Six GU-rich (6GUR) FUS-binding motifs detected by normalization of CLIP-seq by Nascent-seq

Jun-ichi Takeda, Akio Masuda, Kinji Ohno

PII: S0378-1119(17)30250-0
DOI: doi: 10.1016/j.gene.2017.04.008
Reference: GENE 41857

To appear in: *Gene*

Received date: 7 September 2016
Revised date: 3 April 2017
Accepted date: 5 April 2017

Please cite this article as: Jun-ichi Takeda, Akio Masuda, Kinji Ohno , Six GU-rich (6GUR) FUS-binding motifs detected by normalization of CLIP-seq by Nascent-seq. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Gene(2017), doi: 10.1016/j.gene.2017.04.008

Six GU-rich ($6GU_R$) FUS-binding motifs detected by normalization of CLIP-seq by Nascent-seq

Jun-ichi Takeda, Akio Masuda and Kinji Ohno*

Division of Neurogenetics, Center for Neurological Diseases and Cancer, Nagoya University Graduate School of Medicine, 65 Tsurumai, Showa-ku, Nagoya 466-8550, Japan

*Corresponding author
Kinji Ohno, Division of Neurogenetics, Center for Neurological Diseases and Cancer, Nagoya University Graduate School of Medicine, 65 Tsurumai, Showa-ku, Nagoya 466-8550, Japan. e-mail: ohnok@med.nagoya-u.ac.jp

Keywords
FUS; RNA-binding protein; RNA-binding motif; CLIP-seq; Nascent-seq

Abbreviations
CIMS, cross-linking-induced mutation sites; CLIP, cross-linking and immunoprecipitation; RBD, RNA-binding domain; RBM, RNA-binding motif; RBP, RNA-binding protein; ROC, receiver operating characteristic; RPKM, reads per kilobase per million mapped reads

Abstract

FUS, an RNA-binding protein (RBP), is mutated or abnormally regulated in neurodegenerative disorders. FUS regulates various aspects of RNA metabolisms. FUS-binding sites are rich in GU contents and are highly degenerative. FUS-binding motifs of GGU, GGUG, GUGGU and CGCGC have been previously reported. These motifs, however, are applicable to a small fraction of FUS-binding sites. As CLIP-seq tags are enriched in genes that are highly expressed, we normalized CLIP-seq tags by Nascent-seq tags or RNA-seq tags of mouse N2a cells. Nascent-seq identifies nascent transcripts before being processed for splicing and polyadenylation. We extracted frequently observed 4-nt motifs from Nascent-seq-normalized CLIP regions, RNA-seq-normalized CLIP regions, and native CLIP regions. Specific GU-rich motifs were best detected in Nascent-seq-normalized CLIP regions. Analysis of structural motifs using Nascent-seq-normalized CLIP regions also predicted GU-rich sequence forming a stem structure. Sensitivity and specificity were calculated by examining whether the extracted motifs were present at the cross-linking-induced mutation sites (CIMS), where FUS was directly bound. We found that a combination of six motifs (UGUG, CUGG, UGGU, GCUG, GUGG, and UUGG), which were extracted from Nascent-seq-normalized CLIP-regions, had a better discriminative power than (i) motifs extracted from RNA-seq-normalized CLIP regions, (ii) motifs extracted from native CLIP regions, (iii) previously reported individual motifs, or (iv) 15 motifs in SpliceAid 2. Validation of the 6 GU-rich ($6GU_R$) motifs using CLIP-seq of the cerebrum and the whole brain showed that the $6GU_R$ motifs were specifically enriched in CIMS. The number of the $6GU_R$ motifs in an uninterrupted region was counted and multiplied by four to calculate the area, which was defined as the $6GU_R$-Score. The $6GU_R$-Score of 8 or more best discriminated CIMS from CIMS-flanking regions. We propose that the $6GU_R$ motifs predict FUS-binding sites more efficiently than previously reported individual motifs or 15 motifs in SpliceAid 2.

1. Introduction

Fused in Sarcoma (FUS, formerly called translocated-in-liposarcoma, TLS) is an RNA-binding protein (RBP), which belongs to the FET family. The FET family is comprised of FUS, the EWS RNA-binding protein 1 (EWSR1), and the TATA-box binding protein-associated factor 15 (TAF15) (Masuda et al., 2016). FUS plays pivotal roles in RNA metabolisms (splicing, mRNA transport, and microRNA processing), DNA repair, and cell proliferation through binding to target RNAs (Fredericks et al., 2015). Mutations and dysregulations of FUS are causally associated with the pathogenesis of amyotrophic lateral sclerosis (ALS) and frontotemporal lobar degeneration (FTLD) (Fujioka et al., 2013).

Recent advances in high-throughput sequencing technology revealed that most RBPs have thousands of RNA target sites. Identification of RBP-binding sites is critical for understanding the mechanism of RNA processing (Van Nostrand et al., 2016). FUS has four RNA-binding domains (RBDs): two arginine-glycine-glycine boxes (RGGs), an RNA-recognition motif (RRM), and a zinc finger domain (ZnF) (Da Cruz and Cleveland, 2011). Presence of multiple RBDs is likely to be one of the major causes that make identification of the FUS-binding motifs challenging (Li et al., 2014). More than 10 years ago, a GGUG motif was first identified by SELEX (systematic evolution of ligands by exponential enrichment) using a bacterially expressed GST-FUS protein (Lerga et al., 2001). An NMR study revealed that ZnF of FUS recognizes GGUG-containing RNA, while RRM of FUS has no observable interaction with the GGUG-containing RNA (Iko et al., 2004). An *in vitro* RNAcompete analysis with the recombinant RRM domain of FUS showed that the FUS motif was CGCGC (Ray et al., 2013). The CGCGC motif, however, is different from GU-rich motifs reported by others. In addition to these *in vitro* analyses, CLIP-seq (cross-linking and immunoprecipitation, followed by high-throughput sequencing) and iCLIP (individual nucleotide-resolution cross-linking and immunoprecipitation) disclosed that FUS-binding motifs were GUGGU and GGU, respectively (Lagier-Tourenne et al., 2012; Rogelj et al., 2012). In concordance with these GU-rich motifs, another CLIP-seq study disclosed that FUS binds to GU- or GGU-containing sequences on intron 6-exon 7-intron 7 of *FUS*, although no consensus motifs were deduced (Zhou et al., 2013). We similarly reported that FUS binds to GU-rich sequences by CLIP-seq analysis, but has no specific motifs (Masuda et al., 2015). We (Ishigaki et al., 2012) and others (Hoell et al., 2011) additionally reported that FUS binds to secondary structures enriched in G/C or U/A nucleotides. However, the secondary structures accounted for less than 10% of the identified FUS-binding sites (Ishigaki et al., 2012). Another study showed that FUS binds to RNA only in a length-dependent manner (Wang et al., 2015b). Other studies demonstrated that there are no specific binding motifs or alternative binding mechanisms like secondary structures (Schwartz et al., 2012; Nakaya et al., 2013). Thus, there are no approved consensus FUS-binding motifs. Nonetheless, FUS regulates RNA metabolisms in a position-specific manner (Ishigaki et al., 2012; Masuda et al., 2015). FUS, however, is not exceptional, because binding motifs have been determined in only 15% of human RBPs (Ray et al., 2013). Dependable prediction of FUS-binding sites is essential for understanding the functions of FUS under physiological and pathological conditions.

CLIP-seq extensively identifies RNA-binding sites in specific cells or tissues by UV cross-linking of an RBP with an RNA fragment (Zhang and Darnell, 2011). Nascent-seq extensively identifies nascent transcripts in the chromatin fraction in the nucleus (Menet et al., 2012). In Nascent-seq, nascent RNA derived from chromatin fraction is treated with DNase I, followed by removal of polyadenylated RNA and ribosomal RNA. Unlike RNA-seq, Nascent-seq identifies transcripts immediately after transcription and before being processed for splicing and polyadenylation. Therefore, Nascent-seq can measure the expression levels of intron-containing nascent RNA in the

nucleus. In contrast, RNA-seq has three drawbacks: (i) mRNAs in both the nucleus and cytoplasm are detectable; (ii) mRNA lacks introns, where an RBP can bind; and (iii) gene-specific stability of mRNA affects individual RNA-seq coverage. CLIP-seq tags are enriched in highly expressed genes, while they are underrepresented in scarcely expressed genes (Kishore et al., 2011; Wang et al., 2015a). Normalization of CLIP-seq by RNA-seq or Nascent-seq is expected to correct for enrichment of FUS-bind sites in highly expressed genes. Indeed, in eCLIP, CLIP-seq is normalized by RNA-seq, which is called size-matched input (SMInput) (Van Nostrand et al., 2016). As far as we know, FUS-binding motifs have not been analyzed by any normalization method. In an effort to identify FUS-binding motifs, we analyzed FUS CLIP-seq, Nascent-seq, and RNA-seq of N2a mouse neuroblastoma cells. We normalized CLIP-seq tags by Nascent-seq tags or RNA-seq tags to unbiasedly extract FUS-binding sites and to increase the diversity of RBP-binding sites. We here show that normalization of CLIP-seq by Nascent-seq best detects FUS-binding motifs.

2. Materials and Methods

2.1. CLIP-seq data of mouse neuroblastoma cell line (N2a cells) for motif extraction

We used FUS CLIP-seq that we previously obtained from N2a cells (N2a_CLIP) (Masuda et al., 2015). CSFASTQ sequences of N2a cells were comprised of 50-bp strand-specific single-end reads by SOLiD 4 System (Thermo Fisher Scientific). N2a_CLIP was previously deposited in the DDBJ Sequence Read Archive (DRA) (Mashima et al., 2016) with an accession number of DRA001190 (Masuda et al., 2015).

Procedures on N2a_CLIP were essentially the same as our previous report (Masuda et al., 2015), and are summarized in Supplementary Table S1. Briefly, the reads were mapped to the mouse reference genome (UCSC mm9) (Speir et al., 2016) using BioScope version 1.3.1 (Thermo Fisher Scientific). Multiply aligned reads, unreliable reads, and PCR duplicates were removed by Avadis NGS version 1.3 (Strand NGS). Removal of PCR duplicates was important, because CLIP-seq involves PCR amplification of cDNA library with limited complexities (Wang et al., 2015a). The peaks of mapped sequences of N2a_CLIP were called by MACS version 1.4.2 (Zhang et al., 2008). MACS was developed for ChIP-seq, but has also been used for CLIP-seq (Chen et al., 2014).

2.2. CLIP-seq data of the mouse cerebrum and the mouse whole brain for validation

We also used two sets of FUS CLIP-seq obtained from the mouse cerebrum by us (cerebrum_CLIP) (Ishigaki et al., 2012), and the mouse whole brain by others (brain_CLIP) (Lagier-Tourenne et al., 2012). CSFASTQ sequences of cerebrum_CLIP and FASTQ sequences of brain_CLIP were comprised of 50-bp and 36-bp strand-specific single-end reads sequenced by SOLiD System 3.0 (Thermo Fisher Scientific) and HiSeq 2000 (Illumina), respectively. The cerebrum_CLIP and brain_CLIP were deposited in the NCBI Gene Expression Omnibus (GEO) (Barrett et al., 2013) with accession numbers of GSE37190 (Ishigaki et al., 2012) and GSE40651 (Lagier-Tourenne et al., 2012), respectively.

Procedures on cerebrum_CLIP were the same as above and are summarized in Supplementary Table S2. Similarly, procedures on brain_CLIP are summarized in Supplementary Table S3. Briefly, the quality of the FASTQ sequences was first checked by FastQC version 0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and the sequences were filtered by Trimmomatic version 0.32 (Bolger et al., 2014). Second, the quality-controlled FASTQ sequences were mapped to the mouse reference genome (UCSC mm9) by BWA version 0.7.10 (Li and Durbin, 2010) to generate a SAM file. Third, the SAM file was converted to a BAM file, and PCR duplicates were removed

by SAMtools version 0.1.19 (Li et al., 2009).

2.3. Nascent-seq data of N2a cells

FASTQ sequences of Nascent-seq of N2a cells (N2a_Nascent) (Masuda et al., 2015) were comprised of 100-bp strand-specific paired-end reads sequenced by HiSeq 2000 (accession number, DRA003231). Procedures on the FASTQ sequences of N2a_Nascent were essentially the same as our previous report (Masuda et al., 2015), and are summarized in Supplementary Table S4. The sequences were mapped to the mouse reference genome (UCSC mm9) by BWA.

2.4. RNA-seq data of N2a cells

To compare the effect of normalization between pre-mature and mature RNA, we also used RNA-seq of native N2a cells (N2a_RNA) (Han et al., 2014). FASTQ sequences were comprised of 100-bp strand-specific paired-end reads sequenced by HiSeq 2000 (accession number, GSE45119). Procedures on the FASTQ sequences of N2a_RNA were the same as N2a_Nascent, and are summarized in Supplementary Table S5.

2.5. Definition of Nascent-seq-normalized CLIP regions, RNA-seq-normalized CLIP regions, and native CLIP regions, as well as search for putative motifs with MEME-ChIP

BAM files of N2a_CLIP, N2a_Nascent, and N2a_RNA were first converted into a Wiggle format by Pyicoteo version 2.0.7 (Althammer et al., 2011) (Supplementary Tables S1, S4 and S5). N2a_CLIP coverage was normalized by N2a_Nascent coverage at a single nucleotide level using the following equation:

Norm_N2a_CLIP_coverage = (N2a_CLIP coverage / genome-wide N2a_CLIP coverage) / (N2a_Nascent coverage / genome-wide N2a_Nascent coverage)

As the numbers of CLIP-seq tags and Nascent-seq tags are proportional to the number of analyzed cells and to the number of high-throughput reads, CLIP-seq tags and Nascent-seq tags were normalized by genome-wide coverage of each dataset. Exonic N2a_CLIP tags arose from both nascent transcripts and mature mRNA. In contrast, exonic N2a_Nascent tags arose mostly from nascent transcripts and little from mature mRNA. We thus analyzed only intronic FUS-binding regions according to the RefSeq annotation (O'Leary et al., 2016), and further applied the following conditions. First, N2a_CLIP coverage ≥ 1 throughout an uninterrupted region. Second, N2a_Nascent coverage is ≥ 10 throughout an uninterrupted region, and spans ≥ 30 nt. Third, the region is within a MACS peak. We thus extracted 12,838 N2a_CLIP regions. Norm_N2a_CLIP_coverage within a region was summed up to calculate the N2a_CLIP_area. The CLIP-seq regions were sorted in descending order of the N2a_CLIP_area. The top 2,000 regions were defined as Nascent-seq-normalized CLIP regions. N2a_CLIP coverage was similarly normalized by N2a_RNA coverage. Following the same procedures for Nascent-seq, 2,000 RNA-seq-normalized CLIP regions were extracted from 5,890 N2a_CLIP regions, which were normalized by RNA-seq. The top 2,000 regions with high CLIP-seq tag coverage (spanning ≥ 30-nt) were defined as the native CLIP regions.

Putative motifs in Nascent-seq-normalized CLIP regions, RNA-seq-normalized CLIP regions, and native CLIP regions were detected by MEME-ChIP version 4.9.1 (Machanick and Bailey, 2011) with "meme-chip -oc output_directory/ -norc -meme-mod anr -meme-minw 5 -meme-maxw 10 input.fa". According to the web site of MEME-ChIP (http://meme-suite.org/doc/meme-chip.html), MEME-ChIP can be used for CLIP-seq

analysis to search for motifs.

2.6. Prediction of a structural motif of Nascent-seq-normalized CLIP regions by GraphProt

To train the model of FUS-binding motifs by GraphProt version 1.1.4 (Maticzka et al., 2014), we randomly extracted 1,000 Nascent-seq-normalized CLIP regions as a positive training dataset, and 1,000 intronic sequences of 30-100 nt, where FUS CLIP-seq tags were never mapped, as a negative training dataset. We then predicted FUS-binding motifs using a testing dataset, which was comprised of the remaining 1,000 Nascent-seq-normalized CLIP regions. For the execution of GraphProt, 100-nt sequences were added on both ends of CLIP regions for positive and negative training datasets, as well as the testing dataset.

2.7. Calculation of Youden's index to evaluate the presence of extracted motifs in the center of cross-linking-induced mutation sites (CIMS)

First, we calculated frequencies of all possible 4-nt motifs in Nascent-seq-normalized CLIP regions, RNA-seq-normalized CLIP regions, and native CLIP regions. For Nascent-seq-normalized CLIP regions, we extracted 17 4-nt motifs with frequency > 0.8%. Second, to identify cross-linking-induced mutation sites (CIMS) of N2a_CLIP, we generated a VCF file from a BAM file by SAMtools. The VCF file was converted into a BED file by BEDOPS version 2.4.16 (Neph et al., 2012). The BED file was used for CIMS software package version 1.0.5 [analysis steps 96-106 in (Moore et al., 2014)]. We obtained 1,974 CIMS. CIMS ± 100 nt regions were excised from the mm9 mouse genome sequence. Third, we generated all possible combinations of k 4-nt motifs (k = 1 to 10) out of the 17 4-nt motifs with frequency > 0.8%. For all the combinations, the number of motifs in the center of CIMS (CIMS ± 10 nt) was compared to the number of motifs out of CIMS (CIMS – 50 nt ± 10 nt). We then calculated the sensitivity and specificity of each combination of k motifs (k = 1 to 10). For each k, the best combination of motifs with the highest Youden's index (= sensitivity + specificity -1) in ${}_{17}C_k$ combinations was selected. Similar calculations were executed using RNA-seq-normalized CLIP regions and native CLIP regions. For the purpose of validation of the extracted motifs from N2a_CLIP, CIMS of cerebrum_CLIP and brain_CLIP were similarly identified.

3. Results

3.1. FUS-binding motifs extracted from Nascent-seq-normalized CLIP regions, RNA-seq-normalized CLIP regions, and native CLIP regions

We tried to identify FUS-binding motifs using CLIP-seq (N2a_CLIP) and Nascent-seq (N2a_Nascent) of untreated N2a mouse neuronal cells that we reported previously (Masuda et al., 2015). As CLIP-seq coverage represents the number of RNA fragments bound to an RBP (Konig et al., 2012), CLIP-seq coverage is over-represented in highly expressed genes (Wang et al., 2015a). Nascent-seq analyzes unprocessed native transcripts, to which FUS binds immediately after transcription. We thus examined the association between the number of Na2_CLIP tags and the number of Na2_Nascent tags at the gene level. Scattered plot of 7,921 RefSeq coding genes with the reads per kilobase per million mapped reads (RPKM) ≥ 1 for both N2a_CLIP and N2a_Nascent showed a correlation coefficient of 0.574 (Fig. 1). As predicted, N2a_CLIP tags were more abundant in highly expressed genes.

After FUS is bound to nascent RNA, exonic FUS may stay on mature mRNA, whereas intronic FUS is removed from mature mRNA. As exonic N2a_CLIP tags are likely to be over-presented compared to intronic N2a_CLIP tags, we eliminated exonic

N2a_CLIP tags from normalization of N2a_CLIP coverage with N2a_Nascent coverage. Using the Na2_Nascent_normalized N2a_CLIP data, we first restricted our analysis to genomic regions where N2a_Nascent coverage was ≥ 10 (average of coverage = 31.44) in an uninterrupted stretch ≥ 30-nt in introns of RefSeq coding genes. Normalization of CLIP-seq by Nascent-seq yielded 12,838 CLIP regions. We then extracted the top 2,000 CLIP regions with the highest coverages to generate Nascent-seq-normalized CLIP regions.

To compare normalization by Nascent-seq and RNA-seq, we similarly generated RNA-seq-normalized CLIP regions. Normalization of CLIP-seq by RNA-seq yielded 5,890 CLIP regions. We then extracted the top 2,000 CLIP regions with the highest coverages. We also made native CLIP regions, where the top 2,000 CLIP regions with the highest coverages without normalization were located.

First, we searched for FUS-binding motifs on Nascent-seq-normalized CLIP regions, RNA-seq-normalized CLIP regions, and native CLIP regions with MEME-ChIP. The *E*-values of motifs detected in Nascent-seq-normalized CLIP regions were better than those detected in RNA-seq-normalized CLIP regions and native CLIP regions (Supplementary Fig. S1). Especially, motifs with *E*-value < 1×10^{-10} were detected only in Nascent-seq-normalized CLIP regions (Supplementary Fig. S2). We then predicted a structural motif of FUS using Nascent-seq-normalized CLIP regions with GraphProt (Maticzka et al., 2014). GraphProt predicted a GU-rich sequence motif (Supplementary Fig. S3A), as well as a structural motif predicting a stem (Supplementary Fig. S3BC). This was consistent with our previous study (Ishigaki et al., 2012).

We also analyzed single nucleotide frequencies of Nascent-seq-normalized CLIP regions, RNA-seq-normalized CLIP regions, and native CLIP regions. We found that G was abundant in Nascent-seq-normalized CLIP regions, whereas A was abundant in native CLIP regions (Supplementary Table S6). Analysis of dinucleotide frequencies revealed that UG, GU and GG dominated in Nascent-seq-normalized CLIP regions over native CLIP regions (Supplementary Table S7). In addition, UG was abundant in Nascent-seq-normalized CLIP regions compared to RNA-seq-normalized CLIP regions (Supplementary Table S7). Dominance of specific 4-nt motifs was conspicuous in descending order of Nascent-seq-normalized CLIP regions, RNA-seq-normalized CLIP regions, and native CLIP regions (Fig. 2).

We next detected 1,974 CIMS in N2a_CLIP. As an RBP-crosslinked site tends to have a mutant nucleotide in high throughput sequencing, CIMS points to the direct FUS-binding site. Assuming that a motif in CIMS ± 10 nt is true-positive and a motif out of CIMS is false-positive, we calculated the sensitivity and specificity of all possible 4-nt motifs by the binary classification test (Fig. 3). We first selected 17 4-nt motifs with frequency > 0.8% in Nascent-seq-normalized CLIP regions (Table 1). We made all possible combinations of *k* 4-nt motifs (*k* = 1 to 10) using the 17 4-nt motifs, and calculated the Youden's index with CIMS. The best combinations of motifs for each *k* are shown in Table 2. Among these, a combination of 6 4-nt motif combination gave rise to the best Youden's index (Table 2 and Fig. 4), and was better than previously reported individual motifs or 15 motifs annotated by SpliceAid 2 (Piva et al., 2012) (Fig. 4 and Supplementary Table S8). The combination of 6 4-nt motifs included GU-rich UGUG, CUGG, UGGU, GCUG, GUGG, and UUGG motifs, and was designated as $6GU_R$ motifs. We performed similar analysis with RNA-seq-normalized CLIP regions and native CLIP regions (Table 1). The best Youden's indices (0.274 and 0.268) of them were not as good as that (0.283) of $6GU_R$ motifs (Table 2).

3.2. Validation of $6GU_R$ motifs with cerebrum_CLIP and brain_CLIP

The $6GU_R$ motifs extracted from Nascent-seq-normalized CLIP regions were validated using CIMS ± 100 nt regions of cerebrum_CLIP and brain_CLIP. The

numbers of CIMS of cerebrum_CLIP and brain_CLIP were 5,642 and 14,530, respectively. The numbers of the 6GU$_R$ motifs were higher in CIMS than those out of CIMS in both cerebrum_CLIP (Fig. 5A) and brain_CLIP (Fig. 5B).

We also validated that a combination of 4-nt motifs is better than those of the other sized motifs. We similarly made combinations of 5- and 6-nt motifs that were comprised of 22 and 75 motifs, respectively (Supplementary Table S9), which yielded the best Youden's indices with N2a_CLIP according to similar analyses shown in Fig. 4. We found that the numbers of the combinations of 5- and 6-nt motifs in CIMS of cerebrum_CLIP (Supplementary Fig. S4A) and brain_CLIP (Supplementary Fig. S4B) were not as high as those of 6GU$_R$ motifs.

3.3. 6GU$_R$-Score to predict FUS-binding sites

To predict FUS-binding sites, the area comprised of 6GU$_R$ motifs was defined as the 6GU$_R$-Score. For example, a stretch of UGUGGUGU has the UGUG, GUGG, and UGGU motifs, and the 6GU$_R$-Score becomes 12. We compared the 6GU$_R$-Score in CIMS (CIMS ± 10 nt) with the 6GU$_R$-Score out of CIMS (CIMS – 50 nt ± 10 nt) of N2a_CLIP (Fig. 6). The 6GU$_R$-Score of 4 or less was set to 0, because most 4-nt stretches covered by a single 6GU$_R$ motif were likely to be artifacts. We applied the 6GU$_R$-Score to *Fus* (Supplementary Fig. S5) and *Mib1* (Supplementary Fig. S6), and found that FUS-binding sites detected by N2a_CLIP were predicted by 6GU$_R$-Score, but not by SpliceAid 2.

We next examined the position of predicted FUS-binding sites on gene structure. Genomic coordinates of four types of exons (5' UTR-containing exons, 3' UTR-containing exons, coding exons, and alternatively spliced cassette exons) and introns were obtained from UCSC Table Browser (Speir et al., 2016) in a BED format. The predicted FUS-binding sites were most prevalent in introns (Supplementary Fig. S7 and Supplementary Table S10), which was consistent with our previous analyses of positions of CLIP-seq tags (Ishigaki et al., 2012; Masuda et al., 2015).

We also looked into the position of predicted FUS-binding sites around the polyadenylation signal (PAS) sites. We used 46,081 PAS sites obtained from PolyA-seq in our previous report (Masuda et al., 2015), and counted the number of FUS-binding sites around PAS (Supplementary Fig. S8). Predicted FUS-binding sites made a peak downstream to PAS, which is consistent with our previous report showing that binding of FUS downstream to PAS enhances polyadenylation (Masuda et al., 2015).

4. Discussion

GU-rich motifs were more enriched in Nascent-seq-normalized CLIP regions than RNA-seq-normalized CLIP regions and native CLIP regions (Fig. 2, and Supplementary Figs. S1 and S2). Normalization by N2a_Nascent was likely to have suppressed dominance of motifs present in highly expressed genes. RBPs have degenerative binding motifs. Degeneracy, however, is highly variable from RBP to RBP. For example, RBFOX1 and RBFOX2 strictly recognize (U)GCAUG (Kuroyanagi, 2009), and CUGBP1 recognizes only a stretch of $(UG)_n$ (Masuda et al., 2012). In contrast, MBNL1, which has a binding motif of YGCY, binds to highly degenerative sequences and its binding sites cannot be readily predicted (Masuda et al., 2012). Similarly, the binding sites of NOVA1, which has a binding motif of YCAY, cannot be precisely predicted (Ule et al., 2006). An RBP with a stringent motif makes distinct clusters of CLIP-seq tags, whereas an RBP with a degenerative motif makes scattered and less distinct clusters of CLIP-seq tags. CLIP-seq studies by us (Ishigaki et al., 2012; Fujioka et al., 2013; Masuda et al., 2016) and others (Hoell et al., 2011; Lagier-Tourenne et al., 2012; Rogelj et al., 2012; Schwartz et al., 2012; Nakaya et al., 2013; Zhou et al., 2013) indicate that FUS makes scattered clusters of CLIP-seq tags. Indeed, motif analysis

demonstrated highly degenerative motifs (Hoell et al., 2011; Ishigaki et al., 2012; Lagier-Tourenne et al., 2012; Rogelj et al., 2012; Fujioka et al., 2013; Masuda et al., 2016) or lack of a distinct motif (Schwartz et al., 2012; Nakaya et al., 2013; Zhou et al., 2013). In contrast to CLIP-seq, distinct motifs of GGUG and CGCGC were extracted with SELEX (Lerga et al., 2001) and RNAcompete (Ray et al., 2013), respectively. Lerga and colleagues additionally showed that one or two mutations in the GGUG motif are sufficient to loose FUS binding by electrophoretic mobility shift assay (EMSA) (Lerga et al., 2001). The GGUG and CGCGC motifs, however, gave rise to low Youden's indices with N2a_CLIP (Fig. 4). The discrepancy between *in cellulo* CLIP-seq analysis and *in vitro* SELEX/RNAcompete analyses is likely due to difference in the binding conditions (Reyes-Herrera and Ficarra, 2014). Although FUS tends to make scattered clusters of CLIP-seq tags on many genes, FUS has distinct binding sites on the other genes, as exemplified on *Fus* itself (Supplementary Fig. S5). Similarly, we previously reported that FUS regulates alternative splicing by position-specific binding to the downstream introns (Ishigaki et al., 2012). We also reported that binding of FUS to the upstream of a cryptic polyadenylation site (cPolyA) suppresses cPolyA, whereas binding of FUS to the downstream of a cPolyA enhances cPolyA by recruiting CPSF160 (Masuda et al., 2016). FUS is thus likely to recognize specific binding sites on some genes, but the underlying mechanisms remain undetermined. We have shown that a combination of six 4-nt FUS-binding GU-rich ($6GU_R$) motifs yields a better Youden's index than previously reported motifs [GGU (Rogelj et al., 2012), GGUG (Lerga et al., 2001), GUGGU (Lagier-Tourenne et al., 2012), and CGCGC (Ray et al., 2013)] or SpliceAid 2 (Supplementary Table S8) (Piva et al., 2012) (Fig. 4). The sensitivity and specificity of $6GU_R$ motifs, however, were 0.726 and 0.557, respectively (Table 2), which were not high enough to our satisfaction. Our analysis also disclosed that previously identified individual motifs and SpliceAid 2 are able to detect FUS-binding sites with a high specificity, although the sensitivities are low. Conversely, a combination of more than six motifs (Table 2) can be used to increase sensitivity at the cost of low specificity. We hope that $6GU_R$ motifs of FUS will help elucidate yet unidentified functions of FUS in physiological and pathological conditions.

Funding

This work was supported by Grants-in-Aid from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT); Ministry of Health, Labour and Welfare of Japan (MHLW); and Japan Agency for Medical Research and Development (AMED).

Acknowledgments

We are grateful to Prof. Yutaka Suzuki and Dr. Hiroyuki Wakaguri at University of Tokyo for providing local installation programs of DBTSS Genome Viewer.

References

Althammer, S., Gonzalez-Vallinas, J., Ballare, C., Beato, M. and Eyras, E., 2011. Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. Bioinformatics 27, 3333-40.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S. and Soboleva, A., 2013. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res 41, D991-5.

Bolger, A.M., Lohse, M. and Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-20.

Chen, B., Yun, J., Kim, M.S., Mendell, J.T. and Xie, Y., 2014. PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. Genome Biol 15, R18.

Da Cruz, S. and Cleveland, D.W., 2011. Understanding the role of TDP-43 and FUS/TLS in ALS and beyond. Curr Opin Neurobiol 21, 904-19.

Fredericks, A.M., Cygan, K.J., Brown, B.A. and Fairbrother, W.G., 2015. RNA-Binding Proteins: Splicing Factors and Disease. Biomolecules 5, 893-909.

Fujioka, Y., Ishigaki, S., Masuda, A., Iguchi, Y., Udagawa, T., Watanabe, H., Katsuno, M., Ohno, K. and Sobue, G., 2013. FUS-regulated region- and cell-type-specific transcriptome is associated with cell selectivity in ALS/FTLD. Sci Rep 3, 2388.

Han, A., Stoilov, P., Linares, A.J., Zhou, Y., Fu, X.D. and Black, D.L., 2014. De novo prediction of PTBP1 binding and splicing targets reveals unexpected features of its RNA recognition and function. PLoS Comput Biol 10, e1003442.

Hoell, J.I., Larsson, E., Runge, S., Nusbaum, J.D., Duggimpudi, S., Farazi, T.A., Hafner, M., Borkhardt, A., Sander, C. and Tuschl, T., 2011. RNA targets of wild-type and mutant FET family proteins. Nat Struct Mol Biol 18, 1428-31.

Iko, Y., Kodama, T.S., Kasai, N., Oyama, T., Morita, E.H., Muto, T., Okumura, M., Fujii, R., Takumi, T., Tate, S. and Morikawa, K., 2004. Domain architectures and characterization of an RNA-binding protein, TLS. J Biol Chem 279, 44834-40.

Ishigaki, S., Masuda, A., Fujioka, Y., Iguchi, Y., Katsuno, M., Shibata, A., Urano, F., Sobue, G. and Ohno, K., 2012. Position-dependent FUS-RNA interactions regulate alternative splicing events and transcriptions. Sci Rep 2, 529.

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M. and Zavolan, M., 2011. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nat Methods 8, 559-64.

Konig, J., Zarnack, K., Luscombe, N.M. and Ule, J., 2012. Protein-RNA interactions: new genomic technologies and perspectives. Nat Rev Genet 13, 77-83.

Kuroyanagi, H., 2009. Fox-1 family of RNA-binding proteins. Cell Mol Life Sci 66, 3895-907.

Lagier-Tourenne, C., Polymenidou, M., Hutt, K.R., Vu, A.Q., Baughn, M., Huelga, S.C., Clutario, K.M., Ling, S.C., Liang, T.Y., Mazur, C., Wancewicz, E., Kim, A.S., Watt, A., Freier, S., Hicks, G.G., Donohue, J.P., Shiue, L., Bennett, C.F., Ravits, J., Cleveland, D.W. and Yeo, G.W., 2012. Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. Nat Neurosci 15, 1488-97.

Lerga, A., Hallier, M., Delva, L., Orvain, C., Gallais, I., Marie, J. and Moreau-Gachelin, F., 2001. Identification of an RNA binding specificity for the potential splicing factor TLS. J Biol Chem 276, 6807-16.

Li, H. and Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589-95.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-9.

Li, X., Kazan, H., Lipshitz, H.D. and Morris, Q.D., 2014. Finding the target sites of RNA-binding proteins. Wiley Interdiscip Rev RNA 5, 111-30.

Machanick, P. and Bailey, T.L., 2011. MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics 27, 1696-7.

Mashima, J., Kodama, Y., Kosuge, T., Fujisawa, T., Katayama, T., Nagasaki, H., Okuda, Y., Kaminuma, E., Ogasawara, O., Okubo, K., Nakamura, Y. and Takagi, T., 2016. DNA data bank of Japan (DDBJ) progress report. Nucleic Acids Res 44, D51-7.

Masuda, A., Andersen, H.S., Doktor, T.K., Okamoto, T., Ito, M., Andresen, B.S. and Ohno, K., 2012. CUGBP1 and MBNL1 preferentially bind to 3' UTRs and facilitate mRNA decay. Sci Rep 2, 209.

Masuda, A., Takeda, J. and Ohno, K., 2016. FUS-mediated regulation of alternative RNA processing in neurons: insights from global transcriptome analysis. Wiley Interdiscip Rev RNA 7, 330-40.

Masuda, A., Takeda, J., Okuno, T., Okamoto, T., Ohkawara, B., Ito, M., Ishigaki, S., Sobue, G. and Ohno, K., 2015. Position-specific binding of FUS to nascent RNA regulates mRNA length. Genes Dev 29, 1045-57.

Maticzka, D., Lange, S.J., Costa, F. and Backofen, R., 2014. GraphProt: modeling binding preferences of RNA-binding proteins. Genome Biol 15, R17.

Menet, J.S., Rodriguez, J., Abruzzi, K.C. and Rosbash, M., 2012. Nascent-Seq reveals novel features of mouse circadian transcriptional regulation. Elife 1, e00011.

Moore, M.J., Zhang, C., Gantman, E.C., Mele, A., Darnell, J.C. and Darnell, R.B., 2014. Mapping Argonaute and conventional RNA-binding protein interactions with RNA

at single-nucleotide resolution using HITS-CLIP and CIMS analysis. Nat Protoc 9, 263-93.

Nakaya, T., Alexiou, P., Maragkakis, M., Chang, A. and Mourelatos, Z., 2013. FUS regulates genes coding for RNA-binding proteins in neurons by binding to their highly conserved introns. RNA 19, 498-509.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R. and Stamatoyannopoulos, J.A., 2012. BEDOPS: high-performance genomic feature operations. Bioinformatics 28, 1919-20.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D. and Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44, D733-45.

Piva, F., Giulietti, M., Burini, A.B. and Principato, G., 2012. SpliceAid 2: a database of human splicing factors expression data and RNA target motifs. Hum Mutat 33, 81-5.

Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., Na, H., Irimia, M., Matzat, L.H., Dale, R.K., Smith, S.A., Yarosh, C.A., Kelly, S.M., Nabet, B., Mecenas, D., Li, W., Laishram, R.S., Qiao, M., Lipshitz, H.D., Piano, F., Corbett, A.H., Carstens, R.P., Frey, B.J., Anderson, R.A., Lynch, K.W., Penalva, L.O., Lei, E.P., Fraser, A.G., Blencowe, B.J., Morris, Q.D. and Hughes, T.R., 2013. A compendium of RNA-binding motifs for decoding gene regulation. Nature 499, 172-7.

Reyes-Herrera, P.H. and Ficarra, E., 2014. Computational Methods for CLIP-seq Data Processing. Bioinform Biol Insights 8, 199-207.

Rogelj, B., Easton, L.E., Bogu, G.K., Stanton, L.W., Rot, G., Curk, T., Zupan, B., Sugimoto, Y., Modic, M., Haberman, N., Tollervey, J., Fujii, R., Takumi, T., Shaw, C.E. and Ule, J., 2012. Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. Sci Rep 2, 603.

Schwartz, J.C., Ebmeier, C.C., Podell, E.R., Heimiller, J., Taatjes, D.J. and Cech, T.R., 2012.

FUS binds the CTD of RNA polymerase II and regulates its phosphorylation at Ser2. Genes Dev 26, 2690-5.

Speir, M.L., Zweig, A.S., Rosenbloom, K.R., Raney, B.J., Paten, B., Nejad, P., Lee, B.T., Learned, K., Karolchik, D., Hinrichs, A.S., Heitner, S., Harte, R.A., Haeussler, M., Guruvadoo, L., Fujita, P.A., Eisenhart, C., Diekhans, M., Clawson, H., Casper, J., Barber, G.P., Haussler, D., Kuhn, R.M. and Kent, W.J., 2016. The UCSC Genome Browser database: 2016 update. Nucleic Acids Res 44, D717-25.

Suzuki, A., Wakaguri, H., Yamashita, R., Kawano, S., Tsuchihara, K., Sugano, S., Suzuki, Y. and Nakai, K., 2015. DBTSS as an integrative platform for transcriptome, epigenome and genome sequence variation data. Nucleic Acids Res 43, D87-91.

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B.J. and Darnell, R.B., 2006. An RNA map predicting Nova-dependent splicing regulation. Nature 444, 580-6.

Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., Stanton, R., Rigo, F., Guttman, M. and Yeo, G.W., 2016. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods 13, 508-14.

Wang, T., Xiao, G., Chu, Y., Zhang, M.Q., Corey, D.R. and Xie, Y., 2015a. Design and bioinformatics analysis of genome-wide CLIP experiments. Nucleic Acids Res 43, 5263-74.

Wang, X., Schwartz, J.C. and Cech, T.R., 2015b. Nucleic acid-binding specificity of human FUS protein. Nucleic Acids Res 43, 7535-43.

Zhang, C. and Darnell, R.B., 2011. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nat Biotechnol 29, 607-14.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W. and Liu, X.S., 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biol 9, R137.

Zhou, Y., Liu, S., Liu, G., Ozturk, A. and Hicks, G.G., 2013. ALS-associated FUS mutations result in compromised FUS alternative splicing and autoregulation. PLoS Genet 9, e1003895.

Figure legends

Fig. 1. Correlation of the numbers of FUS CLIP-seq tags and Nascent-seq tags at the gene level in N2a cells. The numbers of tags are indicated in reads per kilobase per million mapped reads (RPKM) for each gene. In the 7,921 RefSeq coding genes, genes with RPKM ≥ 1 in both CLIP-seq and Nascent-seq are plotted. As calculation of RPKM in a sliding window of a fixed size yielded a large number of datasets, we calculated RPKM at the gene level for both CLIP-seq tags and Nascent-seq tags. The linear regression line is RPKM_Nascent-seq = 0.607 × RPKM_CLIP-seq + 1.583. Pearson's correlation coefficient = 0.574.

Fig. 2. Comparison of frequencies of all possible 4-nt motifs in Nascent-seq-normalized CLIP regions, RNA-seq-normalized CLIP regions, and native CLIP regions. (A) The horizontal axis indicates 256 4-nt motifs in descending order of frequency. **(B)** The top-ranked 17 4-nt motifs of Nascent-seq-normalized CLIP regions in **(A)**.

Fig. 3. Sensitivity and specificity of a specific motif is calculated using the number of motifs on 1,974 CIMS ± 100 nt segments of N2a_CLIP. (A) An example of UGUG motif. The number of UGUG motifs is plotted from CIMS – 100 nt to CIMS + 100 nt. **(B)** Matrix presentation of the number of motifs to calculate sensitivity and specificity. The number of CIMS ± 100 nt segments with a specific motif within CIMS ± 10 nt [flanked by solid lines in **(A)**] is defined as true-positive (TP). The number of CIMS ± 100 nt segments with a specific motif within CIMS – 50 nt ± 10 nt [flanked by dotted lines in **(A)**] is defined as false-positive (FP). False-negative, FN, and true-negative, TN, are similarly defined for CIMS ± 10 nt and CIMS – 50 nt ± 10 nt without a specific motif, respectively. Sensitivity, specificity, and Youden's index are calculated using the indicated equations.

Fig. 4. Receiver operating characteristic (ROC) plots (A) and Youden's indices (B) of combinations of the best 1 to 10 motifs (Table 2), as well as previously reported motifs. Among them, a combination of 6 4-nt motifs (6GU$_R$ motifs) yielded the best Youden's index of 0.283. For SpliceAid 2, a combination of 15 motifs is indicated.

Fig. 5. Validation of the identified FUS-binding 6GU$_R$ motifs using cerebrum_CLIP and brain_CLIP. The number of 6GU$_R$ motifs are plotted for 5,642 CIMS ± 100 nt segments of cerebrum_CLIP **(A)** and for 14,530 CIMS ± 100 nt segments of brain_CLIP **(B)**. Using Kruskal-Wallis rank sum test and Steel-Dwass post-hoc test with Bonferroni correction (R version 3.2.1), we calculated *p*-values to evaluate the difference in the number of 6GU$_R$ motifs in CIMS (a single position) and that out of CIMS (each point of 200 positions). The 200 *p*-values were all less than 1 x 10^{-12} for both the cerebrum_CLIP and brain_CLIP (not indicated in figure).

Fig. 6. Difference in 6GU$_R$-Score spanning and outside of CIMS of N2a_CLIP. The 6GU$_R$-Scores spanning CIMS (CIMS ± 10 nt) are counted, and the ratio of each bin (0 to 32) of 6GU$_R$-Score is calculated. Similarly, the 6GU$_R$-Score out of CIMS (CIMS – 50 nt ± 10 nt) are counted, and the ratio in each bin (0 to 32) of 6GU$_R$-Score is calculated. For each bin (0 to 32), the ratio of 6GU$_R$-Score out of CIMS is subtracted from that spanning CIMS. The difference is plotted for each bin. A positive value in the difference indicates that the specific 6GU$_R$-Score is enriched in CIMS.

Figure 1

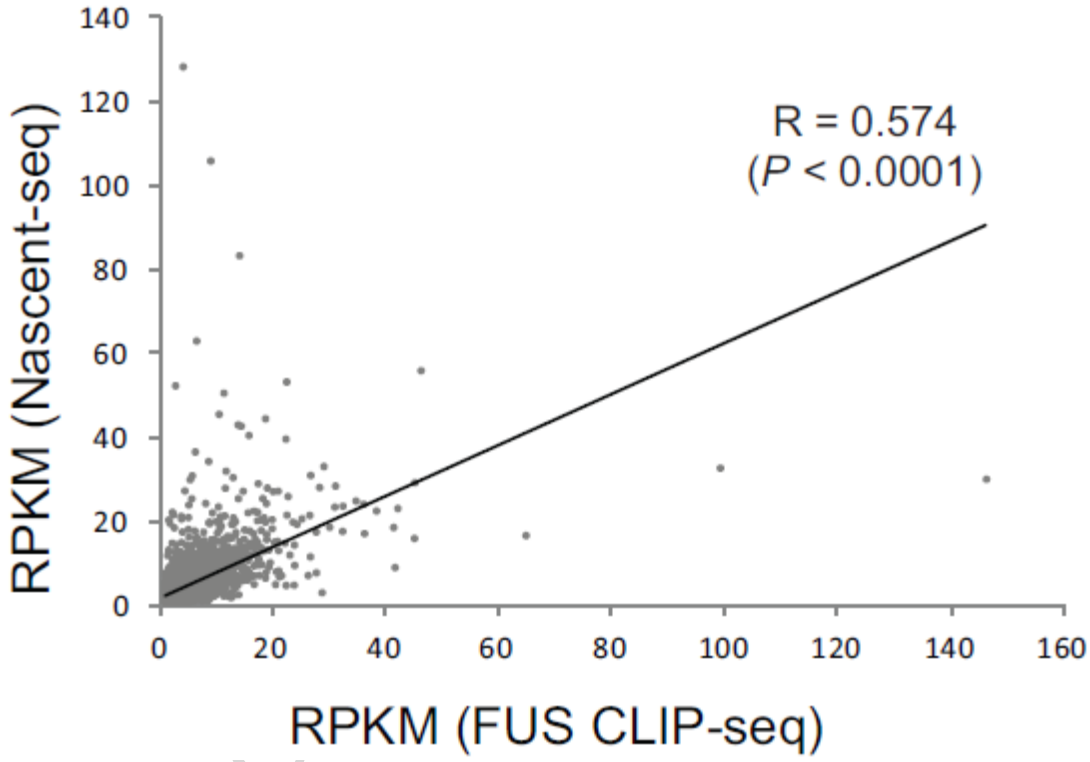


Figure2

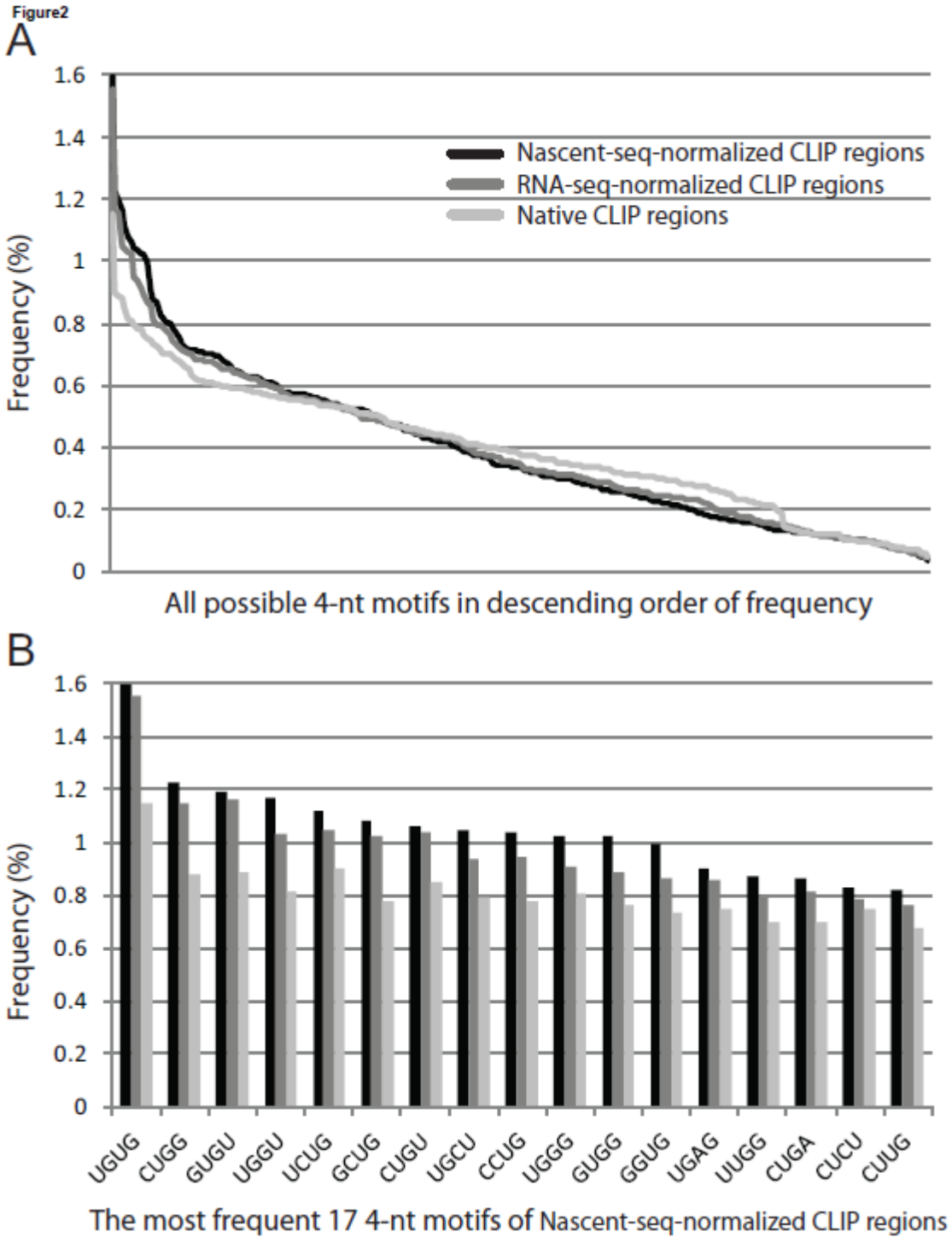
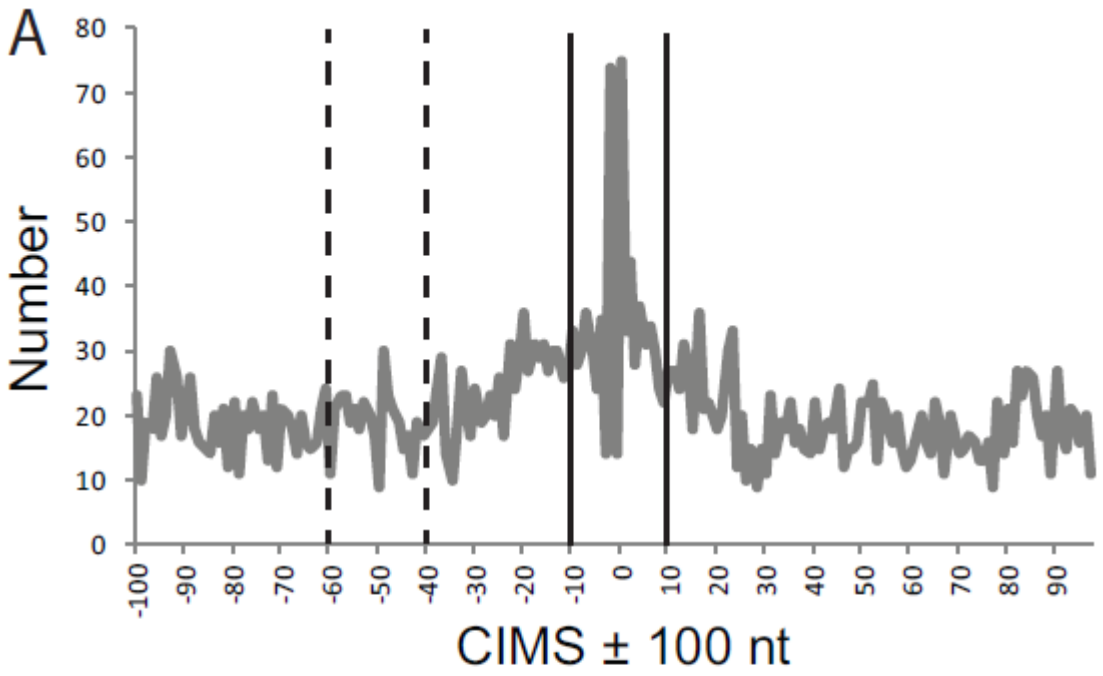


Figure 3

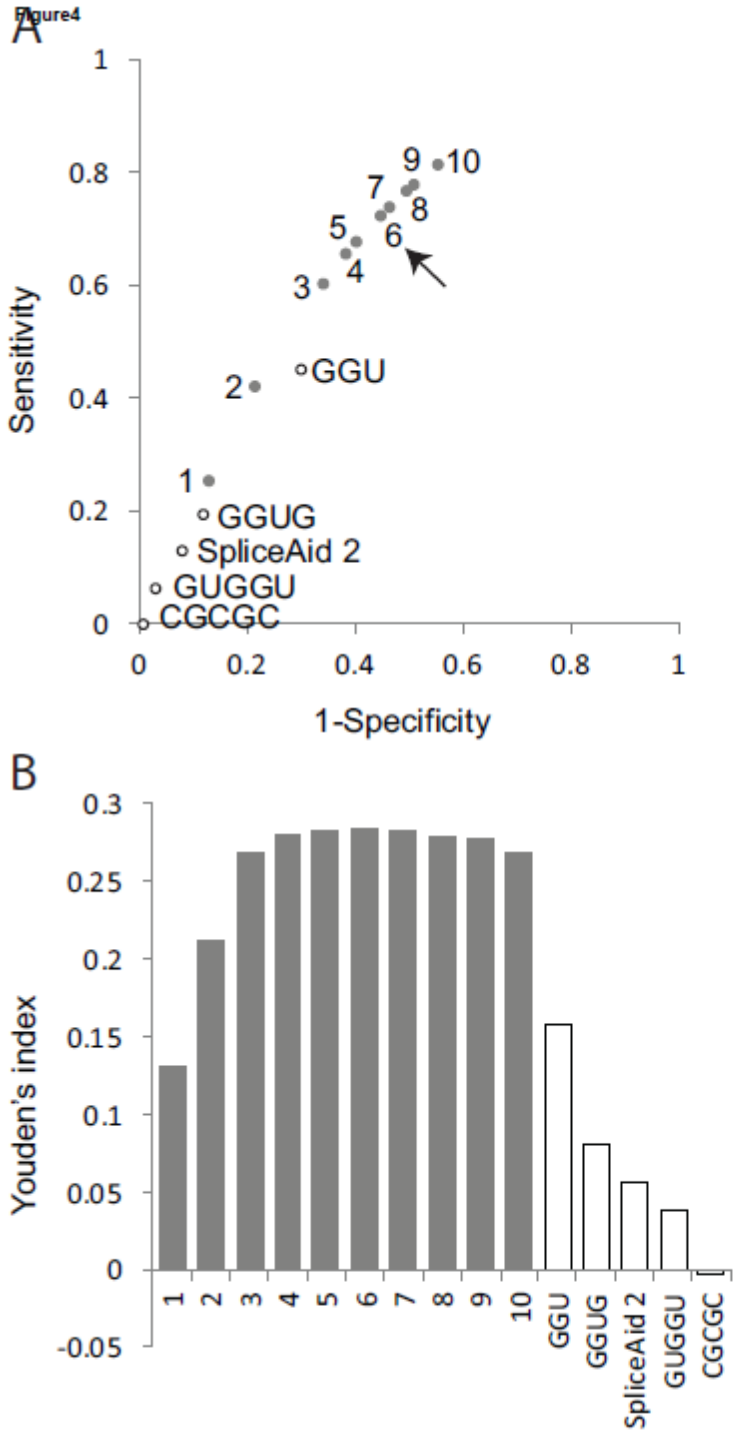


B

	CIMS ± 10	CIMS-50 ± 10
# of seg. with a motif	TP	FP
# of seg. without a motif	FN	TN

Sensitivity = TP/(TP + FN)
Specificity = TN/(TN + FP)
Youden's index = Sensitivity + Specificity -1

Figure4



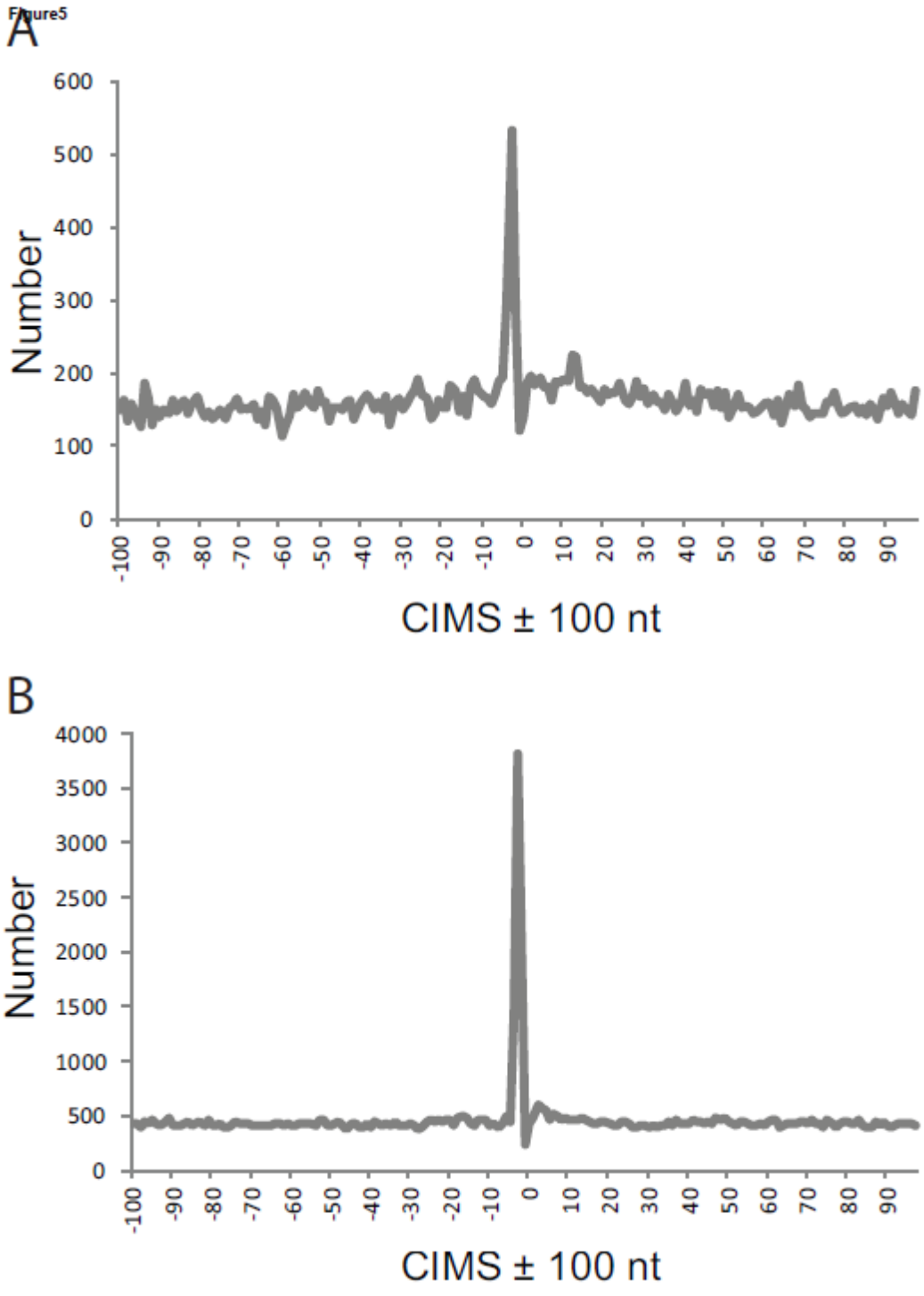


Figure 6

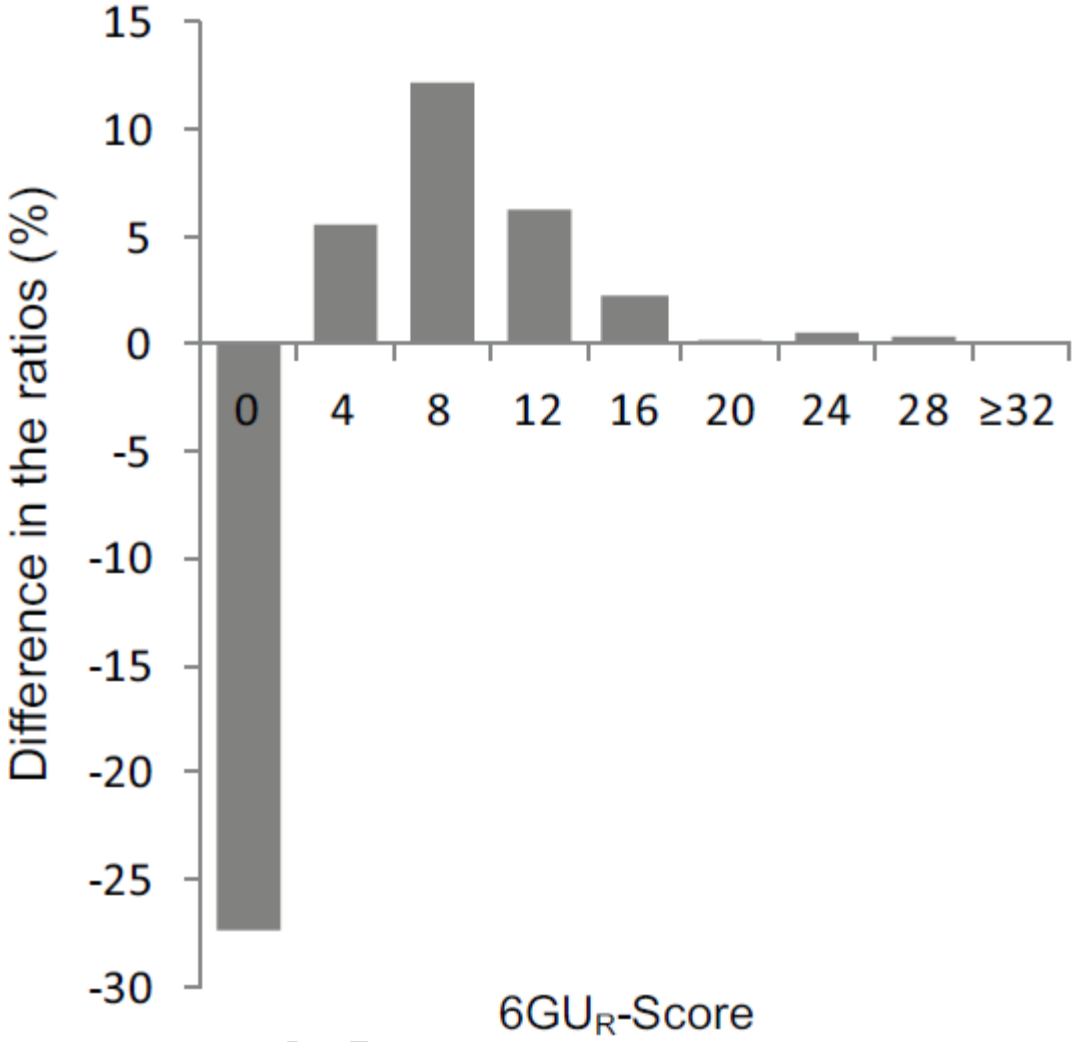


Table 1. Motifs (4-nt) with frequency > 0.8% in the Nascent-seq-normalized CLIP regions, RNA-seq-normalized CLIP regions, and native CLIP regions

A. Nascent-seq-normalized CLIP regions

Motif	Frequency (%)
UGUG	1.598
CUGG	1.228
GUGU	1.194
UGGU	1.168
UCUG	1.118
GCUG	1.080
CUGU	1.060
UGCU	1.044
CCUG	1.037
UGUG	1.037
UGGG	1.026
GUGG	1.026
GGUG	1.000
UGAG	0.902
UUGG	0.873
CUGA	0.867
CUCU	0.828

B. RNA-seq-normalized CLIP regions

Motif	Frequency (%)
UGUG	1.557
GUGU	1.165
CUGG	1.150
UCUG	1.046
CUGU	1.043
UGGU	1.031
GCUG	1.023
CCUG	0.948
UGCU	0.937

UGGG	0.911
GUGG	0.891
GGUG	0.868
UGAG	0.856
CUGA	0.815

C. Native CLIP regions

Motif	Frequency (%)
UGUG	1.152
UCUG	0.899
GUGU	0.888
CUGG	0.879
CUGU	0.852
UGGU	0.813
UGGG	0.805

Table 2. Combinations of 4-nt motifs with the best Youden's indices

A. Nascent-seq-normalized CLIP regions

Combinations	Motifs	1 - Specificity	Sensitivity	Youden's index
1	CUGG	0.125	0.256	0.131
2	CUGG, UUGG	0.210	0.424	0.213
3	UGUG, CUGG, UUGG	0.337	0.605	0.268
4	UGUG, CUGG, GUGG, UUGG	0.378	0.659	0.280
5	UGUG, CUGG, UGGU, GUGG, UUGG	0.398	0.680	0.282
6	UGUG, CUGG, UGGU, GCUG, GUGG, UUGG	0.443	0.726	0.283
7	UGUG, CUGG, UGGU, GCUG, UGGG, GUGG, UUGG	0.459	0.741	0.282
8	UGUG, CUGG, GUGU, UGGU, GCUG, GUGG, GGUG, UUGG	0.490	0.770	0.280
9	UGUG, CUGG, GUGU, UGGU, GCUG, UGGG, GUGG, GGUG, UUGG	0.504	0.781	0.277
10	UGUG, CUGG, GUGU, UGGU, GCUG, CCUG, UGGG, GUGG, GGUG, UUGG	0.549	0.817	0.268

B. RNA-seq-normalized CLIP regions

Combinations	Motifs	1 - Specificity	Sensitivity	Youden's index
1	CUGG	0.125	0.256	0.131
2	UGUG, CUGG	0.263	0.477	0.213
3	UGUG, CUGG, UGGU	0.319	0.570	0.251
4	UGUG, CUGG, UGGU, UGGG	0.368	0.636	0.268
5	UGUG, CUGG, UGGU, UGGG, GUGG	0.384	0.657	0.273
6	UGUG, CUGG, UGGU, GCUG, UGGG, GUGG	0.432	0.705	0.274
7	UGUG, CUGG, UGGU, GCUG, CCUG, UGGG, GUGG	0.484	0.755	0.271
8	UGUG, GUGU, CUGG, UGGU, GCUG, UGGG, GUGG, GGUG	0.480	0.751	0.271
9	UGUG, GUGU, CUGG, UGGU, GCUG, CCUG, UGGG, GUGG, GGUG	0.526	0.794	0.268
10	UGUG, GUGU, CUGG, UGGU, GCUG, CCUG, UGGG, GUGG, GGUG, UGAG	0.571	0.832	0.261

C. Native CLIP regions

Combinations	Motifs	1 - Specificity	Sensitivity	Youden's index
1	CUGG	0.125	0.256	0.131
2	UGUG, CUGG	0.263	0.477	0.213
3	UGUG, CUGG, UGGU	0.319	0.570	0.251
4	UGUG, CUGG, UGGU, UGGG	0.368	0.636	0.268
5	UGUG, GUGU, CUGG, UGGU, UGGG	0.401	0.664	0.263
6	UGUG, UCUG, GUGU, CUGG, UGGU, UGGG	0.473	0.719	0.247
7	UGUG, UCUG, GUGU, CUGG, CUGU, UGGU, UGGG	0.519	0.747	0.228

Youden's index = sensitivity + specificity -1

Highlights

- As CLIP-seq tags are enriched in highly expressed genes, we normalized CLIP-seq tags of N2a cells by Nascent-seq tags of N2a cells.
- We extracted frequently observed 4-nt motifs from the normalized CLIP-seq regions.
- Six GU-rich motifs of UGUG, CUGG, UGGU, GCUG, GUGG and UUGG ($6GU_R$ motifs) were enriched in FUS-binding sites of CLIP-seq.
- We propose that an area covered by $6GU_R$ motifs (the $6GU_R$-Score) of 8 or more efficiently predicts FUS-binding sites.