Splicing regulation of large exons secures phase-separation of transcription factors in vertebrates

Toshihiko Kawachi[1,2], Akio Masuda[1,2,*], Yoshihiro Yamashita[1], Jun-ichi Takeda[1], Bisei Ohkawara[1], Mikako Ito[1], Kinji Ohno[1]

[1]Division of Neurogenetics, Center for Neurological Diseases and Cancer, Nagoya University Graduate School of Medicine, Nagoya, Japan

[2]These authors contributed equally

*Corresponding author

Akio Masuda, MD, PhD

Division of Neurogenetics, Center for Neurological Diseases and Cancer,

Nagoya University Graduate School of Medicine

65 Tsurumai, Showa-ku, Nagoya 466-8550, Japan

Phone: +81-52-744-2447, Fax: +81-52-744-2449, e-mail: amasuda@med.nagoya-u.ac.jp

Abstract

Large exons cannot be readily recognized by the spliceosome. Nevertheless, many large exons are evolutionarily conserved and constitutively spliced. Furthermore, the proteomic significance of large exons remains elusive. Here, we identified a set of nearly 3,000 SRSF3-dependent large constitutive exons (S3-LCEs) in human and mouse. The enriched C-nucleotides in S3-LCEs recruit two splicing factors, hnRNP K and SRSF3. hnRNP K induces the splicing suppression of S3-LCEs, which is mitigated by SRSF3 to achieve constitutive splicing of S3-LCEs. SRSF3 depletion deletes intrinsically disordered regions of transcription factors by skipping S3-LCEs and disrupts their phase-separated assemblies, which leads to cell death. Enrichment of C-nucleotides in large exons to code for proline and serine in intrinsically disordered regions of transcription factors was evolutionarily acquired in vertebrates. Layered splicing regulation by hnRNP K and SRSF3 secures their proper phase separation in vertebrates.

Keywords

Evolution/Large exon/Splicing/Intrinsically disordered region/Transcription factors

Introduction

In vertebrates, exons are considerably shorter than introns (Lander *et al*., 2001). The lengths of exons have decreased during evolution (Yandell *et al*., 2006). In the human genome, the median length of internal exons is 122 nt, and only approximately 5% of the exons exceed 300 nt (Haeussler *et al*., 2019). Large exonic sizes perturb recognition by spliceosome complexes and require splicing-enhancing mechanism(s) (Bolisetty and Beemon, 2012; Bruce and Peterson, 2001). Nevertheless, a substantial number of large exons are evolutionarily conserved and constitutively spliced in mammals (Bolisetty and Beemon, 2012).

RNA-binding proteins (RBPs) are implicated in splicing regulation, and spatiotemporally interact with pre-mRNA and spliceosome (Ule and Blencowe, 2019). Advances in high-throughput sequencing technologies have identified the transcriptome-wide landscape of protein–RNA interaction sites of more than 100 RBPs and revealed context-dependent regulation of alternative splicing (Fu and Ares, 2014; Lee and Ule, 2018). For example, serine/arginine-rich splicing factors (SRSFs) bind to specific regulatory sequences in exons to facilitate splicing, while binding of heterogenous ribonuclear proteins (hnRNPs) to specific regulatory sequences inhibit splicing (Fu and Ares, 2014). In addition, binding of RBPs to upstream and downstream intronic regions has opposite effects on splicing (Fu and Ares, 2014). A recent report described that RBPs preferentially bind to exonic regions of constitutive exons and flanking intronic regions of alternatively spliced (AS) exons (Van Nostrand *et al*., 2020). However, splicing *trans*-acting factors that enable the recognition of large constitutive exons remain unidentified.

Phase separation has emerged as a key mechanism for the organization of highly dynamic, membrane-less compartments regulating diverse cellular processes (Banani *et al*., 2017; Silveira and Bilodeau, 2018). The formation of phase-separated condensates is often driven by intrinsically disordered regions (IDRs) (Banani *et al*., 2017; Silveira and Bilodeau, 2018). IDRs are characterized by low sequence complexity with biased amino acid compositions (Kim and Shendure, 2019). IDR-bearing exons are often AS exons (Barbosa-Morais *et al*., 2012; Buljan *et al*., 2012; Ellis *et al*., 2012; Romero *et al*., 2006) and are enriched with exonic splicing enhancers (Smithers *et al*., 2015).

The interactions between upstream and downstream introns have a role in the splicing of IDR-bearing exons in genes for hnRNP A/D families (Gueroussov *et al*., 2017). However, splicing *trans*-acting factors to regulate most IDR-bearing AS exons remain to be identified.

Analysis of over 800 RNA-seq data derived from RBP-depletion experiments revealed that SRSF3 affected the splicing of the largest exons among other RBPs. We named nearly 3,000 such exons as SRSF3-dependent large constitutive exons (S3-LCEs). S3-LCEs were enriched with C-nucleotides, which serve as binding sites for SRSF3 and hnRNP K. SRSF3 depletion caused the skipping of S3-LCEs, which was rescued by hnRNP K depletion. hnRNP K depletion alone had no substantial effect on the splicing of S3-LCEs. Furthermore, S3-LCEs encode the IDRs of transcription factors. Loss of an IDR in these molecules by SRSF3 depletion disrupted the assembly of transcription factors, including the mediator complex. Evolutionary analyses revealed that large exons are extensively enriched in C-nucleotides encoding proline and serine, which constitute the IDRs of transcription factors in vertebrates. SRSF3-mediated mitigation of the effects of hnRNP K on C-enriched large exons makes transcription steadily sustainable.

Results

SRSF3 regulates splicing of large exons

To extensively analyze the roles of RBPs in pre-mRNA splicing, we extracted 804 publicly available RNA-seq data from the Gene Expression Omnibus (GEO) (Fig 1A). Of these, 42 canonical RBPs were individually knocked down or knocked out in various types of cells (Table EV1). The use of all annotated splice sites was quantified using the MISO algorithm (Katz *et al*., 2010). The difference in the percent spliced in (PSI) values (ΔPSI) between RBP-depleted and control cells was calculated. Depletion of 37 RBPs caused a total of 10,221 exon skipping events. Depletion of 32 RBPs caused a total of 6,790 exon inclusion events. Among them, the depletion of SRSF3 noticeably caused skipping of the largest number of exons (Fig 1B and Table EV2). In contrast, depletion of RBPs caused the inclusion of similar numbers of exons (Fig EV1A). We then calculated the median length of splicing-modulated exons by each RBP depletion, and found that SRSF3 depletion led to

skipping of the longest exons (Fig 1B). Consistent with a previous report showing SRRM4-dependent splicing regulation of microexons (Irimia *et al*., 2014), SRRM4-depletion caused the skipping of the shortest exons. In the ENCODE project, 440 RNA-seq analyses were performed in HepG2 or K562 cells, where 244 RBPs were individually knocked down (Consortium, 2012; Davis *et al*., 2018). The ENCODE datasets similarly showed the dependence of splicing of large exons on SRSF3 (Fig EV1B). The collective findings indicate that SRSF3 is an essential RBP for the splicing of large exons.

Global features of SRSF3-dependent exons

Inspection of 12 RNA-seq datasets of SRSF3-depletion experiments, in which six mouse cells and six human cells were examined, showed significant skipping of large exons in all the datasets (Table EV3 and Fig EV1C), suggesting a conserved role for SRSF3 in the splicing of large exons across cell types and species.

In total, SRSF3-depletion induced skipping of 650 and 2,428 exons in mouse and human cells, respectively (ΔPSI>0.2, Bayes factor>10; SRSF3-dependent constitutive large exons, S3-LCEs). Nearly half of them were annotated as constitutive exons (Fig 1C). The average ratio of constitutive exons in the skipped exons was the highest with SRSF3 in the 37 analyzed RBPs (Fig 1C and Appendix Fig S1A). Further RNA-seq analysis of 53 human tissues from the Genotype-Tissue Expression (GTEx) Consortium (Consortium, 2015) revealed that SRSF3-dependent exons formed a single peak of PSI values at 100% PSI. In contrast, AS exons displayed a bimodal PSI distribution with peaks at approximately 0% and 100% PSI (Fig 1D and EV1D). We also analyzed the maximum difference in PSI values (max-ΔPSI) for each -RBP-dependent exon in the 53 tissues. The average max-ΔPSI values of S3-LCEs were significantly lower than those of the other 36 analyzed RBPs (Fig EV1E). This finding suggested infrequent skipping of S3-LCEs in normal human tissues. As MaxEntScan scores of S3-LCEs, representing the splice-site (SS) strength, were not remarkably high compared to those of the other RBP-dependent exons, S3-LCEs may require splicing-enhancing SRSF3 to be constitutively spliced (Fig EV1F). Although a previous report

showed that large exons often found in the second or penultimate exons (Bolisetty and Beemon, 2012), the location of S3-LCEs showed no such preference compared to other RBP-dependent exons (Appendix Figs S1B and C). We also analyzed the enrichment of 5-mer nucleotides in S3-LCEs and found enrichment of C-nucleotides, including a previously reported SRSF3-binding CNNC motif (Auyeung *et al*., 2013) (Fig 1E).

SRSF3 overrides splicing-suppressive activity of hnRNP K on SRSF3-dependent exons

To explore the direct associations between SRSF3 and S3-LCEs, we performed targeted RNA immunoprecipitation sequencing (tRIP-seq) analysis that we recently developed to detect direct *in cellulo* RBP-RNA interactome (Masuda *et al*., 2020). As expected, the CNNC SRSF3-binding motif (Auyeung *et al*., 2013) was enriched at the SRSF3–RNA interaction sites (Fig EV2A). SRSF3-RNA interaction sites were enriched in exons, particularly near 5' splice sites of SRSF3-dependent large exons (Fig EV2B), suggesting a role of SRSF3 in an exonic splicing enhancer, as previously reported (Corbo *et al*., 2013).

To further dissect SRSF3-dependent regulation of S3-LCE-splicing, we made a minigene carrying *Cpsf6* exons 5 to 8 (Fig EV2C). Similar to endogenous *Cpsf6* mRNA, silencing of *Srsf3* caused the skipping of *Cpsf6* exon 7 (505 nt) of the wild-type minigene, (Fig EV2E). We then sequentially deleted 99-nt blocks excluding the first and last five nucleotides of the exon (Fig EV2C). Deletion of blocks 1 and 5 (Δ1 and Δ5), where SRSF3-binding was enriched (Fig EV2D), mostly (Δ1) and partially (Δ5) mitigated *Srsf3* silencing-mediated exon skipping, respectively (Fig EV2E). Although the disruption of binding of splicing-enhancing SRSF3 was predicted to induce exon skipping, both Δ1 and Δ5 caused the inclusion of the exon (Fig EV2E). These results prompted us to examine the existence of a splicing silencer that should counteract SRSF3.

To investigate candidate splicing silencers, we looked into the enhanced version of the crosslinking and immunoprecipitation (eCLIP) datasets of 116 RBPs in the ENCODE (Van Nostrand *et al*., 2020), although SRSF3-eCLIP was not included in the ENCODE datasets. We found that hnRNP K most abundantly accumulated on S3-LCEs (Fig 2A and EV2G). The expression levels of

hnRNP K and SRSF3 were regulated in parallel across tissues (Fig EV2H), which also suggested a strong relationship between these two factors. hnRNP K functions as a splicing repressor by binding to exonic C-stretches (Feng *et al*., 2019; Yamamoto *et al*., 2016). Motif analysis of hnRNP K-tRIP (see below) indicated that the C-stretch is a binding motif of hnRNP K (Fig EV2I). RT-PCR analysis showed that overexpression of hnRNPK induced skipping of exon 7 of the *Cpsf6* minigene, as expected (Fig EV2F, lane 2). Additional expression of SRSF3 mitigated the skipping of exon 7 in a dose-dependent manner (Fig EV2F, lanes 3 and 4).

RNA-seq analyses of *Srsf3*/*Hnrnpk*-silenced cells revealed that silencing of *Srsf3* alone induced the extensive skipping of large exons (Figs 2C, D, and EV2K). *Hnrnpk* silencing alone had a minimal effect on the splicing of S3-LCEs (Fig 2D). In contrast, *Hnrnpk* silencing in *Srsf3*-silenced cells mitigated the exon skipping caused by *Srsf3* silencing (Figs 2B, C, and D). Co-silencing of *Srsf3* and *Hnrnpk* increased PSI values in 76% of the S3-LCEs compared to the results with *Srsf3* silencing alone (Fig EV2L). Furthermore, the increase in PSI values by the co-silencing of *Srsf3* and *Hnrnpk* was more prominent in longer S3-LCEs (Fig 2E). Similarly, the increases in PSI values were more remarkable in S3-LCEs containing more C-nucleotides (Fig EV2M) and a greater number of triple C-nucleotide stretches (Fig 2F). We also confirmed that the knockdown efficiency of *Srsf3* mRNA was similar between the *Srsf3* silencing alone and the co-silencing of *Srsf3* and *Hnrnpk* (Fig EV2J and EV2O). These results indicated that hnRNP K suppresses the splicing of S3-LCEs.

hnRNP K-tRIP analysis of cells treated with control or *Srsf3* siRNA showed that the binding sites of hnRNP K overlapped with those of SRSF3 in control cells (Fig EV2N). *Srsf3* depletion did not affect the hnRNP K-binding profile to S3-LCEs (Fig 2G) or the overlapping binding sites with SRSF3 (Fig EV2N), indicating that SRSF3 does not interfere with hnRNP K bindings to S3-LCEs.

Next, we performed co-immunoprecipitation and mass spectrometric (CoIP-MS) analysis (Fig 3A) to identify molecules associated with SRSF3 and hnRNP K. A high molecular weight (HMW) fraction in the nuclei was extracted from cells overexpressing FLAG-tagged SRSF3 or

FLAG-tagged hnRNP K, and the proteins were subjected to CoIP-MS analysis. hnRNP K was minimally associated with SRSF3 in the HMW fraction (Figs 3B, C, D, and Table EV4). Instead, SRSF3 was associated with various molecules in the spliceosome, such as U1A, U1C, U1-70k, Sm proteins, U2AF1, and CPSF6 ($p < 0.05$, paired *t*-test), and to a lesser extent with hnRNPs (Figs 3B and D). In contrast, the association of hnRNP K was limited to hnRNPs (Figs 3C and D).

The collective findings indicated that SRSF3 and hnRNP K regulate the splicing of S3-LCEs in opposite directions. Constitutive splicing of S3-LCEs in most cells is achieved by SRSF3 in cooperation with its associated diverse spliceosome molecules. hnRNP K binds to the C-stretches to suppress the splicing of S3-LCEs. However, its splicing-suppressive effect is functionally masked by SRSF3 and its associated molecules.

The SRSF3–hnRNP K axis regulates the global selection of polyadenylation sites

SRSF3 is involved in alternative polyadenylation (APA) (Muller-McNicoll *et al*., 2016) in addition to alternative splicing. We analyzed the selection of APA sites in *Srsf3*/*Hnrnpk*-silenced cells using the DaPars algorithm (Masamha *et al*., 2014). *Srsf3* silencing caused the shortening of the 3′ untranslated regions (UTRs) of more than 1,000 genes (Figs 4A, B, and Appendix Fig S2A). *Hnrnpk* silencing alone had little effect on APA, whereas *Hnrnpk* silencing in *Srsf3*-silenced cells mostly ameliorated the shortening of the 3′ UTRs caused by the *Srsf3* silencing (Figs 4A and B). Thus, the effect of co-silencing of *Srsf3* and *Hnrnpk* on APA was similar to that observed on the splicing of S3-LCEs.

CPSF6 is a member of the cleavage factor I (CFIm) complex. The depletion of CPSF6 causes global 3′ UTR shortening (Zhu *et al*., 2018). The expression level of CPSF6, but not of other CPSFs, was prominently reduced by *Srsf3* silencing (Fig 4C). This reduction was mostly recovered by the additional silencing of *Hnrnpk* (Figs 4C and D), consistent with the changes in the APA sites. *Cpsf6* exon 7 (505 nt) was skipped by *Srsf3* silencing (Fig EV2E and Appendix Fig S2B), resulting in a frameshift in the downstream region. The levels of the exon 7-skipped transcripts were increased by cycloheximide treatment, suggesting that nonsense-mediated mRNA decay (NMD) degraded the

exon-skipped transcripts (Appendix Fig S2B). Consistent with a previous report (Zhu *et al*., 2018), tRIP analysis revealed the enrichment of CPSF6 binding upstream to the affected distal APA sites (Fig 4E and Appendix Fig S2C). The finding indicated the direct role of CPSF6 in the SRSF3-dependent regulation of APA. The collective findings indicated that *Srsf3* silencing induces the skipping of *Cpsf6* exon 7 and reduces the expression level of CPSF6, which causes global shortening of the 3′ UTRs.

SRSF3-dependent exons are required for the formation of transcription complexes

We next explored the functional significance of S3-LCE-bearing genes. Gene Ontology (GO) analysis revealed that S3-LCE-bearing genes were most enriched in the GO term, "positive regulation of transcription" (Figs 5A and B). Clustering analysis of GO terms showed that "positive regulation of transcription" constituted the core of a large GO cluster related to transcriptional regulation. Additionally, the sizes of exons in the four GO clusters associated with transcriptional regulation were significantly higher than those in another GO cluster comprising "organelle organization" (Fig 5C). These results suggested the involvement of S3-LCEs in transcriptional regulation.

Protein–protein interaction network analysis of the transcription factors encoded by genes carrying the SRSF3-dependent exons revealed that these exons were extensively enriched in the genes encoding the components of transcription machineries (Fig 6A). We focused our analysis on the mediator complex, which is a multi-protein complex that plays essential roles in transcription (Allen and Taatjes, 2015). We first analyzed the cellular distribution of MED1 and MED4. *Med1* has an alternative last exon activated by *Srsf3* silencing (Fig 6B). In normal cells, the last exon is extended (isoform-1), while *Srsf3* silencing introduces an intron in the extended region (isoform-2, Fig 6B). The anti-N-terminal MED1 antibody detected the protein expression of a large isoform-1, but not of a small isoform-2, suggesting that isoform-2 is not efficiently translated or the translated protein is rapidly degraded (Appendix Fig S3A, left panel). The last exon of isoform-1 encodes an IDR (Fig 6B) that can phase-separate to form nuclear puncta (Sabari *et al*., 2018). Confocal

immunofluorescence imaging with anti-C-terminal MED1 antibody showed that MED1 in control small interfering RNA (siRNA)-treated C2C12 cells (siCont) was mostly distributed in the nuclei (Fig 6C). In contrast, *Srsf3* silencing significantly reduced nuclear MED1 and increased its cytoplasmic localization (Fig 6C). The decrease in nuclear MED1 was recovered by the co-silencing of *Srsf3* and *Hnrnpk* (Fig 6C). The shift of MED1 from the nuclei to the cytoplasm by *Srsf3* silencing and its recovery by the co-silencing of *Srsf3* and *Hnrnpk* was detected with a different anti-N-terminal MED1 antibody (Appendix Fig S3A) and with a different siRNA set (Appendix Fig S3B). Additionally, we observed a mild decrease in nuclear MED4 by *Srsf3* silencing, although *Srsf3* silencing had no effect on the splicing of *Med4* (Figs 6D and E). Super resolution confocal imaging revealed the formation of distinct nuclear puncta of MED1 (Fig 6F) as previously reported (Sabari *et al*., 2018). These puncta overlapped with the MED4 puncta (Fig 6F). In accordance with the nuclear localization of MED1, *Srsf3* silencing reduced the number of MED1 puncta in the nuclei, and the co-silencing of *Srsf3* and *Hnrnpk* recovered it (Fig 6G). Co-existence of MED1-MED4 in puncta was also greatly reduced in *Srsf3*-silenced cells, which was recovered by the co-silencing of *Srsf3* and *Hnrnpk* (Fig 6H).

We further analyzed MED15 and MED12, both of which have S3-LCEs (Fig EV3A and B). In particular, Med15 was reported to form phase-separated condensates (Boija *et al*., 2018). Similar to MED1, *Srsf3* silencing reduced the number of MED15 and MED12 puncta as well as their nuclear localization (Fig EV3C and D). *Hnrnpk* silencing marginally rescued the exon skipping of MED15 and MED12, and failed to rescue their nuclear localization in *Srsf3*-silenced cells (Fig EV3C and D). Nevertheless, *Hnrnpk* silencing recovered the number of MED15 and MED12 puncta (Figs EV3C and D), which was likely due to the marked recovery of MED1 puncta (Fig 6G). To assess the size of the mediator complexes, the HMW fractions of these cells were resolved by native PAGE. Immunoblotting with the antibodies against MED1, MED4, and MED15 identified the complexes distributed from 800 kDa to several MDa (Fig EV3E). *Srsf3* silencing particularly decreased the megadalton complexes (Fig EV3E). These results indicated the essential involvement of S3-LCEs in the assembly of the mediator complex.

We also examined whether similar regulations were observed in the BAF complex, which has a significant role in chromatin remodeling (Mittal and Roberts, 2020). Exon 3 of *ARID1A* encoding a component of the canonical BAF complex is skipped by *Srsf3* silencing (Fig EV3F). Similar to MED1, ARID1A puncta were reduced in response to *Srsf3* silencing, and were recovered by the co-silencing of *Srsf3* and *Hnrnpk* (Fig EV3G). In addition, *Srsf3* silencing affected the APA of *Smarcc1* and the alternative splicing of *Smarcc2* (Figs EV3H and J), both of which encode the core components of the BAF complex (Mittal and Roberts, 2020). Although these molecules do not apparently form nuclear puncta, nuclear expression of SMARCC1 and SMARCC2 was greatly decreased by *Srsf3* silencing (Figs EV3I and K). Thus, S3-LCEs are also involved in the formation of the BAF complex.

S3-LCEs encode IDRs in transcription factors

We noticed that S3-LCEs frequently encode IDRs (Figs 6B, EV3A, B, F, H, and J). IDRs have essential roles in molecular interaction networks in cells (Oldfield and Dunker, 2014). S3-LCEs are also found in the IDRs of several transcription factors, such as BRD4, SP1, and FUS (Fig EV4A) (Chong *et al*., 2018; Patel *et al*., 2015; Sabari *et al*., 2018). We searched for amino acid features encoded by S3-LCEs. The search frequently revealed S3-LCEs in IDRs. The IUpred2 algorithm (Mészáros *et al*., 2018) determined that S3-LCEs encode higher percentages of disordered amino acids in IDRs compared to the other RBP-responsive exons (Fig 7A). The IDRs of S3-LCEs were enriched, especially in molecules related to transcription (Fig 7B). In addition, the protein coding sequences of S3-LCEs were found to be rich in proline and serine (Fig 7C), which are classified as disorder-promoting amino acids and are encoded by C-rich codons (Oldfield and Dunker, 2014).

IDRs in proteins drive phase separation (Banani *et al*., 2017). We chose eight IDRs encoded by S3-LCEs in transcription factors (Fig EV4A) and purified the recombinant IDRs fused with mCherry. Confocal imaging revealed that all these fusion proteins, but not mCherry alone, formed spherical droplets under crowding conditions (Fig 7D and EV4B). In addition, the IDR encoded on CPSF6 exon 7 and the full-length CPSF6 protein, but not the CPSF6 protein lacking the

IDR (Δexon 7), formed droplets (Fig 7D and EV4B). The droplets were smaller and less numerous at lower protein concentrations, but were resistant to increased ionic strengths (Fig EV4B). In addition, the droplets underwent fusion and exhibited rapid fluorescence recovery following photobleaching (Fig EV4C and movies EV1-3). These results are consistent with liquid-liquid phase-separated condensates (Harlen and Churchman, 2017).

We next examined the effects of *Srsf3*-sliencing-mediated aberration of transcription factors on cell proliferation. *Srsf3*-silenced cells stopped proliferation and began to die after day 2 (Fig 7E). *Hnrnpk* silencing alone had no essential effect on cell proliferation. However, the co-silencing of *Srsf3* and *Hnrnpk* significantly ameliorated *Srsf3* silencing-induced cell death. GO analysis showed that the genes downregulated by *Srsf3*-silencing and those recovered by co-silencing of *Hnrnpk* and *Srsf3* were extensively associated with DNA replication and cell division (Appendix Fig S4A, B and Table EV5). These results provided evidence for a critical role of S3-LCEs in cell proliferation.

C-nucleotide enrichment of vertebrate large exons during evolution secures IDRs of transcription factors in splicing

In the human genome, analysis of all internal coding exons revealed that large exons (> 180 nt) were enriched with C-nucleotides (Fig 8A) that generated codes for proline and serine (Fig 8B). Similarly, in mid-sized exons (60–180 nt), C-nucleotide content also increased with increasing exon length (Fig 8A). In contrast to large exons, the increase in C-nucleotide content in mid-sized exons was not accounted for by changes in amino acid compositions, but rather by the increase in the number of C-nucleotides at the third position of a codon (Fig EV5A). Large exons (> 180 nt) and, to a lesser extent, small exons (< 60 nt) were determined to preferentially encode IDRs in the human genome (Fig 8C and EV5B). Among the large exons, IDR-encoding exons were enriched with proline and serine (Fig EV5C) and C-nucleotides (Fig EV5D). Large exons enriched with C-nucleotides were markedly observed in genes associated with transcriptional regulation (Fig 8E). The collective findings indicated that S3-LCEs constitute a subset of the C-rich large exons encoding

IDRs.

Evolutionary analysis revealed the enrichment of vertebrate large exons with C-nucleotides (Fig EV5E), coding for IDRs (Figs 8D, EV5F and EV5G), and their concentrations in genes associated with transcriptional regulation (Fig 8E). We further analyzed the preferred codons in large exons in 265 species. Neighboring species displayed similar codon preferences in large exons (> 180 nt) (Fig 8F and Appendix Fig S5). Codons enriched in vertebrate large exons preferentially harbored C-nucleotides at the second position of codons, which mostly encode proline and serine (Cluster 1 in Fig 8G). Codons enriched in vertebrate large exons also displayed C-nucleotides at the second and third positions of codons, which encode variable amino acids (Cluster 2 in Fig 8G). A similar analysis using all exons showed that 20 codons increased with increasing exon length in vertebrates (Cluster 2 in Fig EV5H). In contrast to large exons (Cluster 1 in Fig 8G), proline or serine was not enriched in Cluster 2 (Fig EV5I). In accordance with the increase in C-nucleotide content at the third position of a codon with increasing exon lengths (Fig EV5A), C-nucleotides were enriched at the third position of a codon in Cluster 2 (Fig EV5I). In contrast to the enrichment of C-nucleotides in large exons (Clusters 1 and 2 in Fig EV5I), short exons are rich in A and T nucleotides at the second position of codons, which encode order-promoting phenylalanine, isoleucine, and leucine (Oldfield and Dunker, 2014) (Cluster 4 in Fig EV5I). To summarize, C-nucleotides at the second position of a codon were acquired in vertebrate large exons (> 180 nt) to code for proline and serine. In addition, C-nucleotides at the third position of a codon were acquired with increasing exon lengths, which were independent of amino acid constraints.

Discussion

We demonstrated that large exons extensively encode PS-rich IDRs specifically in transcription factors. These large exons are enriched with C-nucleotides to encode P and S, which serve as binding sites for SRSF3 and hnRNP K. hnRNP K suppresses the splicing of the large exons, which is masked by SRSF3. An interesting example of the splicing-suppressive effect of hnRNP K and its masking effect by SRSF3 is observed in *Cpsf6*. hnRNP K suppresses the splicing of a large

exon in *Cpsf6*, and makes an exon-skipped CPSF6 degraded by NMD. CPSF6 is an accessory molecule that activates the distal polyadenylation sites. Depletion of CPSF6 activates proximal polyadenylation sites (Zhu *et al*., 2018). hnRNP K shortens mRNAs and SRSF3 counteracts this shortening. Furthermore, the absence of SRSF3 results in the loss of the IDRs of transcription factors and disrupts their assembly. The collective findings clarify the essential role of SRSF3-dependent splicing regulation in sustainable transcription.

SRSF3-depndent exons were mostly large and constitutively spliced (Figs 1B, C, and D). They were designated S3-LCEs. SRSF3 and hnRNP K can bind to adjacent or overlapping sites, but do not compete for binding to RNA (Fig 2G and EV2N). SRSF3 and hnRNP K are unlikely to directly interact with each other (Figs 3B, C and D). Instead, the interactome analysis showed that SRSF3 associates with diverse spliceosome molecules, while hnRNP K associates only with hnRNPs (Figs 3B, C and D). These observations point to a layered splicing regulation by SRSF3 and hnRNP K. In the concealed layer, hnRNP K binds to C-rich segments in the exons to cause exon skipping (Figs 2C and D). In contrast, in the overriding layer, SRSF3 masks the splicing-suppressive activity of hnRNP K in cooperation with the associated splicing molecules (Figs 2C and D). Constitutive splicing of S3-LCEs was extensively observed in 53 adult human cells/tissues in GTEx (Fig EV1D). Although alternative splicing was scarcely observed in S3-LCEs, SRSF3 may be suppressed in specific cells/tissues at a specific developmental stage to skip the IDR-bearing S3-LCEs to modulate transcription.

Splicing of S3-LCEs was more efficient in naïve cells than in cells silenced for both *Srsf3* and *Hnrnpk* (Fig 2D). These findings suggest that SRSF3 does not simply counteract the splicing-suppressive activity of hnRNP K, and that SRSF3 may also counteract another splicing-suppressing RBP. SRSF3 is thus required for the constitutive splicing of S3-LCEs even in the absence of hnRNP K. The large sizes of S3-LCEs may require SRSF3, since exons exceeding 300 nt are inefficiently recognized by the exon definition mechanism (Robberson *et al*., 1990). SRSFs generally promote exon definition by facilitating the recruitment of spliceosomal components (Busch and Hertel, 2012). Although SRSF3 recognizes C-rich elements (Fig 1E and EV2A), most SRSFs recognize AG-rich

elements (Piva *et al*., 2012; Van Nostrand *et al*., 2020). SRSF11 is an unusual SRSF protein that recognizes CU-rich elements and participates in the splicing regulation of microexons (Gonatopoulos-Pournatzis *et al*., 2018). SRSF3 is another unusual SRSF protein that recognizes C-rich elements and facilitates the constitutive splicing of large exons (Fig 1B).

We showed that SRSF3 overrides the splicing-suppressive activity of hnRNP K on large exons. hnRNP K is a multifunctional RNA/DNA-binding protein implicated in a wide range of biological processes, including chromatin remodeling, nuclear localization of RNAs, and translational control (Lubelsky and Ulitsky, 2018; Wang *et al*., 2020). *Hnrnpk* silencing alone had no effect on the splicing of MED1, MED4, MED15, and MED12 (Figs 6B, D, EV3A and B), but marginally increased the nuclear concentration of MED4 (Fig 6E) as well as the nuclear puncta of MED15 (Fig EV3C).

Our analysis revealed the close relationship between SRSF3 and CPSF6. SRSF3 maintains the expression level of CPSF6 protein through alternative splicing (Figs 4C, D and Appendix Fig S2B), which has also been reported by another group recently (Schwich, Blumel *et al*., 2021). In addition, SRSF3 and CPSF6 make a complex in the HMW fraction of cells (Fig 3B). The SRSF3-CPSF6 interaction is involved in the global regulation of APA (Figs 4A and B). A previous study identified a large complex comprising U1 snRNP and cleavage/polyadenylation factors (U1-CPAFs), which includes CPSF6 and is essential for transcription elongation (So, Di *et al*., 2019). SRSF3 is abundantly found in U1-CPAFs (So *et al*., 2019), suggesting that SRSF3 participates in the regulation of APA through interacting with cleavage/polyadenylation factors in U1-CPAFs.

The biochemical properties of amino acids, such as charge and hydrophobicity, are important determinants of protein function. The large exons encoding IDRs are prominently enriched with codons for uncharged and disorder-promoting amino acids, including proline and serine (Figs 8B, 8C, and EV5C) (Oldfield and Dunker, 2014). We showed that SRSF3 enables the constitutive splicing of large exons encoding IDRs to sustain the assembly of transcription factors (Figs 5, 6, 7 and EV4). In contrast, another SR-related protein, SRRM4, regulates neuron-specific alternative splicing of microexons (Irimia *et al*., 2014). These microexons are enriched with codons for charged

amino acids to modulate the protein interactions involved in neurogenesis. A recent report showed that exons under the control of SRSF1, SRSF2, SRSF3, and SRSF10 have biased amino acid compositions with distinct properties (Fontrodona *et al*., 2019). Similar to SRSF3 and SRRM4, other SRSFs may participate in specific cellular functions through the biased amino acid compositions encoded in their target exons.

Disordered segments are found in 2.0% of Archaean proteins, 4.2% of eubacterial proteins, and 33.0% of eukaryotic proteins (Ward *et al*., 2004). Among eukaryotes, evolutionarily young genes have more IDRs than old genes (Banerjee and Chakraborty, 2017), indicating rapid evolution of IDRs (Brown *et al*., 2002). IDRs are frequent targets of positive selection, in which adaptive substitutions of amino acids are preferentially observed (Afanasyeva *et al*, 2018). IDRs are "disordered regions" where amino acid constraints are not as stringent as catalytic cores or structural scaffolds. This low stringency might have allowed the generation of exonic splicing-enhancing *cis*-elements for SRSF3. The introduction of splicing *cis*-elements for SRSF3 might have prevented elimination of large exons in the course of evolution.

The current study revealed that a set of large exons bearing IDRs has emerged in vertebrates (Fig 8D and EV5F). These exons preferentially encode proline and serine (Figs 8B, 8G, and EV5C), and are essential for the assembly of transcription factors (Fig 6 and EV3). Another set of IDR-bearing exons is comprised of mammalian-specific AS exons (Gueroussov, Weatheritt *et al*., 2017). These exons preferentially encode glycine and tyrosine, and are required for the assembly of hnRNPs. Thus, at least two sets of IDR-bearing exons have distinctly evolved, having different amino acid preferences and distinct functions. Further analyses will disclose diverse sets of distinct IDR-bearing exons in the future.

Codons for proline and serine are rich in C-nucleotides. The enrichment of these amino acids in large exons increases the content of C-nucleotides (Fig 8A). The enrichment of C-nucleotides in large exons recruits SRSF3 and hnRNP K. The increasing ratios of C-nucleotides with increasing exon lengths without affecting amino acid compositions should have further accelerated the recruitment of SRSF3 and hnRNP K in large exons (Fig 8H). Evolutionary acquisition of the

synergistic acceleration of C-nucleotides in large exons to code for proline and serine might have unexpectedly recruited hnRNP K, which needs to be extensively overridden by SRSF3 to secure the formation of IDRs of transcription factors.

Materials and Methods

Cell culture

C2C12 cells and HEK293 cells were grown in DMEM with 10% f*et al* bovine serum at 37°C in 5% CO_2.

siRNA and transfection

Two sets of siRNA duplexes against mouse *Srsf3* and *Hnrnpk* were synthesized by Fasmac. The sense sequences of the siRNAs were as follows: set A-siSrsf3 (5′-CGAUCUAGGUCAAAUGAAA-3′), set A-siHnrnpk (5′-GGGGAGAUCUAAUGGCUUA-3′), set B-siSrsf3 (5′-GGAACUGUCGAAUGGUGAA-3′), and set B-siHnrnpk (5′-GGAGAAAUUCUGAAGAAAA-3′). Set A of siRNAs were used unless otherwise indicated. C2C12 cells were transfected with siRNA using Lipofectamine RNAiMAX (Thermo Fisher Scientific) according to the manufacturer's instructions. We purchased the AllStar Negative Control siRNA (1027281) from Qiagen.

In the *Srsf3*/*Hnrnpk* double-silencing experiments, we constantly introduced the same amounts of specific siRNAs to cells. For example, when a total of 100 pmol of siRNAs was introduced to cells, 50 pmol each of control siRNA and *Hnrnpk*-siRNA, 50 pmol each of control siRNA and *Srsf3*-siRNA, and 50 pmol each of *Hnrnpk*-siRNA and *Srsf3*-siRNA were introduced to *Hnrnpk*-silenced cells, *Srsf3*-silenced cells, and both *Hnrnpk*- and *Srsf3*-silenced cells, respectively. The efficiency and specificity of these siRNA treatments are indicated in Figs EV2J and O.

Minigene and transfection

The minigene spanning exons 5 to 8 of *Cpsf6* gene (Fig EV2C) was constructed by insertion of a PCR-amplified mouse genomic fragment into HindIII and BamHI sites of pcDNA3.1 (+) vector (Thermo Fisher Scientific). C2C12 cells were transfected with the *Cpsf6* minigene using Fugene6 (Promega) according to the manufacturer's instructions. The serial deletions of 99-bp blocks in *Cpsf6* exon 7 were engineered into the *Cpsf6* minigene using the QuikChange site-directed

mutagenesis kit (Agilent Technologies).

Expression vectors and transfection

The expression vectors of hnRNP K and SRSF3 (3XFLAG-HNRNPK and 3XFLAG-SRSF3, respectively) were constructed as follows. First, *Hnrnpk* and *Srsf3* cDNAs were PCR-amplified and inserted into p3XFLAG-CMV10 vector (Sigma-Aldrich). Then, cDNAs of 3XFLAG-tagged SRSF3 and 3XFLAG-tagged HNRNPK were further PCR-amplified and inserted into pEF-BOS vector (Mizushima & Nagata, 1990), an *Eef1a*-promoter driven mammalian expression vector. For the splicing assays, C2C12 cells were transfected with these plasmids using Lipofectamine 3000 (Thermo Fisher Scientific) according to the manufacturer's instructions. For the immunoprecipitation assays, HEK293 cells were transfected using the NEPA21 electroporation system (NEPAGENE). The NEPA21 electroporator was operated with poring pulses of voltage = 125 V, pulse length = 2.5 ms, pulse interval = 50 ms, the number of pulses = 2, decay rate = 10%, and polarity = +; followed by transfer pulses of voltage = 20 V, pulse length = 50 ms, pulse interval = 50 ms, the number of pulses = 5, decay rate = 40%, and polarity = +/-. One million HEK293 cells were transfected with 10 μg of 3XFLAG-HNRNPK, 3XFLAG-SRSF3, or the empty pEF-BOS vector for 48 h prior to the analysis.

RT-PCR

Total RNA was isolated at 40 h after transfection using the Trizol reagent (Thermo Fisher Scientific) followed by treatment with DNase I (Qiagen). cDNAs were synthesized with an oligo-dT primer (Thermo Fisher Scientific) and ReverTra Ace reverse transcriptase (Toyobo). PCR was performed with GoTaq polymerase (Promega) using the following primer sets: *Cpsf6* minigene (forward 5'-GTGGGGACAGATTTCCTGGG-3'; reverse 5'-AACAACAGATGGCTGGCAAC-3') and *Cpsf6* mRNA (forward 5'-GTGGGGACAGATTTCCTGGG-3'; reverse 5'-ACTGCTTGAGATTGCCCGAT-3').

RNA-seq

Total RNA was extracted from C2C12 cells transfected with the indicated siRNA using TRIZOL (Thermo Fisher Scientific), and was further purified with Quick-RNA Miniprep Kit (Zymo research) according to the manufacturer's instructions. The extracted RNA was subjected to RNA-seq at Macrogen, Japan. Briefly, a sequencing library was prepared using the TruSeq Stranded mRNA kit (Illumina), and the library was read on Illumina NovaSeq 6000 (150 bp paired-end reads). Adaptor sequences were trimmed by Cutadapt (version 1.18) (Martin, 2011). Reads were mapped to the mouse reference genome (mm10) using STAR version 2.5.2b (Dobin *et al*., 2013). PSIs of all internal exons in the RefSeq annotation were calculated by MISO (Katz *et al*., 2010).

tRIP-seq

tRIP is an improved version of CLIP methodology, which identifies protein-RNA interactions in living cells with ~100-times higher sensitivity and similar specificity to CLIP (Masuda, Kawachi *et al*., 2021, Masuda, Kawachi *et al*., 2020). C2C12 cells were UV-irradiated at 400 mJ, and whole cell lysates were harvested from the cells. tRIP was performed as previously described (Masuda *et al*., 2020). Samples were sequenced on the Illumina NovaSeq6000 with 150 bp paired-end read (Macrogen, Japan) or Miseq with 150 bp single-read at the core facility of the Nagoya University. Mapping of sequenced reads were performed as previously described (Masuda *et al*., 2020). For paired-end read data, only P5 reads were used for analysis. Briefly, after standard HiSeq demultiplexing, reads were adapter-trimmed and reads less than 18 bp were discarded using cutadapt (v1.10) (Martin, 2011). Mapping was first performed against the mouse repetitive elements in RepBase (Bao *et al*., 2015) with STAR (v2.5.2b) (Dobin *et al*., 2013). Repeat-mapped reads were removed, and all others were then mapped against the mouse genome (mm10) with STAR (v 2.5.2b). Multiply mapped reads were filtered out. Duplicates of reads uniquely mapped to the human or mouse genome were removed by Picard (v2.0.1, http://broadinstitute.github.io/picard/). To identify tRIP-tag clusters, we used MACS (v1.4.2) (Zhang *et al*., 2008) with the following parameters "–f BAM —nomodel —shiftsize 25". Motifs enriched in tRIP-tag clusters were generated by MEME-

ChIP (v 4.11.2) (Machanick and Bailey, 2011), as previously described (Masuda *et al*., 2020).

Subcellular fractionation

Subcellular fractionation was performed by modifying two previously reported methods (Damianov *et al*., 2016; Nojima *et al*., 2016). Cells cultured in 15-cm dish were washed with ice-cold PBS twice and harvested by scraping in ice-cold PBS. Then, cells were pelleted by centrifugation at 500×*g* for 5 min at 4°C, and the supernatant was discarded. Cells were resuspended in 4 ml of ice-cold HLB+N buffer (10 mM HEPES-KOH pH 7.6, 10 mM KCl, 2.5 mM $MgCl_2$, and 0.5% (vol/vol) NP-40) and incubated on ice for 5 min. To separate the cytoplasmic fraction, 1 ml of ice-cold HLB+NS (10 mM HEPES-KOH pH 7.6, 10 mM KCl, 2.5 mM $MgCl_2$, 0.5% (vol/vol) NP-40, and 10% (wt/vol) sucrose) was slowly laid under the cells, and the cells were centrifuged at 1000×*g* for 5 min. Then, the supernatant was removed as a cytoplasmic fraction, and the pelleted nuclei were washed with 4 ml of ice-cold HLB (10 mM HEPES-KOH pH 7.6, 10 mM KCl, and 2.5 mM $MgCl_2$), followed by centrifugation at 500×*g* for 5 min. Then, the supernatant was discarded, and the nuclei were lysed with 10-volumes of nucleus lysis buffer [20 mM HEPES-KOH pH 7.6, 150 mM NaCl, 1.5 mM $MgCl_2$, 0.1% NP-40, Protease inhibitor, and PhosSTOP (Sigma-Aldrich)]. The sample was transferred to a new 1.5-ml tube, and incubated on ice for 5 min. For high molecular weight fraction, the sample was centrifuged at 16,000×*g* for 5 min, and the supernatant was collected as a soluble fraction. The high molecular weight fraction was washed with 1 ml of ice-cold MNase buffer (20 mM Tris-HCl pH 8.0 and 2.5 mM $CaCl_2$), centrifuged at 16,000×*g* for 2 min and the supernatant were discarded. Ice-cold 1x MNase buffer (80 µl) was added to the pellet, and the pellets were separated into small pieces by pipetting. The MNase buffer (20 µl) supplemented with 1 µl of MNase (NEB) was added to the sample (final 20 gel units/µl) and incubated at 37°C in a Thermomixer (Eppendorf) at 1000 rpm for 90 sec. Then, 10 µl of 0.2 M EGTA was immediately added to the sample to stop the nuclease digestion. The sample was centrifuged at 16,000×*g* for 10 min, and the supernatant was collected as HMW MNase-extract.

Immunoprecipitation

HMW MNase-extracts of HEK 293 cells transiently expressing 3XFLAG-SRSF3 or 3XFLAG-HNRNPK were collected as described above (3XFLAG-IP). HMW MNase-extracts of HEK 293 cells transfected with an empty pEF-BOS vector were used as a control (Cont-IP). The extracts were incubated with 5 µl of Anti-FLAG M2 Affinity Gel (Sigma-Aldrich) for 2 h at 4°C. Beads were washed 3 times in wash buffer (50 mM HEPES-KOH pH 7.6, 150 mM NaCl, and 0.05% NP-40) on a spin column (Thermo Fisher Scientific). To remove remaining detergent for subsequent mass spectrometry, beads were washed 2 times in detergent-free wash buffer (50 mM HEPES-KOH pH 7.6 and 150 mM NaCl). Immunoprecipitants were eluted using 50 µl of elution buffer [50 mM HEPES-KOH pH 7.6, 150 mM NaCl, and 1 mg/mL 3XFLAG peptide (Sigma-Aldrich)] by incubation for 1 h at 4°C.

For Mass spectrometry, the proteins were digested by trypsin for 16 h at 37°C after reduction and alkylation. The peptides were analyzed by LC-MS using an Orbitrap Fusion mass spectrometer (Thermo Fisher Scientific) coupled to an UltiMate30000 RSLCnano LC system (Dionex) using a nano HPLC capillary column, 150 µm×75 µm i.d (Nikkyo Technos) via a nanoelectrospray ion source.

The raw data was processed using Proteome Discoverer 1.4 (Thermo Fisher Scientific) in conjunction with MASCOT search engine, version 2.6.0 (Matrix Science Inc.) for protein identification. Peptides and proteins were identified against human protein database in UniProt (release 2019_07), with a precursor mass tolerance of 10 ppm, a fragment ion mass tolerance of 0.8 Da. Fixed modification was set to carbamidomethylation of cysteine, and variable modifications were set to oxidation of methionine. Two missed cleavages by trypsin were allowed.

Three independently co-immunoprecipitated 3XFLAG-IP and Cont-IP samples were subjected to mass spectrometric analyses (CoIP-MS). The identified proteins were sorted according to fold-changes in MASCOT scores of 3XFLAG-IP vs Cont-IP. Then, 25% of the proteins from the top and the bottom were removed from the list. The sum of the MASCOT scores of the remaining 50% proteins was calculated to obtain a normalization factor for each sample. The MASCOT scores

of individual proteins were divided by the normalization factor. Then, a fold-change of the normalized MASCOT scores of 3XFLAG-IP vs Cont-IP was calculated for each sample. In addition, the normalized score of each protein was calculated using the following formula:

Normalized score = $\log_2$ (1,000,000 × MASCOT score / normalization factor).

Statistical difference of three repeated experiments was calculated by paired *t*-test of the normalized scores of each protein.

Recombinant proteins

Plasmids for recombinant proteins that were used for *in vitro* droplet assays were generated by cloning the indicated SRSF3-dependent large constitutive exons (S3-LCE) at EcoRI and XhoI sites of the pGEX-6P-1 vector (Cytiva) harboring the GST and the full length mCherry at the N-terminal end of the insert. All expression constructs were sequenced to ensure lack of PCR artifacts and were transformed into BL21 (DE3) cells (NEB).

The transformed bacteria were grown in LB medium with 100 μg/ml ampicillin. At absorbance OD600 = 0.6, protein expression was induced with 0.1 mM IPTG, and the bacteria were grown additional 18 h at 20°C at 150 rpm. After centrifugation, cell pellets were resuspended in 10 ml PBS and sonicated (MySonic, Power 60, 10 cycles of 30-sec burst and 30-sec rest) at 4°C. GST-fusion proteins were isolated from the soluble lysate using glutathione-Sepharose 4B beads (Cytiva). Then, the recombinant proteins were cleaved on the beads from GST by incubation with PreScission protease (Cytiva) in PreScission buffer (50 mM Tris-HCl pH 8.0, 100 mM NaCl, 1 mM EDTA, and 1 mM DTT) overnight at 4°C. The cleaved proteins were concentrated, and the buffer was exchanged to 100 mM Tris-HCl, pH 7.5 and 20% glycerol using Amicon Ultra centrifugal filters (Millipore).

***In vitro* droplet assay**

Recombinant protein was added to solutions at varying concentrations with indicated final salt concentrations and 15% PEG-8000 as a crowding agent in Droplet formation buffer (50mM

Tris-HCl pH 7.5, 10% glycerol, and 1 mM DTT). The protein solution was placed on a slide glass, covered with a cover slip, and sealed with nail polish to prevent evaporation. Slides were then imaged with an inverted Nikon A1R laser scanning confocal microscope equipped with a 100× NA 1.45 oil Objective. Images were captured in solution away from the surfaces of the slide glass and the cover slip.

FRAP assays were performed using the Nikon A1Rsi laser scanning confocal microscope equipped with a Plan Apo 100× NA 1.45 oil Objective, and Nikon Elements software. A circular region of interest was bleached once using 20% laser power.

Immunostaining

Cells grown on No. 1 glass coverslips (Matsunami) in a 6-well plate were fixed in 4% paraformaldehyde in PBS (Wako) for 10 min at RT. After three washes in PBS for 5 min, cells were permeabilized with 0.25% Triton X-100 in PBS for 5 min at RT. Following three washes in PBS containing 0.1% Tween 20 (PBST) for 5 min, cells were blocked with 5% goat serum albumin (Vector Laboratories) for 30 min at RT and incubated with primary antibodies in 5% goat serum albumin overnight at RT. After three washes in PBST, primary antibody was recognized by secondary antibodies (Goat anti-mouse IgG Alexa Fluor 488 Thermo Fisher Scientific A11001 1:1000 dilution, and Goat anti-rabbit IgG Alexa Fluor 546 Thermo Fisher Scientific A11010 1:1000 dilution) along with 1 mg/ml Hoechst 33258 (Dojindo) in the dark for 1 hour. Cells were washed three times with PBST. Glass slides were mounted onto slides with Fluoromount (Diagnostic BioSystems). Coverslips were sealed with transparent nail polish and stored at 4°C. Images were acquired with the SpinSR10 spinning disk confocal microscope with a 100× UPLAPO OHR (NA 1.5) oil Objective using a Hamamatsu ORCA-Flash4.0 CMOS camera (Hamamatsu Photonics) and OLYMPUS cellSens Dimension software, or with the inverted Nikon A1Rsi laser scanning confocal microscope system equipped with a 60× Pla Apo (NA 1.40) oil Objective and a 20× Plan Apo (NA, 0.75) DIC Objective using Nikon Elements software.

Image analysis

Images acquired with the Nikon A1Rsi confocal microscope were analyzed using ImageJ (1.53c) software (Schindelin *et al*., 2012; Schneider *et al*., 2012). Nuclear boundaries and puncta were automatically identified using the custom macro programs based on a method developed at Duke University Light Microscopy Core Facility (https://microscopy.duke.edu/guides/count-nuclear-foci-ImageJ). Briefly, nuclear masks were created using images stained with Hoechst 33258 by applying the Particle analysis function and the Watershed function of ImageJ, after adjusting parameters on 4-5 images. Similarly, cytoplasmic masks were created using cytoplasmic standard staining such as ACTB. The masks were applied to the acquired images to measure fluorescence intensities in individual nucleic/cytoplasmic regions. To count the number of puncta in each nuclear region, the acquired images were processed with the Find maxima function of ImageJ with a threshold of 22,000, after isolating nuclear regions using the nuclear mask. For the analysis of the overlay between MED1-puncta and MED4-puncta, nuclear regions in MED4-images were first isolated using the nuclear mask as stated above. Next, the fluorescent intensities of the images were adjusted between 0-65,535 arbitrary units using the Enhance contrast function (saturated = 0.35) of ImageJ. As MED4-puncta were stained with anti-MED4 antibody prominently in the nuclei (Figs 5E and 5F), masks of MED4-pancta were created using the Particle analysis function with thresholds of intensity = 65,500 and size = 2-to-infinity. Then, the masks were applied to the staining of MED1, to measure the intensities of MED1 overlapping with individual MED4-puncta. More than 50 nuclei in more than five randomly selected visual fields of at least two independent immunostainings were analyzed in each experiment.

Western blotting

Whole cell lysates were harvested as follows. Cells were washed with ice-cold PBS twice and harvested by scraping in ice-cold PBS. After centrifugation at 1,500×*g* for 5 min, the pellets were resuspended in buffer A (10 mM HEPES-NaOH pH 7.8, 10 mM KCl, 0.1 mM EDTA, 1 mM DTT, 0.5 mM PMSF, 0.1% NonidetP-40, and 1xProteaseInhibitor Cocktail) and kept for 30 min on

ice. After sonication, samples were centrifuged at 20,000×*g* for 5 min to remove cell debris. Western blotting was performed as previously described (Masuda *et al*., 2015).

MTS assay

Cell numbers were estimated using MTS assay (Promega) according to the manufacturer's instructions. Briefly, cells reverse-transfected with the indicated siRNAs were plated in 96-well plates (1,000 cells per well). At the indicated time points after the transfection, 20 μl of CellTiter 96 AQueous One Solution Reagent was added to each sample. After additional 3 h of incubation at 37 °C, the conversion of MTS into the aqueous soluble formazan was measured by the absorbance at 490 nm. The average 490 nm absorbance of the "no cell" control wells was subtracted from all other absorbance values to yield corrected absorbances, which are proportional to the number of living cells.

Antibodies

Anti-ACTB (sc-47778), anti-ARID1A (sc-32761), anti-HNRNPK (sc-32307), anti-MED4 (sc-398179), and control mouse IgG were purchased from Santa Cruz Biotechnology. Anti-CPSF6 (A301-356) and anti-MED1 (A300-793A) were purchased from Bethyl Laboratories. Anti-SRSF3 was purchased from MBL. Anti-MED12, anti-SMARCC1 (D7F8S), and anti-SMARCC (D8O9V9) were purchased from Cell Signaling Technology. Anti-MED1 (HPA052818) and anti-FLAG (M2) was purchased from Sigma-Aldrich. AntiMED17 (GTX115241) was purchased from GeneTex.

Meta analysis

RNA-seq datasets of RBP-depleted cells were extracted from the GEO database using the approved names of RBPs as keywords, such as SRSF1 to 11, hnRNP A to M, RBFOX1/2/3, FUS, SFPQ, PTBP1/2, SRRM4 and TDP-43 (Table EV1). When two or more RBPs were simultaneously depleted in a dataset, the dataset was excluded from the analysis. The extracted RNA-seq datasets were downloaded in sra format by prefetch command in SRA Toolkit (v2.8.1,

http://ncbi.github.io/sra-tools/), and were converted to fastq format using fastq-dump (v 2.8.1). Adaptor sequences were auto-detected and trimmed by Trim Galore (v0.6.4, http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and Cutadapt (version 1.18) (Martin, 2011) with parameters: –j 8. Reads were mapped to the referene genome sequences (mm10 or hg19) using STAR (version 2.5.2b) (Dobin *et al*., 2013) and uniquely mapped reads were used for subsequent analysis using following parameters: --outFilterMultimapNmax 1. To detect and quantify AS events, Percent-Spliced-Ins (PSIs) of all internal exons in the RefSeq annotation were calculated by MISO (version 0.5.3.) (Katz *et al*., 2010). Each RNA-seq sample was compared to its corresponding control RNA-seq. Exons with $|\Delta PSI| > 0.2$, Bayes factor > 10 and the number of exclusion reads > 1 were considered to be differentially spliced, using filter_events implemented in MISO (--delta-psi 0.20 --bayes-factor 10 --num-exc 1). Datasets including 20 or more differentially spliced exons were used for further analysis. The exon lengths differentially skipped or included by a given RBP were estimated as follows. First, the number and the median-length of differentially spliced exons in individual dataset were calculated. Then, an average value of the median exon-lengths of the RBP-depleted datasets was calculated. All the analyzed RBPs were ranked by the calculated average exon lengths and plotted with the bubble sizes representing the number of the exons.

Gene set enrichment analysis (GSEA)

For the calculation of statistical significances in Figs 2A, 7A, EV1E, EV1F, Appendix Fig S1A, S2B, and S2C, we used GSEA (Subramanian *et al*., 2005) software as follows; All RBP-depleted datasets were ranked according to the indicated value (e.g. %constitutive exons, fraction of S3-LCEs in exons with eCLIP-peaks, average fraction of disordered residues, maximum difference of PSI, MaxEntScan score). We made our own “gene sets” that were comprised of pairs of RNA-seq data with and without RBP depletion, and conducted GSEAPreranked analysis with default parameters.

ENCODE data analysis

The datasets of the differential splicing in the RNA-seqs of shRNA-treated cells were downloaded from the ENCODE portal (https://www.encodeproject.org/). Exons with Diff > 0.2 and $q < 0.01$ were regarded as significant. Datasets that contained more than 50 significantly skipped exons were used for the analysis.

The datasets of the significant peaks detected in eCLIPs were downloaded from the ENCODE portal. In Figures 2A and EV2G, the number of S3-LCEs with one or more eCLIP-peaks was divided by the number of all the other exons with one or more eCLIP-peaks in each eCLIP-seq data. The ratios obtained from a replicate of eCLIP-seqs of an RBP were averaged.

GTEx data analysis

To analyze PSI values of exons across human tissues (Figs 1D and EV1D), we downloaded the junction read counts on mapped junctions from the V7 release of the GTEx Consortium (2015) (https://www.gtexportal.org/home/, dbGap accession phs000424.v7.p2). GENCODE.v19 annotation from GTEx portal (https://storage.googleapis.com/gtex_analysis_v7/reference/gencode.v19.genes.v7.patched_contigs.exons.txt) was used for exonic annotation. We treated the reads mapped on an exon and the junction reads bridging to the upstream or downstream neighboring exon as the reads supporting exon inclusion. Similarly, we treated the reads spanning the upstream and downstream neighboring exons as the reads supporting exon exclusion. We first calculated the PSI of an exon in each RNA-seq data, according to the following equation:

PSI = (0.5× inclusion-supporting reads)/(exclusion-supporting reads + 0.5×inclusion-supporting reads).

PSI values calculated based on more than 10 reads (inclusion-supporting reads + exclusion-supporting reads) were used for the analysis. Then, we computed the median value of PSI within each tissue. We used PSI values of exons in genes with transcripts per million reads (TPM) > 1 in individual tissues for the analysis.

TPM values of genes in publicly available 11,688 RNA-seq data of 53 human tissues were obtained from GTEx portal (https://storage.googleapis.com/gtex_analysis_v7/rna_seq_data/GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_tpm.gct.gz). To calculate correlation between *SRSF3*-expression levels and *HNRNP*s-expression levels, TPM values of each gene were averaged for each tissue and Spearman's correlation coefficients between the averaged TPM values of *SRSF3* and those of an *HNRNP* across the 53 tissues were calculated.

Motif analysis

In Fig 1E, we first made 12 sets of S3-LCEs that were specifically skipped by SRSF3-depletion ($|\Delta PSI| > 0.2$ and Bayes factor > 10) from 12 RNA-seq datasets of SRSF3-depleted cells. We also made 12 sets of control exons that were two exons away from the skipped exon. We next counted the number of all possible 5-mer nucleotides in each exon. The frequency of a 5-mer nucleotide was calculated by dividing the number of the 5-mer nucleotide by the total number of all 5-mer nucleotides in each set of S3-LCEs or the control exons. Then, we averaged the frequencies of a 5-mer nucleotide in 12 sets of S3-LCEs and in 12 sets of the control exons. Then, a fold enrichment and a *P*-value by *t*-test of the average values between S3-LCEs and control exons were calculated. In Fig 7C, fold enrichments and statistical significances of all possible dipeptides were similarly calculated.

GO analysis

GO term enrichment analysis of S3-LCEs was performed using g:Profiler (Raudvere *et al.*, 2019). First, genes with S3-LCEs were sorted in order of the length of the exons. Then, the enrichment analysis was conducted with ordered query option using the ordered genes. Only Molecular Functions and Biological Process with no Electronic GO annotations were used. The GO network was generated with the Enrichment Map (Merico *et al.*, 2010) plugin for Cytoscape 3.8.0 (Shannon *et al.*, 2003) with the following parameters: *P*-value cut-off = 0.001; FDR Q value cut-off

= 0.1; Jaccard + Overlap Combined option with cut-off = 0.375; and Combined Constant = 0.5. GO terms were clustered by the Markov cluster algorithm with similarity coefficient using Auto Annotate 1.3.3 plugin, and assigned a label based on their representative GO terms with the lowest *P*-value within the cluster.

GO analysis shown in Fig 8E were conducted on genes with the internal coding exons that have more than 30% of C-nucleotide and 180-nt length for each species (see below). Genes were sorted in order of the length of the exons. Then, the enrichment analysis was conducted with ordered query option using the ordered genes using g:Profiler. Only Molecular Functions were used. The top 30 GO terms enriched in human genes were selected and the hierarchical clustering analysis was performed using Ward's method. The $-\log_{10}$ *P*-values of GO terms are proportionally adjusted between 0 to 1 in each species using the JMP software (relative *P*-values).

Prediction of disordered regions

Human protein sequences were obtained from Ensembl (ftp://ftp.ensembl.org/). Disordered regions in the entire sequences of all protein isoforms were predicted using IUPred2a (Mészáros *et al*., 2018) with long option and a cutoff of 0.5. To calculate the fraction of disordered residues in an exon, the amino acid sequence of a specific protein was segmented into each exon according to the Ensembl annotation. Then, the number of amino acids in disordered regions in an exon was divided by the number of amino acids in the exon to calculate the fraction of disordered residues. PONDR (Romero *et al*., 2001) and DisEMBL-hot loops (Linding *et al*., 2003) were also used in Figs EV4A and EV5B, respectively, to predict disordered regions in the indicated proteins.

Comparisons of multiple species

Genomic sequences, proteomic sequences, and gene transfer format (GTF) files for the analyzed 10 species were downloaded from the Ensembl database (*e.g.* Homo sapiens GRCh38, Macaca mulatta Mmul8.0.1, Mus musculus GRCm38, Monodelphis domestica monDom5, Gallus gallus GRCg6a, Anolis carolinensis AnoCar2.0, Xenopus tropicalis JGI4.2, Danio rerio GRCz11,

Drosophila melanogaster BDGP6, and Caenorhabditis elegans WBcel235). Disordered regions in the entire protein sequences were predicted as described above. An exon length, the composition of nucleic acids, the composition of amino acids, and a fraction of disordered residues were calculated for each internal coding exon. All internal coding exons in individual species were divided into 50 groups according to exon lengths as follows. First, the 171,182 human internal exons were ordered by the exon length and divided into 50 groups so that each group has nearly equal number of exons. Then, all internal exons in the other species were divided into 50 groups with the same ranges of exon lengths as the human groups. The divergence times indicated in Fig 8D were extracted from the TimeTree database (http://www.timetree.org/) (Kumar *et al*., 2017).

To estimate enrichment of codons in large exons across multiple species, we used 265 species of which genomic sequence and transcripts annotation are available in the Ensembl database. Coding sequence files and GTF files of 265 analyzed species were downloaded from the Ensembl database (release-101). We calculated the ratio of each codon in each exon and the exon length including single-exon genes, because *Saccharomyces* has only ~300 introns. To evaluate the relations between the codon ratios and the lengths, a regression coefficient in a linear regression model was calculated with statistical analysis for each codon in large (> 180nt) exons (Fig 8F) and all coding exons (Fig EV5H). Regression coefficients of 61 codons were ranked in descending order. Regression coefficients with $p \geq 0.01$ were excluded from the ranks, and the ranks were normalized so that the values ranged from 1 to 61 in each species. The 265 species and 61 codons were clustered by Ward's method using the JMP software with the imputation functionality.

Data availability

The accession number for the RNA-seq and tRIP-seq data reported in this paper is Gene Expression Omnibus GSE161601 and GSE161602, respectively. The files listing PSI values and Bayes factors of all internal exons obtained by the MISO-analysis of GEO datasets were deposited in BioStudies (accession number, S-BSST654).

Acknowledgments

We wish to acknowledge Mr. Kentaro Taki at the Division for Medical Research Engineering, Nagoya University Graduate School of Medicine for technical support of the mass spectrometry analysis. This study was supported by Grants-in-Aids from the Japan Society for the Promotion of Science [JP18K06058, JP20K06925, JP18K06483, JP16H06279 (PAGS), and JP18K14684]; the Ministry of Health, Labour, and Welfare of Japan (20FC1036); the Japan Agency for Medical Research and Development (JP19gm1010002, JP20ek0109488, and JP19bm0804005); the Naito Foundation; and the Intramural Research Grant for Neurological and Psychiatric Disorders of NCNP (29-4).

Author contributions

T.K., A.M. and K.O. designed the experiments and wrote the manuscript. T.K. and A.M. executed experiments and analyzed the data. T.K. and J.T. designed and performed the computational analyses. Y.Y., B.O. and M.I. helped with the experiments.

Conflict of interest

The authors declare no competing financial interests.

References

Afanasyeva A, Bockwoldt M, Cooney CR, Heiland I, Gossmann TI (2018) Human long intrinsically disordered protein regions are frequent targets of positive selection. *Genome Res* 28: 975-982.

Allen, B.L., and Taatjes, D.J. (2015). The Mediator complex: a central integrator of transcription. *Nat Rev Mol Cell Biol* 16: 155-166.

Auyeung, V.C., Ulitsky, I., McGeary, S.E., and Bartel, D.P. (2013). Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. *Cell* 152: 844-858.

Banani, S.F., Lee, H.O., Hyman, A.A., and Rosen, M.K. (2017). Biomolecular condensates: organizers

of cellular biochemistry. *Nat Rev Mol Cell Biol* 18: 285-298.

Banerjee, S., and Chakraborty, S. (2017). Protein intrinsic disorder negatively associates with gene age in different eukaryotic lineages. *Mol Biosyst* 13: 2044-2055.

Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* 6: 11.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R.*, et al.* (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science* 338: 1587-1593.

Boija, A., Klein, I.A., Sabari, B.R., Dall'Agnese, A., Coffey, E.L., Zamudio, A.V., Li, C.H., Shrinivas, K., Manteiga, J.C., Hannett, N.M.*, et al.* (2018). Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* 175: 1842-1855 e1816.

Bolisetty, M.T., and Beemon, K.L. (2012). Splicing of internal large exons is defined by novel cis-acting sequence elements. *Nucleic Acids Res* 40: 9244-9254.

Brown, C.J., Takayama, S., Campen, A.M., Vise, P., Marshall, T.W., Oldfield, C.J., Williams, C.J., and Dunker, A.K. (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol* 55: 104-110.

Bruce, S.R., and Peterson, M.L. (2001). Multiple features contribute to efficient constitutive splicing of an unusually large exon. *Nucleic Acids Res* 29: 2292-2302.

Buljan, M., Chalancon, G., Eustermann, S., Wagner, G.P., Fuxreiter, M., Bateman, A., and Babu, M.M. (2012). Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Molecular cell* 46: 871-883.

Busch, A., and Hertel, K.J. (2012). Evolution of SR protein and hnRNP splicing regulatory factors. *Wiley Interdiscip Rev RNA* 3: 1-12.

Chong, S., Dugast-Darzacq, C., Liu, Z., Dong, P., Dailey, G.M., Cattoglio, C., Heckert, A., Banala, S.,

Lavis, L., Darzacq, X.*, et al.* (2018). Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science* 361: eaar2555.

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57-74.

Consortium, G. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* 348: 648-660.

Corbo, C., Orru, S., and Salvatore, F. (2013). SRp20: an overview of its role in human diseases. *Biochem Biophys Res Commun* 436: 1-5.

Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. *Genome research* 14: 1188-1190.

Damianov, A., Ying, Y., Lin, C.H., Lee, J.A., Tran, D., Vashisht, A.A., Bahrami-Samani, E., Xing, Y., Martin, K.C., Wohlschlegel, J.A.*, et al.* (2016). Rbfox Proteins Regulate Splicing as Part of a Large Multiprotein Complex LASR. *Cell* 165: 606-619.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K.*, et al.* (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* 46: D794-D801.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15-21.

Ellis, J.D., Barrios-Rodiles, M., Çolak, R., Irimia, M., Kim, T., Calarco, J.A., Wang, X., Pan, Q., O'Hanlon, D., and Kim, P.M. (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. *Molecular cell* 46: 884-892.

Feng, H., Bao, S., Rahman, M.A., Weyn-Vanhentenryck, S.M., Khan, A., Wong, J., Shah, A., Flynn, E.D., Krainer, A.R., and Zhang, C. (2019). Modeling RNA-Binding Protein Specificity In Vivo by Precisely Registering Protein-RNA Crosslink Sites. *Mol Cell* 74: 1189-1204 e1186.

Fontrodona, N., Aube, F., Claude, J.B., Polveche, H., Lemaire, S., Tranchevent, L.C., Modolo, L., Mortreux, F., Bourgeois, C.F., and Auboeuf, D. (2019). Interplay between coding and exonic splicing regulatory sequences. *Genome Res* 29: 711-722.

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

Gonatopoulos-Pournatzis, T., Wu, M., Braunschweig, U., Roth, J., Han, H., Best, A.J., Raj, B., Aregger, M., O'Hanlon, D., Ellis, J.D.*, et al.* (2018). Genome-wide CRISPR-Cas9 Interrogation of Splicing Networks Reveals a Mechanism for Recognition of Autism-Misregulated Neuronal Microexons. *Mol Cell* 72: 510-524 e512.

Gueroussov, S., Weatheritt, R.J., O'Hanlon, D., Lin, Z.Y., Narula, A., Gingras, A.C., and Blencowe, B.J. (2017). Regulatory Expansion in Mammals of Multivalent hnRNP Assemblies that Globally Control Alternative Splicing. *Cell* 170: 324-339 e23.

Haeussler, M., Zweig, A.S., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Hinrichs, A.S., and Gonzalez, J.N. (2019). The UCSC genome browser database: 2019 update. *Nucleic acids research* 47: D853-D858.

Harlen, K.M., and Churchman, L.S. (2017). The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nat Rev Mol Cell Biol* 18: 263-273.

Hu, H., Miao, Y.-R., Jia, L.-H., Yu, Q.-Y., Zhang, Q., and Guo, A.-Y. (2019). AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic acids research* 47: D33-D38.

Irimia, M., Weatheritt, R.J., Ellis, J.D., Parikshak, N.N., Gonatopoulos-Pournatzis, T., Babor, M., Quesnel-Vallieres, M., Tapial, J., Raj, B., O'Hanlon, D.*, et al.* (2014). A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* 159: 1511-1523.

Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing

experiments for identifying isoform regulation. *Nat Methods* 7: 1009-1015.

Kim, S., and Shendure, J. (2019). Mechanisms of interplay between transcription factors and the 3D genome. *Molecular cell* 76: 306-319.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* 34: 1812-1819.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.

Lee, F.C.Y., and Ule, J. (2018). Advances in CLIP Technologies for Studies of Protein-RNA Interactions. *Mol Cell* 69: 354-369.

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., and Russell, R.B. (2003). Protein disorder prediction: implications for structural proteomics. *Structure* 11: 1453-1459.

Lubelsky, Y., and Ulitsky, I. (2018). Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature* 555: 107-111.

Mészáros, B., Erdős, G., and Dosztányi, Z. (2018). IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic acids research* 46: W329-W337.

Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* 27: 1696-1697.

Martin, M. (2011). Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads. *EMBnet. journal* 17: 10-12.

Masamha, C.P., Xia, Z., Yang, J., Albrecht, T.R., Li, M., Shyu, A.-B., Li, W., and Wagner, E.J. (2014). CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature* 510: 412-416.

Masuda A, Kawachi T, Ohno K (2021) Rapidly Growing Protein-Centric Technologies to Extensively

Identify Protein-RNA Interactions: Application to the Analysis of Co-Transcriptional RNA Processing. *Int J Mol Sci* 22: 5312.

Masuda, A., Kawachi, T., Takeda, J.I., Ohkawara, B., Ito, M., and Ohno, K. (2020). tRIP-seq reveals repression of premature polyadenylation by co-transcriptional FUS-U1 snRNP assembly. *EMBO Rep* 21: e49890.

Masuda, A., Takeda, J., Okuno, T., Okamoto, T., Ohkawara, B., Ito, M., Ishigaki, S., Sobue, G., and Ohno, K. (2015). Position-specific binding of FUS to nascent RNA regulates mRNA length. *Genes Dev* 29: 1045-1057.

Merico, D., Isserlin, R., Stueker, O., Emili, A., and Bader, G.D. (2010). Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PloS one* 5: e13984.

Mittal, P., and Roberts, C.W.M. (2020). The SWI/SNF complex in cancer - biology, biomarkers and therapy. *Nat Rev Clin Oncol* 17: 435-448.

Mizushima S, Nagata S (1990) pEF-BOS, a powerful mammalian expression vector. *Nucleic Acids Res* 18: 5322.

Muller-McNicoll, M., Botti, V., de Jesus Domingues, A.M., Brandl, H., Schwich, O.D., Steiner, M.C., Curk, T., Poser, I., Zarnack, K., and Neugebauer, K.M. (2016). SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export. *Genes Dev* 30: 553-566.

Nojima, T., Gomes, T., Carmo-Fonseca, M., and Proudfoot, N.J. (2016). Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. *Nat Protoc* 11: 413-428.

Oldfield, C.J., and Dunker, A.K. (2014). Intrinsically disordered proteins and intrinsically disordered protein regions. *Annual review of biochemistry* 83: 553-584.

Patel, A., Lee, H.O., Jawerth, L., Maharana, S., Jahnel, M., Hein, M.Y., Stoynov, S., Mahamid, J., Saha, S., Franzmann, T.M*., et al.* (2015). A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell* 162: 1066-1077.

Piva, F., Giulietti, M., Burini, A.B., and Principato, G. (2012). SpliceAid 2: a database of human splicing factors expression data and RNA target motifs. *Hum Mutat* 33: 81-85.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research* 47: W191-W198.

Robberson, B.L., Cote, G.J., and Berget, S.M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. *Molecular and cellular biology* 10: 84-94.

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., and Dunker, A.K. (2001). Sequence complexity of disordered protein. *Proteins: Structure, Function, and Bioinformatics* 42: 38-48.

Romero, P.R., Zaidi, S., Fang, Y.Y., Uversky, V.N., Radivojac, P., Oldfield, C.J., Cortese, M.S., Sickmeier, M., LeGall, T., and Obradovic, Z. (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proceedings of the National Academy of Sciences* 103: 8390-8395.

Sabari, B.R., Dall'Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A.V., Manteiga, J.C.*, et al.* (2018). Coactivator condensation at super-enhancers links phase separation and gene control. *Science* 361: eaar3958.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B.*, et al.* (2012). Fiji: an open-source platform for biological-image analysis. *Nat Methods* 9: 676-682.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature methods* 9: 671-675.

Schwich OD, Blumel N, Keller M, Wegener M, Setty ST, Brunstein ME, Poser I, Mozos IRL, Suess B, Munch C, McNicoll F, Zarnack K, Muller-McNicoll M (2021) SRSF3 and SRSF7 modulate 3'UTR length through suppression or activation of proximal polyadenylation sites and regulation of CFIm

levels. *Genome Biol* 22: 82.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* 13: 2498-2504.

Silveira, M.A.D., and Bilodeau, S. (2018). Defining the Transcriptional Ecosystem. *Mol Cell* 72: 920-924.

Smithers, B., Oates, M.E., and Gough, J. (2015). Splice junctions are constrained by protein disorder. *Nucleic acids research* 43: 4814-4822.

So BR, Di C, Cai Z, Venters CC, Guo J, Oh JM, Arai C, Dreyfuss G (2019) A Complex of U1 snRNP with Cleavage and Polyadenylation Factors Controls Telescripting, Regulating mRNA Transcription in Human Cells. *Mol Cell* 76: 590-599.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102: 15545-15550.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P.*, et al.* (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 47: D607-D613.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* 28: 511-515.

Ule, J., and Blencowe, B.J. (2019). Alternative Splicing Regulatory Networks: Functions, Mechanisms,

and Evolution. *Mol Cell* 76: 329-345.

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.* (2020). A large-scale binding and functional map of human RNA-binding proteins. *Nature* 583: 711-719.

Wang, Z., Qiu, H., He, J., Liu, L., Xue, W., Fox, A., Tickner, J., and Xu, J. (2020). The emerging roles of hnRNPK. *J Cell Physiol* 235: 1995-2008.

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 337: 635-645.

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.

Yandell, M., Mungall, C.J., Smith, C., Prochnik, S., Kaminker, J., Hartzell, G., Lewis, S., and Rubin, G.M. (2006). Large-scale trends in the evolution of gene structures within 11 animal genomes. *PLoS Comput Biol* 2: e15.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 11: 377-394.

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

Zhu, Y., Wang, X., Forouzmand, E., Jeong, J., Qiao, F., Sowd, G.A., Engelman, A.N., Xie, X., Hertel, K.J., and Shi, Y. (2018). Molecular mechanisms for CFIm-mediated regulation of mRNA alternative polyadenylation. *Molecular cell* 69: 62-74. e64.

Main figure titles and legends

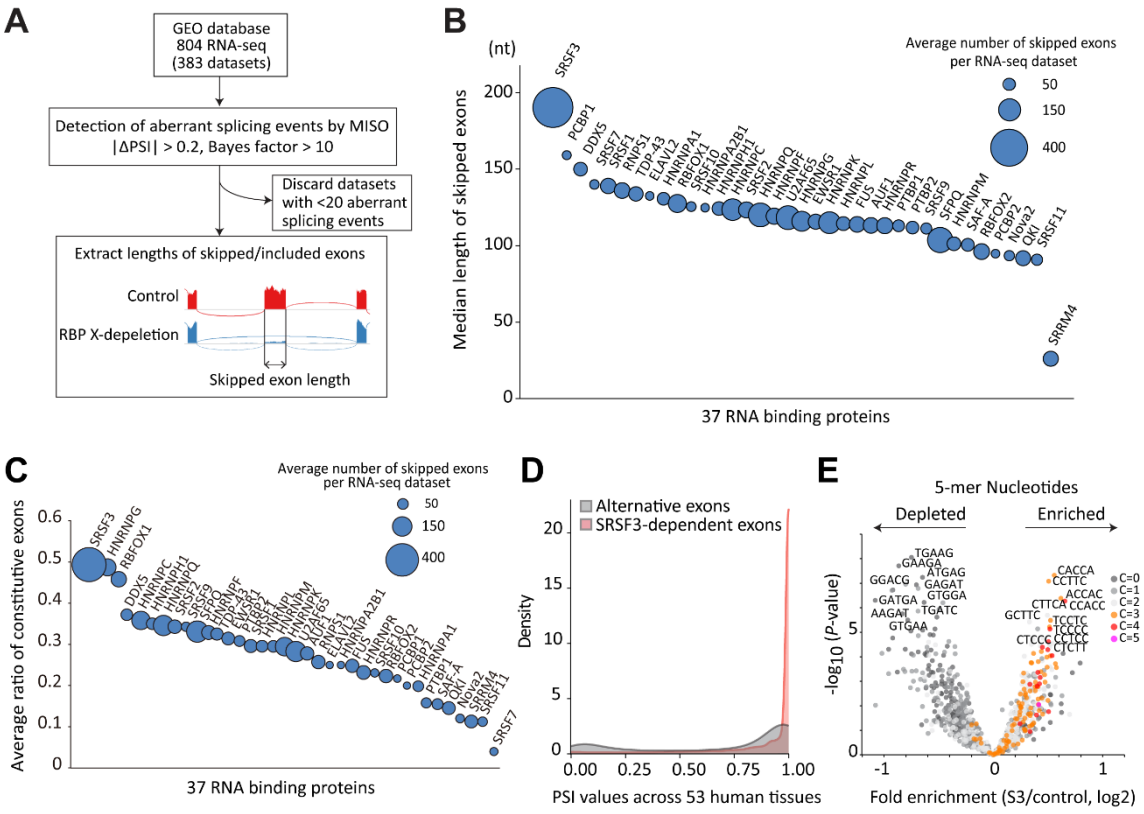

Figure 1. SRSF3 regulates splicing of large exons

(A) Workflow of the meta-analysis for RNA-seq of RNA binding protein (RBP)-depleted cells obtained from the GEO database. Each dataset consisted of RNA-seq of a pair of single RBP-depleted cells and control cells.

(B) Median length and number of exons skipped by each RBP depletion. The median length of the skipped exons in each dataset was calculated. The average of the median lengths was plotted in the ordinate for each RBP. Bubble sizes indicate the average number of skipped exons per dataset. The 37 RBPs were aligned in order of the average lengths of the skipped exons. The detailed data are shown in Table EV2.

(C) Average ratio of constitutive exons in the skipped exons by each RBP depletion. Bubble sizes indicate the average number of skipped exons per dataset. The 37 RBPs were aligned in order of the average ratio of constitutive exons in the skipped exons.

(D) Distributions of PSI values of SRSF3-dependent large constitutive exons (S3-LCEs) and UCSC known alternative exons across 53 human tissues. PSI values of an exon in individual RNA-seq were calculated, followed by calculation of their median value in each tissue. The calculated median values of S3-LCEs (red) or the alternative exons (gray) were extracted and their kernel densities are plotted in the graph.

(E) Frequencies of all possible 5-mer sequences in S3-LCEs. Fold changes in the frequencies between S3-LCEs and control exons are shown in the abscissa. The *P*-values are shown in the ordinate. The color code indicates a number of C-nucleotides in a 5-mer sequence.

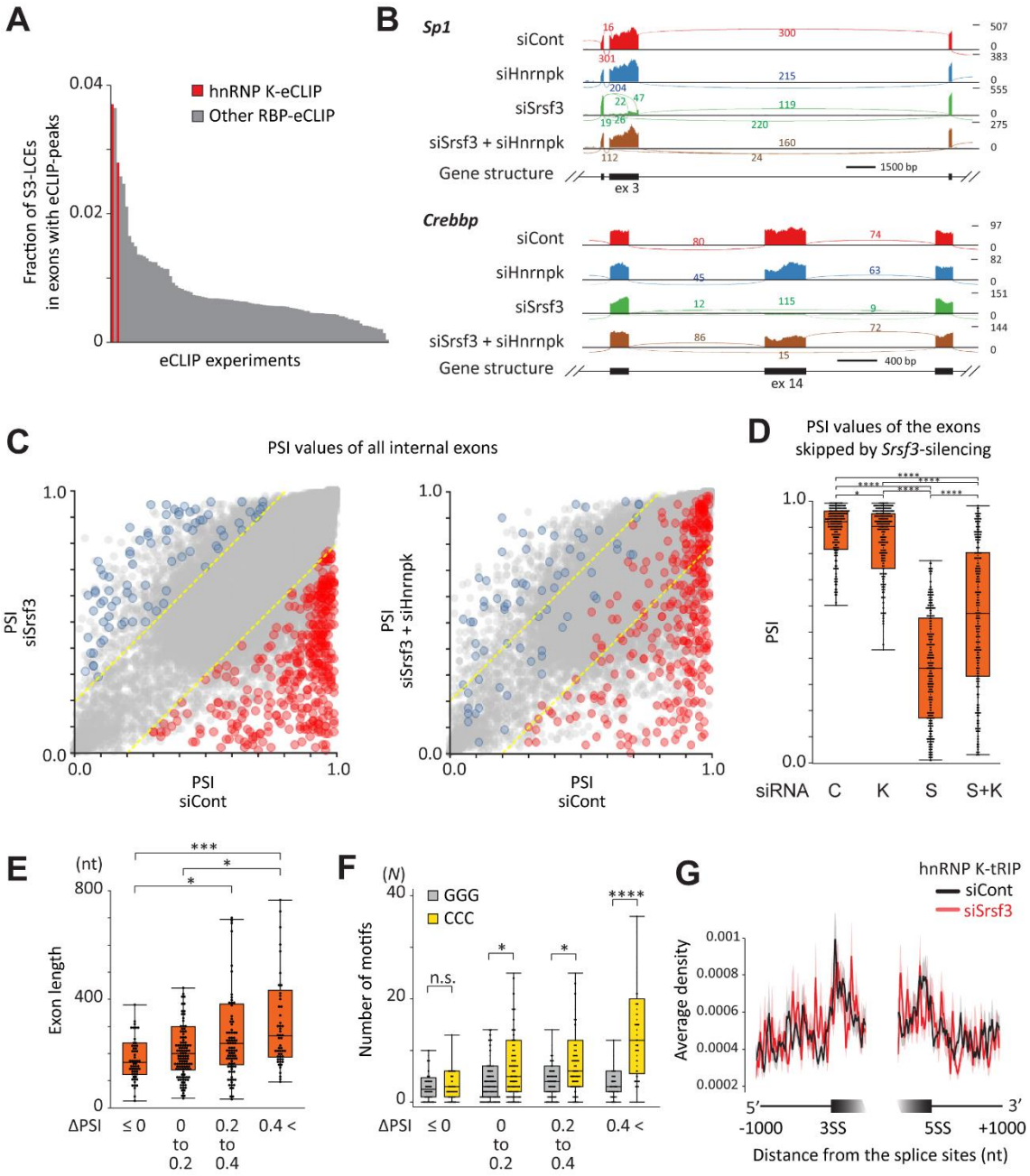

Figure 2. SRSF3 overrides splicing-suppressive activity of hnRNP K

(A) Enrichment of RBP–RNA interaction sites in the SRSF3-dependent large constitutive exons (S3-LCEs). Fraction of S3-LCEs in exons having eCLIP peaks in 334 eCLIP data in ENCODE. hnRNP K-eCLIPs are indicated in red.

(B) Sashimi plots of representative SRSF3-dependent exons. C2C12 cells were treated with control siRNA (siCont), *Srsf3* siRNA (siSrsf3), *Hnrnpk* siRNA (siHnrnpk), or both *Srsf3* and *Hnrnpk*

siRNAs (siSrsf3+siHnrnpk), and RNA-seq was performed. Lines indicate junction-spanning reads. Gene structures are shown below the Sashimi plots.

(C) Scatter plots showing PSI values of control cells (siCont), *Srsf3*-silenced cells (siSrsf3; left panel), and both *Srsf3*- and *Hnrnpk*-silenced cells (siSrsf3 + siHnrnpk; right panel). The PSI values of all internal exons are plotted. Yellow dotted lines indicate ΔPSI = 0.2 and -0.2. Skipped and included exons (|ΔPSI| > 0.2 and Bayes factor > 10) by siSrsf3 are indicated in red and blue, respectively. Note that exon skipping and inclusion are partly mitigated by an additional knockdown of hnRNP K (right panel).

(D) PSI values of the exons skipped by *Srsf3* silencing (red circles in Fig 2C) in cells treated with the control siRNA (C), *Srsf3* siRNA (S), *Hnrnpk* siRNA (K), or both *Srsf3* and *Hnrnpk* siRNAs (S+K). The box plot shows the interquartile range (boxes), the median (central band) and the minimum and maximum except for the outliers at the ends of whiskers (n = 344).

(E) Relationship between exon lengths and recovery of exon skipping in *Srsf3*-silenced cells by additional *Hnrnpk* silencing. The exons skipped by *Srsf3* silencing (red circles in Fig 2C) are classified into four groups according to the degree of recovery by the additional *Hnrnpk* silencing [$\Delta PSI\ (PSI_{siSrsf3+siHnrnpk} - PSI_{siSrsf3})$]. The box plot shows the interquartile range (boxes), the median (central band) and the minimum and maximum except for the outliers at the ends of whiskers (n = 60 for "ΔPSI ≤ 0", n = 131 for "ΔPSI 0 to 0.2", n = 95 for "ΔPSI 0.2 to 0.4", and n = 57 for "ΔPSI > 0.4").

(F) Relationship between the number of CCC motifs in exons and recovery of exon skipping in *Srsf3*-silenced cells by the additional *Hnrnpk* silencing. The exons skipped by *Srsf3* silencing (red circles in Fig 2C) are classified as in (E). The box plot shows the interquartile range (boxes), the median (central band) and the minimum and maximum except for the outliers at the ends of whiskers (n = 60 for "ΔPSI ≤ 0", n = 131 for "ΔPSI 0 to 0.2", n = 95 for "ΔPSI 0.2 to 0.4", and n = 57 for "ΔPSI > 0.4").

(G) Distributions of hnRNP K–RNA interactions around 3′ (left) or 5′ (right) splice sites (SS) of the SRSF3-dependent exons in control cells (black) and *Srsf3*-silenced cells (red). The standard

error of the average density of hnRNP K-tRIP reads is shown as a semi-transparent shade around the average curve.

Data information: *$p < 0.05$, ***$p < 0.001$, ****$p < 0.0001$ by Steel–Dwass test.

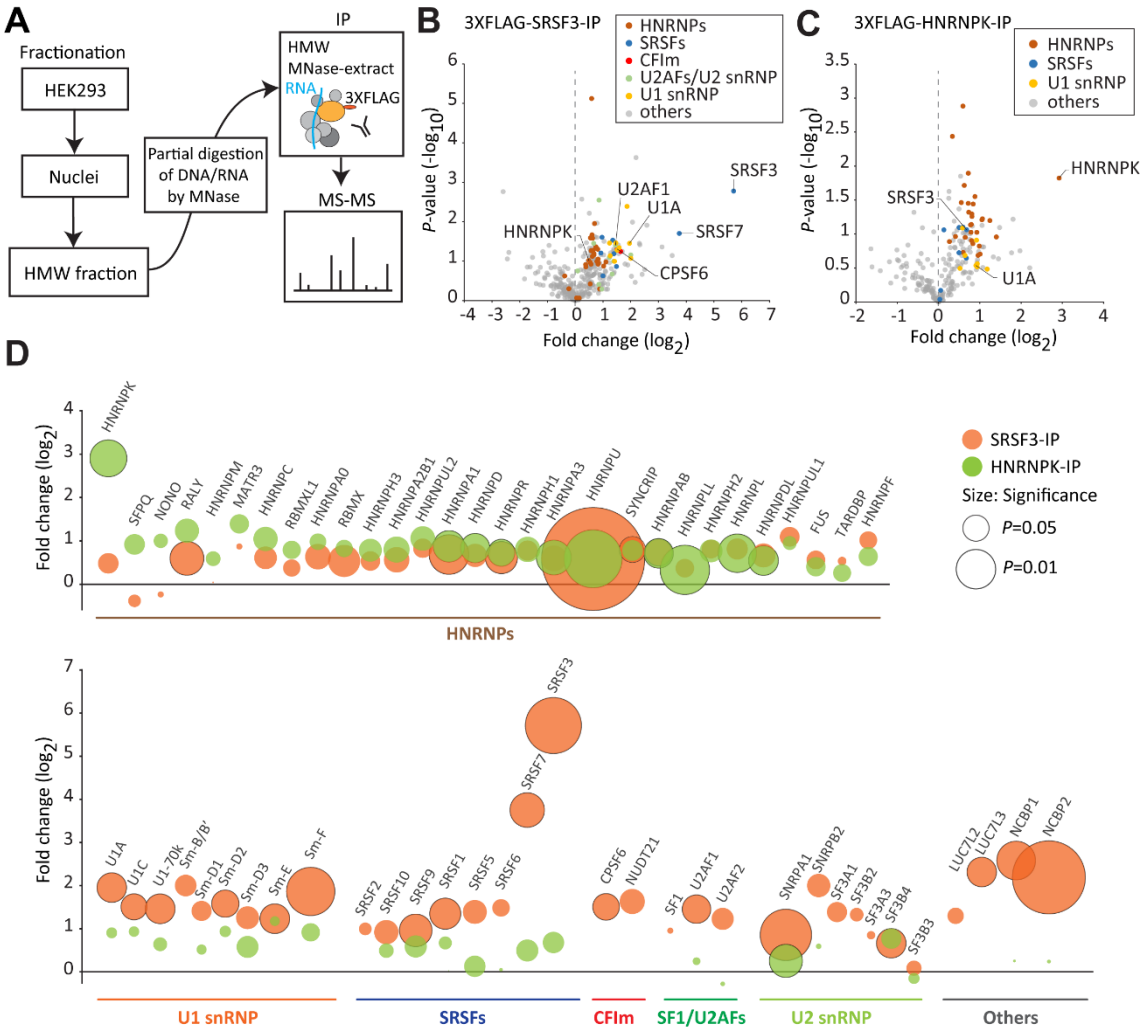

Figure 3. Interactome analysis for SRSF3 and hnRNP K

(A) Schematic diagram of the identification of protein complexes associated with SRSF3 and hnRNP K. Nuclei were extracted from HEK 293 cells overexpressing FLAG-tagged SRSF3 (3XFLAG-SRSF3-IP) or FLAG-tagged hnRNP K (3XFLAG-HNRNPK-IP), followed by isolation of the high molecular weight (HMW) fraction (Damianov *et al.*, 2016). After mild treatment with MNase to release proteins from chromatin, immunoprecipitation (IP) was performed using anti-FLAG antibody and mass spectrometry analysis was performed. Naïve cells not expressing a FLAG-tagged protein were used for control IP (Cont-IP).

(B) Volcano plot showing fold changes versus *P*-values of the normalized MASCOT scores (see Materials and Methods) of identified proteins between 3XFLAG-SRSF3-IP and Cont-IP. Each experiment was triplicated. The members of the representative families of splicing factors are

highlighted by colors (right upper box).

(C) Volcano plot showing fold changes versus *P*-values of the normalized MASCOT scores of identified proteins between 3XFLAG-HNRNPK-IP and Cont-IP. Each experiment was triplicated. The members of the representative families of splicing factors are highlighted by colors (right upper box).

(D) Bubble plot of MASCOT scores of representative proteins identified in 3XFLAG-SRSF3-IP (red) and 3XFLAG-HNRNPK-IP (green). *P*-values by paired *t*-test are indicated by bubble sizes.

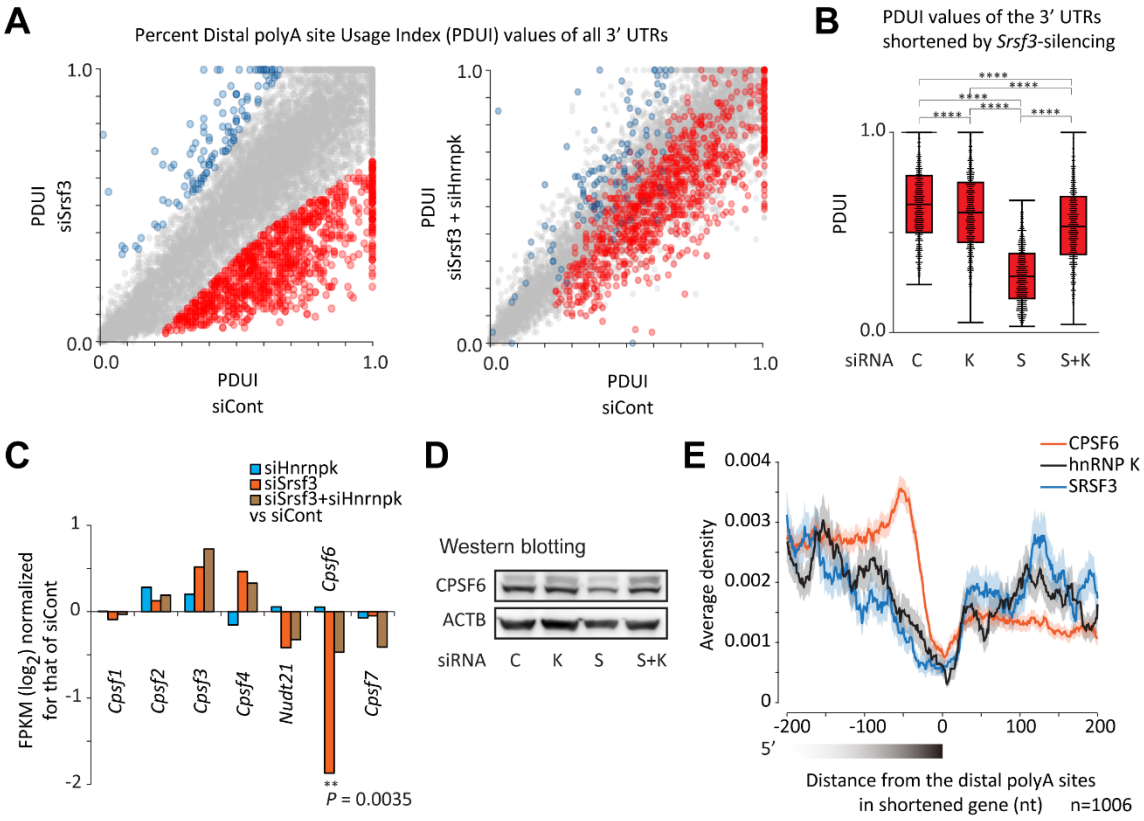

Figure 4. The SRSF3–hnRNP K axis regulates global selection of polyadenylation sites

(A) Scatter plots of percent distal polyA site usage indices (PDUIs) in control cells (siCont), *Srsf3*-silenced cells (siSrsf3; left panel), and both *Srsf3*- and *Hnrnpk*-silenced cells (siSrsf3 + siHnrnpk; right panel). PDUIs of all 3′ UTRs are plotted. The 3′ UTRs shortened and extended by siSrsf3 exceeding the thresholds of $|\Delta PDUI| > 0.2$ and $p < 0.01$ are indicated in red and blue, respectively.

(B) 3′ UTRs that were shortened by *Srsf3* silencing were first selected. PDUIs of these 3′ UTRs in C2C12 cells treated with control siRNA (C), *Srsf3* siRNA (S), *Hnrnpk* siRNA (K), or both *Srsf3* and *Hnrnpk* siRNAs (S+K) are plotted. ****$p < 0.0001$ by Steel–Dwass test. The box plot shows the interquartile range (boxes), the median (central band) and the minimum and maximum except for the outliers at the ends of whiskers ($n = 965$).

(C) mRNA expression levels of seven CPSF factors in C2C12 cells treated with the indicated siRNA normalized for those treated with the control siRNA. **$p < 0.01$ by CuffDiff (Trapnell *et all.*, 2010).

(D) Western blots showing protein expression levels of CPSF6 and β-actin (ACTB). Whole cell lysates were harvested 42 h after siRNA transfection.

(E) Distributions of interactions between RNA and CPSF6 (red), hnRNP K (black), or SRSF3 (blue) around the distal polyA sites that are downregulated in *Srsf3*-silenced cells. The standard error of the average density of tRIP reads is shown as a semi-transparent shade around the average curve.

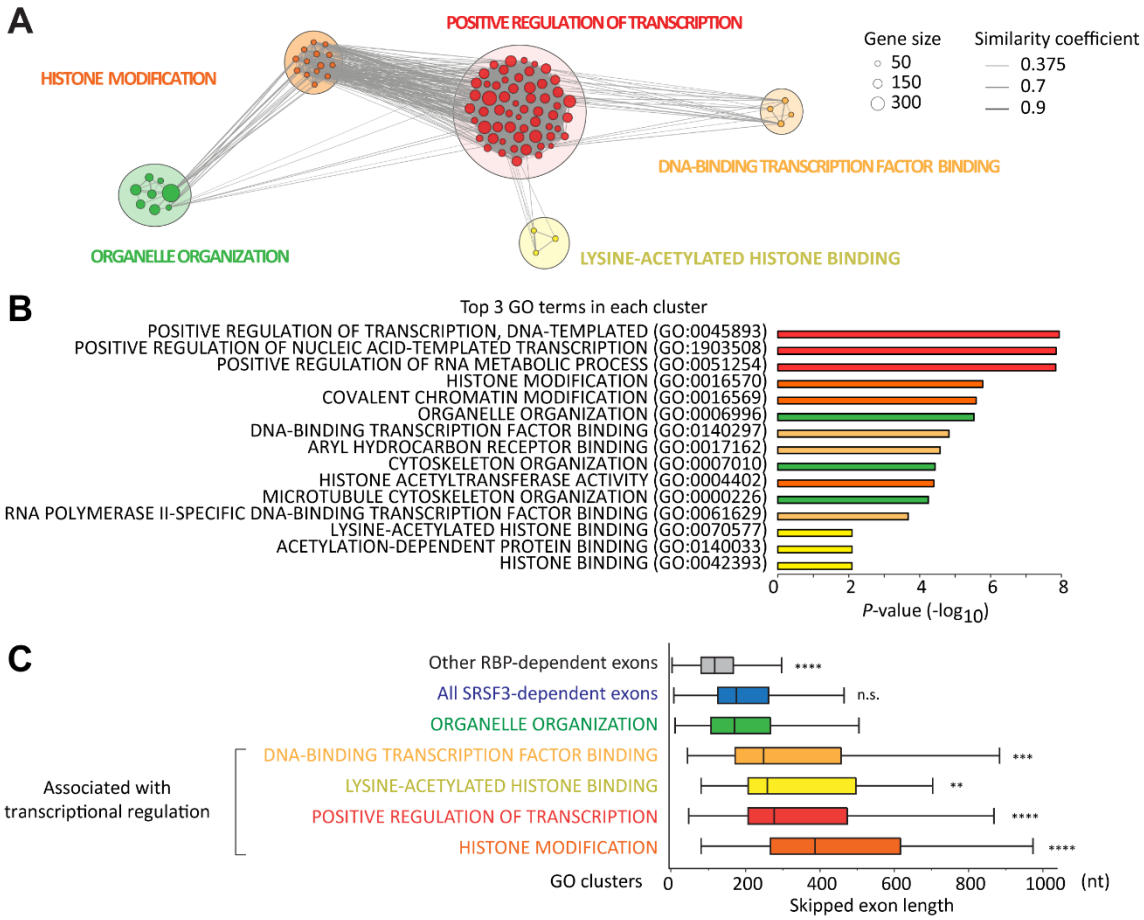

Figure 5. Genes for transcription factors are enriched in SRSF3-dependent large exons

(A) Clustering of GO terms enriched in S3-LCEs. GO terms were clustered using the Markov cluster algorithm. GO clusters are named by the GO term with the lowest *P*-value in each cluster. Node size represents the number of genes in the GO term, and the edge width represents the similarity coefficient between two GO terms.

(B) *P*-values of the top 3 GO terms contained in each GO cluster indicated in (A). GO terms were color-coded according to the corresponding colors in (A).

(C) Box plot showing lengths of S3-LCEs in each GO cluster. $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$ compared to "organelle organization" by Steel–Dwass test. The box plot shows the interquartile range (boxes), the median (central band) and the minimum and maximum except for the outliers at the ends of whiskers (n = 8,115 for "Other RBP-dependent exons", n = 3,078 for "All SRSF3-dependent exons", n = 595 for "ORGANELLE ORGANIZATION", n = 57 for

"DNA-BINDING TRANSCRIPTION FACTOR BINDING", n = 27 for "LYSINE-ACETYLATED HISTONE BINDING", n = 513 for "POSITIVE REGULATION OF TRANSCRIPTION", and n = 75 for "HISTONE MODIFICATION").

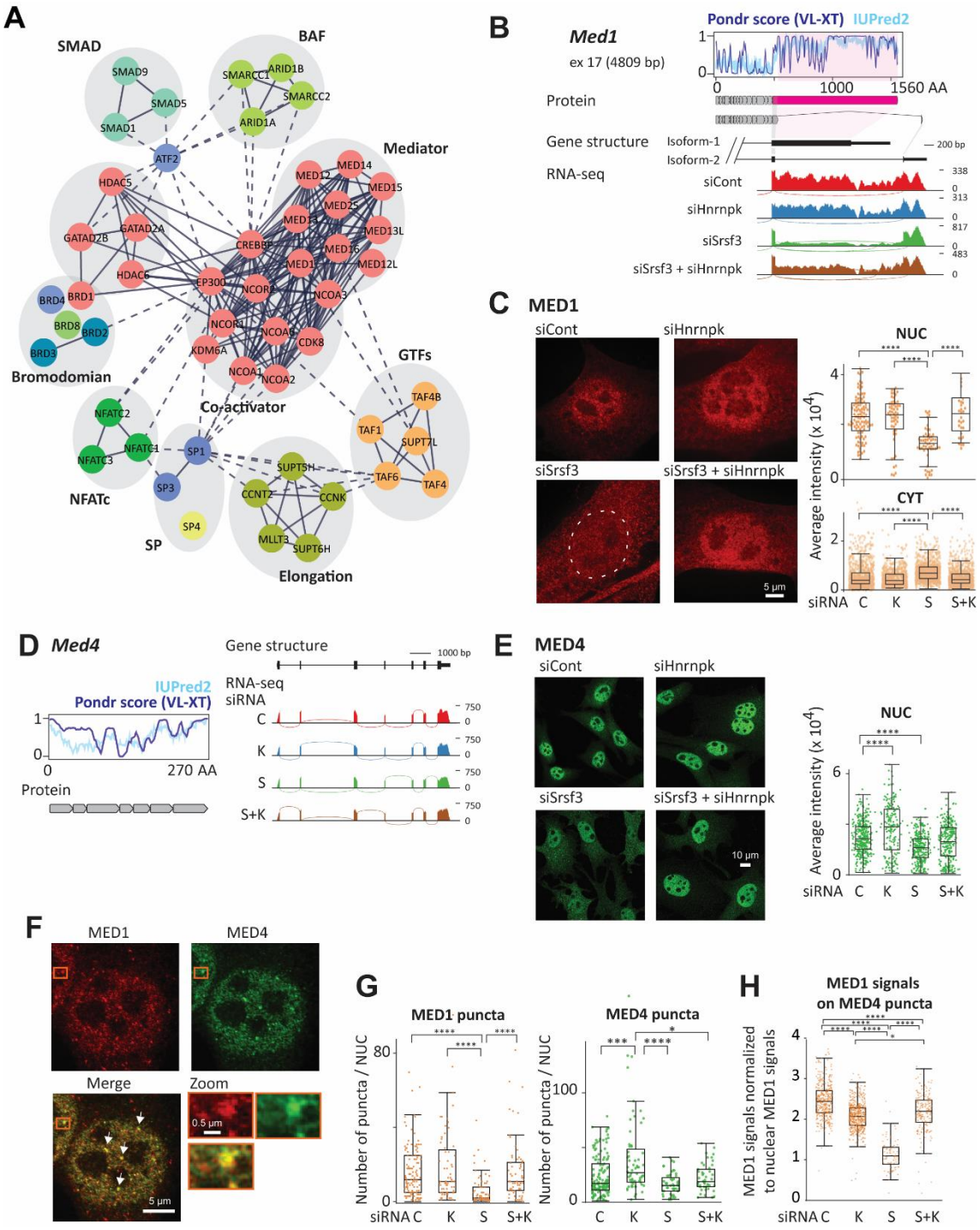

Figure 6. *Srsf3* silencing disrupts transcription assemblies

(A) Representative transcriptional factor genes containing the SRSF3-dependent exons. Edges show protein–protein interaction (PPI) as annotated within the STRING database (Szklarczyk *et al.*, 2019). Solid lines and dotted lines indicate PPIs within a cluster and over a cluster, respectively.

Genes are clustered by the Markov cluster algorithm.

(B) SRSF3-responsive alternative 3′ ends of the *Med1* gene. (Top panel) Disorder analysis for MED1 using PONDR VL-XT (deep blue) and IUPred2 (light blue). The scores in the ordinate indicate disordered tendencies between 0 and 1 (a score of more than 0.5 indicates disordered). (Middle panel) Protein and gene structures of MED1. The protein structure was segmented into individual exonic regions. The SRSF3-responsive last exon is indicated in pink. (Bottom panel) Sashimi plots of RNA-seqs from *Srsf3*- and/or *Hnrnpk*-silenced cells.

(C) Immunofluorescence images of MED1 in C2C12 cells treated with siCont (C), siSrsf3 (S), siHnrnpk (K), or siSrsf3+siHnrnpk (S+K). A white dotted contour outlines the nucleus. Box plots (right) show quantification of nuclear (NUC) and cytoplasmic (CYT) localization of MED1. The average intensity of one nucleus or a cytoplasmic segment is individually plotted.

(D) The intrinsically disordered region (IDR) in MED4. Disorder analysis, protein and gene structures, and Sashimi plots are shown in (B). Splicing of these genes was not affected by the siRNA treatment.

(E) Immunofluorescence images of MED4 in C2C12 cells treated with the indicated siRNAs. Box plots show quantification of nuclear localization of MED4. Each dot represents one nucleus.

(F) Immunofluorescence images of MED1 and MED4 in C2C12 cells The arrows denote representative puncta, including both MED1 and MED4.

(G) The number of MED1 puncta (left) and MED4 puncta (right) per nucleus in cells treated with the indicated siRNA.

(H) MED1 signals on MED4 puncta normalized to nuclear MED1 signals. MED1 signal intensity on each MED4 punctum was individually plotted.

Data information: *$p < 0.01$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$ by Steel–Dwass test. In (C, D, G, and H), box plots show the interquartile range (boxes), the median (central band) and the minimum and maximum except for the outliers at the ends of whiskers. More than 50 nuclei in more than five randomly selected visual fields of at least two independent immunostainings were analyzed in each experiment.

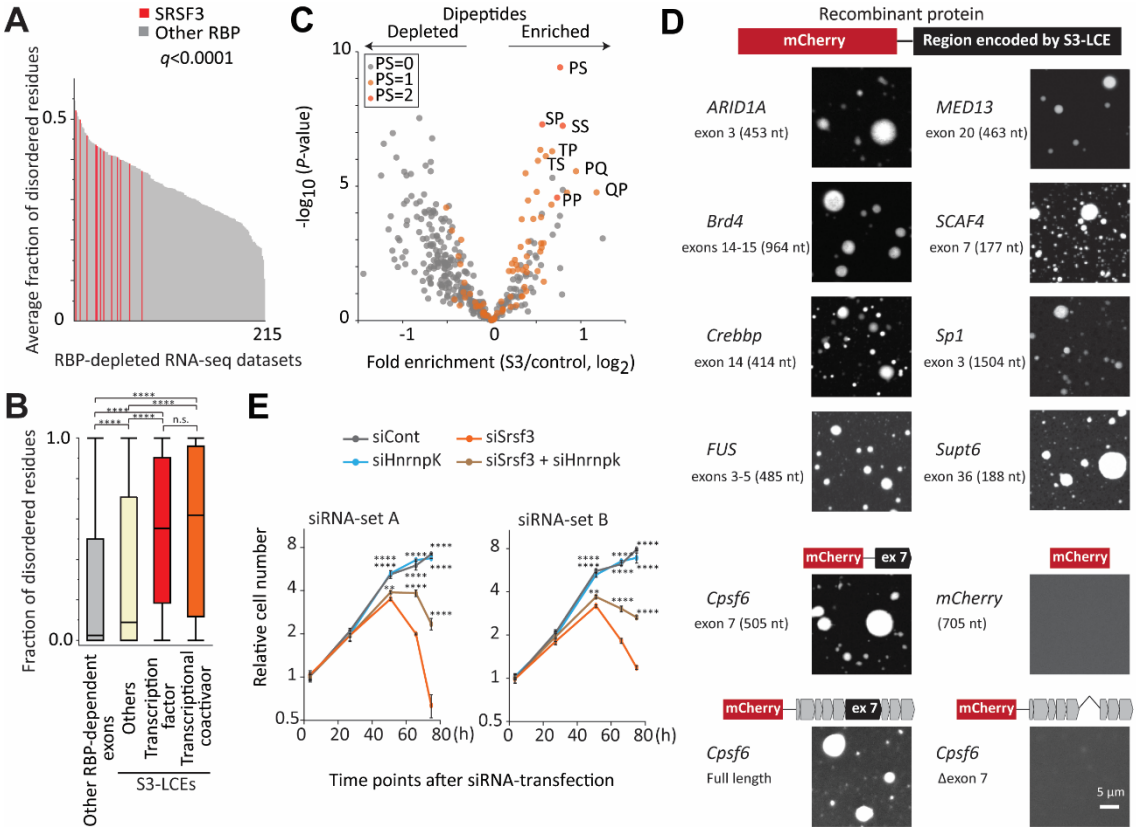

Figure 7. SRSF3-dependent exons encode IDRs of molecules involved in transcriptional regulations

(A) Average fractions of disordered residues in the skipped exons in each RBP-depleted RNA-seq dataset. SRSF3-depleted datasets are indicated in red. The *q*-value between SRSF3-depleted datasets and other datasets was calculated according to the gene set enrichment analysis (GSEA).

(B) Average fractions of disordered residues in S3-LCEs were included in the indicated categories. The genes containing S3-LCEs are classified into 3 groups according to the AnimalTFDB (Hu *et al*., 2019). The fraction in other RBP-dependent exons (gray box) is shown for comparison. ****$p < 0.0001$ by Steel–Dwass test. The box plot shows the interquartile range (boxes), the median (central band) and the minimum and maximum except for the outliers at the ends of whiskers (n = 6,005 for "Other RBP-dependent exons", n = 1,883 for "Others", n = 238 for "Transcription factor", and n = 251 for "Transcriptional coactivator").

(C) Frequencies of all possible dipeptide sequences in S3-LCEs. Fold changes and *P*-values are

shown as in Fig 1E. The color code indicates the number of P or S in a dipeptide.

(D) Droplet formation of mCherry-fused recombinant proteins of the S3-LCEs. Representative images are shown below the schematic of the mCherry fusion protein. Detailed information about the indicated S3-LCEs is shown in Fig EV4B. The recombinant proteins (20 nM) were incubated with 75 mM NaCl in droplet formation buffer.

(E) Temporal profiles of the MTS assay representing the number of C2C12 cells after transfection with the indicated siRNA. The targeted siRNA sites were different between siRNA sets A and B (Fig EV2O). Relative cell number was normalized to that of control cells (siCont) at 3 h. Means and SD are indicated (n = 3 wells each). Data were analyzed with two-way ANOVA and Tukey's post-hoc test (**p < 0.01, ****p < 0.0001 compared to siSrsf3).

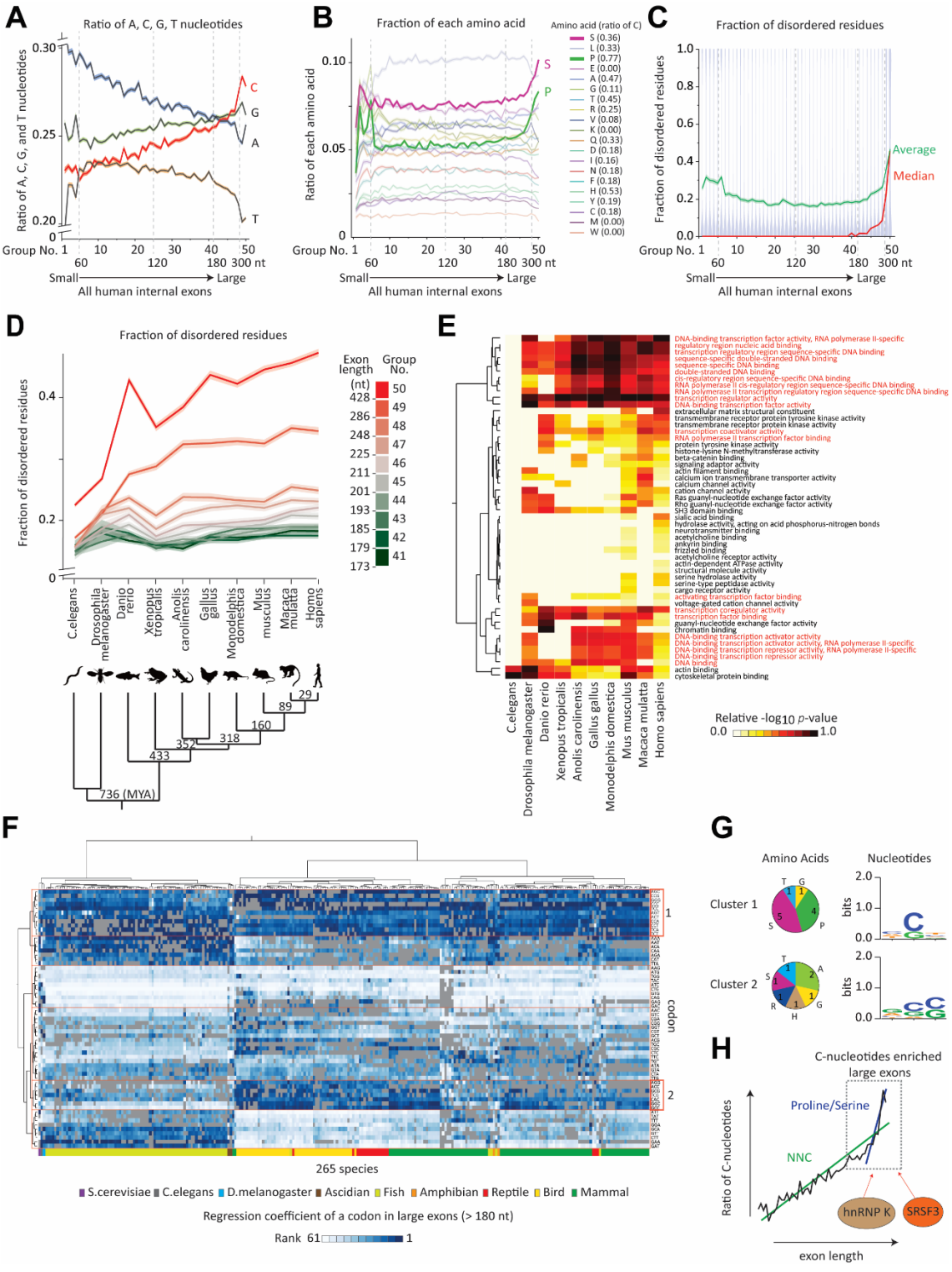

Figure 8. Vertebrate large exons evolutionarily enriched in C-nucleotides retain IDRs of transcription factors in splicing

(A) Average ratios of A, C, G, and T nucleotides in the human internal exons, which are evenly

divided into 50 groups according to their lengths. The mean and standard error (SE, semi-transparent shade) are plotted.

(B) Average ratios of amino acids in all human internal exons, which are evenly divided as in (A). The numbers in parentheses indicate the ratio of C-nucleotides in the codons considering codon usage in humans. The mean and SE (semi-transparent shade) are indicated. Note that proline and serine are frequently used in IDRs and are rich in C-nucleotides in their codons.

(C) Violin plots showing the fractions of disordered residues in all human internal exons, which are evenly divided as in (A). The mean (green) with standard error (SE, semi-transparent shade) and the median (red) of the fractions of disordered residues are indicated.

(D) Average fractions of disordered residues in large internal exons in different species. All internal exons are divided into 50 groups as in (A) based on their lengths. The last 10 groups (> 173 nt) are shown. The mean and SE are indicated. The phylogenetic tree of analyzed species with distance from humans is shown at the bottom.

(E) Clustering analysis and heatmap showing *P*-values of GO terms enriched in genes containing large (> 180 nt) and C-enriched (> 30%) internal exons in each species. The top 30 GO terms enriched in the human genes carrying large and C-enriched exons were selected. GO terms associated with transcription are indicated in red.

(F) Clustering and heatmap showing the relationship between the codon ratios and exon lengths in large exons (> 180 nt) in 265 species. The codon enrichment was estimated by linear regression between the codon ratios and exon lengths in large exons (> 180 nt). A regression coefficient with $p < 0.01$ is indicated by a heatmap, whereas a regression coefficient without statistical significance is indicated in gray. Codons preferentially used in large exons are color-coded in blue. Enlarged view with species names is shown in Appendix Fig S5.

(G) Pie charts showing the number of codons with amino acids in Clusters 1 and 2 in (F). Sequence logos created by the codons in each cluster are also shown.

(H) Schematic of C-nucleotide enrichment in exons and splicing regulation of large exons in vertebrates. The C-nucleotide content is biphasically increased with increasing exon length. The

first part is accounted for by an increase in C-nucleotide number at the third position of a codon (NNC, N stands for any nucleotide), which has marginal effects on amino acid compositions. The second part is accounted for by an increase in C-nucleotide number at the second position of a codon, which increases the IDR-constituting proline and serine content to secure the phase separation of transcription factors. The splicing-suppressive effect by hnRNP K is masked by SRSF3 so that C-rich large exons are constitutively spliced.

Figure EV legends

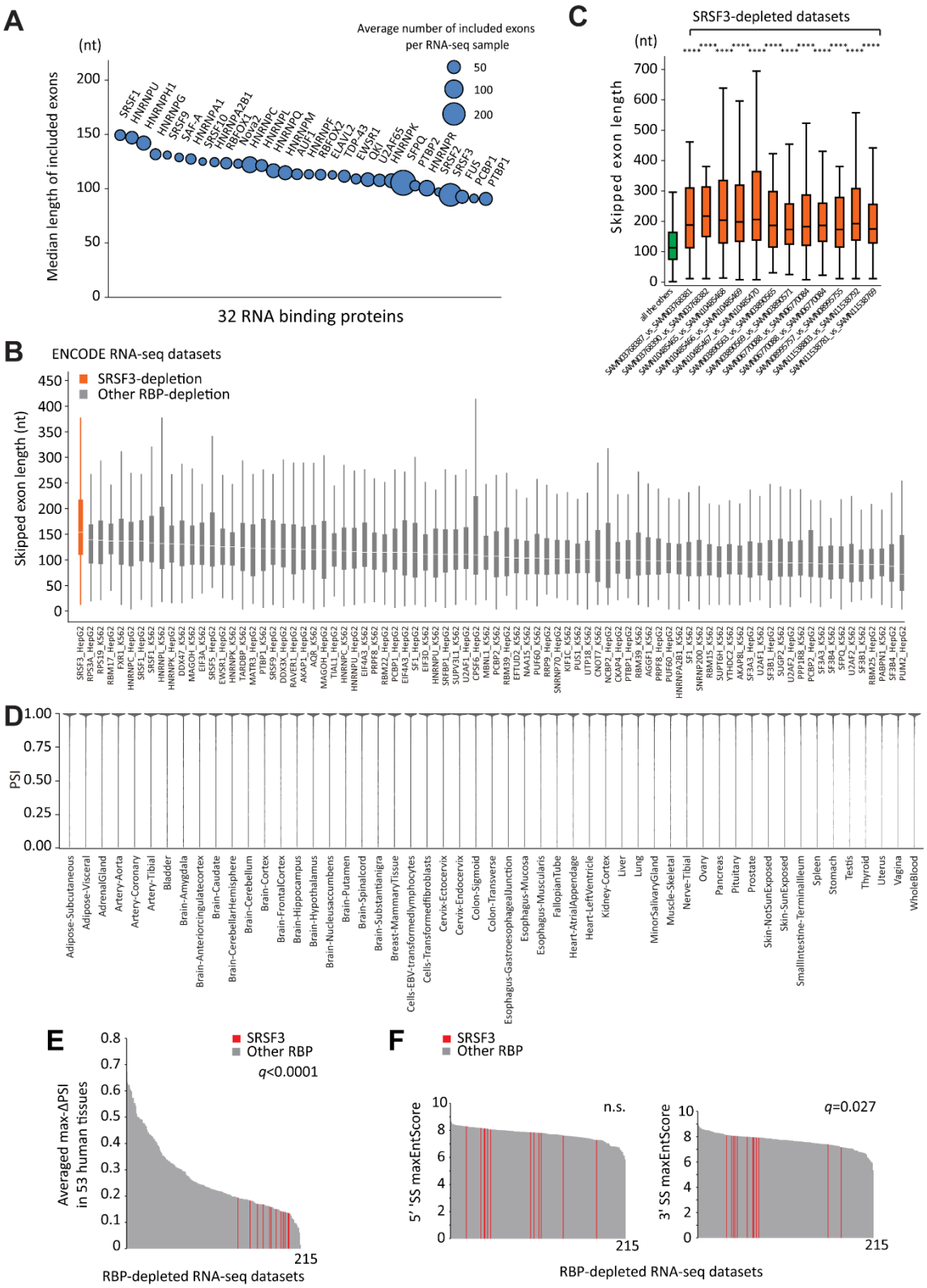

Figure EV1. SRSF3 is required for the constitutive splicing of large exons

(A) Median length and the number of exons included in each RBP depletion. In contrast to Fig 1B, which shows the analysis of exon skipping events, the analysis of exon inclusion events is shown in this panel. The median length of included exons was calculated as shown in Fig 1B. Bubble size indicates the average number of included exons per dataset.

(B) Box plot showing the length of exons skipped by the depletion of each RBP in 82 ENCODE datasets, in which more than 50 exons are skipped by RBP depletion. RBP-depleted RNA-seqs are aligned in order of the median lengths of the skipped exons. SRSF3 depletion is indicated in orange. The box plot shows the interquartile range (boxes), the median (central band) and the minimum and maximum except for the outliers at the ends of whiskers.

(C) Box plot showing lengths of skipped exons in individual RNA-seqs of SRSF3-depleted cells/tissues (orange). Those in all the other RBP-depleted datasets (green) are shown for comparison. All the lengths of the skipped exons in the SRSF3-depleted datasets were significantly larger than those in the other datasets (****$p < 0.0001$ by Steel–Dwass test). The box plot shows the interquartile range (boxes), the median (central band) and the minimum and maximum except for the outliers at the ends of whiskers.

(D) Violin plots showing the distributions of the PSI values of the S3-LCEs in the 53 human tissues shown in Fig 1D.

(E) Maximum differences in PSI values (max-ΔPSI) in 53 human tissues for individual exons. For every internal exon, a max-ΔPSI among the RNA-seqs of 53 human tissues was calculated using the GTEx database. The max ΔPSIs for each RBP-dependent exon in 53 human tissues were averaged for each dataset. The averaged max ΔPSIs are plotted in the graph, where the datasets are aligned in order of their values. SRSF3-depleted datasets (SRSF3-dependent exons) are shown in red. The *q*-value between SRSF3-depleted datasets and the other datasets was calculated according to GSEA.

(F) Mean strengths of 5′ splice sites (upper) and 3′ splice sites (lower) of skipped exons detected in each RNA-seq dataset for RBP depletion. Strengths of splice sites were estimated using MaxEntScan (Yeo and Burge, 2004). SRSF3-depleted datasets are indicated in red. The *q*-value

between SRSF3-depleted datasets and other datasets was according to GSEA.

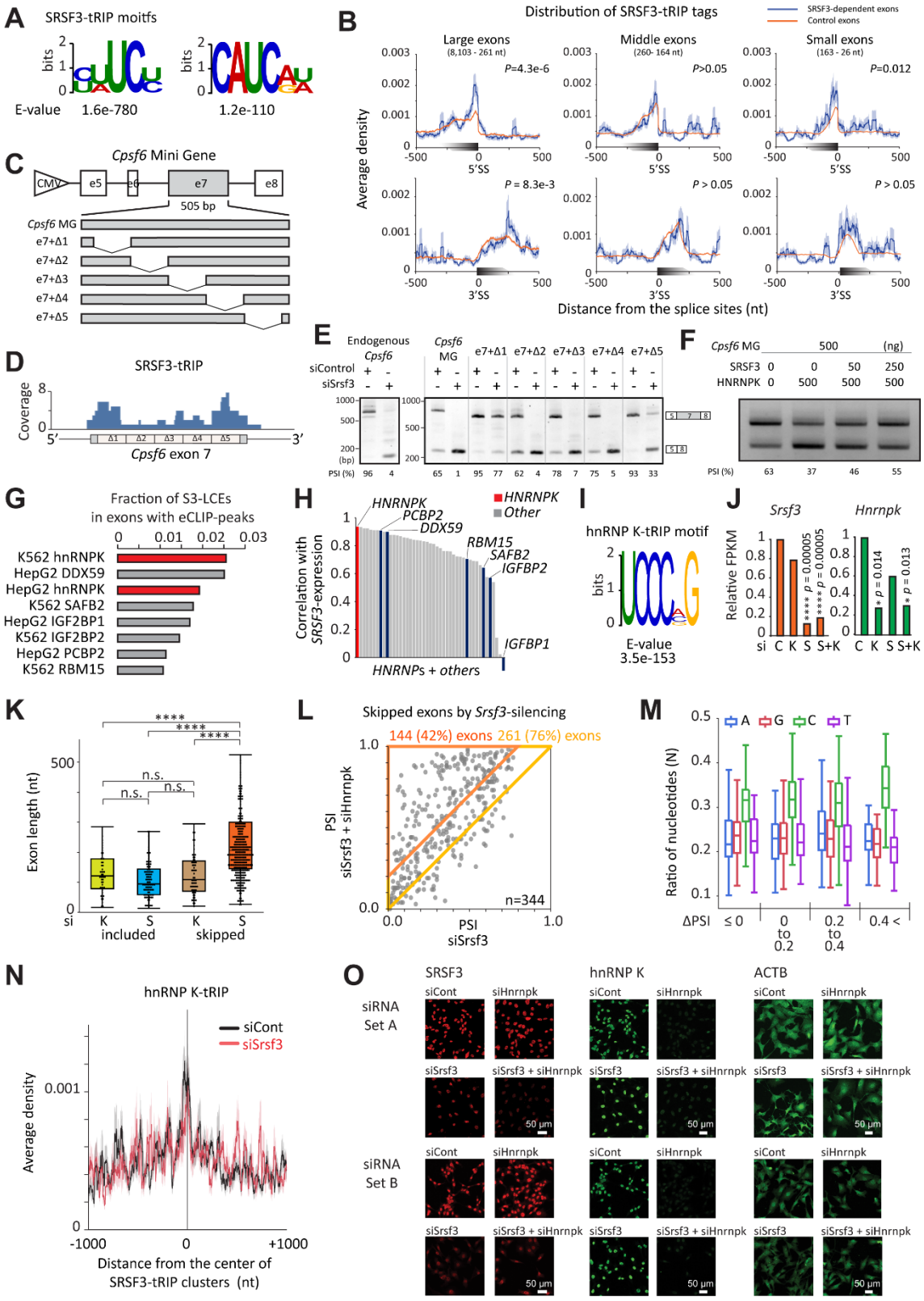

Figure EV2. SRSF3 overrides splicing-suppressive activity of hnRNP K

(A) The top two motifs identified by MEME-ChIP (Machanick and Bailey, 2011) around the peaks of SRSF3-tRIP-seq.

(B) Distribution of SRSF3–RNA interaction sites around the 5′ (upper) and 3′ (lower) splice sites (SS) of the SRSF3-dependent exons. The exons skipped in *Srsf3*-silenced cells (ΔPSI > 0.2 and Bayes factor > 10; SRSF3-dependent exons, blue lines) are evenly divided into three groups according to their lengths. The size-matched constitutive exons spliced irrespective of SRSF3 are used as controls (Control exons, red lines). The standard error of the average density of the SRSF3-tRIP reads is shown as a semi-transparent shade around the read coverage curve. *P*-value was calculated by Wilcoxon test with Bonferroni correction.

(C) The structure of the *Cpsf6* minigene (MG) construct spanning exons 5 to 8, which is driven by a cytomegalovirus promoter. A schematic of serial deletions of 99-bp blocks in *Cpsf6* exon 7 is shown below the structure.

(D) Distribution of SRSF3–RNA interaction sites in and around *Cpsf6* exon 7.

(E) RT-PCR of the endogenous *Cpsf6* mRNA as well as the wild-type and deletion minigene constructs in control-silenced (siCont) and *Srsf3*-silenced (siSrsf3) C2C12 cells. Amplified *Cpsf6* fragments are shown on the right side of the gel images. PSI values (%) of the exon 7 are indicated at the bottom of the panel.

(F) Effects of hnRNP K- and SRSF3-overexpressions on splicing of the wild-type *Cpsf6* minigene (*Cpsf6* MG) construct. C2C12 cells were transfected with 500 ng of *Cpsf6* MG along with the indicated amounts (ng) of SRSF3- and- hnRNP K-expression vectors. A total of 1500 ng DNA was introduced to the cells in each group by adding an empty vector (not indicated). PSI values (%) of the exon 7 are indicated at the bottom of the panel.

(G) The top eight eCLIP datasets with high fractions of SRSF3-dependent large constitutive exons (S3-LCEs) in exons with eCLIP peaks in the indicated cells and RBP.

(H) Correlation coefficients between the expression levels of SRSF3 and those of another RBP in 53 human tissues. The expression levels of hnRNP K were most correlated with those of SRSF3 in the six RBPs shown in Fig EV2G (blue) and 38 hnRNP proteins (gray). The expression

levels of RBPs were estimated by calculating the transcripts per million reads (TPM) using RNA-seq of 53 human tissues deposited in the GTEx database.

(I) A motif enriched by MEME-ChIP (Machanick and Bailey, 2011) around the peaks of hnRNP K-tRIP-seq.

(J) FPKM values of *Srsf3* and *Hnrnpk* in cells treated with the control siRNA (C), *Srsf3* siRNA (S), *Hnrnpk* siRNA (K), or both *Srsf3* and *Hnrnpk* siRNAs (S+K). FKPM values are normalized to the control siRNA (C) (relative RKPM). *$p < 0.05$ and ****$p < 0.0001$ by CuffDiff.

(K) Exon lengths of skipped and included exons ($|\Delta PSI| > 0.2$ and Bayes factor > 10) in *Srsf3*-silenced and *Hnrnpk*-silenced C2C12 cells. ****$p < 0.0001$ by Steel–Dwass test. The box plot shows the interquartile range (boxes), the median (central band) and the minimum and maximum except for the outliers at the ends of whiskers ($n = 49$ for "included by siHnrnpk", $n = 101$ for "included by siSrsf3", $n = 63$ for "skipped by siHnrnpk", and $n = 374$ for "skipped by siSrsf3").

(L) Scatter plot of PSI values of exons in *Srsf3*-silenced cells (siSrsf3) and both *Srsf3*- and *Hnrnpk*-silenced cells (siSrsf3 + siHnrnpk). Exons that were skipped by *Srsf3* silencing in Fig 2C (red circles) are plotted. *Hnrnpk* silencing mitigated 76% of the exon skipping events by *Srsf3* silencing with ΔPSI ($PSI_{siSrsf3+siHnrnpk} - PSI_{siSrsf3}$) > 0 (yellow) and 42% of them with $\Delta PSI > 0.2$ (orange).

(M) Relationship between the A, C, G, and T nucleotides in exons and the recovery of exon skipping in *Srsf3*-silenced cells by the additional *Hnrnpk* silencing. The skipped exons were classified into four groups as shown in Fig 2E. The Jonckheere–Terpstra trend test shows that only the ratios of C-nucleotides increase with increasing ΔPSI values ($p = 0.024$). The box plot shows the interquartile range (boxes), the median (central band) and the minimum and maximum except for the outliers at the ends of whiskers ($n = 60$ for "$\Delta PSI \leq 0$", $n = 131$ for "ΔPSI 0 to 0.2", $n = 95$ for "ΔPSI 0.2 to 0.4", and $n = 57$ for "$\Delta PSI > 0.4$").

(N) Distributions of hnRNP K-tRIP reads around the center of a cluster of SRSF3-tRIP reads on S3-

LCEs to show shared interaction sites of SRSF3 and hnRNP K. The standard error of the average density of hnRNP K-tRIP reads is shown as a semi-transparent shade around the average curve.

(O) Immunofluorescence images of SRSF3, hnRNP K, and β-actin (ACTB) in C2C12 cells treated with control siRNA (siCont), *Srsf3* siRNA (siSrsf3), *Hnrnpk* siRNA (siHnrnpk), or both *Srsf3* and *Hnrnpk* siRNAs (siSRSF3 + siHnrnpk) to show the efficiency of siRNA treatments. Cells were transfected with the indicated siRNAs, and the targeted sites of siRNAs were different between sets A and B. Where the set of siRNAs is not indicated in the figures, the siRNAs of set A have been used for the experiments.

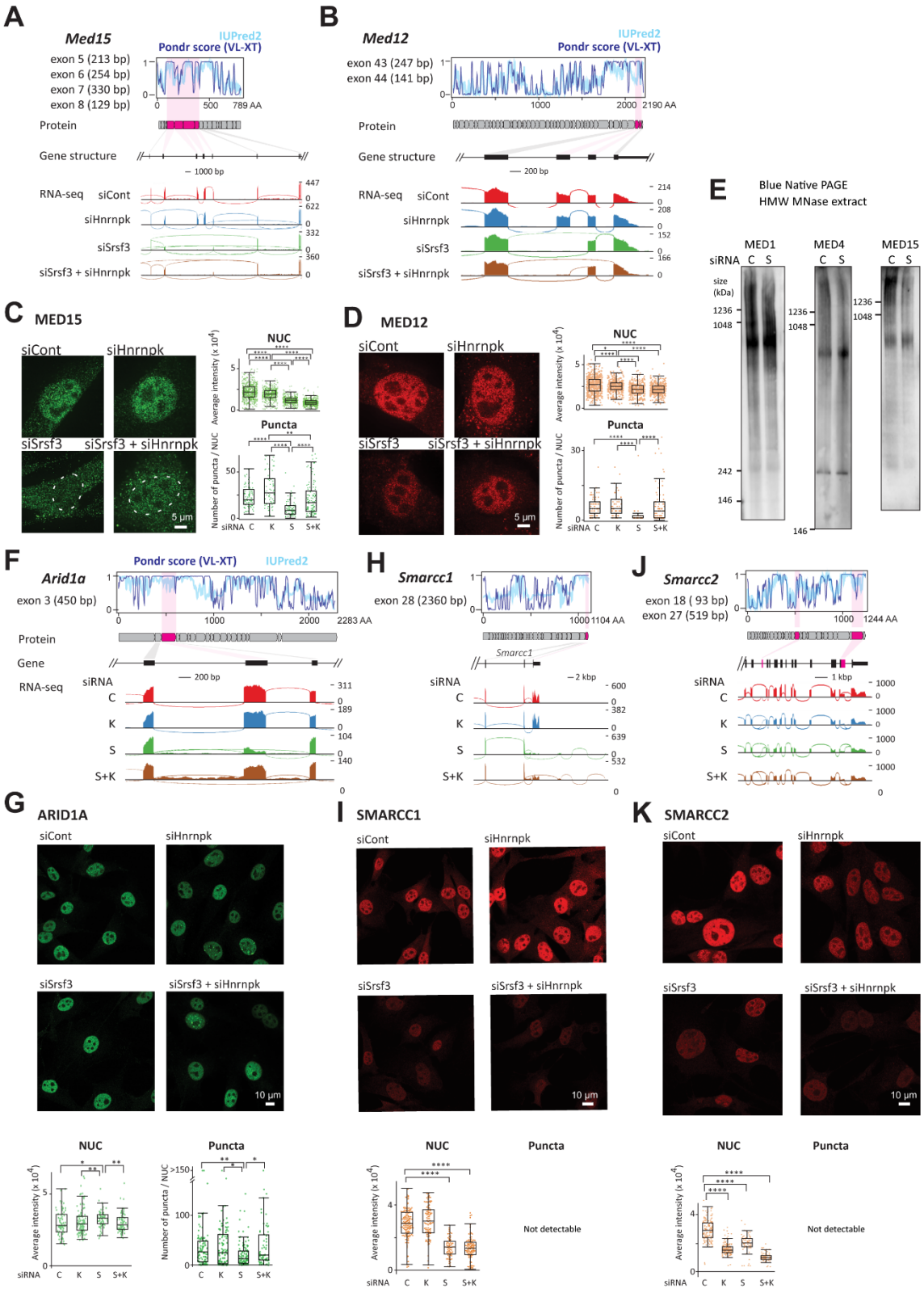

Figure EV3. *Srsf3* silencing disrupts the mediator complex and the BAF complex

(A, B) IDR and S3-LCEs in MED15 (A) and MED12 (B). Disorder analysis (top), protein and gene

structures (middle), and Sashimi plots (bottom) are shown as in Fig 6B. The scores in the ordinates of the top graphs indicate disordered tendencies between 0 and 1 (a score of more than 0.5 indicates disordered).

(C, D) Immunofluorescence images of MED15 (C) and MED12 (D) in C2C12 cells treated with siCont (C), siSrsf3 (S), siHnrnpk (K), or siSrsf3+siHnrnpk (S+K). The white dotted contours outline the nuclei. Box plots (right) show quantification of nuclear localization (NUC) and the number of puncta per nucleus (Puncta) in each siRNA treatment. Each dot represents the average intensity in each nucleus and the number of puncta in each nucleus.

(E) Native page analysis of mediator complexes formed in C2C12 cells treated with siCont (C) or siSrsf3 (S). High molecular weight (HMW) fractions of these cells were resolved by blue native page and were immunoblotted with the antibody against MED1, MED4, or MED15.

(F) The IDR and S3-LCEs in ARID1A. Disorder analysis (top), protein and gene structures (middle), and Sashimi plots (bottom) are shown as in Fig 6B.

(G) Immunofluorescence images of ARID1Ain C2C12 cells treated with the indicated siRNA. Box plots show quantification of nuclear localization (NUC) in each siRNA treatment. A box plot showing the number of puncta is also indicated. Each dot represents one nucleus.

(H) The IDR and S3-LCEs in SMARCC1. Disorder analysis (top), protein and gene structures (middle), and Sashimi plots (bottom) are shown as in Fig 6B.

(I) Immunofluorescence images of SMARCC1 in C2C12 cells treated with the indicated siRNA. Box plots show quantification of nuclear localization (NUC) in each siRNA treatment. Each dot represents one nucleus.

(J) The IDR and S3-LCEs in SMARCC2. Disorder analysis (top), protein and gene structures (middle), and Sashimi plots (bottom) are shown as in Fig 6B.

(K) Immunofluorescence images of SMARCC2 in C2C12 cells treated with the indicated siRNA. Box plots show quantification of nuclear localization (NUC) in each siRNA treatment. Each dot represents one nucleus.

Data information: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$ by Steel–Dwass test. In (C,

D, G, I, and K), box plots show the interquartile range (boxes), the median (central band) and the minimum and maximum except for the outliers at the ends of whiskers. More than 50 nuclei in more than five randomly selected visual fields of at least two independent immunostainings were analyzed in each experiment.

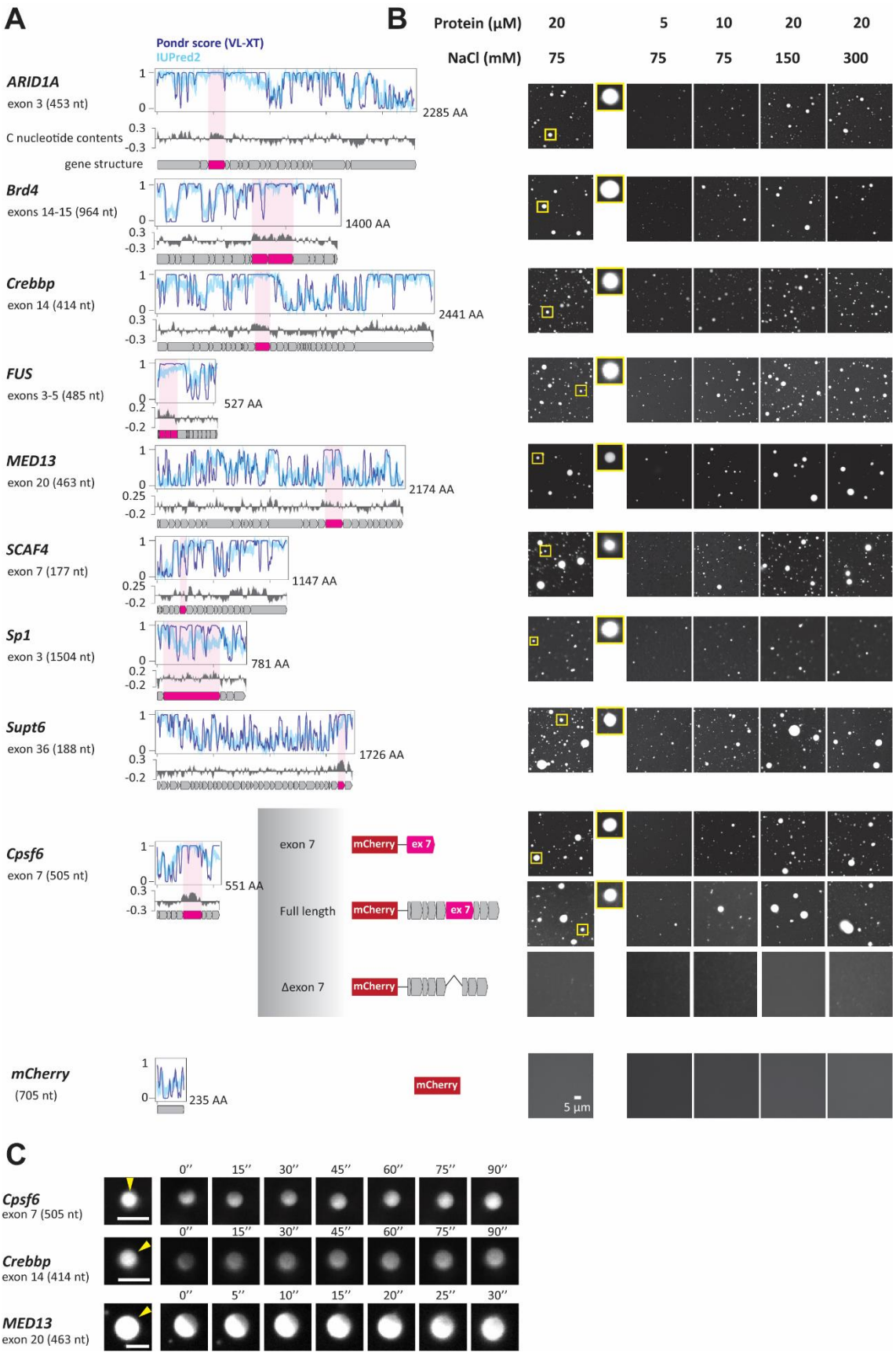
A
Pondr score (VL-XT)
IUPred2
ARID1A
exon 3 (453 nt)
2285 AA
C nucleotide contents
gene structure
Brd4
exons 14-15 (964 nt)
1400 AA
Crebbp
exon 14 (414 nt)
2441 AA
FUS
exons 3-5 (485 nt)
527 AA
MED13
exon 20 (463 nt)
2174 AA
SCAF4
exon 7 (177 nt)
1147 AA
Sp1
exon 3 (1504 nt)
781 AA
Supt6
exon 36 (188 nt)
1726 AA
Cpsf6
exon 7 (505 nt)
551 AA
exon 7
mCherry
ex 7
Full length
Δexon 7
mCherry
(705 nt)
235 AA
B
Protein (μM) 20 5 10 20 20
NaCl (mM) 75 75 75 150 300
5 μm
C
0'' 15'' 30'' 45'' 60'' 75'' 90''
Cpsf6
exon 7 (505 nt)
Crebbp
exon 14 (414 nt)
0'' 5'' 10'' 15'' 20'' 25'' 30''
MED13
exon 20 (463 nt)

Figure EV4. Phase separation of mCherry-fused recombinant proteins of S3-LCEs

(A) IDRs and S3-LCEs in the genes involved in transcriptional regulation. IDRs were predicted by PONDR VL-XT (dark blue) and IUPred2 (light blue). (Top) Disordered tendencies from 0 to 1 are indicated in the ordinate, where a score > 0.5 indicates disordered. (Middle) The difference between the ratio of C-nucleotides in a 50-nt sliding window is shown in the entire coding region. (Bottom) cDNAs, which are segmented into individual exons, are shown. S3-LCE is indicated in red.

(B) Representative images of droplet formation at different protein and salt concentrations. mCherry-fused recombinant proteins of the S3-LCEs were added to the indicated droplet formation buffer.

(C) FRAP recovery images of mCherry-fused recombinant proteins encoded by S3-LCEs. The arrows indicate the sites of photobleaching. Scale bar = 5 μm.

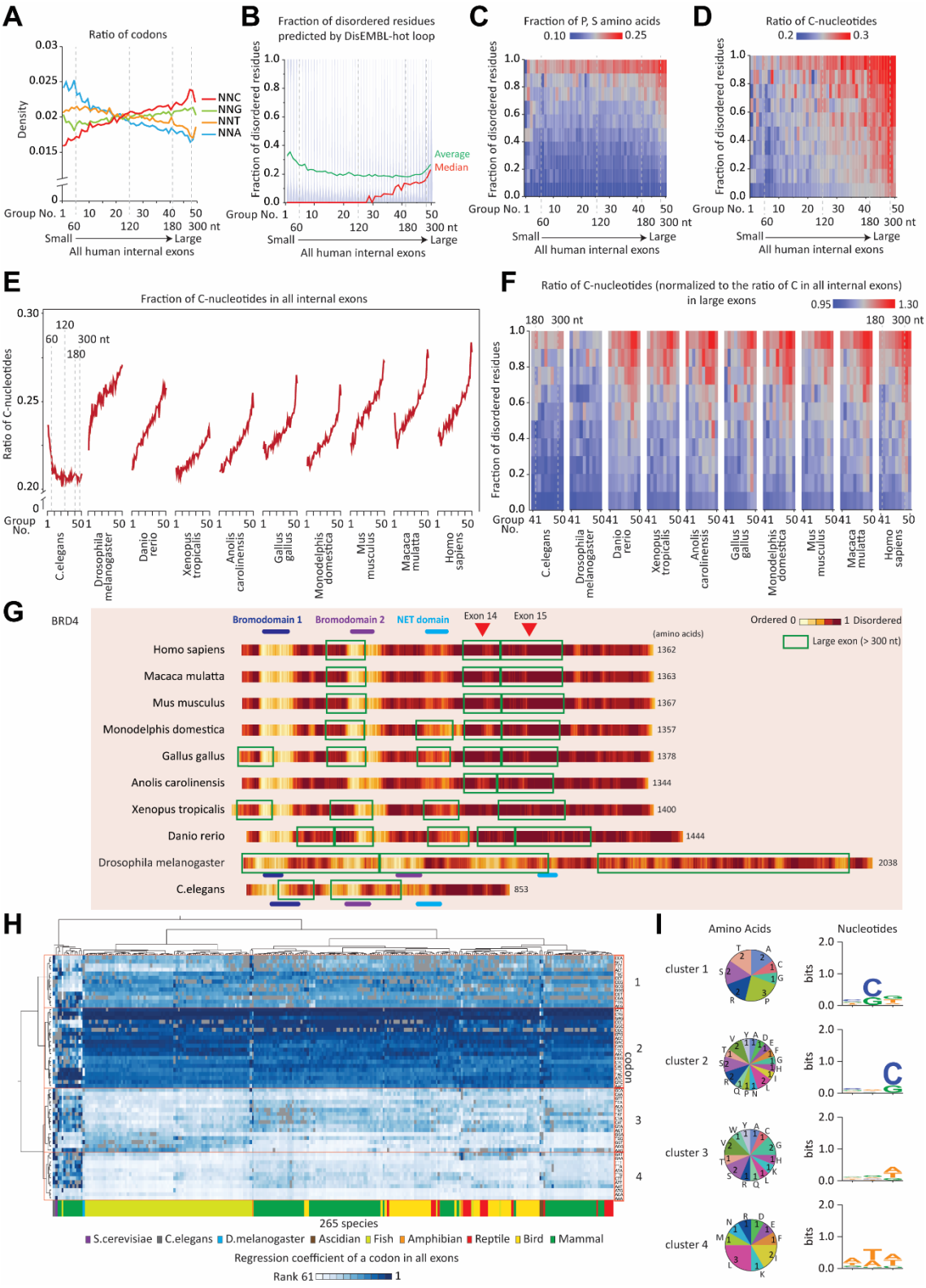

Figure EV5. Vertebrate large exons evolutionarily acquired enrichment of C-nucleotides to retain proline/serine-rich IDRs in splicing

(A) Ratios of NNC, NNG, NNT, and NNA codons in relation to the exon lengths that were divided into fifty groups, as shown in Fig 8A. The ratios were normalized for each of the NNC, NNG, NNT, and NNA codons to calculate the density so that the sum of the ratios became 1.

(B) Violin plots showing fractions of disordered residues in all human internal exons estimated using the DisEMBL-hot loop (Linding *et al*., 2003). A similar graph in Fig 8C used the IUpred2 algorithm to predict disordered residues. All human internal exons were evenly divided into 50 groups, as shown in Fig 8A. The mean (green) with standard error (SE, semi-transparent shade) and the median (red) are indicated.

(C) Heatmap showing average ratios of P and S amino acids in all human internal exons, which are evenly divided as in Fig 8A and divided into ten categories according to their fractions of disordered residues at 0.1 interval.

(D) Heatmap showing the average ratios of C-nucleotides in the human internal exons, which are evenly divided as shown in Fig 8A, and divided into ten groups based on their fractions of disordered residues at 0.1 intervals, as shown in Fig EV5C.

(E) The average ratios of C-nucleotides in all the internal exons of ten species, which were evenly divided, as shown in Fig 8A.

(F) Heatmap showing the average ratios of C-nucleotides in the large internal exons in 10 species, which are divided as in Fig EV5D. The average ratio of C-nucleotides in each group was normalized to that of all internal exons in each species.

(G) A representative gene containing large exons enriched with C-nucleotides and coding for IDRs. IDRs (IUpred2 scores) of BRD4 in 8 vertebrates, those for *C. elegans* orthologs *Bet1*, and those for *Drosophila* orthologs *fs(l)h* are indicated by heatmap and are drawn to scale. Large exons (> 300 nt) are indicated by green boxes. *BRD4* exons 14 and 15 (red arrowheads) are S3-LCEs enriched with C-nucleotides (see Fig EV4A). Note that large exons in vertebrates tend to have high IUpred2 scores, whereas large exons in *C. elegans* and *Drosophila* do not.

(H) Clustering and heatmap showing the relationship between the codon ratios and exon lengths in all coding exons in 265 species. The relationship was analyzed as shown in Fig 8F, except that

all coding exons were used. Enlarged view with species names is shown in Appendix Fig S6.

(I) Pie chart showing the number of codons with amino acids in Clusters 1 to 4 in Fig EV5H. Sequence logos created by the codons in each cluster are also shown.