iScience, Volume *23*

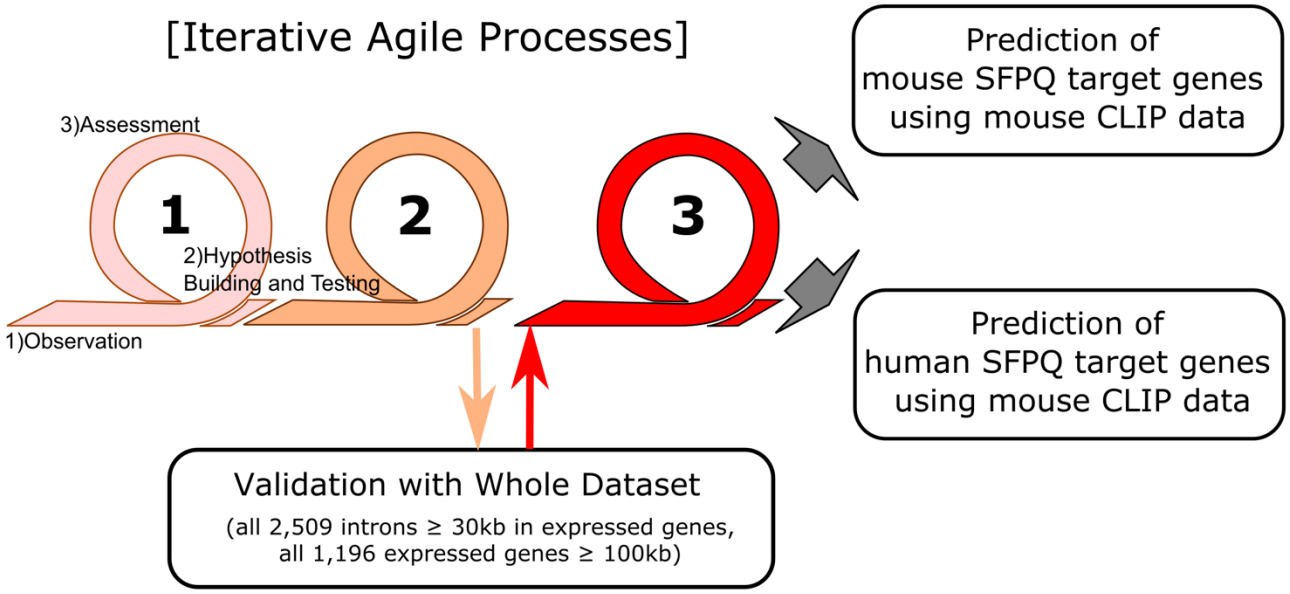
Supplemental Information

Multilateral Bioinformatics Analyses Reveal the Function-Oriented Target Specificities and Recognition of the RNA-Binding Protein SFPQ

Kei Iida, Masatoshi Hagiwara, and Akihide Takeuchi

Supplemental Figures and Legends

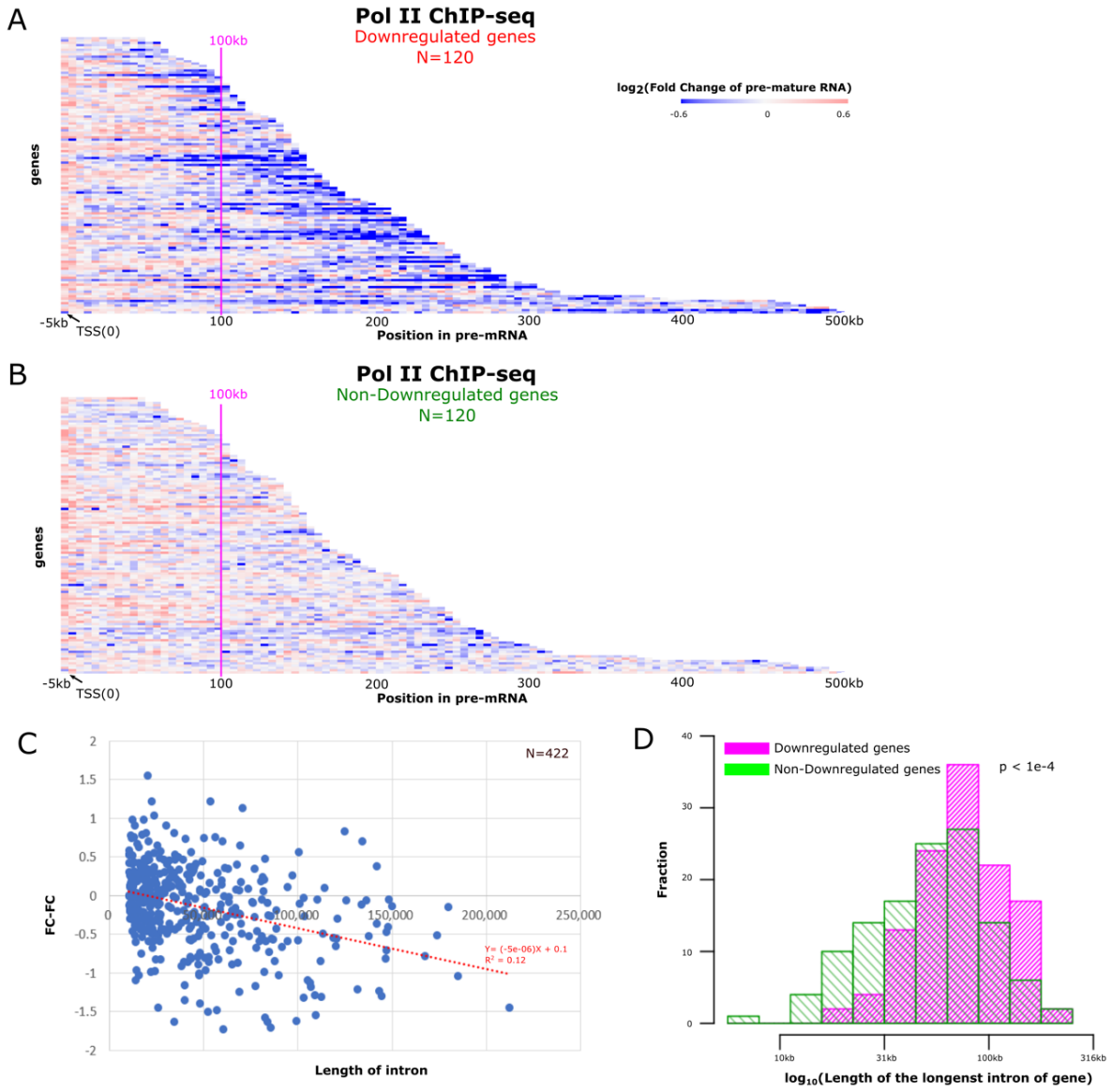
Supplemental Figure S1



Focus	Analysis Set
1. Gene Length	120 pairs of downregulated and non-downregulated genes in Sfpq-KD
2. Intron Features	7 features for 31 severely downregulated introns and 172 non-downregulated introns
3. "Primary Bindings"	Classfied SFPQ-target, and non-target introns

Supplemental Figure S1 Schematic view for iterative agile processes employed in this study, Related to Figure 1, 2, 3, and 4

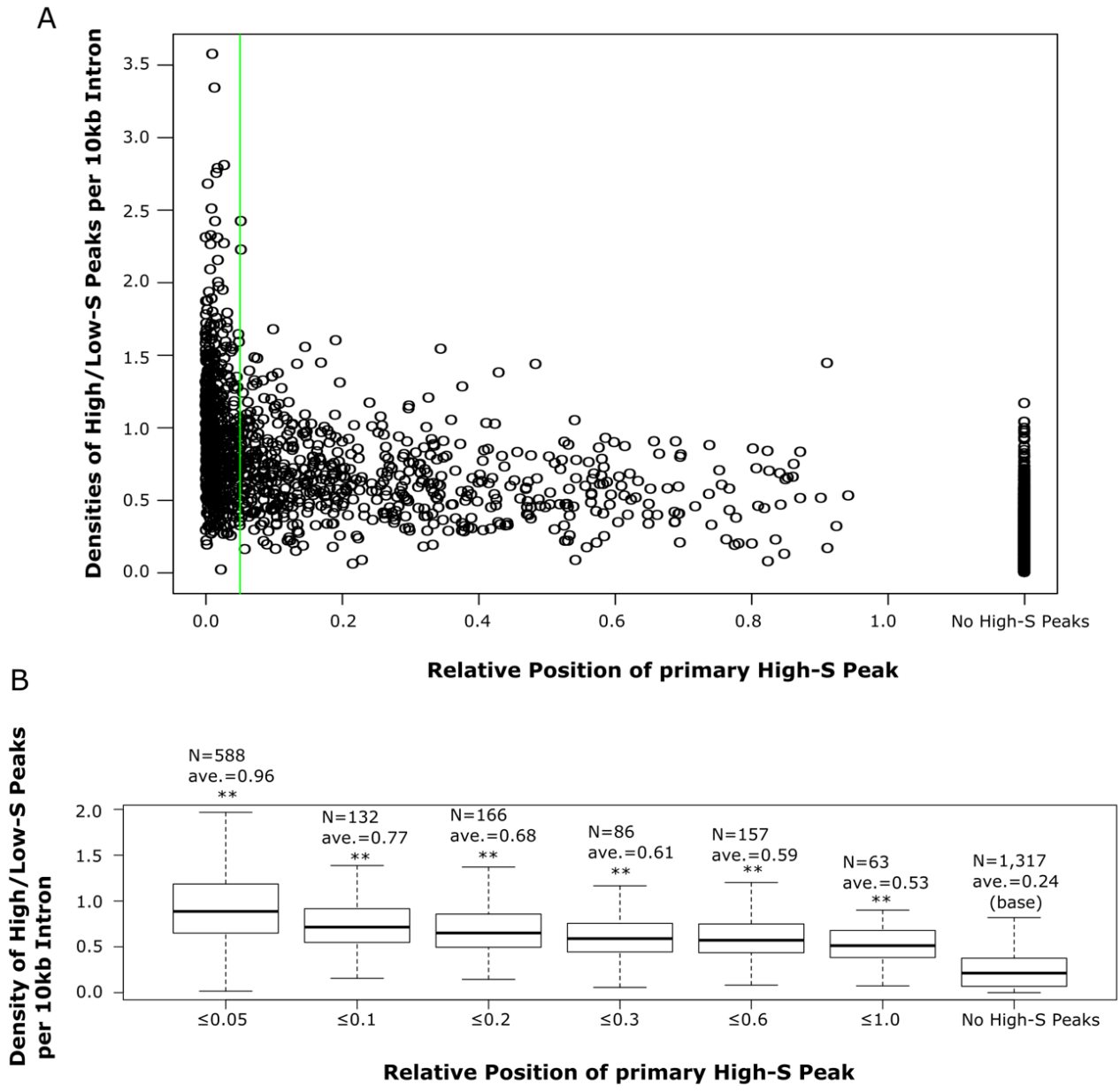
Supplemental Figure S2



Supplemental Figure S2 Pol II ChIP-seq and FC-FC value for length matched downregulated and non-downregulated gene pairs, Related to Figure 1

(A, B) Heatmaps showing pol II ChIP-seq profiles for gene bodies of 120 similar-length pairs of downregulated genes (A) and non-downregulated genes (B). Each block showed 5 kb regions of the gene bodies. Blue and red colors showed down- and up-regulated in *Sfpq*-KD conditions, respectively; (C) A scatter plot for length of introns (x) and ratio of $\log_2$ fold change at 3' 5kb region over long2 fold change at 5' 5kb region, called FC-FC values(y). Red, dotted line showed linear regression result; (D) Histograms for $\log_{10}$ intron lengths. From each gene, intron with the maximum length was selected. Red and green colored graphs showed downregulated gene set and gene sets without large expression changes.

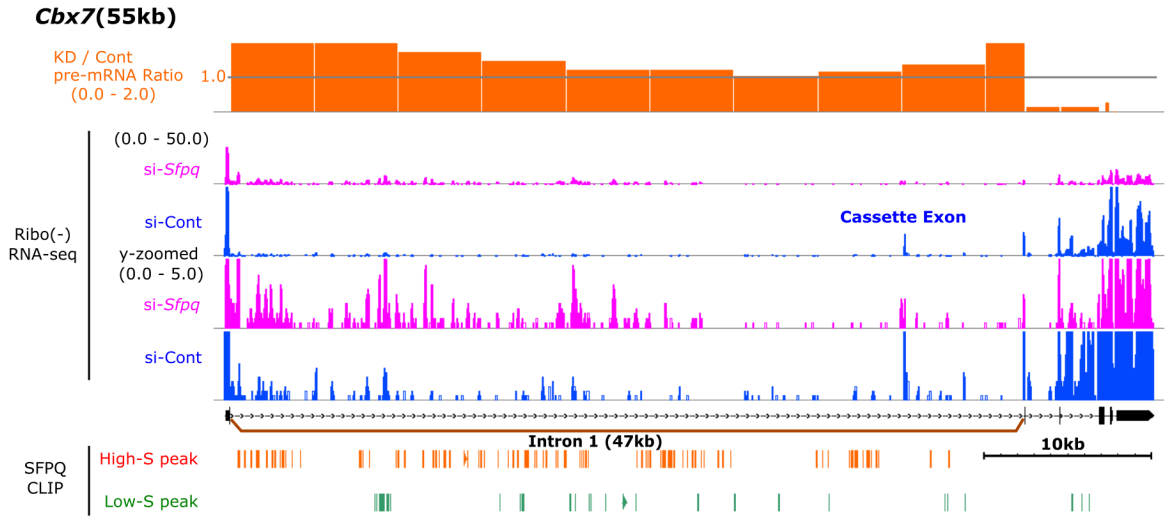
Supplemental Figure S3



Supplemental Figure S3 Impacts of primary High-S peaks on the elevation of total SFPQ-binding on introns, Related to Figure 2

(A, B) A scatter plot (A) and box plots (B) showing relationship between relative position of primary high-S peaks (x) and densities of all (high-S and low-S) SFPQ binding peaks among the introns (i.e. counts per 10 kb intron, y-axis). The differences were statistically tested with Mann–Whitney U test against an intron group with no high-S peaks (located right end). If $p < 1e\text{-}10$, double asterisks were shown.

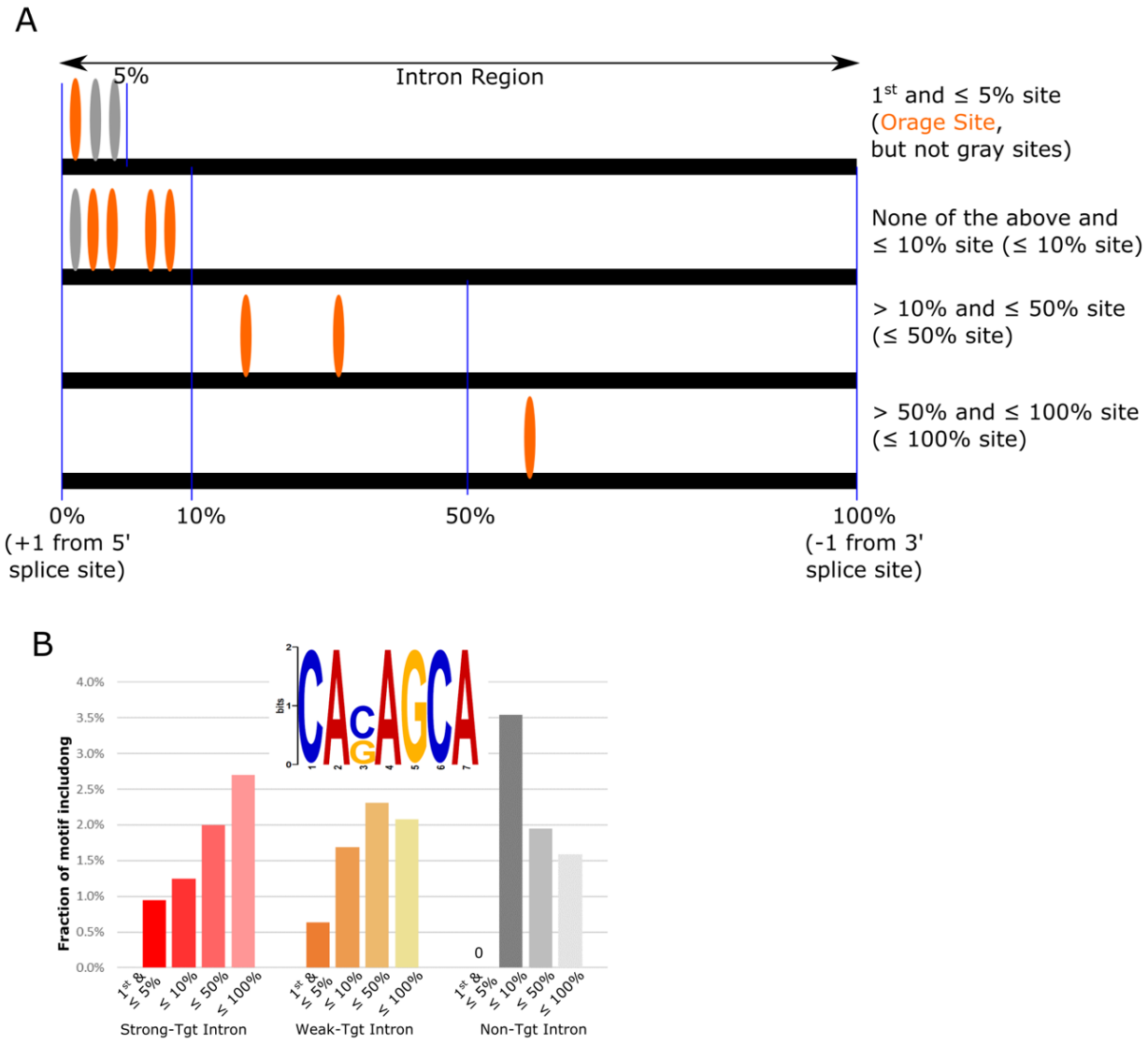
Supplemental Figure S4



Supplemental Figure S4 Genomic view for *Cbx7* locus showing the distributions of Ribo(-) RNA-Seq tags, KO/KD versus Cont pre-mRNA expression ratio, Related to Figure 3

Pre-mRNA length is 55 kb and possessing an intron fulfilled the loose SFPQ target intron criteria. Detailed information for data in this genomic view is described in Figure 1E.

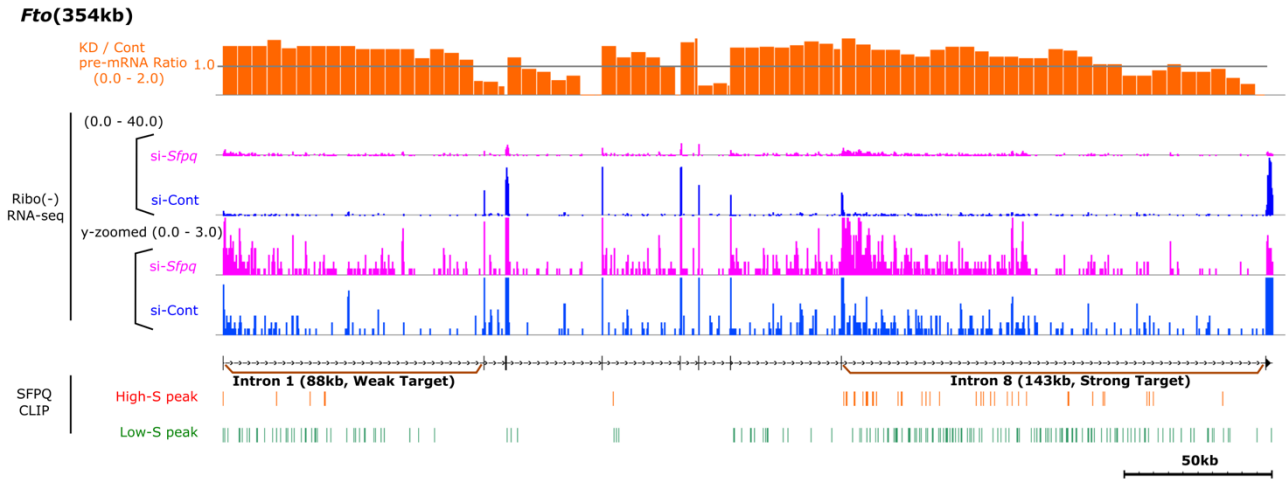
Supplemental Figure S5.



Supplemental Figure S5 Sequence feature analysis of SFPQ binding sites suggested an association between SFPQ binding and upstream exon recognition, Related to Figure 4

(A) Schema showing SFPQ binding positions within introns classification; (B) Bar graph showing fractions of intron regions having the AG-repeat motif for SFPQ-binding found in our previous study.

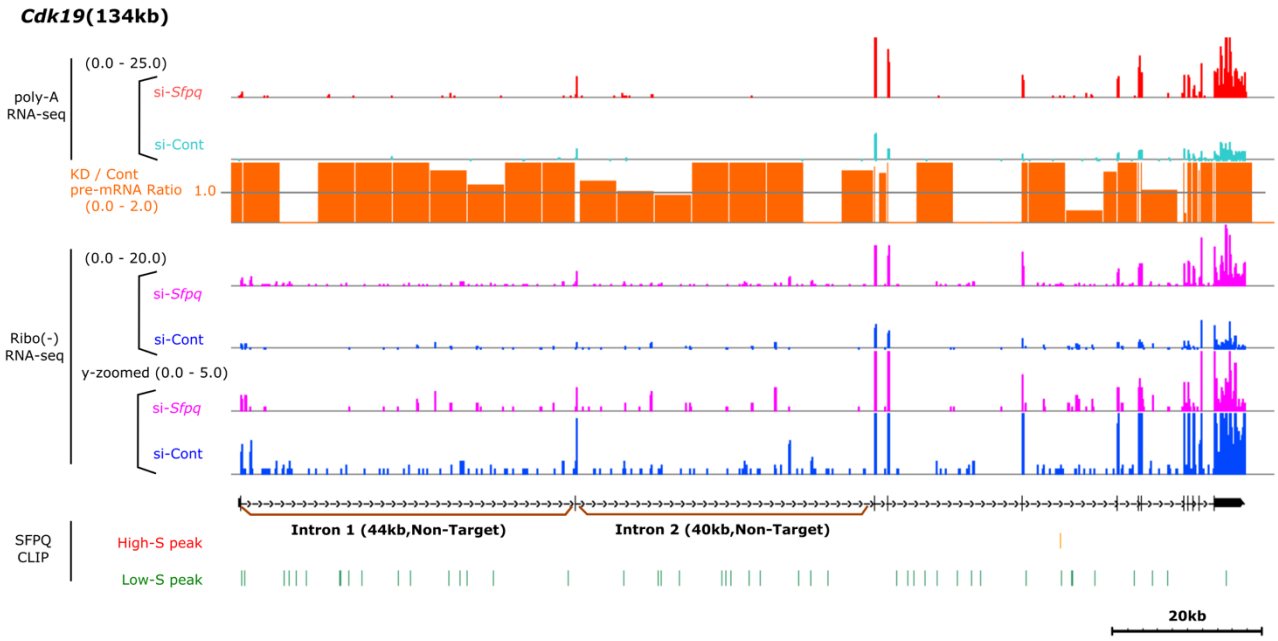
Supplemental Figure S6.



Supplemental Figure S6 Genomic view for *Fto* locus showing the distributions of Ribo(-) RNA-seq tags, KO/KD versus Cont pre-mRNA expression ratio, Related to Figure 6

Detailed information for data in this genomic view is described in Figure 1E.

Supplemental Figure S7.



Supplemental Figure S7 Genomic view for *Cdk19* locus showing the distributions of Ribo(-) RNA-seq tags, KO/KD versus Cont pre-mRNA expression ratio, Related to Figure 1

Poly-A RNA-seq data represents up-regulation of *Cdk19* mature mRNAs in *Sfpq*-KD conditions. Detailed information for data in this genomic view is described in Figure 1E.

Supplemental Table

Supplemental table S1 Presence of strong-, weak-, and non-target introns in severely- and non-downregulated introns, Related to Figure 2

	Strong-target introns	Weak-target introns	Non-target introns	Total
Siverely downregulated introns	14	13	4	31
Non-downregulated introns	3	43	126	172
Total	17	56	130	203

STAR Methods

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
RNA-seq, CLIP-seq, and pol II ChIPseq in Neuro2a Cells	GEO	GEO:GSE60246
Mouse Genome and gene models	NCBI, Refseq	GRCm38, ver.4
proteome dataset obtained from postnatal human brains	PRIDE Archive	PXD005445
Software and Algorithms		
calc_RPKM (original scripts for calculating RPKM)	https://github.com/keiiida/calc_RPKM	06/04/2020
count_Peaks (original scripts for counting peaks)	https://github.com/keiiida/count_Peaks	06/04/2020
Integrated Genome Browser	https://bioviz.org/	9.1.2
MEME Suite	http://meme-suite.org/	4.9.0
SpliceAID 2	http://www.introni.it/splicing.html	Feb. 2013
Metascape	https://metascape.org/	Apr. 2018

Transparent Methods

Analysis dataset

The data obtained in our previous study (Takeuchi et al., 2018), including mRNA-seq (polyA+RNAs) for mature mRNA, Ribo(-) RNA-Seq (rRNA depleted, polyA+/- RNA) for premature mRNA (pre-mRNA), CLIP-seq, and chromatin immunoprecipitation (ChIP) for RNA polymerase II-seq (pol II ChIP-seq) from mouse Neuro2a cells, were used in the current study. *Sfpq*-knockdown (KD) was conducted using small interfering RNA (siRNA) targeting *Sfpq* mRNA (si-*Sfpq*), and control (si-Cont) was used as negative control siRNA or siRNA for *Luciferase* mRNA. Data were deposited in the NCBI Gene Expression Omnibus (GEO) with accession number GSE60246. *Mus musculus* genome ver. GRCm38 and RefSeq gene models (ver. 4) were used for analysis (O'Leary et al., 2016). In expression analysis, the longest transcripts were used as representative models for genes having several isoforms.

Selection of SFPQ regulatory target and nontarget genes

The analysis groups of SFPQ target and non-target genes were selected according to the SFPQ-binding and fold change (FC) data under *Sfpq* KD conditions and used in Fig. 1. Genes meeting the following criteria were considered to exhibit significant SFPQ binding: 1) at least one high-stringent (High-S) SFPQ binding peak, and 2)

more than 32 low-stringent (Low-S) SFPQ binding peaks among pre-mRNA regions. These criteria were defined in our previous study (Takeuchi et al., 2018). Peak calling was performed according to the eCLIP method (Van Nostrand et al., 2016), and peaks with *p* values< 0.01 that were consistently called in duplicated experiments were selected. Entire peaks were divided into High-S and Low-S peaks according to the FCs relative to size-matched (SM) input; peaks with FCs of 2 or more were defined as High-S peaks, and the remaining peaks were defined as Low-S peaks (Takeuchi et al., 2018). SFPQ regulatory target genes were selected using previously described criteria, i.e., transcripts per million transcripts (TPM) ≥ 2 as expression checks, TPM value of FC < one-third under KD conditions, and *q* value < 0.01 (calculated with DEseq2) (Love et al., 2014). For genes that were not downregulated, FCs ranged from 0.83 to 1.20, with TPM values ≥ 2 in either *Sfpq*-KD or si-Cont conditions. In total, 120 length-matched gene pairs were selected from target and nontarget genes (length difference < 10 kb).

Comparison between SFPQ regulatory target and nontarget genes

For visualization of Ribo(-) RNA-Seq or pol-II ChIP-seq data, we separated pre-mRNAs into 5-kb windows, calculated the reads per kilo-bases of the region and per million mapped reads (RPKM) and relative RPKM values as KD/Cont (Takeuchi et al., 2018). Relative RPKM values were adjusted such that the mean of ratios among all windows of all expressed genes was zero, as previously described (Takeuchi et al., 2018). Reads mapped to exons were excluded to detect changes in pre-mRNA levels. Genome views were drawn with Integrated Genome Viewer (Freese et al., 2016). The length distribution of the longest introns between regulatory target and nontarget genes was analyzed by Mann-Whitney U tests.

Criteria for identifying SFPQ target introns

Using all introns in SFPQ target genes, target and nontarget introns were selected and characterized (Fig. 2). In total, 31 introns were selected as SFPQ target introns using the following criteria from SFPQ target genes: intron length ≥ 10 kb and FC-FC values < -1. FC-FC values were defined as the differences in $\log_2$ FC values between

the terminal 5-kb intronic region and the initial 5-kb region. To calculate FC-FC values, we used calc_RPKM scripts (https://github.com/keiiida/calc_RPKM). We selected 172 introns as nontarget introns (control) using the following criteria from SFPQ target genes: length ≥ 10 kb and FC-FC values between -0.25 and 0.25. Cumulative curves were plotted with the factors for intron length and the number/density/relative position of primary SFPQ binding within introns for both High-S and Low-S peaks. Using cumulative plots, two values were calculated for each factor; one value maximized the differences of occupancy between SFPQ-target and nontarget introns (designated as "high-stringent criteria"), whereas the other value provided high occupancy (0.90–0.95) of SFPQ target introns (designated as "low-stringent criteria"). Number of SFPQ binding peaks on each introns were counted with count_Peaks script (https://github.com/keiiida/count_Peaks).

Gene classification analysis

All expressed genes were classified according to the presence of the identified SFPQ target introns (schematically summarized in Fig. 3B). Among long genes (≥ 100 kb), genes whose longest introns were less than 30 kb were classified as "genes without long introns", and the remaining genes having introns ≥ 30 kb were further classified as follows: genes containing at least one intron fulfilling the high-stringent criteria were classified as "genes with strong-target introns" (genes w/ Strong-Tgt introns); genes not having introns that fulfilled the high-stringent criteria but had more than one intron that fulfilled the low-stringent criteria were "genes with weak-target introns" (genes w/ Weak-Tgt introns); and the remaining genes were classified as "genes without non-target introns" (genes w/ Non-Tgt introns).

Identifying specific consensus sequences/features in SFPQ target introns

For analysis of consensus sequences/features in SFPQ target introns, we used Strong-, Weak-, and Non-Tgt introns. Two previously reported SFPQ-binding motifs were used as known motifs (Takeuchi et al., 2018). For the novel motif search, we used 703 High-S SFPQ binding regions located close to the 5′ splice sites and within the upstream

10% regions from the 5′ end among strong and weak SFPQ-target introns. Windows that were 50 nucleotides upstream and downstream to peak summit positions were used for the motif search. Seven-base motifs were searched with MEME Suite. We employed the MAST program from the MEME Suite with a 1×10^{-4} threshold (with parameters "-hit_list -mt 1e-4") (Bailey et al., 2009). For sequence feature analyses, we counted exonic splicing enhancers (ESEs) and exonic splicing silencers (ESSs) with SpliceAID 2 (latest ver. Feb. 2013) (Piva et al., 2009). If ESE and ESS consensus sequences were found in intronic regions, we counted them as intronic splicing silencers (ISSs) and intronic splicing enhancers (ISEs), respectively.

Gene Ontology (GO) analysis

GO enrichment analysis for each gene group was performed using the Metascape web tool (Tripathi et al., 2015).

Proteome analysis data set and processing

The proteome dataset obtained from postnatal human brains (PXD005445, PRIDE Archive, and EBI) (Carlyle et al., 2017) was used for analyzing the co-expression of SFPQ and protein products from extra-long genes in human brains. Protein expression was counted for each sample and was then normalized as peptides per 30,000 peptides (PP30K). Human gene symbols were converted to mouse symbols using a Perl script according to homologene tables (Sayers et al., 2019). Over-representation between co-expressed genes with SFPQ and predicted SFPQ target genes was evaluated. Genes commonly expressed both in Neuro2a and human brains were used as the denominator of expressed genes. We calculated fractions of genes possessing Strong-, Weak- and Non-target introns of SFPQ with the fraction of genes lacking long introns against expressed genes. Moreover, the fractions of genes showing high Pearson correlation coefficients (PCC ≥ 0.5 to SFPQ) were also calculated. We compared the enrichment between SFPQ target genes and high PCC genes, and significant differences were analyzed using Mann-Whitney U tests.

Supplemental References

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202-8.

Carlyle, B.C., Kitchen, R.R., Kanyo, J.E., Voss, E.Z., Pletikos, M., Sousa, A.M.M., Lam, T.T., Gerstein, M.B., Sestan, N., and Nairn, A.C. (2017). A multiregional proteomic survey of the postnatal human brain. Nat. Neurosci. *20*, 1787–1795.

Freese, N.H., Norris, D.C., and Loraine, A.E. (2016). Integrated genome browser: visual analytics platform for genomics. Bioinformatics *32*, 2089–2095.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

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., et al. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat. Methods *13*, 508–514.

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., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733-45.

Piva, F., Giulietti, M., Nocchi, L., and Principato, G. (2009). SpliceAid: a database of experimental RNA target motifs bound by splicing proteins in humans. Bioinformatics *25*, 1211–1213.

Sayers, E.W., Beck, J., Brister, J.R., Bolton, E.E., Canese, K., Comeau, D.C., Funk, K., Ketter, A., Kim, S., Kimchi, A., et al. (2019). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res.

Takeuchi, A., Iida, K., Tsubota, T., Hosokawa, M., Denawa, M., Brown, J.B., Ninomiya, K., Ito, M., Kimura, H., Abe, T., et al. (2018). Loss of Sfpq Causes Long-Gene Transcriptopathy in the Brain. Cell Rep. *23*, 1326–1341.

Tripathi, S., Pohl, M.O., Zhou, Y., Rodriguez-Frandsen, A., Wang, G., Stein, D.A., Moulton, H.M., DeJesus, P., Che, J., Mulder, L.C.F., et al. (2015). Meta- and Orthogonal Integration of Influenza "OMICs" Data Defines a Role for

UBR4 in Virus Budding. Cell Host Microbe *18*, 723–735.