Supplementary Notes for "DeepRCI: Predicting RNA-chromatin interactions via deep learning with multi-omics data"

Yuanpeng Xiong[1,2] Xuan He[3] Dan Zhao[3] Tao Jiang[4,1,2,*]
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, China
[4]Department of Computer Science and Engineering, University of California, Riverside, CA, USA
* Corresponding authors. Email: jiang@cs.ucr.edu, zengjy321@tsinghua.edu.cn

1. Application of DeepRCI in analyses of fusion genes

As described in [1, 2], RNA-chromatin interactions are related to the fusion genes, the fusion transcripts that can serve as the core indicators of cancers. To demonstrate the ability of our model on the application related to this aspect, we designed a computational experiment to measure the capability of our model to recall fusion genes from all the interaction pairs following the same idea as in [3]. More specifically, we first filtered all the predicted RNA-DNA pairs at the whole genome scale, and only kept those interaction pairs with the prediction score higher than a given threshold. Then we calculated the number of fusion genes that had been experimentally validated in the remaining interaction pairs. As shown in Table S3, our model outperformed all the baseline methods under all three given thresholds, which demonstrated the high recall rate of our model.

2. Intensities of RNA-chromatin interactions

The intensities of RNA-chromatin interactions represent the numbers of RNA fragments that interact with specific DNA fragments, which can be written as

$$V_i = \sum_{s_j \in G} I\left(P_{s_i,s_j} > 0.5\right), (1)$$

where *I* stands for the indicator function, s_i stands for the DNA fragments and s_j stands for all the RNA sequences transcripted from the whole genome *G*, and $P_{si,sj}$ stands for the interaction score of s_i and s_j predicted by DeepRCI.

3. Extended analyses on more datasets collected from other sequencing techniques

To further illustrate that our model can be generalized to data from other technologies, we have also applied our model to predict the RNA-chromatin interactions measured from GRID-seq [4] and RADICL-seq [5], denoted as grid17 and radicl20, respectively. We believe these two datasets could be representative to prove the generalizability of our model to different techniques, though we did not compare CHAR-seq techniques [6] due to the lack of training data. The interaction read pairs of RNAs and DNAs in grid18 and radicl20 were downloaded from the GEO database with accession numbers GSE132192 and GSE82312, respectively. The Hi-C data and the ATAC-seq data were also integrated into these two datasets (accession numbers GSM2375122, and GSM5117383 from the GEO database, respectively).

As shown in Fig. S1, our model still outperformed the the-start-of-the-art methods significantly on these two new datasets. To further validate the role of Hi-C data in the prediction task, we also performed extra ablation studies on grid17 and radicl20 datasets. More specifically, we first trained a model with sequence and Hi-C data, denoted as DeepRCI-hic, and compared it to all other baseline methods. Consequently, we found that DeepRCI-hic achieved an accuracy of 0.792, compared to 0.77 for DeepRCI-seq (Fig. S1). These results demonstrated that solely integrating Hi-C data can also enhance the performance of the model trained with only sequences information.

4. Edge swapping strategy for generating negative samples

Here, we describe another sampling strategy that was used for generating negative samples, denoted as edge swapping, which can preserve the representativeness of DNA/RNA sequences in the original data. More specifically, for two positive pairs (i, j), (p, q), we generated new pairs (i, p) and (j, q), and kept them only if the new pairs were not positive pairs. We compared the performance of our models with the baseline methods. The comparison results (Fig. S5) indicated that our model also performed much better than the baselines on predicting RNA-chromatin interaction with the new sampling strategy.

5. The effect of read length in variformer

To investigate the effect of the length of short reads that are split from each sequence in variformer, we conducted an additional fivefold cross-validation test where we split sequences of variable lengths into short reads with different choices of read length. As shown in Figure S6, DeepRCI achieved the best performance when the read length was set as 1000. Therefore, we kept this hyperparameter setting in our model.

6. Bayesian network module

6.1 Motivation for the Bayesian graph

Our motivation for building a Bayesian graph model to model the underlying mechanisms of RNA-chromatin interactions was mainly inspired by the following experimental observations.

(1) Calandrelli et al. [1, 2] reported that the numbers and strength of chromatin loops are influenced by RNA-chromatin interactions. In particular, the global suppression of caRNAs can increase the number of chromatin loops and enhancer-promoter interactions. This result indicates that the three-dimensional organizations of chromatin are highly associated with RNA-chromatin interactions.

(2) Yan et al. [3] proposed an "RNA-pose" model to represent the features of fusion transcripts based on a validation cohort of 96 lung cancer samples. In this model, the transcripts of a gene interact with the transcripts of another nearby gene through spatial proximity, which subsequently forms a fusion RNA. This "RNA-pose" model indicates that the RNA-chromatin interactions can potentially be detected in the chromatin openness regions and related to chromatin conformation.

(3) With the development of deep learning methods, more and more prediction tools have demonstrated that raw DNA/RNA sequences can contain valuable features that are associated with chromatin open accessibility [7], RNA-chromatin interactions [1, 8] and chromatin conformation [9, 10].

6.2 Deriviation of the Bayesian joint probability equation

Equation:

$$log\ P\ (I, D, O, S_{DNA}, S_{RNA}) = m(I|D, O, S_{RNA}, S_{DNA};\ \Theta_1) + f\ (D|S_{DNA}, S_{RNA};\ \Theta_2)\ +\ g(O|S_{DNA}, S_{RNA};\ \Theta_3), (2)$$

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, *I* stand for the interactions states of RNA-DNA pairs, *P* stands for the joint probability of the five random variables, and *m*, *f* and *g* stand for the neural networks that map the original features to the corresponding conditional probabilities, while Θ_1, Θ_2, and Θ_3 represent their learnable weight parameters.

Proof : Based on the Bayesian product rule, the log value of the joint probability distribution can be written as:

$$\begin{aligned} log\ P\ (I, D, O, S_{DNA}, S_{RNA}) &= log[P\ (I|D, O, S_{DNA}, S_{RNA}) P\ (D, O, S_{DNA}, S_{RNA})] \\ &= log[P\ (I|D, O, S_{DNA}, S_{RNA}) P\ (D|O, S_{DNA}, S_{RNA}) P\ (O, S_{DNA}, S_{RNA})] \\ &= log[P\ (I|D, O, S_{DNA}, S_{RNA}) P\ (D|O, S_{DN}A, S_{RNA}) \\ &\quad P\ (O|S_{DNA}, S_{RN}A) P\ (S_{DNA}, S_{RNA})] \\ &= log[P\ (I|D, O, S_{DNA}, S_{RNA}) P\ (D|O, S_{DNA}, S_{RNA}) \\ &\quad P\ (O|S_{DNA}, S_{RNA}) P\ (S_{DNA}) P\ (S_{RNA}|S_{DNA})]. (3) \end{aligned}$$

Based on the derived Bayesian network shown in Fig.1b and the assumption of conditional independence, the log value of the joint probability can be written as:

$$log\ P\ (I, D, O, S_{DNA}, S_{RNA}) = log[P\ (S_{DNA})P\ (S_{RNA})P\ (D|\ S_{DNA}, S_{RNA})$$
$$P\ (O|S_{DNA}, S_{RNA})P\ (I|D, O, S_{DNA}, S_{RNA})]$$
$$= log[P\ (S_{DNA})] + log[P\ (S_{RNA})] + log[P\ (D|\ S_{DNA}, S_{RNA})]$$
$$log[P\ (O|\ S_{DNA}, S_{RNA})] + log[P\ (I|D, O, S_{DNA}, S_{RNA})]. (4)$$

Our model is trained by optimizing the binary cross entropy loss, that is,

$$\arg\min_{\Theta} E[-I\ log(P\ (I, D, O, S_{DNA}, S_{RNA}|\Theta))], (5)$$

where E stands for the expectation of the cross entropy and Θ stands for the learnable weight parameters of the model. Given the fact that input sequences S_{DNA} and S_{RNA} are the observation data, we can safely drop $P\ (S_{DNA})$ and $P\ (S_{RNA})$ during the optimization process according to the instructions in [14] and then rewrite $\log(P\ (I, D, O, S_{DNA}, S_{RNA}|\Theta))$ as:

$$log\ P\ (I, D, O, S_{DNA}, S_{RNA}|\Theta) = log[P\ (D|S_{DNA}, S_{RNA}|\Theta 1)] + log[P\ (O|S_{DNA}, S_{RNA}|\Theta 2)]$$
$$+ log[P\ (I|D, O, S_{DNA}, S_{RNA}|\Theta 3)]. (6)$$

Inspired by the normalization flow theory [11-13], we replaced these three probability distributions with neural networks m, f and g (see the architectures of the neural networks in Fig. S7), which are trained simultaneously using an end-to-end strategy [14].

7. Supplementary figures

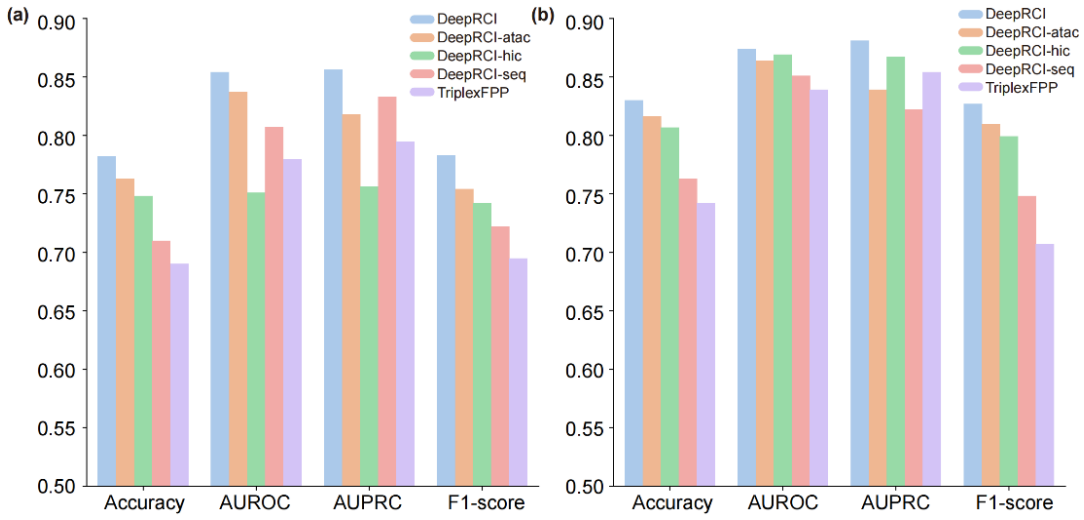

Figure S1: The performance of different models on multi-omics data for the RADICL-seq dataset and the GRID-seq dataset. (a) Performance comparisons on the RADICL-seq dataset among DeepRCI-seq, DeepRCI-atac, DeepRCI-hic, DeepRCI, and TriplexFPP, measured in terms of Accuracy, AUROC score, AUPRC score and F1-score. DeepRCI-atac represents our model trained using multi-omics information except for Hi-C data. DeepRCI-seq represents our DeepRCI model trained with only sequence data. DeepRCI-hic represents our model trained using multi-omics information except for ATAC-seq data. (b) Performance comparisons on the GRID-seq dataset among DeepRCI-seq, DeepRCI-atac, DeepRCI-hic, DeepRCI, and TriplexFPP, measured in terms of Accuracy, AUROC score, AUPRC score and F1-score.

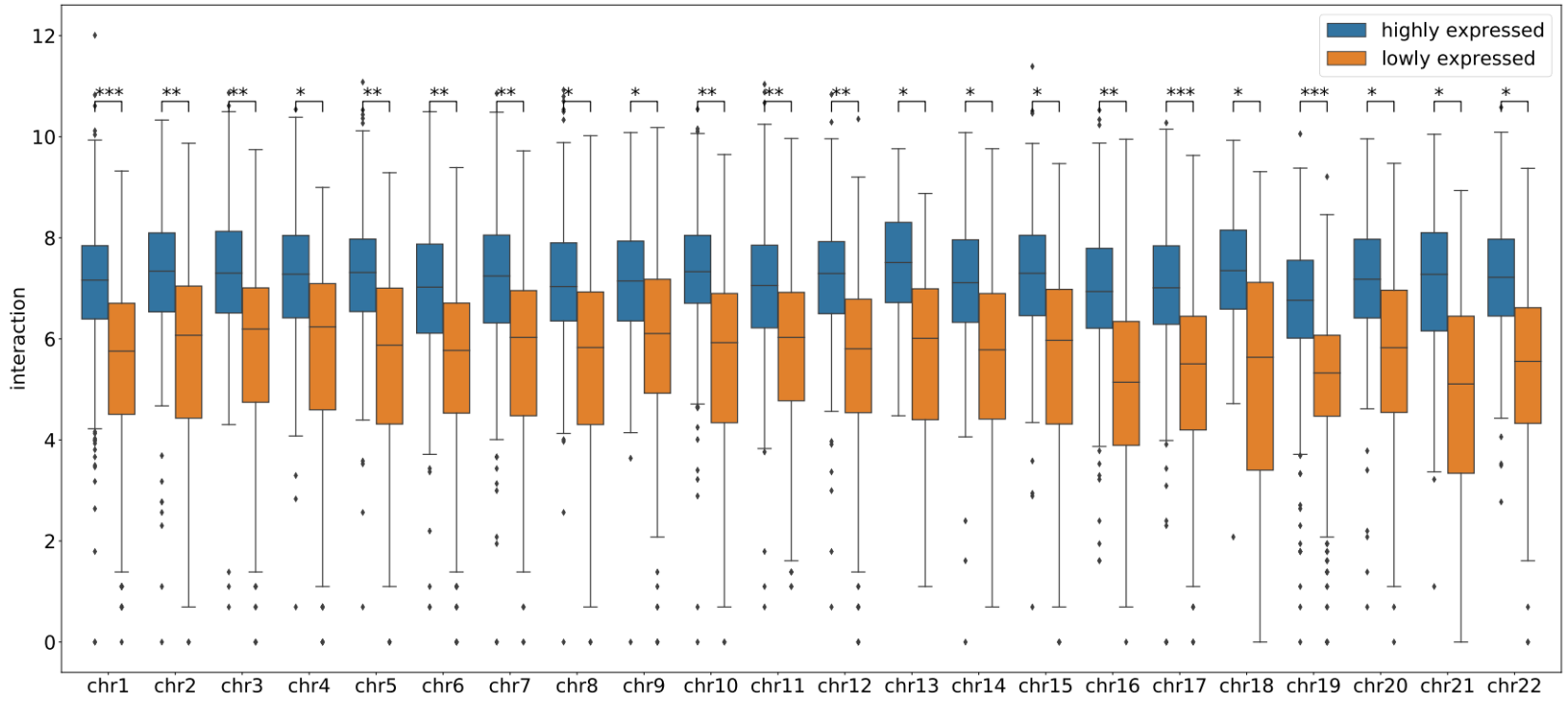

Figure S2: The box plot of RNA-chromatin interaction intensities between highly expressed genes (top 25%) and lowly expressed genes (bottom 25%) on euchromosomes. Highly expressed genes tend to have high interaction intensities (*: $P=1 \times 10^{-30}$~10^{-1},**: $P=1 \times 10^{-60}$ ~10^{-30}, ***: $P=1 \times 10^{-90}$~10^{-60}, Wilcoxon rank-sum test).

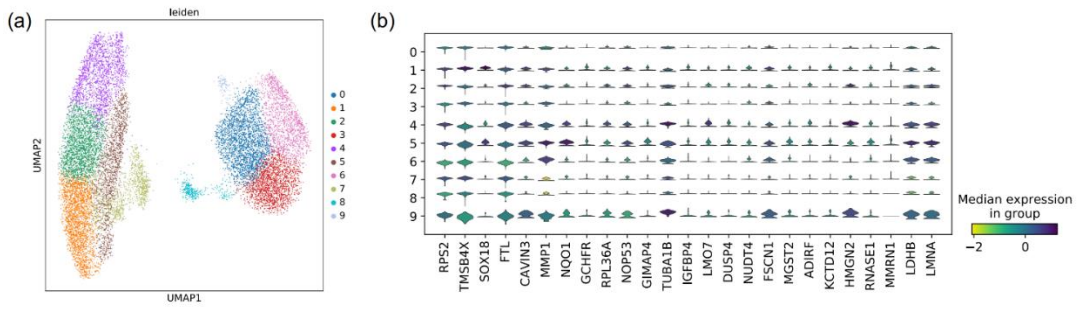

Figure S3: Gene regulation and differential expression analyses of normal and abnormal human umbilical vein endothelial cells (HUVECs). (a) The low dimensional map of the 10× single-cell RNA-seq data (clustered by the leiden algorithm [3]), indicated that there mainly existed two types of cells in our samples (i.e., normal and abnormal cells). (b) The stacked violin plot of differentially expressed genes (top 25 genes according to the log foldchanges) between normal and abnormal HUEVCs.

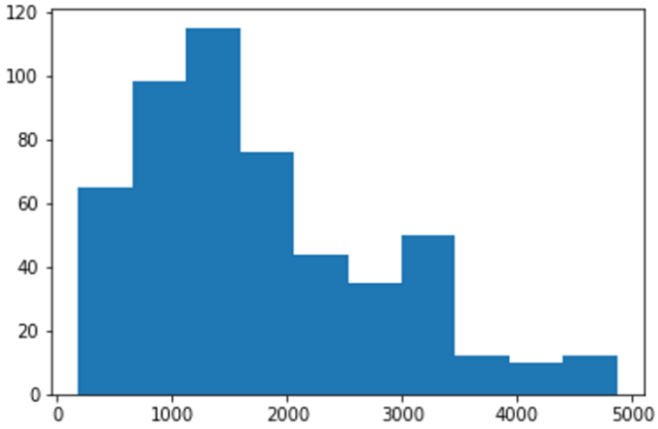

Figure S4: The histogram of the lengths of lncRNA sequences in the zhang2020 dataset.

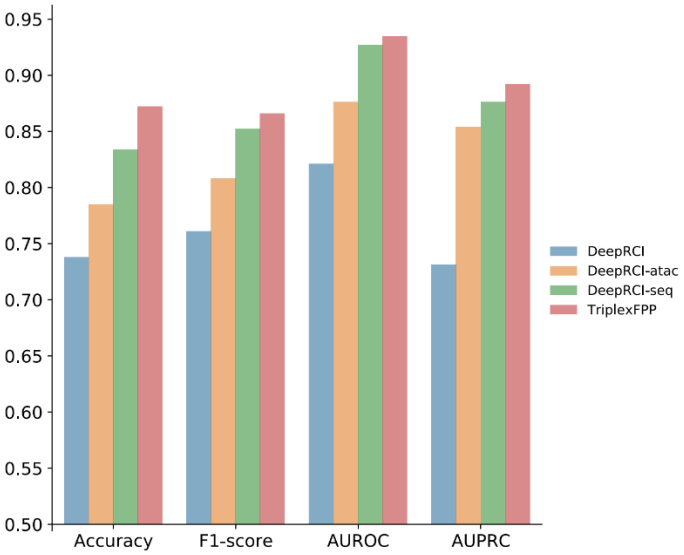

Figure S5: Performance comparisons on the stress20 dataset sampled with the “edge swapping” strategy among DeepRCI, DeepRCI-seq, DeepRCI-atac, and TriplexFPP, measured in terms of accuracy, F1-score, AUROC score, and AUPRC score. DeepRCI-atac represents a version of DeepRCI trained using multi-omics information except for Hi-C data.

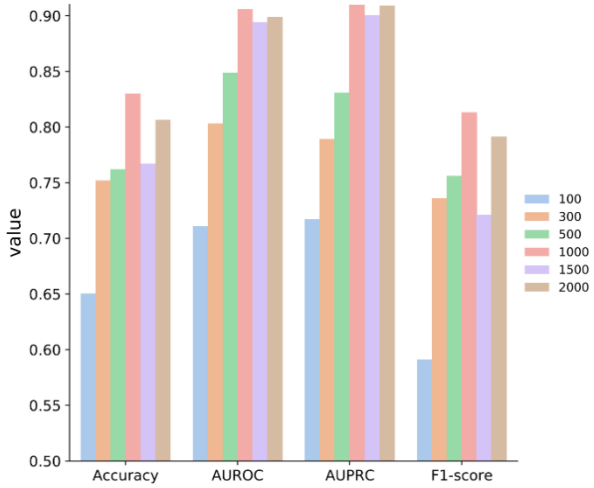

Figure S6: Performance comparison among different read lengths in variformer on the zhang2020 dataset in terms of accuracy, F1-score, AUROC, and AUPRC.

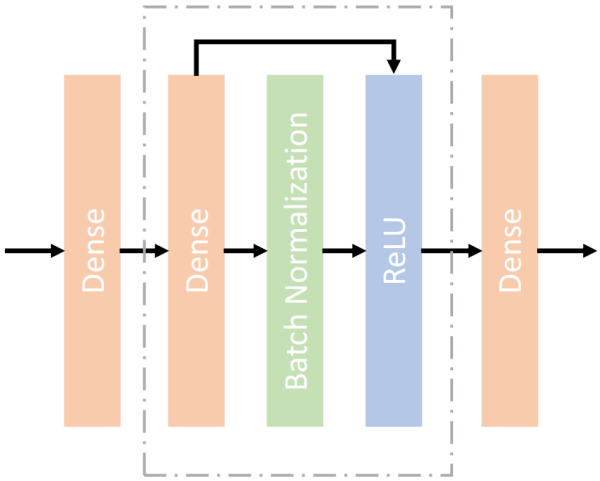

Figure S7: Architecture of the feature mapping layer. The grey dash box means that this part is repeated multiple times. Here, “Dense” denotes fully connected layers and ”ReLU” denotes the linear rectification function.

8. Supplementary tables

Query_ID	Target_ID	Optimal_offset	p-value	E-value	q-value	Overlap	Query_consensus	Target_consensus	Orientation
WDNMWNNWKNHW	M5934_1.02	0	3.24E-07	0.000237382	0.000234873	12	TGACAGCTGTCA	TGACAGCTGTCA	+
WDNMWNNWKNHW	M5935_1.02	0	8.22E-07	0.000602395	0.000298014	12	TGACAGCTGTCA	TGACAGCTGTCA	+
WDNMWNNWKNHW	M6344_1.02	0	4.78E-06	0.00350547	0.00115614	12	TGACAGCTGTCA	TGACAGCTGTCAA	+
WDNMWNNWKNHW	M5722_1.02	0	1.69E-05	0.0124223	0.00307275	12	TGACAGCTGTCA	TGACACCTGTCA	+
DCWDNNNNHWGH	M6402_1.02	-3	3.55E-05	0.0260254	0.0260254	9	ACATTTAAATGT	TGTAACTGT	+
NWNNWDNNNNHWNNWN	M5616_1.02	-2	6.95E-05	0.0509191	0.0505661	12	CTGCTGTATACAGCAG	GCTATAAATAGC	+
WDNHWNWNWNWDNHW	M4572_1.02	-1	7.42E-05	0.0544122	0.054238	14	ATGCATTGAATGCAT	TGCTGACTCAGCAAA	+
WNHHWBHNDVWDDNW	M0632_1.02	-2	8.85E-05	0.0648531	0.0424239	10	TGAATGATTCATTCA	AATGTATCAA	+
DNHDNHWDHWDNHDNH	M5735_1.02	-1	8.87E-05	0.0649975	0.0646111	13	TAATTCTGCAGAATTA	AAATTAGCATAAT	+
DCWDNHWGH	M6436_1.02	1	9.14E-05	0.0669964	0.053234	9	ACAGTCTGT	TACAGACTGTTCT	+
WNNHWBHNDVWDNNW	M0632_1.02	-2	9.23E-05	0.0676231	0.0676231	10	TGACAGTTACTGTCA	AATGTATCAA	+
WNHHWSHNDSWDDNW	M0632_1.02	-2	9.54E-05	0.0699034	0.0453299	10	TGAATGATTCATTCA	AATGTATCAA	+
WNHHWBHNDVWDDNW	M0630_1.02	-2	0.000115754	0.0848479	0.0424239	10	TGAATGATTCATTCA	AATGTATCAA	+
WNHHWSHNDSWDDNW	M0630_1.02	-2	0.00012374	0.0907014	0.0453299	10	TGAATGATTCATTCA	AATGTATCAA	+
DNWDNHDNNHDNHWNH	M6417_1.02	-4	0.000131876	0.0966652	0.05749	12	TCATGATGCATCATGA	ATATATTCATGAG	+
DCWDNHWGH	M1841_1.02	2	0.00014525	0.106468	0.053234	9	ACAGTCTGT	GAACATTCTGTTCTT	+
WDNMWNNWKNHW	M6343_1.02	1	0.000154622	0.113338	0.022428	12	TGACAGCTGTCA	CATAAAACTGTCA	+
DNWDNHDNNHDNHWNH	M6426_1.02	-1	0.000156934	0.115033	0.05749	10	TCATGATGCATCATGA	CATAAATAAT	+
WNHNDSHNDSHNDNW	M0630_1.02	-2	0.000172649	0.126551	0.0845497	10	TGTATGATTCATACA	AATGTATCAA	+

Table S1: The top 20 DNA sequence motifs matched to the motifs obtained from CIS-BP [14] with TOMTOM (version 5.0.5) [15]. They were sorted according to the reported p-values. The column names of this table are described at tomtom website

(doc/tomtom-output-format.html). More details of the results from TOMTOM can be found at github website (mlcb-thu/DeepRCI).

Query_ID	Target_ID	Optimal_offset	p-value	E-value	q-value	Overlap	Query_consensus	Target_consensus	Orientation
DNHDDSHNDSHHDNH	M256_0.6	-5	0.00018	0.0180866	0.017992	6	TCATGCAGTGCATGA	CAGTGC	+
HNNSWNNNNNWSNND	M256_0.6	-3	0.00026	0.0251163	0.024985	6	AGTCAGGGCCTGACT	CAGTGC	+
NNHNDSHNDSHNDNN	M074_0.6	-8	0.00063	0.0613189	0.0609982	7	TCATTCATTGAATGA	TGAATGAG	+
MWGNNCWK	M061_0.6	1	0.00095	0.0930866	0.0925998	6	CTGGCCAG	GCTGGAC	+
CWGDNHCWG	M256_0.6	0	0.00118	0.115263	0.11466	6	CAGGGCCTG	CAGTGC	+
DNHDBNNNNNNVHDNH	M234_0.6	-1	0.00119	0.116969	0.116357	7	TGCAGATATATCTGCA	GATGATT	+
NDNSHNDSHNDSNHN	M256_0.6	-3	0.00132	0.129141	0.125053	6	CTGCAGTGACTGCAG	CAGTGC	+
NNDSHNNHDNNDSHNN	M231_0.6	-4	0.00134	0.131775	0.129734	7	CATCAGCATGCTGATG	AGCATGC	+
HDDSHNDNHNDSHHD	M256_0.6	-3	0.00140	0.137524	0.133854	6	ATGCAGAGTCTGCAT	CAGTGC	+
NHDNHNDSHNDNHDN	M256_0.6	-7	0.00148	0.145305	0.142219	6	CTGGACTGAGTCCAG	CAGTGC	+
DSHDNNNNNNNNHDSH	M159_0.6	-9	0.00157	0.153402	0.151027	7	TCATGCAGCTGCATGA	TGCATGC	+
NWGHHDNNNHDDCWN	M037_0.6	-1	0.00192	0.187795	0.183703	7	CTGCATGGCATGCAG	AGCTTGC	+
CWKDNHMWG	M256_0.6	0	0.00225	0.220797	0.219642	6	CTGGGCCAG	CAGTGC	+
DNHDVHNNNNDBHDNH	M234_0.6	-1	0.00228	0.223676	0.222506	7	TGCTGATGCATCAGCA	GATGATT	+
HNNSHNDSHNDSNND	M159_0.6	-6	0.00229	0.224355	0.176099	7	TGTCATTGAATGACA	TGCATGC	+
NDBNHWKNMWDNVHN	M140_0.6	0	0.00249	0.243896	0.24262	7	CTGACAGTCTGTCAG	CAGACAG	+
NNDSHNNHDNNDSHNN	M159_0.6	-4	0.00276	0.270274	0.133044	7	CATCAGCATGCTGATG	TGCATGC	+
CWKDNHMWG	M256_0.6	0	0.00280	0.274243	0.272809	6	CAGGGCCTG	CAGTGC	+
NDDMWNDNHNWKHHN	M256_0.6	-3	0.00298	0.292133	0.290605	6	ATGCAGTGACTGCAT	CAGTGC	+
NMWGNNNNCWKN	M085_0.6	-3	0.00300	0.294174	0.292635	8	CCTGAATTCAGG	GAATTAAG	+

Table S2: The top 20 RNA sequence motifs matched to the motifs obtained from CISBP-RNA [14] with TOMTOM (version 5.0.5) [15]. They were sorted according to the reported p-values. The column names of this table are described tomtom website (doc/tomtom-output-format.html). More details of the results from TOMTOM can be found at github website (mlcb-thu/DeepRCI).

Threshold	DeepRCI-seq	DeepRCI	TriplexFPP
0.6	21	**27**	14
0.7	17	**22**	8
0.8	10	**15**	5

Table S3: DeepRCI detected more fusion genes under different threshold settings than the baselines.

Settings of hyperparameters			Performance			
Dropout[1]	Learning rate	Weight decay[2]	Accuracy	AUROC	AUPRC	F1-score
1	1.00E-03	0	0.75(±0.043)	0.824(±0.049)	0.832(±0.052)	0.748(±0.043)
1	1.00E-04	0	0.76(±0.038)	0.84 (±0.038)	0.84(±0.05)	0.759(±0.053)
0.4	1.00E-04	0	0.768(±0.043)	0.845(±0.038)	0.85(±0.042)	0.76(0.039)
0.5	**1.00E-04**	**0**	**0.781**(±0.041)	**0.852**(±0.027)	**0.858**(±0.035)	0.765(0.045)
0.6	1.00E-04	0	0.756(±0.072)	0.831(±0.058)	0.831(±0.049)	0.767(±0.051)
0.5	1.00E-04	1.00E-03	0.77(±0.052)	0.832(±0.062)	0.839(±0.059)	**0.771**(±0.048)
0.5	1.00E-04	1.00E-05	0.75(±0.041)	0.822(±0.0465)	0.82(±0.047)	0.742(±0.04)

1 Dropout: the probability of keeping input elements from a Bernoulli distribution.

2 Weight decay: coefficient for L2 penalty.

Table S4: Hyperparamter search for DeepRCI using a fivefold cross-validation procedure. We applied a line-search strategy [16] to determine the settings of hyperparameters, including dropout rate, learning rate, and weight decay rate. We also adopted a repeating process to yield a robust selection of hyperparameters. More specifically, for each group of hyper-parameters, we repeated the training process five times with different random numbers. The best hyperparameter settings are marked in bold.

References

1. R. Calandrelli, X. Wen, T. C. Nguyen, C.-J. Chen, Z. Qi, W. Chen, Z. Yan, W. Wu, K. Zaleta Rivera, R. Hu, et al. Three-dimensional organization of chromatin associated rnas and their role in chromatin architecture in human cells. bioRxiv, 2021.

2. R. Calandrelli, L. Xu, Y. Luo, W. Wu, X. Fan, T. Nguyen, C.-J. Chen, K. Sriram, X. Tang, A. B. Burns, et al. Stress-induced rna–chromatin interactions promote endothelial dysfunction. Nature communications, 11(1):1– 13, 2020.

3. V. A. Traag, L. Waltman, and N. J. Van Eck. From louvain to leiden: guaranteeing well-connected communities. Scientific reports, 9(1):1– 12, 2019.

4. B. Zhou, X. Li, D. Luo, D.-H. Lim, Y. Zhou, and X.-D. Fu. Grid-seq for comprehensive analysis of global rna–chromatin interactions. Nature protocols, 14(7):2036–2068, 2019.

5. A. Bonetti, F. Agostini, A. M. Suzuki, K. Hashimoto, G. Pascarella, J. Gimenez, L. Roos, A. J. Nash, M. Ghilotti, C. J. Cameron, et al. Radicl-seq identifies general and cell type–specific principles of genome-wide rna-chromatin interactions. Nature communications, 11(1):1– 14, 2020.

6. M. D. Simon. Capture hybridization analysis of rna targets (chart). Current protocols in molecular biology, 101(1):21–25, 2013.

7. Q. Liu, F. Xia, Q. Yin, and R. Jiang. Chromatin accessibility prediction via a hybrid deep convolutional neural network. Bioinformatics, 34(5):732–738, 2018.

8. S. He, H. Zhang, H. Liu, and H. Zhu. Longtarget: a tool to predict lncrna dna-binding motifs and binding sites via hoogsteen base-pairing analysis. Bioinformatics, 31(2):178– 186, 2015.

9. H. Tao, H. Li, K. Xu, H. Hong, S. Jiang, G. Du, J. Wang, Y. Sun, X. Huang, Y. Ding, et al. Computational methods for the prediction of chromatin interaction and organization using sequence and epigenomic profiles. Briefings in Bioinformatics, 2021.

10. P. Farr´e, A. Heurteau, O.Cuvier, and E.Emberly. Dense neural networks for predicting chromatin conformation. BMC bioinformatics, 19(1):1– 12, 2018.

11. I. Kobyzev, S. Prince, and M. Brubaker. Normalizing flows: An introduction and review of current methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

12. L. Dinh, J. Sohl-Dickstein, and S.Bengio.Density estimation using real nvp. arXiv preprint arXiv:1605. 08803, 2016.

13. D. Rezende and S. Mohamed. Variational inference with normalizing flows. In International conference on machine learning, pages 1530– 1538. PMLR, 2015.

14. S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch. The human transcription factors. Cell, 172(4):650–665, 2018.

15. S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble. Quantifying similarity between motifs. Genome biology, 8(2):1–9, 2007.

16. J. Bergstra, R. Bardenet, Y. Bengio, and B. K´egl. Algorithms for hyper-parameter optimization. Advances in neural information processing systems, 24, 2011.