Supporting Information

Yao et al. 10.1073/pnas.1211101109

SI Materials and Methods

Cell Culture and Transfection. HeLa cells were grown in DMEM plus 10% FBS. For CstF64 RNAi, a pSuperior.puro plasmid was constructed to express shRNAs targeting CstF64 mRNA (target sequence: GTTAGATGCCAGAGGATTA). Transfections were carried out using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. Stable CstF64 RNAi cell lines were obtained by selection with puromycin and expansion of single colonies. To knock down CstF64τ in CstF64 RNAi cell lines, predesigned siRNAs (Ambion s23471) were transfected into a stable CstF64 RNAi cell line using Lipofectamine 2000. Knockdown efficiencies were determined by Western blot analysis using antibodies against CstF64 (mAb 6A9) and CstF64τ (Bethyl A301-487A).

Gel Shift Assay. RNA substrates were synthesized by T7 transcription in the presence of α-^{32}P UTP. RNAs (~1.5 nM) were incubated with 0–60 μM GST-CstF64-RRM fusion protein in 10.6 μL of binding buffer [10 mM Hepes (pH 7.9), 50 mM NaCl, 0.5 mM $MgCl_2$, 0.1 mM EDTA, 5% glycerol, 1 mM ATP, 10 mM creatine phosphate, 5 mM β-mercaptoethanol, 0.25 mM PMSF, 0.7 μg of *Escherichia coli* tRNA, and 1.4 μg of BSA) at 30 °C for 10 min. Reaction mixtures were resolved on 5% nondenaturing PAGE gels.

In Vitro Cleavage/Polyadenylation Assay. In vitro cleavage/polyadenylation assays were carried out as described previously (1). For the competition assays shown in Fig. 4*B*, 0–50 μM GST-CstF64-RRM fusion protein was added in the reaction.

Individual Nucleotide Resolution UV Crosslinking and Immunoprecipitation Sequencing. CstF64 individual nucleotide resolution UV crosslinking and immunoprecipitation sequencing (iCLIP-seq) was performed as described previously with minor modifications using a polyclonal antibody (Bethyl A301-092A) (2). Three replicate iCLIP libraries were prepared and sequenced using the Illumina HiSeq platform. A total of ~43 million reads that could be uniquely mapped to the human genome were obtained. By incorporating a random trinucleotide barcode, iCLIP-seq allows elimination of PCR artifacts and enables direct counting of cDNAs (2). Once reads that truncate at the same genomic location and have the same random nucleotide barcode were combined, ~33 million reads remained (~11 million reads for each replicate), each representing a uniquely crosslinked RNA. Because the nucleotide immediately upstream of the start of each iCLIP tag corresponds to a protein crosslinking site, only the positions of the crosslinking nucleotides were retained in the final mapping results. The height of each peak, termed the cDNA count, reflects the amount of CstF64 crosslinking detected at this position.

Luciferase Assays. HeLa cells were transfected with pPASPORT plasmids, and cells were harvested at 24 h posttransfection. Luciferase assays were performed using the Promega Dual-Luciferase Reporter Kit following the manufacturer's instructions.

Bioinformatic Analysis. ***iCLIP-seq data analyses. Filtering and mapping.*** Raw reads were demultiplexed using the sequencing barcode unique to each replicate, and an additional random trinucleotide identifying individual DNA molecules was clipped but kept as metadata. Reads with quality <20 for 10% or more of the bases were removed. The remaining reads were mapped to build human genome assembly 19 (hg19) using bowtie with parameters "bowtie -n 2 -m 1 -s 1" (up to two mismatch and only unique match to the genome allowed) (3). Mapped reads that truncated at the same sites and had the same trinucleotide barcodes were combined. After mapping, the base upstream of the 5′ end of each read was retained as the CLIP binding site, and the total number of reads sharing the same CLIP binding site on the same strand was used as the cDNA count at that position.

Data quality. To assess data quality, we calculated the fraction of sites that had a minimum cDNA count in at least two of the three replicates. We also compared pentamer frequencies among the three replicates. For each pentamer, we counted the number of cross-linking sites in each replicate with which the pentamer overlapped and compared the overall frequencies of each pentamer across replicates. We found excellent agreement among the replicates (R^2 = 0.9999, 0.9994, and 0.9994) (Fig. 1*C*). To determine the overall quality of our data signal and identify an acceptable threshold for considering a clip site a true positive, we estimated the false discovery rate (FDR) for our data as described previously (4). For each gene, we counted the number of reads aligning to it and randomly placed an equal number of reads along the gene's length 100 times. For a particular cDNA count, h (height), we calculated the FDR at that count as (%background height ≥ h)/(%foreground height ≥ h). We used a minimum cDNA count of three, which had an estimated FDR of 0.12%. We also required that each replicate be represented by at least one read at the retained binding sites.

Motif analysis. We first ranked all of the reproducible CstF64 crosslinking sites according to their cDNA counts. We then iteratively chose the crosslinking site with the highest cDNA count that did not overlap with the 21-nt region spanning a site that had been chosen previously. We analyzed a total of 36,859 nonoverlapping CLIP sites by checking for the presence of 6mer in the 21 bp surrounding all sites. For each binding site, we randomly sampled 100 21-bp windows from the similar regions (i.e., 5′ UTRs, exons, introns, 3′ UTRs, and intergenic regions). We used the mean and SD of the background site motif to calculate a *z*-score for motif enrichment. We further classified those CstF64 binding sites into two groups based on the existence of AWTAAA (representing AATAAA or ATTAAA) within the 40 nt upstream. We aligned the 20 most greatly enriched motifs using a previously published method (5), and generated sequence logos using WebLogo 3 (http://weblogo.threeplusone.com/) from their alignment. We used the same approach to perform motif analysis of the CstF64 CLIP$^+$ and CLIP$^-$ PASs.

Distribution within genes. We assessed the overlap of CLIP binding sites with different gene regions. Our hierarchical classification first checked for overlap with 3′UTRs, then with 5′ UTRs, then with coding exons, then with introns, and finally with noncoding genes. All sites that did not overlap one of these categories were considered intergenic. Noncoding genes were derived from four separate data sources accessible from (*i*) the University of California Santa Cruz (UCSC) Genome Browser's Refseq noncoding genes; (*ii*) the wgRNA table (6) consisting of C/D and H/ACA box snoRNAs, scaRNAs, and microRNAs; (*iii*) the lincRNATranscripts table (7), consisting of large intergenic noncoding RNAs; and (*iv*) tRNAs (8).

Conservation. We used the phyloP scores (9) from UCSC to summarize base-level conservation around CLIP binding sites (Figs. 3*A* and 4*A*). For each CLIP binding site, we determined the average conservation in a window surrounding the site and plotted the SEM as a gray envelope. Also for each CLIP binding site, we sampled the overlapping 3′UTR (Fig. 3*A*) or intron (Fig. 4*A*) 100 times to create a control distribution (gray line below).

Analysis of the relationship between CstF64 crosslinking and RNA sequence. To analyze the differences between CstF64 CLIP$^+$ and CLIP$^-$ PASs, we took 6,122 high-confidence PAS sites (i.e, with a read count >20) from our PAS-seq dataset and counted the total number of reproducible iCLIP reads in the 30-bp region downstream of the cleavage/polyadenylation site. We sorted all of the sites by iCLIP cDNA count and grouped the top 1,000 sites as CstF64 CLIP$^+$ PASs and the bottom 1,000 sites as CstF64 CLIP$^-$ PASs.

Analysis of direct RNA sequencing data. Sequencing and reads mapping. Direct RNA sequencing (DRS) was performed by Helicos BioSciences, and DRS reads were aligned to hg19 using the indexDPgenomic tool in Helisphere (Helicos BioSciences). The uniquely mapped reads with a minimum mapped length of 25 and an alignment score of 4.0 were kept for further analysis. We first filtered all mapped reads for those arising from internal poly(A) priming as described previously (10). We next identified individual poly(A) sites (PASs) by reversing 5′ ends of the non–internal-priming reads. To construct a consensus poly(A) annotation for downstream analysis, we used pooled data from both HeLa-Mock and CstF64-RNAi cells to iteratively cluster all individual PASs within 40 nt to its nearest PAS on the same chromosome strand. The weighted coordinate, calculated as the sum of the product of the coordinate of an individual poly(A) and its percentage of use in the whole cluster, was taken as the representative coordinate of the corresponding poly(A) cluster. The frequencies of poly(A) clusters in the different samples were calculated according to the above consensus coordinates of poly(A) clusters in the pooled data. Next, the poly(A)s residing in the whole gene region, including exons, introns, and the downstream 100-nt region of the terminal exon, were collected as possible poly(A)s of a certain gene [UCSC genes (hg19) and Ensembl genes (release 61)].

Alternative polyadenylation analysis. To compare the alternative polyadenylation (APA) profiles in HeLa and CstF64-RNAi cells or CstF64&τ-RNAi cells using DRS data, we first removed PASs that overlapped with snoRNA/scaRNA/snRNA regions and those that had none read in two of the three samples. For the remaining PASs, we used the Fisher exact test to compare the ratio of the DRS read counts of one PAS to the sum of the read counts of all of the other PASs within the same gene. The *P* values were adjusted by the Benjamini–Hochberg method for calculating the FDR. PASs with an FDR <0.05 were defined as significantly changed PASs. To create the scatterplot shown in Fig. 4*B*, we selected two PASs with the smallest *P* values for each gene with multiple PASs and calculated the corresponding proximal/distal ratio. In the figure, PAS pairs with an FDR <0.05 and log10(proximal/distal PAS read count) >0.2 are highlighted in red for proximal-to-distal switches and in blue for distal-to-proximal switches.

Comparing DRS and iCLIP-seq data. For the analysis shown in Fig. 4*D*, for all genes in the group, we first normalized the iCLIP signals detected within 200 nt downstream of both the proximal and distal sites, by dividing the cDNA counts at each position by the total cDNA counts within this 400-nt region for each gene. We then summed the normalized iCLIP signals at each position for all of the genes within each group, and plotted these values on the *y*-axis.

Primer sequences for all qRT-PCR and plasmid constructions are available on request.

1. Shi Y, et al. (2009) Molecular architecture of the human pre-mRNA 3′ processing complex. *Mol Cell* 33(3):365–376.
2. König J, et al. (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* 17(7):909–915.
3. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.
4. Yeo GW, et al. (2009) An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat Struct Mol Biol* 16(2):130–137.
5. Hu J, Lutz CS, Wilusz J, Tian B (2005) Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *RNA* 11(10): 1485–1493.
6. Griffiths-Jones S (2004) The microRNA Registry. *Nucleic Acids Res* 32(Database issue): D109–D111.
7. Cabili MN, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 25(18):1915–1927.
8. Lowe TM, Eddy SR (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25(5):955–964.
9. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 20(1):110–121.
10. Fu Y, et al. (2011) Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res* 21(5):741–747.

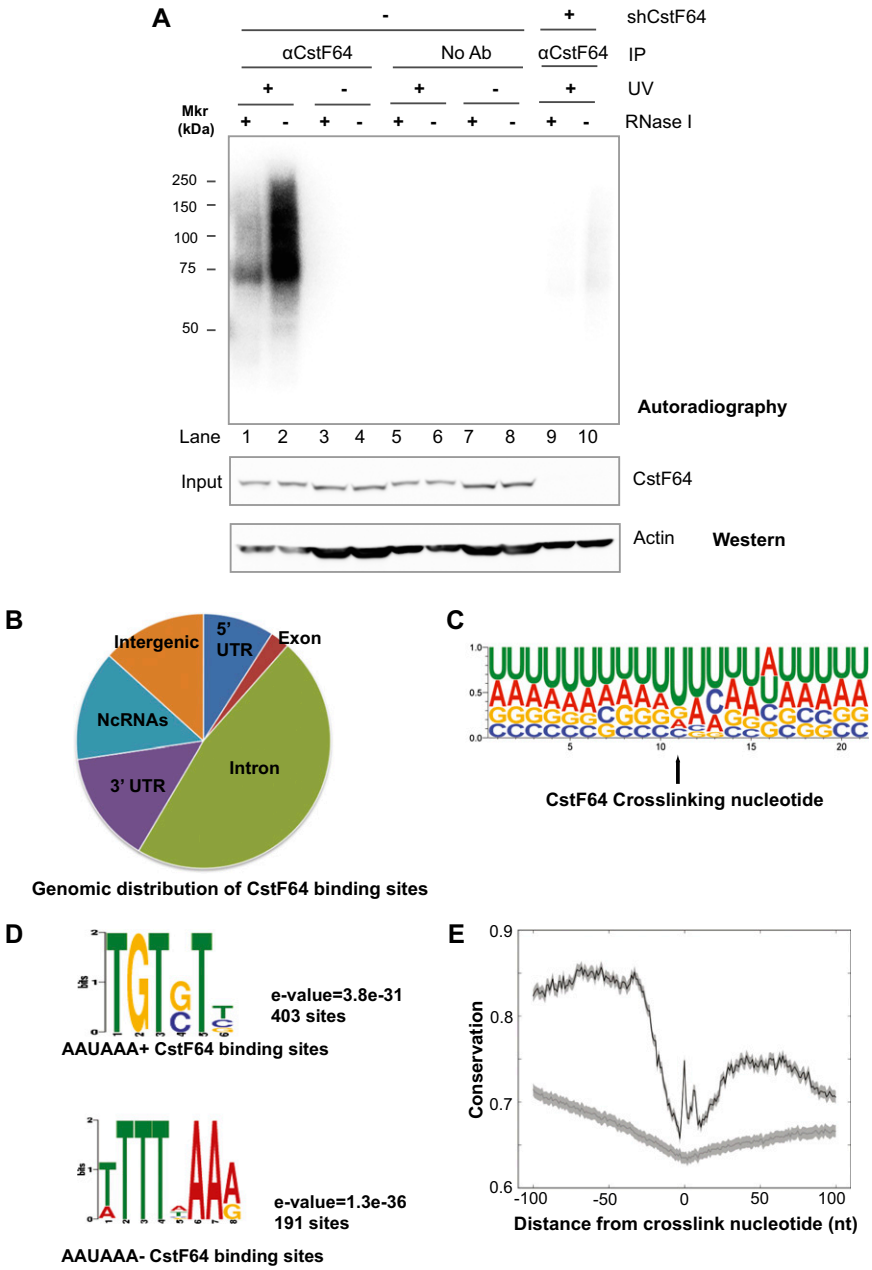
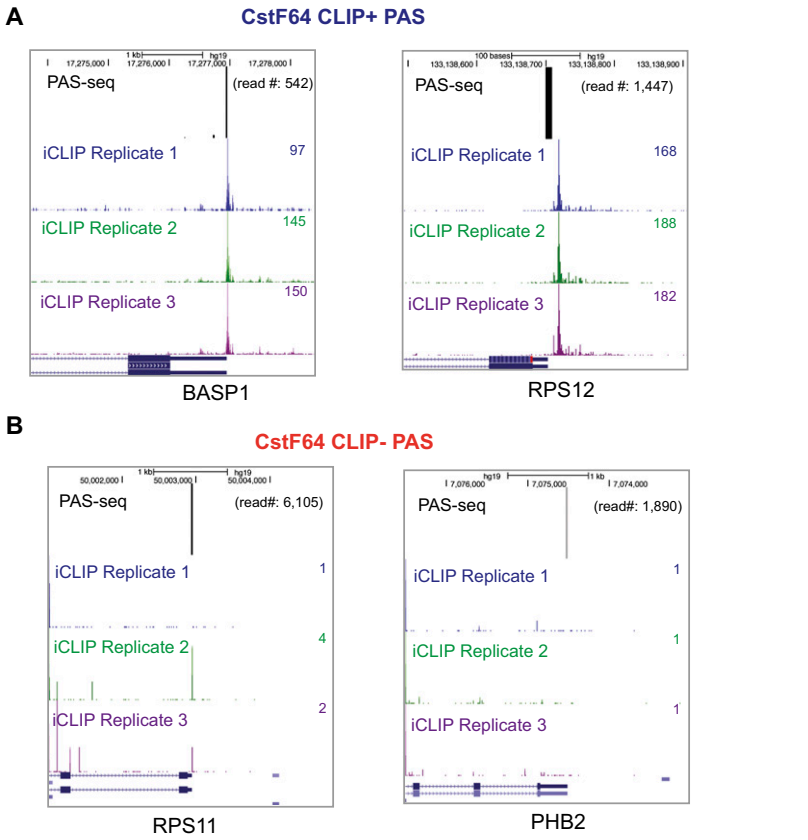


Fig. S1. Summary of CstF64 iCLIP-seq analysis. (*A*) Specificity of CstF64 iCLIP-seq. (*Upper*) Autoradiograph of the 5′ 32P-labeled RNA–protein complexes from immunoprecipitation using no antibody (No Ab) or anti-CstF64 antibodies (αCstF64) with cell lysates from control HeLa (–shCstF64) or a HeLa cell line stably expressing shRNAs targeting CstF64 mRNA (+shCstF64). (*Lower*) CstF64 or actin Western blots of cell lysates used in iCLIP experiments shown in the upper panel. (*B*) Pie chart of CstF64 binding site distribution in the genome. The 3′ UTRs shown are annotated 3′ UTRs in Refseq plus a 200-nt downstream sequence. (*C*) Sequence logo based on all nonoverlapping crosslinking nucleotides and 10 nt on either side. (*D*) Multiple Em for Motif Elicitation analysis of the top 1,000 AAUAAA^{+} and AAUAAA^{-} CstF64 binding sites. The most greatly enriched motifs from both groups are shown. (*E*) Conservation of CstF64 binding sites in 3′ UTRs and neighboring sequences. The *y*-axis is the average PhyloP conservation score of a nucleotide at a given distance from the crosslinking site. Random positions in the same 3′ UTRs were used as controls (lower line, labeled). The gray envelope represents the SEM.



C

BASP1: GUAAUAAAGACUAUUGGGUUUUUUGUUCUGUUUGUUUUACUUUUCUAGUCCCCUUCAAUAA
RPS12: UUCCUCAUGUCUGGCUUUUUAUUUGGGGCAGUAAAAUAAGGUCCCUGUUAGCAAAGUAAAA
RPS11: CAGGCCAGACUUGGGAUCUUCCGCGCCUUUACCAGAGCAUUGGUGGGGGUGGGGGUGUGCC
PHB2: GUCAAGGCUUCACGUGUUCAUGAACUGGGUAACUGGCAGCAAGCAUGCGCACGUUCACAUG

Fig. S2. PAS-seq and CstF64 iCLIP-seq results at CstF64 CLIP$^+$ and CLIP$^-$ PASs. (*A*) Two examples of CstF64 CLIP$^+$ PASs. (*B*) Two examples of CstF64 CLIP$^-$ PASs with data for PAS-seq and all three iCLIP-seq replicates. (*C*) Nucleotide sequences of the 60-nt fragment downstream of the cleavage sites (CSs) for all four PASs used for gel shift assays, as shown in Fig. 3*A*.

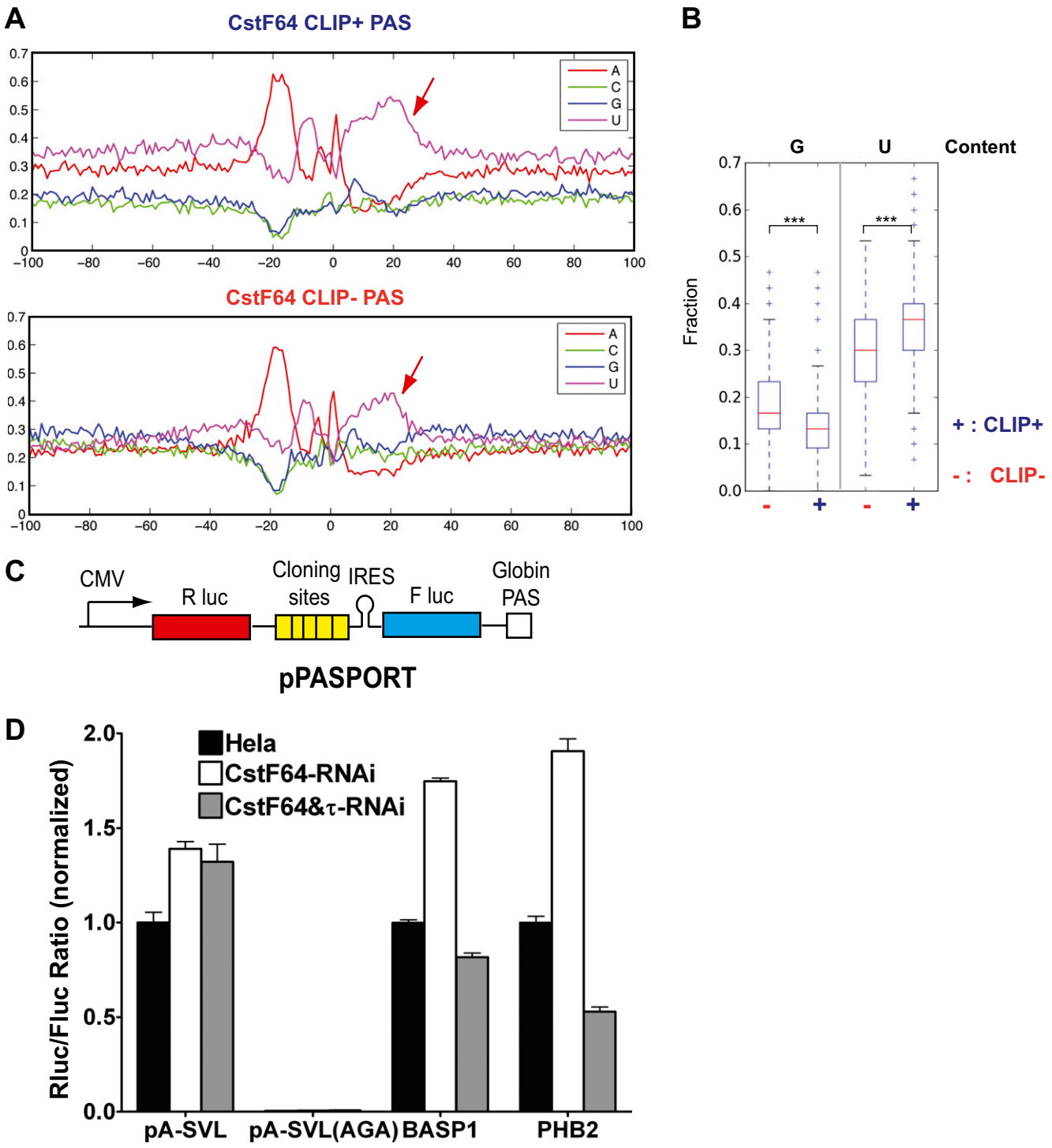


Fig. S3. Comparison of CstF64 CLIP$^+$ and CLIP$^-$ PASs. (*A*) Nucleotide composition for 1,000 CstF64 CLIP$^+$ PASs (*Upper*) and CLIP$^-$ PASs (*Lower*). The percentages of each nucleotide from 100 nt upstream to 100 nt downstream of the CSs (0 nt) are shown. Red arrows indicate the region (0–40 nt) in which CstF64–RNA interactions occur in CstF64 CLIP$^+$ PASs. (*B*) Comparison of G and U content in 1,000 CstF64 CLIP$^+$ and CLIP$^-$ PASs within the 0- to 40-nt region downstream of the CSs. ***P value <0.001. (*C*) Schematic of the pPASPORT reporter construct. *Renilla* (R luc) and firefly (F luc) luciferase genes are expressed in one bicistronic mRNA. An encephalomyocarditis virus internal ribosome entry site (IRES) upstream of the *Fluc* gene drives cap-independent translation of *Fluc*. Sequences to be tested are inserted into multiple cloning sites between the end of *Rluc* and the IRES. (*D*) Reporter assays with *SVL*, *SVL* AGA (*SVL* mutant with the AAUAAA hexamer mutated to AAGAAA), *BASP1*, and *PHB2* in control HeLa, CstF64-RNAi, and CstF64&τ-RNAi cells. Rluc/Fluc ratio values (*y*-axis) were normalized against those in control HeLa cells.

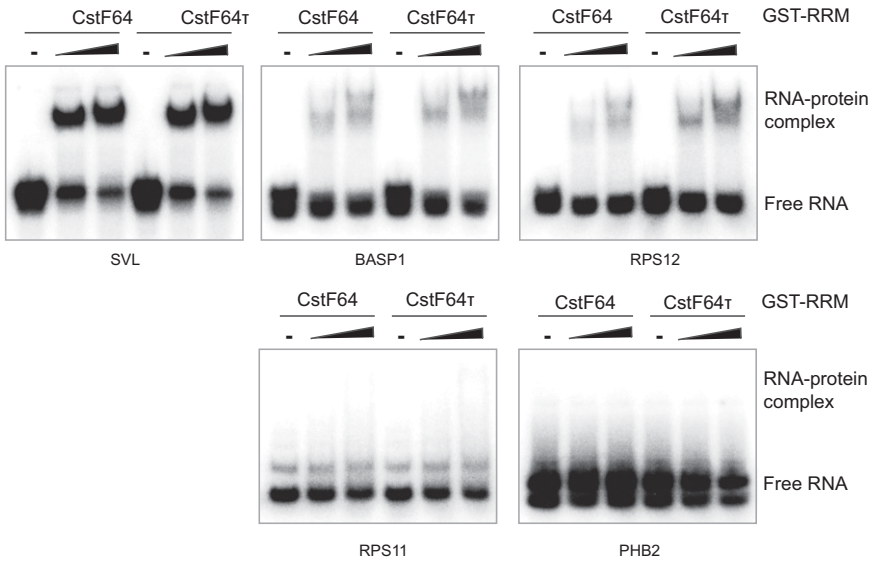


Fig. S4. Comparison of the RNA-binding specificities of CstF64 and CstF64τ. Gel shift assays using recombinant GST-CstF64-RRM or GST-CstF64τ and the specified RNA substrates are shown. Assay conditions are the same as described in Fig. 3*A*.

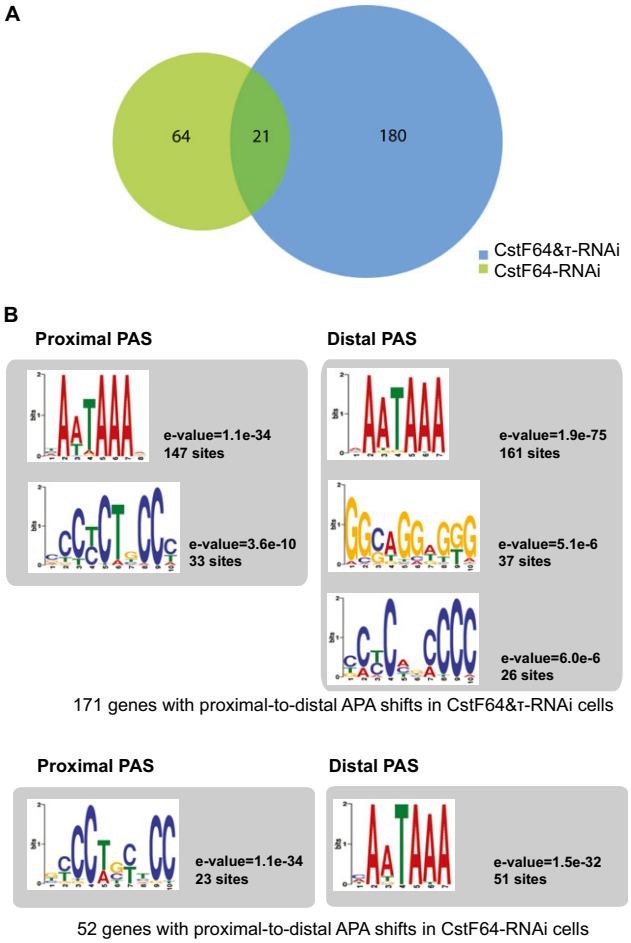


Fig. S5. Comparison of APA changes in CstF64-RNAi and CstF64&τ-RNAi cells. (*A*) Venn diagram comparing the genes with two PASs showing significantly different uses in CstF64-RNAi and CstF64&τ-RNAi cells. (*B*) Multiple Em for Motif Elicitation analysis of the proximal and distal PASs (200-nt sequence centering on the CSs) of genes with proximal-to-distal shifts in CstF64&τ-RNAi (*Upper*) and CstF64-RNAi (*Lower*) cells.

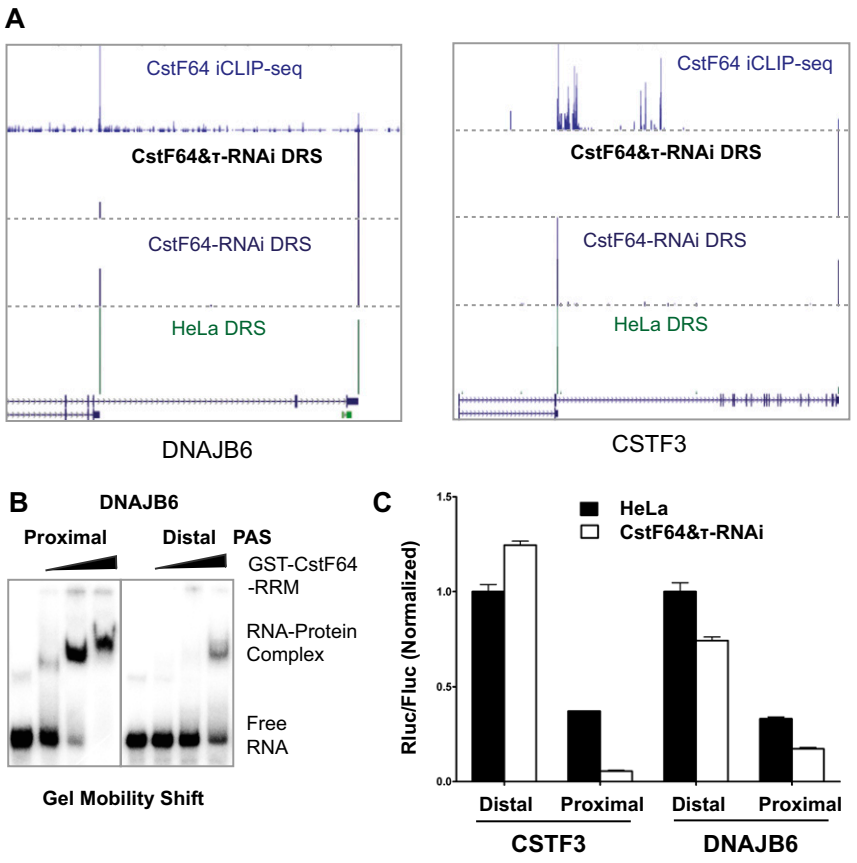


Fig. S6. Mechanisms of CstF64-mediated APA regulation. (*A*) iCLIP-seq and DRS mapping results for *DNAJB6* and *CSTF3*, with each track specified. Two major PASs were observed. (*B*) Gel shift assays using GST-Cstf64-RRM and the 60-nt fragment immediately downstream of the CSs of the proximal and distal PASs of *DNAJB6*. (*C*) Reporter assays for *CSTF3* and *DNAJB6* proximal and distal PASs.The proximal and distal PASs of *CSTF3* and *DNAJB6* were cloned into pPASPORT and transfected into control HeLa or CstF64&τ-RNAi cells. The Rluc/Fluc ratio of each reporter construct was normalized to that of *CSTF3* distal PAS.

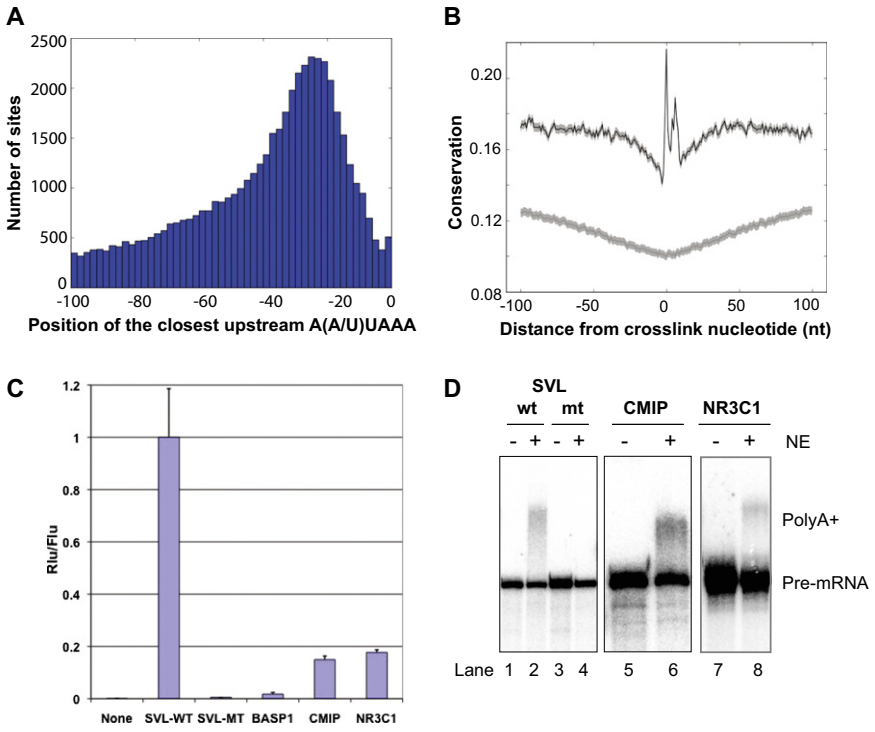


Fig. S7. Characterization of intronic CstF64 binding sites. (*A*) Distribution of the closest upstream A(A/U)UAAA relative to intronic CstF64 crosslinking sites. Position 0 on the *x*-axis represents the CstF64 crosslinking site. The *y*-axis shows the number of CstF64 crosslinking sites that have A(A/U)UAAA at a specific position. (*B*) Conservation of intronic CstF64 binding sites and neighboring sequences (similar to Fig. S1*E*). (*C*) Rluc/Fluc ratio from dual luciferase assays for different sequences (specified on the *x*-axis) inserted into pPASPORT. (*D*) In vitro cleavage/polyadenylation using intronic PASs. Radiolabeled RNAs were extracted after incubation in the presence (+) or absence (–) of HeLa nuclear extract (NE) under APA conditions, resolved on a denaturing 8% gel, and visualized by phosphorimaging. Pre-mRNA and poly(A)$^+$ RNAs are labeled.

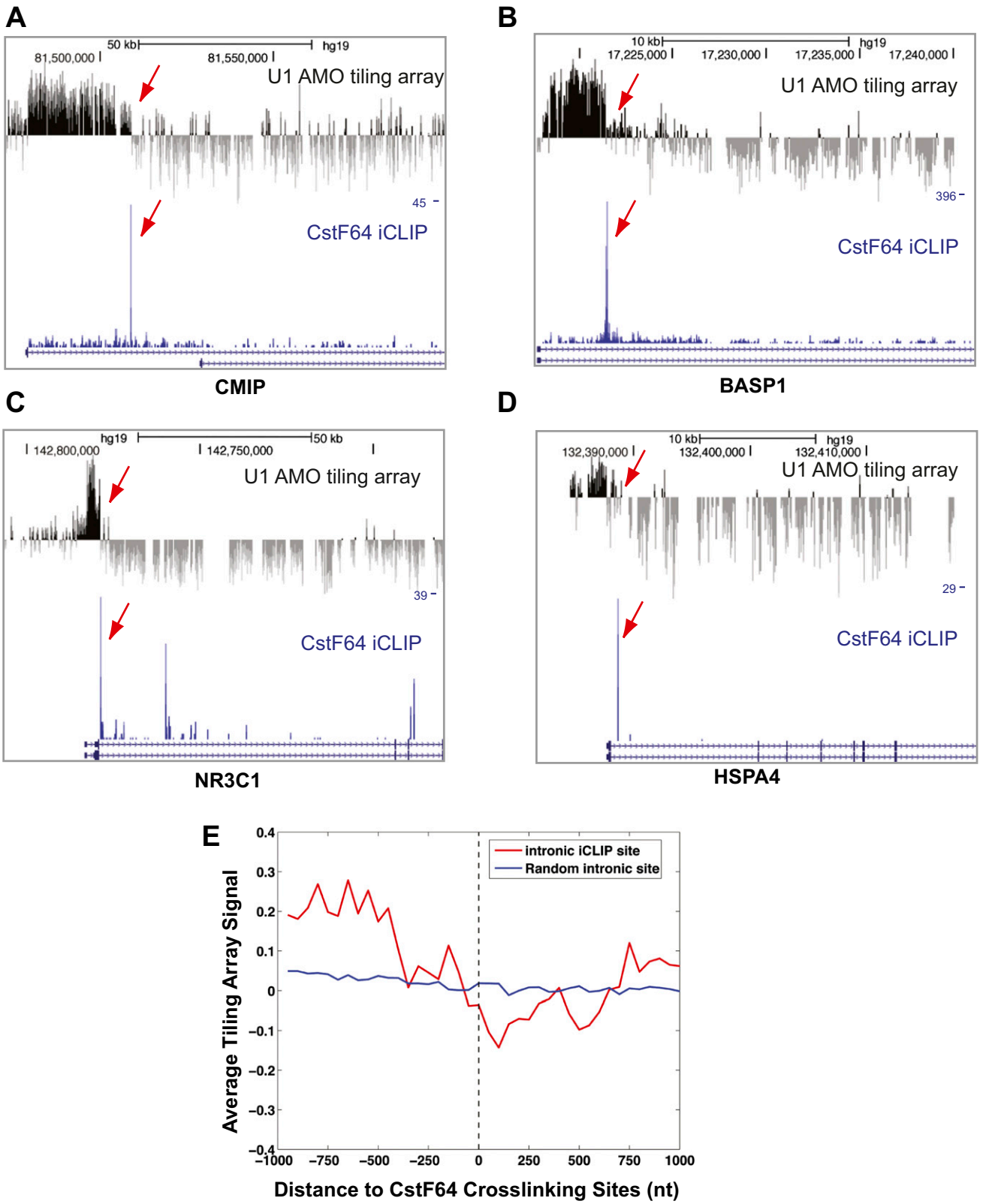


Fig. S8. U1 suppression of CstF64-bound intronic PASs. (*A–D*) The upper tracks show Affymetrix tiling array data from Kaida et al. (1) comparing total RNAs prepared from HeLa cells treated with control or U1-specific antisense morpholino oligos (AMOs). Sites where the signal intensity abruptly decreases represent premature termination sites and are marked by red arrows. The lower tracks show CstF64 iCLIP-seq mapping results. (*E*) Average U1:00 AMO tiling array data surrounding intronic CstF64 binding sites (red line) or random intronic sites (blue line).

1. Kaida D, et al. (2010) U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* 468(7324):664–668.

Other Supporting Information Files

Dataset S1 (XLSX)