CELF1 preferentially binds to exon-intron boundary and regulates alternative splicing in HeLa cells

Heng Xia, Dong Chen, Qijia Wu, Gang Wu, Yanhong Zhou, Yi Zhang, Libin Zhang

PII: S1874-9399(17)30087-1
DOI: doi:10.1016/j.bbagrm.2017.07.004
Reference: BBAGRM 1166

To appear in: *BBA - Gene Regulatory Mechanisms*

Received date: 10 March 2017
Revised date: 30 June 2017
Accepted date: 17 July 2017

Please cite this article as: Heng Xia, Dong Chen, Qijia Wu, Gang Wu, Yanhong Zhou, Yi Zhang, Libin Zhang, CELF1 preferentially binds to exon-intron boundary and regulates alternative splicing in HeLa cells, *BBA - Gene Regulatory Mechanisms* (2017), doi:10.1016/j.bbagrm.2017.07.004

CELF1 preferentially binds to exon-intron boundary and regulates alternative splicing in HeLa cells

Heng Xia,[1] Dong Chen[2], Qijia Wu[3#], Gang Wu[1], Yanhong Zhou[1], Yi Zhang[2,3*] and Libin Zhang[1*]

[1]School of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China

[2]Center for Genome Analysis, ABLife Inc., Optics Valley International Biomedical Park, Building 18-1, East Lake High-Tech Development Zone, 388 Gaoxin 2nd Road, Wuhan, Hubei 430075, China.

[3]Laboratory for Genome Regulation and Human Health, ABLife Inc., Optics Valley International Biomedical Park, Building 18-2, East Lake High-Tech Development Zone, 388 Gaoxin 2nd Road, Wuhan, Hubei 430075, China.

* To whom correspondence should be addressed. Tel: +86-27-87792217; Fax: +86-27-97792170; Email: yizhang@ablife.cc; libinzhang@hust.edu.cn

#Present address: Seqhealth Technology Co., Ltd. Building C2, No.666 Gaoxin Road, Wuhan East Lake High-tech Development Zone, Wuhan 430074，China.

Abstract

The current RIP-seq approach has been developed for the identification of genome-wide interaction between RNA binding protein (RBP) and the bound RNA transcripts, but still rarely for identifying its binding sites. In this study, we performed RIP-seq experiments in HeLa cells using a monoclonal antibody against CELF1. Mapping of the RIP-seq reads showed a biased distribution at the 3'UTR and intronic regions. A total of 15,285 and 1,384 CELF1-specific sense and antisense peaks were identified using the ABLIRC software tool. Our bioinformatics analyses revealed that 5' and 3' splice site motifs and GU-rich motifs were highly enriched in the CELF1-bound peaks. Furthermore, transcriptome analyses revealed that alternative splicing was globally regulated by CELF1 in HeLa cells. For example, the inclusion of exon 16 of LMO7 gene, a marker gene of breast cancer, is positively regulated by CELF1. Taken together, we have shown that RIP-seq data can be used to decipher RBP binding sites and reveal an unexpected landscape of the genome-wide CELF1-RNA interactions in HeLa cells. In addition, we found that CELF1 globally regulates the alternative splicing by binding the exon-intron boundary in HeLa cells, which will deepen our understanding of the regulatory roles of CELF1 in the pre-mRNA splicing process.

Keywords: RNA binding proteins; RIP-seq; RNA sequencing; alternative splicing; splice site

1. Introduction

Regulation of genome expression at the level of transcription, pre-mRNA processing, as well as mRNA turnover and translation usually depends on specific interactions between RNA binding proteins (RBPs) and their RNA targets [1-7]. RBPs bind to RNA target sequences and participate in forming ribonucleoprotein complexes [8, 9]. Therefore, numerous softwares have not only been applied to probe DNA-protein interaction in ChIP-seq data [10-13], but also to identify the RNA-protein binding sites in CLIP-seq and RIP-seq data [14-19].

Accurate identification of the RNA targets of an RBP is very important for understanding its regulatory roles in gene expression. The combination of the RNA immunoprecipitation and microarray analysis was once a high-throughput approach for systematical identification of the transcripts to which an RBP binds [20]. However, this method relies on a large number of probes and high hybrid efficiency of microarray, without which there will be a loss of some crucial and novel RNA targets. Fortunately, this limitation has been overcome by the CLIP-seq technology, which sequences cDNA libraries constructed from RBP-bound RNA fragments eluted from either membrane or gel after *in vivo* crosslinking immunoprecipitation [21-23]. However, CLIP-seq methods are challenged by high experimental failure rates and extremely low library complexity [24]. RIP-seq technology, although similar to CLIP-seq, lacks several operational steps including enzymatic digestion of immuno-precipitated RNA and gel separation of protein-RNA complexes, and thus avoids these problems [25, 26]. Nevertheless, RIP-seq technology has lower accuracy in identification of the binding

sites. Jack Keene's lab recently reported a DO-RIP-seq method to include the enzymatic digestion procedure, which is applied to quantify HuR binding sites with high coverage on the whole human transcriptome [27]. CELF1 (previously called as CUGBP1) is a multifunctional RNA binding protein, which was reported to be associated with various mRNA metabolism processes, including pre-mRNA splicing [28-33], mRNA decay [34-37] and translation [38, 39]. CELF1 binds to U/G rich element in pre-mRNA and affects splice site selection. For example, aberrant expression of nuclear CELF1 in Myotonic Dystrophy I has been linked to disrupted alternative splicing patterns of several transcripts including cardiac troponin T [30], the insulin receptor [28] and the muscle-specific chloride channel [29]. CELF1 was also found to recognize "UGUUUGUUUGU" element (GU-rich element, GRE) or GU-repeat sequences in the 3'UTR [32, 40-42] to regulate mRNA stability.

Very importantly, recent studies showed that CELF1 protein is associated with several different types of cancer including esophageal cancer and breast cancer [43-45]. For example, CELF1 regulates GRE (GU-rich element)-containing epithelial-mesenchymal transition EMT driver mRNAs and correlates with increased metastasis in human breast cancer. CELF1 protein is significantly overexpressed in breast cancer tissues and drives *in vivo* metastatic colonization in the mouse xenograft model [45]. These studies prompt further investigations of the CELF1 targets and its regulatory roles in cancer cells.

Recently, a study of CELF1 RNA targets by CLIP-seq in human HeLa cells has been reported [18]. In that report, two million of unique reads and 2,972 CELF1-

binding clusters are reported. Limited analysis and biological relevance are performed. In this study, we for the first time utilized the RIP-seq approach to define the genome-wide CELF1-RNA interaction and alternative splicing events regulated by CELF1 in HeLa cells. The results showed that RIP-seq reads were strongly enriched at splice sites and GU-rich motifs. The ABLIRC algorithm identified 15,285 and 1,384 CELF1-specific sense and antisense peaks, which were distributed in over five thousand genes and enriched in the intronic and 3'UTR regions. Totally, 19,466 alternative splicing events were identified (*p*-value cutoff < 0.05), including 1,122 regulated alternative splicing events (RASEs) and 451 CELF1-bound and -affected alternative splicing events were identified. We also revealed the alternative splicing pattern of the LMO7 gene, a marker gene of breast cancer specifically expressed in metastatic breast cancer cells [46], which is positively regulated by CELF1. Taken together, this work clearly defined a complex genome-wide CELF1-RNA interaction map in human cancer cell line and indicated that CELF1 binds to exon-intron boundary and regulates various alternative splicing events, which will be helpful for understanding the regulation mechanism of CELF1 at the pre-mRNA splicing level and speculating the novel functions of CELF1 in multiple biological processes.

2. Materials and Methods

2.1 siRNA transfection and qRT-PCR

HeLa cells were cultured at 37℃, 5% CO_2 in DMEM containing 10% fetal bovine serum, penicillin (10 units / μl) and streptomycin (10 units / ml). For transient

knockdown of CELF1 expression, HeLa cells were incubated at 37° C in a CO_2 incubator until the cells were 60-80% confluent. CELF1 siRNA (5'-GAGCCAACCUGUUCAUCUA-3' [40], GenePharma) or control siRNA with final concentration of 50 nmol/L was transfected into HeLa cells using Lipofectamine™ 2000 (Life Technologies) according to the manufacturer's instructions. After 24 hours, the cells were collected in TRIzol (Invitrogen), and the total RNA was isolated according to the manufacturer's instructions. qRT-PCR was performed to confirm the knockdown of CELF1 mRNA. Meanwhile, RT-PCR assay was used to evaluate the effect of decreased CELF1 expression level on LMO7 pre-mRNA splicing in HeLa cells. qPCR primers for CELF1 and internal control GAPDH and primers for detecting the pre-mRNA splicing are shown in Supplemental Table 1.

2.2 Western Blot Analysis

CELF1 western blots assay was performed as described previously[34]. In brief, for the preparation of total cell lysates, the normal and CELF1-knockdown HeLa cells were lysed in RIPA buffer containing 50 mM Tris-HCl (pH 7.4), 150 mM NaCl, 1.0% deoxycholate, 1% Triton X-100, 1mM EDTA and 0.1% SDS. The samples were centrifuged (12,000rpm, 5min) and the supernatants were further analyzed on a 10% SDS-PAGE gel and subsequently transferred to a PVDF membrane (Millipore). CELF1 was detected using monoclonal antibody diluted in TBST (1:15,000, Millipore) and GAPDH was used as a loading control (1:30,000, Santa Cruz).

2.3 RIP-seq

RNA immunoprecipitation was performed as previously described [47]. Briefly, HeLa

cells were first lysed in ice-cold lysis buffer (10 mM HEPES, pH 7.0, 100 mM KCl, 5 mM $MgCl_2$, 0.5% NP-40, 10 mM DTT) with 200 U/ml RNase inhibitor (Promega) and a protease inhibitor (Roche) on ice for 5 min. The mixture was then vibrated vigorously and centrifuged at 13,000 x g at 4 °C for 20 min to remove cell debris. The supernatant was pre-cleared with 100 μl of DynaBeads protein G (Life Technologies) at 4 °C for 30 minutes. The pre-cleared supernatant was incubated with DynaBeads protein G conjugated with anti-CELF1 antibody (Millipore) or normal IgG at 4 °C for 6 hours. The beads were washed 6 times with lysis buffer and then divided into two groups, one for RNA isolation from CELF1-RNA complexes and another for the western blotting assay for CELF1 immunoprecipitation. The CELF1-bound RNAs were isolated from the immunoprecipitate of anti-CELF1 using TRIzol (Invitrogen), followed by the preparation of the Illumina Truseq pair-end libraries. In brief, the collected RNAs were fragmented at 95ºC, followed by end repair and 5' adaptor ligation. The reverse transcription was performed with RT primers harboring 3’ adaptor sequence and random hexamer. The generated cDNAs were PCR-amplified and the 200-500 bp products were purified. For high-throughput sequencing, the libraries were prepared according to the manufacturer's instructions and applied to Hi-seq 2000 system for 100 nt pair-end sequencing (ABlife. Inc, Wuhan, China).

2.4 RIP-qPCR

TRIzol (Invitrogen) was used to extract total RNAs from the immunoprecipitate of CELF1 according to the manufacturer’s instructions. Random primer was then used for the cDNA synthesis. In order to detect whether CELF1 target genes were significantly

and specifically enriched in the CELF1 immunoprecipitate, we used normal PCR as a confirmation. Meanwhile, we used input RNA as a reference and performed quantitative RT-PCR as described [25] to determine the relative level of specific RNAs in the IgG and CELF1 immunoprecipitates. Q-PCR data represents the mean values from at least three independent experiments. Various genes were selected for PCR-amplification both in CELF1 and IgG immunoprecipitates. Gene-specific PCR primer pairs are shown in Supplemental Table 1.

2.5 RNA sequencing

WT and siRNA-transfected HeLa cells were used for RNA-seq assay in this study. Moreover, the cultured WT HeLa cells were divided into two groups, one for RIP-seq assay and another for the RNA-seq assay. In short, the cDNA libraries of RNA-seq were prepared using TruSeqTM RNA Sample Preparation Kit (Illumina, Inc.) according to the manufacturer's instructions. Briefly, oligo (dT) magnetic beads (NEB) were used to purify poly-A mRNAs from 10 μg of total RNA. The extracted poly (A) mRNAs were further fragmented into 200-500 bp in size, followed by cDNA synthesis and ligation to adaptors (Illumina). The ligated cDNAs were amplified to generate the final cDNA libraries. The cDNA libraries were subsequently quantified and sequenced on the Illumina NextSeq 500 platform using the pair-ends protocol to generate 2x150 nt reads.

2.6 Data processing and mapping

For RNA-seq data, adaptors and low quality bases were trimmed from raw sequencing reads using FASTX-Toolkit (Version 0.0.13), and reads less than 16 nt were discarded.

Clean Reads were aligned to the human- GRCH38 genome using tophat2 [48] with 4 mismatches. For RIP-seq data, data processing method was the same to the RNA-seq data. After processing, we merged the two technical replicates and aligned the combined reads to the human-GRCH38 genome using tophat2 [48] with 2 mismatches.

2.7 RIP-seq Peak Calling analysis

After reads were aligned onto the genome, we discarded the reads with multiple genomic locations due to the ambiguous origination. Identical aligned reads were counted and merged as unique reads. The binding regions of CELF1 on genome were identified using "ABLIRC" strategy. Reads with at least 1 bp overlap region on the locus of genome were clustered as peaks. The unique reads number (N), the lengths of each reads and the observed max peak height were then counted for each peak. In the meanwhile, for each gene, the reads with the same number (N) and lengths were generated using computational simulation. The outputting reads were further mapped to the same genes to generate random max peak height from overlapping reads. The whole process was repeated for 500 times. All the observed peaks with heights higher than those of random max peaks (p-value < 0.05) were selected. The CELF1 and IgG samples were analyzed by the simulation independently. After simulation, the CELF1 peaks that have overlap with IgG peaks were removed. The target genes of CELF1 were finally determined by analyzing the locations of all the CELF1 binding peaks on the human genome and the binding motifs of CELF1 protein were called by Homer software [49].

2.8 RNA-seq data DEG and AS analysis

After mapping reads onto the genome, we discarded the reads that were with multiple genomic locations due to the ambiguous origination. Reads with only one genome location were preserved to calculate read number and RPKM value (RPKM represents reads per kilobase and per million) for each gene. DEGs between the si-CELF1 and Control sample were analyzed by using edgeR [50], one of R packages. For each gene, the *p*-value was computed and the significance threshold to control FDR at a given value was calculated.

To define and quantify the alternative splicing events (ASEs) in the RNA-seq samples and the regulated alternative splicing events (RASEs) between Control and si-CELF1 cells, we developed a AS method called ABLas (under submission). From the TopHat2 mapping result, we obtained the splicing junction (SJ) reads that have gaps while aligning to the genome. In this study, the known SJs were defined by joining splice sites at the exon-intron boundaries of all annotated introns in human GRCh38 genome. All other SJs involving only one or none of the known splice site were considered as novel junction. For each junction site, we defined the boundary reads as reads walking through the site and each side of intron boundary region with no less than 8 nt. The junctions located inside the coordinates of annotated genes were regarded as genic SJs. The genic SJs can be classified into one of the nine types of ASE. Seven of the canonical AS events were skipped exons (ES), cassette exon (CE), alternative 5'-splice sites (A5SS), alternative 3'-splice sites (A3SS), mutually exclusive exons (MXE), alternative first exons (AFE or 5'MXE) and alternative last exons (ALE or 3'MXE) according to the models described previously[51]. The ABLas algorithm is based on a

given gene model and calculates the ratio of reads supporting the alternative SJs to the total sequence reads supporting model SJs and alternative SJs. The known ASEs were composed of all known SJs, while novel ASEs involved a novel junction. For the known exon skipping (ES) splicing events, we required the total SJ reads should be no less than 10, among which SJ reads supporting the alternative and model SJs were all at least 2. To be a qualified candidate novel ES event, the novel SJs containing at least two support reads were selected for analysis, and the ratio of reads supporting the alternative SJs to the total was at least 15%. We next defined the exon skipping ratio. SJ reads supporting ES were treated as **a**, SJ reads supporting exon inclusion were treated as **b** and **c**, from 5' to 3' of mRNAs. The exon skipping ratio (ESR) was defined as: ESR = a/(a+(b+c)/2). The other ASEs were verified with the same method and threshold as ES. The splicing ratio was calculated with reads mapped on specific splice junctions.

Intron retention (IR) is caused by reduced usage of the candidate splice sites, which cannot be predicted effectively by considering splice junctions. This class of alternative splicing event was identified according to the border reads spanning exon-intron junction and the mean of local reads depth. We selected the intron retention (IR) splicing events with the following conditions: (1) having boundary reads at either the 5' or 3' splice site of the candidate; (2) the mean base depth in the candidate intron should be at least 20% that of the flanking exon and twice that of the intronic depth in the model gene. The ratio of mean base depth in the candidate intron to the flanking exon was considered as the IR intensity.

After detecting the ASEs in each RNA-seq sample, we identified regulated ASEs (RASEs) between si-CELF1 and Control samples. Fisher's exact test was chosen as the candidate statistical model to calculate the significant *p*-value. The input data for statistical model was the alternative reads and model reads of samples, respectively. We also obtained the changed ratio of alternatively spliced reads and constitutively spliced reads between si-CELF1 and Control samples, which is referred as the RASE ratio. *P*-value < 0.05 and RASE ratio > 0.1 were set as the threshold for RASEs detection.

2.9 Other statistical analysis

For the random peak calculation of CELF1 RIP-seq, we used rand function of Perl language to generate the random peaks. Random and real peaks had the same length and had arisen from the same gene locus, but the random peak location was randomly calculated. Statistical analysis was performed using the R software (https://www.r-project.org/) unless otherwise stated. Significance of differences was evaluated with either Student's t-test when only two groups were compared, or hypergeometric test for Venn diagram and functional term enrichment analysis, or Fisher exact test for genomic region enrichment analysis. Gene Ontology annotation and KEGG pathways analysis of CELF1 targets were further performed using DAVID on line platform (http://david.abcc.ncifcrf.gov/).

3. Results

3.1 CELF1 RIP-seq reads are preferentially enriched in 3'UTR and intronic

regions

We initially used the RIP-seq approach to identify the CELF1-interacting transcripts in HeLa cells. In order to assess the specificity of CELF1 pull-downs, we set IgG as the control of immunoprecipitation. As previous studies have showed that there is no obvious expression of CELF2 in HeLa cells [33], the monoclonal CELF1 antibody was only used for the immunoprecipitation of the CELF1-RNA complexes from the total HeLa cells lysate.

As shown in Fig. 1A, CELF1 was observed by western blotting in the total cell lysate and CELF1 IP fraction, while no CELF1 was detected in the IP fraction of the negative control IgG. The cDNA libraries of RNAs from anti-CELF1 and IgG immunoprecipitates were sequenced using the Hi-seq 2000 platform. After removing adaptor sequences and low-quality reads, a total of 27,030,650 and 12,300,623 reads were recovered from the CELF1 and IgG immunoprecipitates (Supplemental Table 2). When these reads were mapped onto the human GRCH38 genome using Tophat2 [48], only about 20% aligned. The majority of the unaligned reads was resulted from the presence of broken adaptor and primer sequences. Most of the anti-CELF1 reads were uniquely mapped, while many fewer IgG reads were either mapped or uniquely mapped (Supplemental Table 2). A total of 3,731,205 and 321,522 uniquely mapped reads from the anti-CELF1 and IgG immunoprecipitates were respectively recovered for further analysis in this study.

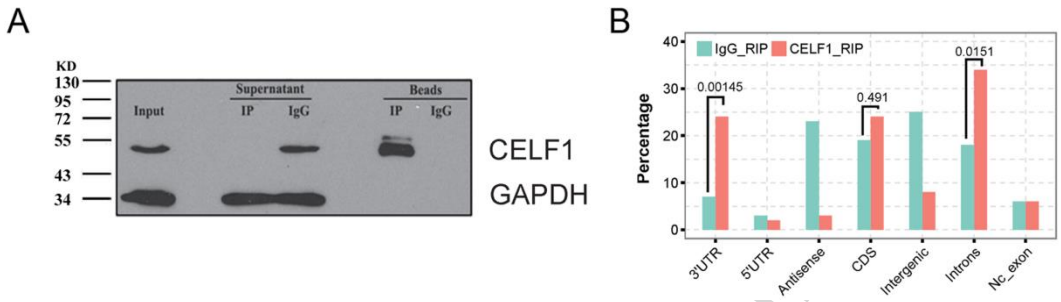

Figure 1. RIP-Seq assay and the profile of sequencing reads aligned to the human genome. (*A*) Immunoprecipitation of CELF1-associated RNAs using CELF1 monoclonal antibody. The efficient immunoprecipitation of CELF1 protein from HeLa extracts was validated by western blots. (*B*) Bar plot of the genomic region distribution of uniquely mapped anti-CELF1 reads. *P*-value was obtained by Fisher's exact test.

Ribonuclease was usually used to digest the immunoprecipitated RNAs in the RBP-RNA complexes in the CLIP-seq protocols to generate short-length reads from the sites protected by RBPs [22]. Theoretically, the RIP protocol lacks an RNase digestion step and recovers intact transcripts in the RBP-RNA complexes; these transcripts are randomized for cDNA library construction for deep sequencing. To our surprise, when we plotted the distribution of uniquely mapped anti-CELF1 reads on the whole human genome, we found that the RIP-seq reads were highly enriched in 3'UTR regions and intronic regions (Fig. 1B), which showed consistent result with CELF1 CLIP-seq data by previous studies [52, 53]. Importantly, we observed that the fractions of clean reads that mapped to 3'UTR and intronic regions were significantly higher in CELF1 immunoprecipitate than in IgG immunoprecipitate (Fig. 1B).

3.2 RIP-seq reads are peaked at splice sites and GU-rich motifs

In silico random CLIP strategy has been successfully used to recover the binding sites of RBPs from CLIP-seq data [22, 54]. Here, we developed a software tool called ABLIRC based on this strategy to recover the CELF1 binding sites from the RIP-seq reads (Materials and Methods). The ABLIRC algorithm identified 15,285 sense peaks and 1384 antisense peaks from CELF1-associated RNAs, and 1,088 sense and 681 antisense peaks from IgG-associated RNAs (control). After peaks shared by both CELF1 and IgG samples were filtered, a total of 15,542 (sense and antisense) peaks distributed in 5,932 genes (Supplemental Dataset 1) resulted. Fig. 2A depicts the reads density landscape of these peaks on *TMEM41B* isoforms. A significant intronic peak was found relative to the IgG sample. To explore the reliability of the RIP-seq results, we downloaded and analyzed the two sets of CELF1 CLIP-seq data from HeLa cells [18]. We found that a significant fraction of the reported CLIP peaks from both sets of data overlapped with our RIP peaks. For dataset B, 16.04% CLIP peaks overlapped. In many cases, although the CLIP and RIP peaks were not directly overlapped, they located in the same genes. For example, dataset B CLIP peaks were distributed in 1910 genes, 64.66% of them were overlapped with RIP-seq genes (Supplemental Fig. 1C, *p*-value= $7.380872e^{-89}$, hypergeometric test). This overlap rate were much higher than CELF1-bound genes shared between other published CLIP data [18] .

Furthermore, we observed that most CELF1-binding peaks were within 200-nt in length, while the distance between peaks was mainly represented within 100-nt (Supplemental Fig. 1A and B). These results indicated that CELF1 binding sites were

closely positioned on target RNAs and likely reflected a coordinated action of multiple CELF1-binding events in regulating RNA metabolism.

Interestingly, we observed more CELF1-binding peaks at CDS than intronic regions (Fig. 2B). Considering that we classified peaks spanning exon-intron boundaries as CDS peaks in the ABLIRC pipeline, it could be possible that a fraction of CDS peaks were exon-intron boundary peaks. To explore this possibility, we plotted CELF1-associated peaks onto exon-intron junctions. The results showed that 1319 (8.49%) and 1391 (8.95%) peaks were overlapped with the 5' and 3' splice sites, respectively (Fig. 2C). To validate the above conclusion, we also analyzed the distribution of CELF1 CLIP-seq peaks on exon-intron junctions and found that 4.42% and 5.38% of libA peaks were overlapped with the 5' and 3' splice sites, respectively. The consistency indicates CELF1 binding of splice sites might be direct.

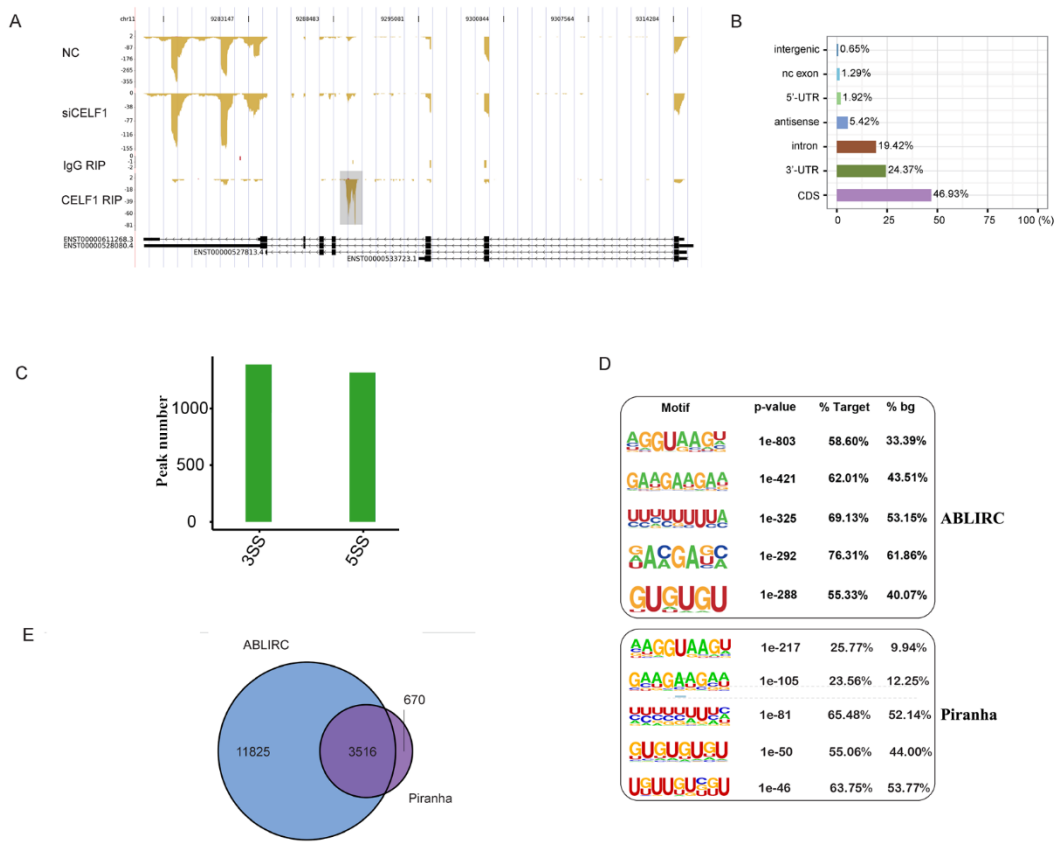

Figure 2. Peak-calling analysis of RIP-seq reads. (*A*) The reads density landscape of CELF1-bound peaks on *TMEM41B* isoforms. (*B*) Genomic distribution of CELF1-bound peaks called by ABLIRC algorithm. (*C*) Statistics analysis of CELF1-bound peaks that have overlap with the 5SS and 3SS. (*D*) Extracted CELF1 peaks motifs using ABLIRC or Piranha. (*E*) The comparative result of ABLIRC and Piranha peak calling methods by Venn diagram analysis.

We then searched for the overrepresented motifs in 15,542 CELF1-binding peaks using Homer (http://homer.salk.edu/homer/motif/index.html). As shown in Fig. 2D, GU-rich motif was among the top 5 CELF1-bound motifs. Consistent with the current understanding, CELF1-binding motifs were enriched at the 3'UTRs. Interestingly, CELF1-bound RNAs are enriched in the conserved sequences of 5' and 3' splice sites, ranking as the top first and third of the overrepresented motifs. The 5'ss motif AGGUAAG was located at the intron-exon boundaries, while the 3'ss only contained

U-rich stretch in intronic regions. The enrichment of 5'ss and 3'ss in the CELF1-binding motifs is highly consistent with the enriched anti-CELF1 reads at the exon-intron junctions.

In order to validate these results, we called CELF1-bound peaks using Piranha, a published software tool for the identification of RNA-protein interaction sites from high-throughput sequencing data [19]. As shown in Fig. 2E, a total of 3,516 peaks called by ABLIRC were validated by Piranha, representing 22.62% and 83.99% of ABLIRC and Piranha peaks, respectively. Moreover, we found that 5'ss motif AGGUAAG, 3'ss motif and GU-rich motif were also highly enriched in Piranha peaks (Fig. 2D), which was similar to those in ABLIRC peaks. We also downloaded the published EZH2 RIP-seq data [25] and analyzed them using ABLIRC algorithm. As a result, a total of 2,639 target genes of EZH2 were obtained (Supplemental Fig. 2A). Among them, 1,039 target genes were identical to the published results [25]. Given that no binding motifs of EZH2 were previously reported, we identified the overrepresented motifs in EZH2-binding peaks. As shown in Supplemental Fig. 2B, the top 1 motif residing in EZH2-binding peaks was CU-enriched (p-value = 1e-34, 20.16% of peaks).

Taken together, it is the first time we showed that ABLIRC algorithm, which was successfully applied to extract the binding motifs of CELF1 RNA binding protein, will be very helpful to understand the regulatory role of protein-RNA interactions in the gene expression processes in living cells.)

3.3 Validation and functional analysis of the CELF1-bound genes in HeLa cells

We determined CELF1-bound genes in HeLa cells by those containing at least one

CELF1-bound peak recovered by ABLIRC. Supplemental Dataset 2 lists a total of 5,123 CELF1-bound genes. We then selected some CELF1 targets for RIP-PCR validation. The quantitative RIP-PCR results showed that all candidate RNA targets had significant enrichment in the anti-CELF1 immunoprecipitate relative to the IgG (Fig. 3A). For instance, significant enrichment was observed in the CELF1 immunoprecipitate relative to IgG immunoprecipitate for NDUFS2, SMAD7 and CELF1, which have been identified to be the targets of CELF1 previously [36, 55]. As a negative control, we showed that PLK2 RNA was not enriched in the CELF1 immunoprecipitate (Supplemental Fig. 3), which is consistent with the RIP-Chip result in the mouse C2C12 cells [55]. In addition, as a marker gene of breast cancer [46], LMO7 RNA was also found to be effectively enriched in the CELF1 immunoprecipitate. Furthermore, the identified CELF1 targets were further assigned to the biochemical pathways in the KEGG database, resulting in 262 KEGG biochemical pathways (Supplemental Dataset 3). Fig. 3B shows top 20 pathways including multiple cancer-related pathways including cell cycle, adherens junction, pathways in cancer, small cell lung cancer, pancreatic cancer, proteoglycan in cancer, p53 signaling pathway and viral carcinogenesis. Interestingly, some RNA metabolic pathways were also enriched including spliceosome, RNA degradation and RNA transport, which suggested the diverse functions of CELF1 in HeLa cells.

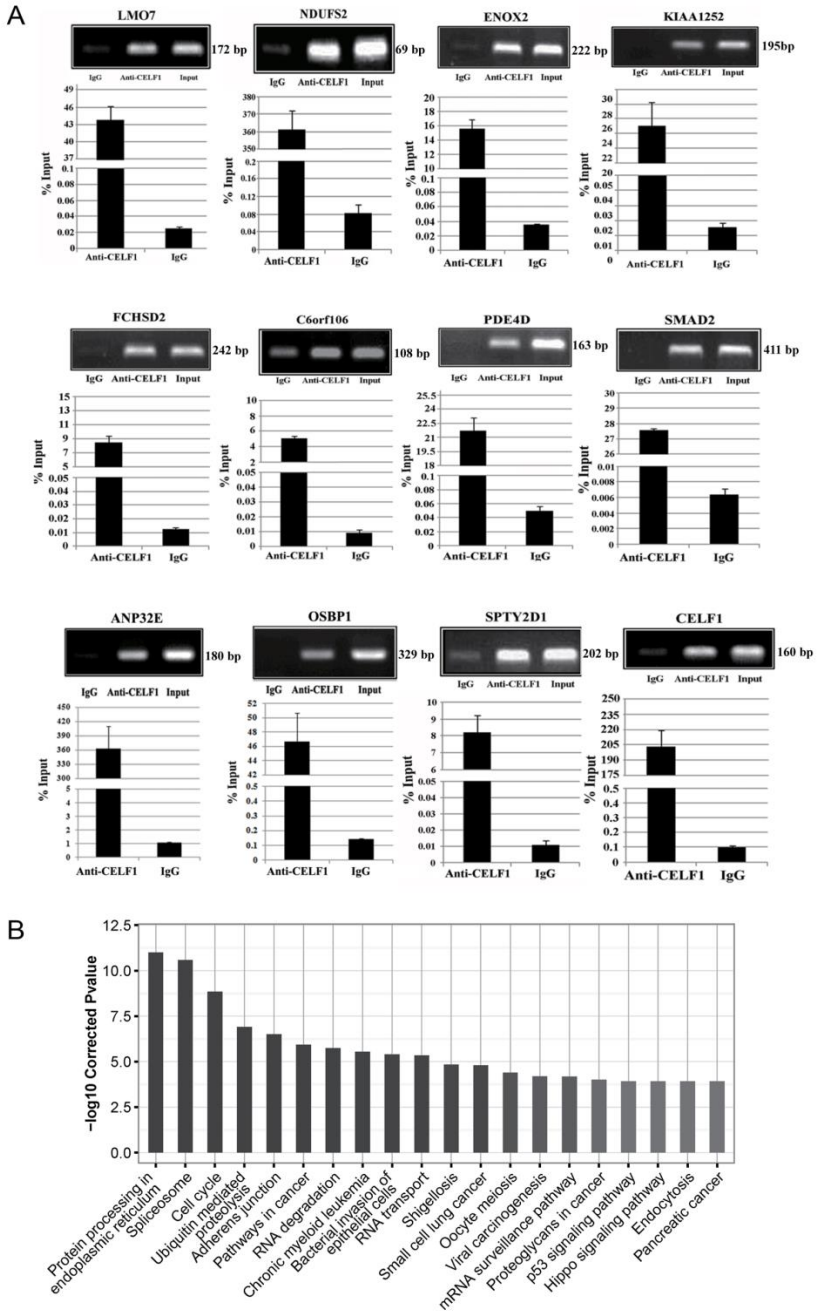

Figure 3. Validation and functional analysis of CELF1-bound genes in HeLa cells. (*A*) RIP-qPCR validation of CELF1-bound genes. (*B*) Bar plot of the top 20 KEGG functionally enriched pathways of CELF1-bound genes.

3.4 Alternative splicing is globally regulated in HeLa cells

Our ABLIRC analysis of RIP-seq data obtained from HeLa cells showed the prevalence of CELF1 binding to 5'ss and 3'ss. Therefore, we further used transcriptome

sequencing (two biological replicates) to explore the potential function of CELF1 in HeLa cells. As shown in Fig. 4A, both western blots and quantitative PCR (q-PCR) experiments showed that CELF1 expression was efficiently downregulated by siRNA in HeLa cells. First transcriptome sequencing generated a total of 14,818,746 mapped reads in NC (Control) and si-CELF1 HeLa cells, in which ~23.5% were junction reads. After running ABLas software tool (under submission), 4,631 alternative splicing events (ASEs) were obtained (Materials and Methods). To obtain the alternative splicing events that changed in response to the reduced CELF1 expression, a cut-off of *p*-value ≤ 0.05 and changed AS ratio ≥ 0.1 were selected, which resulted in 618 regulated alternative splicing events (RASEs) between si-CELF1 and control cells. Furthermore, 10 randomly selected cassette exon/skipped exon events regulated by CELF1 were confidently validated by semi-quantitative RT-PCR (Supplemental Fig. 4). The validated splicing events were located in the following genes, YIPF1 (Yip1 domain family, member 1), CD46 (CD46 molecule, complement regulatory protein), PARD3 (par-3 family cell polarity regulator), GIPC1 (GIPC PDZ domain containing family, member 1), L1CAM (L1 cell adhesion molecule), PPIP5K2 (diphosphoinositol pentakisphosphate kinase 2), ZDHHC16 (zinc finger, DHHC-type containing 16), SUN1 (Sad1 and UNC84 domain containing 1), LMO7 (LIM domain 7), and ITGA6 (integrin, alpha 6). As shown in Supplemental Fig. 4, we observed distinct CELF1 binding peaks on the exon-intron junctions of *LMO7*, *PPIP5K2*, *IGTA6*, *LICAM*, *PARD3* and *CD46* pre-mRNA splicing regions, which indicated CELF1 might directly bind to these targets and regulate alternative splicing

in HeLa cells. In contrast, no obvious CELF1 binding peaks were enriched on *GIPC1*, *YIPF1*, *ZDHHC16* and *SUN1* pre-mRNA splicing regions, indicating these pre-mRNA splicing events could be indirectly regulated by CELF1 in HeLa cells.

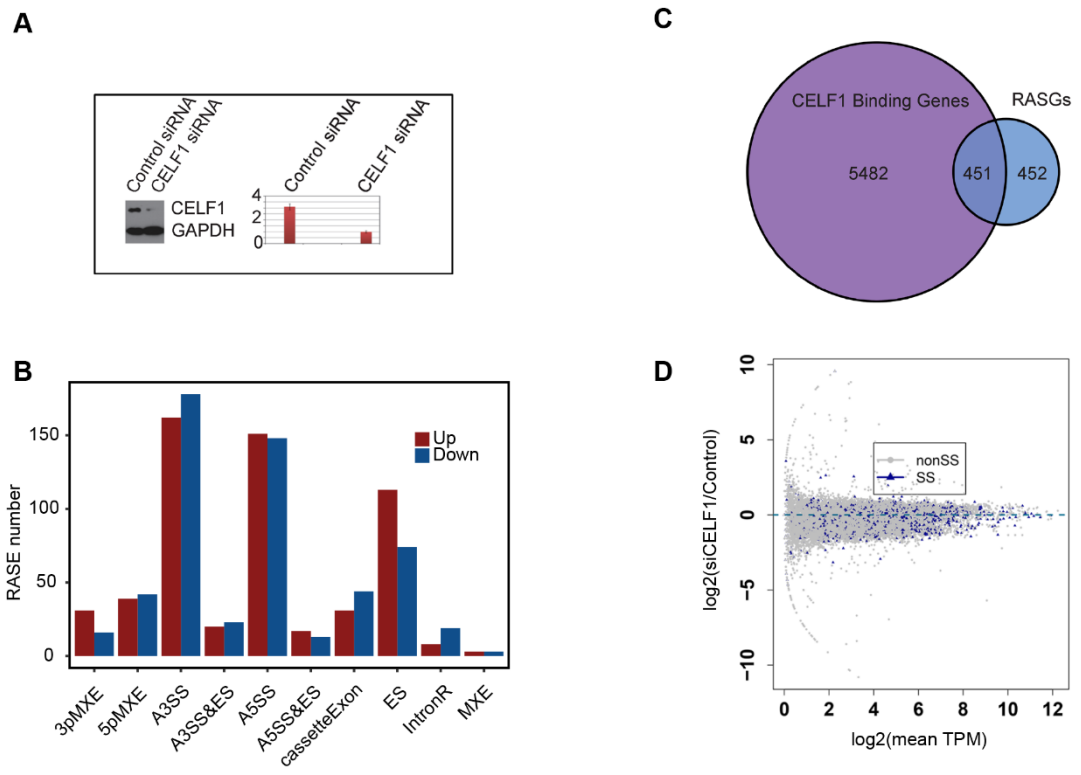

Figure 4. Identification of CELF1-bound and -regulated splicing events using RNA-seq analysis. (*A*) Both western blots and q-PCR experiments showed that CELF1 expression was efficiently down-regulated in HeLa cells by siRNA. (*B*) Classification of different AS types regulated by CELF1 protein. (*C*) The overlap analysis between CELF1-bound genes and the CELF1-regulated AS genes (RASGs) using Venn diagram (*p-value* = $8.71996e^{-32}$). (D). Dot plot represents the expression level of exons bound by CELF1. X-axis represents the mean TPM (in log2) of CELF1-bound exons in the siCELF1 and Control samples. Y-axis represents the log2 fold change between the siCELF1 and Control. Blue points represent exons bound by CELF1 at the splice sites.

The above results greatly extended the role of CELF1 in regulating alternative

splicing. In order to explore more reliable alternative splicing events regulated by CELF1, we performed deeper transcriptome sequencing of Control and si-CELF1 cells, which generated 52,071,459 and 59,136,067 clean reads, respectively. The ABLas program identified a total of 19,466 alternative splicing events in HeLa cells. Fig. 4B showed that 1,122 alternative splicing events changed significantly in the siCELF1 HeLa cells in comparison with WT HeLa cells. Among them, a majority of the splicing events (Fig. 4B) belonged to A3SS (alternative 3' splice site, 340), A5SS (alternative 5' splice site, 299) and exon-skipping (187) categories. The other CELF1-bound and -affected splicing events included CassetteExon (number: 75), 5pMXE (mutually exclusive 5'UTRs, 81), 3pMXE (mutually exclusive exons, 47) and IntronR (Intron retention, number: 27). Collectively, these results showed that A3SS, A5SS and ES are main splicing events regulated by CELF1, which indicated that CELF1 globally regulates alternative splicing events in HeLa cells.

About half (451) of the CELF1-regulated alternative splicing genes overlapped the CELF1-bound genes (Fig. 4C), indicating that CELF1 binding might regulate a lot of alternative splicing events by direct binding to RNA targets. To further explore the relationship between CELF1 binding and alternative splicing regulation, we analyzed the expression and/or splicing inclusion of CELF1-bound exons. The results showed that CELF1 could associate splice sites regardless of the expression level of the exon. Globally, the expression of CELF1-associated exons was reduced when CELF1 expression was down-regulated, indicating a positive correlation between CELF1 association and the exon expression and/or inclusion (Fig. 4D). Consistent with this

hypothesis, analysis of the differentially expressed genes upon siRNA silencing of CELF1 expression showed 264 down-regulated genes, but only 148 upregulated genes (edgeR pipeline, *p*-value≤0.01, fold-change ≥2). Moreover, we observed an increased number of exon skipping events (ES) and decreased number of exon inclusion (cassette exon) when CELF1 expression is reduced (Fig. 4B). All these lines of evidence supported a model that CELF1 binding positively regulates exon inclusion.

3.5 CELF1 regulates the alternative splicing of LMO7

To verify if these identified alternative splicing events are really affected by CELF1 in HeLa cells, we used RT-PCR to detect the changes of the alternative splicing patterns between Control and CELF1-depleted HeLa cells. The quantification of AS pattern was measured as the inclusion/ exclusion (In/Ex) ratio. Supplemental Fig. 4 shows some representative affected alternative splicing events with two alternative splicing modes including exon-skipping, alternative 5' splice site. Here, we further selected LMO7 as an example to study the effect of CELF1 binding on the pre-mRNA splicing. LMO7 is up-regulated in the metastatic stage of multiple cancers and has been characterized as a breast cancer marker gene [46, 56-58]. In this study, we first analyzed the interaction between CELF1 and LMO7 pre-mRNA by RIP-seq technology. As shown in the up panel of Fig. 5A, a large number of CELF1-bound peaks (20 sense peaks and 6 antisense peaks) were found across the exonic locations of *LMO7*. Previous CLIP-seq data [18] also detected 21 peaks from the *LMO7* genomic location (libA), from which 6 peaks were overlapped with our RIP-seq peaks. The RIP-seq and CLIP-seq peaks both implied that LMO7 is a direct target of CELF1 protein in HeLa cells. We also

designed RIP-PCR experiment to confirm the direct interaction between CELF1 and LMO7 pre-mRNA. As shown in Fig. 3, in contrast with IgG control, LMO7 RNA was significantly enriched in the CELF1 IP fraction, which showed that CELF1 binds to LMO7 pre-mRNA in vivo.

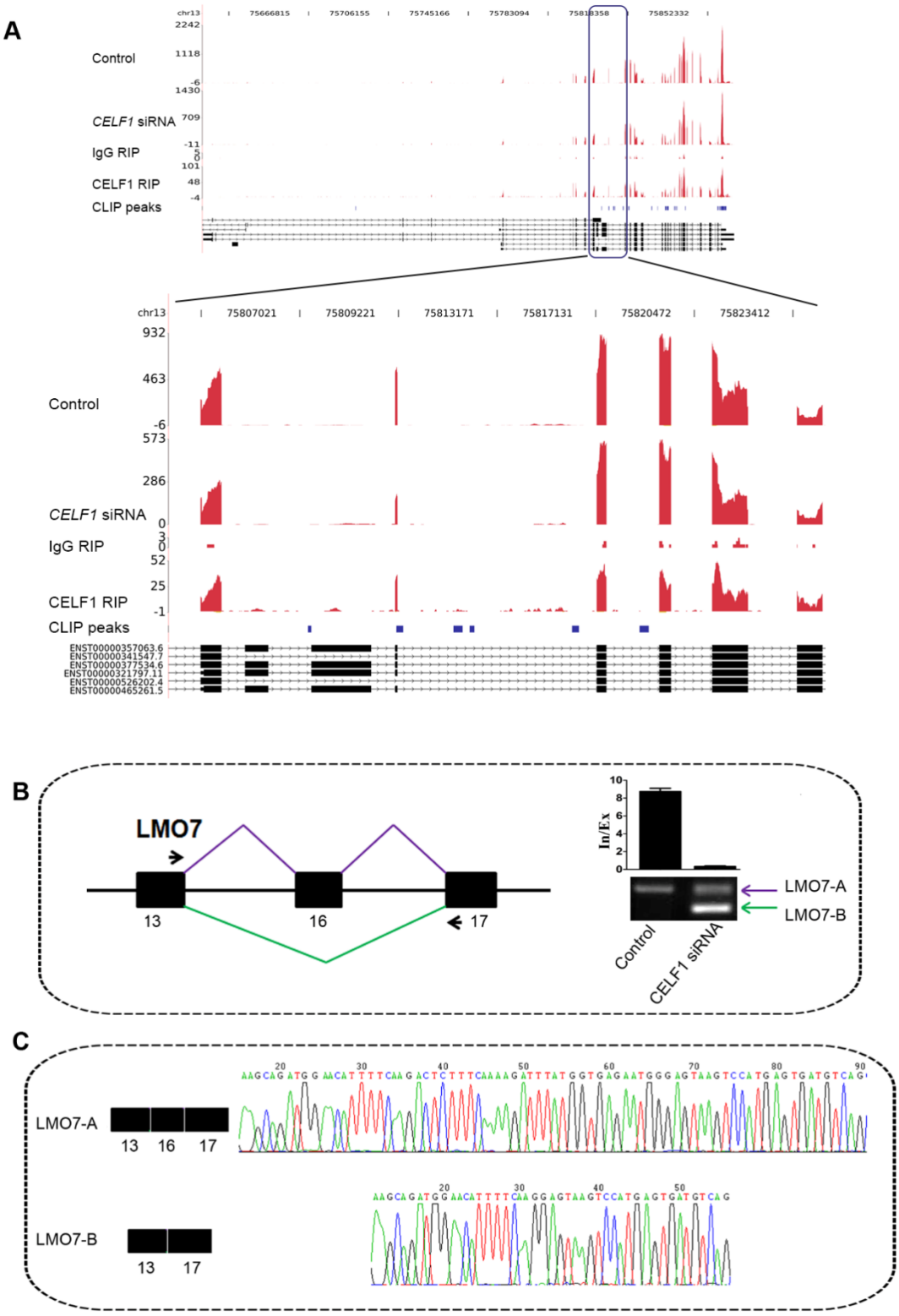

Figure 5. Validation of CELF1-affected alternative splicing events. (*A*) The distribution of reads across the whole region (top panel) and around exon 13, exon 16 and exon 17 regions (bottom panel) in the LMO7 genomic location. CLIP peaks track (Blue) represents the published CLIP peaks of

CELF1 [18]. (*B*) (Left panel) Schematic diagram showing the exon skipping pattern of LMO7 pre-mRNA. (Right panel) RT-PCR validation of the LMO7 pre-mRNA alternative splicing regulation by CELF1. (*C*) The sequencing results of LMO7-A and LMO7-B isoforms.

To explore if the splicing of LMO7 pre-mRNA was regulated by the binding of CELF1, we reduced CELF1 expression by siRNA treatment and observed the splicing pattern change of *LMO7* pre-mRNA. As shown in Fig. 4A, both western blots and q-PCR experiments showed that CELF1 expression was efficiently downregulated in HeLa cells by siRNA. The *LMO7* pre-mRNA splicing pattern was examined by the RT-PCR shown in Fig. 5B. The exon 16 of *LMO7* pre-mRNA is alternatively spliced in normal HeLa cells. the isoform A (LMO7-A), in which exon 16 is included, is the predominant splicing pattern in contrast with isoform B (LMO7-B) in which exon 16 is excluded. RT-PCR experiments revealed that treatment of control siRNA did not change the splicing pattern of exon16. However, LMO7-B increased from 10% in the wild-type HeLa cells to 75% in the HeLa cells treated with CELF1 siRNA. Furthermore, we cut the LMO7-A and LMO7-B isoforms from agarose gel and sequenced them (Fig. 5C). The result clearly showed that the depletion of CELF1 induced the exon 16 exclusion of LMO7 pre-mRNA, which indicated that LMO7 pre-mRNA splicing is directly regulated by CELF1 in HeLa cells.

4. Discussion

As a multifunctional RNA binding protein, CELF1 is involved in the regulation of mRNA metabolism[59, 60] and the altered CELF1 function contributes to the pathology of Myotonic Dystrophy type 1 [61-65]. In addition, various lines of evidence proved

that CELF1 is linked with the processes of cancer and apoptosis [43, 44]. Therefore, some high-throughput approaches have been performed to globally identify CELF1 targets [36, 55, 66], which greatly broaden the understanding of CELF1 functions in various biological processes. Here, for the first time we used RIP-seq approach to globally decipher the CELF1-RNA interactions in human cancer cell line. The RIP-seq results showed that sequencing reads of CELF1-associated RNAs were mainly mapped to 3'UTR region, intron region, CDS region and intergenic region, which suggested CELF1 has versatile RNA targets *in vivo*. Interestingly, ～14.8% reads were distributed on intergenic region in the genome where various non-coding RNAs including miRNA are generated. This result gave us a hint that CELF1 might be involved in the process of non-coding RNA metabolism, which has not been reported before and would be worthwhile to be further explored in future.

Analysis of both the RIP-seq data obtained in this study and the previously published CLIP-seq data, showed that a significant fraction of CELF1-bound peaks was enriched at 5' and 3' splice sites in HeLa cells. The recognition of 5' and 3' splice sites in pre-mRNA are critical events in an alternative splicing decision. Our results showed that alternative splicing is globally regulated by CELF1 in HeLa cells. Specifically, CELF1 binding correlates with an increased exon level and increased exon inclusion. It could be possible that CELF1 regulates the splice site selection by recognizing both 5' and 3' splice sites in HeLa cells. These results may deepen our understanding of the alternative splicing regulation mechanism of CELF1 at the pre-mRNA splicing level.

As each gene has a different expression level, the number of RIP-seq reads mapped

to the genome is not related to the binding specificity of CELF1. Therefore, in order to analyze RIP-seq data, the noise signal brought by gene expression level should be first removed. For instance, the study by the Philip Sharp lab showed that the interference of expression abundance on specific mRNA binding sites of Ago2 can be eliminated by setting transcript abundance of exon array as a reference [67]. Here, we used ABLIRC algorithm to distinguish the enriched binding sites and binding motifs of CELF1 protein from background noise.

CLIP-seq usually generates short-length reads and identifies exact binding sites, which are used to deduce the sequence motifs recognized by RBPs. On the contrary, RIP-seq method is generally thought to identify the intact RNA molecules bound by RBPs. In this study, we have used ABLIRC algorithm to effectively identify the binding peaks and binding motifs of CELF1 protein from RIP-seq data. For example, the GU-rich motif is overrepresented in CELF1-binding peaks. This result is highly consistent with the previous reports that CELF1 regulates the stability of multiple mRNAs by binding to GU-rich sequences in 3'UTRs, which also proved the feasibility of ABLIRC algorithm in analyzing RIP-seq data. We suggest that, during the RIP operation in the lysis buffer, endogenous RNases could lead to RNA fragmentation. Although CLIP-seq is successful in the precise identification of RBP binding sites in various RNA species, its high failure rate and low reproducibility has limited its application. Here we show that deep analysis of RIP-seq data could be an efficient alternative to resolve the CLIP-seq problems. CLIP and RIP techniques of the same RBPs under similar conditions can also generate a large number of complementary data, which deepen our understanding

of the regulatory roles of RBPs in multiple biological processes.

We also investigated the global alternative splicing events regulated by CELF1 using RNA sequencing. A total of 1,122 RASEs were identified from 19,466 alternative splicing events. Among them, about half (451) of the pre-mRNA targets were regulated by direct binding of CELF1in HeLa cells. Importantly, we found that alternative splicing regulation of CELF1 is global, which may represent the flexibility of RNA binding protein to regulate various alternative splicing events in *vivo*. Finally, we introduced LMO7 pre-mRNA as a novel target of CELF1 protein. We showed that the downregulation of CELF1 induced by siRNA led to an increase of exon16 exclusion of LMO7 in HeLa cells, which indicated that the splicing pattern of LMO7 pre-mRNA is regulated by CELF1. Very importantly, it was widely reported that LMO7 was upregulated in the metastatic stage of multiple cancers [56-58]. Moreover, LMO7 was specifically expressed in the metastatic stage of breast cancer cells and has been characterized as a breast cancer marker gene [46]. Therefore, this finding provides a new proof that CELF1 could be implicated in cancer-related pathological processes, which is helpful to discover novel biological functions of CELF1 in the process of tumorigenesis and provide new clues for the anti-tumor therapy in the future.

Acknowledgements

We would like to thank Chao Cheng for technical contributions with RIP-seq data analysis. We are grateful to Dr. Michael J. Leibowitz (University of California, Davis) and Hong Wu for the language polishing. We also thank the members of ABLife Inc.

for their helpful discussions and critical reading of the manuscript. This study was supported by the National Natural Science Foundation of China (31540083, 31000570) to L.Z.. Funding for open access charge: National Natural Science Foundation of China (31540083, 31000570). This study was also supported by ABLife (ABL2013-07010) granted to Y. Zhang.

Accession number

RIP-seq and RNA-seq data have been deposited in NCBI Sequence Read Archive (SRA) Sequence Database with accession number SRP093556, SRP093557, SRP093566, SRP093568 and SRP093570, SRP093571, SRP093573, SRP093574, respectively.

Author Contributions

Y. Zhang and L.Z. designed and coordinated the project. H.X., G.W. and Y. Zhou performed the experimental research and analyzed the data. D.C. and Q.W. performed bioinformatics analysis. Y. Zhang and L.Z. wrote the manuscript. All authors commented on or contributed to the final manuscript.

References:

[1] G. Dreyfuss, V.N. Kim, N. Kataoka, Messenger-RNA binding proteins and the messages they carry, Nat. Rev. Mol. Cell. Biol., 3 (2002) 195–205.

[2] M.J. Moore, From birth to death: the complex lives of eukaryotic mRNAs, Science, 309 (2005) 1514-1518.

[3] J.D. Keene, RNA regulons: coordination of post-transcriptional events, Nat. Rev. Genet., 8 (2007) 533–543.

[4] B.P. Lewis, C.B. Burge, D.P. Bartel, Conserved seed pairing, often flanked by

adenosines, indicates that thousands of human genes are microRNA targets, Cell, 120 (2005) 15-20.

[5] D.P. Bartel, MicroRNAs: target recognition and regulatory functions, Cell, 136 (2009) 215-233.

[6] R.W. Carthew, E.J. Sontheimer, Origins and Mechanisms of miRNAs and siRNAs, Cell, 136 (2009) 642-655.

[7] O.A. Sofola, P. Jin, Y. Qin, R. Duan, H. Liu, M. de Haro, D.L. Nelson, J. Botas, RNA-binding proteins hnRNP A2/B1 and CUGBP1 suppress fragile X CGG premutation repeat-induced neurodegeneration in a Drosophila model of FXTAS, Neuron., 55 (2007) 565-571.

[8] J.D. Keene, S.A. Tenenbaum, Eukaryotic mRNPs may represent posttranscriptional operons, Mol. Cell., 9 (2002) 1161-1167.

[9] P.L. Boutz, P. Stoilov, Q. Li, C.H. Lin, G. Chawla, K. Ostrow, L. Shiue, M.J. Ares, D.L. Black, A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons, Genes Dev, 21 (2007) 1636-1652.

[10] H. Ji, An integrated software system for analyzing ChIP-chip and ChIP-seq data, Nat. Biotechnol., 26 (2008) 1293-1300.

[11] N.U. Rashid, ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions, Genome Biol., 12 (2011) R67.

[12] J. Rozowsky, PeakSeq enables systematic scoring of ChIP-seq experiments

relative to controls, Nat. Biotechnol., 1 (2009) 66-75.

[13] Y. Zhang, Model-based analysis of ChIP-Seq (MACS), Genome Biol., 9 (2008) R137.

[14] S. Kishore, A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins, Nat. Methods, 8 (2011) 559-564.

[15] Y. Li, RIPSeeker: a statistical package for identifying protein-associated transcripts from RIP-seq experiments, Nucleic Acids Res., 41 (2013) e94.

[16] Kucukural A, Özadam H, Singh G, Moore MJ, C. C., ASPeak: an abundance sensitive peak detection algorithm for RIP-Seq, Bioinformatics, 29 (2013) 2485-2486.

[17] Y. Sugimoto, J. Konig, S. Hussain, B. Zupan, T. Curk, M. Frye, J. Ule, Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions, Genome Biol, 13 (2012) R67.

[18] O. Le Tonqueze, B. Gschloessl, V. Legagneux, L. Paillard, Y. Audic, Identification of CELF1 RNA targets by CLIP-seq in human HeLa cells, Genomics data, 8 (2016) 97-103.

[19] 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.

[20] S.A. Tenenbaum, C.C. Carson, P.J. Lager, J.D. Keene, Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays, Proc. Natl Acad. Sci., 97 (2000) 14085-14090.

[21] G.W. Yeo, N.G. Coufal, T.Y. Liang, G.E. Peng, X.D. Fu, F.H. Gage, An RNA code

for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells, Nat Struct Mol Biol., 16 (2009) 130-137.

[22] Y. Xue, Y. Zhou, T. Wu, T. Zhu, X. Ji, Y.S. Kwon, C. Zhang, G. Yeo, D.L. Black, H. Sun, Genome-wide Analysis of PTB-RNA Interactions Reveals a Strategy Used by the General Splicing Repressor to Modulate Exon Inclusion or Skipping, Mol Cell., 36 (2009) 996-1006.

[23] R. Xiao, P. Tang, B. Yang, J. Huang, u.Y. Zho, C. Shao, H. Li, H. Sun, Y. Zhang., X.D. Fu, Nuclear matrix factor hnRNP U/SAF-A exerts a global control of alternative splicing by regulating U2 snRNP maturation, Mol Cell., 45 (2012) 656-668.

[24] 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.

[25] J. Zhao, T.K. Ohsumi, J.T. Kung, Y. Ogawa, D.J. Grau, K. Sarma, J.J. Song, R.E. Kingston, M. Borowsky, J.T. Lee, Genome-wide Identification of Polycomb-Associated RNAs by RIP-seq, Mol. Cell., 40 (2010) 939-953.

[26] Z. Nie, F. Zhou, D. Li, Z. Lv, J. Chen, Y. Liu, J. Shu, Q. Sheng, W. Yu, W. Zhang, RIP-seq of BmAgo2-associated small RNAs reveal various types of small non-coding RNAs in the silkworm, Bombyx mori, BMC genomics, 14 (2013) 661.

[27] C.O. Nicholson, M. Friedersdorf, J.D. Keene, Quantifying RNA binding sites transcriptome-wide using DO-RIP-seq, RNA, 23 (2017) 32-46.

[28] R.S. Savkur, A.V. Philips, T.A. Cooper, Aberrant regulation of insulin receptor

alternative splicing is associated with insulin resistance in myotonic dystrophy, Nat. Genet., 29 (2001) 40-47.

[29] N. Charlet-B, R.S. Savkur, G. Singh, A.V. Philips, E.A. Grice, T.A. Cooper, Loss of the muscle-specific chloride channel in type 1 myotonic dystrophy due to misregulated alternative splicing, Mol. Cell., 10 (2002) 45-53.

[30] A.V. Philips, L.T. Timchenko, T.A. Cooper, Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy, Science, 280 (1998) 737-741.

[31] T.H. Ho, D. Bundman, D.L. Armstrong, T.A. Cooper, Transgenic mice expressing CUG-BP1 reproduce splicing mis-regulation observed in myotonic dystrophy, Hum. Mol. Genet., 14 (2005) 1539-1547.

[32] E.T. Wang, A.J. Ward, J.M. Cherone, J. Giudice, T.T. Wang, D.J. Treacy, N.J. Lambert, P. Freese, T. Saxena, T.A. Cooper, Antagonistic regulation of mRNA expression and splicing by CELF and MBNL proteins, Genome Res., 25 (2015) 858-871.

[33] V.A. Barron, H. Zhu, M.N. Hinman, A.N. Ladd, H. Lou, The neurofibromatosis type I pre-mRNA is a novel target of CELF protein-mediated splicing regulation, Nucleic Acids Res, 38 (2010) 253-264.

[34] L. Zhang, J.E. Lee, J. Wilusz, C.J. Wilusz, The RNA-binding Protein CUGBP1 Regulates Stability of Tumor Necrosis Factor mRNA in Muscle Cells: implications for Myotonic Dystrophy, J. Biol. Chem., 283 (2008) 22457-22463.

[35] I.A. Vlasova, N.M. Tahoe, D. Fan, O. Larsson, B. Rattenbacher, J.R. Sternjohn, J. Vasdewani, G. Karypis, C.S. Reilly, P.B. Bitterman, Conserved GU-rich elements

mediate mRNA decay by binding to CUG-binding protein 1, Mol Cell., 29 (2008) 263-270.

[36] B. Rattenbacher, D. Beisang, D.L. Wiesner, J.C. Jeschke, M. von Hohenberg, I.A. St Louis-Vlasova, P.R. Bohjanen, Analysis of CUGBP1 targets identifies GU-repeat sequences that mediate rapid mRNA decay, Mol. Cell. Biol., 30 (2010) 3970-3980.

[37] J. Russo, J.E. Lee, C.M. Lopez, J. Anderson, T.P. Nguyen, A.M. Heck, J. Wilusz, C.J. Wilusz, The CELF1 RNA-Binding Protein Regulates Decay of Signal Recognition Particle mRNAs and Limits Secretion in Mouse Myoblasts, PloS one, 12 (2017) e0170680.

[38] N.A. Timchenko, P. Iakova, Z.J. Cai, J.R. Smith, L.T. Timchenko, Molecular basis for impaired muscle differentiation in myotonic dystrophy, Mol. Cell Biol., 21 (2001) 6927-6938.

[39] N.A. Timchenko, R. Patel, P. Iakova, Z.J. Cai, L. Quan, L.T. Timchenko, Over expression of CUG triplet repeat-binding protein-CUGBP1 in mice inhibits myogenesis, J. Biol. Chem., 279 (2004) 13129-13139.

[40] L.I.a.B. Vlasova-St, P.R. , Coordinate regulation of mRNA decay networks by GU-rich elements and CELF1, Current Opinion in Genetics & Development., 21 (2011) 444-451.

[41] Edwards. J., E. Malaurie, A. Kondrashov, J. Long, de Moor, C.H., M.S. Searle, E. J., Sequence determinants for the tandem recognition of UGU and CUG rich RNA elements by the two N--terminal RRMs of CELF1, Nucleic Acids Res., 39 (2011) 8638-8650.

[42] M. Teplova, J. Song, H.Y. Gaw, A. Teplov, D.J. Patel, Structural insights into RNA recognition by the alternate-splicing regulator CUG-binding protein 1, Structure., 18 (2010) 1364-1377.

[43] E.T. Chang, J.M. Donahue, L. Xiao, Y. Cui, J.N. Rao, D.J. Turner, W.S. Twaddell, J.Y. Wang, R.J. Battafarano, The RNA-binding protein CUG-BP1 increases survivin expression in oesophageal cancer cells through enhancedmRNA stability, Biochem. J., 446 (2012) 113-123.

[44] F.J. Rodriguez, C. Giannini, Y.W. Asmann, M.K. Sharma, A. Perry, K.M. Tibbetts, R.B. Jenkins, B.W. Scheithauer, S. Anant, S. Jenkins, Gene expression profiling of NF-1-associated and sporadic pilocytic astrocytoma identifies aldehyde dehydrogenase 1 family member L1 (ALDH1L1) as an underexpressed candidate biomarker in aggressive subtypes, J. Neuropathol. Exp. Neurol., 67 (2008) 1194-1204.

[45] A. Chaudhury, S. Cheema, J.M. Fachini, N. Kongchan, G. Lu, L.M. Simon, T. Wang, S. Mao, D.G. Rosen, M.M. Ittmann, S.G. Hilsenbeck, C.A. Shaw, J.R. Neilson, CELF1 is a central node in post-transcriptional regulatory programmes underlying EMT, Nature communications, 7 (2016) 13362.

[46] C.M. Perou, T. Sørlie, M.B. Eisen, R.M. van de, S.S. Jeffrey, C.A. Rees, J.R. Pollack, D.T. Ross, H. Johnsen, L.A. Akslen, Molecular portraits of human breast tumours, Nature., 406 (2000) 747-752.

[47] S. Jayaseelan, F. Doyle, S. Currenti, S.A. enenbaum, RIP: an mRNA localization technique, Methods Mol Biol., 714 (2011) 407-422.

[48] 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 biology, 14 (2013) R36.

[49] 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, Molecular cell, 38 (2010) 576-589.

[50] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics, 26 (2010) 139-140.

[51] E.T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S.F. Kingsmore, G.P. Schroth, C.B. Burge, Alternative isoform regulation in human tissue transcriptomes, Nature, 456 (2008) 470-476.

[52] E.T. Wang, A.J. Ward, J.M. Cherone, J. Giudice, T.T. Wang, D.J. Treacy, N.J. Lambert, P. Freese, T. Saxena, T.A. Cooper, C.B. Burge, Antagonistic regulation of mRNA expression and splicing by CELF and MBNL proteins, Genome Res, 25 (2015) 858-871.

[53] A. Masuda, H.S. Andersen, T.K. Doktor, T. Okamoto, M. Ito, B.S. Andresen, K. Ohno, CUGBP1 and MBNL1 preferentially bind to 3' UTRs and facilitate mRNA decay, Scientific reports, 2 (2012) 209.

[54] S.W. Chi, J.B. Zang, A. Mele, R.B. Darnell, Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps, Nature, 460 (2009) 479-486.

[55] J.E. Lee, J.Y. Lee, J. Wilusz, B. Tian, C.J. Wilusz, Systematic Analysis of Cis-

Elements in Unstable mRNAs Demonstrates that CUGBP1 Is a Key Regulator of mRNA Decay in Muscle Cells, PLoS One., 5 (2010) e11201.

[56] M. Furuya, N. Tsuji, T. Endoh, R. Moriai, D. Kobayashi, A. Yagihashi, N. Watanabe, A novel gene containing PDZ and LIM domains, PCD1, is overexpressed in human colorectal cancer, Anticancer Res., 22 (2002) 4183-4186.

[57] S. Kang, H. Xu, X. Duan, J.J. Liu, Z. He, F. Yu, S. Zhou, X.Q. Meng, M. Cao, G.C. Kennedy, PCD1, a novel gene containing PDZ and LIM domains, is overexpressed in several human cancers, Cancer Res.,, 60 (2000) 5296-5302.

[58] M. Sasaki, N. Tsuji, M. Furuya, K. Kondoh, C. Kamagata, D. Kobayashi, A. Yagihashi, N. Watanabe, PCD1, a novel gene containing PDZ and LIM domains, is overexpressed in human breast cancer and linked to lymph node metastasis, Anticancer Res., 23 (2003) 2717-2721.

[59] P. Iakova, G.L. Wang, L. Timchenko, M. Michalak, O.M. Pereira-Smith, J.R. Smith, N.A. Timchenko, Competition of CUGBP1 and calreticulin for the regulation of p21 translation determines cell fate, EMBO J., 23 (2004) 406-417.

[60] Y. Zheng, W.K. Miskimins, CUG-binding protein represses translation of p27Kip1 mRNA through its internal ribosomal entry site, RNA Biol., 8 (2011) 365-371.

[61] J.D. Brook, M.E. McCurrach, H.G. Harley, A.J. Buckler, D. Church, H. Aburatani, K. Hunter, V.P. Stanton, J.P. Thirion, T. Hudson, Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member, Cell., 69 (1992) 385.

[62] G. Sicot, G. Gourdon, M. Gomes-Pereira, Myotonic dystrophy, when simple

repeats reveal complex pathogenic entities: new findings and future challenges, Hum. Mol. Genet., 20 (2011) R116-R123.

[63] B. Schoser, L. Timchenko, Myotonic dystrophies 1 and 2: complex diseases with complex mechanisms, Curr. Genomics., 11 (2010) 77-90.

[64] A.J. Ward, M. Rimer, J.M. Killian, J.J. Dowling, T.A. Cooper, CUGBP1 overexpression in mouse skeletal muscle reproduces features of myotonic dystrophy type 1, Hum. Mol. Genet., 19 (2010) 3614-3622.

[65] M. Koshelev, S. Sarma, R.E. Price, X.H. Wehrens, T.A. Cooper, Heart-specific overexpression of CUGBP1 reproduces functional and molecular abnormalities of myotonic dystrophy type 1, Hum. Mol. Genet., 19 (2010) 1066-1075.

[66] A. Graindorge, O. Le Tonquèze, R. Thuret, N. Pollet, H.B. Osborne, Y. Audic, Identification of CUG-BP1/EDEN-BP target mRNAs in Xenopus tropicalis, Nucleic Acids Res., 36 (2008) 1861-1870.

[67] A.K. Leung, A.G. Young, A. Bhutkar, G.X. Zheng, A.D. Bosson, C.B. Nielsen, P.A. Sharp, Genome-wide identification of Ago2 binding sites from mouse embryonic stem cells with and without mature microRNAs, Nat. Struct. Mol. Biol., 18 (2011) 237-244.

Figure Legends:

Figure 1. RIP-Seq assay and the profile of sequencing reads aligned to the human genome. (*A*) Immunoprecipitation of CELF1-associated RNAs using CELF1 monoclonal antibody. The efficient immunoprecipitation of CELF1 protein from HeLa

extracts was validated by western blots. (*B*) Bar plot of the genomic region distribution of uniquely mapped anti-CELF1 reads. P-value was obtained by Fisher exact test.

Figure 2. Peak-calling analysis of RIP-seq reads. (*A*) The reads density landscape of CELF1-bound peaks on *TMEM41B* isoforms. (*B*) Genomic distribution of CELF1-bound peaks called by ABLIRC algorithm. (*C*) Statistics analysis of CELF1-bound peaks that have overlap with the 5SS and 3SS. (*D*) Extracted CELF1 peaks motifs using ABLIRC or Piranha. (*E*) The comparative result of ABLIRC and Piranha peak calling methods by Venn diagram analysis.

Figure 3. Validation and functional analysis of CELF1-bound genes in HeLa cells. (*A*) RIP-qPCR validation of CELF1-bound genes. (*B*) Bar plot of the top 20 KEGG functionally enriched pathways of CELF1-bound genes.

Figure 4. Identification of CELF1-bound and -regulated splicing events using RNA-seq analysis. (*A*) Both western blots and q-PCR experiments showed that CELF1 expression was efficiently down-regulated in HeLa cells by siRNA. (*B*) Classification of different AS types regulated by CELF1 protein. (*C*) The overlap analysis between CELF1-bound genes and the CELF1-regulated AS genes (RASGs) using Venn diagram (*p-value* = $8.71996e^{-32}$). (D). Dot plot represents the expression level of exons bound by CELF1. X-axis represents the mean TPM (in log2) of CELF1-bound exons in the siCELF1 and Control samples. Y-axis represents the log2 fold change between the siCELF1 and Control. Blue points represent exons bound by CELF1 at the splice sites.

Figure 5. Validation of CELF1-affected alternative splicing events. (*A*) The distribution of reads across the whole region (top panel) and around exon 13, exon 16

and exon 17 regions (bottom panel) in the LMO7 genomic location. CLIP peaks track (Blue) represents the published CLIP peaks of CELF1[18]. (*B*) (Left panel) Schematic diagram showing the exon skipping pattern of LMO7 pre-mRNA. (Right panel) RT-PCR validation of the LMO7 pre-mRNA alternative splicing regulation by CELF1. (*C*) The sequencing results of LMO7-A and LMO7-B isoforms.

Highlights:

1. RIP-seq identifies the binding peaks and motifs of CELF1 protein on RNA targets in HeLa cells.

2. 5’ and 3’ splice site motifs are highly enriched in the CELF1-bound peaks in HeLa cells.

3. Transcriptome analysis reveals that alternative splicing is globally regulated by CELF1 in HeLa cells.

4. The inclusion of exon 16 of LMO7 gene, a breast cancer signature gene, is positively regulated by CELF1.