A novel protein domain in an ancestral splicing factor drove the evolution of neural microexons

Torres-Mendez, Antonio; Bonnal, Sophie; Marquez, Yamile; Roth, Jonathan; Iglesias, Marta; Permanyer, Jon; Almundi, Isabel; O'Hanlon, Dave; Guitart, Tanit; Soller, Matthias; Gingras, Anne-Claude; Gebauer, Fátima; Rentzsch, Fabian; Blencowe, Benjamin J.; Valcárcel, Juan; Irimia, Manuel

DOI:
10.1038/s41559-019-0813-6

A novel protein domain drove the emergence of neural microexons in bilaterian animals

Antonio Torres-Méndez [1,2,†], Sophie Bonnal [1,2,†], Yamile Marquez [1,2], Jonathan Roth [3,4], Marta Iglesias [5], Jon Permanyer [1,2], Isabel Almudí [6], Dave O'Hanlon [3], Tanit Guitart [1,2], Matthias Soller [7], Anne-Claude Gingras [4,8], Fátima Gebauer [1,2], Fabian Rentzsch [5], Benjamin J. Blencowe [3], Juan Valcárcel [1,2,9] and Manuel Irimia [1,2,9,*]

[1] Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08003, Spain

[2] Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain

[3] Donnelly Centre, University of Toronto, Toronto, ON M5S 3E1, Canada

[4] Department of Molecular Genetics, University of Toronto, Toronto, Ontario, M5S 3E1, Canada

[5] Sars Centre for Marine Molecular Biology, University of Bergen, Thormøhlensgt. 55, 5006 Bergen, Norway

[6] Centro Andaluz de Biología del Desarrollo (CABD), CSIC-Universidad Pablo de Olavide-Junta de Andalucía, Seville, Spain.

[7] School of Biosciences, College of Life and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK.

[8] Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto M5G 1X5, Canada.

[9] ICREA, Pg. Lluís Companys 23, Barcelona, Spain

[†] Co-first authors

[*] Correspondence to: MI, mirimia@gmail.com

Short title: The evolutionary origin of neural microexons

Keywords: complexity, genome evolution, transcriptomics, nervous system

Abstract

The mechanisms by which entire programs of gene regulation emerged during evolution are poorly understood. Neuronal microexons represent the most conserved class of alternative splicing in vertebrates and are critical for proper brain development and function. Here, we discover neural microexon programs in non-vertebrate species and trace their origin to bilaterian ancestors through the emergence of a previously uncharacterized ‘enhancer of microexons' (eMIC) protein domain. The eMIC domain originated as an alternative, neural-enriched splice isoform of the pan-eukaryotic *Srrm2*/*SRm300* splicing factor gene, and subsequently became fixed in the vertebrate and neuronal-specific splicing regulator *Srrm4/nSR100* and its paralog *Srrm3*. Remarkably, the eMIC domain is necessary and sufficient for microexon splicing, and functions by interacting with the earliest components required for exon recognition. The emergence of a novel domain with restricted expression in the nervous system thus resulted in the evolution of splicing programs that qualitatively expanded the neuronal molecular complexity in bilaterians.

Main text

Neural microexons are 3-27 nucleotide-long cassette exons that are activated during neuronal differentiation and generate protein isoforms with one to several additional amino acids [1,2]. These amino acids are often located on protein surfaces and can impact protein-protein interactions. Given their very short length, which is expected to preclude standard spliceosomal interactions needed for exon definition, the presence of extensive microexon programs in vertebrates was surprising and mechanistically challenging. For most mammalian neural microexons, a single factor, the Ser/Arg repetitive matrix protein 4 (SRRM4; also known as nSR100) is the major factor responsible for their inclusion [1]. Homozygous mutant mice deficient in *Srrm4* show numerous neurodevelopmental defects affecting both the central and peripheral nervous system and most die at birth [3], and a mutation in *Srrm4* has been linked to deafness in mice [4]. In contrast, heterozygous mutant mice expressing reduced levels of Srrm4 are viable, but show deficient neuronal excitability and synaptic transmission, and display several autistic-like neurodevelopmental and behavioral abnormalities [5]. Misregulation of many of the orthologous microexons has also been linked to autism spectrum disorder (ASD) in human subjects, correlating with reduced expression of *SRRM4* [1]. Consistent with their functional importance, among all classes of alternatively spliced exons, neural microexons are the most evolutionarily conserved and most likely to preserve coding frame [1,2]. However, their evolutionary origin is not known.

Results and discussion

Srrm4 was previously defined as a vertebrate-specific neural splicing regulator [6] and target microexons were subsequently identified in all major vertebrate groups [1,7]. To initially investigate the evolutionary origins of *Srrm4* and neural microexons, we analyzed tissue-specific RNA sequencing (RNA-seq) data from amphioxus (*Branchiostoma lanceolatum*). This non-vertebrate species shares a general chordate bodyplan with vertebrates, but lacks many of their key traits. Remarkably, this analysis revealed a microexon program comparable in size and neural specificity to that of zebrafish and mammals (Fig. 1a). Furthermore, RNA-Seq data from multiple tissues from centipede (*Strigamia maritima*) and fruitfly (*Drosophila melanogaster*) also revealed neural-specific microexons programs in both species (Fig. 1a). Zebrafish and mammalian neural microexons were largely shared, highlighting the unusually high conservation of this splicing program within vertebrates (Fig. 1b). Moreover, although most microexons are not conserved between vertebrates and non-vertebrates, several are shared

among phyla, spanning 600 million years (MY) of evolution (Supplementary Dataset 1). These include the autism-linked microexon in human *CPEB4* [8], which is conserved across vertebrates and in the amphioxus and centipede *Cpeb2/3/4* ortholog. These findings thus indicate that neural microexon programs predate the last common ancestor of bilaterian animals.

To assess whether the non-vertebrate microexon programs are regulated by a mechanism that is homologous to that of their vertebrate counterparts, we first sought evidence of Srrm4-dependent regulation by searching for UGC motifs, which are known to mediate direct Srrm4-dependent regulation of microexons in mammals [1,4,7,9]. In all cases, we found enrichment of UGC-containing motifs in the upstream intron, near the 3′ splice site (ss) (10-25 nt) (Fig. 1c and Supplementary Fig. 1a). Moreover, both vertebrate and non-vertebrate microexons were associated with strong 5' splice sites (ss), branch points and polypyrimidine tracts, but with weak 3′ss AG contexts (Fig. 1d and Supplementary Fig. 1b), although some of these signals were less pronounced for *Drosophila* and *Strigamia*, probably due to the differences in gene structure that favor intron definition mechanisms in these species. Given these sequence similarities between vertebrate and non-vertebrate species, we next asked whether the inclusion of *Drosophila* microexons could be promoted by heterologous expression of human SRRM4. Strikingly, overexpression of SRRM4 in SL2 *Drosophila* cells promoted inclusion of two-thirds (18/27) of probed neural microexons (Fig. 1e). Furthermore, among all alternative exons enhanced by SRRM4 overexpression in SL2 cells, microexons were strongly enriched (26/66 vs 31/611, $p=4.9x10^{-14}$, Fisher's exact test). Thus, expression of human SRRM4 is sufficient to activate microexon splicing in non-vertebrate cells, as it is in non-neural mammalian cells (Supplementary Fig. 1c, [1]).

These results prompted us to next search for factors with SRRM4-like activity in non-vertebrates. First, within vertebrates, based on phylogenetic and syntenic analyses, we identified *Srrm2*, *Srrm3* and *Srrm4* as paralogs resulting from the two rounds of whole genome duplication that occurred at the base of the vertebrate clade [10] (Fig. 2a and Supplementary Fig. 2a). The three paralogs present as a mix-and-match of domain architecture, with domains ranging from shared across all three paralogs to paralog-specific (Supplementary Fig. 2a). *Srrm2* (aka *SRm300* in humans and *Cwc21* in yeast) is a pan-eukaryotic, broadly expressed, splicing factor involved in both constitutive and alternative splicing [11-14]. It is composed of an N-terminal 'cwf21 domain' that interacts with the core U5 snRNP spliceosomal protein Prp8, and a C-terminal region with an Arg/Ser-repeat (RS) domain and other repetitive features

(Supplementary Fig. 2a and [14,15]). In contrast, *Srrm4* contains an RS domain and an annotated domain of unknown function: PFAM15230 (hereafter "enhancer of microexons" or eMIC domain; Supplementary Fig. 2a). *Srrm3*, which shares ~30% identity with *Srrm4*, contains both cwf21 and eMIC domains found in *Srrm2* and *Srrm4*, respectively (Supplementary Fig. 2a).

By contrast, in most non-vertebrate bilaterian animals, we found a single gene orthologous to the three vertebrate loci, which contains all *Srrm2/3/4* domains (Fig. 2b and Supplementary Fig. 2a). This gene shows greater sequence similarity to *Srrm2* than to other family members (Fig. 2a). However, in contrast to the three vertebrate paralogs, non-vertebrate *Srrm2/3/4* genes nearly always generate multiple isoforms derived from several alternative initiation, splicing and polyadenylation events (Fig. 2b-d). Two of these events were of particular interest given their tissue-specificity and level of evolutionary conservation: (i) an alternative promoter, which leads to the inclusion of a short motif conserved with vertebrate *Srrm4*, and that is neurally enriched both in centipede and amphioxus (Fig. 2c and Supplementary Fig. 2b); and (ii) an alternative last exon, which encodes the second half of the eMIC domain found in vertebrate Srrm3 and Srrm4, and that is primarily expressed in neural tissues in all three studied non-vertebrate species (Fig. 2d and Supplementary Fig. 2c,d). In contrast to the *Srrm3* and *Srrm4* vertebrate paralogs, no tissue-specificity was observed for total *Srrm2/3/4* transcripts at the level of steady-state mRNA expression in non-vertebrates (Fig. 2e). Altogether, these data suggest that neural regulation of microexons in non-vertebrates might be associated with alternative isoform usage of the single *Srrm2/3/4* splicing factor.

To test this hypothesis, we performed heterologous expression experiments in HEK293 cells (which lack *SRRM4* expression), using a measure of the inclusion levels of endogenous human neural microexons as readout of SRRM4-dependent activity (Fig. 3a). Expression of vertebrate (zebrafish) *Srrm2* and *Srrm4* served as negative and positive heterologous controls, respectively. Overexpression of four mRNA isoforms of the amphioxus *Srrm2/3/4* locus combining the two alternative promoters (with and without cwf21 domain) and the two last exons encoding the alternative C-termini (Supplementary Fig. 3a) revealed that the microexon regulatory activity depends exclusively on the presence of the full eMIC domain in the C-terminus of the protein, as indicated by the inclusion pattern of *MEF2C*, *CERS6*, *APBB1* and other analyzed microexons (Fig. 3a and Supplementary Fig. 3b). Transfection of CG7971-A and CG7971-C, the equivalent isoforms of the *Drosophila Srrm2/3/4* ortholog, confirmed that only the neural C-terminus isoform (i.e. containing the eMIC domain) promotes microexon

inclusion (Fig. 3a and Supplementary Fig. 3b). To exclude a potential contribution of the alternative N-termini of the *Srrm2/3/4* proteins to microexon regulation, we performed RNA-seq of stable HEK293 cell lines overexpressing the two different microexon-regulating amphioxus isoforms (with different N-termini). Both isoforms commonly regulated all detected SRRM4-dependent microexons in HEK293 cells (Fig. 3b). Furthermore, co-transfection of the amphioxus ortholog with minigenes containing vertebrate microexons and flanking intronic sequences with either wild type or mutated UGC *Srrm4*-binding sites revealed the same dependency on this motif to mediate microexon inclusion as previously described for the human protein (Fig. 3c and (4,7)). RNA-seq from stable HEK293 cell lines overexpressing *Srrm2/3/4* orthologs from zebrafish and fruitfly confirmed the ability of eMIC-containing proteins to enhance microexon inclusion genome-wide, irrespective of their N-termini (Fig. 3d and Supplementary Fig. 3c,d). Remarkably, this increased exon inclusion by eMIC-containing proteins affected up to 82% (138/169) of all neural microexon events with sufficient read coverage (Fig. 3d). Moreover, unlike other neural splicing factors, the splicing effect of eMIC-containing proteins was extremely specific for short exons (Fig. 3e and Supplementary Fig. 1c).

To understand the history of the eMIC domain and its connection with neural microexon programs, we next studied the phylogenetic distribution of this domain across metazoans. This analysis identified two independent loses since the bilaterian ancestors, in the nematode and platyhelminth clades (Supplementary Fig. 4a). Tissue- and cell-type specific RNA-seq data from *Caenorhabditis elegans* revealed only few neural microexons with no enrichment of the intronic UGC-motif, consistent with the secondary loss of the eMIC domain (Fig. 3f and Supplementary Fig. 4b,c). Moreover, the absence of the eMIC domain in all non-bilaterian groups, as well xenacoelomorphs, precisely located its origin in nephrozoan ancestors. Consistently, RNA-seq of the cnidarian *Nematostella vectensis* showed few microexons overall, and no enrichment in *Elav1*::mOrange positive cells (i.e. neuronal cells [16]) versus other cell populations from transgenic polyps (Fig. 3g and Supplementary Fig. 4d,e). As expected, transfection of the *Nematostella Srrm2*-like ortholog into HEK293 cells did not promote microexon inclusion (Fig. 3h third lane).

The results thus far suggest an evolutionary scenario in which the ancient *Srrm2*-like gene acquired the eMIC domain in bilaterian ancestors, conferring the ability to promote microexon inclusion. To test this hypothesis, we fused the region encoded by the last two exons of the

human *SRRM4* gene to the *Nematostella* Srrm2-like protein, and tested its regulatory activity in HEK293 cells (Supplementary Fig. 3e). Strikingly, this evolutionary chimera was able to promote microexon inclusion (Fig. 3h). Furthermore, fusing the C-terminal region of human SRRM4 to other unrelated SR proteins (i.e. SRRM1, SRSF1 and SRSF4) was sufficient to convert all of them into microexon regulators (Supplementary Fig. 3f,g). In fact, truncated versions of SRRM4 and SRRM3 consisting only of the two exons encoding the eMIC domain (referred to as 'C4' and 'C3' peptides, respectively) were sufficient to promote microexon inclusion genome-wide (Fig. 4a-c and Supplementary Fig. 5c-e), and a minimal construct of 78 aminoacids including the eMIC domain and 8 SR repeats (referred to as 'Min4') had limited but detectable microexon regulatory activity (Fig 4 and Supplementary Fig. 5b). In contrast, deleting the 39 most conserved amino acids of the eMIC domain in the full-length SRRM4 protein completely abrogated its function ('FL_M' in Fig. 4a-c and Supplementary Fig. 5c-e). Single point mutations in this region dramatically reduced the activity of the C-terminal peptides (Supplementary Fig. 5a) and combining only two point mutations resulted in the complete loss of function of the mutant peptides ('C3_M' and 'C4_M' in Fig. 4 and Supplementary Fig. 5).

To elucidate how the eMIC domain exerts its function at the biochemical level, we stably expressed different Flag epitope-tagged domains of SRRM4 in HEK293 cell lines and analyzed their interaction partners by affinity purification coupled to mass-spectrometry (AP-MS) (Fig. 4d and Supplementary Fig. 5h). The mutant lacking the eMIC domain (FL_M) interacts mainly with members of the exon junction complex (EJC) and associated factors, such as EIF4A3, PNN, ACIN1 or RNPS1. The latter has been recently identified as a major SRRM4 interactor in a genome-wide screen for factors regulating microexon splicing [9]. On the other hand, the C-terminal region interacts specifically with early spliceosomal factors, particularly proteins that function in the recruitment of U2 snRNP to the branch site region of pre-mRNA. Indeed, only SF1 and U2AF, which are among the earliest factors known to function in spliceosome assembly, were found to interact with all constructs containing an eMIC domain (Fig. 4d and Supplementary Fig. 5f,h), which we validated by co-immunoprecipitation and western-blot analysis (Fig. 4e and Supplementary Fig. 5g). These interactions thus suggest a role for the eMIC domain at the initial steps of spliceosome assembly. Consistent with this, *in vitro* spliceosome assembly assays using HeLa cell nuclear extracts and a pre-mRNA transcript containing the 6nt-long *apbb1* microexon revealed an ATP- and U2 snRNP-dependent complex (Fig. 4f and Supplementary Fig. 6a) that was enhanced in nuclear extracts obtained

from cells expressing either the full-length SRRM4 or C4 as well as by addition of purified recombinant C4 / C3 peptides in a branch point- and UGC-dependent manner (Fig. 4g and Supplementary Fig. 6b,c). The effect of the peptide was independent of the 3' splice site AG (Fig. 4g), and was also detected on a substrate lacking a functional 5' splice site (Supplementary Fig. 6c,d), suggesting that the eMIC domain promotes early steps of 3' splice site recognition. Consistent with the results in cell culture (Fig. 3 and 4), different wild-type versions of both human- and zebrafish-derived C3 and C4 peptides were active in promoting spliceosomal A complex formation *in vitro* (Supplementary Fig. 6e), but mutant peptides displayed reduced or no activity (Supplementary Fig. 6f). Mass-spectrometry analysis of complexes assembled *in vitro* on wild-type and UGC-mutated *apbb1* RNAs confirmed that eMIC-containing peptides enhanced formation of early stages of splicing complex formation by promoting the recruitment of SF1, U2AF, and U2 snRNPs (Fig. 4h, Supplementary Fig. 7, and Supplementary Dataset 2) [7,17]. The effect of C3 peptide requires SF1 (Supplementary Fig. 6h), indicating that the peptide does not overcome early BP recognition by this factor. Collectively, these experiments provide biochemical evidence that the eMIC domain is crucial for microexon inclusion by promoting the earliest stages of spliceosome assembly.

In summary, our comparative approach has uncovered a genomic novelty in the *Srrm2*-like locus of bilaterian ancestors by which an alternative splicing event in a core splicing factor generated a new protein domain whose associated biochemical function enabled the specific recognition and inclusion of microexons. Therefore, unlike most reported cases (e.g. [18-20]), this neofunctionalization occurred not via gene duplication but through novel domain emergence and alternative splicing of an ancestral locus. Intriguingly, we have not found sequence similarity between the eMIC domain and any other characterized protein domains in any species, suggesting *de novo* evolution as a specialized type of RS domain. Besides, secondary structure predictions of this domain indicate an unstructured conformation. Describing its precise activity in contacting early spliceosomal components thus remains an outstanding question.

The 600 MY-old biochemical function provided by the eMIC domain qualitatively expanded the regulatory landscapes of animal proteomes by enabling the evolution of transcriptomic elements – microexons –, most of which could have not been previously recognized by the cellular machinery. The restriction of the eMIC-encoding isoforms to the nervous system facilitated the expansion and specialization of neuronal proteomes by innovations in a set of

genes involved in diverse cellular functions, including vesicle transport and release, cytoskeleton organization, and chromatin regulation, among others [1,5,21,22]. During vertebrate evolution, duplication and subfunctionalization of a single ancestral *Srrm2/3/4* gene gave rise to three splicing factors, *SRRM2*, *SRRM3* and *SRRM4*. The latter two became specifically expressed in the nervous system and evolved to provide critical roles in nervous system development and function, a feature that is underscored by *SRRM4* disruption causing nervous system developmental defects and neurological disorders.

Materials and Methods

RNA-seq datasets

All RNA-Seq samples used in the current study are listed in Supp. Dataset 3. For human and mouse, we used publicly available data provided in VastDB (http://vastdb.crg.eu). For zebrafish (*Danio rerio*), amphioxus (*Branchiostoma lanceolatum*), worm (*Caenorhabditis elegans*) and fruifly (*Drosophila melanogaster*), we collected publicly available RNA-seq data from the NCBI Short Read Archive (SRA). For the centipede *Strigamia maritima*, we performed tissue dissections in adult specimens for nerve cords, heads, muscles, Malpighian tubes, male gonads and salivary glands. RNA was extracted using Trizol (Ambion, following manufacturer's instructions). Ribosomal RNA depleted, strand-specific libraries were built, and sequenced using Illumina HiSeq2500 at the Genomics Unit at CRG. An average of 140 million 75-nt paired-end reads were generated per sample. Raw data was submitted to SRA (accession number SRP149913).

Computational analysis of tissue-specific microexons

vast-tools [1,23] was used for identification of microexons and quantification of their inclusion levels (using the metric Percent Spliced In [PSI]) from RNA-seq data from different cell and tissue types in each of the seven species analyzed (Supplementary Dataset 3). In brief, *vast-tools* uses different modules to identify and quantify all major types of alternative splicing, including simple and complex cassette exon events, retained introns, alternative acceptor and donor site choices and microexons [1,23]. It has been used to analyze alternative splicing in multiple species, providing high validation rates by RT-PCR [1,3,5,7,23-29], especially for microexons [1]. In particular, the microexon module compiles all annotated and *de novo* identified 3-15 nucleotide microexons and profiles their inclusion levels using exon-microexon-exon junction read counts as support for inclusion, instead of the standard exon-

exon junctions (details in [1]). Associated VASTDB files to run *vast-tools* are available to download for each species (https://github.com/vastgroup/vast-tools): human (species key "Hsa", Hg19 or Hg38, http://vastdb.crg.eu/libs/vastdb.hsa.16.02.18.tar.gz), mouse ("Mmu", mm9 or mm10, http://vastdb.crg.eu/libs/vastdb.mmu.16.02.18.tar.gz), zebrafish ("Dre", danRer10, http://vastdb.crg.eu/libs/vastdb.dre.01.12.18.tar.gz), amphioxus ("Bla", Bl71nemr, http://vastdb.crg.eu/libs/vastdb.bla.01.12.18.tar.gz), centipede ("Sma", Smar1, http://vastdb.crg.eu/libs/vastdb.sma.01.12.18.tar.gz), fruitfly ("Dme", BDGP6, http://vastdb.crg.eu/libs/vastdb.dme.01.12.18.tar.gz) and nematode ("Cel", WBcel235, http://vastdb.crg.eu/libs/vastdb.cel.01.12.18.tar.gz). For simplicity, only MIC_S (simple events) from the microexon pipeline were used for all species but Hsa and Mmu.

Following previous studies [23], we defined as alternatively spliced (Fig. 1b) those microexons with $10 \leq PSI \leq 90$ in at least 10% of the samples with sufficient read coverage (*vast-tools* score LOW or higher; for details, see https://github.com/vastgroup/vast-tools/blob/master/README.md) or a range of PSIs ≥ 25 across the same samples. To identify microexons that were specifically enriched (or depleted) in neural tissues, we required a $\Delta PSI \geq 15$ (or $\Delta PSI \leq -15$) between the neural and every other tissue with sufficient read coverage. Microexons that presented equivalent enriched inclusion patterns for any other tissue were grouped together according to the tissue type (Fig. 1a): muscle (skeletal and/or heart), visceral (liver, kidney, lung, digestive tract, Malpighian tube or salivary gland), skin (epidermis or skin), gonads (testis, ovary or female/male gonads) or other (early development or immune/blood tissues). For all analyses, we required that the microexon had sufficient read coverage in at least four different tissues (from a total of 12 tissues in vertebrates and 6 to 8 in non-vertebrates; Supplementary Dataset 3).

Microexon orthology analysis

To identify orthologous microexons among the analyzed species, we followed a similar methodology as previously described [29], with some modifications in the gene clustering step. In brief, we generated clusters of homologous genes for human, mouse, zebrafish, amphioxus, centipede and fruitfly taking into account the different number of whole genome duplication events experienced by each group (two for human and mouse, and three for zebrafish). For this purpose, we used orthology information from OMA [30], Multiparanoid [31] and pairwise BlastP between all six species, employing the longest protein-coding isoform for each gene. Then, for

each gene from species A versus each gene of species B, a combined orthology score was derived for the gene pair based on the information provided by the different sources:

(i) OMA: +1 if they belonged to the same OMA cluster.

(ii) Multiparanoid: +1 if they belonged to the same Multiparanoid cluster, and both genes had a Multiparanoid score ≥0.5.

(iii) BlastP: +1 if $hit_{spA=>spB} \leq R_{spB}/min(R_{spA},R_{spB})$ & $hit_{SpB=>SpA} \leq R_{spA}/min(R_{spA},R_{spB})$, or +0.5 if $hit_{spA=>spB} \leq R_{spB}/min(R_{spA},R_{spB})$ & $hit_{SpB=>SpA} = 1+(R_{spA}/min(R_{spA},R_{spB}))$, where $hit_{spA=>spB}$ is the rank of the gene pair in the BlastP output of species A against species B, $hit_{spB=>spA}$ the rank in the reciprocal blast (i.e. species A against species B), R_{spA} and R_{spB} the levels of ancestral ploidy in species A and B (1x in non-vertebrates, 4x in mammals and 8x in zebrafish), and $min(R_{spA},R_{spB})$ the lowest of these two values.

Therefore, each pair of genes from any two species may have a maximum orthology score of 3. Next, gene orthology clusters were built in an iterative manner with subsequently less stringent cut-offs ('ConfLevel' in Supplementary Dataset 4), and aiming at minimizing the number of false positive orthology assignments.

(L1) Only pairs of genes with an orthology score of 3 were considered orthologs and assigned to the same cluster. When a pair of orthologous genes belonged to two different clusters, the two clusters were merged. However, to avoid artifactual merging, the two clusters were merged only if the total ploidy complement of the merged cluster was closer to the idealized situation (4-4-8-1-1-1 for human, mouse, zebrafish, amphioxus, fruitfly and centipede, respectively) than each of the clusters individually. Note that this approach is conservative with regards to false positive orthology calls. At this stage, we also grouped clusters based on previously reported ohnology relationships ([32], using their intermediate confidence level).

(L2) Next iterations are aimed at filling the potential gaps in the idealized ploidy complement. For each clusters generated in step (1), we searched for potential orthologs only in the species without an assigned ortholog or with a number of members lower than its ancestral ploidy (in the case of vertebrates). To assign novel orthologs to a gene cluster, we required an orthology score against any of the members of the clusters ≥2. Two different clusters could be merged if the missing ortholog already belonged to another cluster, with the limitations described above.

(L3) A similar process as in (2) to fill potential gaps was performed again, but requiring an orthology score ≥1.5 and not allowing cluster merges.

Finally, to generate clusters based on less supported orthology relationships, genes that were not assigned to any cluster were reevaluated using iterations (L1) and (L3), but using a minimal orthology score ≥2.5 (L4) and 1.5 (L5), respectively. Clusters of gene orthologs are provided as Supplementary Dataset 4.

Then, for each gene cluster in which at least one human gene hosts a neural-enriched microexon, we aligned all pairs of proteins from each species using MAFFT [33]. Intron positions and their corresponding phases for each species' protein were integrated into the resulting pairwise protein alignment to introduce the exon-intron structure information [34]. A microexon in a given species A was considered to be conserved in species B, if: (i) their upstream and downstream constitutive exons had at least 20% of protein similarity with the corresponding exons in species B, and (ii) their corresponding flanking upstream and downstream intron positions showed the same phase and an offset of no more than 5 residues in the protein alignment. Moreover for human and mouse microexons, additional homology information was obtained using liftOver (http://genome.ucsc.edu/cgi-bin/hgLiftOver) and curated manually. Finally, homologous microexon clusters were built using a guilt-by-association approach with all the homologous microexons identified between each pair of species (i.e. if the microexon was considered conserved between species A and B and A and C, a cluster for species A, B and C was produced). Clusters for human neural microexons that were found conserved in at least another studied species are provided in Supplementary Dataset 1.

Analysis of sequence features associated with neural microexons

For each species, we compared different sequence features of neural-enriched microexons against a set of 1,000 random longer (>27 nts) non-neural alternative exons, as control. We calculated the strength of the donor and acceptor splice sites (5′ss and 3′ss/AG, respectively) according to maximum entropy score models [35]. For branch point and polypyrimidine tract score calculation, we employed the best predicted branch point for each exon [36]. To test for differences in the median of the scores for each feature between the two groups (neural microexons vs. longer non-neural alternative exons), we used Mann-Whitney U tests. To generate the RNA maps in Fig. 1c, Extended Data Fig. 1a and 4c, we used rna_maps function from Matt [37], using sliding windows of 17 nts. Searches were restricted to the microexon, the first and last 150 nts of the upstream and downstream intron (or total intron length if shorter than 150 nts) and 13 nts into the upstream and downstream exons. Below the main RNA map,

we also displayed the percentage of the data that is used at every nucleotide position, to graphically depict the number of exons being analysed for each region.

Stable heterologous expression of human SRRM4 in *Drosophila* SL2 cells

Myc-tagged human *SRRM4* was cloned into MT-STABLE1 vector [38] (kindly provided by Dr. James D. Sutherland) in the position of the puromycin resistance gene by NheI/XhoI restriction enzyme digestion. MT-STABLE1-myc-SRRM4 and empty MT-STABLE1 (used as a negative control) were co-transfected into *D. melanogaster* SL2 cells, together with the resistance plasmid Ac5-puro in a 9:1 ratio using Effectene transfection reagent (Qiagen), following manufacturer's instructions. Cells were maintained under selection for 48h after transfection by adding 5ug/ml puromycin, reaching a stable population of >90% GFP-positive cells. Myc-SRRM4 expression was induced with 500μM copper sulphate for 4 days. Total RNA was extracted from SRRM4-expressing and control cells using RNeasy® Plus Mini kit (Qiagen), following manufacturer's instructions, and subjected to polyA-selected RNA-sequencing at the CRG Genomics Unit. An average of 66 million 125-nt paired reads were generated for each sample (Supplementary Dataset 3). The impact of *SRRM4* expression on exon inclusion levels was calculated using *vast-tools* as the ΔPSI between the test and control samples. Only exons and microexons with coverage scores of LOW or higher in both samples were represented in Fig. 1e.

Search for *Srrm4* orthologs and generation of multi-species alignment

Srrm4 orthologs for all organisms included in Fig. 2 and Extended Data Fig. 2a-c were identified by tblastn searches against NCBI and EnsemblMetazoa databases (Supplementary Dataset 5). We follow the conventional gene nomenclature from the community studying whole genome duplications in vertebrates, i.e. naming the non-vertebrate orthologous genes to the three vertebrate ohnologs (*Srrm2*, *Srrm3* and *Srrm4*) as *Srrm2/3/4*. To generate the protein alignment to build the phylogenetic tree in Fig. 2a, *Srrm2/3/4* orthologs from representative species of major metazoan groups were selected, and *Srrm1* orthologs from the same species were used as outgroups. Protein sequence alignment was performed using MUSCLE [39]. The phylogenetic tree was built using Mr. Bayes [40], with VT matrix with +G improvement as evolutionary model as determined by ProtTest [41], and visualized with FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree). Branch support values (posterior probabilities) are calculated for half million iterations in Mr. Bayes. For ancestral gene structure reconstruction, exon-intron structure of *Srrm2/3/4* orthologs was assessed using GeneWise [42], and the resulting

protein sequences aligned taking into consideration intron position and phases [34]. For alignment of the alternative N- and C-terminal regions in bilaterians, we searched for the presence of the conserved motifs in the genomic regions of *Srrm4* orthologs using tblastn, since they are often not present in the annotation. Protein sequence alignments of these protein regions (Fig. 2c,d) were done using MUSCLE and visualized using MSAViewer [43]. Information-content sequence logos were generated using Skylign [44].

Quantification of expression and alternative isoform usage of *Srrm2/3/4* orthologs

To estimate the percentage of transcripts from the *Srrm2/3/4* locus that use the alternative first and last exons in non-vertebrate bilaterians, we mapped RNA-seq data from different tissues to isoform-specific exon-exon junctions. For quantification of amphioxus alternative C-terminus usage, we used the PSIs as quantified by *vast-tools*, since this is a simple exon skipping event in this species that can be readily measured with this pipeline. For quantification of the alternative promoter in amphioxus and *Strigamia*, and the alternative last exon of *Drosophila*, the number of reads mapping to all the potential donor or acceptor sites from the reference upstream or downstream exon, respectively, were obtained from the eej2 output of *vast-tools*, which reports the read counts for all possible exon-exon junction combinations [1]. In the case of the alternative C-terminal isoform of *Strigamia*, we manually mapped reads to all the possible exon-exon junctions that could take place within this terminal region. In all cases, we displayed the percentage of reads mapping the exon-exon junction supporting the neural isoform (blue lines in Fig. 2c,d) over the total number of reads across all competing junctions. For gene expression analysis of *Srrm2/3/4* homologs, gene counts reported by *vast-tools* were used as input for edgeR [45], which provided the total counts per million (CPM) as a measure of gene expression per species. To normalize the expression across the different orthologs, CPM values were converted to Z-scores within each gene and species.

Generation of constructs used for HEK293 heterologous expression experiments

Full-length open reading frames for each tested isoform were amplified from cDNA from various sources, 6xHistidine tagged at the N-terminus, cloned into pcDNA3.1(-), and sequence-verified for transient overexpression in HEK293 cells. Protein chimeras fusing the C-terminal region of human SRRM4 to several proteins (*N. vectensis* Srrm2, zebrafish Srrm1 and Srrm2, and human SRSF1 and SRSF4) were performed using standard restriction/ligation methods (all primers provided in Supplementary Dataset 6). Point mutants of C4 and C3 peptides and the full-length SRRM4 lacking the eMIC domain were generated by site-directed

mutagenesis (inverse PCR of plasmids encoding the wild-type sequences) using oligos as recommended by NEBaseChanger (http://nebasechanger.neb.com).

Cloning of splicing reporters

Minigene encompassing Exon – *Intron* – Microexon – *Intron* – Exon - First 25 nucleotides of the downstream intron were constructed for *asap1* and *apbb1* zebrafish microexons. The corresponding genomic regions were cloned using KpnI and NotI restriction sites in between PT1 and PT2 sequences, for detection by RT-PCR after transfection [46], and under a CMV promoter. To facilitate cloning, introns were shortened, resulting in constructs with the following structure (italics indicate intronic sequences):

- *asap 1* : 87 nts – *250 nts XhoI 250 nts* – 9 nts – *163 nts HindIII 66 nts* – 101 nts – *25 nts* (the endogenous upstream and downstream introns are 2032 and 1578 nts-long, respectively).
- *apbb1* : 128 nts – *279 nts* – 6 nts – *250 nts PstI 230 nts* – 115 nts – *25 nts* (the endogenous upstream and downstream introns are 279 and 1784 nts-long, respectively).

The splicing profile of the minigenes was assessed using PT1 and a reverse primer in the last exon for the *asap1* minigene, and with zebrafish specific primers in the first and last exons for the *appb1* minigene.

Transient heterologous expression of *Srrm4* orthologs

HEK293 cell transfections were performed using Lipofectamine 2000 (Invitrogen, following manufacturer's instructions) in 6-well plates with 1 ug of expression plasmid. For minigene analysis, 10 ng of plasmid bearing the minigene were co-transfected with the expression construct. Cell pellets were collected 24 h after transfection. RNA was extracted using RNeasy Plus Mini kit (Qiagen) and cDNA generated using SuperScript III (Thermo Fisher Scientific, following manufacturer's instructions). RT-PCRs of microexon events were performed using primers annealing to the flanking upstream and downstream constitutive exons (all primer sequences are provided in Supplementary Dataset 6) or using PT1/PT2 to detect the pattern of alternative splicing of the minigene.

Generation of HEK293 Flp-In T-REx Stable Cell Lines used for RNA-seq

Flp-In™ (Thermo Fisher Scientific) expression plasmids: pcDNA5 and pcDNA5 containing GFP, 3xFlag-tagged human SRRM4, FL_M and C4; zebrafish Srrm2, Srrm3 and Srrm4;

Drosophila CG7971-C or amphioxus Srrm2/3/4 isoforms containing the eMIC domain, were generated using standard restriction/ligation methods (all primers provided in Supplementary Dataset 6). For each construct, tetracycline inducible Flp-In T-Rex 293 stable cell lines were generated by transfecting 400 ng of the pcDNA5 plasmid with 3.6 μg of pOG44 Flp-recombinase expression plasmid (Thermo Fisher Scientific, V600520) using lipofectamine 2000 (Invitrogen). Cells containing stably integrated constructs were selected in 100 μg/mL Hygromycin B until visible colonies were formed, and individual colonies were expanded and stored. For each protein, two clones from independent colonies were screened for transgene expression by qPCR before and 24 h after induction with 1 μg/mL doxycycline. For each protein, we selected the clone with the closest transcript levels to the other lines for RNA-seq (RNA-seq-based quantifications are provided in Extended Data Fig. 3d). RNA coming from a well of a 6-well plate for every line was extracted after 24 h induction with doxycycline, using RNeasy Plus Mini kit (Qiagen). RNA quality was checked using Bioanalyzer (Agilent) and strand-specific Illumina libraries were prepared and sequenced at the CRG Genomics Unit. An average of 60 million 125-nt paired-end reads were generated for each sample (Supplementary Dataset 3). Raw RNA-seq data was processed with *vast-tools* and the impact on the endogenous human microexons was compared between the control and the different lines using the module *compare*.

Western blot analysis of FLAG-tagged SRRM4 constructs

Total cellular proteins were extracted from HEK293 Flp-In cells growing in a 6-well plate after 24h induction with 1 μg/mL doxycycline, using 2x Laemmli buffer (125mM Tris-HCl pH 6.8, 20% glycerol, 4% SDS). Cell lysates were subjected to sonication (5 x 30 sec pulses with 30 sec between them). BCA assay (Thermo Fisher Scientific) was used for determining protein concentration. Western blots were performed using 10% SDS-polyacrylamide gels and the nitrocellulose membranes were probed with antibodies against FLAG-tag anti-FLAG M2 (F1804, Sigma) and α-tubulin (3873S, Sigma) at dilution 1:1000 in TBST x% milk. Immunolabelling was detected by enhanced chemiluminiscence (Luminata Forte Western HRP substrate WBLUC0100, Merck) and visualized with a digital luminescent image analyzer (Amersham Imager 600).

Analysis of target exon length distribution for tissue-specific splicing factors

Published RNA-seq datasets for *Nova1/2*, *Rbfox1/2/3*, *Srrm4* and *Esrp1/2* knockout or knockdown experiments (Supplementary Dataset 3) were analysed using *vast-tools* to calculate exon inclusion changes dependent on the depletion of these splicing factors. The top 100 exons regulated by each protein family (i.e. having the highest absolute ΔPSI compared to the control condition), were used for exon size distribution calculation as depicted in Fig. 3e.

Cell dissociation and RNA isolation of sorted populations from *N. vectensis*

Neuron-specific transgenic *NvElav1::mOrange* polyps of *N. vectensis* [16] were spawned and reared as previously described [47]. Several hundreds of mOrange positive primary polyps (12-14 days old) were dissociated at 37 ºC for ~30 min in calcium- and magnesium-free *Nematostella* medium (CMF/NM) containing EDTA and 0.25% of trypsin. Single cell suspensions were then stained at room temperature with Hoechst 33342 and 7-aminoactinomycin D (7-AAD) to exclude debris and non-viable cells by Fluorescent-Activated Cell Sorting (FACS). Doublet/multiplet exclusion was performed by plotting Hoechst-H vs Hoechst-A. Approximately 650,000 viable mOrange positive or negative cells were sorted into CMF/NM containing 0.5% of BSA (pH=7.8) using a BD FACSAria that was maintained at 4 ºC throughout the procedure. Cells were collected by centrifugation (800 x g) for 10 min at 4 ºC and resuspended into TRIzol® LS reagent (3:1 reagent to sample ratio) in four independent sorts. Total RNA from the positive and negative populations was isolated using Direct-zol™ RNA Microprep columns (Zymo Research), following manufacturer's instructions, and RNA quality was assessed by Bioanalyzer (Agilent). Two replicates of the *elav*+ and *elav*- samples with RNA Integrity Number ≥ 9.0, each consisting of pooled RNA extractions from two different sorting experiments, were sequenced in a Illumina HiSeq2500 at the CRG Genomics Unit, generating an average of 65 million 126-nt paired-end strand-specific reads per sample (Supplementary Dataset 3). Raw fastq data was processed using *vast-tools*, and the number of differentially regulated microexons between the *elav*+ and *elav*- fractions obtained employing *vast-tools compare* with default parameters (ΔPSI ≥ 15 and a pairwise differences between all replicates of ΔPSI ≥ 5). VASTDB libraries for *N. vectensis* can be downloaded in the *vast-tools* web site ("Nve", GCA_000209225, http://vastdb.crg.eu/libs/vastdb.nve.10.05.18.tar.gz).

Generation of HEK293 Flp-In T-REx Stable Cell Lines used for mass-spectrometry

Flp-In™ (Thermo Fisher Scientific) compatible expression plasmids (5' pcDNA5-BirA*-FLAG, modified from Thermo Fisher Scientific V6010-20) containing full length and mutant human *SRRM4* and *SRRM3* cDNA constructs were generated using Gateway cloning (Thermo

Fisher Scientific). For each bait protein, tetracycline inducible Flp-In T-Rex 293 stable cell lines were generated by transfecting 200 ng of the pcDNA5-BirA*-FLAG bait construct with 2 μg of pOG44 Flp-recombinase expression plasmid (Thermo Fisher Scientific, V600520) using jetPRIME transfection reagent (Polyplus, CA89129-924). Cells containing stably integrated constructs were selected in 200 μg/mL Hygromycin B until visible colonies were formed and passaged to a 150 mm plate. Stable HEK293 Flp-In T-REx cell lines expressing BirA*-FLAG, BirA*-FLAG-NLS, or FLAG without a bait cDNA were generated in parallel to be used as negative controls. For FLAG affinity purification experiments, all HEK293 Flp-In stable cell lines, including controls, were grown to ~70% confluency and induced for 24 h with 1 μg/mL tetracycline to express the BirA*-FLAG tagged bait protein. After induction, each 150 mm plate was harvested by scraping in 1 mL ice-cold PBS, pelleted, and then frozen at -80°C until purification.

Sample Preparation for FLAG AP-MS

FLAG Affinity Purification Coupled with Mass Spectrometry (AP-MS) experiments were performed in two biological replicates, essentially as described in [48], with slight modifications. In brief, cell pellets for each bait protein were lysed in ice cold TAP lysis buffer containing 50 mM HEPES-NaOH (pH 8.0), 100 mM NaCl, 2 mM EDTA, and 10% glycerol with freshly added 0.1% NP-40, 1 mM DTT, 1 mM PMSF, and 1:500 Protease inhibitor cocktail (Sigma-Aldrich, P8340) at a 1:4 pellet weight to volume ratio (for 0.1 g add 0.4 mL of lysis buffer). Resuspended pellets were placed on dry ice for 5 minutes and thawed in a 37 °C water bath until only a small ice pellet remained. Samples were quickly moved back on to ice and sonicated with three 10 sec bursts with 2 sec rest at an amplitude of 35%. To solubilize chromatin and reduce the detection of interactions mediated by RNA or DNA, 250U of Benzonase Nuclease (Sigma-Aldrich, E8263) was added to each tube and incubated 30 minutes with rotation at 4°C. Lysates were cleared by centrifugation at 20,000 x g for 20 minutes at 4°C and the lysates transferred to tubes containing 25 μL of 50% magnetic anti-FLAG M2 slurry beads (Sigma-Aldrich, M8823) pre-washed in lysis buffer. FLAG immunoprecipitation was allowed to proceed for 3 h at 4°C with rotation. After incubation, beads were pelleted by centrifugation (1,000 rpm for 1 min) and magnetized to aspirate the unbound lysate. The beads were then demagnetized and washed with 1 mL of lysis buffer and the total volume (with beads) transferred to a new tube. Beads were washed once more with 1 mL of lysis buffer followed by one wash with 50 mM ammonium bicarbonate (ABC) at pH 8. All wash steps

were performed on ice using cold lysis buffer and ABC. After the final wash, any residual ABC was aspirated from the beads, and 1 μg of trypsin (Sigma-Aldrich, T6567) in 10 μL of ABC was added to each tube. The samples were incubated at 37°C overnight with rotation. The following morning, beads were magnetized, and the supernatant transferred to a new tube. Another 250 ng of trypsin was added to each supernatant in 5 μL of ABC (total volume of 15 μL) and further digested with rotation for another 4 h at 37°C. Samples were acidified with formic acid to a final concentration of 2.5% and dried in a centrifugal evaporator.

AP-MS Data Acquisition

Data acquisition from AP-MS was done following [49]. Digested peptides were dissolved in 5% formic acid in a volume in which 6 μL contained purified material from one-quarter of a 150 mm plate. For each sample, 5 μL of each sample was directly loaded at 800 nL/min onto a 15 cm 100 μm ID emitter tip packed in-house with 3.5 μm Reprosil C18 (Dr. Maische). The peptides were eluted from the column at 400 nL/min over a 90 min gradient generated by a 425 NanoLC (Eksigent, Redwood, CA) and analyzed on a TripleTOF^{TM} 6600 instrument (AB SCIEX, Concord, Ontario, Canada). The gradient started at 2% acetonitrile with 0.1% formic acid and increased to 35% acetonitrile over 90 min followed by 15 min at 80% acetonitrile, and then 15 min at 2% acetonitrile for a total of 120 min. To minimize carryover between each sample, the analytical column was flushed for 1 h at 1500 nl/min with an alternating sawtooth gradient from 35% acetonitrile to 80% acetonitrile, holding each gradient concentration for 5 min. Analytical column and instrument performance were verified after each sample by analyzing 30 fmol bovine serum albumin (BSA) tryptic peptide digest with 60 fmol α-casein tryptic digest with a short 30 min gradient. MS mass calibration was performed on BSA reference ions between each sample. Acquisition was in Data Dependent mode and consisted of one 250 millisec MS1 TOF survey scan from 400-1250 Da followed by twenty 100 millisec MS2 candidate ion scans from 100–2000 Da in high sensitivity mode. Only ions with a charge of 2+ to 4+ that exceeded a threshold of 200 counts per second were selected for fragmentation, and former precursors were excluded for 10 seconds after 1 occurrence.

AP-MS Data Analysis

MS data were stored, searched and analyzed using the ProHits laboratory information management system (LIMS) platform [50]. The WIFF data files were converted to MGF format using WIFF2MGF and subsequently converted to an mzML format using ProteoWizard

(3.0.4468) and the AB SCIEX MS Data Converter (V1.3 beta). The mzML files were searched using Mascot (v2.3.02) and Comet (2014.02 rev.2). The results from each search engine were jointly analyzed through the Trans-Proteomic Pipeline (TPP) [51] via the iProphet pipeline [52]. The spectra were searched against a total of 72,482 proteins consisting of the NCBI human RefSeq database (v57, Aug 9th, 2016, forward and reverse sequences) supplemented with "common contaminants" from the Max Planck Institute (http://141.61.102.106:8080/share.cgi?ssid=0f2gfuB) and the Global Proteome Machine (http://www.thegpm.org/crap/index.html), as well as sequences from common fusion proteins and epitope tags. The database parameters were set to search for tryptic cleavages, allowing up to 2 missed cleavage sites per peptide, MS1 mass tolerance of 40 ppm with charges of 2+ to 4+ and an MS2 mass tolerance of ± 0.15 amu. Asparagine/glutamine deamidation and methionine oxidation were selected as variable modifications. A minimum iProphet probability of 0.95 was required for protein identification with a minimum number of 2 unique peptides required for protein interaction scoring (Supplementary Dataset 2). Significance Analysis of INTeractome (SAINTexpress version 3.6.1 [53]) was used as a statistical tool to calculate the probability value of each potential interaction from background contaminants. Briefly, our experimental design included specific negative controls (BirA*-FLAG, BirA*-FLAG-NLS, and FLAG), each run in several biological replicates (12 total). Each biological replicate of a bait was analyzed independently against these controls before averaging of the score values and assessment of the Bayesian False Discovery Rates (FDR; [53]). High-confidence interactions are those with FDR ≤ 1%.

Co-immunoprecipitation experiments

HA-tagged constructs were transiently transfected into HEK293 Flp-In cells grown in 6 well plates using Lipofectamine 2000 (Invitrogen, following manufacturer's instructions). After 24 hours, cells were treated with 0.2 µg/mL (C4 and Min4) or doxycycline 10µg/mL (C4_M) and 24 h later cells were harvested in cold phosphate buffered saline (PBS) and cells lysed in 250 µL cold 1xTAPS buffer (50 mM Hepes pH 8.0, 100 mM NaCl, 10% glycerol, 2mM EDTA, 0.1% NP-40, 1mM DTT and protease inhibitors). Lysates were subject to sonication (7x 1 sec pulses with 1 sec in between at 30% power). For nuclease digestion, 250U Benzonase was added and lysates were incubated at 4°C for 60 minutes with rotation. Lysates were cleared in a microcentrifuge by spinning at 15,000 x g for 10 minutes at 4°C. Clarified lysates were pre-cleared with 10 uL Dynabeads protein G (Thermo Fisher Scientific) for 30 minutes at 4°C. Anti-HA (for SF1 and U2AF2) or anti-Flag (to test for U2AF1 interaction)

immunoprecipitation were performed using magnetic Dynabeads protein G (Thermo Fisher Scientific) complexed with Rat anti-HA antibody (Roche) or mouse anti-Flag M2 antibody (Sigma). Antibody was incubated with lysates for 1 hour at 4°C followed by incubation with washed Dynabeads protein G for 1 hour at 4°C with rotation. Following the incubation step, the complexes were washed 3 times with cold 1x TAPS buffer. Elution was performed by boiling in 1x Laemmli buffer at 95°C for 5 min.

In vitro spliceosome assembly assays

Cy5-CTP/Cy5-UTP-labelled RNAs were *in vitro* transcribed from PCR products using T7 Megascript kit (Ambion). Sequences of *apbb1* and *asap1* RNAs used are provided in Supplementary Dataset 6. IgM bears the 3' 70 nucleotides and exon 2 containing an exonic enhancer. Spliceosome assembly assays were performed as described previously [54] with 15 ng/µL of fluorescently labelled RNA and the indicated recombinant proteins. Mixes with recombinant proteins were incubated for 18 minutes either on ice or at 30°C before heparin treatment. Complexes were resolved by native agarose/polyacrylamide gel electrophoresis. After electrophoresis, fluorescence was detected using a PhosphorImager Typhoon. To perform DNA-directed RNAse H inactivation of U1 and U2 snRNAs, 40 µl of HeLa cell nuclear extracts were supplemented with 1.25 mM ATP, 6.25 mM CP, 2.8 mM MgCl2 and 1 µL RNAse H and 14.25 µM oligonucleotide complementary to U1 or U2 snRNAs (U1 position 1 to 15: CTGCCAGGTAAGTAT, U2 position 28 to 42 : CAGATACTACACTTG) and incubated for 15 minutes at 37ºC and 30 minutes at 30ºC. Depletion of U2 snRNP extracts was achieved following [55] with an oligonucleotide targeting U2 snRNA (position 1 to 14, [56]). Nuclear extracts from HeLa Flp-In cells were prepared as described in [57].

Generation of HeLa Flp-In T-REx Stable Cell Lines

Flp-In T-Rex HeLa cells were kindly provided by Stephen Taylor (University of Manchester, Manchester, England, United Kingdom). Flp-In T-Rex tetracycline transactivator HeLa cells were transfected with 0,5 µg of pcDNA5 vector and 4,5 µg of pOG44 vector in a 10 cm plate (10^6 cells) using Lipectamine 2000 (Invitrogen). Colonies were selected as for HEK293 Flp-In T-Rex but 150 □g/mL Hygromycin B was used.

Expression and purification of recombinant peptides

The optimized cDNAs were cloned in pEMT-11 vector to facilitate the expression of His-tagged SRRM3/4 derived peptides. The vectors were transformed in BL21 DE3 bacterial strain,

one colony was grown overnight in 10 mL LB culture under antibiotic selection, the culture was diluted in 200 mL and the induction of expression was started at OD (600 nm) 0.6 with 1 mM IPTG for 4 hours at 37°C. The cells were sedimented and resuspended in 1 mL Cell Lytic B Cell Lysis buffer (Sigma) supplemented with 0.4 mg/mL lysozyme. The samples were incubated 20 minutes at room temperature, sonicated for 10 minutes (30 sec ON / 30 sec OFF cycles, Bioruptor) and centrifuged 20 minutes at 13000 rpm at 4ºC. 7.5 mL of buffer A (50 mM Tris pH 8.0, 500 mM NaCl, 1% Triton-X100) and 1.5 ml slurry TALON Metal affinity resin (BD Biosciences) were added to the supernatant and incubated for 45 minutes with rotation at 4ºC. Beads were sedimented and washed 3 times with 8 ml of buffer A, once with 8 ml buffer A without Triton-X100 and eluted with 500 mM imidazole in buffer A without Triton-X100. Fractions were analysed on a 15% SDS-PAGE by Coomassie staining, the most concentrated one was dialyzed against buffer D (20 mM Hepes-KOH 8.0, 20% glycerol, 0.2 mM EDTA, 150 mM KCl, 1 mM DTT) and the protein concentration was estimated by Bradford quantification.

MS2-MBP was expressed in BL21 DE3 cells, and after centrifugation of the induced cells, the pellet was resuspended in 20 mM Tris-HCl pH 7.6, 200 mM NaCl, complete EDTA-free protease inhibitor tablet and sonicated. After centrifugation, the supernatant was incubated with amylose resin pre-equilibrated in the same buffer for 2 hours with rotation at 4ºC. The resin was washed with 20 mM Hepes-KOH pH 7.9, 150 mM NaCl, 0.05% NP40 and the protein eluted with 5 mM Na2HPO4, 15 mM maltose. The eluate was loaded onto a Heparin-Sepharose column (Amersham), the beads were washed with 5 mM Na2HPO4 and eluted with 20 mM Hepes-KOH pH7.9, 100 mM KCl, 15% glycerol, 0.5 mM DTT, 0.2 mM PMSF.

Purification of the RNP associated to *apbb1* substrate

To allow the formation of protein-RNA complexes, 6 µg of RNA bearing 3 MS2 stem loops at their 3' end were boiled for 1 min, cooled on ice for 5 minutes, mixed with 20-fold excess of MS2-MBP recombinant protein in 20 mM Hepes-KOH pH 7.9 and further incubated for 30 minutes on ice. The complexes were then mixed with HeLa cells nuclear extracts (CilBiotech) in a final volume of 250 µl containing 24 mM Hepes-KOH pH 7.9, 65 mM KCl, 0,0024 mM MgCl2, 2 mM ATP, 20 mM CP, 40% nuclear extracts and complemented or not with 875 ng His-tagged C3 peptide (3.5 ng/µL) and incubated for 9 minutes at 30ºC. From this, 240 µL were loaded on a 4 mL 10-30% glycerol gradient (Beckman Coulter Polypropylene ultra centrifugation tubes) and run for 3 hours at 55,000 rpm (SW60Ti rotor). The gradient was fractionated in three parts (top-middle-bottom), the middle and bottom parts were mixed and

incubated with 90 µl slurry amylose resin (NEB) pre-equilibrated in 20 mM Hepes-KOH pH 7.9, 200 mM NaCl. The samples were incubated with rotation for 10 minutes at 4ºC, beads were washed 5 times with 1 mL of 20 mM Hepes-KOH pH 7.9, 0.05 % NP40, 150 mM NaCl and eluted with 12 mM maltose added to this last buffer. Eluted proteins were precipitated in ethanol for 20 minutes at -20°C and centrifuged at 13,000 rpm for 30 minutes at 4°C, one third of the precipitated proteins were loaded on a 10% SDS-PAGE and silver stained using the Silver Quest Staining kit (Invitrogen). The remaining two thirds of each replicate were sent to LC-MSMS mass spectrometry analysis for identification at the CRG Proteomics Unit. The sequences of the peptides were added in the search for identification and quantification.

RNP Mass Spectrometry Acquisition

Samples were analyzed using a LTQ-Orbitrap Velos Pro mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) coupled to an EasyLC (Thermo Fisher Scientific (Proxeon), Odense, Denmark). Peptides were loaded onto a 2-cm pre-column and separated by reverse-phase chromatography using a 25-cm column with an inner diameter of 75 µm, packed with 5 µm C18 particles (Nikkyo Technos Co., Ltd. Japan). Chromatographic gradients started at 97% buffer A and 3% buffer B with a flow rate of 300 nL/min, and gradually increased to 93% buffer A and 7% buffer B in 1 min, and to 65% buffer A / 35% buffer B in 60 min. After each analysis, the column was washed for 10 min with 10% buffer A / 90% buffer B. Buffer A: 0.1% formic acid in water. Buffer B: 0.1% formic acid in acetonitrile.

The mass spectrometer was operated in positive ionization mode with nanospray voltage set at 2.2 kV and source temperature at 250°C. Ultramark 1621 for the FT mass analyzer was used for external calibration prior the analyses. Moreover, an internal calibration was also performed using the background polysiloxane ion signal at m/z 445.1200. The instrument was operated in DDA mode and full MS scans with 1 micro scans at resolution of 60,000 were used over a mass range of m/z 350-2000 with detection in the Orbitrap. Auto gain control (AGC) was set to 1e6, dynamic exclusion (60 seconds) and charge state filtering disqualifying singly charged peptides was activated. In each cycle of DDA analysis, following each survey scan the top ten most intense ions with multiple charged ions above a threshold ion count of 5000 were selected for fragmentation at normalized collision energy of 35%. Fragment ion spectra produced via collision-induced dissociation (CID) were acquired in the Ion Trap, AGC was set to 5e4, isolation window of 2.0 m/z, activation time of 0.1 ms and maximum injection time of 100 ms was used. All data were acquired with Xcalibur software v2.2.

RNP Mass Spectrometry Data Analysis

Raw MS/MS files were processed using Proteome Discoverer version 1.4 (Thermo Fisher Scientific, Bremen). Peak lists were searched using Mascot software version 2.5.1 (Matrix Science, UK) against the human swissProt database (version of April 2017) containing 20,797 protein entries, a list of 600 common contaminants, and all the corresponding decoy entries. The precursor ion mass tolerance was set to 7 ppm, and the fragment ion mass tolerance was set to 0.5 Da. Up to three missed cleavages were allowed, and Oxidation (M) and Acetylation (Protein N-term) were defined as variable modifications, whereas carbamidomethylation (C) was set as fixed modification. Significance Analysis of INTeractome (SAINTexpress version 3.6.1 [53] was used as a statistical tool to calculate the probability value of each potential interaction compared with background samples with no RNA, requiring a minimum of 2 unique peptides. Each biological replicate was analyzed independently against no-RNA controls before averaging of the score values and assessment of the Bayesian False Discovery Rates (Supplementary Data S5, [53]). High-confidence interactions are those with FDR ≤ 1%. For Fig. 4g and Extended Data Fig. 4b-d, background (i.e. no-RNA samples) numbers of spectral counts were subtracted from the average spectral counts between replicates, and log2 fold-change enrichment between each pair of conditions was calculated.

Code availability

All software used to analyze the data is publicly available and listed in the Reporting Summary. VASTDB files to run *vast-tools* are available to download for each species (https://github.com/vastgroup/vast-tools) as indicated in the methods section. Custom code to generate orthologous gene clusters and figure plots are available upon request.

Data availability

Raw RNA-seq data was submitted to the Sequence Read Archive (SRP149913). Mass spectrometry data was submitted to ProteomeXchange Consortium: AP-MS through MassIVE (massive.ucsd.edu, accession codes: MSV000082361, PXD009779) and RNP-MS via PRIDE[58] (PXD010034). All other RNA-seq datasets used in the study are listed in Supplementary Dataset 3.

References

1 Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**, 1511-1523 (2014).
2 Li, Y. I., Sanchez-Pulido, L., Haerty, W. & Ponting, C. P. RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. *Genome Res* **25**, 1-13 (2015).
3 Quesnel-Vallières, M., Irimia, M., Cordes, S. P. & Blencowe, B. J. Essential roles for the splicing regulator nSR100/SRRM4 during nervous system development. *Genes Dev* **29**, 746-759 (2015).
4 Nakano, Y. *et al.* A mutation in the Srrm4 gene causes alternative splicing defects and deafness in the Bronx waltzer mouse. *PLoS Genet* **8**, e1002966 (2012).
5 Quesnel-Vallieres, M. *et al.* Misregulation of an Activity-Dependent Splicing Network as a Common Mechanism Underlying Autism Spectrum Disorders. *Mol Cell* **64**, 1023-1034 (2016).
6 Calarco, J. A. *et al.* Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. *Cell* **138**, 898-910 (2009).
7 Raj, B. *et al.* Global regulatory mechanism underlying the activation of an exon network required for neurogenesis. *Mol Cell* **56**, 90-103 (2014).
8 Parras, A. *et al.* Autism-like phenotype and risk gene mRNA deadenylation by CPEB4 mis-splicing. *Nature* **560**, 441-446 (2018).
9 Gonatopoulos-Pournatzis, T. *et al.* Genome-wide CRISPR-Cas9 Interrogation of Splicing Networks Reveals a Mechanism for Recognition of Autism-Misregulated Neuronal Microexons. *Mol Cell* **72**, 510-524 (2018).
10 Putnam, N. *et al.* The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064-1071 (2008).
11 Blencowe, B. J., Issner, R., Nickerson, J. A. & Sharp, P. A. A coactivator of pre-mRNA splicing. *Genes Dev* **12**, 996-1009 (1998).
12 Eldridge, A. G., Li, Y., Sharp, P. A. & Blencowe, B. J. The SRm160/300 splicing coactivator is required for exon-enhancer function. *Proc Natl Acad Sci USA* **96**, 6125-6130 (1999).
13 Khanna, M. *et al.* A systematic characterization of Cwc21, the yeast ortholog of the human spliceosomal protein SRm300. *RNA* **15**, 2174-2185 (2009).
14 Grainger, R. J., Barrass, J. D., Jacquier, A., Rain, J. C. & Beggs, J. D. Physical and genetic interactions of yeast Cwc21p, an ortholog of human SRm300/SRRM2, suggest a role at the catalytic center of the spliceosome. *RNA* **15**, 2161-2173 (2009).
15 Blencowe, B. J. *et al.* The SRm160/300 splicing coactivator subunits. *RNA* **6**, 111-120 (2000).
16 Nakanishi, N., Renfer, E., Technau, U. & Rentzsch, F. Nervous systems of the sea anemone Nematostella vectensis are generated by ectoderm and endoderm and shaped by distinct mechanisms. *Development* **139**, 347-357 (2012).
17 Wongpalee, S. P. *et al.* Large-scale remodeling of a repressed exon ribonucleoprotein to an exon definition complex active for splicing. *Elife* **5**, e19743 (2016).
18 McKeown, A. N. *et al.* Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. *Cell* **159**, 58-68 (2014).
19 Logeman, B. L., Wood, L. K., Lee, J. & Thiele, D. J. Gene duplication and neo-functionalization in the evolutionary and functional divergence of the metazoan copper transporters Ctr1 and Ctr2. *J Biol Chem* **292**, 11531-11546 (2017).
20 Arnegard, M. E., Zwickl, D. J., Lu, Y. & Zakon, H. H. Old gene duplication facilitates origin and diversification of an innovative communication system--twice. *Proc Natl Acad Sci USA* **107**, 22172-22177 (2010).

21 Wang, J. *et al.* LSD1n is an H4K20 demethylase regulating memory formation via transcriptional elongation control. *Nat Neurosci* **18**, 1256-1264 (2015).
22 Matsushita, M., Yamamoto, R., Mitsui, K. & Kanazawa, H. Altered motor activity of alternative splice variants of the mammalian kinesin-3 protein KIF1B. *2009* **10**, 1647-1654 (2009).
23 Tapial, J. *et al.* An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Res* **27**, 1759-1768 (2017).
24 Braunschweig, U. *et al.* Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res* **24**, 1774-1786 (2014).
25 Giampietro, C. *et al.* The alternative splicing factor Nova2 regulates vascular development and lumen formation. *Nat Commun* **6**, 8479 (2015).
26 Gueroussov, S. *et al.* An alternative splicing event amplifies evolutionary differences between vertebrates. *Science* **349**, 868-873 (2015).
27 Han, H. *et al.* MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498**, 241-245 (2013).
28 Solana, J. *et al.* Conserved functional antagonism of CELF and MBNL proteins controls stem cell-specific alternative splicing in planarians. *eLife* **5**, e16797, doi:10.7554/eLife.16797 (2016).
29 Burguera, D. *et al.* Evolutionary recruitment of flexible Esrp-dependent splicing programs into diverse embryonic morphogenetic processes. *Nat Commun* **8**, 1799 (2017).
30 Altenhoff, A. M., Gil, M., Gonnet, G. H. & Dessimoz, C. Inferring hierarchical orthologous groups from orthologous gene pairs. *PLoS One* **8**, e53786 (2013).
31 Alexeyenko, A., Tamas, I., Liu, G. & Sonnhammer, E. L. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* **22**, e9-15 (2006).
32 Singh, P. P., Arora, J. & Isambert, H. Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. *PLoS Comput Biol* **11**, e1004394 (2015).
33 Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**, 511-518 (2005).
34 Irimia, M. & Roy, S. W. Spliceosomal introns as tools for genomic and evolutionary analysis. *Nucleic Acids Res.* **36**, 1703-1712 (2008).
35 Yeo, G. W. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**, 377-394 (2004).
36 Corvelo, A., Hallegger, M., Smith, C. W. & Eyras, E. Genome-wide association between branch point properties and alternative splicing. *PLoS Comput Biol* **6**, e1001016 (2010).
37 Gohr, A. & Irimia, M. Matt: Unix tools for alternative splicing analysis. *Bioinformatics*, doi: 10.1093/bioinformatics/bty1606 (2018).
38 Gonzalez, M. *et al.* Generation of stable Drosophila cell lines using multicistronic vectors. *Sci Rep* **1**, 75 (2011).
39 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).
40 Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-755 (2001).
41 Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164-1165 (2011).

42 Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988-995 (2004).
43 Yachdav, G. *et al.* MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics* **32**, 3501-3503 (2016).
44 Wheeler, T. J., Clements, J. & Finn, R. D. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics* **15**, 7 (2014).
45 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
46 Sakamoto, H., Inoue, K., Higuchi, I., Ono, Y. & Shimura, Y. Control of Drosophila Sex-lethal pre-mRNA splicing by its own female-specific product. *Nucleic Acids Res* **20**, 5533-5540 (1992).
47 Fritzenwanker, J. H. & Technau, U. Induction of gametogenesis in the basal cnidarian Nematostella vectensis (Anthozoa). *Dev Genes Evol* **212**, 99-103 (2002).
48 Lambert, J. P., Tucholska, M., Go, C., Knight, J. D. & Gingras, A. C. Proximity biotinylation and affinity purification are complementary approaches for the interactome mapping of chromatin-associated protein complexes. *J Proteomics* **118**, 81-94 (2015).
49 Youn, J. Y. *et al.* High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Mol Cell* **69**, 517-532 (2018).
50 Liu, G. *et al.* Data Independent Acquisition analysis in ProHits 4.0. *J Proteomics* **149**, 64-68 (2016).
51 Deutsch, E. W. *et al.* A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **10**, 1150-1159 (2010).
52 Shteynberg, D. *et al.* iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol Cell Proteomics* **10**, M111.007690 (2011).
53 Teo, G. *et al.* SAINTexpress: improvements and additional features in Significance Analysis of INTeractome software. *J Proteomics* **100**, 37-43 (2014).
54 Mackereth, C. D. *et al.* Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. *Nature* **475**, 408-U174, doi:10.1038/nature10171 (2011).
55 Blencowe, B. J. & Lamond, A. I. Purification and depletion of RNP particles by antisense affinity chromatography. *Methods Mol Biol* **118**, 275-287 (1999).
56 Barabino, S. M., Blencowe, B. J., Ryder, U., Sproat, B. S. & Lamond, A. I. Targeted snRNP depletion reveals an additional role for mammalian U1 snRNP in spliceosome assembly. *Cell* **63**, 293-302 (1990).
57 Dignam, J. D., Lebovitz, R. M. & Roeder, R. G. Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res* **11**, 1475-1489 (1983).
58 Vizcaino, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* **32**, 223-226 (2014).

Acknowledgements

The authors thank Ben Lehner and Scott W. Roy for critical reading of the MS, Michael Akam and Ken Siggens for providing access to Strigamia specimens, Michael Sattler and Peijian Zou for pETM11 clones, Stephen Taylor for providing HeLa Flp-In cell line, Brith Bergum (Flow

cytometry core facility at Univ. of Bergen) for assistance with Nematostella FACS, and the CRG Genomics Unit. Animal silhouettes were obtained from PhyloPic. This work has been funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (ERC-StG-LS2-637591 to MI, and ERC-AdvG-670146 to JV), the Spanish Ministry of Economy and Competitiveness (BFU2014-55076-P to MI, BFU2014-005153 to JV, and the 'Centro de Excelencia Severo Ochoa 2013-2017', SEV-2012-0208), AGAUR, Fundación Botín (to JV), and the Canadian Institutes of Health Research (to BJB). RNP Mass spectrometric analyses were performed at the CRG/UPF Proteomics Unit which is part of the Proteored-PRB3 supported by PE I+D+i 2013-2016 (PT17/0019) of the ISCIII and ERDF, by "Programa CERCA Generalitat de Catalunya" and "Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement" (2017SGR595). ATM held a FPI-SO fellowship, YM a Marie Skłodowska-Curie IF.

Contributions

A.T-M. performed bioinformatic analyses, molecular biology and cell culture experiments. S.B. performed cell culture, spliceosome assembly and RNP-MS experiments with the supervision of J.V. Y.M. performed bioinformatics analyses. J.R. did and analysed AP-MS experiments under A-C.G. and B.J.B supervision, M.Ig. did *Nematostella* experiments with supervision from F.R., J.P. did *Strigamia* dissections and molecular biology work and D.O'H. did co-immunoprecipitation experiments under the supervision and B.J.B. T.G., I.A., F.G., M.S. did *Drosophila* and molecular biology experiments. M.Ir conceived and supervised the study, provided resources and performed bioinformatic analyses. A.T-M., S.B., M.Ir. and B.J.B. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing financial interests.

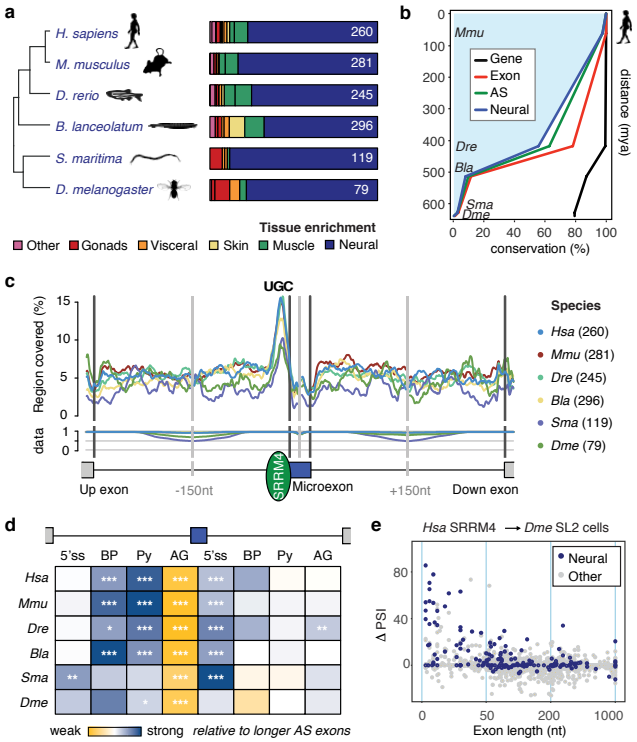


Figure 1 | Neural microexon programs in bilaterian animals

a, Phylogenetic tree of the main species used in the study and presence of tissue-specific microexons in bilaterians. Alternatively spliced microexons are classified based on their enrichment in a particular tissue type: neural, muscle (skeletal and/or heart), visceral (liver, kidney, lung, salivary gland, digestive tract or Malpighian tube), skin (or epidermis), gonads (female or male) and others (blood or early development). The number of neural microexons is indicated for each species. **b**, Evaluation of the conservation of human neural microexons in five other bilaterian species at different levels. Gene: presence of the gene ortholog in the second species; Exon: presence of the microexon ortholog in the same intron; AS: the orthologous microexon is also alternatively spliced; Neural: the orthologous microexon is also enriched in neural tissues. **c**, Enrichment of UGC motifs in the upstream intronic region of neural microexons. In brackets, number of exons analyzed per species. 'Data' subpanel refers to the proportion of total sequences used for the RNA map at each position as implemented by

Matt[37]. **d**, Splicing regulatory features of neural microexons compared to longer alternatively spliced exons. 5′ss and AG: splicing donor and acceptor MaxEnt scores, respectively; BP: best predicted branch-point sequence score; pY: polypyrimidine tract sequence score. Median score differences per feature were normalized to the maximum difference observed per feature type across species. P-values correspond to Mann-Whitney U tests: * <0.01, ** <0.001, *** < 0.0001. See Supplementary Fig. 1b and Supplementary Materials for details. **e**, Change in exon inclusion levels (ΔPSI) upon heterologous expression of human SRRM4 in *Drosophila* SL2 cells for all endogenous alternative spliced exons with sufficient read coverage. Blue dots indicate neural-enriched exons. X-axis scales linearly between every indicated exon length mark. Species abbreviations: *Bla, B. lanceolatum; Dme, D. melanogaster; Dre, D. rerio; Hsa, H. sapiens; Mmu, M. musculus; Sma, S. maritima.*

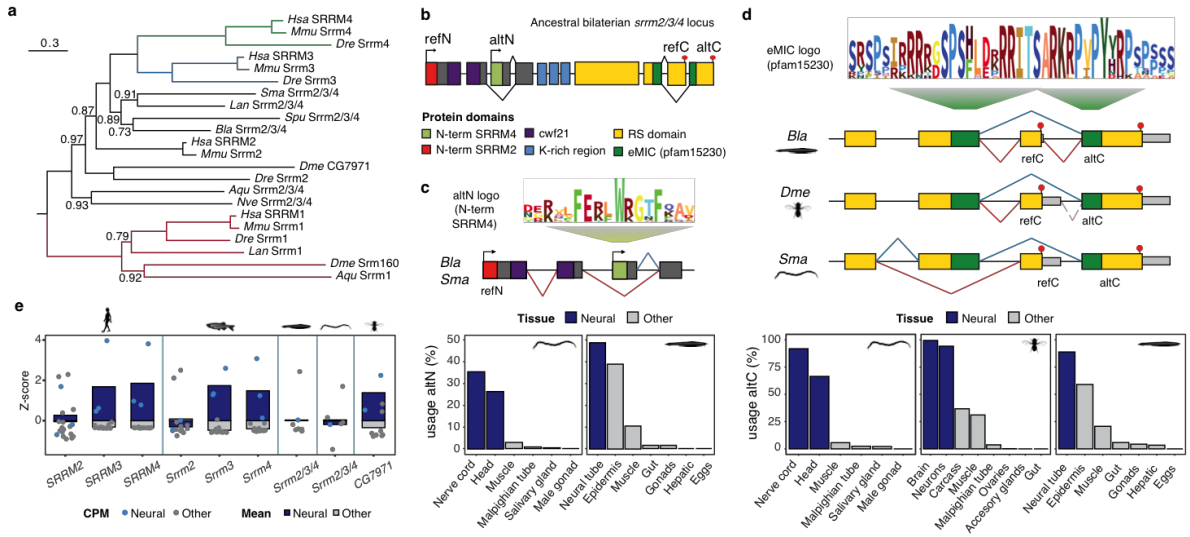


Figure 2 | Evolution of the *Srrm2/3/4* locus in metazoans

a, Bayesian phylogenetic tree of Srrm proteins in metazoans. Blue and green lines mark the vertebrate-specific duplication of the locus. Posterior probabilities lower than 1 are indicated for each node. Scale bar: aminoacid substitution frequency. Species abbreviations: *Aqu, Amphimedon queenslandica; Nve, N. vectensis; Dme, D. melanogaster; Dre, D. rerio; Hsa, H. sapiens; Mmu, M. musculus; Sma, S. maritima; Lan, Lingula anatina; Spu, Strongylocentrotus purpuratus; Bla, B. lanceolatum*. **b**, Prototypic gene structure of the *Srrm2/3/4* locus in non-vertebrate bilaterian animals highlighting the main protein features (refN/altN: reference (containing the cwc21 domain)/alternative N-terminus; refC/altC: reference/alternative (containing the full eMIC domain) C-terminus; RS domain: arginine/serine-rich domain). Arrows, red circles and non-horizontal lines represent promoters, stop codons and conserved alternative spliced isoforms, respectively. **c**, Alternative N-terminus of *Srrm2/3/4* in non-vertebrate bilaterians. Usage of the alternative promoter 'altN' (illustrated by blue lines) leads to the inclusion in neural tissues of a protein motif with high sequence similarity to vertebrate *Srrm4*. **d**, Alternative C-terminal region of *Srrm2/3/4* in non-vertebrate bilaterians. Usage of a different last exon by alternative splicing and/or alternative polyadenylation (altC, illustrated by blue lines) leads to the incorporation of the full eMIC domain (PFAM15230). **e**, Expression profile of *Srrm2/3/4* orthologs across tissues, quantified in counts per million (CPM) from RNA-seq data, and Z-score normalized per gene. Boxes represent the average value of the normalized CPM per tissue group (neural or non-neural). Other: non-neural tissues.

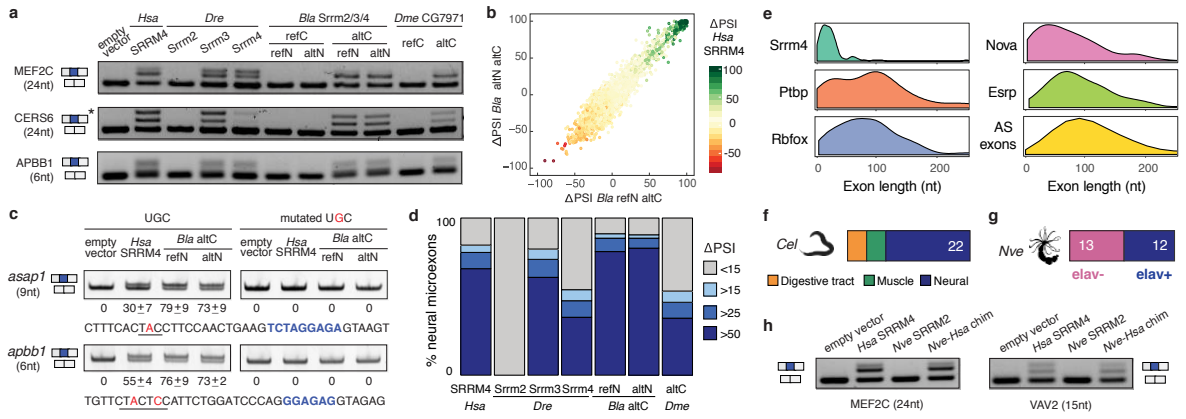


Figure 3 | The eMIC domain is key for neural microexon splicing

a, Effect on the inclusion of endogenous neural microexons of heterologous expression of *Srrm2/3/4* orthologs in human non-neural HEK293 cells. refN/altN, reference/alternative N-terminus; refC/altC, reference/alternative C-terminus. Band descriptions: constitutive flanking exons in grey and neural microexons in blue, asterisk in CERS6 gel marks a heteroduplex. **b**, Both eMIC-containing amphioxus isoforms produce similar transcriptome-wide effects in microexon inclusion (ΔPSI) when stably heterologously expressed in HEK293 cells irrespectively of their N-terminus. **c**, Co-transfection of amphioxus *Srrm2/3/4* eMIC-containing isoforms (refN-altC and altN-altC) with wild-type and UGC-mutated microexon minigenes from zebrafish in HEK293 cells. Numbers indicate quantifications from three replicates (average ± standard deviation). **d**, Effect of the heterologous expression of *Srrm2/3/4* orthologs on endogenous neural microexons as quantified by RNA-seq, grouped by the magnitude of ΔPSI. **e**, Distribution of exon lengths for the top 100 most affected exons by the knock-down of several tissue-regulated splicing factors, as quantified from public mouse RNA-seq datasets. AS exons: all mouse alternatively spliced exons. **f**, Tissue-enriched microexons in *C. elegans*. The number of neural microexons is indicated. **g**, Number of microexons in *N. vectensis* enriched in each of the two cell populations sorted from transgenic *Elav1*::mOrange polyps (*Elav1*+ is a marker of neuronal identity, Supplementary Figure 4d,e). **h**, Heterologous expression in HEK293 cells of *N. vectensis Srrm2*-like and a chimera fusing the C-terminal region of human *SRRM4* to the *N. vectensis* ortholog. For details on the chimeric construct, see Supplementary Fig. 3e. Species abbreviations: *Hsa, H. sapiens; Dre, D. rerio; Bla, B. lanceolatum; Dme, D. melanogaster; Cel, C. elegans; Nve, N. vectensis.*

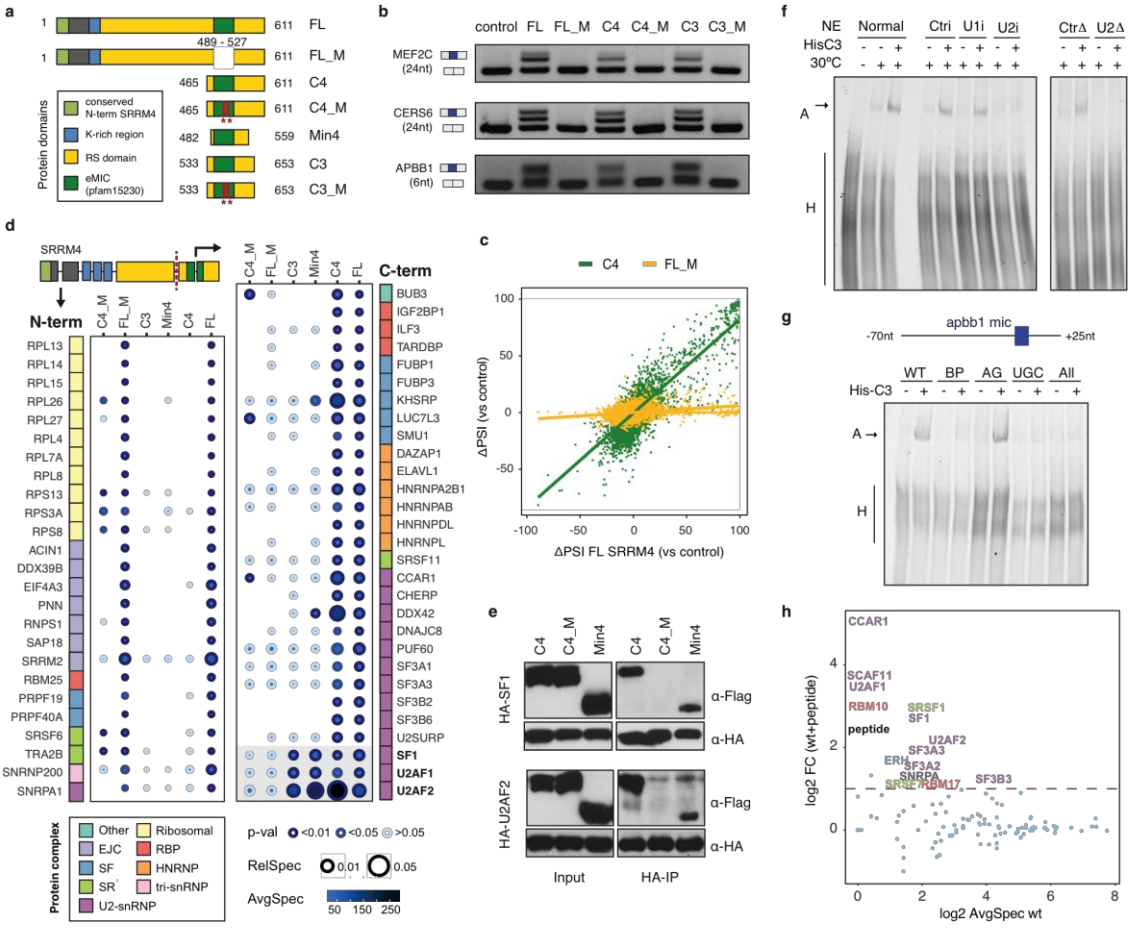


Figure 4 | The eMIC domain interacts with early spliceosomal factors to promote microexon inclusion

a, Schematic representation of protein constructs used to interrogate eMIC function. FL, full-length SRRM4; FL_M, SRRM4 with eMIC domain deleted; C4, C-terminal peptide of SRRM4; C4_M, loss-of-function mutant of C4 (T510A, K514Q); C3, C-terminal peptide of SRRM3; C3_M, loss-of-function mutant of C3 (T577A, K581Q); Min4: minimal construct with microexon regulatory capacity in HEK293 cells (see Supp. Fig. 5). End and start amino acid positions are relative to full-length SRRM4 or SRRM3. **b**, Overexpression of human SRRM4- and SRRM3-derived constructs in HEK293 cells. **c**, Genome-wide effect of C4 (containing the eMIC domain, green) compared to FL_M (no eMIC domain, yellow) on microexon inclusion measured as ΔPSI respect to the control (stable line overexpressing GFP). **d**, AP-MS identification of protein interaction partners for specific regions of SRRM3 and SRRM4. Green box: set of proteins interacting with all constructs containing the eMIC domain. RelSpec, relative spectral counts across each bait; AvgSpec, average number of spectral counts across replicates. **e**, Co-immunoprecipitation assay showing interaction of Flag-tagged eMIC-

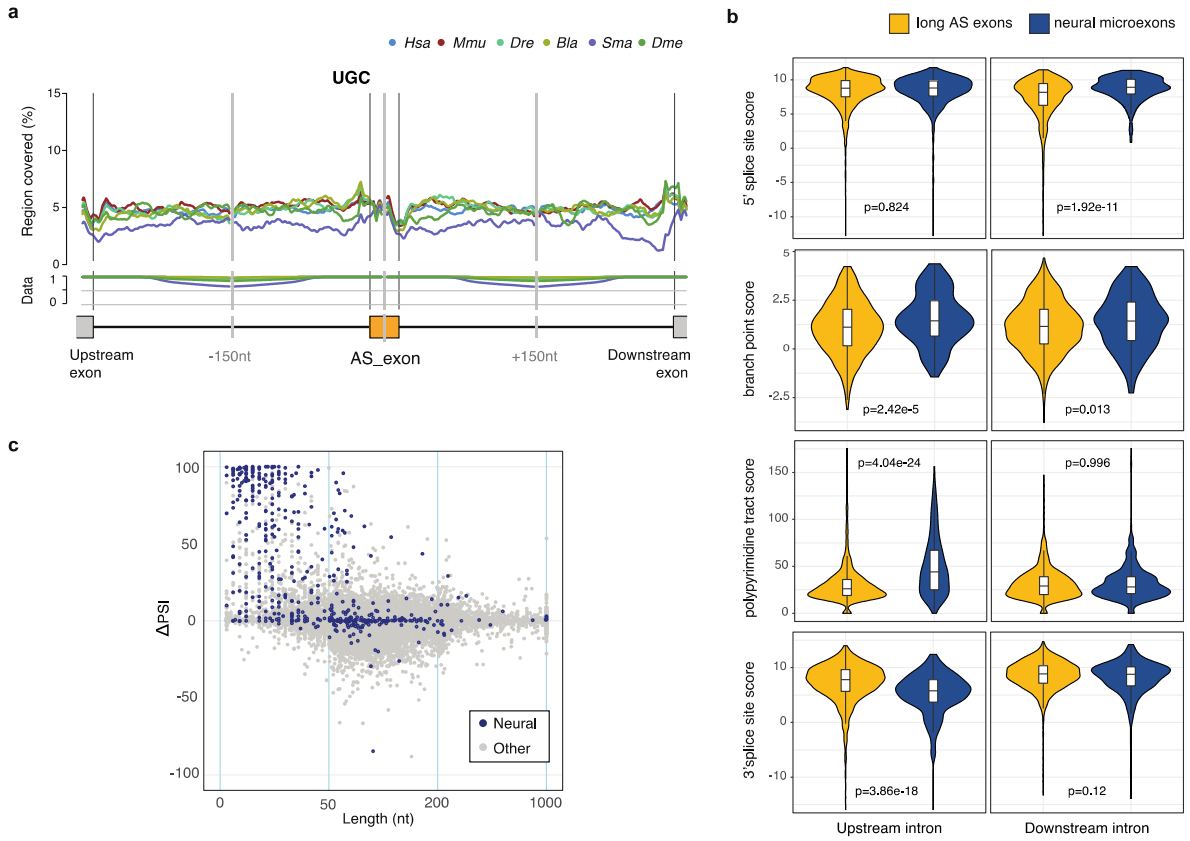
containing constructs with HA-tagged SF1 and U2AF2. **f**, C3 peptide promotes pre-spliceosome -complex A- assembly on *apbb1* RNA in functional U2 snRNP containing nuclear extracts. HeLa cell nuclear extracts were RNAse-H inactivated with a control (Ctr) oligonucleotide or with oligonucleotides complementary to the 5' end of U1 snRNA (U1i) or to the branch point recognition sequence of U2 snRNA (U2i); or depleted of U2 snRNP (U2Δ) and its control (CtrΔ); His-C3: His-tagged C3 peptide. The position of the H and A complexes are indicated. **g**, C3 peptide promotes pre-spliceosome -complex A- assembly on *apbb1* microexon minigene RNA in a sequence-dependent manner. WT: wild type RNA; BP, AG, UGC: branch point, 3' splice site AG and UGC element mutants, respectively; All: mutation of all three sequence elements. **h**, Effect of adding C3 peptide on the protein composition of complexes assembled onto the *apbb1* microexon RNA, as quantified by MS. AvgSpec: average number of spectral counts per protein in complexes formed in the absence of C3 peptide. FC: fold-change after addition of C3 peptide. Proteins are colored by protein complex, as per Fig. 4d.

Supplementary Materials for:

A novel protein domain drove the emergence of neural microexons in bilaterian animals

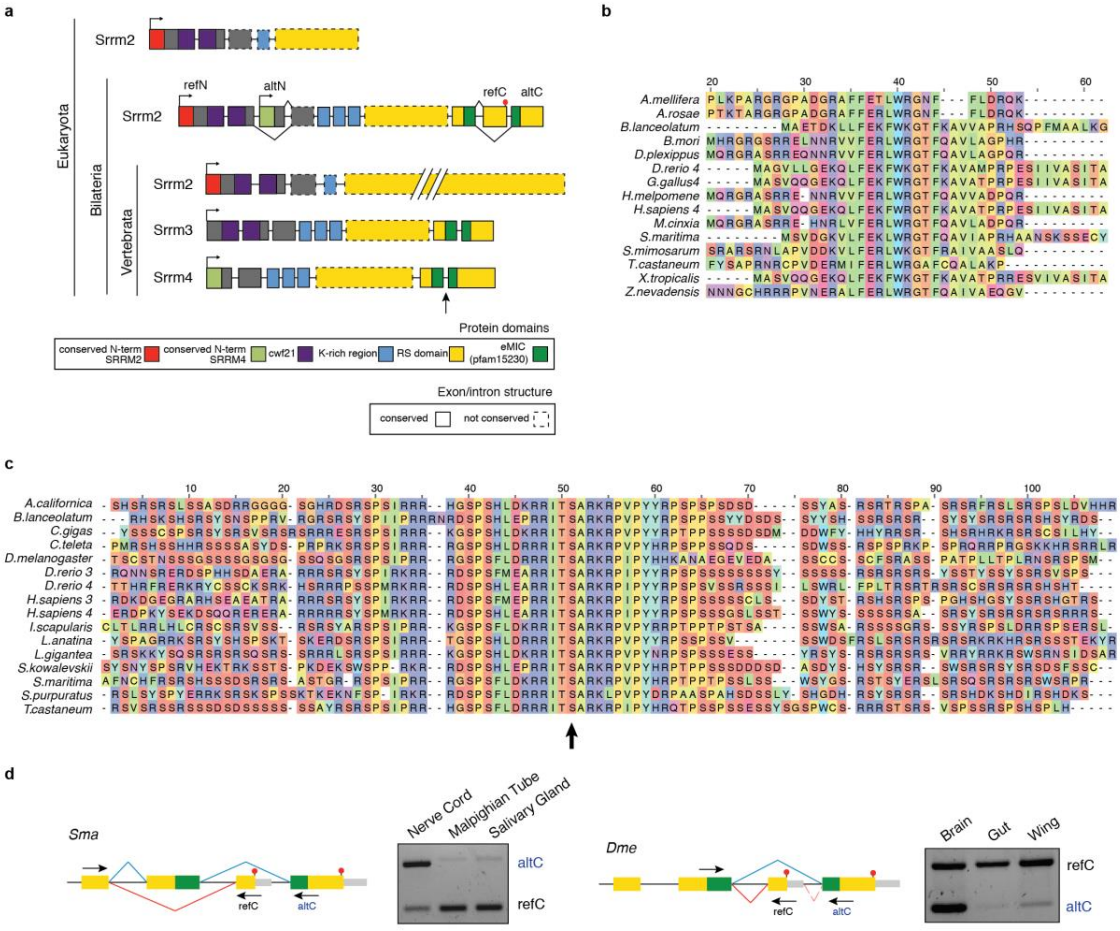
Antonio Torres-Méndez †, Sophie Bonnal †, Yamile Marquez, Jonathan Roth, Marta Iglesias, Jon Permanyer, Isabel Almudí, Dave O'Hanlon, Tanit Guitart, Matthias Soller, Anne-Claude Gingras, Fátima Gebauer, Fabian Rentzsch, Benjamin J. Blencowe, Juan Valcárcel, and Manuel Irimia *

Correspondence to: mirimia@gmail.com



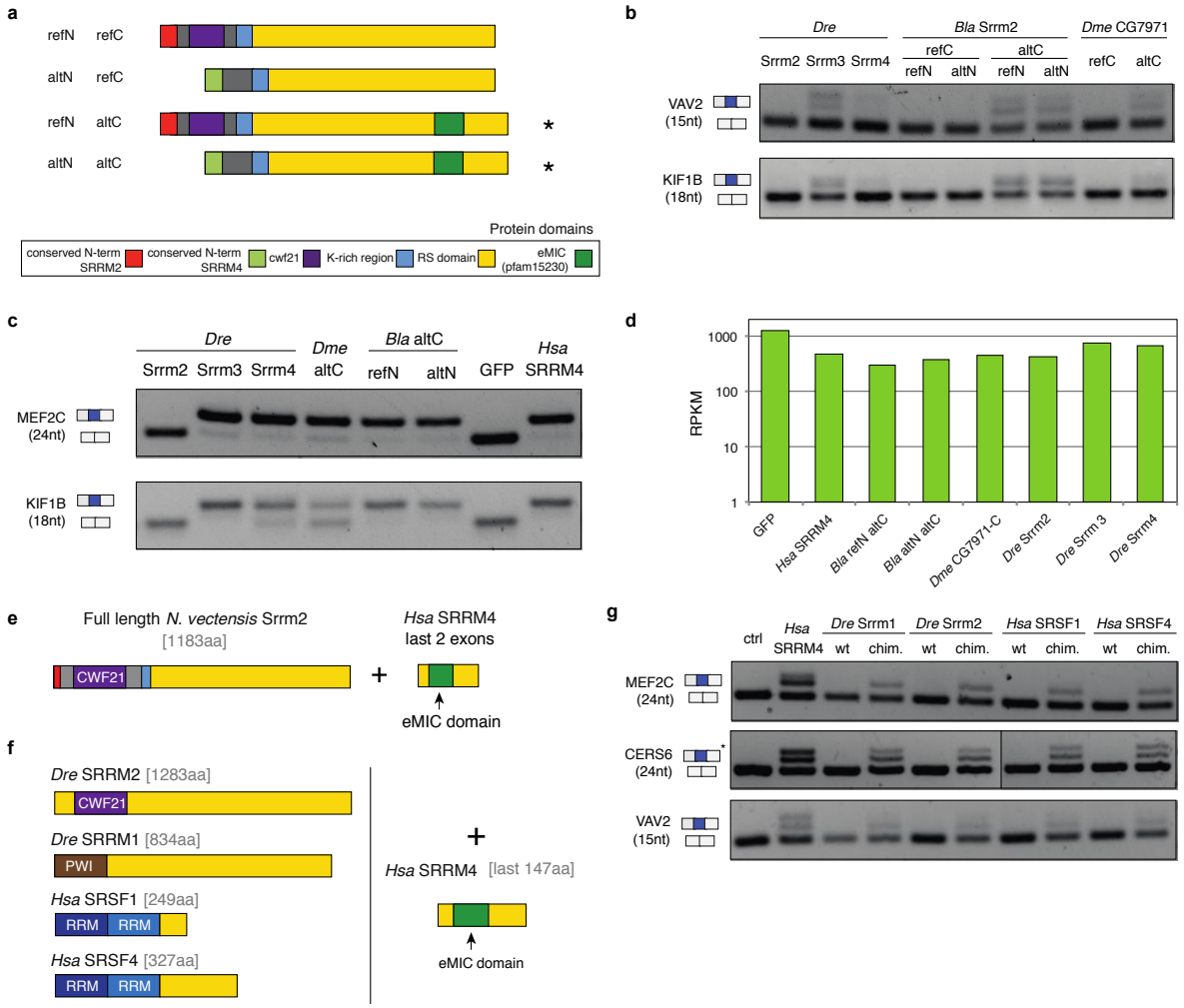
Supplementary Figure 1 | Neural microexon programs in bilaterian animals

a, Absence of UGC-motif enrichment in the intronic regions surrounding long (>27 nts) alternatively spliced exons. 1,000 randomly sampled exons were used per species. 'Data' subpanel refers to proportion of total sequences used for the RNA map at each position as implemented in Matt. **b**, Related to Fig. 1D, comparison of score distributions for several splicing features between neural microexons and longer alternatively spliced exons. Boxplot elementes: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range. P-value from Mann-Whitney U-tests. **c**, Effect on inclusion levels of heterologously expressing human *SRRM4* in non-neural HEK293 cells for all endogenous alternatively spliced exons with sufficient read coverage, grouped by tissue-specificity and sorted by exon length. x-axis scales linearly between each indicated exon length mark. Exon groups: Neural, higher inclusion in neural samples; Other, alternatively spliced exons with no neural enrichment. Species abbreviations: *Hsa, H. sapiens; Mmu, M. musculus; Dre, D. rerio; Bla, B. lanceolatum; Sma, S. maritima; Dme, D. melanogaster.*



Supplementary Figure 2 | Evolution of *Srrm2/3/4* loci in bilaterians

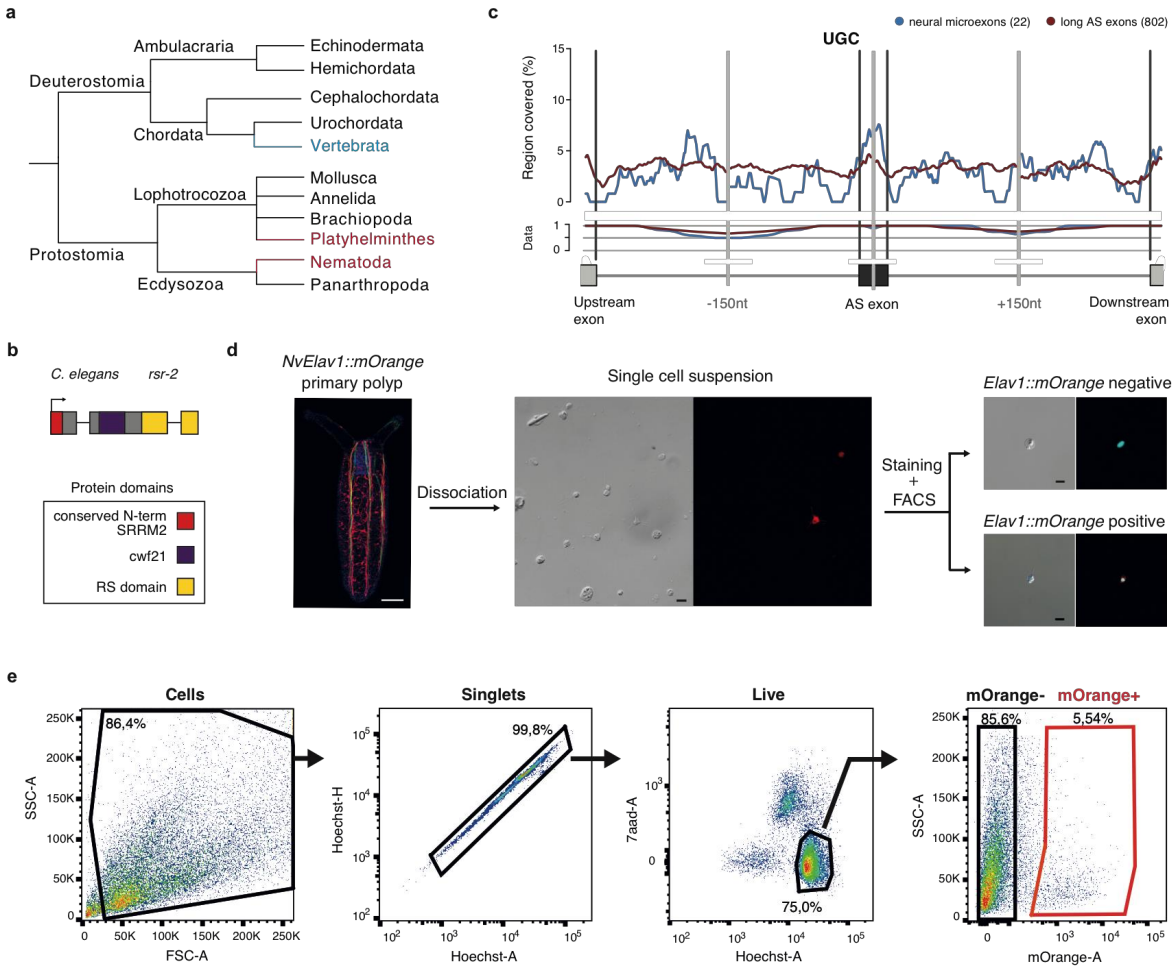
a, Exon-intron structure of *Srrm2/3/4* genes across metazoans. Conservation of the structure (solid line) requires that the introns are in the same position of the alignment (+/- 5 residues) and share the same phase. In vertebrates, duplication of the locus resulted in three paralogs with different protein domain composition. refN: reference N-terminus (containing the cwc21 domain), altN: alternative N-term, refC: reference C-terminus, altC: alternative C-term (containing the full eMIC domain), K: lysine, RS domain: arginine/serine-rich domain. Arrows at the start of the gene represent promoters; red circles, open reading frame ends; non-horizontal lines represent possible alternative spliced isoforms. The arrow in last intron of *Srrm4* matches the arrow in (c). **b**, Protein alignment of the N-terminal sequence of vertebrate Srrm4 (indicated with a '4' after the species name) and the conserved altN in non-vertebrate bilaterians. **c,** Protein alignment of the eMIC domain at the C-terminus of vertebrate Srrm3 and Srrm4 and of neural non-vertebrate Srrm2/3/4 isoforms (altC), used for the protein logo in Fig. 2d. The arrow indicates the position of the last intron (as in panel (a)). **d**, RT-PCR assays showing the alternative usage of the eMIC domain (altC) in different tissues from *S. maritima* and *D. melanogaster*. Arrows indicate the position of the oligonucleotides used for RT-PCR.



Supplementary Figure 3 | Microexon regulatory activity of *Srrm4* orthologs

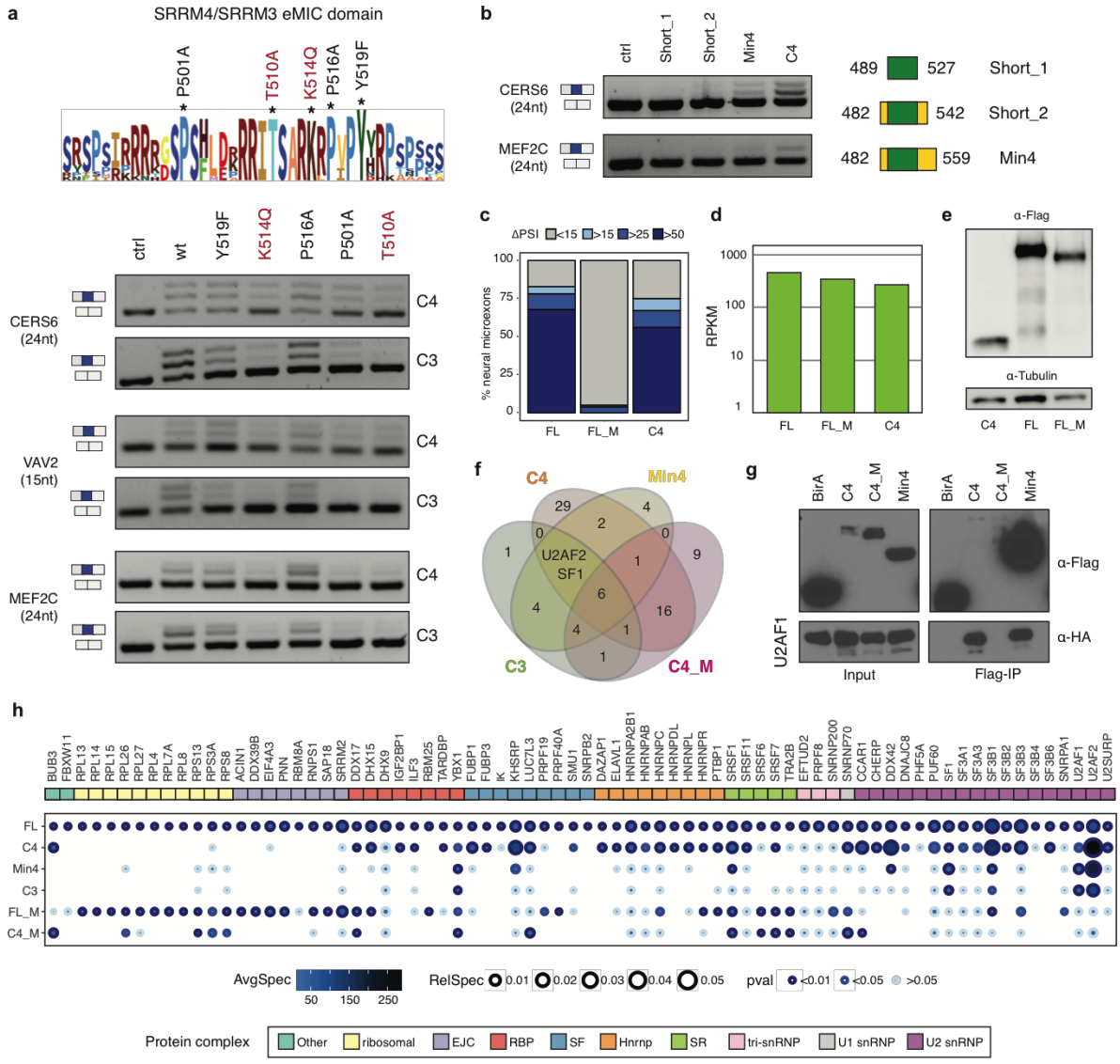
a, Schematic representation of the four main isoforms of the *Srrm2/3/4* gene from amphioxus that were heterologously expressed in human HEK293 cells (related to Fig. 3a). Stars indicate the two isoforms with microexon regulatory activity, i.e. containing the eMIC domain. refN: reference N-terminus (containing the cwc21 domain), altN: alternative N-term, refC: reference C-terminus, altC: alternative C-term (containing the full eMIC domain), K: lysine, RS domain: arginine/serine-rich domain. **b**, Effect of the transient heterologous expression of different *Srrm2/3/4* orthologs in non-neural HEK293 cells on *VAV2* and *KIF1B* endogenous microexons. Band descriptions: constitutive flanking exons in grey and neural microexons in blue. **c**, RT-PCR assays of two SRRM4-dependent microexons in the stable lines heterologously expressing different *Srrm2/3/4* orthologs or green fluorescent protein (GFP). **d**, Expression levels of *Srrm2/3/4* orthologs and GFP in the doxycycline-inducible HEK293 cell lines after 24h induction, as quantified from RNA-seq reads in RPKMs (reads per kilobase per million mapped reads). **e**, Protein domain composition of *N. vectensis Srrm2*-like and the chimera generated by fusing the cnidarian protein with the region

encoded by the last two exons of human *SRRM4* (containing the eMIC domain). In brackets, protein length in number of amino acids. **f**, Protein domain composition of other SR proteins tested for microexon regulatory activity and schematic representation of their corresponding protein chimeras with the last two exons of human *SRRM4*. **g**, Microexon regulatory activity of several SR proteins and chimeras fusing the eMIC domain of *SRRM4*, as indicated by the inclusion level of *MEF2C*, *CERS6* and *VAV2* microexons. wt: wild-type protein, chim: chimera. In brackets, microexon length. Asterisk marks a heteroduplex in *CERS6* RT-PCR. Species abbreviations: *Hsa, H. sapiens; Dre, D. rerio; Bla, B. lanceolatum; Dme, D. melanogaster.*



Supplementary Figure 4 | Absence of eMIC domain correlates with absence of neural microexon programs

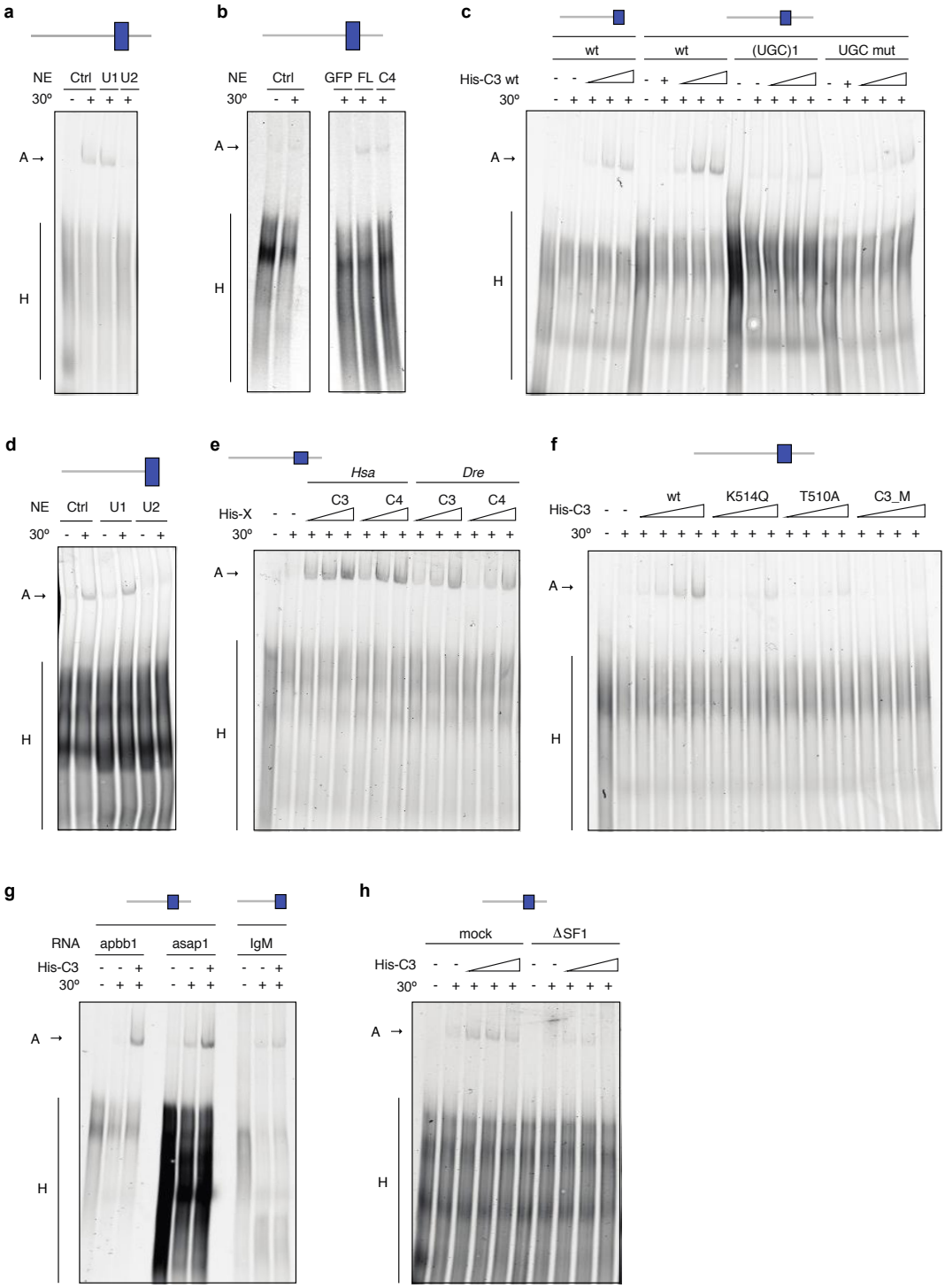
a, Phylogenetic tree of major bilaterian clades, indicating branches where eMIC domain loss or duplication occurred (red and blue, respectively). **b**, Exon-intron structure and protein domain composition of the *Srrm2/3/4* ortholog in *C. elegans* (*rsr-2*) shows a secondary loss of the eMIC domain. **c**, Absence of UGC motif enrichment in the upstream intronic regions of *C. elegans* neural microexons, comparable to longer alternatively spliced exons. In brackets, number of exons analyzed per group. 'Data' subpanel refers to proportion of total sequences used for the RNA map at each position as implemented in Matt [1]. **d**, Experimental workflow of fluorescence-activated cell sorting (FACS) of *Nematostella Elav1::mOrange* positive and negative cell populations. Representative confocal image of the neuron-specific transgenic *NvElav1::mOrange* used (left): red, immunostaining against mOrange; blue, DAPI; green, phalloidin; scale bar 100 μm. Representative bright-field and fluorescence images obtained after tissue dissociation (middle) and FACS-sorting (right): red, live mOrange-fluorescence; cyan, Hoechst3342-labeled nuclei; scale bar: 10 μm. **e**, FACS gating strategy applied to sort *Elav1::mOrange* positive and negative cell populations after excluding most of the debris from the cell suspension (cells gate), doublet/multiplets (singlets gate) and non-viable cells (live gate). Representative FACS plots showing the percentage of events gated in each step.



Supplementary Figure 5 | The eMIC domain is necessary and sufficient for neural microexon inclusion

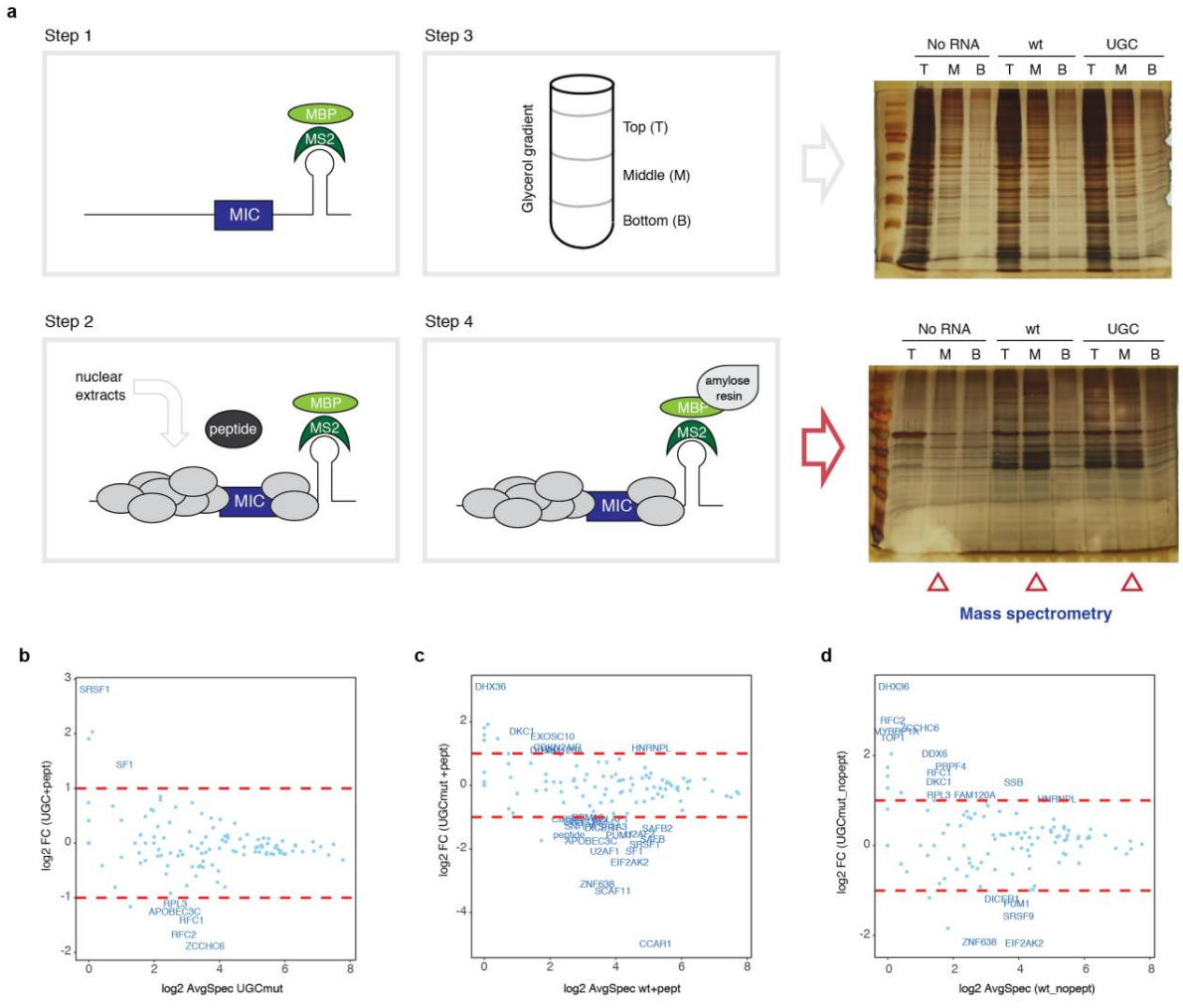
a, Single point mutations in eMIC domain largely disrupt the microexon regulatory activity of C4 and C3 (SRRM4 and SRRM3 C-terminal peptides, respectively). Amino acid positions relative to SRRM4 protein sequence. The relative effect of each mutation is comparable between C4 and C3, as indicated by the inclusion levels of endogenous microexons in HEK293 cells. **b**, Minimal peptides derived from SRRM4 (containing the eMIC domain) that have detectable microexon regulatory activity. Left: inclusion levels of two endogenous microexons, right: start and end positions of the peptide in the full-length protein. Min4: minimal peptide used for affinity-purification mass-spectrometry experiments depicted in Fig. 4d-e and Extended Data Fig. 5f-h. **c**,

Genome-wide effect of heterologously expressing full length SRRM4 (FL), SRRM4 lacking the eMIC domain (FL_M) and SRRM4 C-terminal peptide (C4) on the inclusion levels of endogenous neural microexons in HEK293 cells. Colors represent the magnitude of ΔPSI. **d**, Expression levels of FL, FL_M and C4 in doxycycline-inducible HEK293 cell lines after 24h induction, as quantified from RNA-seq reads in RPKMs (reads per kilobase per million mapped reads). **e**, Western-blot with anti-Flag M2 antibody shows similar expression levels of Flag-tagged FL_M and C4 proteins in doxycycline-inducible HEK293 stable cell lines used for RNA-seq. **f**, Venn-diagram showing number of protein partners (BFDR<=0.01) identified by affinity-purification mass-spectrometry (AP-MS) for the C-terminal peptides of SRRM3 and SRRM4 (C3 and C4), the minimal functional SRRM4 peptide (Min4) and the loss-of-function mutant of SRRM4 (C4_M). Only SF1 and U2AF2 interact with all and only functional eMIC-containing constructs, which we validated by co-immunoprecipitation (Fig. 4e). **g**, Co-immunoprecipitation assay showing interaction of Flag-tagged eMIC-containing constructs with HA-tagged U2AF1. BirA tag alone is used as negative control. **h**, Full list of protein partners (BFDR<=0.01) identified by AP-MS interacting with full-length SRRM4 (FL), complementary to Fig. 4d. RelSpec, relative spectral counts across each bait; AvgSpec, average number of spectral counts across replicates.


a
NE Ctrl U1 U2
30º - + + +
A →
H
b
NE Ctrl GFP FL C4
30º - + + + +
A →
H
c
wt wt (UGC)1 UGC mut
His-C3 wt
30º
A →
H
d
NE Ctrl U1 U2
30º - + - + - +
A →
H
e
Hsa Dre
C3 C4 C3 C4
His-X
30º
A →
H
f
wt K514Q T510A C3_M
His-C3
30º
A →
H
g
RNA apbb1 asap1 IgM
His-C3
30º
A →
H
h
mock ΔSF1
His-C3
30º
A →
H

Supplementary Figure 6 | The eMIC domain promotes sequence-specific spliceosomal A-complex formation

a, A complex formation on the *apbb1* microexon and flanking intronic sequences (including 3' 70 nucleotides of the upstream intron and 5' 25 nucleotides of the downstream intron) is ATP- and U2 snRNA-dependent. HeLa cells nuclear extracts were RNAse-H inactivated with a control (Ctr) oligonucleotide or with oligonucleotides complementary to the 5' end of U1 snRNA (U1) or to the branch point recognition sequence of U2 snRNA (U2), as indicated. The position of H and A complexes are indicated. **b**, Nuclear extracts from doxycycline-inducible HeLa cells overexpressing SRRM4 (FL) and C4 peptide promote pre-spliceosomal -complex A- formation on the *apbb1* microexon RNA substrate. **c**, His-tagged C3 peptide promotes pre-spliceosome -complex A- assembly on *apbb1* RNA in a UGC sequence-dependent manner. **d**, His-tagged C3 peptide promotes A complex formation on the *apbb1* substrate in the absence of a 5' splice site. Assays were as in panel 5SA using an RNA substrate lacking sequences of the downstream intron in the presence or absence of 50 ng/ul of His-tagged C3 peptide. **e**, A complex formation on the *apbb1* substrate is enhanced by His-tagged SRRM3/SRRM4-derived peptides from both human (*Hsa* C3/C4) and zebrafish (*Dre* C3/C4) proteins. Assays were carried out as in Extended Data Fig. 5a using 16,7; 50 and 150 ng/ul of the indicated peptides. **f**, A complex formation on the *apbb1* substrate is sensitive to mutations of the His-tagged C3 peptide that inactivate SRRM function (Extended Data Fig. 4b). Assays were carried out as in Extended Data Fig. 5a using 5,5; 16,7; 50 and 150 ng/ul of the wild type or mutant peptides described in Extended Data Fig. 4a,b. **g**, Activity of His-tagged C3 peptide on A complex formation is substrate dependent. Pre-spliceosomal -- complex A -- was assembled on RNAs comprising *apbb1* and *asap1* SRRM4-dependent neural microexons and flanking intronic sequences (70 nucleotides upstream and 25 nucleotides downstream), or the 3' 70 nucleotides exon 2 of *IgM*. **h**, Spliceosome assembly reactions using either mock- or SF1-immunodepleted HeLa cells nuclear extracts[2] and increasing amounts of His-tagged C3 peptide. All experiments were done on *apbb1* microexon derived substrates, including or not the first 25nt of the downstream intron, as indicated in each panel. Position of H and A complexes are indicated.



Supplementary Figure 7 | Characterization of ribonucleoprotein complexes (RNPs) assembled on a microexon

a, Purification of complexes assembled on the MS2-tagged *apbb1* RNA. Two successive purification steps involved gradient fractionation and affinity purification of MS2-MBP-bound transcripts. MS2-tagged RNAs (bearing either WT or UGC mutated sequences) (as well as a condition without RNA used as a negative control) were incubated with MS2-MBP tagged protein (step 1) and further incubated under splicing condition in HeLa cells nuclear extracts (step 2). Complexes were fractionated on a glycerol gradient (step 3) and further purified on an amylose resin (step 4). The protein composition of the different fractions was visualized by silver staining of a SDS-PAGE after steps 3 and 4. The fractions used for the LC-MSMS of the pilot experiment are labeled with triangles. **b**, Effect of adding His-tagged C3 peptide on the protein composition of complexes assembled onto the UGC-mutated *apbb1* microexon RNA, as quantified by MS. **c**, Effect of the UGC mutation on the RNP composition in presence of His-tagged C3 peptide. **d**,

Effect of the UGC mutation on RNP composition in the absence of His-tagged C3 peptide. Panels b-d: AvgSpec, average number of spectral counts per protein; FC, fold-change in the number of spectral counts between conditions. Only proteins (dots) with a fold enrichment or depletion over the reference sample higher than two (marked by dotted red lines), and with at least 5 average spectral counts have a text descriptor.