SOFTWARE

Open Access

RBPsuite 2.0: an updated RNA-protein binding site prediction suite with high coverage on species and proteins based on deep learning

Xiaoyong Pan[1*†], Yi Fang[1†], Xiaojian Liu[1†], Xiaoyu Guo[1] and Hong-Bin Shen[1*]

Abstract

Background RNA-binding proteins (RBPs) play crucial roles in many biological processes, and computationally identifying RNA-RBP interactions provides insights into the biological mechanism of diseases associated with RBPs.

Results To make the RBP-specific deep learning-based RBP binding sites prediction methods easily accessible, we developed an updated easy-to-use webserver, RBPsuite 2.0, with an updated web interface for predicting RBP binding sites from linear and circular RNA sequences. RBPsuite 2.0 has a higher coverage on the number of supported RBPs and species compared to the original RBPsuite, supporting an increased number of RBPs from 154 to 353 and expanding the supported species from one to seven. Additionally, RBPsuite 2.0 replaces the CRIP built into RBPsuite 1.0 with iDeepC, a more accurate RBP binding site predictor for circular RNAs. Furthermore, RBPsuite 2.0 estimates the contribution score of individual nucleotides on the input sequences as potential binding motifs and links to the UCSC browser track for better visualization of the prediction results.

Conclusions RBPsuite 2.0 is an updated, more comprehensive webserver for predicting RBP binding sites in both linear and circular RNA sequences. It supports more RBPs and species and provides more accurate predictions for circular RNAs. The tool is freely available at http://www.csbio.sjtu.edu.cn/bioinf/RBPsuite/.

Keywords Deep learning, RNA-binding proteins, Linear RNAs, Circular RNAs

Background

RNA-binding proteins (RBPs) are involved in many biological processes [1], and their dysregulation may result in various diseases [2]. With the high-throughput technology developing, a large volume of RBP binding sites derived from sequencing data have been generated, i.e., the ENCODE project [3] and starBase [4]. However, the high-throughput experimental methods for RBP binding sites on RNAs still have some limitations, i.e., system noises and low cross-linking efficiency, resulting in the missing of some true RBP–RNA interactions. Fortunately, experimental data can serve as the training data for machine learning, including deep learning, models to predict RBP binding sites with promising performance [5–14], and these models can be further applied to screen new RBP binding sites for follow-up wet-lab experiments [15].

From a protein-centric viewpoint, many deep learning-based methods have been developed for predicting RBP binding sites from RNA sequences, employing a single model trained per RBP [9, 16–22]. For example, the first convolutional neural network (CNN) based method DeepBind [16] uses CNNs to classify RBP binding sites from non-binding sites. iDeepS [17] applies CNNs to

[†]Xiaoyong Pan, Yi Fang and Xiaojian Liu shared co-first author.

*Correspondence:
Xiaoyong Pan
2008xypan@sjtu.edu.cn
Hong-Bin Shen
hbshen@sjtu.edu.cn
[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China

extract high-level features derived from sequences and predicted secondary structures, followed by long-short temporary network (LSTM) to capture the dependency between sequences and structures to predict RBP binding sites from sequences. Similarly, pysster encodes the sequence and predicted secondary structure into one-hot encode vector with an extended alphabet, which is fed into CNNs for RBP binding sites prediction [9]. Instead of using predicted secondary structures, PrismNet combines experimental secondary structure data and sequences to improve the prediction performance for 168 RBPs [18]. Considering that some RBPs have a limited number of known binding targets, the Siamese neural network-based model iDeepC is designed to yield good prediction performance for the poorly characterized RBPs [19]. DeepPN combines convolutional neural networks and graph convolutional networks in a deep parallel neural network framework to predict RNA–protein binding sites [22]. BERT-RBP fine-tunes the bidirectional encoder representations from transformer (BERT) architecture pre-trained on a human reference genome to predict RNA–RBP interactions using sequence information [21]. HDRNet proposed an end-to-end deep learning-based framework that utilizes both sequence information and in vivo RNA secondary structure profiles to accurately predict dynamic RBP binding events across various cellular conditions [20]. These prediction methods based on deep learning have shown good performance in identifying RBP binding sites on RNAs.

However, most of the existing deep learning-based methods only provide source codes of the prediction methods, only a few online websevers have been developed, like DeepCLIP [23], PrismNet [18], and catRAPID omics v2.0 [24], although there exist some non-machine-learning-based methods like RBPmap [25]. DeepCLIP integrates convolutional layers and bidirectional long short-term memory (BiLSTM) to train the model on a human dataset [26]. The PrismNet architecture uses a convolutional layer, a two-dimensional residual block, and a one-dimensional residual block connected by max pooling to predict protein–RNA interactions, with training data that includes both human and mouse data [27]. It is worth noting that although catRAPID omics v2.0 can predict protein–RNA interaction propensities in eight model organisms, it does not use deep learning methods and cannot tell where are the predicted binding sites located. Instead, catRAPID omics v2.0 computes the secondary structure of sequences and combines it with physicochemical features, including hydrogen bonding, hydrophobicity, and van der Waals forces, to predict protein-RNA interaction [24]. Previously, we developed an online webserver RBPsuite, here denoted as RBPsuite 1.0 [28], which integrates our previously developed deep learning method CRIP and iDeepS for predicting RBP binding sites on circular RNAs (circRNAs) and linear RNAs, respectively.

RBPsuite 1.0 has been used in numerous studies to identify novel RBP binding sites, and the accuracy of its predictions has been proven in these studies. [29–31]. For example, alterations in the E2f1 UTR sequences in vivo expression tests reveal a region around eighth–ninth uORFs conserved among several species of Drosophila, potentially serving as translational enhancer sites, which is consistent with the prediction of RBPsuite 1.0 [29]. Moreover, RBPsuite 1.0 was used to screen the RBPs binding to non-coding RNA regions of SARS-COV-2 [31], which gives some insights into SARS-CoV-2 RNA interactomes. In addition, the binding sites predicted by the RBPsuite 1.0 were successfully validated by wet-lab experiments [32, 33]. For example, RBP suite 1.0 is used to predict potential IGF2BP1 binding sites on the sense and antisense strands of LINC02428 [32], where the interactions are validated with western blotting. RBPsuite 1.0 was used to infer the binding RBPs of circTmeff1, where RNA immunoprecipitation (RIP) validated the interaction between TDF-43 and circTmeff1 [33]. All the results indicate that RBPsuite 1.0 is able to support biological scientists in investigating the regulation between RBPs and RNAs with fast prediction, which guides the design of web-lab experiments with low cost and high efficiency.

Similarly, RBPsuite 1.0 only covers 154 human RBPs derived from ENCODE eCLIP data and is limited to only the human species. In this study, we updated RBPsuite 1.0 to RBPsuite 2.0 (Fig. 1) with a new web interface and the following new functional features: (1) RBPsutie 2.0 supports RBP sites prediction for seven species (human, mouse, zebrafish, fly, worm, yeast, and *Arabidopsis*) instead of only human species; (2) For human species, RBPsuite 2.0 covers the number RBPs from 154 to 223. (3) For circRNAs, RBPsuite 2.0 updates the prediction engine from CRIP [34] to iDeepC, offering improved performance. (4) RBPsuite 2.0 is able to view the genomic context of any RBP binding site in the UCSC browser, which can be further analyzed with other sources of data, like conservation. (5) RBPsuite 2.0 provides model interpretation to obtain the contribution to RBP-RNA interactions of individual nucleotides on the RNAs, which can be used to locate the potential binding motifs.

Methods

Benchmark dataset construction

RBPsuite 1.0 covers 154 human RBPs with binding sites derived from the ENCODE eCLIP profile [35]. These narrow peaks were generated by the eCLIP-seq processing pipeline v2.0 of ENCODE and can be downloaded from

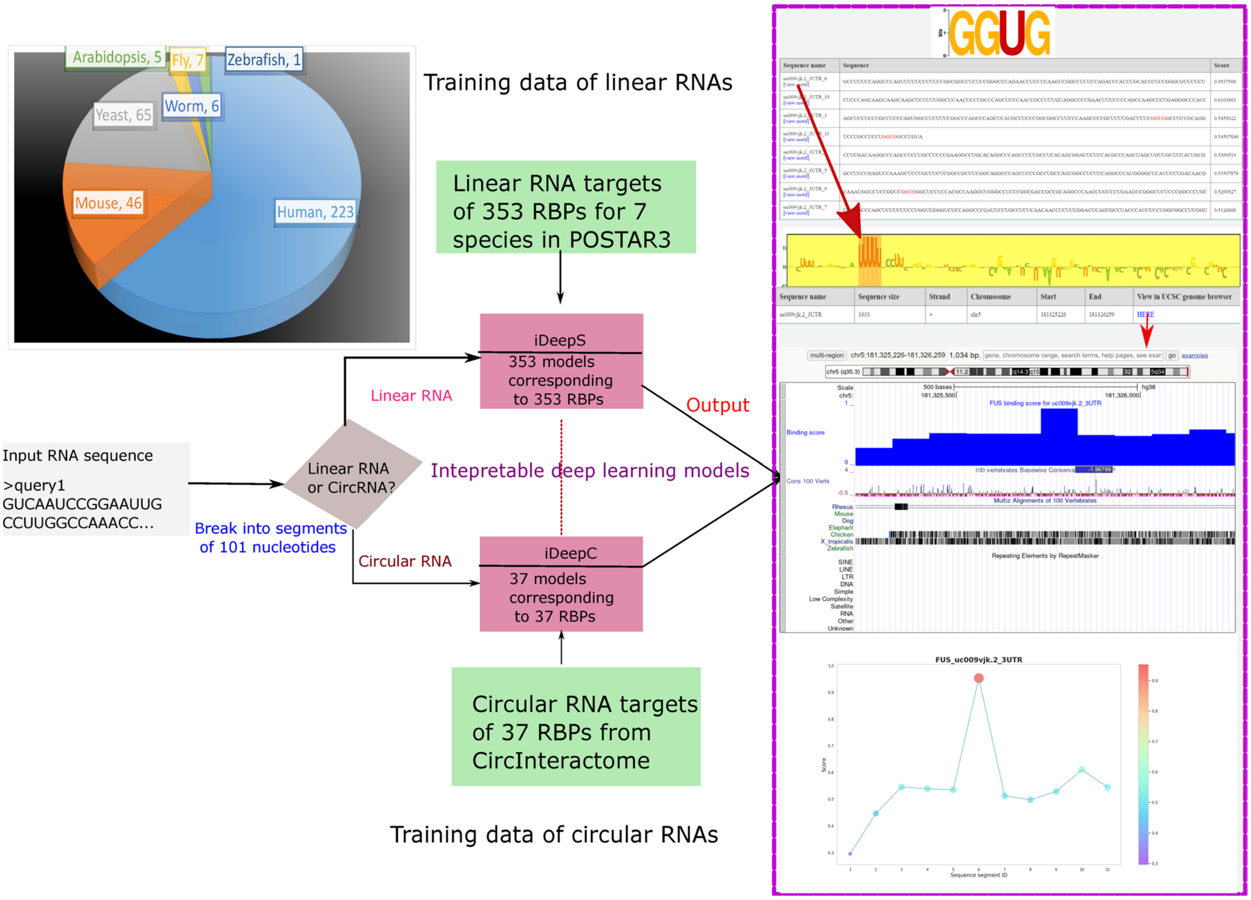

Arabidopsis, 5
Fly, 7
Zebrafish, 1
Yeast, 65
Worm, 6
Mouse, 46
Human, 223
Training data of linear RNAs
Linear RNA targets of 353 RBPs for 7 species in POSTAR3
iDeepS
353 models corresponding to 353 RBPs
Linear RNA
Output
Input RNA sequence
>query1
GUCAAUCCGGAAUUG
CCUUGGCCAAACC...
Linear RNA or CircRNA?
Intepretable deep learning models
Break into segments of 101 nucleotides
Circular RNA
iDeepC
37 models corresponding to 37 RBPs
Circular RNA targets of 37 RBPs from CircInteractome
Training data of circular RNAs
GGUG
Sequence name
Sequence
Score
Sequence size
Strand
Chromosome
Start
End
View in UCSC genome browser
HERE
FUS binding score for uc009vjk.2_3UTR
Binding score
FUS_uc009vjk.2_3UTR
Score
Sequence segment ID

Fig. 1 The flowchart of RBPsuite 2.0. RBPsuite 2.0 first breaks the input RNA sequence into 101-nt fragments, which are fed into deep models to predict binding scores for a specified RBP. RBPsuite 2.0 lists the predicted binding scores and contribution scores of nucleotides for all the 101nt fragments, links the results to the UCSC browser for better visualization, and shows the binding scores for individual fragments

ENCODE in BED format [3]. To expand the coverage of RBP binding predictions of RBPsuite, we update RBPsuite 2.0 by systematically integrating binding sites of 351 RBPs of seven species (i.e., human, mouse, zebrafish, fly, worm, *Arabidopsis*, yeast) [36, 37]. These binding sites are downloaded from the CLIPdb module of POSTAR3 databases [36, 37]. POSTAR3 identified the RBP binding sites from 1499 CLIP-seq datasets, which has covered 10 various CLIP-seq technologies (i.e., HITSCLIP, PAR-CLIP, iCLIP, eCLIP, iCLAP, urea-iCLIP, 4sUiCLIP, BrdU-CLIP, Fr-iCLIP and PIP-seq) [37]. The genome version of the seven species corresponds to hg38 (human), mm10 (mouse), danRer11 (zebrafish), dm6 (fly), ce11 (worm), TAIR10 (*Arabidopsis*), and sacCer3 (yeast), respectively. The RBP binding sites downloaded from POSTAR3 CLIPdb include information about the chromosome name, start and end positions of the feature in standard chromosomal coordinates, site name, genome strand, RBP name, CLIP-seq technologies, analysis software, and GEO accession number.

To prepare the positive and negative sequence for model training and evaluation, the RBP binding sites were processed as follows. (1) We split RBP binding sites set to obtain peaks of each RBP. (2) For each RBP, we select the sites completely contained by a transcript by intersect of pybedtools [38, 39]. As one RBP binding site may correspond to multiple transcripts, the transcript with the maximum length of each RBP binding site was retained. (3) The peaks are extended to 101 nt by introducing random padding on both sides to ensure that the position of the binding sites sequence on the 101 nt segment is not fixed. (4) Negative RBP binding regions were produced by implementing a shuffle of pybedtools, these negative sites are those regions without any identified binding peak located within the same transcript, and we randomly select the same number of negative regions as the positive samples. (5) The sequences of positive and negative regions were retrieved by the fetch function of pysam, which is a wrapper for htslib [40] and the samtools [41] package. (6) After removing the existing RBPs in RBPsuite 1.0, we obtain 69 novel human RBPs, and the RBPs of the other six species are all kept in RBPsuite 2.0. (7) Finally, RBPsuite 2.0 now supports RBP binding sites prediction on linear RNAs for a total of 353 RBPs (i.e., 223 human RBPs, 46 mouse RBPs, 1 zebrafish RBPs, 7 fly RBPs, 6 worm RBPs, 5 *Arabidopsis* RBPs, and 65 yeast RBPs).

Moreover, we collect experimentally verified binding motifs of RBPs from ATtRACT [42] and CISBP-RNA [43] databases, and they are further scanned against the sequence segments to locate the binding sites using FIMO in MEME suite [44] with$p$-value < 0.01.

Model training and evaluation

We train the backend prediction models in the same way as RBPsuite 1.0. For each RBP, we split its positive and negative binding sequences into the training set, validation set and test set with a ratio 7:1:2, where the model is trained on the training set, optimized on the validation set, and evaluated on the test set. For linear RNAs, iDeepS is trained on RBP-binding linear RNAs. In total, 223, 46, 65, 7, 6, 5, and 1 models are trained for human, mouse, yeast, fly, *Arabidopsis*, and zebrafish, respectively. For circRNAs, iDeepC is trained on RBP-binding circRNAs, where 37 models for 37 human RBPs are trained. The detailed hyper-parameters are given in Table 1. The main built-in models of RBPsuite 2.0, iDeepS and iDeepC, were trained on 4 NVIDIA GeForce RTX 3090 GPUs and took approximately 3 days to complete.

iDeepS for predicting RBP binding sites on linear RNAs

iDeepS integrates RNA sequences and predicted secondary structures to infer RBP binding sites on linear RNAs. It first uses CNNs to extract abstract features, which are fed into LSTM to capture long dependency between sequences and structures. In addition, iDeepS mines the binding sequence and structure motifs from the learned filters of CNNs, where these motifs align well with experimentally verified motifs. Moreover, as benchmarked in PrismNet [18], iDeepS achieves the best performance among sequence-based methods, which demonstrate that iDeepS is still a competitive method for predicting RBP binding sites on linear RNAs. The network module of iDeepS is defined as below:

$$f_\theta(\mathrm{x}) = \mathrm{LSTM}\left(\mathrm{Concat}\left(\mathrm{Conv}\left(x_{seq}\right),\ \mathrm{Conv}(x_{stru})\right)\right) \quad (1)$$

where Conv is the convolution operation, LSTM is the LSTM layer, Concat is the concatation operation, x_{seq} and x_{stru} is the one-hot encoding of the input sequence and predicted secondary structure.

Table 1 Hyper-parameters of CRIP, iDeepS, and iDeepC in RBPsuite 1.0 and 2.0

	CRIP	iDeepS	iDeepC
Batch size	50	50	128
Epoch	30	30	20
Loss function	Categorical cross-entropy loss	Categorical cross-entropy loss	Cross-entropy loss
Early stopping	5	5	5
Learning rate	0.0004	0.01	0.0004

iDeepC for predicting RBP binding sites on circRNAs

In RBPsuite 1.0, it uses CRIP as the backend predictor for RBP sites on circRNAs. However, it performs well only on the RBPs with a sufficient number of known binding circRNAs. To make the predictor work well for the RBPs with only a limited number of known binding circRNAs, RBPsuite 2.0 uses iDeepC as the backend predictor. iDeepC designs a Siamese neural network-based model to predict RBP binding sites from sequences with a pre-training strategy. The model mainly consists of a CNN module with an attention mechanism and a metric module, which captures the mutual information between circRNAs with pairwise metric learning. For model training, the model is first initialized with a pre-trained model of C22ORF28, which is further trained with the RBP-specific binding circRNAs. Given a test sequence, iDeepC calculates its similarities based on learned embeddings to multiple known positive sequences in the support set, and the pairwise similarity differences are inputted into the learned metric module to estimate the RBP binding probability. The benchmark results demonstrate that iDeepC achieves promising performance on the poorly characterized RBPs with a limited number of binding circRNAs.

The network module of the Siamese network in iDeepC consists of two-layer CNNs and a lightweight attention layer [45]. $f_\theta(\mathrm{x})$ is formulated as below:

$$f_\theta(x) = Conv(Conv(x)) \otimes a \quad (2)$$

where Conv is the convolution operation, a is the function of the attention,$\otimes$ is the multiplication.

Integrated gradient for highlighting key nucleotides for binding to RBPs

In addition to locating the verified motifs in the fragments using the MEME tool, we use an integrated gradient to highlight the key nucleotides contributing to the binding of RBPs. We apply an attribution method based on integrated gradient [46], which aims to explain the relationship between the model's predictions in terms of its features. Feature weights or nucleotide importance can be calculated by the integrated gradient method. We consider nucleotides with weights greater than the threshold as key nucleotides and highlight them in the visualization, where the threshold is set as a half of the maximum weight.

$$\mathrm{IMP}_i(\mathrm{x}) = \left(x_i - x_i'\right) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \tag{3}$$

where x is the input feature, i is the index of the i th feature, F is the trained model, and $x\prime$ is the baseline input.

In this study, we change the size of input features through linear interpolation and then calculate the backpropagation gradient of each feature. The larger the gradient is, the more important the feature is. An important operation of IG is that it does not draw a conclusion only based on one interpolation, but based on the integral value of the gradients within the interpolation range. Using IG, we can detect important nucleotides associated with RBP binding, which are further used to detect RBP binding motifs.

Linking to UCSC browser track for prediction result visualization

We used BLAT [47], a sequence alignment tool to search the input sequence for four species including human, mouse, fly, and yeast against their genomes as shown in Table 2. BLAT finds sequences with 95% similarity quickly with low memory usage and high performance. The input sequence will be located to the best match in the genomes and visualized in the UCSC genome browser [48] with an additional binding score track, which is convenient for users to obtain other relevant information such as adjacent genes, conservation information, and so on.

Implementation

All models based on neural networks, including iDeepS and iDeepC, were developed using Keras with a TensorFlow backend. RBPsuite 2.0 integrates two deep learning-based methods: the iDeepS for linear RNAs and iDeepC for circRNAs. The partial training parameters for iDeepS and iDeepC are listed in Table 2. Additionally, iDeepS uses the SGD optimizer to update model parameters during the backpropagation process, while iDeepC uses the Adam optimizer for the same purpose. All three models were trained with early stopping based on validation set loss after 5 epochs to prevent overfitting. For specific implementation details, please refer to the original code repositories for these models.

Table 2 The species that can be visualized in the UCSC browser supported by RBPsuite 2.0

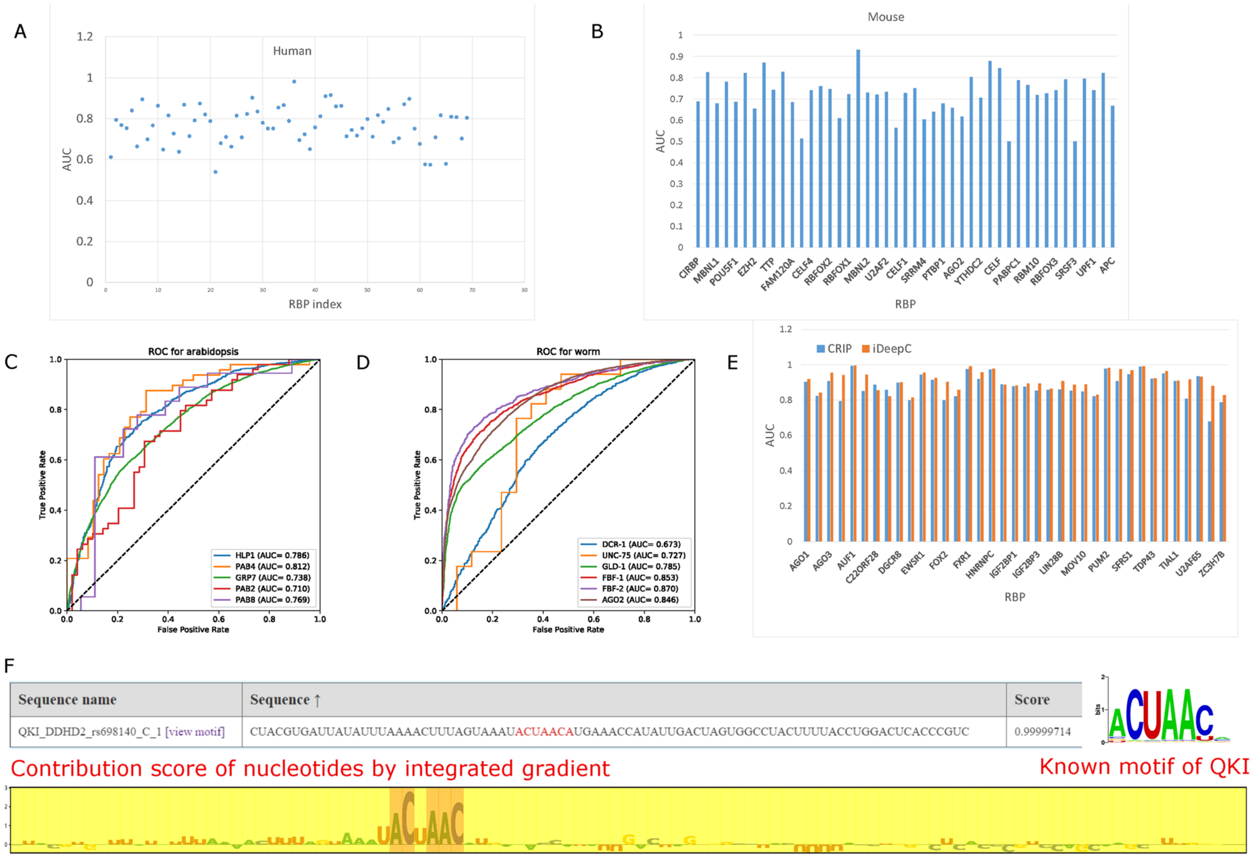
Species	Genome	Abbreviation
Human	GRCh38	hg38
Mouse	GRCmm39	mm39
Fly	BDGP Release 6 + ISO1 MT	dm6
Yeast	SacCer_Apr2011	sacCer3

Results

For each RBP, one special model is trained. RBPsuite 2.0 uses iDeepS as the RBP binding site prediction engine for linear RNAs, and the newly developed iDeepC for circular RNAs. We evaluate the prediction performance on newly constructed benchmark datasets for seven species, whereas existing methods have primarily focused on human data, neglecting other species. As benchmarked on 168 human RBP datasets for linear RNAs in the PrismNet study [18], iDeepS outperforms other seven sequence-based methods, i.e., DeepCLIP, pysster, and so on. Thus, here we only evaluate the performance of iDeepS for the newly added RBPs and species in RBPsuite 2.0. As shown in Fig. 2, RBPsutie 2.0 yields similar performance for seven species. It achieves an average area under the ROC curve (AUC) of 0.766 across 69 new human RBPs for linear RNAs, an average AUC of 0.722 across 47 mouse RBPs, an average AUC of 0.697 across seven fly RBPs, an average AUC of 0.763 across five *Arabidopsis* RBPs, an average AUC of 0.792 across six worm RBPs, an average AUC of 0.674 across 65 yeast RBPs and an AUC of 0.714 for one zebrafish RBP. The AUC distribution of human and mouse are shown in Fig. 2A and B, and the ROC curve for *Arabidopsis* and worm is illustrated in Fig. 2C and D. We can see that iDeepS in RBPsuite 2.0 achieves an average AUC of about 0.7 for all seven species, and for some RBPs, it yields an AUC of over 0.9.

Additionally, the performance of RBPsuite 2.0 across seven species, and the variation in AUC values across species, especially for yeast and zebrafish. Previous works used training sets from uniform sequencing technologies and analysis processes, whereas the updated data of other species in RBPSuite 2.0 comes from POSTAR3, which involves significant changes in sequencing technology and analysis processes, which raises the challenge for the generalizability of the RBPsuite 2.0.

As shown in Fig. 2E, the iDeepC in RBPsutie 2.0 increase the average AUC of 0.881 to 0.912 across 37 human RBPs for circRNAs, especially for some RBPs with a small number of binding circRNAs. For WTAP with a limited number of 496 binding circRNAs, iDeepC in RBPsuite 2.0 yields an AUC of 0.880, which is a relative increase of 29.8% compared to the AUC 0.678 of CRIP. The results validate the reason that we



Sequence name	Sequence ↑	Score
QKI_DDHD2_rs698140_C_1 [view motif]	CUACGUGAUUAUAUUUAAAACUUUAGUAAAUACUAACAUGAAACCAUAUUGACUAGUGGCCUACUUUUACCUGGACUCACCCGUC	0.99999714

Fig. 2 The performance of RBPsuite 2.0 on the benchmark datasets. **A** AUCs of 69 new human RBPs; **B** AUCs of 46 mouse RBPs; **C** ROC curve of five *Arabidopsis* RBPs; **D** ROC curve for six worm RBPs. **E** The AUC comparison between the original CRIP and the updated iDeepC in RBPsuite 2.0 for circRNAs of 37 human RBPs. **F** The detected motifs with integrated gradient in RBPsuite 2.0 for QKI binding DDHD2, where the detected potential motif aligns well with the known motifs of QKI

update for the backend predictor for circRNA in RBPsuite 2.0.

In addition, we investigate the detected motifs by an integrated gradient in RBPsuite 2.0. We use RBPsuite 2.0 to predict the binding score for DDHD2 with QKI, which has a known motif in the CISBP-RNA database. As shown in Fig. 2F, RBPsuite 2.0 is able to predict this RNA sequence binding with QKI with a high score over 0.999. In addition, the integrated gradient in RBPsuite 2.0 calculates the contribution scores of individual nucleotides for DDHD2, and the region with the highest contribution scores aligns well with the known motif of QKI. The results show that RBPsuite 2.0 is able to detect some biological motifs to locate the binding positions on the input sequence using an integrated gradient. In addition, the backend predictors iDeepS and iDeepC have been proven to successfully construct RNA-binding protein motifs based on learned sequence features, and these motifs are consistent with the experimentally verified motifs.

Input of RBPsuite 2.0

Given a RNA transcript sequence, RBPsuite 2.0 first break the input sequence into non-overlapping 101-nt fragments. Then, users need to specify the RNA type, "Linear RNA" or "Circular RNA" to choose iDeepS or iDeepC for predicting the RBP binding sites. Furthermore, for linear RNAs, users need to specify which species, i.e., human, mouse, yeast, fly, *Arabidopsis*, and zebrafish, should be used. In addition, we provide two prediction modes: RBP-specific models and the general model will run all available RBP-specific models for prediction.

Output of RBPsuite 2.0

For RBP binding site prediction, RBPsuite 2.0 provides two modes for the prediction output. One is the RBP-specific mode, where one model is trained per RBP. The other is the general model where RBPsuite 2.0 predicts the RBP binding sites of all available RBPs on RNAs.

Output of the RBP-specific model in RBPsuite 2.0

In addition to the specific RBPs with the trained models, we add specific models for unseen RBPs, where the prediction model of this RBP uses the specific model of the most similar RBP in the training set. The RBP-specific model will provide the following four types of outputs.

Motif logo

If the chosen RNA-binding protein has a verified motif, RBPsuite2.0 will provide the motif logo, which will be visualized along the input sequence.

Binding score table with motifs and by integrated gradient

In order to improve the efficiency of the prediction, RBPsuite2.0 divides the input sequence into segments of a length 101 without overlap and displays the predicted binding score of each segment with the RBP. These segments are arranged in descending order of binding scores, and segments with a score less than 0.5 are filtered out. When the RBP has a verified motif, RBPsuite2.0 marks the potential motifs in RED according to the position on the segment using the MEME FIMO tool. In addition, RBPsuite2.0 provides a sort function. You can click the name of each column to sort the results. RBPsuite2.0 also visualizes the nucleotide importance in the sequence segments calculated by integrated gradient and highlights the key nucleotides.

The best match in hg38 and visualization in the UCSC browser

For human sequence, RBPsuite2.0 provides the best match of the input sequence in hg38 by BLAT. The detailed information includes sequence length, chromosome, strand, start index, and end index. In addition, RBPsuite2.0 visualizes the input sequence with the binding score track (prediction scores on the sequence) linking to the UCSC genome browser.

Binding score visualization

RBPsuite2.0 provides the visualization for the binding scores in the segments along the input sequence, where one point refers to one 101-nt segment.

Output of the general mode in RBPsuite 2.0

The general mode has a directory interface that lists all the RBPs, which predict binding scores between the input RNA sequence and the model trained for this RBP. For more details about the prediction results of individual RBPs, the users can click the RBP of interest to see the predicted RBP binding sites of this RBP for the input sequence. The output of each RBP is the same as the specific model.

Case study on predicting RBP binding sites and SNP impact with RBPsuite 2.0

To demonstrate the usage of RBPsuite 2.0, we apply it to predict RBP binding sites on Myosin light chain 6 (MYL6) for RBFOX2. In pancreatic ductal adenocarcinoma, RBFOX2 serves as a crucial metastatic suppressor by regulated alternative splicing [49]. As RBFOX2 target gene, the isoform of MYL6 exhibits exon 6 skipping increased in RBFOX2-depleted cells [49]. To gain insights into the binding interaction between RBFOX2 and MYL6, as well as to identify the specific binding sites on MYL6. We utilize RBPsuite 2.0 to predict the binding sites between RBFOX2 and the nearby region of exon 6 on MYL6. Based on ENCODe eCLIP data on HepG2 and K562 [50], we obtain the RBFOX2 binding profile and the mean binding profile with a window size of 101 of MYL6 transcript. Notably, in HepG2 cells, the binding region of RBFOX2 is mainly located to the left of the exon 6. In K562 cells, exon 6 is located in the RBFOX2 binding region. To inspect the region surrounding exon 6, the first 1999 bp is trimmed, and the RBFOX2 binding sites on the remaining sequence are predicted by RBPsuite 2.0. As shown in Fig. 3A, RBPsuite 2.0 is able to identify two true binding sites, but it also predicts some false positive binding sites, which should be further improved in future updates of RBPsuite. One potential reason is that positive and negative samples are constructed using a fixed ratio such as 1:1. In reality, protein binding patterns on an RNA molecule are quite sparse, and the number of unbinding sites is much more than the number of binding sites.

The variants in RBP binding site can significantly impact RBP binding and cause diseases [51]. For instance, the variant at SNP rs6981405 C > A in DDHD2 3′ UTR has been shown to disrupt the binding of RBP protein quaking (QKI), which alters the abundance of mature DDHD2 mRNA, and is associated with an increased schizophrenia risk [51, 52]. As shown in Fig. 3B, the rs6981405 genotype is AA in K562 and is CC in HepG2 confirmed by RNAseq data from ENCODE (ENCSR366YOG and ENCSR570WLM). To evaluate the influence of rs6981405 on the QKI–DDHD2 3′ UTR interaction by RBPsuite 2.0, we expand the sequence by 50 bp around rs6981405 to obtain a segment with a length of 101 bp. with the variant changing from C to A, the predicted binding score with QKI decreased from 0.536 to 0.429 (Fig. 3C). The results show that RBPsuite

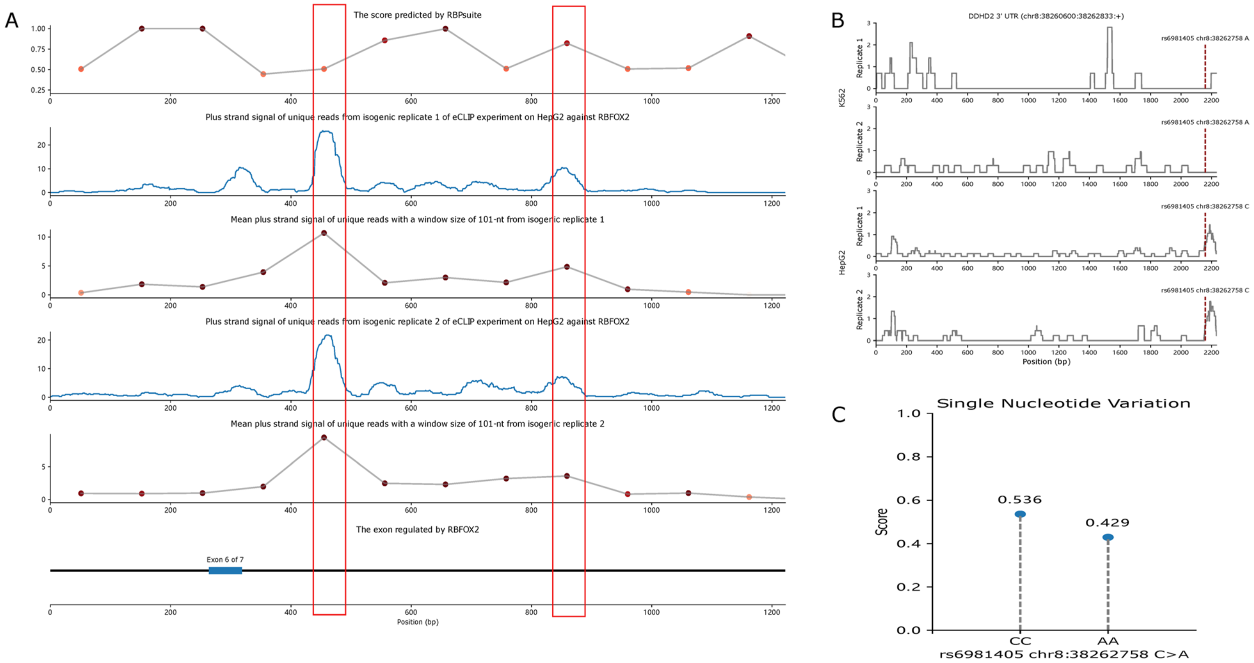


Fig. 3 Applying RBPsuite 2.0 on RBP binding site and SNV impact prediction. **A** The prediction results of RBPsuite 2.0 for predicting RBFOX2 binding sites in the region around exon 6 of MLY6. **B** The eCLIP reads of QKI on DDHD2 before and after the rs6981405; **C** The predicted binding scores with QKI before and after the rs6981405 on DDHD2 using RBPsuite 2.0, where DDHD2 without rs6981405 mutation binds to QKI

2.0 may be applied to infer the SNV impact on the RBP binding sites.

Discussion

RNA-binding proteins (RBPs) are highly involved in various regulatory processes, e.g., gene splicing and localization, and provide important functional information for patient care. In this study, we updated our previous webserver RBPsuite for predicting RBP binding sites from RNA sequences. In this study, updated RBPsuite 2.0 has a higher coverage on the number of supported RBPs and species than the original RBPsuite, supporting an increased number of RBPs from 154 to 353 and expanding the supported species from one to seven. In addition, RBPsuite 2.0 replaces the CRIP built into RPBsuite 1.0 with iDeepC, a more accurate RBP binding site predictor for circular RNAs. RBPsuite 2.0 currently supports only seven species for linear RNAs and one species for circRNAs. The main reason is due to the lack of training data derived from high-throughput data for protein-RNA interactions. The acquisition of high-throughput data for protein-RNA interactions is costly, and obtaining data on multiple protein-RNA binding events from a single experiment is even more challenging. For instance, in the commonly used method for detecting protein-RNA interactions, eCLIP-seq, the ENCODE project has thus far only generated 252 eCLIP-seq datasets involving 168 RBPs derived from K562, HepG2, and human adrenal gland tissues. Although other studies have also explored various sequencing techniques to obtain high-throughput protein-RNA interaction data, differences in experimental methods and data analysis approaches often result in significant variability in the data. Therefore, acquiring comprehensive high-throughput data on protein-RNA interactions across different species will require the accumulation of more sequencing data through ongoing biological research.

In future updates, we will incorporate more species and RBPs if we can collect training binding sites for them. To date, most of the existing deep learning-based methods for RBP binding site prediction are still limited to those RBPs with some known binding targets, and they cannot make predictions for those RBPs without any verified binding targets. To address this challenge, zero-shot learning may be applied to enable RBP binding site prediction of RBPs without any known binding targets. In addition, some domain knowledge can be incorporated, which can guide the predictors to learn true binding patterns without requiring many training samples.

We provide comprehensive introductions and usage instructions for RBPsuite 2.0. In the introduction section, users can gain a detailed understanding of the principles behind RBPsuite 2.0’s predictions for linear and circular RNA interactions with proteins. In the dataset section, we provide a thorough explanation of the training data used in the RBPsuite 2.0, along with convenient download links. The Help page offers a clear, step-by-step guide for using RBPsuite, which can be easily followed with the example sequences in the prediction interface. As a result, RBPsuite 2.0 has a low learning curve and is easy to use. Additionally, we respond promptly to user inquiries via email to address any questions.

Conclusion

In this work, we have updated our previous webserver, RBPsuite, to predict RBP binding sites from RNA sequences with enhanced functionality. The upgraded RBPsuite 2.0 now supports seven species (human, mouse, zebrafish, fly, worm, yeast, and Arabidopsis), expanding beyond the original focus on humans. It also includes a comprehensive collection of 223 human RBPs. For circRNAs, RBPsuite 2.0 replaces the CRIP prediction engine with iDeepC, achieving improved prediction performance. To further enhance the interpretability of RBPsuite 2.0 predictions, the platform enables users to visualize the genomic context of any RBP binding site through the UCSC Genome Browser. It also provides detailed insights into the contributions of individual nucleotides to RBP-RNA interactions. These advancements are expected to deepen our understanding of circRNA-protein interactions and uncover the intricate mechanisms of functional regulation in organisms.

Abbreviations

AUC	Area under the ROC curve
CNN	Convolutional neural network
LSTM	Long-short temporary network
BERT	Bidirectional encoder representations from transformer
BiLSTM	Bidirectional long short-term memory
CircRNAs	Circular RNAs
RIP RNA	Immunoprecipitation
MYL6	Myosin light chain 6
QKI	Quaking
SGD	Stochastic gradient descent

Acknowledgements

We would like to thank the anonymous reviewers and editors, whose constructive feedback has been very helpful in enhancing the presentation of this paper.

Authors' contributions

HBS and XP conceptualized the project. XP and XL collected and processed the datasets. XP trained the models, and updated RBPsuite 2.0. YF and XG developed the web application for RBPsuite 2.0. HBS and XP supervised and guided the project. XP drafted the manuscript with assistance from YF and XL, and received feedback from HBS. All authors reviewed and approved the final manuscript.

Funding

This work is supported by the National Natural Science Foundation of China (No. 62473257, 62073219), and the Science and Technology Commission of Shanghai Municipality (24ZR1435300, 22511104100).

Data availability

All data generated or analyzed during this study are included in this published article, its supplementary information files and publicly available repositories. The online webserver and data are available at http://www.csbio.sjtu.edu.cn/bioinf/RBPsuite/ and https://doi.org/10.5281/zenodo.14216847. The source codes are available at https://github.com/hehew321/iDeepC, https://github.com/xypan1232/iDeepS2 or https://doi.org/10.5281/zenodo.14949530. The original data of POSTAR3 CLIPdb is at http://111.198.139.65/RBP.html, where POSTAR3 CLIPdb stores RNA-protein interaction datasets for seven species, all of which are available for download.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 20 May 2024 Accepted: 3 March 2025
Published online: 11 March 2025

References

1. Hentze MW, Castello A, Schwarzl T, Preiss T. A brave new world of RNA-binding proteins. Nat Rev Mol Cell Bio. 2018;19(5):327–41.
2. Mackenzie IR, Rademakers R, Neumann M. TDP-43 and FUS in amyotrophic lateral sclerosis and frontotemporal dementia. Lancet Neurol. 2010;9(10):995–1007.
3. Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004;306(5696):636–40.
4. Li JH, Liu S, Zhou H, Qu LH. Yang JH: starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res. 2014;42(Database issue):D92-97.
5. Pan X, Yang Y, Xia CQ, Mirza AH, Shen HB. Recent methodology progress of deep learning for RNA-protein interaction prediction. Wiley Interdiscip Rev RNA. 2019;10(6):e1544.
6. Ghanbari M, Ohler U. Deep neural networks for interpreting RNA-binding protein target preferences. Genome Res. 2020;30(2):214–26.
7. Maticzka D, Lange SJ, Costa F, Backofen R. GraphProt: modeling binding preferences of RNA-binding proteins. Genome Biol. 2014;15(1): R17.
8. Zhang S, Zhou J, Hu H, Gong H, Chen L, Cheng C, Zeng J. A deep learning framework for modeling structural features of RNA-binding protein targets. Nucleic Acids Res. 2016;44(4):e32.
9. Budach S, Marsico A. pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks. Bioinformatics. 2018;34(17):3035–7.
10. Yu H, Wang J, Sheng Q, Liu Q, Shyr Y. beRBP: binding estimation for human RNA-binding proteins. Nucleic Acids Res. 2019;47(5):e26.
11. Kazan H, Ray D, Chan ET, Hughes TR, Morris Q. RNAcontext: A New Method for Learning the Sequence and Structure Binding Preferences of RNA-Binding Proteins. Plos Comput Biol. 2010;6(7):e1000832.
12. Strazar M, Zitnik M, Zupan B, Ule J, Curk T. Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins. Bioinformatics. 2016;32(10):1527–35.
13. Horlacher M, Wagner N, Moyon L, Kuret K, Goedert N, Salvatore M, Ule J, Gagneur J, Winther O, Marsico A. Towards in silico CLIP-seq: predicting protein-RNA interaction via sequence-to-signal learning. Genome Biol. 2023;24(1):180.
14. Wang YX, Chen Z, Pan ZQ, Huang SJ, Liu J, Xia WQ, Zhang HN, Zheng MY, Li HL, Hou TJ, et al. RNAincoder: a deep learning-based encoder for RNA and RNA-associated interaction. Nucleic Acids Res. 2023;51(W1):W509–19.

15. Horlacher M, Cantini G, Hesse J, Schinke P, Goedert N, Londhe S, Moyon L, Marsico A: A Systematic Benchmark of Machine Learning Methods for Protein-RNA Interaction Prediction. Brief Bioinform. 2023;24(5):bbad307.
16. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831-+.
17. Pan X, Rijnbeek P, Yan J, Shen HB. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. BMC Genomics. 2018;19(1):511.
18. Xu Y, Zhu J, Huang W, Xu K, Yang R, Zhang QC, Sun L. PrismNet: predicting protein-RNA interaction using in vivo RNA structural information. Nucleic Acids Res. 2023;51(W1):W468–77.
19. Wu H, Pan X, Yang Y, Shen HB. Recognizing binding sites of poorly characterized RNA-binding proteins on circular RNAs using attention Siamese network. Briefings in Bioinformatics. 2021;22:bbab279.
20. Zhu H, Yang Y, Wang Y, Wang F, Huang Y, Chang Y, Wong KC, Li X. Dynamic characterization and interpretation for protein-RNA interactions across diverse cellular conditions using HDRNet. Nat Commun. 2023;14(1):6824.
21. Yamada K, Hamada M. Prediction of RNA-protein interactions using a nucleotide language model. Bioinform Adv. 2022;2(1):vbac023.
22. Zhang J, Liu B, Wang Z, Lehnert K, Gahegan M. DeepPN: a deep parallel neural network based on convolutional neural network and graph convolutional network for predicting RNA-protein binding sites. BMC Bioinformatics. 2022;23(1):257.
23. Gronning AGB, Doktor TK, Larsen SJ, Petersen USS, Holm LL, Bruun GH, Hansen MB, Hartung AM, Baumbach J, Andresen BS. DeepCLIP: predicting the effect of mutations on protein-RNA binding with deep learning. Nucleic Acids Res. 2020;48(13):7099–118.
24. Armaos A, Colantoni A, Proietti G, Rupert J. Tartaglia GG: catRAPID omics v2.0: going deeper and wider in the prediction of protein-RNA interactions. Nucleic Acids Res. 2021;49(W1):W72–9.
25. Paz I, Kosti I, Ares M, Cline M, Mandel-Gutfreund Y. RBPmap: a web server for mapping binding sites of RNA-binding proteins. Nucleic Acids Res. 2014;42(W1):W361–7.
26. Gronning AGB, Doktor TK, Larsen SJ, Petersen USS, Holm LL, Bruun GH, Hansen MB, Hartung AM, Baumbach J, Andresen BS. DeepCLIP: predicting the effect of mutations on protein-RNA binding with deep learning. Nucleic Acids Res. 2020;48(13):7099–118.
27. Sun L, Xu K, Huang W, Yang YT, Li P, Tang L, Xiong T, Zhang QC. Predicting dynamic cellular protein-RNA interactions by deep learning using in vivo RNA structures. Cell Res. 2021;31(5):495–516.
28. Pan X, Fang Y, Li X, Yang Y, Shen HB. RBPsuite: RNA-protein binding sites prediction suite based on deep learning. BMC Genomics. 2020;21(1):884.
29. Ovrebo JI, Bradley-Gill MR, Zielke N, Kim M, Marchetti M, Bohlen J, Lewis M, van Straaten M, Moon NS, Edgar BA. Translational control of E2f1 regulates the Drosophila cell cycle. Proc Natl Acad Sci U S A. 2022;119(4):e2113704119.
30. Forester CM, Oses-Prieto JA, Phillips NJ, Miglani S, Pang X, Byeon GW, DeMarco R, Burlingame A, Barna M, Ruggero D. Regulation of eIF4E guides a unique translational program to control erythroid maturation. Sci Adv. 2022;8(51):eadd3942.
31. Jiang LYQ, Xiao M, Liao QQ, Zheng LQ, Li CY, Liu YM, Yang B, Ren AM, Jiang C, Feng XH. High-sensitivity profiling of SARS-CoV-2 noncoding region-host protein interactome reveals the potential regulatory role of negative-sense viral RNA. Msystems. 2023;8(4):e0013523.
32. Du X, Zhou P, Zhang H, Peng H, Mao X, Liu S, Xu W, Feng K, Zhang Y. Downregulated liver-elevated long intergenic noncoding RNA (LINC02428) is a tumor suppressor that blocks KDM5B/IGF2BP1 positive feedback loop in hepatocellular carcinoma. Cell Death Dis. 2023;14(5):301.
33. Chen R, Yang TT, Jin B, Xu WR, Yan YW, Wood N, Lehmann HI, Wang SQ, Zhu XL, Yuan WL, et al. CircTmeff1 Promotes Muscle Atrophy by Interacting with TDP-43 and Encoding A Novel TMEFF1–339aa Protein. Adv Sci. 2023;10(17):e2206732
34. Zhang K, Pan X, Yang Y, Shen HB. CRIP: predicting circRNA-RBP-binding sites using a codon-based encoding and hybrid deep neural networks. RNA. 2019;25(12):1604–15.
35. Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, Kaul R, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583(7818):699-+.
36. Yang YC, Di C, Hu B, Zhou M, Liu Y, Song N, Li Y, Umetsu J, Lu ZJ. CLIPdb: a CLIP-seq database for protein-RNA interactions. BMC Genomics. 2015;16(1):51.
37. Zhao W, Zhang S, Zhu Y, Xi X, Bao P, Ma Z, Kapral TH, Chen S, Zagrovic B, Yang YT, et al. POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. Nucleic Acids Res. 2022;50(D1):D287–94.
38. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
39. Dale RK, Pedersen BS, Quinlan AR. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics. 2011;27(24):3423–4.
40. Bonfield JK, Marshall J, Danecek P, Li H, Ohan V, Whitwham A, Keane T, Davies RM. HTSlib: C library for reading/writing high-throughput sequencing data. Gigascience. 2021;10(2):giab007.
41. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. Genome Project Data Processing S. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
42. Giudice G, Sanchez-Cabo F, Torroja C, Lara-Pezzi E. ATtRACT-a database of RNA-binding proteins and associated motifs. Database (Oxford). 2016;2016:baw035.
43. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013;499(7457):172–7.
44. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res. 2015;43(W1):W39-49.
45. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. pp. 4510–4520.
46. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. Proceedings of International Conference on Machine Learning. 2017;70:3319–28.
47. Kent WJ. BLAT - The BLAST-like alignment tool. Genome Res. 2002;12(4):656–64.
48. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002;12(6):996–1006.
49. Jbara A, Lin KT, Stossel C, Siegfried Z, Shqerat H, Amar-Schwartz A, Elyada E, Mogilevsky M, Raitses-Gurevich M, Johnson JL, et al. RBFOX2 modulates a metastatic signature of alternative splicing in pancreatic cancer. Nature. 2023;617(7959):147–53.
50. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis C, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.
51. Park CY, Zhou J, Wong AK, Chen KM, Theesfeld CL, Darnell RB, Troyanskaya OG. Genome-wide landscape of RNA-binding protein target site dysregulation reveals a major impact on psychiatric disorder risk. Nat Genet. 2021;53(2):166–73.
52. Horlacher M, Wagner N, Moyon L, Kuret K, Goedert N, Salvatore M, Ule J, Gagneur J, Winther O, Marsico A: Towards In-Silico CLIP-seq: Predicting Protein-RNA Interaction via Sequence-to-Signal Learning. Genome Biol . 2023;24(1):180

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.