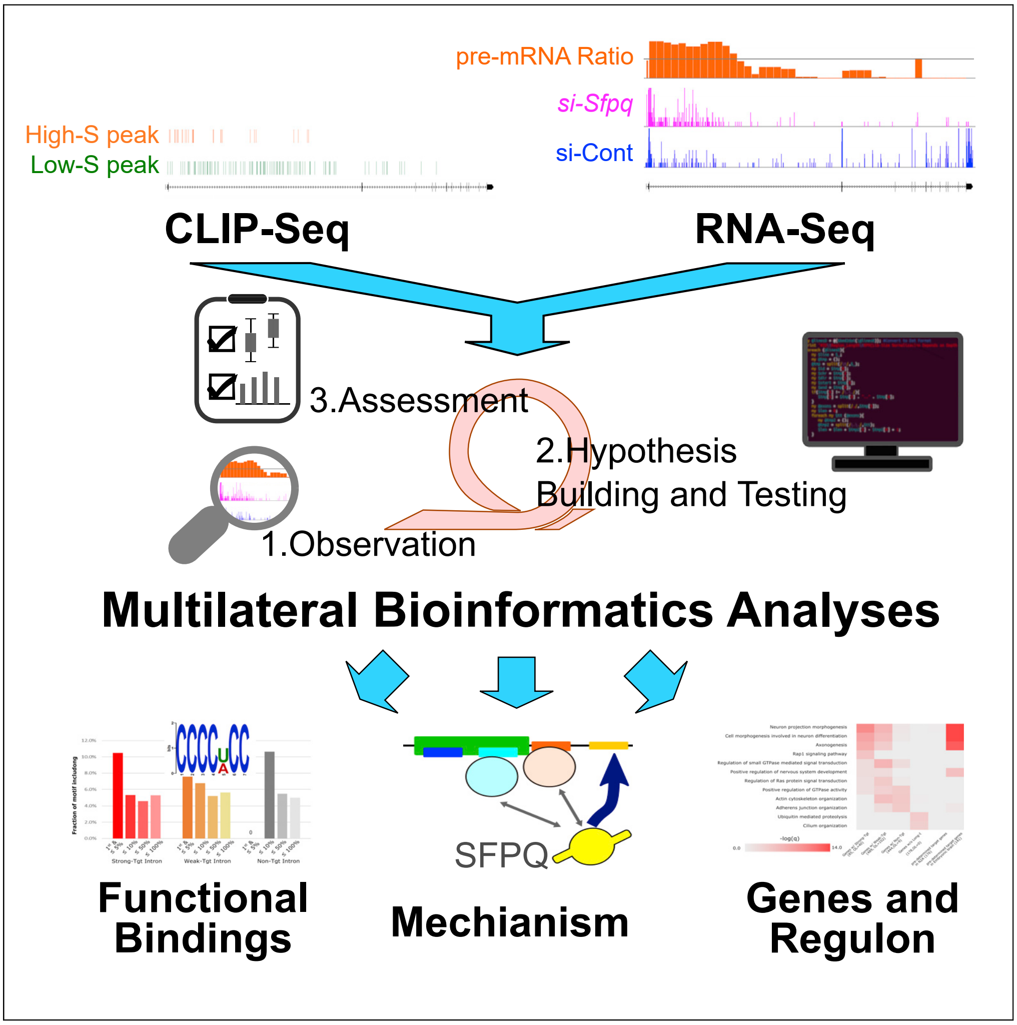
iScience

CellPress OPEN ACCESS

Article

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



Kei Iida, Masatoshi Hagiwara, Akihide Takeuchi

takeuchi.akihide.8r@kyoto-u.ac.jp

HIGHLIGHTS

We developed a multilateral bioinformatics (MLBI) analysis for RNA biology study

MLBI analyses successfully identified functionally essential binding from CLIP-seq

MLBI analyses revealed the target recognition mechanism of RBP on mRNAs

MLBI analyses identified RBP regulon and predicted target genes in close species

Iida et al., iScience *23*, 101325
July 24, 2020 © 2020 The Author(s).
https://doi.org/10.1016/j.isci.2020.101325

CellPress OPEN ACCESS

Article

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

Kei Iida,[1] Masatoshi Hagiwara,[2] and Akihide Takeuchi[2,3,*]

SUMMARY

RNA-binding proteins (RBPs) recognize consensus sequences and regulate specific target mRNAs. However, large-scale CLIP-seq revealed loose and broad binding of RBPs to larger proportion of expressed mRNAs: e.g. SFPQ binds to >10,000 pre-mRNAs but distinctly regulates <200 target genes during neuronal development. Identification of crucial binding for regulation and rules of target recognition is highly anticipated for systemic understanding of RBP regulations. For a breakthrough solution, we developed a bioinformatical method for CLIP-seq and transcriptome data by adopting iterative hypothesis testing. Essential binding was successfully identified in C-rich sequences close to the 5′ splice sites of long introns, which further proposed target recognition mechanism via association between SFPQ and splicing factors during spliceosome assembly. The identified features of functional binding enabled us to predict regulons and also target genes in different species. This multilateral bioinformatics approach facilitates the elucidation of functionality, regulatory mechanism, and regulatory networks of RBPs.

INTRODUCTION

In mammals, approximately 1,500 genes encode RNA-binding proteins (RBPs), which directly bind to specific target mRNAs by recognizing consensus sequences and regulate RNA metabolism (Castello et al., 2012; Gerstberger et al., 2014; Hentze et al., 2018; Sundararaman et al., 2016). The consensus sequences of each RBP are called cis-elements; these sequences define the target specificities and functions of RBPs. For example, Rbfox and Nova bind to UGCAUG and YACY motifs close to splice sites, respectively and regulate tissue-specific alternative splicing (Licatalosi et al., 2008; Weyn-Vanhentenryck et al., 2014). Recent studies by the ENCODE project are attempting to reveal the binding properties and functions of entire RBPs by genome-wide identification of target RNAs, binding sites, and consensus sequences using CLIP-seq (ultraviolet [UV]-crosslinking and immunoprecipitation with high-throughput sequencing) (Van Nostrand et al., 2018). However, RBP binding is not necessarily restricted to unique consensus sequences but is broadly observed in a larger proportion of expressed gene transcripts. In their study, totally 26 RBPs were analyzed and more than 30% (8 RBPs including SFPQ) did not showe distinct consensus sequences. For example, SRSF5 (serine and arginine-rich splicing factor 5), PABPN1L (PABPN1-like, cytoplasmic), RBP22, and RBM15B (RNA-binding motif protein 15B) were shown to have more than 10 nonoverlapping motifs in each RBP; however, these motifs accounted for less than 50% of all binding sites (Van Nostrand et al., 2018). These results suggest that RBP binding is not simply defined by consensus sequences but that binding specificities are also regulated by more complex rules. Binding of RBPs is known to be altered by the flexibility of RNA-binding domains to sequence variants of cis-elements, by adjacent sequences and mRNA structures, and by interactions with other RBPs (Fu and Ares, 2014). Furthermore, the recruitment mechanism, binding portion in pre-mRNAs, and co-binding partners allow RBPs to play multiple context-dependent roles, such as roles in transcription, splicing, or stabilization of pre-mRNAs (Hentze et al., 2018). Therefore, an imminent issue is to elucidate the regulatory rules that define the binding specificity and functions of RBPs with regard to distinct subsets of mRNA.

SFPQ (Splicing factor proline and glutamine rich) is a multifunctional RBP that regulates mRNA processing, transcription, and DNA repair (Knott et al., 2016; Yarosh et al., 2015). In a recent study, we found that SFPQ regulates the transcriptional elongation of extra-long genes (≥ 100 kb) in the brain and muscles (Hosokawa et al., 2019; Takeuchi et al., 2018). Mechanistically, SFPQ co-transcriptionally binds to target pre-mRNAs,

[1] Medical Research Support Center, Kyoto University, Graduate School of Medicine, Kyoto University, Konoecho Yoshida Sakyo-ku, Kyoto 606-8501, Japan

[2] Department of Anatomy and Developmental Biology, Graduate School of Medicine, Kyoto University, Konoecho Yoshida Sakyo-ku, Kyoto 606-8501, Japan

[3] Lead Contact

*Correspondence: takeuchi.akihide.8r@kyoto-u.ac.jp

https://doi.org/10.1016/j.isci.2020.101325

recruits cyclin-dependent kinase 9 (CDK9), and activates RNA polymerase II (Pol II). Notably, SFPQ binding has been detected on pre-mRNAs from 11,286 genes in CLIP-seq, and loss of *Sfpq* disrupts the expression of 141 genes, representing only 1.2% of all SFPQ-bound genes, and 95.6% of these 141 genes have pre-mRNAs > 100 kb in length (Takeuchi et al., 2018). In this study, we aimed to decipher the binding specificity of RBPs linked to unique functions using a newly developed analytical method with SFPQ as a model. Conventional bioinformatic analysis methods for determining RBP functions have adopted pre-settled or so-called ''waterfall'' analytical designs; however, these approaches have not succeeded in addressing the issues described earlier. Thus, we employed an ''agile data science method'' (Russell, 2017) as an iterative approach for observation point installation and hypothesis-driven mathematical modeling (Figure S1). We applied this method to elucidate how SFPQ recognizes and regulates specific long pre-mRNAs among the huge spectrum of potential binding target mRNAs and revealed the underlying molecular mechanisms that define the target specificities of SFPQ.

RESULTS

SFPQ Regulates Long Intron-Containing Genes

To identify functional SFPQ binding, which is essential for regulating transcriptional elongation of extra-long genes, we integrated data of genome-wide SFPQ-binding mapping (CLIP-seq) with RNA-seq data under *Sfpq*-KD and si-control (Cont) conditions. We then crosschecked the relationships between SFPQ binding and fold change (FC) of expression. In Neuro2a cells, 176 genes (1.2% of total 11,434 SFPQ-bound genes) were identified as SFPQ regulatory targets that had significant SFPQ-binding peaks, and their expression levels were significantly downregulated under *Sfpq*-KD condition (FC of *Sfpq*-KD/Cont ratio <1/3, see Methods for detailed regulatory target gene identification criteria) (Figure 1A, genes indicated by red dots). It is accounted for 13.7% of genes passing the same SFPQ-binding criteria (1,289 genes), whereas 432 (33.5%) were identified as nontarget genes with expression levels not affected under *Sfpq*-KD condition ($0.83 \le FC \le 1.20$; Figure 1A, genes indicated as green dots). These results indicated that total SFPQ binding is not the only factor to identify the SFPQ target genes. Because SFPQ was shown to be required for long gene expression, we next evaluated the gene lengths of SFPQ target genes. The results showed that 85.8% (151/176) were longer than 100 kb, suggesting that only longer genes with significant SFPQ binding were affected by *Sfpq*-KD. However, the distribution of gene lengths in these two groups overlapped (42 kb–1.9 mb for downregulated genes and 13 kb to 1.1 mb for genes that were not downregulated; Figure 1A), indicating that gene length was informative, but not sufficient, for defining SFPQ target genes.

To determine the SFPQ-binding characteristic to SFPQ target genes, we selected size-matched gene pairs from SFPQ target and nontarget genes, both of which had substantial SFPQ bindings. We identified 120 gene pairs that exhibited pre-mRNA length differences less than 10 kb and fulfilled the criteria for SFPQ binding (Figure 1B). We next compared the transcriptional elongation statuses in these groups using Ribo(−) RNA-seq and Pol II ChIP-seq. SFPQ target genes showed a gradual decrease in pre-mRNA level and Pol II distribution, particularly beyond 100 kb downstream of transcription start sites (TSSs), which was a symptom of transcriptional elongation defects (Figures 1C and S2A). Although SFPQ nontarget genes had a similar number of SFPQ-binding peak as SFPQ target genes, neither pre-mRNA expression nor Pol II density was decreased for these genes (Figures 1D and S2B), indicating that all but part of SFPQ binding was functionally essential for regulating long pre-mRNA expression.

To identify functionally essential SFPQ binding, we closely examined SFPQ binding in target genes. In typical SFPQ target genes, dense SFPQ binding was detected within extra-long introns. *Sfpq*-KD caused a gradual decrease of pre-mRNAs expression within these introns, as shown for the *Samd4* gene, which possessed a distinctive long intron 2 (131 kb; Figure 1E). A gradual decrease of the pre-mRNA level from the 5′–3′ region in *Sfpq*-KD was observed as the stepwise decline of the pre-mRNA ratio between *Sfpq*-KD and control conditions. In contrast, the non-SFPQ target gene *Ralgapa1* had a pre-mRNA length similar to that of *Samd4* but did not possess extra-long introns (Figure 1F). SFPQ binding on *Ralgapa1* pre-mRNA was broadly detected on the entire pre-mRNA, although the density was much lower than that of the *Samd4* gene. In addition, the levels of *Ralgapa1* pre-mRNA did not gradually decrease, and its KD/Cont pre-mRNA ratio was approximately 1.0 over the entire length. These data suggested that dense SFPQ binding to long introns was functionally essential for long pre-mRNA expression, and regulatory targets of SFPQ showed gradual decrease of their pre-mRNA under *Sfpq*-KD. To detect target introns of SFPQ, changes in pre-mRNA levels were examined by calculating the FC-FC value as the difference in the KD/Cont pre-mRNA ratio (FC) between the 3′- and 5′-regions. The FC-FC values of SFPQ target introns showed negative values owing to the smaller FC values in 3′-regions than in the 5′-regions. Using this FC-FC value, we analyzed its correlation with intron length for all 422 introns ≥10 kb among the 120 SFPQ

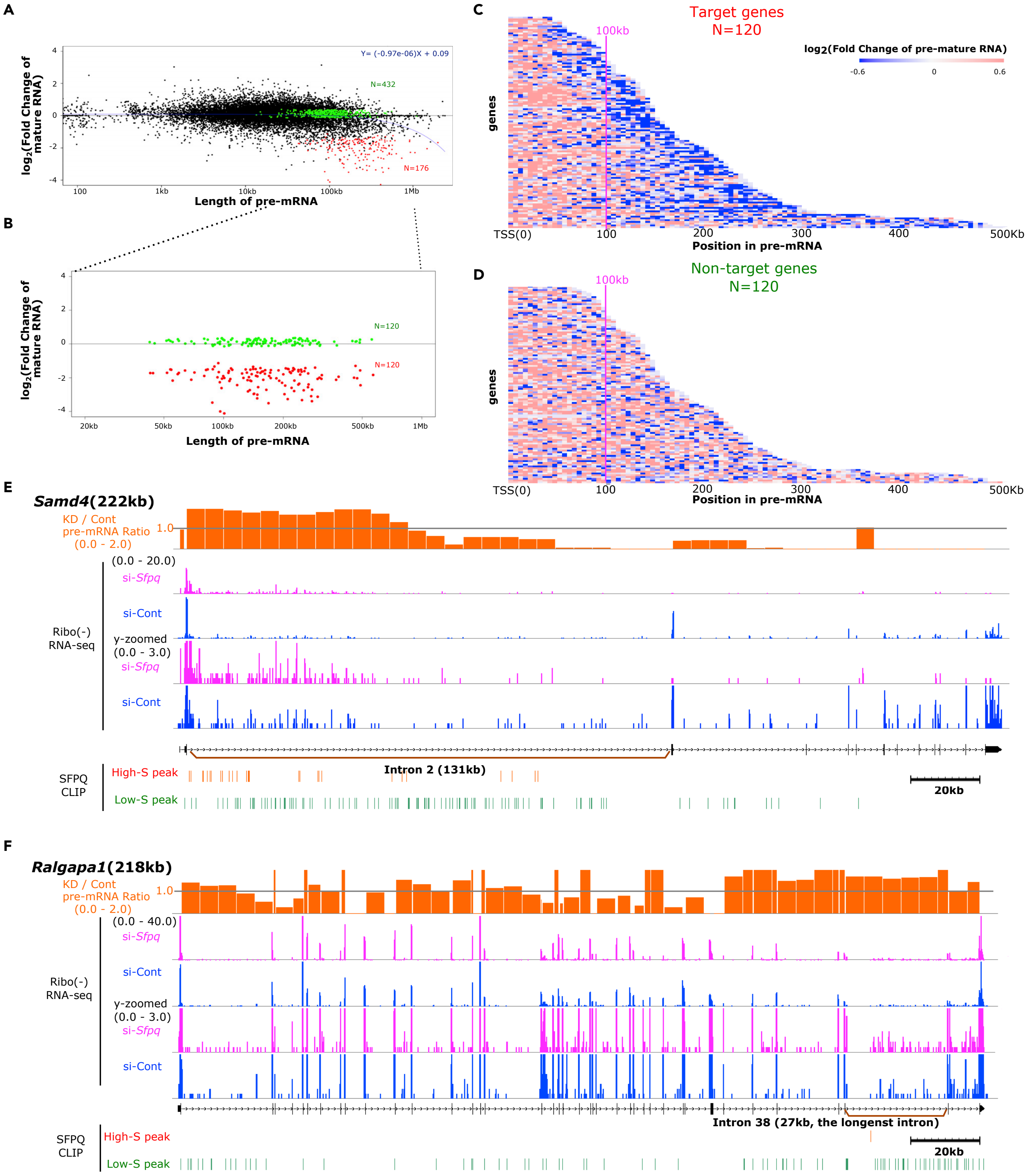


Figure 1. SFPQ Regulated Genes Containing Long Introns

(A) Relationships between pre-mRNA lengths and $\log_2$ FC in *Sfpq*-KD and control (Cont-si) Neuro2a cells. The red dots show 176 genes that were significantly downregulated (FC of *Sfpq*-KD/Cont ratio <1/3) with significant SFPQ binding peaks. The green dots show 432 genes meeting the same SFPQ binding criteria, but without changes in FC values (0.83 < FC < 1.20).

(B) Higher magnification of the graph in A from the range of 20 kb to 1 Mb, showing similar distributions of 120 gene pairs with pre-mRNA lengths (length difference <10 kb).

Figure 1. *Continued*

(C and D) Heatmaps showing pre-mRNA expression profiles for the 120 downregulated genes, captioned as target genes (C) and 120 non-downregulated genes, captioned as non-target genes. (D). Each block shows a 5-kb window for the gene bodies. Down- or upregulation of pre-mRNA expression levels is shown with color according to the Log_2 FC of *Sfpq*-KD/Cont.

(E) Genomic view for *Samd4*, which was shown to be an SFPQ target gene in a previous study. Smoothed Ribo(−) RNA-seq profiles of *Sfpq*-KD/si-Cont (orange), *Sfpq*-KD Ribo(−) RNA-seq profiles (pink), and si-Cont (blue) positions of high-S (red) and low-S peaks (green) are shown.

(F) Genomic view of *Ralgapa1*, a size-matched gene pair partner of *Samd4*, among non-target genes.

See also Figures S1, S2, and S7.

target genes. A weakly negative correlation was detected between these two factors (Figure S2C), indicating that intron length was an important feature of SFPQ targets. We further compared the intron lengths of SFPQ target and nontarget genes. The results showed that SFPQ target genes tended to have longer introns than nontarget genes; however, their distribution overlapped (Figure S2D). Together, these results indicated that SFPQ regulated longer introns as its targets and that other factors were expected to delineate target and nontarget genes.

Identification of Factors Defining SFPQ Target Specificity

Next, we attempted to identify the factors that defined the target specificity of SFPQ, particularly the rules that determined SFPQ binding and functionality in specific introns. We attempted to identify the features of SFPQ binding on SFPQ target introns. We selected introns from SFPQ target genes for which the FC-FC values were significantly negative (≤−1). These introns were designated as "severely downregulated introns." For comparison, we selected introns from SFPQ target genes for which the length was ≥10 kb and the FC-FC value was between −0.25 and +0.25 as controls as "non-downregulated introns." With these criteria, 31 and 172 introns were identified as severely downregulated and non-downregulated introns, respectively (Figure 2A). In these two intron groups, the following four potential factors were examined for segregation of target and nontarget introns: (1) intron length; (2) total number of SFPQ binding; (3) density of SFPQ binding using normalized SFPQ binding counts per 10 kb; and (4) position of SFPQ binding closest to the 5′ splice site (termed the primary binding position). Previous studies have demonstrated the significance of SFPQ binding in the 5′ intron region as the seed-binding event mediating the polymerization of SPFQ on mRNA via its coiled-coil domain. This primary binding further allows serial SFPQ binding from 5′ to 3′ on pre-mRNAs and facilitates transcriptional elongation (Huang et al., 2018; Lee et al., 2015; Takeuchi et al., 2018). Therefore, the position of 5′ SFPQ binding was examined. In these analyses, we considered the stringency of SFPQ binding, which indicates the specificity and exclusiveness of SPFQ binding at certain RNA loci compared with size-matched input controls (SMI) (Van Nostrand et al., 2016). Among all SFPQ binding peaks with p values <0.01, those with CLIP-tag/SMI-tag FC ≥ 2 were designated as "High-S peaks" (high-stringent binding peaks compared with SMI), and the remaining peaks were termed as "Low-S peaks" (low-stringent binding peaks compared with SMI), as defined in our previous study (Takeuchi et al., 2018). These peaks accounted for 27,732 High-S peaks and 208,828 Low-S peaks. Then, factors (1)–(4) were analyzed and compared between severely downregulated and non-downregulated intron considering High-S- and Low-S SFPQ-binding peaks, except factor (1) (seven factors were investigated in total). The seven factors described earlier were examined to determine whether these factors could delineate severely downregulated and non-downregulated introns by drawing cumulative curves (Figures 2B–2H). Six features clearly segregated the 2 groups: (1) intron length, (2) High-S peak count, (3) Low-S peak count, (4) High-S peak density, (5) Low-S peak density, and (6) primary High-S peak binding position. Next, we checked whether the combination of these six features could be used for identifying functional SFPQ binding. For this purpose, we selected the following two thresholds for each feature: (1) a threshold giving the maximum difference ratio between two intron groups (designated as "high-stringent criteria," indicated with black arrows), and (2) a threshold including 90% of highly downregulated introns (termed as "low-stringent criteria," indicated with gray arrows; Figures 2B–2H). The distribution of severely downregulated and non-downregulated introns separated by these criteria was summarized in Table 1. Combinations of the six features were tested using these two thresholds with 31 downregulated and 172 non-downregulated introns in SFPQ target genes as following. Total 203 introns were classified into three groups: (1) introns fulfilling high-stringent criteria for all six features (designated as "strong-target introns"); (2) introns not meeting at least one high-stringent criterion but fulfilling low-stringent criteria for all six features (termed as "weak-target introns"); and (3) the remaining introns (termed as "nontarget introns"; Tables S1 and S2). We checked whether these criteria could clearly identify downregulated introns by plotting the FC-FC value for each intron in these three groups (Figure 2I). Fourteen of 17 strong-target introns (82%) were included in the severely downregulated intron group (defined above and in Figure 2A), indicating that the high-stringent criteria successfully enriched downregulated introns under *Sfpq*-KD conditions.

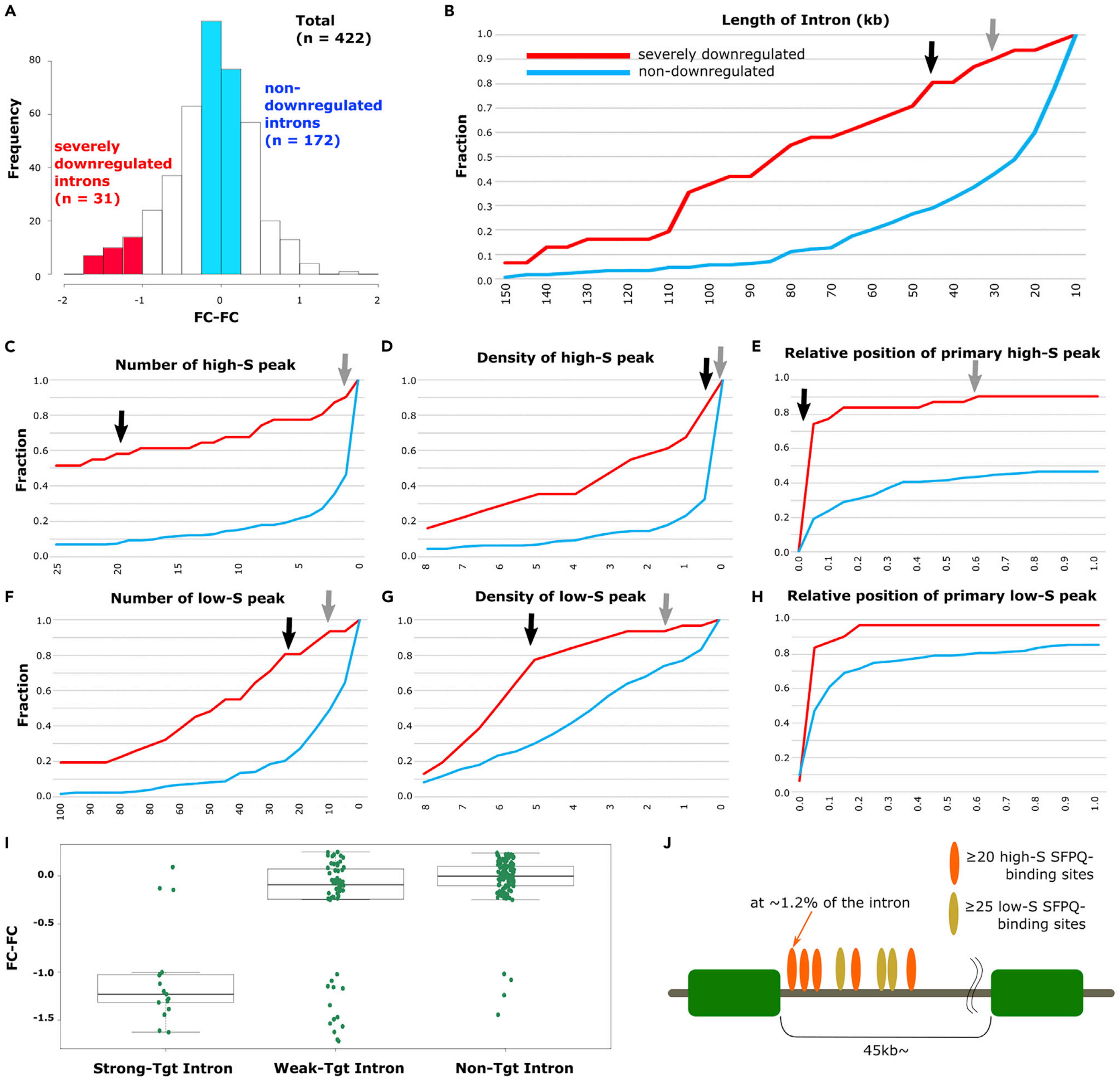


Figure 2. Identification of Factors Defining SFPQ Target Specificity

(A) Histogram of FC-FC values of introns in downregulated genes and having lengths ≥20 kb. In total, 31 severely downregulated introns (FC-FC values ≤ −1, indicated with red) and 172 non-downregulated introns (−0.25 < FC-FC < +0.25, indicated with blue) were compared.

(B–H) Cumulative plots for intron length (B), number of high-S peaks (C), density of high-S peaks (counts per 10 kb) (D), relative position of primary high-S peaks (E), number of low-S peaks (F), density of low-S peaks (counts per 10 kb) (G), and relative position of low-S peaks (H). Black and gray arrows indicate the values for the high-stringent and low-stringent criteria, respectively.

(I) Boxplot of FC-FC values of introns fulfilling all of the high-stringent criteria, introns fulfilling all of the low-stringent criteria, and the remaining introns. Upper, middle, and lower lines indicate the first quartile, median, and third quartile, respectively. Whiskers show the furthest point within 1.5 interquartile ranges (IQRs) from the upper and lower quartiles.

(J) Graphic abstract summarizing the identified features for strong SFPQ target introns. Detailed criteria are shown in Table 1.

See also Figure S1 and S3 and Tables S1 and S2.

High Stringent Criteria	# Of Intron Passed	Fraction of Severely Downregulated Intron	Fraction of Non-downregulated Intron	Difference of Fractions	Low Stringent Criteria	# Of Intron Passed	Fraction of Severely Downregulated Introns
Intron length ≥45kb	25	0.807	0.291	0.516	Intron length ≥30kb	28	0.903
Count of High-S peaks ≥20	18	0.581	0.076	0.505	Count of High-S peaks ≥1	28	0.903
Density of High-S peaks (counts/10kb) ≥ 0.5	26	0.839	0.326	0.513	Density of High-S peaks (counts/10kb) ≥ 0.03	28	0.903
Relative position of primary High-S peak ≤0.0128	15	0.484	0.052	0.432	Relative position of primary High-S peak ≤0.6	28	0.903
Count of Low-S peaks ≥25	25	0.807	0.204	0.603	Count of Low-S peaks ≥10	29	0.936
Density of Low-S peaks (counts/10kb) ≥ 5	24	0.774	0.302	0.472	Density of Low-S peaks (counts/10kb) ≥ 1.5	29	0.936

Table 1. Definition of High- and Low-Stringent Criteria Used for Defining Strong- and Weak-Target Introns of SFPQ and the Number of Introns Passing Each Criterion

Introns fulfilling all six high-stringent criteria were designated as "strong-target introns" and introns not classified into strong-target introns but fulfilling low-stringent criteria for all six features were termed as "weak-target introns."

Moreover, 13 of 56 weak-target introns (23%) were included in the severely downregulated intron group. Enrichment of SFPQ target introns in the weak-target intron group was lower than that in the strong-target intron group, but some severely downregulated introns that did not meet the strong-target criteria were included. The nontarget introns contained only four severely downregulated introns (3%) out of 130, indicating that the low-stringent criteria efficiently segregated SFPQ nontarget introns (Table S1). We also assessed the relationships between SFPQ binding close to the 5′-end of the intron and total SFPQ binding within introns. The relative High-S binding peaks position close to 5′ introns and total SFPQ binding density within introns were examined for all 2,509 introns ≥30 kb in length within the expressed genes (Figure S3). The presence of upstream SFPQ binding, with relative position in the total intron length within 5% from its 5′ end, was significantly correlated with elevated SFPQ-binding density within entire introns, consistent with a model showing that primary SFPQ binding close to the 5′ end of the intron causes serial SFPQ binding on entire introns by polymerization (Huang et al., 2018; Lee et al., 2015; Takeuchi et al., 2018).

In summary, the identified features that defined SFPQ target introns by high-stringent criteria were as follows: long introns ≥45 kb, total High-S SFPQ binding peaks in introns ≥20 with a density ≥0.5 per 10 kb intron length, total Low-S binding peaks ≥25 with a density ≥5 per 10 kb intron length, and primary High-S SFPQ binding within 1.28% of the relative position in the total intron length. These identified criteria are schematically summarized in Figure 2J. A typical example of SFPQ target introns could be observed in the shorter gene *Cbx7*. The length of *Cbx7* pre-mRNA was 55 kb, which was much shorter than 100 kb; however, this gene had a strong-target intron with a length of 47 kb (>45 kb). Loss of *Sfpq* caused a gradual decrease of *Cbx7* pre-mRNA within this intronic region and subsequently decreased its mature mRNA expression (Figure S4). Thus, our multimodal analysis combining Ribo(−) RNA-seq and CLIP-seq data successfully identified the features of SFPQ target introns.

The Identified Features Could Segregate Significant SFPQ Binding to Regulate the Transcriptional Elongation of Long Genes

The features of functional SFPQ binding in target introns were identified using a limited number of target and nontarget genes by integrating CLIP-seq and transcriptome data. Therefore, we next tested whether the identified features could be expanded to detect target introns from entire transcripts using CLIP-seq data. According to the

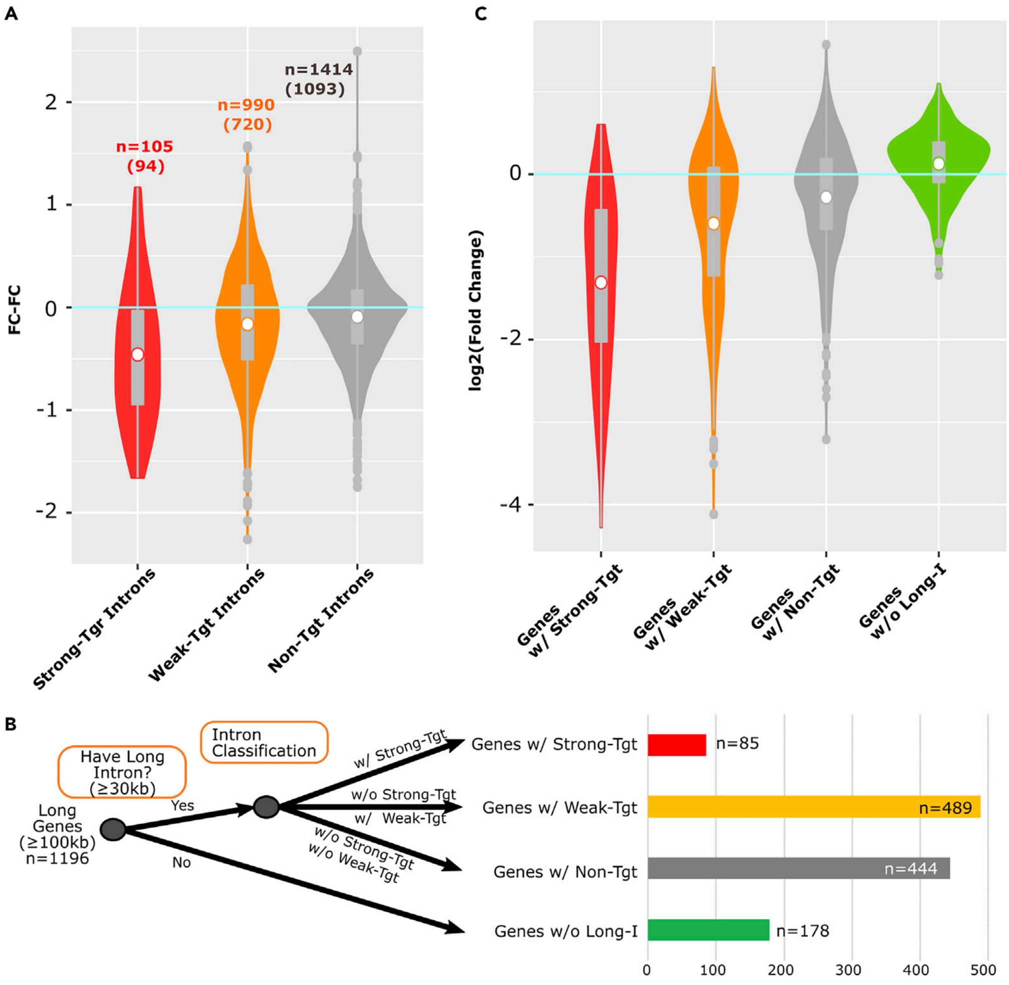


Figure 3. Identified Features Segregated Functional SFPQ Binding to Regulate Transcriptional Elongation of Long Genes

(A) Violin plots for FC-FC values of all expressed genes ≥30 kb, which were classified according to the identified SFPQ target intron criteria. Numbers in brackets shows intron counts in long genes (≥100 kb).

(B) Flowchart for long gene classification results by intron length and SFPQ target intron criteria. Genes for which pre-mRNA lengths ≥100 kb were classified into four groups: genes with strong-target introns (Genes w/Strong-Tgt), genes with weak-target introns (Genes w/Weak-Tgt), genes with non-target introns (Genes w/Non-Tgt), and genes without long introns (Genes w/o Long-Intron).

(C) Violin plots for fold changes (si-*Sfpq*/si-Cont) of classified long genes.

See also Figures S1 and S4 and Table S3.

identified seven features with high and low stringencies, we sorted all introns (n = 2,509) in all expressed genes into strong-target, weak-target, and nontarget intron groups. When the FC-FC value was examined, strong-target introns showed significantly lower FC-FC values than weak-target or nontarget introns, indicating successful separation of SFPQ target introns from all expressed genes (Figure 3A). Then, we further examined whether identification of the SFPQ target intron could be expanded to predict SFPQ target genes. First, we classified the expressed long genes (≥100 kb) according to the presence of long introns (≥30 kb) as a minimum criterion for SFPQ target introns. Among 1,196 expressed long genes (≥100 kb), 1,018 genes (85.1%) possessed long introns (≥30 kb), indicating a positive correlation between total gene length and long introns. Genes with more than one long intron were further classified into three groups according to the presence of strong-target, weak-target, and nontarget introns (Figures 3B and Table S3). We examined changes in the expression levels of these gene groups under *Sfpq*-KD conditions. As expected, genes possessing strong-target introns showed the lowest mean FCs under *Sfpq*-KD conditions (Figure 3C, red plot), and the mean FCs of genes possessing weak-target and nontarget introns increased stepwise (Figure 3C, orange and gray plots, respectively). In addition, the mean FC of long genes without long introns was 0.13 (linear value: 1.09), indicating that these genes were not downregulated under *Sfpq*-KD conditions (Figure 3C, green plot). Taken together, these results indicated that multilateral bioinformatics analyses could successfully identify SFPQ target genes that were defined by SFPQ target introns.

Sequence Feature Analysis of SFPQ-Binding Sites Suggested an Association between SFPQ Binding and Upstream Exon Recognition

Next, we attempted to identify the molecular mechanisms through which functional SFPQ binding formed to regulate its specific target introns. Our results indicated that SFPQ binding close to the 5′ intronic region was essential for subsequent serial binding of SFPQ toward the 3′ end (Figures 2E and S3). We expected that 5′ regions of SFPQ target introns, particularly primary binding sites close to the 5′ end, should have stronger and/or unique consensus SFPQ-binding sequences; therefore, we attempted to identify specific features of primary binding sites. SFPQ-binding sites in strong-target, weak-target, and nontarget introns were classified according to their relative binding portions in introns as follows. The first binding site located within the upstream region (5% of the total intron length from the 5′ intron end) was classified into the first (≤5%) group. The remaining sites were classified according to relative position within the introns, i.e., within 10% region (≤10%), within 10%–50% region (≤50%), and the remaining regions (≤100%). This classification is schematically outlined in Figure S5A.

Our previous study identified C-rich and AG-rich sequences as cis-elements of SFPQ (Takeuchi et al., 2018). Accordingly, we next examined the presence of these motifs. CU-rich sequences with a core UCC sequence had significantly higher occupancy exclusively in the first and ≤5% SFPQ binding sites of strong-target introns compared with remaining regions of strong-target introns and compared with the same regions in weak-target and nontarget introns (Figure 4A), whereas AG-rich sequences were not (Figure S5B). Next, we searched for another consensus sequence that was specific to the first position of strong- or weak-target introns. Then, we identified the C-rich sequence that was significantly enriched in the first and ≤10% SFPQ-binding sites of strong-target introns compared with the remaining regions of strong-target introns and the same regions of weak-target introns (Figure 4B). These results indicated that CU- or C-rich sequences contributed to the formation of SFPQ binding close to the 5′ intronic region. This and previous studies have demonstrated that SFPQ binds to target mRNAs using rather non-strict-binding sequences. Accordingly, we then attempted to identify additional factors that regulate the target specificities of SFPQ. Because previous studies have demonstrated the interactions between SFPQ and splicing factors (Patton et al., 1993; Snijders et al., 2015; Talukdar et al., 2011), we performed quantitative analyses focusing on sequences related to splice site recognition, i.e., intronic splicing enhancer/silencer sequences (ESEs, ISEs, ESSs, and ISSs). SFPQ target introns tended to possess significantly more ESEs, ESSs, and ISEs but fewer ISSs per 100 bases (Figures 4C–4F). Significant numbers of ESE, ESS, and ISE sequences adjacent to the 5′ splice sites in strong-target introns suggested that exon recognition of upstream target introns was regulated by several splicing factors and that target recognition of SFPQ close to 5′ splice sites was conducted via association. The mechanistic insight of SFPQ target recognition at 5′ splice sites and sequential binding on target introns is schematically summarized in Figure 4G.

Prediction of SFPQ Regulatory Target Genes by Identifying Functional SFPQ Binding

Our previous study indicated that SFPQ target genes in mouse embryonic brains are highly enriched in gene ontology (GO) categories related to nervous system development, specifically axonal guidance, dendrite projections, and neuronal projections (Takeuchi et al., 2018). This previous result supported the idea that RBPs comprehensively regulate clusters of genes having specific functions and that these functional gene clusters are called regulons. In this study, using all expressed genes sorted according to the presence of strong-target, weak-target, and nontarget introns (Figure 3B and Table S3), we conducted GO analysis. As a result, GO categories related to neural development, including neuron projection morphogenesis, cell morphology involved in neuro differentiation, and axonogenesis, were identified as enriched terms in genes with strong- and weak-target introns (Figure 5 and Table S4). In our previous study, 176 genes were identified as SFPQ targets in differentiated Neuro2a cells (Figure 5, predetermined target genes in N2a), and GO categories related to neural development were not identified, presumably because brain-specific gene expression was much weaker *in vitro* than *in vivo*, which hindered SFPQ target gene identification. However, using the combined criteria, we detected only 85 genes harboring strong-target introns of SFPQ and 489 genes harboring weak-target introns, and these genes were highly enriched in GO categories related to neural development. These results indicated that SFPQ comprehensively regulated essential genes involved in neural development and that prediction of SFPQ target genes from the identified functional SFPQ-binding events detected unique regulons.

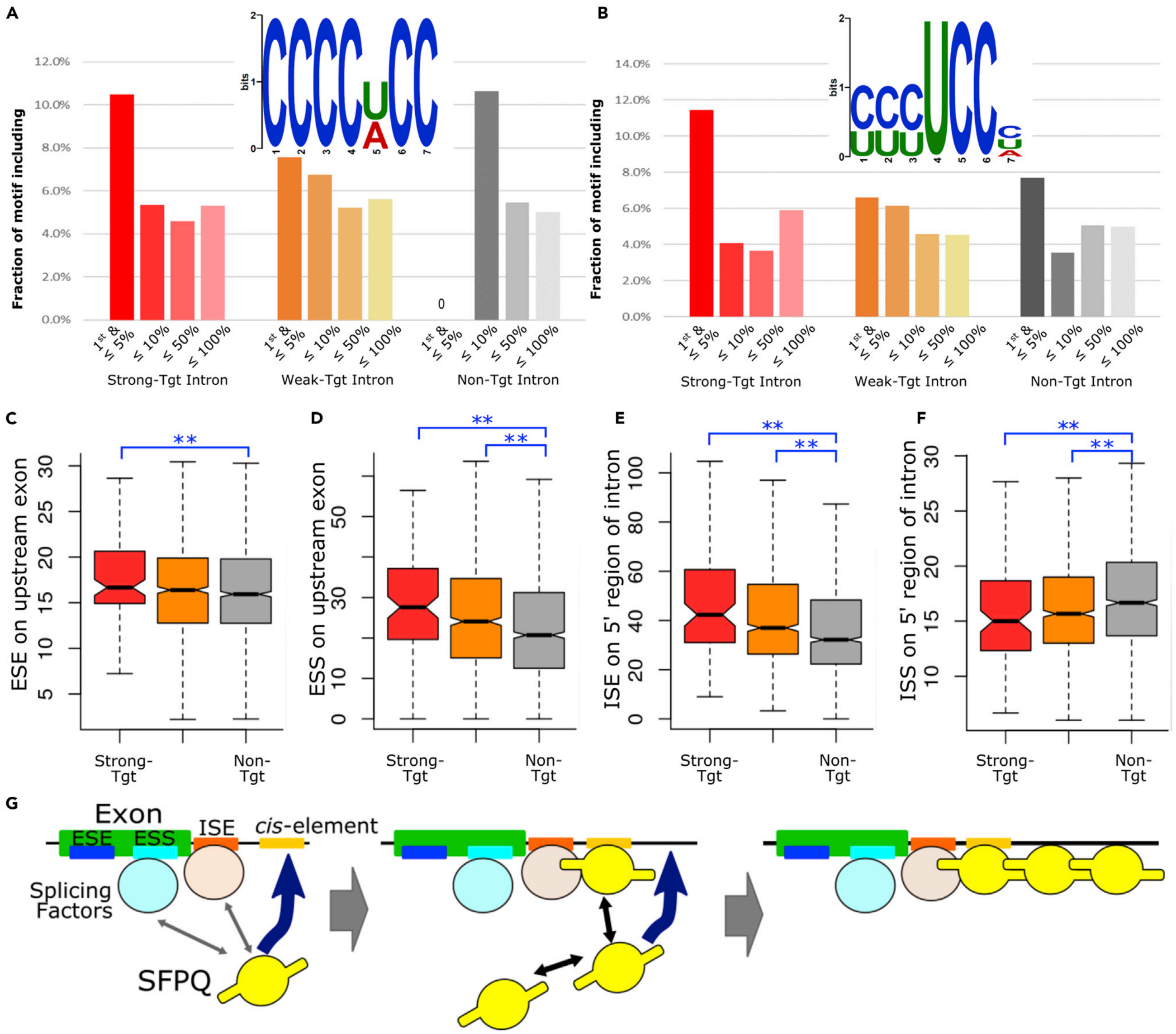


Figure 4. Sequence feature analysis of SFPQ-binding sites suggested an association between SFPQ binding and upstream exon recognition

(A and B) Bar graphs showing the fraction of SFPQ target introns possessing SFPQ binding motifs. The C-rich motif found in our previous study (A) and the CU-rich motif newly identified in primary SFPQ-binding sites among strong- or weak-target introns (B).

(C–F) Box plots showing the numbers of splicing enhancer and silencer sequences in upstream exons of SFPQ target introns. Counts of ESE (C), ESS (D), ISE (E), and ISS (F) were determined. $*p < 0.05$, $**p < 0.01$. The box plot format is similar to that in Figure 2I.

(G) Working model for SFPQ binding to its target intron.

See also Figures S1 and S5.

SFPQ expression is spatially and temporally controlled in developing brains (Takeuchi et al., 2018). This observation implies that the presence or absence of SFPQ can modulate SFPQ target gene expression. Thus, we examined the correlations between SFPQ and predicted SFPQ target genes using public proteome data in human brains (Carlyle et al., 2017). First, we compared SFPQ protein expression in different brain regions, including the dorsolateral prefrontal cortex (DFC), primary visual cortex (V1C), hippocampus (HIP), amygdala (AMY), mediodorsal thalamic nucleus (MD), striatum (STR), and cerebellar cortex (CBC), in brain samples from patients aged 0–60 years. The CBC showed the highest expression of SFPQ, and an age-dependent decrease in SFPQ expression was observed (Figure 6A). Among different brain regions, the expression levels of 492 proteins were positively correlated with SFPQ protein levels (Pearson correlation coefficient ≥0.5, Table S5). We further analyzed whether these positively correlated genes possessed

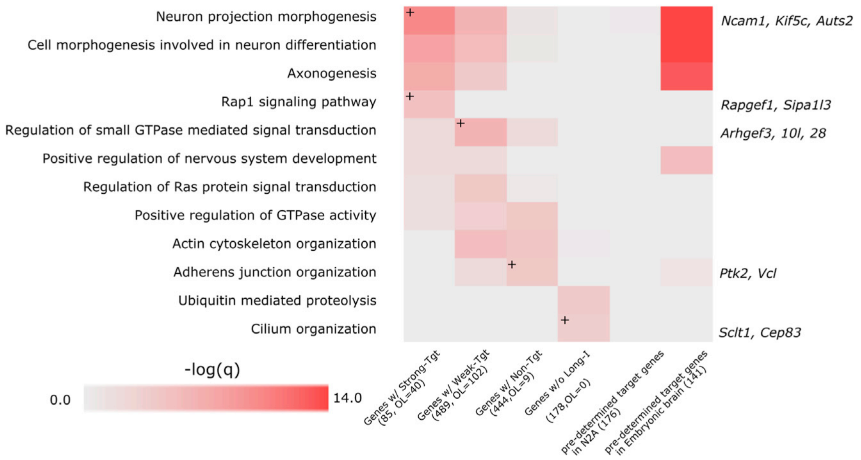


Figure 5. Prediction of SFPQ Regulatory Target Genes via Identification of Functional SFPQ Binding in Neuro2a Cells

Heatmap showing *q*-values of enriched GO terms for SFPQ target long genes and predetermined SFPQ target genes in Neuro2a (N2a) cells and embryonic mouse brain tissues. Numbers in parentheses show the count of genes belonging to each gene group and the counts of genes showing overlap (OL) with predetermined SFPQ target genes in Neuro2a cells. See also Table S4.

SFPQ target introns in mouse homologous genes. Genes possessing strong-target introns showed the highest enrichment in positively correlated genes, followed by genes with weak-target introns, although these enrichments were not statistically significant (Figure 6B). Genes with nontarget introns and genes without long introns did not show enrichment.

To further identify relationships between SFPQ protein levels and global long gene expression, we selected brain regions in which SFPQ protein levels were relatively high (SFPQ-high) or low (SFPQ-low). In total, 7 samples had higher SFPQ expression (>25 PP30K), and 14 have lower SFPQ expression (7 PP30K). As expected, genes possessing strong-target introns showed elevated expression in SFPQ-high samples compared with that in SFPQ-low samples (Figure 6C). Protein expression from genes with weak-target introns was not different between these two groups, and genes with nontarget introns showed slightly decreased protein levels in SFPQ-high samples. *Fto* is a long gene with one strong-target intron (intron 8) and one weak-target intron (intron 1; Figure S6). SFPQ and FTO protein levels were highly positively correlated (PCC = 0.9; Figure 6D). These results indicated that the expression of genes possessing strong-target introns in the human brain was significantly correlated with SFPQ expression. Taken together, our multilateral bioinformatics analyses identified functional SFPQ binding and facilitated prediction of SFPQ target genes.

DISCUSSION

By integrating CLIP-seq with transcriptome data and carrying out multilateral bioinformatics analyses, we successfully segregated functional SFPQ binding from total SFPQ deposition on a large proportion of expressed pre-mRNAs. We found that SFPQ binding close to 5′ splice sites of introns was a crucial factor distinguishing target and nontarget introns and could regulate mRNA expression of long introns (≥30 kb) (Figures 2 and 3). The presence of High-S SFPQ binding close to the 5′ intron within 1.2% of the total intron length provided clear segregation between severely downregulated introns and non-downregulated introns (Figure 2E) In addition, High-S SFPQ binding close to 5′ intron within ~5% of the total intron length increased total SFPQ binding in the entire intron (Figure S3), indicating that SFPQ binding close to exon-intron junctions was critical for regulating downstream introns. These results were consistent with findings in a structural study of SFPQ showing that SFPQ deposition at high-affinity binding sites acted as the seed for multimerization of SFPQ via SFPQ-SFPQ binding on target nucleic acids using its coiled-coil domains (Lee et al., 2015). From these observations, we proposed a model in which High-S SFPQ binding peaks close to the 5′ intron end facilitated SFPQ recruitment toward downstream pre-mRNAs, which caused co-transcriptional SFPQ binding to entire long introns and serially activated transcriptional elongation (Takeuchi et al., 2018). In addition to High-S peaks, we also assessed the effects of Low-S SFPQ binding peaks near 5′ splice sites. However, these peaks did not segregate between severely downregulated and

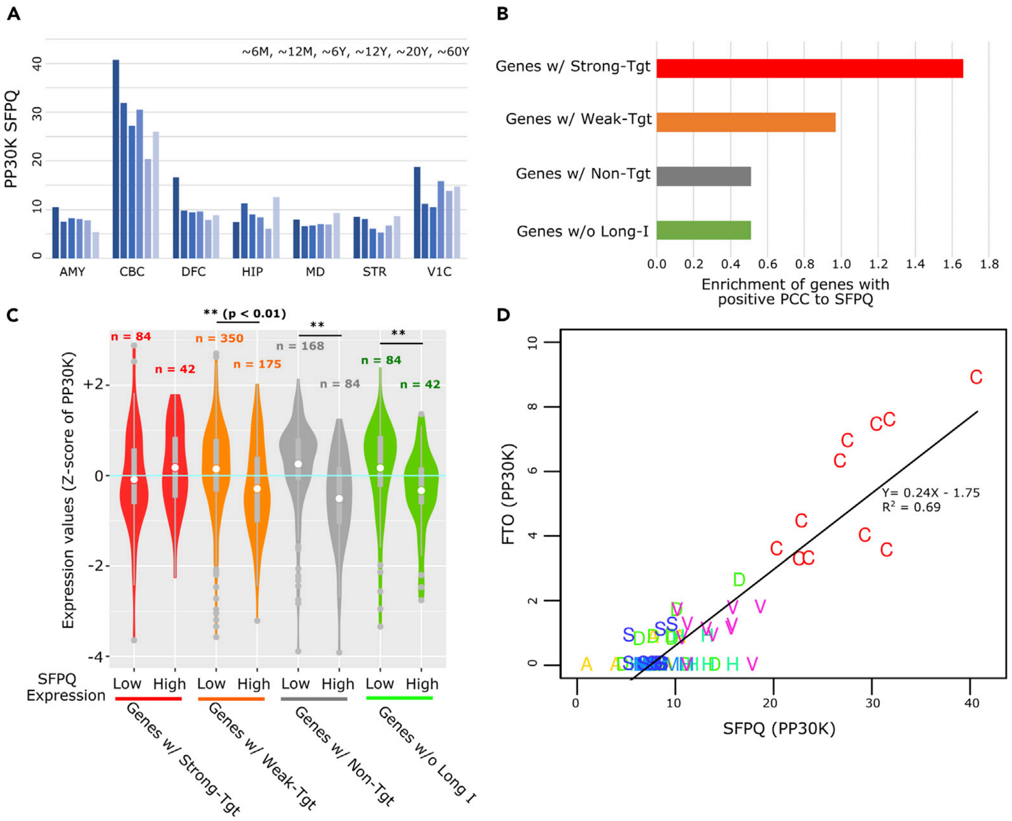


Figure 6. Prediction of SFPQ Regulatory Target Genes in Human Brains

(A) Protein expression levels of SFPQ in the indicated brain regions at different ages (peptides per 30 million peptides [PP30M]). Abbreviations of postmortem brain regions: dorsolateral prefrontal cortex (DFC), primary visual cortex (V1C), hippocampus (HIP), amygdala (AMY), mediodorsal thalamic nucleus (MD), striatum (STR), and cerebellar cortex (CBC).

(B) Over-representation analysis between classified long genes and genes positively correlated with SFPQ (PCC $\geq$0.5). Long genes ($\geq$100 kb) were classified into four groups according to the presence of SFPQ target introns, similar to that in Figure 3B.

(C) Violin plots for protein expression levels from long genes. Five postmortem brain samples with low SFPQ expression and another five samples with high SFPQ expression were analyzed.

(D) Scatterplot showing the correlation between SFPQ (x) and FTO (y) protein expression levels. Abbreviations: DFC (D); V1C (V); HIP (H); AMY (A); MD (M); STR (S); CBC (C).

See also Figure S6 and Table S5.

non-downregulated introns (Figure 2H). These results indicated that selective SFPQ binding in the 5′-end of introns detected as high-S peaks were significantly involved in the formation of functional binding of SFPQ.

SFPQ binding sites close to the 5′ splice sites were enriched with nonstrict CU- or C-rich sequences (Figures 4A, 4B, and S5B). We found that enrichment of ESEs, ESSs, and ISEs within 300 bp of 5′ splice sites in SFPQ target introns was compared with that in nontarget introns (Figures 4C–4F). These results indicated that functional binding of SFPQ and target recognition was facilitated by splicing factors through exon recognition (Figure 4G). To unravel the mechanism of target recognition by SFPQ, we analyzed the presence of consensus sequences of different RBPs in the SFPQ target introns and found that binding sites of hnRNP-H1/2/3, -F, and -K were enriched (data not shown). Because these hnRNPs have been reported to interact with SFPQ (Talukdar et al., 2011; Yamamoto et al., 2016), it is possible that these RBPs facilitate SFPQ recruitment to 5′ splice sites. For proteins that interact with SFPQ, it is necessary to consider its family members NONO (non-POU domain-containing octamer binding protein) and PSPC1 (paraspeckle component 1). SFPQ and these proteins belong to the *Drosophila* behavior/human splicing (DBHS) family (Knott et al., 2016), and NONO and PSPC1 are known to interact with SFPQ and each other to form heterodimers (Huang et al., 2018, 2020; Knott et al., 2016; Lee et al., 2015). Thus, it is also possible that DBHS family members contribute to the target specification and co-transcriptional binding of SFPQ. Further studies to identify

RBPs that share target RNAs or binding sites could also improve our understanding of cooperative regulation of RBPs for target specification. Although our study successfully identified and extracted functional bindings of SFPQ among entire deposition, mechanism of the sequence and structure dependent-recognition of RBPs to target RNAs remained to be elucidated. Several studies provided useful computational methods. For example, CapR surveyed specific secondary structures of target RNAs based on CLIP-seq data (Fukunaga et al., 2014). A web-tool beRPB estimates target RNAs based on single or multiple cis-elements (Yu et al., 2018). Further studies using existing computational method would help our understanding about the molecular mechanism for target RNA recognition of extracted functional bindings of RBPs.

In a previous study, SFPQ target genes were identified by the combination of substantial SFPQ binding on pre-mRNAs and downregulation of genes under SFPQ-KO or KD conditions (FC < 1/3). In this study, we integrated CLIP-seq and Ribo(−) RNA-seq data, identified functional SFPQ binding from SFPQ target introns, and used criteria to determine SFPQ target genes (Figures 3A and 3C). Indeed, 45 of 85 strong-SFPQ target genes were newly identified based on the current criteria and were subjected to GO enrichment evaluation. For example, the GO term "neuron projection morphogenesis" contained 22 SFPQ strong-target genes, 15 of which, including *Kif5c*, *Auts*, and *Kalrn*, were newly identified (Figure 5 and Table S4). The downregulation of these genes in *Sfpq*-KD Neuro2A cells was moderate (their $\log_2$(FC) values ranged from 0.92 to 0.03), and this could explain the difference in target genes. Because SFPQ target genes possessing strong-target introns showed high enrichment in nervous system development in GO analysis (Figure 5), we concluded that RBPs regulated specific functional gene clusters and that the methods developed in this study could successfully capture SFPQ regulons. In proteome analysis in human brains, SFPQ protein expression and the protein expression levels of genes possessing SFPQ target introns were significantly correlated (Figure 6), demonstrating that functional SFPQ binding enabled prediction of SFPQ target genes. We identified the m^6A eraser gene FTO as the gene with the highest PCC among SFPQ strong-target genes (Figure 6D). Moreover, because FTO has been reported to bind SFPQ and work together with SFPQ to eliminate m^6A modification (Song et al., 2020), identification of SFPQ regulatory target genes may further enable identification of SFPQ interacting partners. Furthermore, SFPQ showed relatively high expression levels in spatiotemporally limited regions, such as the young cerebral cortex (Figures 6B and 6C), and SFPQ strong-target genes, i.e., *CAMK1D*, *KALRN*, and *FMNL2*, which are essential genes for neuron development, showed co-expression patterns with SFPQ, with high PCC values of ≥ 0.5 (Table S5). Thus, our results indicated that such a specific transcriptome would be formed via RBPs, including SFPQ, and that prediction of RBP target genes could greatly improve our understanding of gene regulatory networks, as demonstrated for nervous system development.

Past studies demonstrated SFPQ's multi-functions to regulate mRNA splicing, microRNA targeting via 3′ UTR, or DNA repair (Knott et al., 2016; Yarosh et al., 2015). Although we had assessed these functions in brains and differentiating neuronal cells, it was unlikely that loss of *Sfpq* directly caused alternative splicing changes, modulated microRNA (miRNA) targeting of mRNAs with long 3′ UTRs, or disrupted DNA repair (Takeuchi et al., 2018). Thus, our observation indicates that a major function of SFPQ in developing brains is to facilitate transcriptional elongation and loss of *Sfpq* caused downregulation of extra-long genes. As the supportive evidence, among genes having ≥ 1 High-S peak and ≥ 32 Low-S peaks, only one gene was upregulated in *Sfpq*-KD Neuro2A cells. The gene, *Cdk19* (*Cdk11* in old name), had 134 kb length with two long introns >40 kb but these long introns did not pass the criteria of SFPQ target introns (Figure S7).

In conclusion, using an integrated data science approach, we provided a solution for identifying the features of functionally essential binding of RBPs among entire deposition on expressed mRNAs. In addition, we showed that these features could be used for predicting regulatory target genes. This approach also provided insights into the molecular mechanisms through which target recognition of RBPs could be formed, i.e., not merely depending on consensus sequences but also through interactions with other RBPs during mRNA processing. Thus, we propose that the multilateral bioinformatics approach employed in this study could be applied as a standard analytical approach in RNA biology to identify functional RBP binding and determine the biological roles of these proteins that lead to understand complicated regulative nature of RBPs on RNA molecules and regulatory network of each RBPs; regulon.

Limitation of the Study

Elucidating precise molecular mechanisms of RBP's target recognition further required determination of the binding and working co-partners. For this purpose, it is highly anticipated to accumulate the information of functional bindings of other RBPs.

Resource Availability

Lead Contact

Further information and requests for resources should be addressed to the Lead Contact, Akihide Takeuchi (takeuchi.akihide.8r@kyoto-u.ac.jp)

Materials Availability

N/A.

Data and Code Availability

The raw data for CLIP-seq, RNA-seq, pol II ChIP-seq were deposited and available in the NCBI Gene Expression Omnibus (GEO) with accession number GSE60246. Computer programs used in this study can be available at GitHub; https://github.com/keiiida.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101325.

ACKNOWLEDGMENTS

We thank Dr. Motoyasu Hosokawa for helpful comments and discussions.

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) from the Ministry of Education, Culture, Sports, Science and Technology of Japan [MEXT, JSPS KAKENHI 19500269, 25500288, 21249013, 15H05721, 19K06907, 20K07310 to M.H., A.T., and K.I.]; Innovative Cell Biology by Innovating Technology [Cell Innovation to M.H. and A.T.]; a Core Research for Evolutional Science and Technology grant from the Japan Science and Technology Agency [to M.H.]; a grant from the Japan Agency for Medical Research and Development [to M.H.]; the Asian CORE Program of JSPS [to M.H.]; iCeMS Cross-Disciplinary Research Promotion Project of Kyoto University [to A.T.]; and the Fujiwara Memorial Foundation [to A.T.].

AUTHOR CONTRIBUTIONS

K.I. conceived the project, developed the bioinformatics analytical method, and performed analysis. A.T. conceived the project and contributed to the bioinformatics analyses. All authors contributed to the writing of the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 6, 2020
Revised: June 8, 2020
Accepted: June 24, 2020
Published: July 24, 2020

REFERENCES

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.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M., et al. (2012). Insights into RNA biology from an Atlas of Mammalian mRNA-binding proteins. Cell *149*, 1393–1406.

Fu, X.-D., and Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. Nat. Rev. Genet. *15*, 689–701.

Fukunaga, T., Ozaki, H., Terai, G., Asai, K., Iwasaki, W., and Kiryu, H. (2014). CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data. Genome Biol. *15*, R16.

Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. Nat. Rev. Genet. *15*, 829–845.

Hentze, M.W., Castello, A., Schwarzl, T., and Preiss, T. (2018). A brave new world of RNA-binding proteins. Nat. Rev. Mol. Cell Biol. *19*, 327–341.

Hosokawa, M., Takeuchi, A., Tanihata, J., Iida, K., Takeda, S., and Hagiwara, M. (2019). Loss of RNA-binding protein Sfpq causes long-gene transcriptopathy in skeletal muscle and severe muscle mass reduction with metabolic myopathy. iScience *13*, 229–242.

Huang, J., Casas Garcia, G.P., Perugini, M.A., Fox, A.H., Bond, C.S., and Lee, M. (2018). Crystal structure of a SFPQ/PSPC1 heterodimer provides insights into preferential heterodimerization of human DBHS family proteins. J. Biol. Chem. *293*, 6593–6602.

Huang, J., Ringuet, M., Whitten, A.E., Caria, S., Lim, Y.W., Badhan, R., Anggono, V., and Lee, M. (2020). Structural basis of the zinc-induced cytoplasmic aggregation of the RNA-binding protein SFPQ. Nucleic Acids Res. *48*, 3356–3365.

Knott, G.J., Bond, C.S., and Fox, A.H. (2016). The DBHS proteins SFPQ, NONO and PSPC1: a multipurpose molecular scaffold. Nucleic Acids Res. *44*, 3989–4004.

Lee, M., Sadowska, A., Bekere, I., Ho, D., Gully, B.S., Lu, Y., Iyer, K.S., Trewhella, J., Fox, A.H., and Bond, C.S. (2015). The structure of human SFPQ reveals a coiled-coil mediated polymer essential for functional aggregation in gene regulation. Nucleic Acids Res. *43*, 3826–3840.

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

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.

Van Nostrand, E.L., Freese, P., Pratt, G.A., Wang, X., Wei, X., Xiao, R., Blue, S.M., Chen, J.-Y., Cody, N.A.L., Dominguez, D., et al. (2018). A large-scale binding and functional map of human RNA binding proteins. BioRxiv, 179648.

Patton, J.G., Porro, E.B., Galceran, J., Tempst, P., and Nadal-Ginard, B. (1993). Cloning and characterization of PSF, a novel pre-mRNA splicing factor. Genes Dev. *7*, 393–406.

Russell, J. (2017). Agile Data Science 2.0: Building Full-Stack Data Analytics Applications with Spark.

Snijders, A.P., Hautbergue, G.M., Bloom, A., Williamson, J.C., Minshull, T.C., Phillips, H.L., Mihaylov, S.R., Gjerde, D.T., Hornby, D.P., Wilson, S.A., et al. (2015). Arginine methylation and citrullination of splicing factor proline- and glutamine-rich (SFPQ/PSF) regulates its association with mRNA. RNA *21*, 347–359.

Song, H., Wang, Y., Wang, R., Zhang, X., Liu, Y., Jia, G., and Chen, P.R. (2020). SFPQ is an FTO-binding protein that facilitates the demethylation substrate preference. Cell Chem. Biol. *27*, 283–291.e6.

Sundararaman, B., Zhan, L., Blue, S.M., Stanton, R., Elkins, K., Olson, S., Wei, X., Van Nostrand, E.L., Pratt, G.A., Huelga, S.C., et al. (2016). Resources for the comprehensive discovery of functional RNA elements. Mol. Cell *61*, 903–913.

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.

Talukdar, I., Sen, S., Urbano, R., Thompson, J., Yates, J.R., and Webster, N.J.G. (2011). hnRNP A1 and hnRNP F modulate the alternative splicing of exon 11 of the insulin receptor gene. PLoS One *6*, e27869.

Weyn-Vanhentenryck, S.M., Mele, A., Yan, Q., Sun, S., Farny, N., Zhang, Z., Xue, C., Herre, M., Silver, P.A., Zhang, M.Q., et al. (2014). HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. Cell Rep. *6*, 1139–1152.

Yamamoto, K., Furukawa, M.T., Fukumura, K., Kawamura, A., Yamada, T., Suzuki, H., Hirose, T., Sakamoto, H., and Inoue, K. (2016). Control of the heat stress-induced alternative splicing of a subset of genes by hnRNP K. Genes Cells *21*, 1006–1014.

Yarosh, C.A., Iacona, J.R., Lutz, C.S., and Lynch, K.W. (2015). PSF.: Nuclear busy-body or nuclear facilitator? Wiley Interdiscip. Rev. RNA *6*, 351–367.

Yu, H., Wang, J., Sheng, Q., Liu, Q., and Shyr, Y. (2018). beRBP: binding estimation for human RNA-binding proteins. Nucleic Acids Res. *47*, e26.

iScience, Volume *23*

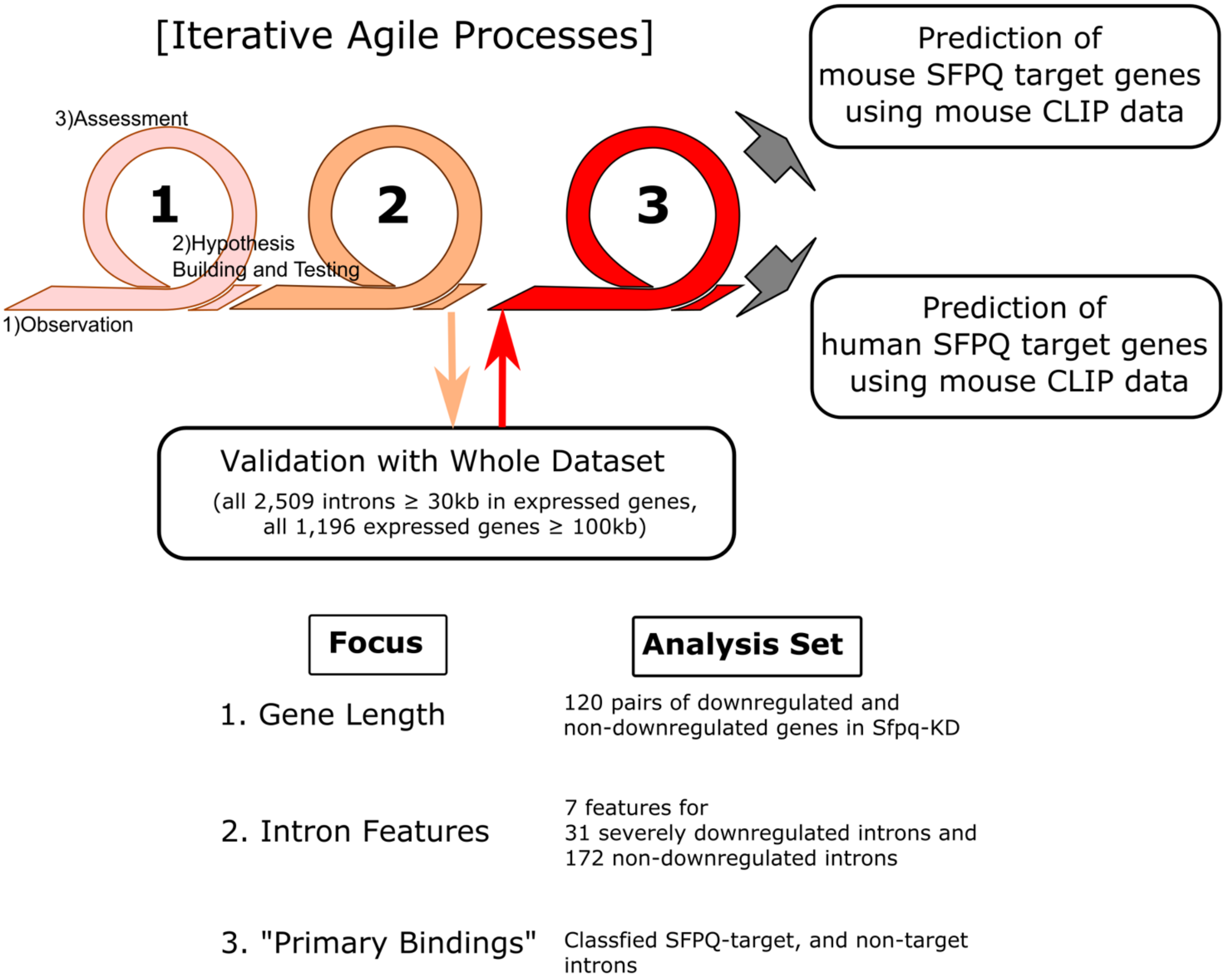
Supplemental Information

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

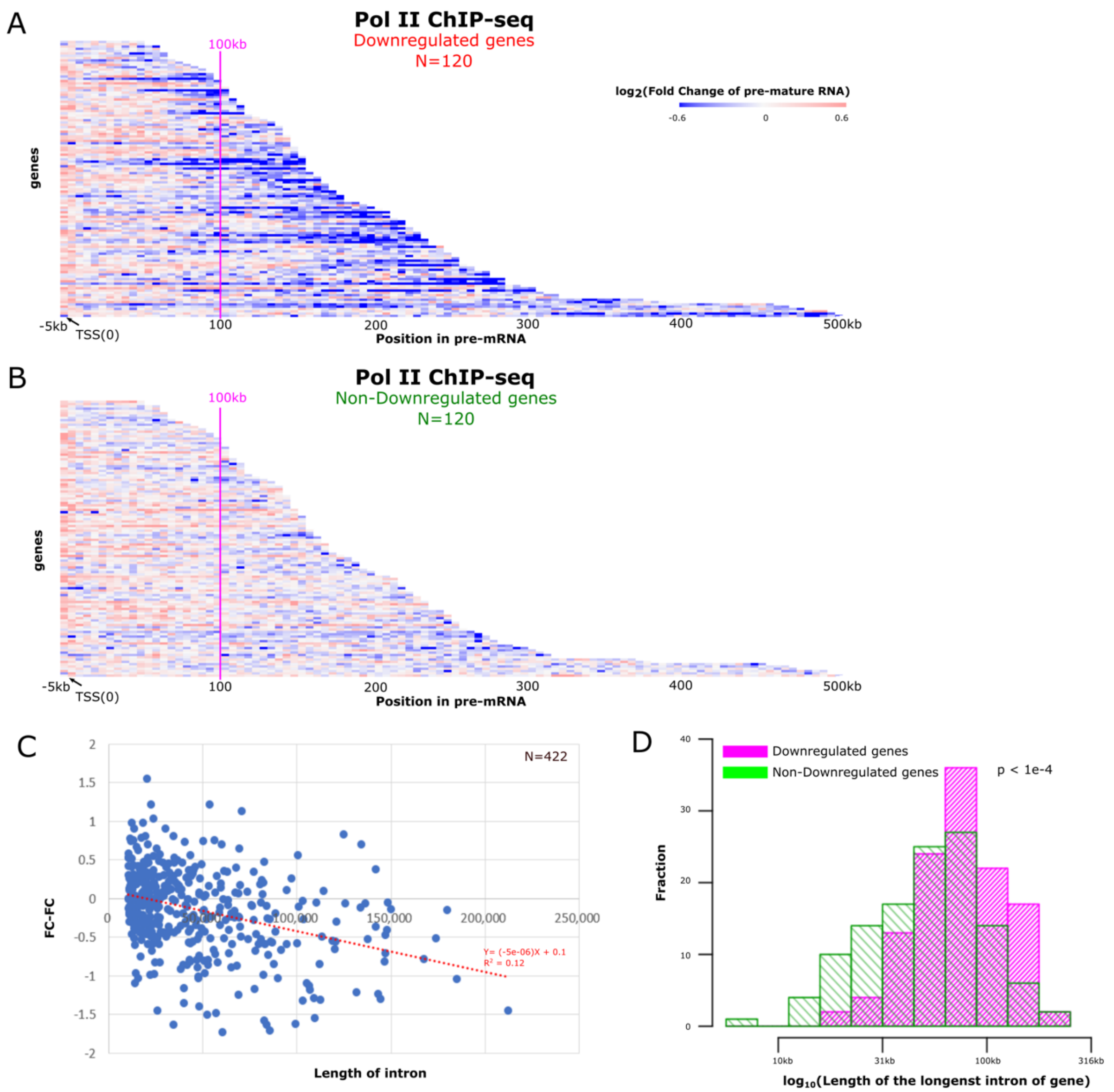
Supplemental Figure S1



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

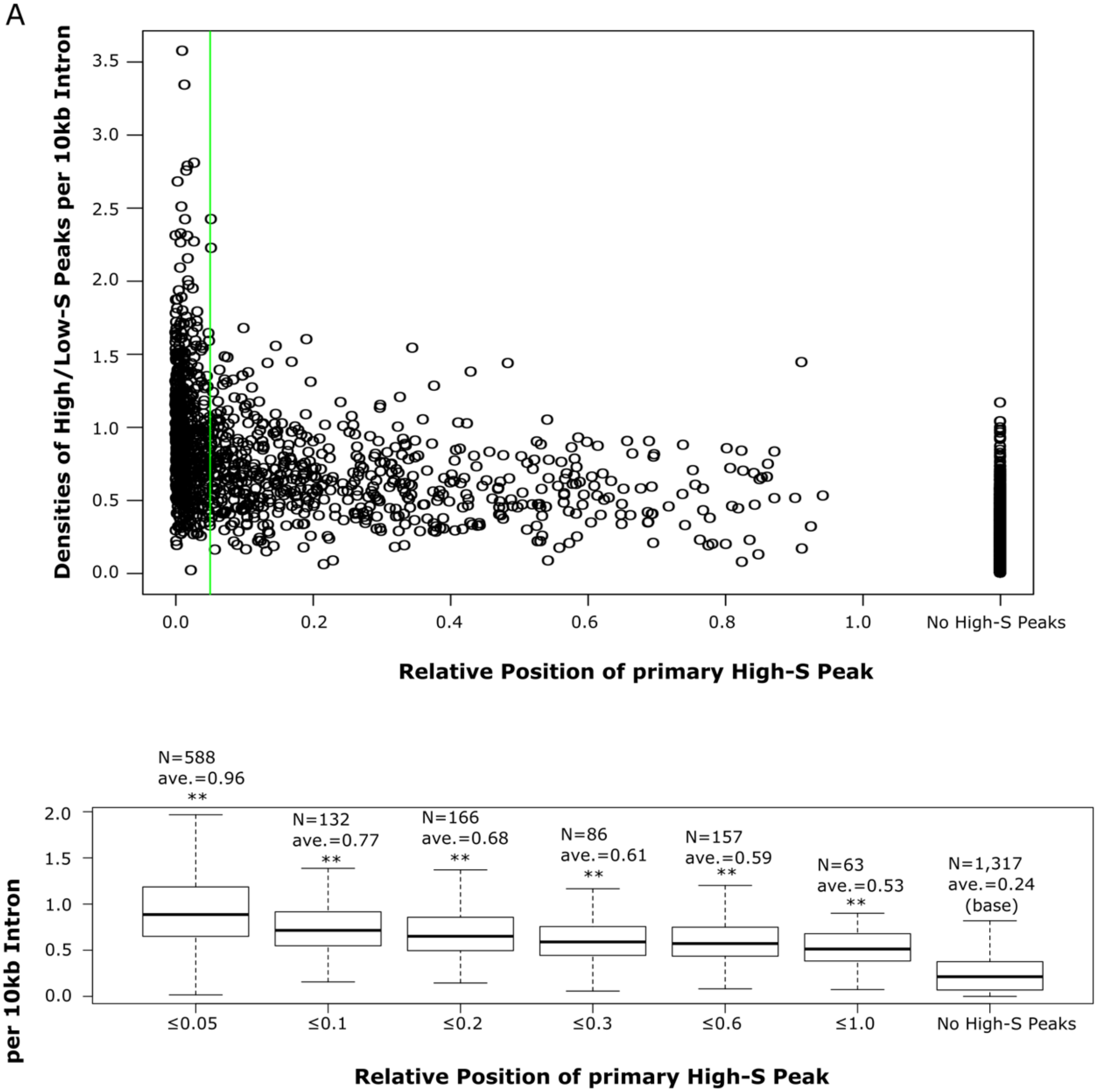
Supplemental Figure S2



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.

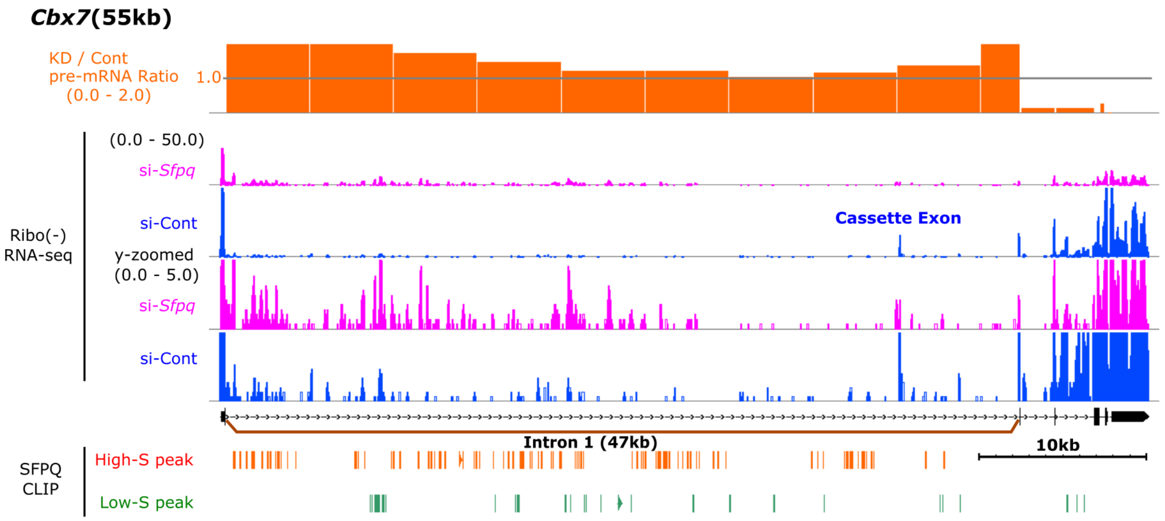
Supplemental Figure S3



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-10, double asterisks were shown.

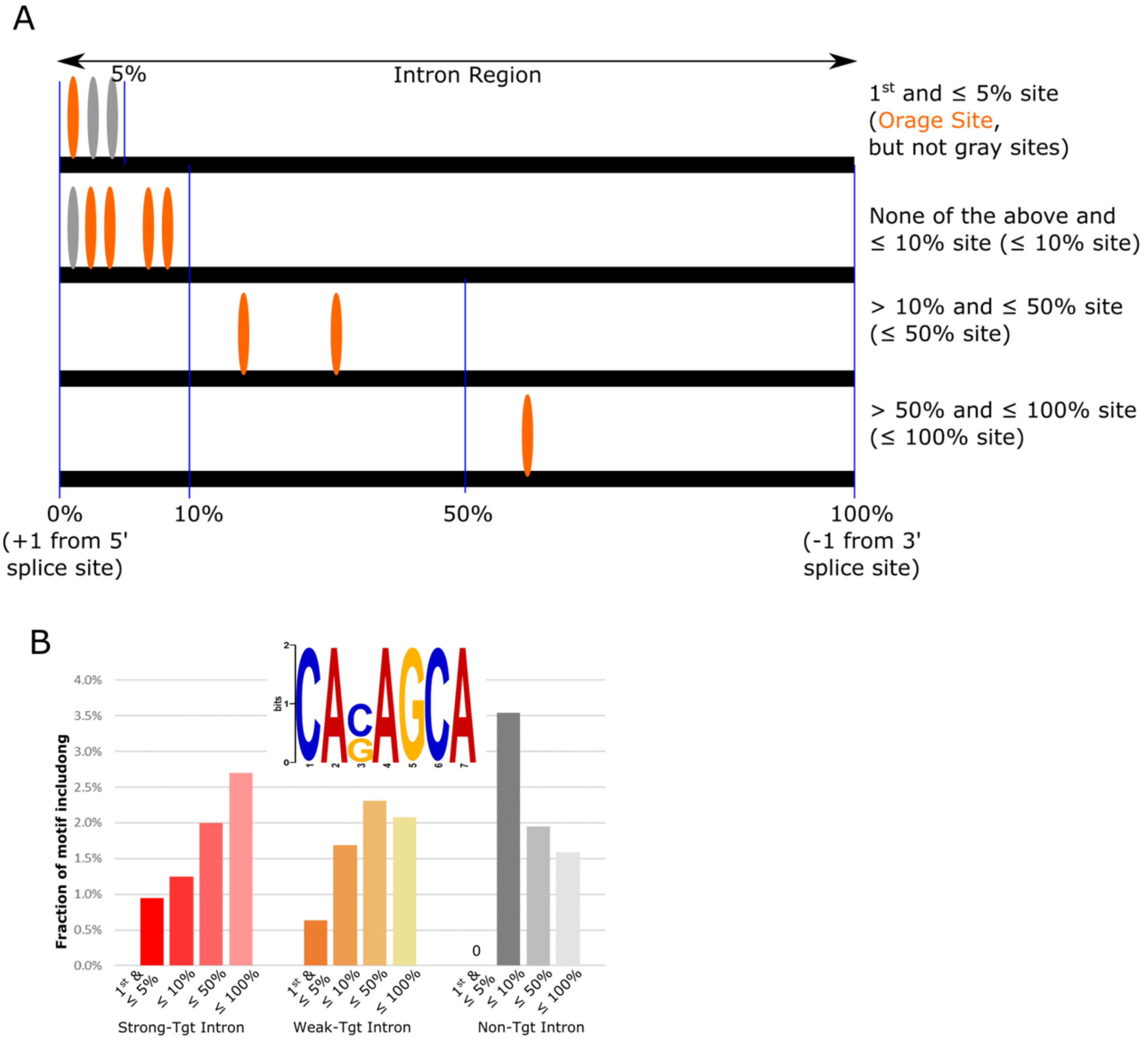
Supplemental Figure S4



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.

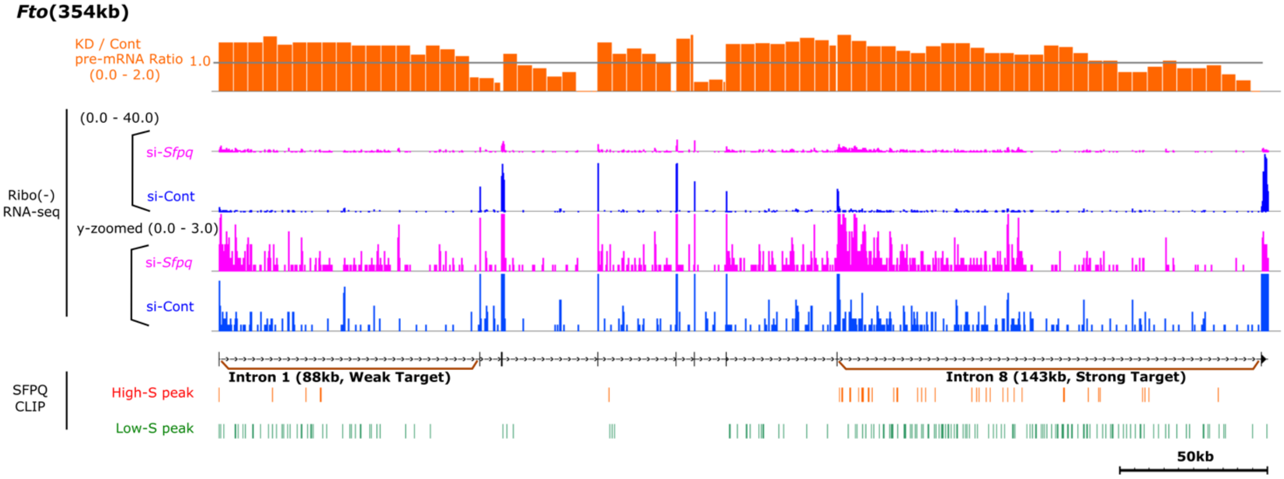
Supplemental Figure S5.



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.

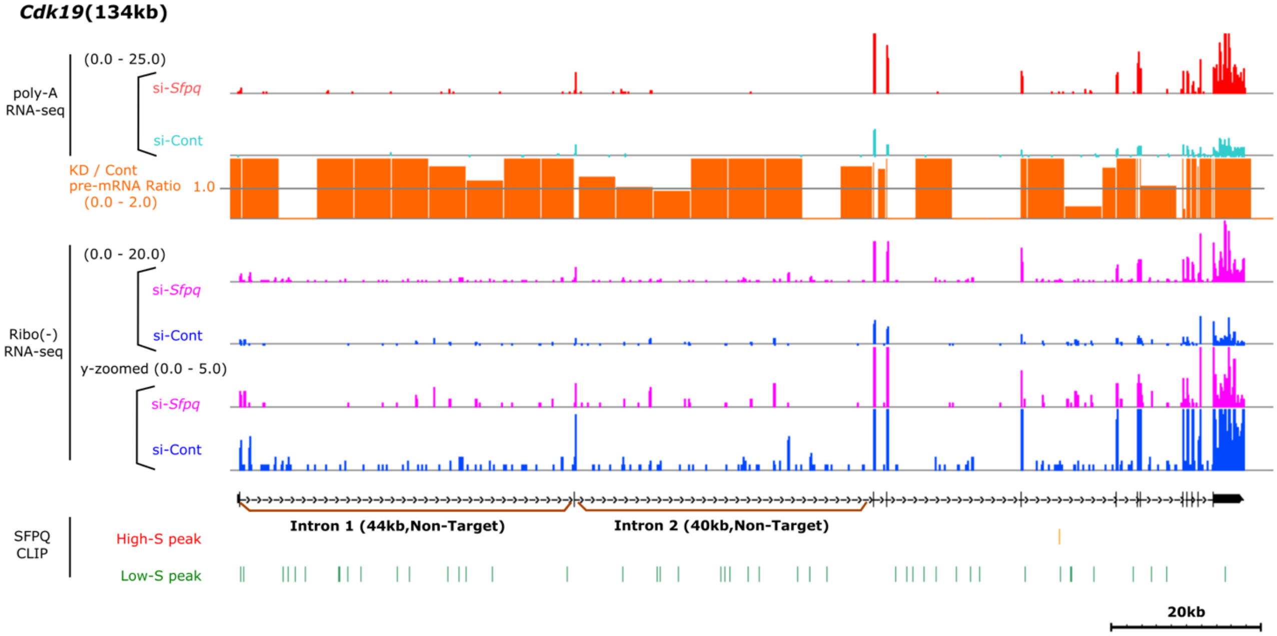
Supplemental Figure S6.



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.