, H. and Zhu, H. (2015) LongTarget: a tool to predict lncRNA DNA-binding motifs and binding sites via Hoogsteen base-pairing analysis. Bioinformatics, 31, 178−186
22. Zhang, Y., Long, Y. and Kwoh, C. K. (2020) Deep learning based DNA: RNA triplex forming potential prediction. BMC Bioinformatics, 21, 522
23. Müller, A. T., Hiss, J. A. and Schneider, G. (2018) Recurrent neural network model for constructive peptide design. J. Chem. Inf. Model., 58, 472−479
24. Sureyya Rifaioglu, A., Doğan, T., Jesus Martin, M., Cetin-Atalay, R., and Atalay, V. (2019) Deepred: automated protein function prediction with multi-task feed-forward deep neural networks. Sci. Rep., 9, 1−16
25. Zeng, Y., Chen, X., Luo, Y., Li, X. and Peng, D. (2021) Deep drug-target binding affinity prediction with multiple attention blocks. Brief. Bioinform., 22, bbab117

26. Lopez-Del Rio, A., Martin, M., Perera-Lluna, A. and Saidi, R. (2020) Effect of sequence padding on the performance of deep learning models in archaeal protein functional prediction. Sci. Rep., 10, 14634
27. Schubert, T., Pusch, M. C., Diermeier, S., Benes, V., Kremmer, E., Imhof, A. and Längst, G. (2012) Df31 protein and snoRNAs maintain accessible higher-order structures of chromatin. Mol. Cell, 48, 434−444
28. Schubert, T. and Längst, G. (2013) Changes in higher order structures of chromatin by RNP complexes. RNA Biol., 10, 175−179
29. Sen, S., Cheng, Z., Sheu, K. M., Chen, Y. H. and Hoffmann, A. (2020) Gene regulatory strategies that decode the duration of NFκB dynamics contribute to LPS-versus TNF-specific gene expression. Cell Syst., 10, 169−182.e5
30. Lanchantin, J., Singh, R., Wang, B. and Qi, Y. (2017) Deep motif dashboard: visualizing and understanding genomic sequences using deep neural networks. In: Pacific Symposium on Biocomputing 2017, pp. 254–265. Singapore: World Scientific
31. 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
32. Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R. and Weirauch, M. T. (2018) The human transcription factors. Cell, 172, 650−665
33. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. and Noble, W. S. (2007) Quantifying similarity between motifs. Genome Biol., 8, R24
34. Zhao, X., Li, D., Pu, J., Mei, H., Yang, D., Xiang, X., Qu, H., Huang, K., Zheng, L. and Tong, Q. (2016) CTCF cooperates with noncoding RNA MYCNOS to promote neuroblastoma progression through facilitating MYCN expression. Oncogene, 35, 3565−3576
35. Guttman, M. and Rinn, J. L. (2012) Modular regulatory principles of large non-coding RNAs. Nature, 482, 339−346
36. Wang, I. X., Grunseich, C., Fox, J., Burdick, J., Zhu, Z., Ravazian, N., Hafner, M. and Cheung, V. G. (2018) Human proteins that interact with RNA/DNA hybrids. Genome Res., 28, 1405−1414
37. Engreitz, J. M., Ollikainen, N. and Guttman, M. (2016) Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. Nat. Rev. Mol. Cell Biol., 17, 756−770
38. Iyengar, B. R., Choudhary, A., Sarangdhar, M. A., Venkatesh, K. V., Gadgil, C. J. and Pillai, B. (2014) Non-coding RNA interact to regulate neuronal development and function. Front. Cell. Neurosci., 8, 47
39. Li, L., Luo, H., Lim, D.-H., Han, L., Li, Y., Fu, X.-D. and Qi, Y. (2021) Global profiling of RNA-chromatin interactions reveals co-regulatory gene expression networks in *Arabidopsis*. Nat. Plants, 7, 1364−1378
40. Chen, X., Sun, Y., Cai, R., Wang, G., Shu, X. and Pang, W. (2018) Long noncoding RNA: multiple players in gene expression. BMB Rep., 51, 280−289
41. Mishra, K. and Kanduri, C. (2019) Understanding long noncoding RNA and chromatin interactions: what we know so far. Noncoding RNA, 5, 54
42. Antonov, I. V., Mazurov, E., Borodovsky, M. and Medvedeva, Y. A. (2019) Prediction of lncRNAs and their interactions with nucleic acids: benchmarking bioinformatics tools. Brief. Bioinform., 20, 551−564
43. Quinodoz, S. A., Jachowicz, J. W., Bhat, P., Ollikainen, N., Banerjee, A. K., Goronzy, I. N., Blanco, M. R., Chovanec, P., Chow, A., Markaki, Y., *et al*. (2021) RNA promotes the formation of spatial compartments in the nucleus. Cell, 184, 5775−5790.e30
44. Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B. and Liu, J. S. (2012) HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics, 28, 3131−3133
45. Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics, 28, 3150−3152
46. Ribeiro, L. S. F., Bui, T., Collomosse, J. and Ponti, M. (2020) Sketchformer: transformer-based representation for sketched structure. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14153–14162
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017) Attention is all you need. In: Advances in neural information processing systems, pp. 5998–6008
48. Ioffe, S. and Szegedy, C. (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning, pp. 448–456
49. Hu, H., Xiao, A., Zhang, S., Li, Y., Shi, X., Jiang, T., Zhang, L., Zhang, L. and Zeng, J. (2019) DeepHINT: understanding HIV-1 integration via deep learning with attention. Bioinformatics, 35, 1660−1667
50. Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H. and Winther, O. (2017) DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics, 33, 3387−3395