Genomics 114 (2022) 149–160

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Two novel RNA-binding proteins identification through computational prediction and experimental validation

Juan Xie [a], Xiaoli Zhang [a], Jinfang Zheng [a], Xu Hong [a], Xiaoxue Tong [a], Xudong Liu [a], Yaqiang Xue [b], Xuelian Wang [c], Yi Zhang [c], Shiyong Liu [a,*]

[a] School of Physics, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China
[b] Laboratory for Genome Regulation and Human Health, ABLife Inc., Wuhan, Hubei 430075, China
[c] ABLife BioBigData Institute, Wuhan, Hubei 430075, China

ARTICLE INFO

Keywords:
RNA-binding protein
RBPPred
CLIP1
DMD
iRIP-seq
CLIP-seq
ClinVar
Cancer

ABSTRACT

Since RBPs play important roles in the cell, it's particularly important to find new RBPs. We performed iRIP-seq and CLIP-seq to verify two proteins, *CLIP1* and *DMD*, predicted by RBPPred whether are RBPs or not. The experimental results confirm that these two proteins have RNA-binding activity. We identified significantly enriched binding motifs UGGGGAGG, CUUCCG and CCCGU for *CLIP1* (iRIP-seq), *DMD* (iRIP-seq) and *DMD* (CLIP-seq), respectively. The computational KEGG and GO analysis show that the *CLIP1* and *DMD* share some biological processes and functions. Besides, we found that the SNPs between *DMD* and its RNA partners may be associated with Becker muscular dystrophy, Duchenne muscular dystrophy, Dilated cardiomyopathy 3B and Cardiovascular phenotype. Among the thirteen cancers data, *CLIP1* and another 300 oncogenes always co-occur, and 123 of these 300 genes interact with *CLIP1*. These cancers may be associated with the mutations occurred in both *CLIP1* and the genes it interacts with.

1. Introduction

RNA plays a vital role in regulation of gene expression and interacts with other molecules such as RNA [1] and RNA-binding proteins (RBPs) [2]. It is worth noting that RBP-RNA interactions participate in numerous biological processes, like alternative splicing [3–5]. The disorder of this process may cause cancer [4]. The discovery of the RBP-RNA interactions will be of great benefit to understanding the mechanism behind this biological process.

Over the past few years, several high-throughput sequencing methods were designed to characterize RBP-RNA interactions. In CLIP-seq technique, the UV irradiation is used to covalently link RNA and RBP, and then high-throughput sequencing is performed on the enriched RNA [6–11]. Unlike CLIP-seq, RIP-seq is specific to a particular target protein. The protein-RNA complex is immunoprecipitated by the antibody of the target protein at first, and then the RBP-binding RNA is sequenced [12,13]. In recent years, scientists have developed other experimental methods for identifying RBP-RNA interaction and RNA-binding region using physicochemical properties of RBP and RNA [14–18], and these techniques of capturing RBPs-RNA interactions are no longer limited to polyadenylated RNAs. RBR-ID identified RNA-binding regions on 803 proteins, in which the cell was treated with 4-thiouridine while cross-linking [14]. In Chemistry-assisted RNA interactome capture (CARIC), alkyne is used to label RNA during photocross-linking making CARIC independent of the polyadenylation state of RNA in recognizing RBP-RNA interactions [15]. By this strategy, CARIC recovered 597 RBPs, in which 130 proteins are novel RBPs. Among these 130 novel RBPs, RBPPred predicts 51 of them are RBPs. Recently, Queiroz et al. exploited orthogonal organic phase separation (OOPS) to capture the proteins that can interact with RNAs and identified 1838 RBPs, in which 926 are putative novel RBPs [16]. A general purification method XRNAX for protein-crosslinked RNA discovered more than 700 RBPs that interact with non-polyadenylated RNA [18]. And uvCLAP does not rely on radioactive materials and can efficiently identify RBPs in vivo [19]. These high-throughput sequence techniques reveal hundreds of RBP-RNA interactions and will undoubtedly allow us to understand the function of the protein better.

In addition to the experimental methods, there are multiple computational approaches that can be applied to predict RNA-binding proteins. These approaches are divided into three categories. One is

* Corresponding author.
E-mail address: liushiyong@gmail.com (S. Liu).

https://doi.org/10.1016/j.ygeno.2021.12.003
Received 14 January 2021; Received in revised form 5 August 2021; Accepted 13 December 2021
Available online 16 December 2021

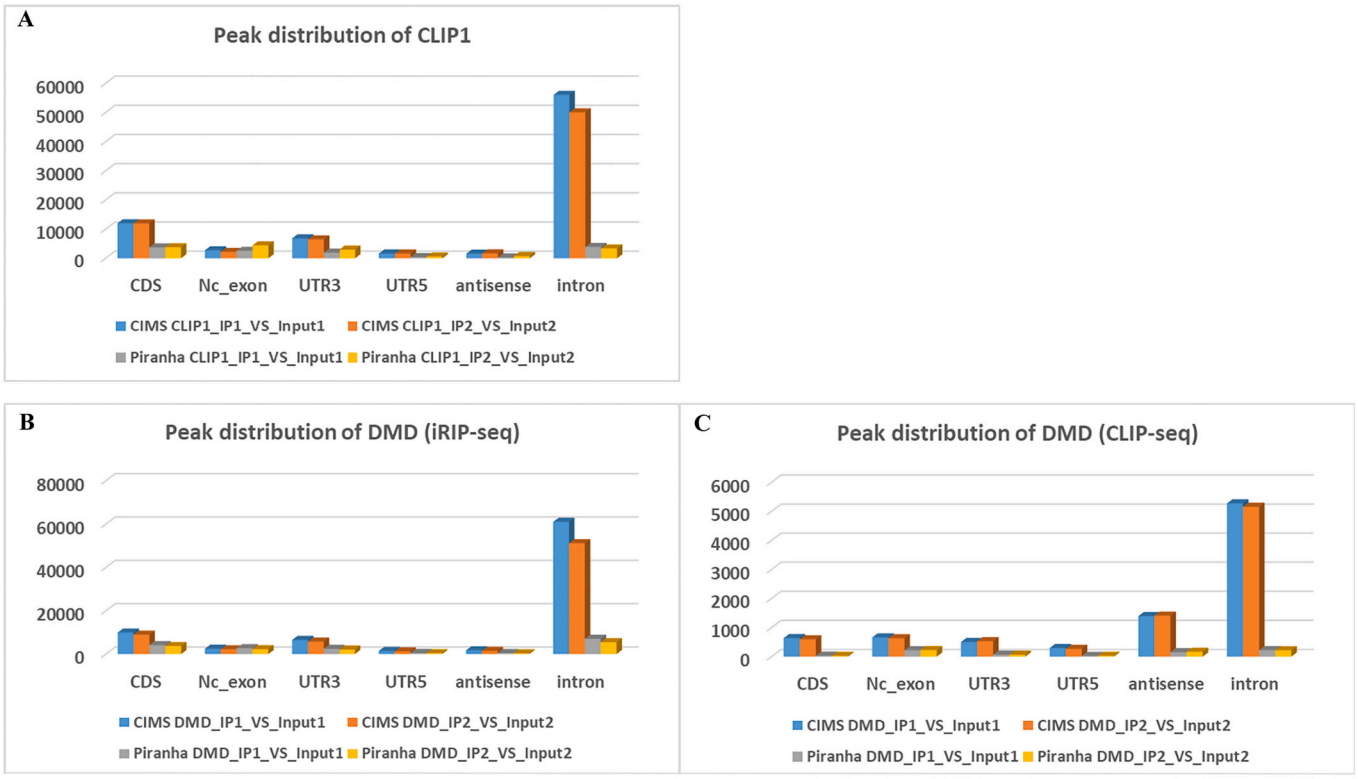

Fig. 1. Peaks annotation. The distribution of peaks is shown across reference genomic regions for *CLIP1* (iRIP-seq)(A), *DMD* (iRIP-seq)(B) and *DMD* (CLIP-seq) (C). The results show that the peaks of *CLIP1* and *DMD* are mainly distributed in the intron.

feature-based [20–26]. These methods select RBPs from non-RBPs by calculating sequences or structural features such as physicochemical properties, secondary structures, and evolutionary information. Another is template-based [27–32]. This type of method mainly determines whether the target binds RNA or not by comparing the difference between the sequence or the structure of the target and the template in the template library. And the other is the network-based SONAR [33], which relies on protein-protein interactions.

Our method RBPPred [21], which was based on the sequence features of given proteins, was proposed for judging whether a protein is an RBP or not. On an independent testing set of human, RBPPred achieved sensitivity of 0.84, specificity of 0.97 and MCC of 0.79. Among the top 500 potential RBPs predicted by RBPPred, 482 proteins have not been experimentally verified. We performed the operations as in Fig. S7. Finally, *CLIP1* and *DMD* were kept for iRIP-seq (improved RNA immunoprecipitation-coupled high-throughput sequencing) [34,35] and CLIP-seq verification experiment to verify the predictive power of RBPPred (see Methods) as previous researches [23,33,36]. iRIP-seq (see Methods) is established to validate whether *CLIP1* (UniProt ID: P30622) is RBP or not, iRIP-seq and CLIP-seq (see Methods) verified whether *DMD* (UniProt ID: P11532) [37] is indeed RBP or not. *CLIP1* was originally identified as a microtubule-binding protein by Rickard et al. in HeLa cells [38]. The C-terminal of this protein is important to combine with microtubule plus-ends. The N-terminal of this protein, which includes two Cap-Gly repeat domains with 57-amino acids, is linked to the C-terminus via the α-helical domain [39]. And CLIP1 is highly expressed in muscle [40]. Besides, Hyosuk Cho et al. found that *ANRIL* regulates monocyte adhesion to endothelial cells, transendothelial monocyte migration and endothelial cells migration by regulating the expression levels of *CLIP1* [41]. Another gene *DMD*, which can encode a muscular dystrophy protein, is one of the longest known human gene containing 2.4 million base-pairs and 79 exons. And the transcription and co-transcription splicing of *DMD* gene will take more than 16 h [37]. The protein dystrophin (encoded by *DMD*) was named in 1987 by Eric P. Hoffman et al. [42]. Although in normal skeletal muscle tissue, dystrophin only accounts for 0.002% of total muscle protein, but its deletion or mutation will lead to diseases such as Duchenne muscular dystrophy, which is no good treatment currently [43]. It was reported that in the presence of premature termination codons, the expression level of *DMD* gene is dropped, which makes the patient Duchenne Muscular Dystrophy (DMD) [44]. We used GEPIA2 [45] to analyze the expression of CLIP1 and DMD in cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC). It is found that the expression of CLIP1 and DMD in CESC will increase and decrease significantly compared with normal people. At different stages of CESC, the expression of CLIP1 and DMD remained basically unchanged (Fig. S9).

CLIP1 is predicted to be a non-RBP by SPOT-sEq. [27] and TriPepSVM [23], besides, Huang et al. [15] reported that *CLIP1* is a non-RBP too. Huang et al. identified that *CLIP1* has two or more uniquely peptides when they are analyzing the CARIC-enriched RBPs by quantitative proteomics. But *CLIP1* was filtered out when uniquely peptides were further measured by quantitative peptides. It may be due to their strict screening conditions, resulting in that partial RBPs are considered as non-RBPs. Conversely, Queiroz et al. [16] reported that *CLIP1* binds to RNA in MCF10A (a cell line from healthy individuals). *DMD* was predicted as a RBP by SPOT-seq and predicted as non-RBP by TriPepSVM.

CLIP1 and *DMD* are predicted as RNA-binding proteins by RBPPred. However, are they real RNA-binding proteins in cell? What are the RNA binding motifs for them? What are they doing when binding to RNAs? Are the mutations in CLIP1/DMD-RNA binding motifs associated with diseases? In this work, we are trying to address these questions by iRIP-seq, CLIP-seq experiment and bioinformatics analysis. *CLIP1* and *DMD* are experimentally validated as RBPs in iRIP-seq. In the CLIP-seq experiment, *DMD* is verified as RBP. By analyzing experimental data outputted by iRIP-seq, we found that the binding-reads/peaks of *CLIP1* and *DMD* are mainly distributed in introns. Motifs analysis indicated that CLIP1 binds the UGGGGAGG motif primarily in iRIP-seq, while

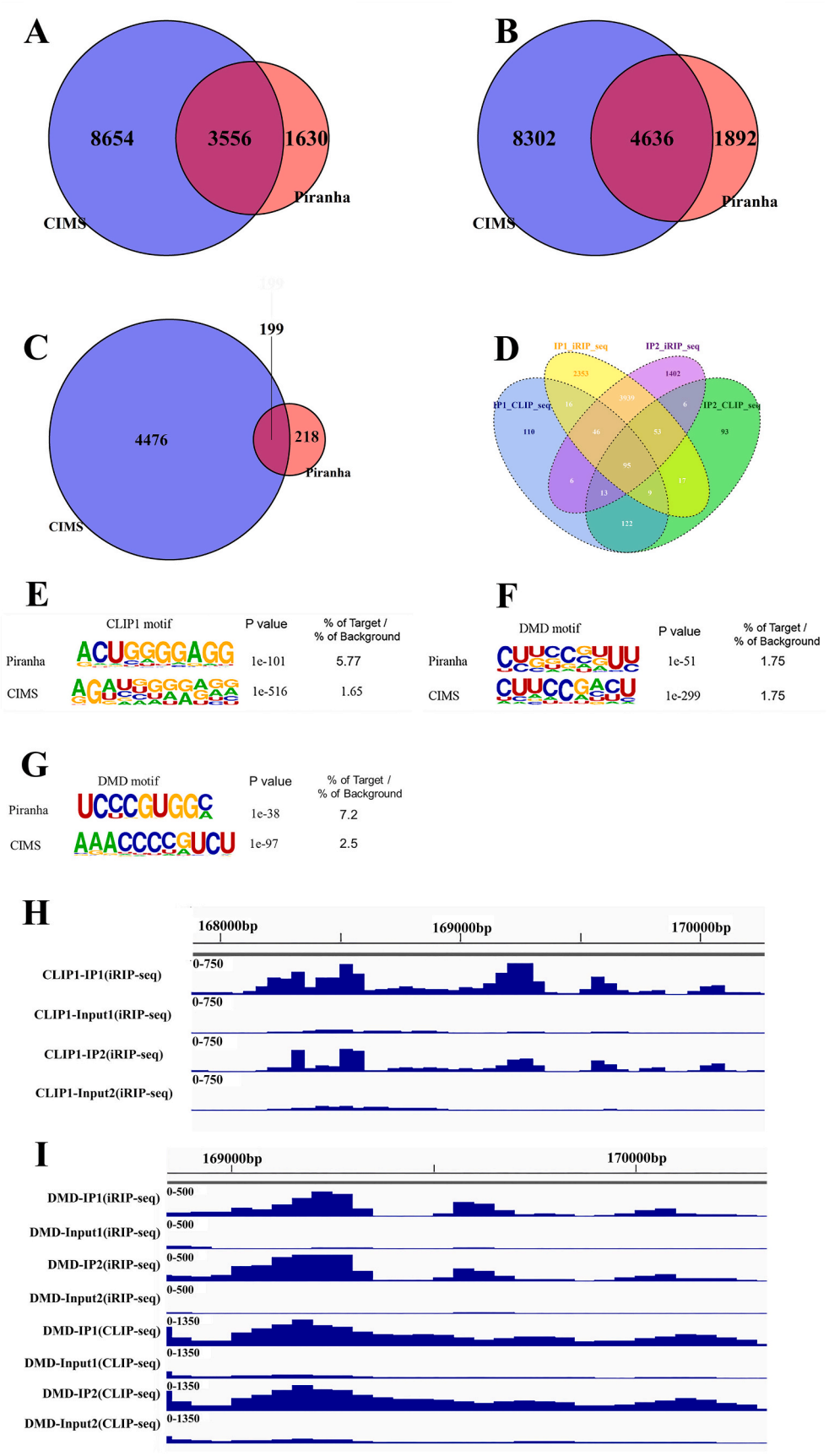

Fig. 2. Peaks gene and motif annotation. The peak genes are obtained from Piranha and CIMS analysis for *CLIP1* (iRIP-seq) (A), *DMD* (iRIP-seq) (B) and *DMD* (CLIP-seq) (C) and their overlapped genes are shown by Venn diagram. (D) shows the overlap of *DMD*-binding gene in repeated experiments of CLIP-seq and iRIP-seq, the results of Piranha are shown. The extracted peaks motifs are enriched by *CLIP1*-binding (iRIP-seq) (E), *DMD*-binding (iRIP-seq) (E) and *DMD*-binding (CLIP-seq) (F). (H) and (I) show the browser tracks of *CLIP1* and *DMD* on the KI270733.1 chromosome, which is shown by IGV.

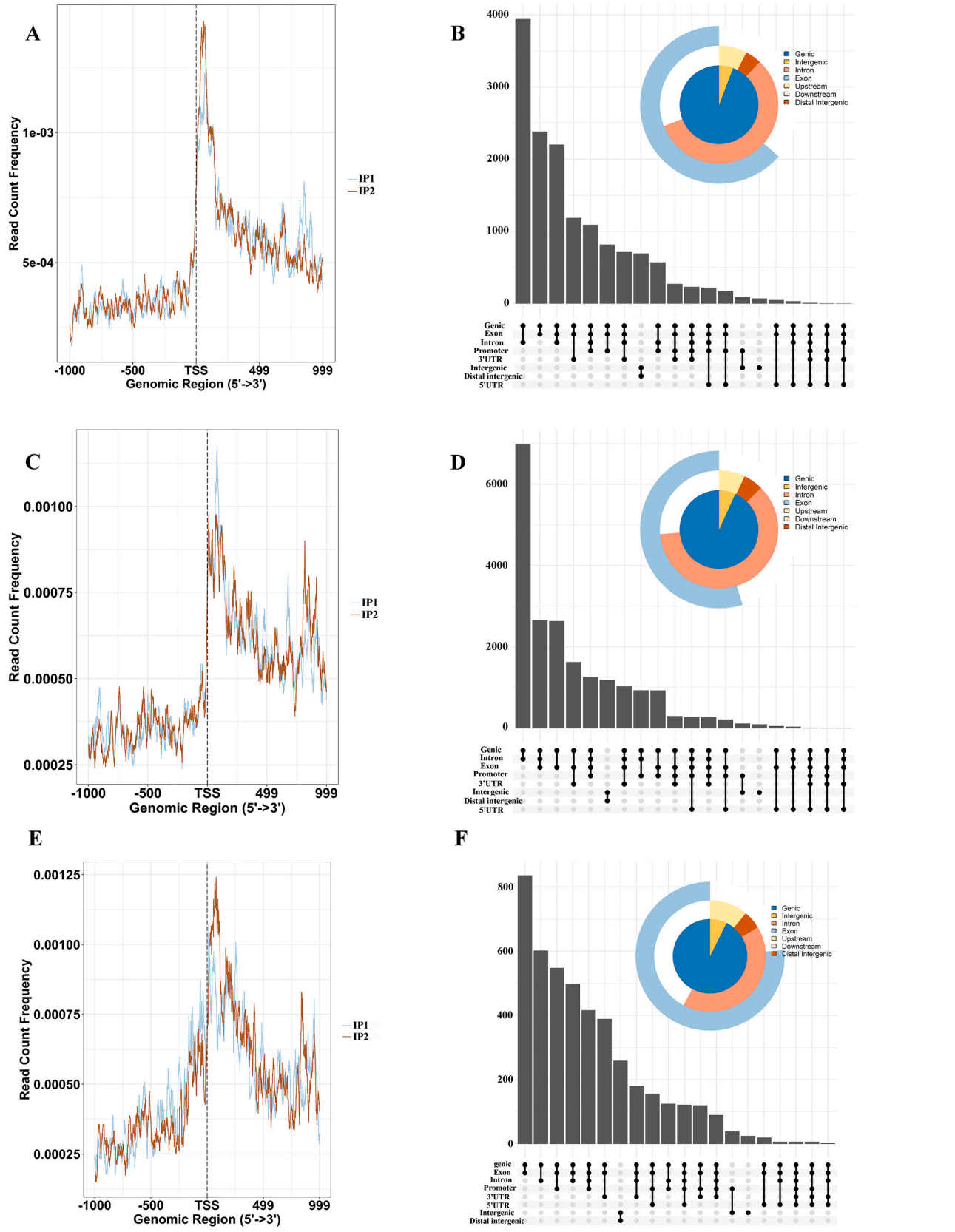

Fig. 3. The distribution of peaks near TSS. The density of the peaks is shown in 1 kb upstream and downstream of the TSS for *CLIP1* (iRIP-seq) (A), *DMD* (iRIP-seq) (C) and *DMD* (CLIP-seq) (E) respectively. (B), (D) and (F) describe the distribution of peaks of *CLIP1* (iRIP-seq), *DMD* (iRIP-seq) and *DMD* (CLIP-seq) respectively on the genome of 1 kb upstream and downstream of TSS. The pie charts in the upper right corner indicate the intersection of peaks on the genome. This corresponds to the black dotted line in the figure, and the number of intersections corresponds to the histogram portion in the figure. Piranha analysis results are shown.

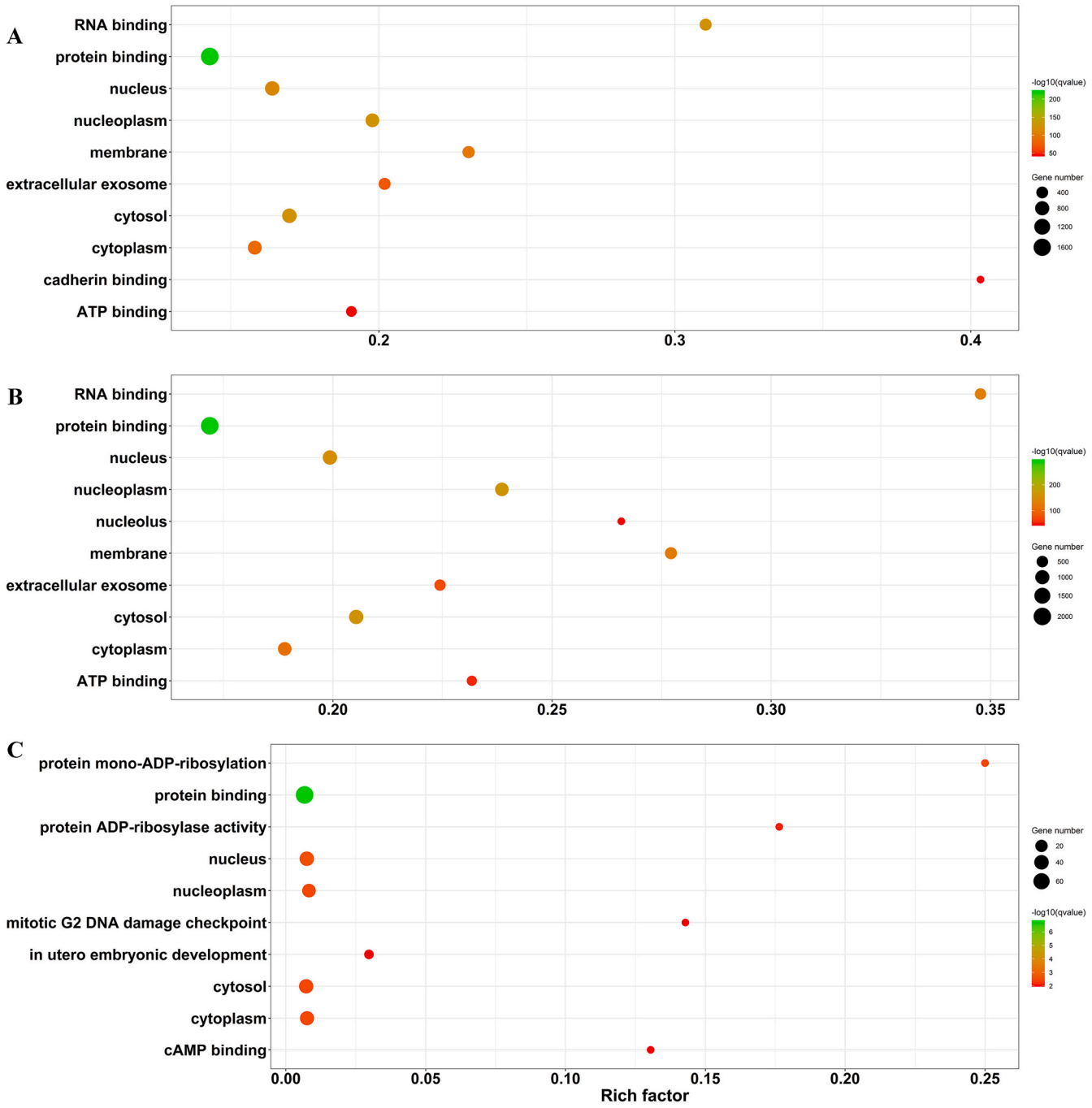

Fig. 4. Computational GO enrichment analysis. *CLIP1/DMD*-binding peak-related genes are used for GO and KEGG analysis, which are calculated by KOBAS 3.0 [76]. (A), (B) and (C) are the results of *CLIP1* (iRIP-seq), *DMD* (iRIP-seq) and *DMD* (CLIP-seq) GO enrichment analysis, respectively. The X axis represents the number of GO terms. The Y axis represents the type of GO term.

DMD binds the CUUCCG and CCCGU motifs mostly in iRIP-seq and CLIP-seq, respectively. In addition, we mapped *CLIP1/DMD*-binding RNAs to Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG). And the results showed that *CLIP1/DMD*-binding RNAs are overlapping in protein binding and nucleus terms. In addition to the above analysis, we analyzed the relationship between *CLIP1/DMD*-RNA and disease. We found that SNPs often appear at the *CLIP1/DMD*-binding sites. Mutations on the *DMD*-binding sites are likely to cause Becker muscular dystrophy, Duchenne muscular dystrophy, Dilated cardiomyopathy 3B and Cardiovascular phenotype and so on. In thirteen cancers, CLIP1 and another 300 genes co-occur and 123 of these 300 cancer genes are the binding partners of the CLIP1, in which genes co-occur refers to those cancers genes that always occur in the thirteen cancers. So, *CLIP1* may have a constant relationship with these diseases.

2. Results

2.1. Identification of DMD-RNA and CLIP1-RNA interaction by iRIP-seq and CLIP-seq

In order to verify the reliability of the RBPPred, we selected two proteins for validation (see Methods). The iRIP-seq (see Methods) and CLIP-seq techniques (see Methods) were utilized to detect the potential RBPs in HeLa cells. To get the more reliable experimental results, target

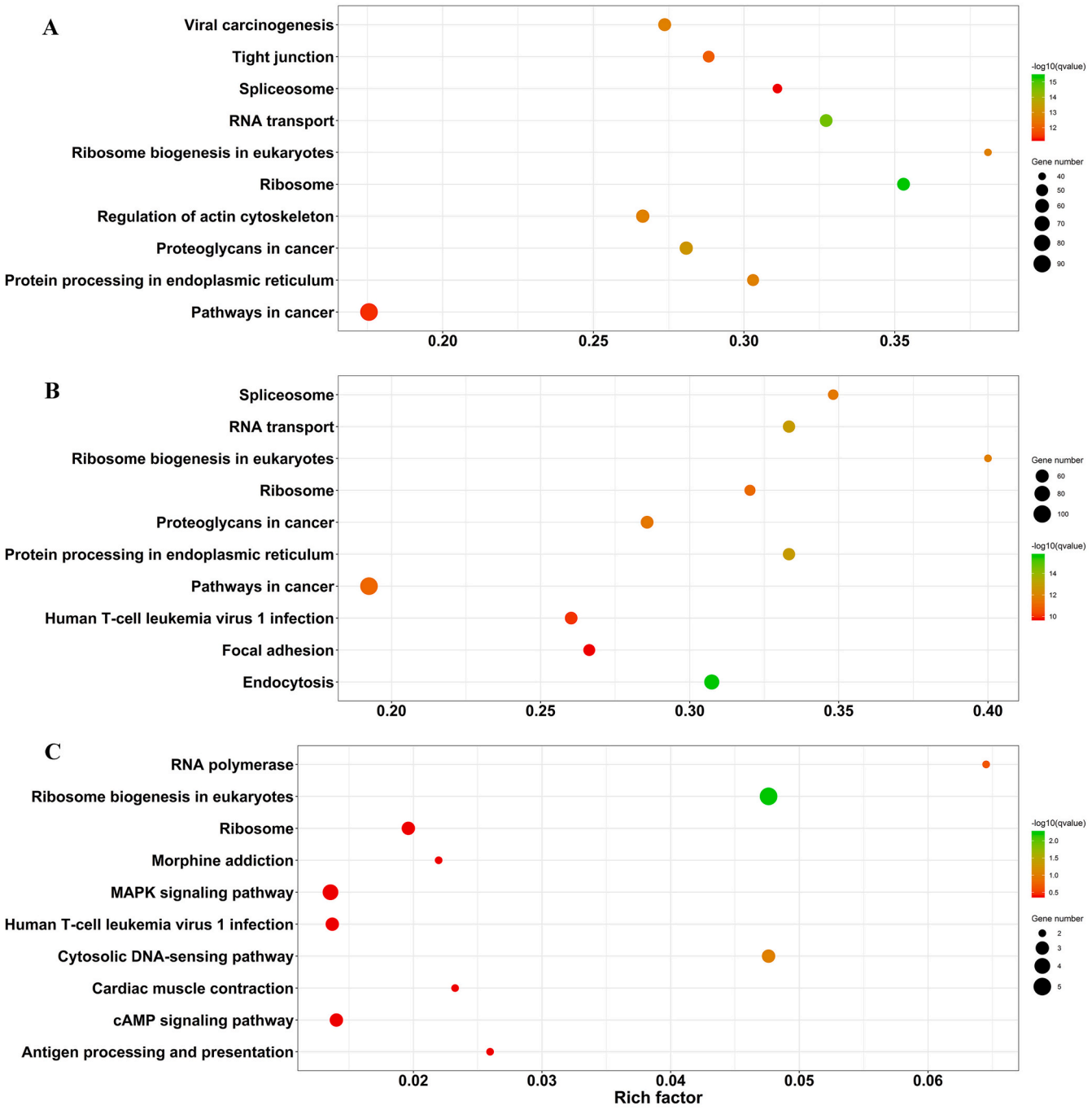

Fig. 5. Computational KEGG Pathway enrichment analysis. *CLIP1/DMD*-binding peak-related genes are used for KEGG analysis, which are calculated by KOBAS 3.0 [76]. (A), (B) and (C) are the results of KEGG pathway analysis for *CLIP1* (iRIP-seq), *DMD* (iRIP-seq) and *DMD* (CLIP-seq), respectively. The X-axis indicates the number of KEGG pathways in which *CLIP1/DMD*-binding peaks-related genes are located divided by the number of all genes in the pathway.

proteins were subjected to two biological replicates and input control in iRIP-seq and CLIP-seq, their western blots (Fig. S1) were used to test the immunoprecipitation efficacy. The experimental data is analyzed by phdRBP (Pipline for High-throughput Data analysis for RNA-Binding Protein), and finally *CLIP1/DMD*-binding peaks can be obtained.

2.2. Data pre-processing and mapping

In the pre-processing data section, adaptors, low quality reads and duplicate reads are trimmed from raw sequencing data (see Methods). The rest reads are called clean reads and are employed for further analysis. The cleaned reads are mapped to the human-GRCH38 genome using TopHat2 [46]. The genes of expressed reads are then counted with the htseq-count in the HTSeq package [47] (Supplementary Table S1). The cleaned reads display a broad range of distribution on the whole genome. We observed that the reads of both *CLIP1* and *DMD* are significantly enriched in the intronic region (Supplementary Table S2). The replicates of *CLIP1* and *DMD* show high correlation, with Pearson score greater than 0.98, which suggests that the iRIP-seq of *CLIP1* and the iRIP-seq and the CLIP-seq of *DMD* are highly reproducible (Fig. S2).

2.3. Identification of CLIP1/DMD-binding sites

To identify the regions which target proteins bind to, the peak calling

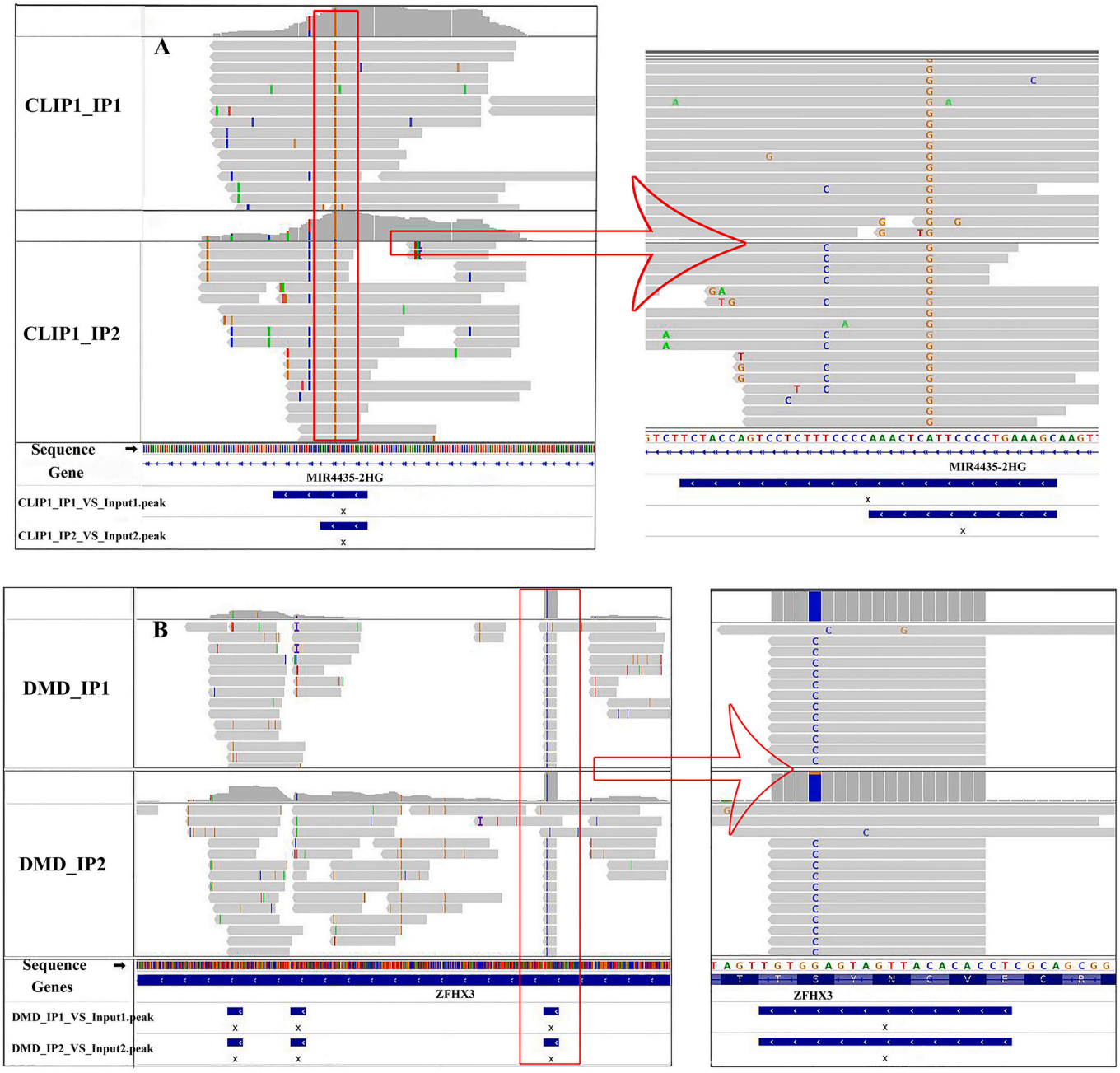

Fig. 6. SNP visualization. (A) shows mutations on *MIR4435-2HG* gene-coded RNA that interacts with *CLIP1*. This picture indicates the mutation from base A to base G at the site of presentation. (B) shows the mutation from base G to base C on the *ZFHX3* gene-coded RNA that interacts with *DMD*.

software Piranha [48] and CIMS [49] are used. The *CLIP1/DMD*-binding sites show that *CLIP1/DMD* bind RNAs in different genic regions but preferential enrichment in the intronic regions (Supplementary Table S2, Fig. 1A, Fig. 1B and Fig. 1C), suggesting that both of them may play roles in regulating pre-mRNA splicing. So, the Hum-mPLoc3.0[50] was used to predict the subcellular localization of *CLIP1* and *DMD*. The prediction results show that both *CLIP1* and *DMD* are located on the cytoplasm and cytoskeleton. For *CLIP1*, the prediction scores for the cytoplasm and cytoskeleton are 1.3078 and 0.6999. For *DMD*, the prediction scores for the cytoplasm and cytoskeleton are 2.5361 and 1.4852. Literatures [51,52] report that the splicing is mostly co-transcriptional in humans, and the splicing events are located in the cytoplasm. So, it may be reasonable that most of the peaks of *CLIP1* and *DMD* in intron may be associated with splicing.

For *CLIP1* (iRIP-seq), 15,457 peaks are mapped to 5186 genes from the result of Piranha analysis, among them 3556 genes are overlapped with the result of CIMS (Fig. 2A), *CLIP1* (iRIP-seq) refers to *CLIP1* using iRIP-seq experiment to verify. And 21,465 peaks refer to 6528 genes from the result of Piranha analysis for *DMD* (iRIP-seq), among them 4636 genes are overlapped with the result of CIMS (Fig. 2B), *DMD* (iRIP-seq) refers to *DMD* using iRIP-seq experiment to verify. Besides, 5052 peaks are related to 417 genes from the result of Piranha analysis for *DMD* (CLIP-seq), among them 199 genes are overlapped with the result of CIMS (Fig. 2C), *DMD* (CLIP-seq) refers to *DMD* using CLIP-seq experiment to verify. In addition, we found that 95 genes appeared in 4 repeated experiments of iRIP-seq and CLIP-seq (Fig. 2D, Fig. S3), and these 95 genes are listed in Supplementary Table S3. Furthermore, to explore the binding preferences of target proteins, we analyzed the RNA sequences of the peaks. We performed motif analysis using the HOMER software [53]. This identified the UGGGGAGG motif bound by *CLIP1*

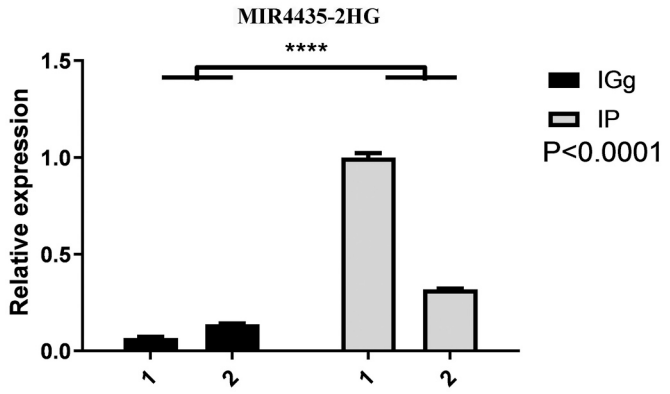

Fig. 7. The qPCR result of *CLIP1* binding to *MIR4435-2HG*. The results show that *MIR4435-2HG* was enriched by *CLIP*1, and Two-way ANOVA has significant difference.

(iRIP-seq) preferentially (Fig. 2E). In addition, *DMD* (iRIP-seq) is significantly crosslinked at the CUUCCG motif (Fig. 2F). The hexanucleotide motif CUUCCG, which is complementary to the 3′- terminal end of the18 S ribosomal RNA, has been suggested as a potential ribosome attachment site on the mRNA [54]. *DMD* (CLIP-seq) is mainly combined with CCCGU motif (Fig. 2G). In Fig. 2H and Fig. 2I, it shows genome browser tracks for *CLIP1* and *DMD* respectively in KI270733.1 chromosome, in which we use RPKM to measure gene expression.

In addition, we analyzed the distribution of peaks in 1 kb upstream and downstream of the transcription start site (TSS) (Fig. 3A, Fig. 3C, Fig. 3E, Fig. S4A, Fig. S4C and Fig. S4E) like previous research [55]. It shows that peaks enrich in downstream of TSS. In order to show the distribution of peaks near TSS intuitively, we used ChIPseeker [56] to plot the peaks distribution of 1 kb upstream and downstream of TSS (Fig. 3B, Fig. 3D, Fig. 3F, Fig. S4B, Fig. S4D and Fig. S4F). As we can see from the these figures, the peaks are mainly distributed in exon and intron, and some peaks are distributed at the exon-intron boundary, in which peaks may be related to alternative splicing.

To obtain the overlapped peaks between two replicate samples, we used BEDTools [57] to calculate the overlap of peaks (Fig. S5). From the Fig. S5, we can see that for *CLIP1* (iRIP-seq), there are 3800 genes overlapped and 6907 peaks overlapped. For *DMD* (iRIP-seq), the overlapped genes and peaks are 4133 and 7781, respectively. In *DMD* (CLIP-seq), the overlapped genes and peaks are 239 and 2114, respectively, which is less than that of iRIP-seq, because the peaks obtained by CLIP-seq are less than that of obtained by iRIP-seq. The overlapping genes and peaks indicate that the two replicate experiments are related, which also further demonstrates that *CLIP1* and *DMD* have RNA binding activity.

Author contributions

Conceptualization, JX and SYL; Investigation JX, XLZ, JFZ, XH, XXT, YZ and SYL; RBPPred software development, XLZ and SYL; CLIP-seq and iRIP-seq experiment, YZ, YQX and XLW; Data Curation, JX and JFZ; Writing - Original Draft, JX; Writing - Review & Editing, JX, XLZ, JFZ, XH, XXT, XDL and SYL; Funding Acquisition, SYL; Supervision, SYL.

Accession codes

All data sets (including *CLIP1* and *DMD* iRIP-seq) have been deposited at the Gene Expression Omnibus (GSE128318).

Our source code phdRBP (Pipline for High-throughput Data analysis for RNA-Binding Protein) and the data for analysis the mutations and diseases (Clinvar_cBioPortal) are freely available at http://rnabinding.com/phdRBP/.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities [2016YXMS017] and the Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase) under Grant No. U1501501.

We thank the National Supercomputer Center in Guangzhou for support of computing resources. We are grateful to ABlife. Inc. (Wuhan, China) for conducting iRIP-seq and CLIP-seq experiments and data analysis guidance.

We thank Xiujuan Ou for helping to improve the picture quality.

Appendix A. Computational function analysis of binding peak-related genes

Computational GO functional enrichment analysis of *CLIP1/DMD*-binding peak associated genes indicates that these genes may bind to protein and participate in cellular process (Fig. 4A, Fig. 4B, Fig. 4C and Fig. S6A). The identified *CLIP1/DMD*-binding peak-related genes are further assigned to the biochemical pathways in the KEGG database, revealing that *CLIP1/DMD*-binding targets share some biochemical pathways, such as ribosome biogenesis in eukaryotes (Fig. 5A, Fig. 5B, Fig. 5C and Fig. S6B). It may imply that *CLIP1* and *DMD* are involved in these biological pathways and may play regulatory roles in these events.

2.4. DMD/CLIP1 are involved in many diseases

The POSTAR2 database [58] collected a large number of SNPs on RBP-binding sites. We ask whether there are SNPs in *CLIP1/DMD*-binding sites or not. If so, will these SNPs cause diseases? To answer these two questions, we analyzed whether the mutations occurred at the *CLIP1/DMD* and genes corresponding to *CLIP1/DMD*-binding RNA or not. *CLIP1/DMD*-binding sites were mapped to ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) [59], which outputted 9267 mutations (Supplementary Table S4). RNABindRPlus [60] was used to predict the RNA-binding site on *DMD/CLIP1*. After prediction, the protein sequence of *DMD/CLIP1* and their binding sites were mapped to the GRCH38 reference genome (see Methods). Finally, we analyzed whether there are mutations in the binding sites of *DMD/CLIP1* or not. The results show that there is no mutation occurred at the RNA-binding site of *CLIP1*, but there are 37 mutations in the RNA-binding site of *DMD* (Supplementary Table S4). And according to ClinVar, these mutations associate with four diseases, Becker muscular dystrophy, Duchenne muscular dystrophy, Dilated cardiomyopathy 3B and Cardiovascular phenotype. Unexpectedly, mutations in some of the RNAs that interact with *DMD* also associate with these four diseases (Supplementary Table S4). So, the occurrence of these four diseases may be related to the mutations in *DMD* and its interaction partners.

Since it is not clear whether these mutations occurred at the binding sites between *DMD/CLIP1* and its partner RNAs or not, therefore, we cannot make sure that there is a direct link between the disease and the mutations of *DMD/CLIP1*-RNA. So, we wondered if *DMD/CLIP1*-RNA interactions are

present in the same disease. Therefore, we queried *CLIP1* and *DMD* in the cBioPortal cancer database (https://www.cbioportal.org/datasets) [61], which contains TCGA and the Cancer Cell Line Encyclopedia (CCLE) data. We analyzed thirteen of twenty-four cancer data with multiple clinical TCGA cancer samples. These thirteen cancers are Adrenal Glandp, Bladder, Urinary Tract, Bowel, Breast, Cervix Esophagus, Stomach, Head and Neck, Liver, Lung Ovary, Fallopian Tube, Pancrease and Skin. We chose these thirteen cancers because *CLIP1* mutated and carcinogenic in all these thirteen cancers. Unexpectedly, *CLIP1* and another 300 cancer genes are always co-occurrence in these thirteen cancers and all of them accompany with mutations. Combining with the results of peak calling, among 300 genes, 123 gene-coded RNAs interact with *CLIP1* (iRIP-seq, IP1, Supplementary Table S5). Therefore, if these 124 genes are pathogenic in these thirteen cancers, *CLIP1* may have a certain regulatory effect on these pathogenic genes (p-value = 8.316e-05, Supplementary Table S5). These findings may clarify the pathogenic mechanism of these related diseases, which may further help in the diagnosis and treatment of these diseases.

To visualize the mutations in the *CLIP1/DMD*-binding RNA, we used Integrative Genomics Viewer (IGV) [62] for visualization. In Fig. 6A, it shows the distribution of *CLIP1*-binding *MIR4435-2HG* gene near the position of 111, 480, 547 on chromosome 2. It is clear that the base A is mutated to the base G at this position. According to dbSNP database (https://www.ncbi.nlm.nih.gov/snp), the mutation ID is rs55784630. This A/G single-nucleotide variation is located in the intronic region of the long non-coding RNA (lncRNA) of *MIR4435-2HG*. In order to verify the direct binding of CLIP1 to MIR4435-2HG, we use the RIP-qPCR experiment. According to the library information, IgG/Input was used as the control. The results show that the MIR4435-2HG gene is consistent with expectations, and it has an enrichment combination with CLIP1, and Two-way ANOVA has significant differences (Fig. 7).

In Fig. 6B, it shows partial results of *DMD*-binding (iRIP-seq) *ZFHX3* gene near the location of 72,957,947 on chromosome 16. We can see that the base G is mutated to base C. Querying the SNP-related databases, we haven't obtained information about this mutation, perhaps it is a new mutation. It is unknown that whether the mutations of CLIP1-binding MIR4435-2HG or DMD-binding ZFHX3 at this position affect certain diseases or not and it requires further research.

Appendix 3. Discussion

More and more studies have shown that RBPs play a vital role in many diseases [4,63], making it urgently to identify RBPs quickly and accurately. Though lots of RBPs are recently identified by experiment, many potential RBPs are to be discovered. 17% of the RBPs predicted by RBPPred were predicted by SPOT-seq at the same time, suggesting that many proteins probably bind RNA but are not experimentally validated or annotated as RBPs yet [21].

For further verify the predictive ability of our proposed method, three potential RBPs are selected by the means introduced in Fig. S7 and are used to be validated with iRIP-seq and CLIP-seq experiment. We performed iRIP-seq and CLIP-seq experiments on two proteins, *CLIP1* and *DMD*, However, *CLIP1* was unsuccessfully overexpressed in the CLIP-seq experiment. The third protein, SMARCA4, due to the coding region is too long to construct plasmid. It is worthy of note that high overlap of peaks and genes of the repeated experiments is in favor of that both proteins have RNA-binding activity. In addition, our new method Deep-RBPPred [22], the deep learning-based version of RBPPred also supports that *CLIP1* and *DMD* are RBPs.

The results of iRIP-seq and CLIP-seq repeat experiments for these two selected potential RBPs with high overlap is in favor of that both proteins are novel RBPs in HeLa cell. It is sufficient to certify that our program, RBPPred, is a powerfully predictor to identify candidate RBPs. The experimental results show that both the target proteins *CLIP1* and *DMD* tend to bind to intronic regions. Computational GO and KEGG analysis show that the RNAs interacts with these two proteins perform some biological functions, such as protein binding and nucleus terms, and may participate in the same biological pathways, such as ribosome biogenesis in eukaryotes.

There are mutations in genes, which transcribe RNA to interact with *DMD/CLIP1* and can cause diseases. We found that 300 mutant genes are always present in thirteen cancers simultaneously with *CLIP1* (iRIP-seq), and 123 of them interact with *CLIP1* (iRIP-seq). Perhaps *CLIP1* have a certain regulatory effect on these genes in these thirteen cancers.

Appendix 4. Materials and methods

4.1. Three RBP candidates to experiment

Top 500 proteins in the list download from the results of RBPPred [21] were employed as targets. After removing 18 proteins (From April 13, 2016 to August 20, 2016) which have been verified to be RBPs [7,64], and then there remains 482 proteins. In Supplementary Table S6, we have updated the RBPs that were experimentally verified from August 20, 2016 to September 25, 2020. And 170 of the top 500 proteins have been experimentally verified as RBPs.

To find potential RBPs for experimentation, we followed the steps in Fig. S7. We calculated the similarity between the target protein and template protein at first. Because the more similar the protein structures are, the more similar their functions are, so we aligned the structure of the target protein with the template protein structure in the protein-RNA complex and then calculated their similarity. 439 RNA-protein complex structures [31] were used as templates and 482 proteins as targets, and then TM-score [65] was applied to assess whether the target can find the template or not. If the target protein has a three-dimensional structure in Protein Data Bank [66], the X-ray structure with the best resolution (at least better than 3 Å) was used. If the target protein in UniProt [37] doesn't have a crystal structure, we adopt the structure from ModeBase [67], whose sequence identity is greater than or equal to 30% with the target protein. The target protein is not considered, if there is no structure with a sequence similarity greater than 30% with the target protein. We used 0.45 as the threshold of TM-score same as PRIME [31]. If TM-score is greater than or equal to 0.45, these two structures are thought to be similar and this target protein is more likely to be an RBP. Then, inspired by Murakawa et al. [68], we picked out the proteins that contain RNA Binding Domain (RBD). The proteins' domains are downloaded from NCBI (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi). If the domain in target protein is a classical RBD (listed in paper [69]) or nonclassical RBD (domain in proteins has been confirmed to be RNA-binding in at least one experiment) or RBD unknown (known RBPs are not annotated as RNA-binding proteins) [69], we consider that this protein is more likely to be an RBP. Additionally, to find RBPs that play an important role, the number of articles in UniProt [37] and PubMed about the function of gene were considered. Finally, 4 proteins meet all screening criterions (April 14, 2016 to October 18, 2016). More details are shown in Supplementary Table S6.

Brannan et al. [33] proposed that if a protein interacts with the more RBPs, then this protein is the more likely to be an RBP. Hence, we inquired the

RBPs which interact with target proteins and its homologous functions in UniProt [37] (Supplementary Table S7). Finally, top 3 proteins (*CLIP1*, *DMD* and *SMARCA*4) are picked out to experiment. Among these three proteins, *CLIP1* and *DMD* were annotated as non-RBP by SOPT-sEq. [27], and Huang et al. [15] thought that *CLIP1* is not an RBP. Another *SMARCA4* is failed in constructing plasmid. Fig. S8 shows the domain patterns of the final targets, which are drawn by IBS [70]. All domains contained in these two target proteins belong to RBD unknown.

4.2. Cell culture and transfection

Human cervical cancer cell line HeLa cells were obtained from CCTCC (China Center for Type Culture Collection, Wuhan, Hubei, China) in 2017. The cell line has been authenticated with STR analysis by Cell Bank, Type Culture Collection, Chinese Academy of Sciences (CBTCCCAS), and tested for the free of mycoplasma contamination. STR analysis was performed as previously described [71].

HeLa cells were cultured in 10 cm vessel with 5% CO2 at 37 °C in MEM medium containing 10% FBS (fetal bovine serum), 100 U/mL penicillin and 100 mg/mL streptomycin. Cells were cultured to a confluence of 50–60% and transfected by pIRES-hrGFP-1a vector containing the DMD gene performed using Lipofectamine 2000 (Invitrogen, Carlsbad, CA, USA, 11668019) according to the manufacturer's protocol.

4.3. iRIP-seq

Cells were cross-linked on ice with UV irradiation type C (254 nm) at 400 mJ per cm^2 in the presence of cold PBS (4 ml per 15-cm dish). Cells were scraped off and pelleted at 1000 g at 4 °C and stored at −80 °C until further use. Cells lysis was performed in cold wash buffer (1× PBS, 0.1% SDS, 0.5% NP-40 and 0.5% sodium deoxycholate) supplemented with a 200 U/ml RNase inhibitor (Takara) and protease inhibitor cocktail (Roche) and incubated on ice for 30 min. The clear cell was lysed by centrifugation at 10,000 rpm for 10 min at 4 °C followed by addition of RQ I (Promega, 1 U/μl) to a final concentration of 1 U/μl and incubation in a water bath for 3 min at 37 °C. Cool reaction was subsequently done for 5 min on ice before proceeding, and the reaction was stopped by adding EDTA.

For immunoprecipitation, the supernatant was incubated overnight at 4 °C with 10 μg Flag-antibody and control IgG-antibody. The immunoprecipitation was further incubated with protein A/G Dynabeads for 2 h at 4 °C. After applying to a magnet and removing the supernatants, the beads were sequentially washed with lysis buffer, high-salt buffer (250 mM Tris 7.4, 750 mM NaCl, 10 mM EDTA, 0.1% SDS, 0.5% NP-40 and 0.5 deoxycholate), and PNK buffer (50 mM Tris, 20 mM EGTA and 0.5% NP-40) for two times, respectively. The beads were resuspended in Elution buffer (50 nM Tris 8.0, 10 mM EDTA and 1% SDS), and the suspension was incubated for 20 min in a heat block at 70 °C to release the immunoprecipitated RBP with crosslinked RNA and vortex. The magnetic beads were removed on the separator. The supernatant was transferred to a clean 1.5 ml microfuge tube. Proteinase K (Roche) was added into the 1% input (without immunoprecipitated) and immunoprecipitated RBP with crosslinked RNA, with a final concentration of 1.2 mg/ml and they were incubated for 120 min at 55 °C. The RNA was purified with Trizol reagent (Life technologies).

The Illumina ScriptSeq™ v2 RNA-Seq Library Preparation Kit (Epicentre) was used as the cDNA libraries. The cDNAs were purified and amplified, PCR products corresponding to 200–500 bps were purified, quantified and stored at −80 °C until it was used for sequencing.

Quantitative polymerase chain reaction (qPCR) for CLIP1 iRIP was carried out according to the protocol of one published study [72].

For high-throughput sequencing, the libraries were prepared following the manufacturer's instructions and applied to Illumina Nextseq500 system for 150 nt paired-end sequencing by ABlife. Inc. (Wuhan, China).

4.4. CLIP-seq

Cells were washed with ice-colded PBS for 3 times and UV cross-linking was performed with UV irradiation type C (254 nm) at 400 ml per cm2. Crosslinked cells were scraped off the plate and were collected by centrifugation at 1000 xg for 5 min. Cells lysis was performed in cold lysis buffer (1× PBS, 0.1% SDS, 0.5% NP-40 and 0.5% sodium deoxycholate) supplemented with a 1% RNase inhibitor (Takara) and 2% protease inhibitor cocktail (Roche) for 30 min. Cell lysates were cleared by centrifugation at 10,000 rpm for 10 min at 4 °C and the supernatants were used for immunoprecipitation.

For DNA digestion, RQ1 (promega) was added to the lysate and was incubated at 37 °C for 3 min.

For immunoprecipitation, 600 μL lysate was incubated with 15 μg antibody or control IgG antibody overnight at 4 °C. The immunuprecipitates were further incubated with protein A/G Dynabeads for 2-3 h at 4 °C. After applying to magnet and removing the supernatants, the beads were sequentially washed with wash buffer (1× PBS, 1% SDS, 0.5% NP-40 and 5% sodium deoxycholate), high-salt wash buffer (5× PBS, 1% SDS, 0.5% NP-40 and 5% sodium deoxycholate), and PNK buffer (50 mM Tris pH = 7.4, 10 mM MgCl2 and 0.5% NP-40) for two times, respectively.

The on-bead digestion was performed by adding MNase (Thermo), followed by incubation at 37 °C for 10 min. After washing with PNK buffer as described above, dephosphorylation and phosphorylation were performed with calf intestinal alkaline phosphatase (CIP, NEB) and polynucleotide kinase (PNK, NEB), respectively.

The immunoprecipitated protein-RNA complex was eluted from the beads by heat denaturing and was resolved on a Novex 4–12% Bis-Tris precast polyacrylamide gel (Invitrogen). The protein-RNA complexes were cut from the gel and RNA was extracted with Trizol after digesting the proteins.

The recovered RNA was used to generate paired-end sequencing library with Balancer NGS Library Preparation Kit for small/microRNA (Gnomegen) following the manufacture instructions. Libraries corresponding to 150–250 bps were purified, quantified and stored at −80 °C until used for sequencing. The libraries were applied to Illumina Novaseq system for 151 nt paired-end sequencing by ABlife. Inc. (Wuhan, China).

Appendix B. PhdRBP for analysis the data of iRIP-seq/CLIP-seq

First, the reads are pre-processed. We first remove adapter sequences with Cutadapt (https://doi.org/10.14806/ej.17.1.209) and set the parameter as "cutadapt -a AGATCGGAAGAGC -u 3 -m 16". Next, we remove bases with quality below 20, and retain reads longer than 16 bases with FASTX-Toolkit package (http://hannonlab.cshl.edu/fastx_toolkit/) by setting parameter "fastq_quality_trimmer -t 20 -l 16 -Q 33". Then, we remove the read if it has more than 30% base with quality below 20 ("fastq_quality_filter -q 20 -p 70 -Q 33"). After that, we discard NN, polyG and polyA in the reads, and the reads are retained which is longer than 16 bases (cutadapt -a N -m 16 —O 1 —N - | cutadapt -a GGGGGGGGGGGGG —O 4 -e 0.1 -n 5 - | cutadapt -a AAAAAAAAAAAAAA -m 16 –label clean —O 4 -e 0.1). Last but not least, end2 in the double-ended sequencing data is performed reverse

complementation by fastx_reverse_complement of FASTX-Toolkit package and then cat to end1 to form single-ended data.

Second, the cleaned reads are mapped to genome. After pre-processing, retained reads are aligned to GRCh38 using TopHat2 [46] and NovoAlign (http://www.novocraft.com/). We reserve the read which is mapped to genome by setting the TopHat2 parameter as "-p 16 -o –G –read-edit-dist 4 -N 4 –b2-N 1 -a 8 -m 0 -g 2 -p 12 –microexon-search –no-coverage-search –report-secondary-alignments." After mapping, the duplicate reads are removed by SAMtools [73]. In addition to TopHat2, we utilize NovoAlign to map reads to genome for CIMS [74]. We use the fastq2fasta in the HOMER [53] to convert the fastq file to the fasta file, and then the duplicates are removed from the obtained fasta file by fasta2collapse in CIMS. Thereafter, we set the NonoAlign parameter as "-F FA -d -f -t 110 -o Native -r None."

Third, the peaks are detected. In order to identify the relatively reliable binding sites, we use Piranha and CIMS to analyze the mapping result. For Piranha, we set the parameters: -b 20 -p 0.001 -s -d ZeroTruncatedNegativeBinomial. For CIMS, parameters are set same as Moore et al. [75], but -p is set to 0.01 to prevent more peaks being filtered out.

Fourth, the binding sites of RBPs are annotate. According to the GFF3 annotation file, which is download from GENCODE (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_23/), we extract the CDS, 5'UTR, 3'UTR, intron, antisense, Nc_exon regions. Nc_exon is defined as: exon doesn't be annotated as protein coding in the GFF3 file. Antisense is defined as: gene_type is antisense in GFF3 file. After getting the peaks, we calculate the distribution of binding peaks in various regions of the genome.

Fifth, the motifs are analyzed. For motif discovery, we run the findMotifs.pl from HOMER software on the peak sequence with the following parameters: -rna -len 5,6,7,8,9,10,11,12 -p 24. Before running findMotif.pl, we need to prepare the fasta of peak (peak.fa) and the fasta of background peak (bg_peak.fa). For the peak.fa, we remove the peaks from the IP samples who share at least one base with the input sample, and then get the peaks of the IP samples and finally extract the peak.fa using the "getfasta" in BEDTools [57]. For the bg_peak.fa, the fasta of gene (gene.fa) is used to generate bg_peak.fa. Gene.fa is the fasta sequence of each gene in GFF3. The bg_peak sequences are randomly selected from gene.fa. If the length of the randomly selected gene.fa is no longer than the length of the peak, then the gene.fa is taken as the bg_peak.fa; if it is longer than the length of the peak, it is randomly intercepted according to the length of the peak sequence, and then output as bg_peak.fa.

4.5. *Mapping the protein sequence and its binding site to the GRCH38 reference genome*

First, linking the sequence of the target protein to the consensus coding sequence (CCDS) of UniProt so that we can get the genomic location of the DNA which encodes the target protein, which only contains the conservative coding region. Then, we obtained the binging sites of the target protein by RNABindRPlus [60]. After these two steps, we obtained the position of the binding sites of the target protein on the genome. Next, we can analyze whether a mutation occurs at the binding site or not according to the ClinVar.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2021.12.003.

References

[1] G. Storz, J. Vogel, K.M. Wassarman, Regulation by small RNAs in bacteria: expanding frontiers, Mol. Cell 43 (2011) 880–891.

[2] A.G. Baltz, M. Munschauer, B. Schwanhausser, A. Vasile, Y. Murakawa, M. Schueler, N. Youngs, D. Penfold-Brown, K. Drew, M. Milek, E. Wyler, R. Bonneau, M. Selbach, C. Dieterich, M. Landthaler, The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts, Mol. Cell 46 (2012) 674–690.

[3] H. Xia, D. Chen, Q. Wu, G. Wu, Y. Zhou, Y. Zhang, L. Zhang, CELF1 preferentially binds to exon-intron boundary and regulates alternative splicing in HeLa cells, Biochim. Biophys. Acta 1860 (2017) 911–921.

[4] E. Sebestyen, B. Singh, B. Minana, A. Pages, F. Mateo, M.A. Pujana, J. Valcarcel, E. Eyras, Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks, Genome Res. 26 (2016) 732–744.

[5] Y. Zhang, L. Gu, Y. Hou, L. Wang, X. Deng, R. Hang, D. Chen, X. Zhang, Y. Zhang, C. Liu, X. Cao, Integrative genome-wide analysis reveals HLP1, a novel RNA-binding protein, regulates plant flowering by targeting alternative polyadenylation, Cell Res. 25 (2015) 864–876.

[6] B.J. Zarnegar, R.A. Flynn, Y. Shen, B.T. Do, H.Y. Chang, P.A. Khavari, irCLIP platform for efficient characterization of protein-RNA interactions, Nat. Methods 13 (2016) 489–492.

[7] E.L. Van Nostrand, G.A. Pratt, A.A. Shishkin, C. Gelboin-Burkhart, M.Y. Fang, B. Sundararaman, S.M. Blue, T.B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, G.W. Yeo, Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP), Nat. Methods 13 (2016) 508–514.

[8] Y. Sugimoto, A. Vigilante, E. Darbo, A. Zirra, C. Militti, A. D'Ambrogio, N. M. Luscombe, J. Ule, hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1, Nature 519 (2015) 491–494.

[9] J. Konig, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D.J. Turner, N. M. Luscombe, J. Ule, iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution, Nat. Struct. Mol. Biol. 17 (2010) 909–915.

[10] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M.J. Ascano, A.C. Jungkamp, M. Munschauer, A. Ulrich, G. S. Wardle, S. Dewell, M. Zavolan, T. Tuschl, Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP, Cell 141 (2010) 129–141.

[11] D.D. Licatalosi, A. Mele, J.J. Fak, J. Ule, M. Kayikci, S.W. Chi, T.A. Clark, A. C. Schweitzer, J.E. Blume, X. Wang, J.C. Darnell, R.B. Darnell, HITS-CLIP yields genome-wide insights into brain alternative RNA processing, Nature 456 (2008) 464–469.

[12] J.D. Keene, J.M. Komisarow, M.B. Friedersdorf, RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts, Nat. Protoc. 1 (2006) 302–307.

[13] G.H. David, D.R. Kelley, D. Tenen, B. Bernstein, J.L. Rinn, Widespread RNA binding by chromatin-associated proteins, Genome Biol. 17 (2016) 28.

[14] C. He, S. Sidoli, R. Warneford-Thomson, D.C. Tatomer, J.E. Wilusz, B.A. Garcia, R. Bonasio, High-resolution mapping of RNA-binding regions in the nuclear proteome of embryonic stem cells, Mol. Cell 64 (2016) 416–430.

[15] R. Huang, M. Han, L. Meng, X. Chen, Transcriptome-wide discovery of coding and noncoding RNA-binding proteins, Proc. Natl. Acad. Sci. U. S. A. 115 (2018) E3879–E3887.

[16] R.M.L. Queiroz, T. Smith, E. Villanueva, M. Marti-Solano, M. Monti, M. Pizzinga, D. M. Mirea, M. Ramakrishna, R.F. Harvey, V. Dezi, G.H. Thomas, A.E. Willis, K. S. Lilley, Comprehensive identification of RNA-protein interactions in any organism using orthogonal organic phase separation (OOPS), Nat. Biotechnol. 37 (2019) 169–178.

[17] E.C. Urdaneta, C.H. Vieira-Vieira, T. Hick, H.H. Wessels, D. Figini, R. Moschall, J. Medenbach, U. Ohler, S. Granneman, M. Selbach, B.M. Beckmann, Purification of cross-linked RNA-protein complexes by phenol-toluol extraction, Nat. Commun. 10 (2019) 990.

[18] J. Trendel, T. Schwarzl, R. Horos, A. Prakash, A. Bateman, M.W. Hentze, J. Krijgsveld, The human RNA-binding proteome and its dynamics during translational arrest, Cell 176 (2019) 391–403 e319.

[19] D. Maticzka, I.A. Ilik, T. Aktas, R. Backofen, A. Akhtar, uvCLAP is a fast and non-radioactive method to identify in vivo targets of RNA-binding proteins, Nat. Commun. 9 (2018) 1142.

[20] M. Kumar, M.M. Gromiha, G.P. Raghava, SVM based prediction of RNA-binding proteins using binding residues and evolutionary information, J. Mol. Recognit. 24 (2011) 303–313.

[21] X. Zhang, S. Liu, RBPPred: predicting RNA-binding proteins from sequence using SVM, Bioinformatics 33 (2017) 854–862.

[22] J. Zheng, X. Zhang, X. Zhao, X. Tong, X. Hong, J. Xie, S. Liu, Deep-RBPPred: predicting RNA binding proteins in the proteome scale based on deep learning, Sci. Rep. 8 (2018) 15264.

[23] A. Bressin, R. Schulte-Sasse, D. Figini, E.C. Urdaneta, B.M. Beckmann, A. Marsico, TriPepSVM: de novo prediction of RNA-binding proteins based on short amino acid motifs, Nucleic Acids Res. 47 (9) (2019) 4406–4417.

[24] I. Paz, E. Kligun, B. Bengad, Y. Mandel-Gutfreund, BindUP: a web server for non-homology-based prediction of DNA and RNA binding proteins, Nucleic Acids Res. 44 (2016) W568–W574.

[25] F. Agostini, A. Zanzoni, P. Klus, D. Marchese, D. Cirillo, G.G. Tartaglia, catRAPID omics: a web server for large-scale prediction of protein-RNA interactions, Bioinformatics 29 (2013) 2928–2930.

[26] M. Sharan, K.U. Forstner, A. Eulalio, J. Vogel, APRICOT: an integrated computational pipeline for the sequence-based identification and characterization of RNA-binding proteins, Nucleic Acids Res. 45 (2017), e96.
[27] H. Zhao, Y. Yang, Y. Zhou, Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction, RNA Biol. 8 (2011) 988–996.
[28] Y. Yang, H. Zhao, J. Wang, Y. Zhou, SPOT-Seq-RNA: predicting protein-RNA complex structure and RNA-binding function by fold recognition and binding affinity prediction, Methods Mol. Biol. 1137 (2014) 119–130.
[29] H. Zhao, Y. Yang, Y. Zhou, Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets, Nucleic Acids Res. 39 (2011) 3017–3025.
[30] Y. Yang, J. Zhan, H. Zhao, Y. Zhou, A new size-independent score for pairwise protein structure alignment and its application to structure classification and nucleic-acid binding prediction, Proteins 80 (2012) 2080–2088.
[31] J. Zheng, P.J. Kundrotas, I.A. Vakser, S. Liu, Template-based modeling of protein-RNA interactions, PLoS Comput. Biol. 12 (2016), e1005120.
[32] J. Zheng, J. Xie, X. Hong, S. Liu, RMalign: an RNA structural alignment tool based on a size independent scoring function, BMC Genomics 20 (2019) 276.
[33] K.W. Brannan, W. Jin, S.C. Huelga, C.A. Banks, J.M. Gilmore, L. Florens, M. P. Washburn, E.L. Van Nostrand, G.A. Pratt, M.K. Schwinn, D.L. Daniels, G.W. Yeo, SONAR discovers RNA-binding proteins from analysis of large-scale protein-protein Interactomes, Mol. Cell 64 (2016) 282–293.
[34] W. Li, Z. Zhang, X. Liu, X. Cheng, Y. Zhang, X. Han, Y. Zhang, S. Liu, J. Yang, B. Xu, L. He, L. Sun, J. Liang, Y. Shang, The FOXN3-NEAT1-SIN3A repressor complex promotes progression of hormonally responsive breast cancer, J. Clin. Invest. 127 (2017) 3421–3440.
[35] W.H. Nie, S. Wang, R. He, Q. Xu, P.H. Wang, Y. Wu, F. Tian, J.H. Yuan, B. Zhu, G. Y. Chen, A-to-I RNA editing in bacteria increases pathogenicity and tolerance to oxidative stress, PLoS Pathog. 16 (2020).
[36] S. Peled, O. Leiderman, R. Charar, G. Efroni, Y. Shav-Tal, Y. Ofran, De-novo protein function prediction using DNA binding and RNA binding proteins as a test case, Nat. Commun. 7 (2016) 13424.
[37] T. UniProt Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 46 (2018) 2699.
[38] J.E. Rickard, T.E. Kreis, Identification of a novel nucleotide-sensitive microtubule-binding protein in HeLa cells, J. Cell Biol. 110 (1990) 1623–1633.
[39] P. Pierre, R. Pepperkok, T.E. Kreis, Molecular characterization of two functional domains of CLIP-170 in vivo, J. Cell Sci. 107 (Pt 7) (1994) 1909–1920.
[40] L. Griparic, T.C. Keller, Identification and expression of two novel CLIP-170/Restin isoforms expressed predominantly in muscle, Biochim. Biophys. Acta 1405 (1998) 35–46.
[41] H. Cho, G.Q. Shen, X.F. Wang, F. Wang, S. Archacki, Y.B. Li, G. Yu, S. Chakrabarti, Q.Y. Chen, Q.K. Wang, Long noncoding RNA ANRIL regulates endothelial cell activities associated with coronary artery disease by up-regulating CLIP1, EZR, and LYVE1 genes (vol 294, pg 3881, 2019), J. Biol. Chem. 294 (2019) 8715.
[42] E.P. Hoffman, R.H. Brown Jr., L.M. Kunkel, Dystrophin: the protein product of the Duchenne muscular dystrophy locus, Cell 51 (1987) 919–928.
[43] D.J. Blake, A. Weir, S.E. Newey, K.E. Davies, Function and genetics of dystrophin and dystrophin-related proteins in muscle, Physiol. Rev. 82 (2002) 291–329.
[44] R. Garcia-Rodriguez, M. Hiller, L. Jimenez-Gracia, Z. van der Pal, J. Balog, K. Adamzek, A. Aartsma-Rus, P. Spitali, Premature termination codons in the DMD gene cause reduced local mRNA synthesis, Proc. Natl. Acad. Sci. U.S.A. 117 (2020) 16456–16464.
[45] Z.F. Tang, B.X. Kang, C.W. Li, T.X. Chen, Z.M. Zhang, GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis, Nucleic Acids Res. 47 (2019) W556–W560.
[46] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S.L. Salzberg, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions, Genome Biol. 14 (2013) R36.
[47] S. Anders, P.T. Pyl, W. Huber, HTSeq–a Python framework to work with high-throughput sequencing data, Bioinformatics 31 (2015) 166–169.
[48] P.J. Uren, E. Bahrami-Samani, S.C. Burns, M. Qiao, F.V. Karginov, E. Hodges, G. J. Hannon, J.R. Sanford, L.O. Penalva, A.D. Smith, Site identification in high-throughput RNA-protein interaction data, Bioinformatics 28 (2012) 3013–3020.
[49] C. Zhang, R.B. Darnell, Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data, Nat. Biotechnol. 29 (2011) 607–614.
[50] H. Zhou, Y. Yang, H.B. Shen, Hum-mPLoc 3.0: prediction enhancement of human protein subcellular localization through modeling the hidden correlations of gene ontology and functional domain features, Bioinformatics 33 (2017) 843–853.
[51] H. Tilgner, D.G. Knowles, R. Johnson, C.A. Davis, S. Chakraborty, S. Djebali, J. Curado, M. Snyder, T.R. Gingeras, R. Guigo, Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs, Genome Res. 22 (2012) 1616–1625.
[52] N. Viphakone, I. Sudbery, L. Griffith, C.G. Heath, D. Sims, S.A. Wilson, Co-transcriptional loading of RNA export factors shapes the human transcriptome, Mol. Cell 75 (2019) 310–323 e318.
[53] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y.C. Lin, P. Laslo, J.X. Cheng, C. Murre, H. Singh, C.K. Glass, Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities, Mol. Cell 38 (2010) 576–589.
[54] G. Inana, T. Shinohara, J.J. Maizel, J. Piatigorsky, Evolution and diversity of the crystallins. Nucleotide sequence of a beta-crystallin mRNA from the mouse lens, J. Biol. Chem. 257 (1982) 9064–9071.
[55] S. Mai, X. Qu, P. Li, Q. Ma, C. Cao, X. Liu, Global regulation of alternative RNA splicing by the SR-rich protein RBM39, Biochim. Biophys. Acta 2016 (1859) 1014–1024.
[56] T.W. Chen, H.P. Li, C.C. Lee, R.C. Gan, P.J. Huang, T.H. Wu, C.Y. Lee, Y.F. Chang, P. Tang, ChIPseek, a web-based analysis tool for ChIP data, BMC Genomics 15 (2014) 539.
[57] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics 26 (2010) 841–842.
[58] Y. Zhu, G. Xu, Y.T. Yang, Z. Xu, X. Chen, B. Shi, D. Xie, Z.J. Lu, P. Wang, POSTAR2: deciphering the post-transcriptional regulatory logics, Nucleic Acids Res. 47 (2019) D203–D211.
[59] M.J. Landrum, J.M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, W. Jang, K. Katz, M. Ovetsky, G. Riley, A. Sethi, R. Tully, R. Villamarin-Salomon, W. Rubinstein, D.R. Maglott, ClinVar: public archive of interpretations of clinically relevant variants, Nucleic Acids Res. 44 (2016) D862–D868.
[60] R.R. Walia, L.C. Xue, K. Wilkins, Y. El-Manzalawy, D. Dobbs, V. Honavar, RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins, PLoS One 9 (2014), e97725.
[61] J. Gao, B.A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S.O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, N. Schultz, Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal, Sci. Signal. 6 (2013) pl1.
[62] H. Thorvaldsdottir, J.T. Robinson, J.P. Mesirov, Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration, Brief. Bioinform. 14 (2013) 178–192.
[63] K.E. Lukong, K.W. Chang, E.W. Khandjian, S. Richard, RNA-binding proteins in human genetic disease, Trends Genet. 24 (2008) 416–425.
[64] A. Castello, B. Fischer, C.K. Frese, R. Horos, A.M. Alleaume, S. Foehr, T. Curk, J. Krijgsveld, M.W. Hentze, Comprehensive identification of RNA-binding domains in human cells, Mol. Cell 63 (2016) 696–710.
[65] Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score, Nucleic Acids Res. 33 (2005) 2302–2309.
[66] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I. N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.
[67] U. Pieper, B.M. Webb, G.Q. Dong, D. Schneidman-Duhovny, H. Fan, S.J. Kim, N. Khuri, Y.G. Spill, P. Weinkam, M. Hammel, J.A. Tainer, M. Nilges, A. Sali, ModBase, a database of annotated comparative protein structure models and associated resources, Nucleic Acids Res. 42 (2014) D336–D346.
[68] Y. Murakawa, M. Hinz, J. Mothes, A. Schuetz, M. Uhl, E. Wyler, T. Yasuda, G. Mastrobuoni, C.C. Friedel, L. Dolken, S. Kempa, M. Schmidt-Supprian, N. Bluthgen, R. Backofen, U. Heinemann, J. Wolf, C. Scheidereit, M. Landthaler, RC3H1 post-transcriptionally regulates A20 mRNA and modulates the activity of the IKK/NF-kappaB pathway, Nat. Commun. 6 (2015) 7367.
[69] S.C. Kwon, H. Yi, K. Eichelbaum, S. Fohr, B. Fischer, K.T. You, A. Castello, J. Krijgsveld, M.W. Hentze, V.N. Kim, The RNA-binding protein repertoire of embryonic stem cells, Nat. Struct. Mol. Biol. 20 (2013) 1122–1130.
[70] W. Liu, Y. Xie, J. Ma, X. Luo, P. Nie, Z. Zuo, U. Lahrmann, Q. Zhao, Y. Zheng, Y. Zhao, Y. Xue, J. Ren, IBS: an illustrator for the presentation and visualization of biological sequences, Bioinformatics 31 (2015) 3359–3361.
[71] Q.L. Song, F.T. Yi, Y.H. Zhang, D.K.J. Li, Y.X. Wei, H. Yu, Y. Zhang, CRKL regulates alternative splicing of cancer-related genes in cervical cancer samples and HeLa cell, BMC Cancer 19 (2019).
[72] Y.F. Tu, X.F. Wu, F.Y. Yu, J.Z. Dang, Y.X. Wei, H. Yu, W.L. Liao, Y. Zhang, J. Wang, Tristetraprolin-RNA interaction map reveals a novel TTP-RelB regulatory network for innate immunity gene expression, Mol. Immunol. 121 (2020) 59–71.
[73] H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, Bioinformatics 27 (2011) 2987–2993.
[74] A. Shah, Y. Qian, S.M. Weyn-Vanhentenryck, C. Zhang, CLIP tool kit (CTK): a flexible and robust pipeline to analyze CLIP sequencing data, Bioinformatics 33 (2017) 566–567.
[75] M.J. Moore, C. Zhang, E.C. Gantman, A. Mele, J.C. Darnell, R.B. Darnell, Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis, Nat. Protoc. 9 (2014) 263–293.
[76] D. Bu, H. Luo, P. Huo, Z. Wang, S. Zhang, Z. He, Y. Wu, L. Zhao, J. Liu, J. Guo, S. Fang, W. Cao, L. Yi, Y. Zhao, L. Kong, KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis, Nucleic Acids Res. 49 (2021) W317–W325.