Recursive splicing in long vertebrate genes

Christopher R. Sibley1,2,*, Warren Emmett†*, Lorea Blazquez1, Ana Faro5, Nejc Haberman1, Michael Briese2,5, Danijah Trabzuni1,6, Mina Ryten1,7, Michael E. Weale7, John Hardy1, Miha Modic2,8, Tomaž Curk9, Stephen W. Wilson4, Vincent Plagnol3 & Jernej Ule1,2

It is generally believed that splicing removes introns as single units from precursor messenger RNA transcripts. However, some long *Drosophila melanogaster* introns contain a cryptic site, known as a recursive splice site (RS-site), that enables a multi-step process of intron removal termed recursive splicing1,2. The extent to which recursive splicing occurs in other species and its mechanistic basis have not been examined. Here we identify highly conserved RS-sites in genes expressed in the mammalian brain that encode proteins functioning in neuronal development. Moreover, the RS-sites are found in some of the longest introns across vertebrates. We find that vertebrate recursive splicing requires initial definition of an ‘RS-exon’ that follows the RS-site. The RS-exon is then excluded from the dominant mRNA isoform owing to competition with a reconstituted 5′ splice site formed at the RS-site after the first splicing step. Conversely, the RS-exon is included when preceded by cryptic promoters or exons that fail to reconstitute an efficient 5′ splice site. Most RS-exons contain a premature stop codon such that their inclusion can decrease mRNA stability. Thus, by establishing a binary splicing switch, RS-sites demarcate different mRNA isoforms emerging from long genes by coupling cryptic elements with inclusion of RS-exons.

Recursive splicing has been validated within the long introns (>24 kilobases (kb)) of three *D. melanogaster* genes1–2. The RS-sites in these introns contain a 3′ splice site followed by a sequence that reconstitutes a 5′ splice site after the first part of the intron is spliced, thereby allowing subsequent splicing of the second part of the intron (Fig. 1a). While one mammalian sequence located at the start of an alternative exon was proposed to function as an RS-site when pre-spliced to an upstream exon in a splicing reporter2, recursive splicing has not been observed in endogenous vertebrate genes. This is despite >8,000 human protein-coding genes containing introns >24 kb, and many vertebrate genes containing motifs similar to the *D. melanogaster* RS-sites1.

Long genes exhibit increased expression in the nervous system, as evident by analysis of human tissues or differentiating cells3 (Fig. 1b and Extended Data Fig. 1b–d), and are enriched in Gene Ontology (GO) terms associated with the nervous system (Extended Data Fig. 1a). We therefore produced 1.5 billion paired-end total RNA sequencing reads in RNA-seq data (Extended Data Fig. 2c–e). We proposed that if some of these were major splicing events, they should cause significant deviations from the expected linear decrease of reads across long introns (Fig. 1c, d). Analysis of our RNA-seq data identified 40,163 unique, unannotated cryptic splice sites in introns >1 kb that contained either 5′ or 3′ splice site motifs, 419 of which conformed to the RS-site motif (Supplementary Table 1, worksheets 1 and 2). We evaluated deviations from the expected saw-tooth pattern by establishing an analysis that computed the fit of linear regression slopes of each intron as a single unit or as two units separated at newly detected intronic junctions (Fig. 1c–e and Extended Data Figs 2a, b and 3). Since intron size is a crucial determinant of our ability to detect unexpected saw-tooth patterns reliably, we restricted our analysis to genes with at least one intron >150 kb. This identified 19 unique cryptic splice sites in the long introns of 14 genes that significantly improved the goodness-of-fit of the regression model in both RNA-seq and FUS iCLIP data sets. Of these, 9 had the RS-site motif while the remainder had a 3′ splice site motif (P < 0.01 in both data sets, Fig. 1d–f and Supplementary Table 1, worksheet 3). The genes containing these 9 RS-sites mostly function in cell adhesion and axon guidance and are linked to neurodevelopmental disorders (Supplementary Table 2).

The 9 RS-sites occurred at transition points of intronic linear regression slopes in all four individuals and all brain regions profiled (Fig. 1d and Extended Data Figs 3 and 4). Reverse transcription PCR (RT–PCR) from a separate human brain validated splice products for 8 RS-sites which were detectable at identical PCR cycle number as the mature mRNA, suggesting equal abundance, while no PCR products were observed when reverse primers were shifted upstream of RS-sites (Fig. 2a and Extended Data Fig. 5a–g).

Notably, an alternative 5′ splice site is present downstream of each RS-site that could lead to inclusion of alternative exons (hereafter termed RS-exons, Fig. 2b). However, RS-exons were not detectable in mRNA transcripts at comparable PCR cycle numbers used to detect RS-site junctions (Fig. 2a and Extended Data Fig. 5a–g), arguing that RS-sites are being used for recursive splicing rather than for RS-exon inclusion. Despite RS-exon skipping, mammalian conservation of both the RS-sites and alternative 5′ splice sites following the RS-exons is comparable to that of canonical 5′ and 3′ splice sites (Fig. 2c, d and Extended Data Fig. 5i). Indeed, mouse FUS iCLIP regression patterns directly match conserved RS-sites (Extended Data Fig. 6a–b).

Splicing of most vertebrate exons requires exon definition4, in which both splice sites flanking an exon are recognized in unison via interactions between U2AF proteins, Ser/Arg-rich (SR) proteins and small nuclear ribonucleoproteins (snRNPs)5 (Supplementary Information). We speculated that RS-exons co-evolved with RS-sites to enable exon definition (Fig. 2e). Accordingly, we masked the 5′ splice site following the CADM1 and ANK3 RS-exons in SH-SY5Y neuroblastoma cells using an antisense oligonucleotide (AON-A1; Fig. 2e). This markedly

---

1Department of Molecular Neuroscience, UCL Institute of Neurology, Queen Square, London WC1N 3BG, UK. 2MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK.
3University College London Genetics Institute, Gower Street, London WC1E 6BT, UK. 4Department of Cell and Developmental Biology, University College London, Gower Street, London WC1E 6BT, UK.
5Institute for Clinical Neurology, University of Würzburg, Versobachestr. 5, 97078, Würzburg, Germany. 6Department of Genetics, King Faisal Specialist Hospital and Research Centre, Riyadh 12111, Saudi Arabia. 7Department of Medical & Molecular Genetics, King’s College London, Guy’s Hospital, London SE1 9RT, UK. 8Institute of Stem Cell Research, German Research Center for Environmental Health, Helmholtz Center Munich, 85764 Neuherberg, Germany. 9Faculty of Computer and Information Science, University of Ljubljana, 1000 Ljubljana, Slovenia.

*These authors contributed equally to this work.
reduced RS-site usage in both genes (Fig. 2f). We subsequently repli-
cated this observation in vivo at the conserved RS-site/RS-exon of the zebras
fish cadm2a gene (Fig. 2g and Extended Data Fig. 5h). The reduced RS-site
usage also led to a ~6-fold increase in abundance of the intronic region upstream of both human RS-sites, indicating a change in the saw-tooth pattern consistent with splicing of the intron as a whole (Fig. 2h). Interestingly, the reduced RS-site usage caused a ~2-fold reduction in zebrafish cadm2a total mRNA (Fig. 2i), an effect not seen for the human CADM1 and ANK3 genes (Fig. 2) and Supplementary Information). Despite RS-exons usually being skipped, our findings demonstrate that RS-exon definition is crucial for the initial step of vertebrate recursive splicing (Fig. 2e).

Since recursive splicing requires initial definition of an RS-exon, we questioned whether some annotated alternative exons might function as RS-exons. We found 99 candidate annotated RS-exons with RS-site sequences located precisely at their starts (Extended Data Fig. 7a). Splice-junction reads from brain RNA-seq data were present at the start of 16 of these exons despite evidence for exon skipping. These included exons in the CADM2 and NTM genes that significantly improved the goodness-of-fit of linear regression in RNA-seq and iCLIP data sets across their >150 kb introns (Supplementary Table 1, worksheet 4 and 5). We confirmed RS-site mediated exon-skipping in both genes by RT–PCR (Extended Data Fig. 7b, f). Thus, the first intron in CADM2 gene contains two RS-sites; the first followed by an unan-
notated RS-exon, and the second by an annotated RS-exon.

To validate the exon definition mechanism further, we established a splicing reporter containing the second CADM2 RS-site, the annotated RS-exon and its 5’ splice site, and the surrounding constitutive exons, each flanked by their nearest ~100 nucleotides of CADM2 intronic sequence (P1; Fig. 3a). Despite the >500-kb long intron being reduced to ~0.5 kb, the reporter replicated the findings of endogenous genes; 79% of mRNA isoforms skipped the RS-exon while RS-site usage was readily detected (Fig. 3b and Extended Data Fig. 8a). As expected given the need for exon definition to recognize RS-sites, mutating the 5’ splice site following the RS-exon greatly reduced RS-site usage, and the intron remained a single unit in most splicing intermediates (P1-
m1; Fig. 3a, b and Extended Data Fig. 8a).

Next, to examine why RS-exons are excluded from the mRNA, we mutated the RS-site of the CADM2 reporter to prevent formation of the reconstituted 5’ splice site after the first splicing event (Fig. 3a). Notably, this resulted in complete inclusion of the RS-exon, suggesting competition exists between the two 5’ splice sites at either end of the
Figure 3a) Agreeing with the splicing reporter, AON-A2 markedly increased RS-exon inclusion in all human and zebrafish experiments (Fig. 3c and Extended Data Fig. 8b). Collectively, this demonstrates that the RS-exon is skipped owing to a splice site competition that leads to use of the reconstituted 5′ splice site instead of the 5′ splice site of the RS-exon (Fig. 3a, and Supplementary Information).

We noticed that RS-exons typically contain one or more in-frame stop codons (Fig. 2b and Extended Data Fig. 5i), inclusion of which should prevent translation of full-length protein and target transcripts with preceding start codons to nonsense-mediated decay. We included inclusion of the RS-exons in CADM1 and ANK3 by masking the 5′ splice site of their RS-sites (Fig. 3c, and Supplementary Information). This increased the proportion of isoforms containing the RS-exon (Fig. 3d), confirming that RS-exon inclusion can target transcripts for nonsense-mediated decay and thus has the potential to regulate transcript stability (Supplementary Information).

Having identified the mechanisms underlying vertebrate recursive splicing, we next explored the functions of RS-sites. Although *D. melanogaster* RS-sites have been proposed to maintain splicing integrity of long introns, the assayed human and zebrafish long introns remained accurately spliced after recursive splicing inhibition with AON-A1 (Extended Data Fig. 8c). We therefore explored an additional hypothesis that RS-sites regulate inclusion of RS-exons under specific contexts. We identified minor isoforms in the CADM2 and *NTM* genes that use a different promoter, and were therefore not detected by our initial RT–PCR reactions. Their detection required 10 more amplification cycles compared to the dominant isoform, confirming that they are minor isoforms (Extended Data Fig. 7c, d, g). Surprisingly, RS-exons are completely included in these minor isoforms that have an alternative exon or promoter preceding the RS-site (Fig. 4a–c and Figure 4).

RS-exon (P1-m2; Fig. 3a, b). To compare with endogenous genes, we designed another antisense oligonucleotide, AON-A2, to mask the section of RS-sites that contributes to the reconstituted 5′ splice site in the human CADM1, ANK3 or zebrafish cadm2a genes (AON-A2; Fig. 3a). Agreeing with the splicing reporter, AON-A2 markedly increased RS-exon inclusion in all human and zebrafish experiments (Fig. 3c and Extended Data Fig. 8b). Collectively, this demonstrates that the RS-exon is skipped owing to a splice site competition that leads to use of the reconstituted 5′ splice site instead of the 5′ splice site of the RS-exon (Fig. 3a, and Supplementary Information).

We noticed that RS-exons typically contain one or more in-frame stop codons (Fig. 2b and Extended Data Fig. 5i), inclusion of which should prevent translation of full-length protein and target transcripts with preceding start codons to nonsense-mediated decay. We included inclusion of the RS-exons in CADM1 and ANK3 by masking the 5′ splice site of their RS-sites (Fig. 3c, and Supplementary Information). This increased the proportion of isoforms containing the RS-exon (Fig. 3d), confirming that RS-exon inclusion can target transcripts for nonsense-mediated decay and thus has the potential to regulate transcript stability (Supplementary Information).

Having identified the mechanisms underlying vertebrate recursive splicing, we next explored the functions of RS-sites. Although *D. melanogaster* RS-sites have been proposed to maintain splicing integrity of long introns, the assayed human and zebrafish long introns remained accurately spliced after recursive splicing inhibition with AON-A1 (Extended Data Fig. 8c). We therefore explored an additional hypothesis that RS-sites regulate inclusion of RS-exons under specific contexts. We identified minor isoforms in the CADM2 and *NTM* genes that use a different promoter, and were therefore not detected by our initial RT–PCR reactions. Their detection required 10 more amplification cycles compared to the dominant isoform, confirming that they are minor isoforms (Extended Data Fig. 7c, d, g). Surprisingly, RS-exons are completely included in these minor isoforms that have an alternative exon or promoter preceding the RS-site (Fig. 4a–c and Figure 4).
Figure 4 | Splice site competition allows a binary splicing switch for RS-exons. a, RNA-seq read density patterns in the CADM2 gene shown in 5-kb windows, with linear regression performed after the first intron is split at the two RS-sites indicated with blue vertical lines. Isomers expressed from the dominant and minor promoters in human frontal cortex tissue are shown, and primer locations used for b indicated by coloured arrows. Grey forward primer is located in the first exon of dominant isoform, blue forward primer is located in the first RS-exon, red forward primer is located in the first exon of alternative isoform (P2). Zoomed area represents the sequence at the start of the second RS-exon. b, c, RT–PCR analysis of RS-exon inclusion in indicated CADM2 isoforms (b) or indicated NTM isoforms (c) \( n = 4 \) and \( n = 3 \) respectively; Extended Data Fig. 7. Values are mean ± s.d. d, Schematic of CADM2 splicing reporter variants P1 and P1-m3, based on the dominant CADM2 isoform (white), and P2 and P2-m1, based on the minor CADM2 isoform (red). Splice site scores for reconstituted and RS-exon 5′ splice sites are indicated. e–f, Qiaxcel analysis of indicated CADM2 splicing reporter products after transfection in SH-SY5Y cells \( n = 3 \) or \( n = 4 \), 2 separate experiments. The expected size of PCR products is shown next to each electropherogram.

Extended Data Fig. 7c–g). Similarly, the RS-exon is also detected in expressed sequence tags of minor OPCML isoforms that contain alternative exons preceding the RS-site (Extended Data Fig. 9a). A related splicing mechanism that coordinates alternative promoters with downstream alternative splicing has been observed in the human EBP41 and EBP41L3 genes, although this involves a reconstituted 3′ splice site to make it distinct from recursive splicing12.

To understand how preceding exons can dictate inclusion of RS-exons in a binary manner, we compared the computationally predicted strengths of the three relevant 5′ splice sites in CADM2 (ref. 12); the 5′ splice sites reconstituted from the RS-site after its splicing to the preceding exon of either the dominant or minor isoforms, and the 5′ splice site of the RS-exon (Fig. 4d). We used the last three nucleotides of the preceding exon and the six nucleotides from the RS-site to calculate the scores of the reconstituted 5′ splice sites12. We found that the reconstituted 5′ splice site had a high score when the first exon is derived from the dominant promoter (10.6), a low score when derived from the minor promoter (5.1), while the 5′ splice site of the RS-exon had an intermediate score (7.0). This indicates that the strength of the reconstituted 5′ splice site dictates whether the RS-exon is included or skipped. Indeed, 5′ splice sites reconstituted from the preceding exon of the dominant isoform in all 9 high-confidence RS-sites had equal or higher splice site scores than the 5′ splice sites of their corresponding RS-exons, in agreement with observed RS-exon skipping (Extended Data Fig. 8d and Supplementary Table 3, worksheet 1).

To evaluate experimentally, we mutated the 5′ splice site of the CADM2 RS-exon in our splicing reporter such that its score was higher (12.2) than the reconstituted 5′ splice site of the dominant isoform (10.6, P1-m3; Fig. 4d). This mutation favoured RS-exon inclusion (Fig. 4e). We then replaced the preceding exon of the dominant isoform with the one from the minor isoform. This led to complete inclusion of the RS-exon, re-capitulating behaviour of the endogenous gene (P2; Fig. 4d, f). Finally, swapping the last three nucleotides of the preceding exon in the minor isoform to the sequence of dominant isoform led to RS-exon skipping, consistent with the higher score of the reconstituted 5′ splice site (10.6, P2-m1; Fig. 4d, f). These results reveal that the binary splicing switch is a consequence of the relative strengths of competing 5′ splice sites present after the RS-exon is spliced to the preceding exon.

Introns containing the high-confidence RS-sites are among the longest introns in all vertebrate species (Fig. 4g and Extended Data Fig. 9b). This includes Tetraodon nigroviridis, which has the shortest known vertebrate genome and otherwise contains very short introns13. Furthermore, 8 out of 9 of our high-confidence RS-sites are located in the long first intron of the gene. We confirmed that long introns generally have an increased incidence of cryptic exons and noisy splicing14,15,16 by observing an increased incidence of cryptic junctions in our RNA-seq data in long first introns (Extended Data Fig. 9c and Supplementary Table 3, worksheet 2 and 3). Because most of the 435 putative RS-sites identified in our study are present in the longest human genes (419 intronic loci, 16 annotated RS-exons; Fig. 4h), RS-
sites are thus well positioned to couple inclusion of cryptic exons with RS-exons. As most RS-exons contain a premature stop codon, this may also allow quality control of the novel mRNA isoforms (Supplementary Information).

In summary, recursive splicing of long vertebrate genes involves two steps (Fig. 4i). First, the RS-exon is defined, which requires its own 5’ splice site. After splicing of the RS-exon to the preceding exon, a new 5’ splice site is reconstituted from the RS-site that competes with the 5’ splice site of the RS-exon. The strength of the reconstituted 5’ splice site determines whether the RS-exon is skipped via recursive splicing or included. Notably, the upstream exons of dominant isoforms reconstitute a strong 5’ splice site that leads to recursive splicing, whereas other alternative exons, which commonly emerge in long introns to produce minor isoforms, generally end in sequences that lead to RS-exon inclusion. In light of studies linking aberrant expression of long genes to neurological diseases, mutations or deletions around RS-exon inclusion. In light of studies linking aberrant expression of long genes to neurological diseases, mutations or deletions around RS-exons may also contribute to human genetic diseases.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 30 May 2014; accepted 9 April 2015.
Published online 13 May 2015.

1. Burnette, J. M., Miyamoto-Sato, E., Schaub, M. A., Conklin, J. & Lopez, A. J. Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. Genetics 170, 661–674 (2005).
2. Hatton, A. R., Subramaniam, V. & Lopez, A. J. Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon–exon junctions. Mol. Cell 2, 787–796 (1998).
3. Grellscheid, S. N. & Smith, C. W. An apparent pseudo-exon acts both as an alternative exon that leads to nonsense-mediated decay and as a zero-length exon. Mol. Cell Biol 26, 2237–2246 (2006).
4. Shepard, S., McCready, M. & Fedorov, A. The peculiarities of large intron splicing in animals. PLoS ONE 4, e7853 (2009).
5. Thakurela, S. et al. Gene regulation and priming by topoisomerase IIs in embryonic stem cells. Nature Commun. 4, 2478 (2013).
6. Ameer, A. et al. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. Nature Struct. Mol. Biol. 18, 1435–1440 (2011).
7. Rogelj, B. et al. Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. Sci. Rep. 2, 603 (2012).
8. Ke, S. & Chasin, L. A. Context-dependent splicing regulation: exon definition, co-occurring motif pairs and tissue specificity. RNA Biol. 8, 384–388 (2011).
9. Robberson, B. L., Cote, G. J. & Berger, S. M. Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol. Cell. Biol. 10, 84–94 (1990).
10. McGinley, N. J. & Smith, C. W. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? Trends Biochem. Sci. 33, 385–393 (2008).
11. Polymenidou, M., Tan, J. S., Mohandas, N. & Conboy, J. G. Intrasplicing coordinates alternative first exons with alternative splicing in the protein 4.1R gene. EMBO J. 27, 122–131 (2008).
12. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comp. Biol. 11, 1384–1394 (2004).
13. Jaillon, O. et al. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431, 946–957 (2004).
14. Roy, M., Kim, N., Xing, Y. & Lee, C. The effect of intron length on exon creation ratios during the evolution of mammalian genomes. RNA 14, 2261–2273 (2008).
15. Pickrell, J. K., Pai, A. A., Gilad, Y. & Pritchard, J. K. Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genet. 6, e1001236 (2010).
16. Lagier-Tourenne, C. et al. Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 interrupt in processing long pre-mRNAs. Nature Neurosci. 15, 1488–1497 (2012).
17. Polymenidou, M. et al. Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. Nature Neurosci. 14, 459–468 (2011).
18. King, I. F. et al. Topoisomerases facilitate transcription of long genes linked to autism. Nature 501, 58–62 (2013).
19. Anders, S. & Huber, W. Differential expression analysis for sequence count data. Genome Biol. 11, R106 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank S. El-Andaloussi for technical support, J. Witten, J. König and U. laboratory members for comments on the manuscript, and remaining members of the UK Brain Expression Consortium: S. Guelli, K. D’Sa, M. Matarin, J. Vandrrovo, A. Ramasamy, J. A. Botia, C. Smith and P. Forabosco. This work was supported by the European Research Council (206776-CLIP and 617837-Translate) to J.U.; Marie Curie Post-doctoral Research Fellowship (267783-NeuroCRYSP) to LB.; the Slovenian Research Agency (J7-5460) to J.U. and T.C.; the UK NIHR Biomedical Research Centre at Moorfields Eye Hospital and UCL Institute of Ophthalmology to V.P. and W.E.; the Wellcome Trust to S.W.A. and F.U.; the UK Medical Research Council (MRC) (U105185858) to J.U.; MRC training fellowships to C.R.S. and M.B.; and MRC project grant (G0901254), MRC training fellowship (G0802462) and MRC Sudden Death Brain Bank.

Author Contributions C.R.S., M.B. and J.U. conceived and designed the project; C.R.S., L.B., A.F., M.B., M.M. and D.T. performed experiments; C.R.S., W.E., L.B., V.P., T.C. and J.U. analysed the data and interpreted results with contributions from M.R., M.E.W. and J.H.; C.R.S. and J.U. wrote the manuscript with contributions from W.E., V.P., L.B. and S.W.W.

Author Information The sequence data and scripts are publicly available from the European Genome-phenome Archive under the accession number EGAS00001001170, ArrayExpress (E-MTAB-3534), http://icount.biobas.is/ and https://github.com/vplagno/recursive_splicing. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.U. (j.u@ucl.ac.uk) and V.P. (v.plagno@ucl.ac.uk).
METHODS
RNA-seq library preparation and sequencing. Brain samples for analysis were provided by the Medical Research Council Sudden Death Brain and Tissue Bank (Edinburgh, UK). Transcriptomic analysis of postmortem human tissue was approved by The National Hospital for Neurology and Neurosurgery & Institute of Neurology Joint Research Ethics Committee, UK (REC reference number 10/H0716/3). All four individuals sampled were of European descent, neurologically normal during life and confirmed to be neuropathologically normal by a consultant neuropathologist using histology performed on sections prepared from paraffin-embedded tissue blocks. Twelve central nervous system regions were sampled from each individual. The regions studied were: cerebellar cortex, frontal cortex, temporal cortex, occipital cortex, hippocampus, the inferior olivary nucleus (sub-dissected from the medulla), putamen, substantia nigra, thalamus, hypothalamus, intrabular white matter and cervical spinal cord.

RNA was extracted using Qiagen tissue kits, and quality controlled as detailed previously. Libraries were prepared by the UK Brain Expression Consortium in conjunction with AROS Applied Biotechnology A/S. In brief, 100 ng total RNA was used as input for cDNA generation using the Ovation RNA-seq System V2 (NuGen Technologies). The RNA was processed according to the manufacturer’s protocol resulting in amplified cDNA from total RNA and concomitant de-selection of rRNA. Notably, reverse transcription in this protocol is carried out using both oligo-dT and random primers. This allowed total RNA profile patterns to be assessed with the latter and locations of splicing to be inferred. One microgram of the cDNA was fragmented using a Covaris S220 Ultrasonicator and the fragmented cdNA was used as the starting point for Illumina’s TruSeq DNA library preparation. Finally, library molecules containing adaptor molecules on both ends were amplified through 10 cycles of PCR. The libraries were sequenced using Illumina’s TruSeq V3 chemistry/HiSeq2000 and 100 base pair paired-end reads. The sequencing data was converted to fastq-files using Illumina’s CASAVA Software.

RNA-seq processing. Paired end RNA-seq data was mapped to the human genome (hg19) using STAR aligner (v.2.3) with default settings and known splice junctions from GENCODE. For high-confidence RS-site junction detection, alignments were processed from all intronic regions >150 kb using an in-house processing pipeline implementing python (v.2.7.2), BEDtools (v.2.17.0) and R (v.3.0.0). This size limit was chosen since linear regression patterns could most readily be evaluated in such long introns (Extended Data Fig. 2a, b), and represented 943 introns in 780 genes (ReFSeq release 60). Alignments from all 48 samples in >150 kb introns were combined and processed together unless indicated in the text. All spliced alignments with minimum flanking overlap of >10 nucleotides (hereafter termed anchor) and junction region exceeding 5 kb were selected and considered for further analysis. Each anchor sequence was then annotated to verify it conformed to a known splicing boundary (hereafter termed exon anchor). All further analysis was done using only those novel junctions that had a single exon anchor (Extended Data Fig. 2c). Novel junctions were then ruled out if they were not detected across either multiple brain regions or in multiple patients. We subsequently asked whether intronic sequences immediately adjacent to the novel junctions contained pentamers found at 1% of all 5′ splice sites genome-wide (Extended Data Fig. 2d), or sequences located at 3′ splice sites (polypyrimidine tract consisting of >11 pyrimidines present in the region of -22 to -1, including YAG as last three positions; Extended Data Fig. 2e). Novel junctions within 418 nucleotides of one another, the ninety-fifth percentile of exon lengths genome-wide, were considered in close enough proximity to have potential for exon formation. This analysis identified 2,981 novel junctions in introns >150 kb; 979 joined an upstream exon to an intronic 3′ consensus splice site, 1,296 joined an intronic 5′ consensus splice site with a downstream exon, and 353 pairs of junctions were proximally spaced in a manner that could form a novel exon (Supplementary Table 1, worksheet 1). For low confidence RS-site junction detection in introns >1 kb, the same process was repeated in which alignments were now processed from all intronic regions >1 kb, and the minimum novel junction span was now 100 bp. RS-sites identified in this analysis were not tested if multiple introns overlapped. Remove higher P-values since RNA-seq has depth to identify most frequently used introns. Remaining: 536. (4) Plot after/before ratios. After/before ratios >1 correspond to increased slope, and <1 to reduced slope of linear regression line across intron. (5) Significance threshold set at P < 0.01 for both FUS and RNA-seq. Remaining 24 junctions. (6) Select junctions with after/before ratio of >1 in both data sets. Remaining 21 junctions representing 19 unique intronic loci; indicated by YES in column AF of Table 5. Each junction was then enumerated and classified as ‘known’ or ‘novel’ using the known-gene UCSC annotations. If a junction was not present in this annotation database and subsequently classified as novel, this was considered limited evidence for recursive splicing. Examples were subsequently considered

or Illumina Miseq. The iCLIP libraries contained an experimental barcode plus a random barcode, which allowed multiplexing and the removal of PCR duplicates, respectively. The iCLIP data were mapped to hg19 using Bowtie2 and further processed as described previously.

Computational analyses. All scripts used for the analyses in this paper are available at the Github repository (https://github.com/vplagnol/recursive_splicing).

Linear regression analysis. To establish the analysis of linear regression, each annotated intron greater than 50 kb (in at least one Ensembl transcript) was first analysed independently (Extended Data Fig. 2a, b). Following evaluation of different sized windows, we ultimately divided the introns into two bins both the RNA-seq and FUS iCLIP data, we then computed the number of reads pairs mapping to each bin using samtools v0.19. We then ran a regression analysis with the number of mapped reads in each bin as a dependent variable. As a test, we first used this to examine genes containing multiple introns >50 kb. This showed that slopes of fitted regression lines were comparable for different long introns of the same gene (Extended Data Fig. 2a, b). Since the slope depends on transcriptional elongation rate, this observation agrees with the finding that transcription rate is relatively constant across individual genes. We therefore assumed a constant (unconstrained) slope across each entire gene. Reducing the 5-kb bin size or the intron length cut-off reduced the reliability in the method, implying individual units of >50 kb are most appropriate for this computational analysis. Accordingly, when splitting introns into two separate parts based on novel junctions, we focused on >150 kb introns to account adequately for this size limit.

Next, for our baseline model, we coded the positions of all potential exons located in the >150 kb-intron-long gene introns (based on Ensembl annotations) using binary dummy variables and let the fitted read count data resel to an arbitrary value at each putative exon. We then considered for each intron a set of augmented models that include the same covariates at the baseline model (constant slope, dummy variable for potential exons) in addition to an additional dummy variable for each of the novel junctions identified by the split read analysis.

We used a standard F-test P-value to compare the fit between the baseline model and the augmented one to quantify the improvement of the goodness-of-fit provided by each additional potential RS-sites. Introns were eventually ranked on the basis of these F-test P-values, with significance threshold for further analysis set at P < 0.01 for both data sets (Supplementary Table 1, worksheet 3). Taken together, the following filtering workflow was used in linear regression analysis for production of Fig. 1.d. (1) Select novel junctions, which connect upstream exon to deep intronic loci. Initial junctions: 1,378. (2) Exclude junctions in which the gradient remains negative after strand correction. Remaining: 1,146. (3) Select lowest P-value for a junction if multiple introns overlap. Remove higher P-values since RNA-seq has depth to identify most frequently used introns. Remaining: 536. (4) Plot after/before ratios. After/before ratios >1 correspond to increased slope, and <1 to reduced slope of linear regression line across intron. (5) Significance threshold set at P < 0.01 for both FUS and RNA-seq. Remaining 24 junctions. (6) Select junctions with after/before ratio of >1 in both data sets. Remaining 21 junctions representing 19 unique intronic loci; indicated by YES in column AF of Supplementary Table 1, worksheet 3.

Alternative GURAG exon analysis. All alternative exons within the UCSC AltEvents track were evaluated for GURAG pentamers in their 3′ UTR. Two lines of evidence were then pursued to evaluate their use as RS-exons. First, we asked whether exons overlapped intronic read transition points despite being skipped. Linear regression analysis was performed on all alternative exons from UCSC AltEvents table that fell within an Ensembl transcript and would have flanking introns both >50 kb (Supplementary Table 1, worksheet 4). Analysis was performed using both RNA-seq and FUS data sets. Identified GURAG exons were matched to these results to determine candidate exons that show high levels of inclusion. These were subsequently followed up through evaluation of junction counts between these exons and both upstream and downstream exons. Within RNA-seq data, and linear regression analysis between both upstream and downstream exons in which the GURAG exon would be skipped. Limited evidence for recursive splicing was considered a double-significance in linear regression analysis, but junction counts indicating that the skipped product dominates.

Second, we asked whether these GURAG exons made regular contact with upstream exons with which they are not expected to junction (based on known gene isoforms). This could indicate that the junction is used, but the GURAG exon is not included, leading to absence of isoform annotation. To identify known or novel junctions between the 99 GURAG alternative exons and upstream exons, we evaluated all junctions in RNA-seq data that were made between the identified 99 candidate exons and any upstream exon (Supplementary Table 1, worksheet 5). Each junction was then enumerated and classified as ‘known’ or ‘novel’ using the known-gene UCSC annotations. If a junction was not present in this annotation database and subsequently classed as novel, this was considered limited evidence for recursive splicing. Examples were subsequently considered
high confidence if splicing patterns inferred from the aforementioned analysis of total RNA-seq read density patterns suggested frequent use of the novel junction. Combined, these analyses identified 16 putative annotated RS-exons, two of which (in the CADM2 and NTM genes) we further experimentally validated.

Cryptic element usage analysis. Introns contain numerous cryptic splicing motifs that do not produce products indicated by current gene annotations, yet might infrequently be used to create low-level alternative isoforms (Supplementary Note). To evaluate the frequency of cryptic element usage in introns of differing lengths (Extended Data Fig. 9c), we performed a search for novel junctions in our RNA-seq data that connected first intron loci to canonical second exons. Second exons were chosen because we observed that 8 out of 9 of our high-confidence sites were located in long first introns. To perform this analysis while limiting duplication of the same exon owing to multiple transcripts, RefSet annotations were refined to include only those transcripts defined as canonical by UCSC known-gene table. Intersection of both annotation databases identified 21,331 second exons common to both databases. Of these, 798 were subsequently removed owing to a lack of evidence of gene expression across all brain regions based on gene-derived RNA-seq FPKM values. For the remaining 20,733 exons, upstream intronic regions were searched for all junctions connecting these exons to any upstream elements (Supplementary Table 3, worksheet 2). Junctions were classified according to the nature of the upstream elements. Specifically we separated into three categories: ‘exon–exon’ represented junctions between the canonical first exon and second exon, ‘isomeric’ represented junctions between an alternative first exon and the second exon that are present in UCSC/RefSeq/Gencode databases, and ‘novel’ represented entirely unexpected junctions between intronic elements in the UCSC/RefSeq/Gencode databases that junction to the second exon. We restricted our final analysis of cryptic upstream elements to the 6,619 genes in which a canonical exon–exon junction was detected which according to the full-length of the canonical first intron. The number of novel junctions to cryptic upstream elements was then counted in these genes, with genes grouped in bins based on the length of the canonical first intron. To avoid overlap with non-canonical minor transcripts, ‘isomeric’ junctions were not considered. Significance between bins was determined using the Mann–Whitney U test with two tails.

To evaluate cryptic element usage to all 142 candidate RS-sites (high-confidence targets, all cassette exons starting with GURAG, and novel junctions detected that were consistent with RS-sites but failed to meet significance in linear regression analysis), the upstream gene body of candidate RS-sites genes were searched for all junctions present within brain RNA-seq libraries that connected these candidate RS-sites to any upstream elements (Supplementary Table 3, worksheet 3). Junctions were then classified according to the nature of the upstream elements. Specifically we asked whether the junction was to an annotated upstream exon or cryptic exon/promoter.

Gene expression comparisons. For tissue-specific gene expression comparisons in Extended Data Fig. 1, RNA-seq data from 16 human tissues obtained by the Illumina Human Body Map Project (GEO series accession number GSE0611) and RNA-seq data from 12 human tissues collected as part of the Genotype Expression (GTEX) Project (http://www.gtexportal.org) were mapped to hg19 genomes (hg19). For the cell line comparisons mapped in the same way to either hg19 or mm9 data, data were collected from the following sources: myoblast differentiation (mm9, GEO series accession number GSE20846), erythropoiesis (hg19, GEO series accession number GSE40243) and motor neuron differentiation (mm9, GEO series accession number GSE20846), erythropoiesis (mm9, GEO series accession number GSE20846). Expression values across replicates was calculated using DESeq2. Tissue-specific comparisons were made based around the brain and all other individual tissues for all protein coding genes. For cell-specific comparisons, differentiated cells were plotted against undifferentiated cells in respective data sets. The log2-fold expression changes were plotted against first intron gene length. In instances in which several gene lengths were reported for a given gene, the maximum gene length was used.

Cross species intron lengths. To determine cross-species intron lengths, all human RefSeq genes were mapped to indicated species using the xenoRefGene track. Corresponding intron lengths were determined using exon start and exons end coordinates from all single-mapping transcripts. Identical introns found across multiple transcripts of the same gene were collapsed into a single unique intron for analysis so not to be counted multiple times.

GO term analysis. The GO term associated with >150 kb human UCSC genes analysed by GOrilla27 using two unranked lists of genes. UCSC genes >150 kb were used as targets, while all UCSC genes were used as background. For visualization, GO terms with >1 × 10−5 FDR q-value or less than twofold enrichment were omitted.

Motif analysis. Sequence analysis around novel junction intronic loci was performed using WebLogo28. Recursive exon maps were generated by string matching consensus 5′ splice sites and stop codons to regions following RS-sites after considering open reading frame of upstream RefSeq exons. Strong consensus splice sites were considered GTAG, GTGG, GTAGG and GTATG (Fig. 2b and Extended Data Fig. 5). Weak consensus splice sites are GTAAA, GTATA, GTGGG, GTAAC, GTGAC and GTACG (Extended Data Fig. 5).

Splice site score calculation. MaxEntScan was used as previously described using the First-order Markov Model setting by adding the last three nucleotides of the exon and the first 6 nucleotides of the 5′ splice site21. Competing splice site scores are presented in Supplementary Table 3, worksheet 1, and Extended Data Fig. 8d.

Conservation scores. For conservation scores, the 46-way placental mammal conservation by PhastCons track on the UCSC genome browser was used (phastCons46wayPlacental). Conservation scores were obtained for a given region using table browser, and mean scores calculated after alignment to specified features. Conservation was calculated at RS-sites (n = 9), at 5′ splice sites downstream of RS-exons (n = 9), at 5′ and 3′ splice sites spanning constitutive exons in genes containing RS-sites (n = 130), and at the next two nearest 5′ splice sites downstream of RS-exons (n = 18).

Cell culture. SH-SY5Y cells (ATTC, CRL-2266) were cultured at 37°C, 5% CO2 in DMEM (Life Technologies) supplemented with 10% FBS, and routinely tested for mycoplasma contamination. For all treatments in this cell line, cells were seeded to be 70–80% confluent at the day of transfection in 6-well plates.

For AON treatment, cells were transfected at 24 h with 10 μM of stated AON using Endo-porter transfection reagent (Gene-tools) as per manufacturer’s instructions. At 48 h after transfection, cell media was removed and cells lysed and RNA extracted with Qiazo. All AONs were purchased from Gene-tools, and carried morpholino modifications. Sequences used to target the listed genes were: CADM1: AON-A1: 5′-AGGACAGTTGGAATGATGCCTAC-3′; CADM1: AON-A2: 5′-ATCCAAAGTAAAGTTGTAATGCCTAC-3′; ANK3: AON-A1: 5′-TTTAAATGGAACACCGATATGCCTAC-3′; ANK3: AON-A2: 5′-AATTGGGCACTGCCTATTGCCTAC-3′.

A non-specific AON-NS that is not complementary to any locus in human genome, but has similar GC content as the AON-A1 and AON-A2, was used as control: AON-NS: 5′-CCTCTTCACCTCAGTCAATTTATA-3′.

For cyclolheximide treatment after AON-A2 transfection, cells were seeded to be 50–70% confluent at the day of transfection and were treated at 48 h (first experiment) or 36 h (second experiment) with either 100 μg ml−1 of cyclolheximide dissolved in DMSO, or an equivalent volume of DMSO alone. At 6 h after treatment, cell media was removed and cells lysed and RNA extracted using Qiazo (Qiagen).

Zebrafish AON treatments. Zebrafish experiments were performed by injecting 1 pg of AON (Gene-tools) into the yolk of 1-cell-stage embryos. Embryos were grown at 28.5 °C and collected at 2 days post-fertilization for RNA extraction. AON NS: 5′-CCTCTTCACCTCAGTCAATTTATA-3′; AON-A1: 5′-GTGGAAAAAATACCCCAAGACTCAC-3′; AON-A2: 5′-AATGGTTCTACGCTAC-3′.

Splicing reporter design. The CADM2 splicing reporter (P1) was designed such that the RS-exon following the second CADM2 RS-site was flanked by two short introns and the surrounding CADM2 constitutive exons (Supplementary Table 4). Introns consisted of the first and last ~100 nucleotides of respective introns separated by multiple cloning sites. Constitutive exons were flanked by HindIII and EcoRI sites, respectively. Constructs were sub-cloned into the pcDNA3 multiple cloning site of the pBluescript plasmid using HindIII and EcoRI sites. Construct P2 was subsequently generated by removing the dominant first CADM2 exon and first ~100 nucleotides of intron present in construct P1 with HindIII and FseI, and subcloning a separate synthetic gene product into the digested plasmid. This synthetic gene product consisted of the alternative first exon and first ~100 nucleotides of the corresponding intron sequence. Synthetic gene products can be found in Supplementary Table 4. Mutations to both reporter variants were made by crossover PCR using construct P1 or P2 as targets and primers listed in Supplementary Table 4.

Cell fractionation. For nuclear-cytoplasmic fractionation of cell lines, samples were suspended in 1 ml cytoplasmic lysis buffer (50 mM Tris-HCl pH 7.4, 10 mM NaCl, 0.5% NP-40, 0.25% Triton X-100, 1 mM EDTA, 1/200 volume of RNAsin and 1/100 volume of protease inhibitor cocktail) and homogenized by pipetting. Sample was spun for 3 min at 3,000g. Supernatant was collected as the cytoplasmic fraction and subjected to a further spin at 10,000g for 10 min. Supernatant was removed and RNA extracted using Trizol LS (Life technologies) and the Zymogen RNAdirect extraction kit as per manufacturer’s instructions. The pellet from the initial spin was retained as the nuclear fraction and lysed using Qiazo before RNA was extracted using the Zymogen RNAdirect extraction kit as per manufacturer’s instructions.

RNA extraction. For cell culture experiments Qiazo (Qiagen) suspended RNA was extracted using the Zymogen RNAdirect extraction kit as per manufacturer’s instructions. For brain total RNA extraction and zebrafish tissue total RNA extrac-
tion, tissue was first suspended in Qiazol (Qiagen) and homogenized using a TissueRuptor (Qiagen). RNA was then extracted using the Zymogen RNAdirect extraction kit as per manufacturer’s instructions.

**RT–PCR analysis.** All RNA was reverse transcribed using the high capacity cDNA synthesis kit (Applied Biosystems) using random primers and standard protocol. A total of 1 μg was used in each reaction and cDNA then diluted according to downstream application. For RT–PCR, samples were diluted 1:5 and 1 μl used for each subsequent PCR reaction. For qPCR samples were diluted 1:10 and 5 μl used for each subsequent PCR reaction.

For RT–PCR analysis, 10 ng cDNA was amplified using 2X Phusion PCR mastermix (Thermo-scientific) as per manufacturer’s instructions and each primer at a final concentration of 0.5 μM. Products were run on pre-cast 6% TBE gels (Life Technologies) using low molecular mass marker (New England Biolabs) or Hyperladder V (Bioline) as a ladder. Where exon inclusion was determined from RT–PCR images, band intensity of expected product sizes were determined using ImageJ software and expressed as a percentage of total intensity for all expected bands with indicated primers.

For Qiaxcel analysis cDNA was amplified with 2X Phusion PCR mastermix (Thermo-scientific) as per manufacturer’s instructions and each primer at a final concentration of 0.5 μM. Samples were subsequently purified using QIAquick PCR Purification Kit and loaded onto a Qiaxcel DNA cartridge (Qiagen) and run next to a 50–800-bp DNA marker (Qiagen) on the Qiaxcel machine (Qiagen) as per manufacturer’s instructions.

For qPCR analysis, 25 ng of cDNA was amplified using SYBR green PCR mastermix (Applied Biosystems) and each primer at a final concentration of 0.165 μM. PCR was carried out using an Applied Biosystems 7900HT machine as per manufacturer’s instructions and quantification assessed according to standard curves generated for each primer. Signal for each interrogated junction in qPCR analysis of human genes is normalized to GAPDH and/or EIF4A2 gene expression, and in zebrafish to actb1 and eif4a gene expression.

Primer sequences used for RT–PCR analysis and expected product sizes can be found in Supplementary Table 4.

No statistical methods were used to predetermine sample size, and experiments were not randomized.

---

20. Trabzuni, D. et al. Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies. J. Neurochem. 119, 275–28 (2011).
21. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).
22. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760–1774 (2012).
23. König, J. et al. iCLIP reveals the function of hnrNP particles in splicing at individual nucleotide resolution. Nature Struct. Mol. Biol. 17, 909–915 (2010).
24. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25 (2009).
25. Singh, J. & Padgett, R. A. Rates of in situ transcription and splicing in large human genes. Nature Struct. Mol. Biol. 16, 1128–1133 (2009).
26. Kim, D. et al. TopHat2: accurate alignment of transciptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36 (2013).
27. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10, 48 (2009).
28. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190 (2004).
29. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnol. 28, 511–515 (2010).
30. Herrera, F. J., Yamaguchi, T., Roelink, H. & Tjian, R. Core promoter factor TAF9B regulates neuronal gene expression. eLife 3, e02559 (2014).
31. Madzo, J. et al. Hydroxymethylation at gene regulatory regions directs stem/early progenitor cell commitment during erythropoiesis. Cell Rep. 6, 231–244 (2014).
Extended Data Figure 1 | Long gene expression is enriched in the brain. a, GO term analysis of genes >150 kb relative to all human genes. All GO terms are associated with enrichment scores >2. b, The log2-fold gene expression ratios following DESeq analysis of all human protein-coding genes between the brain and all other tissues. Data are represented as Loess smoothing curves after the genes by their maximum length in kilobases. Hashed vertical line indicates 150 kb gene length. RNA-seq data was obtained from the GTEx consortium. c, Individual scatterplots used to create Fig. 1b and representing DESeq analysis of individual genes within indicated tissues compared to the brain. Red dots indicate genes that contain RS-sites, blue dots indicate dystrophin, and black dots indicate titin (two long genes most highly expressed in muscle tissues). Grey dots are all remaining genes. d, DESeq analysis of individual gene expression after vs before differentiation of C2C12 mouse myoblasts (GSM521256) into myogenic lineage (GSM521259), after vs before differentiation of mouse embryonic stem cells (GSM1346027) into motor neurons (GSM1346035), or after vs before differentiation of haematopoietic stem cells (GSM992931) into erythroid lineage (GSM992934). Loess smoothing curves are shown after sorting the genes by their maximum length in kilobases. Hashed vertical line indicates 150 kb gene length.
a.

![Graph showing nucleotide frequencies](image_a)

b.

![Graph showing intra-gene variation in gradients of >50kb introns](image_b)

c.

Pooled RNA-Sequencing
4 brains, 12 regions
(1.58 x 10^9 paired-end reads)

Restrict to genes with end of exon
150kb+ introns

Regression analysis
before/after junction consideration

10nt overlap at start of exon
1649 reads

10nt overlap at end of exon
1332 junctions

Second part of junction maps to preceding intron
>10nt

Second part of junction maps to preceding intron
>10nt

10nt overlap at end of exon
10nt overlap at start of exon

Restrict to genes with 150kb+ introns

Regression analysis before/after junction consideration

10nt overlap at start of exon
1649 reads

10nt overlap at end of exon
1332 junctions

Second part of junction maps to preceding intron
>10nt

Second part of junction maps to preceding intron
>10nt

d.

| Motif   | Frequency |
|---------|-----------|
| GTAAG   | 26.20%    |
| GTGAG   | 24.64%    |
| GTAGG   | 5.67%     |
| GTATG   | 5.47%     |
| GTAAA   | 4.55%     |
| GTAAT   | 3.47%     |
| GTGGG   | 3.19%     |
| GTAAC   | 2.25%     |
| GTGAG   | 1.86%     |
| GTAGC   | 1.52%     |
| GTACA   | 1.30%     |
| GTATT   | 1.30%     |
| GTACT   | 1.29%     |
| GTGTG   | 1.16%     |
| GTGCG   | 1.14%     |
| GTACC   | 1.02%     |

e.

![Graph showing nucleotide frequencies](image_e)
Extended Data Figure 2 | Linear regression analysis and novel junction sequence considerations used to identify mammalian recursive splice sites.

a, Examples of RNA-seq read density patterns for three genes together with their calculated gradients across the (1) first intron >50 kb, and (2) the average across all other >50-kb long introns within the same gene. Gradients represent the change in summated read count every 5 kb since RNA-seq reads are grouped in 5-kb windows and linear regression performed on resulting histograms. b, Density plot indicating the ratio of gradients of all other >50 kb introns within the same gene: the gradient of the first intron >50 kb. Blue hashed line represents ratio of 1. This would indicate that gradients for long introns within the same gene are comparable and transcription is proceeding at a largely constant rate. c, Schematic of the bioinformatics pipeline used to identify novel junctions. d, Ranking of human 5’ splice site pentamer usage genome-wide. e, Nucleotide usage frequency at human 3’ splice sites genome-wide, and branch-point positioning relative to 3’ splice site genome-wide.
Extended Data Figure 3 | Inferred splicing patterns identify recursive splice sites within mammalian >150 kb intron genes. a–g, RNA-seq (red) read density patterns and normalized FUS iCLIP (green) cross-link density patterns for the OPCML (a), ROBO2 (b), HSST3 (c), ANK3 (d), CADM2 (e), NCAM1 (f) and PDE4D (g) genes within human brains. RNA-seq reads and normalized FUS iCLIP cross-links are grouped in 5-kb windows. RefSeq introns >150 kb were searched for novel junctions and linear regression performed on all Ensembl introns >50 kb in which novel junctions were located. Gene isoforms displayed are those including introns within which significant junctions were identified. Red novel junctions represent significant improvements in goodness-of-fit in both RNA-seq and FUS regression analysis ($P < 0.01$ in both data sets, $F$-test). Blue novel junctions contact RS-exons. Grey novel junctions were not deemed significant following regression analysis. Zoomed area represents sequence at deep intronic loci surrounding novel junction. Phylo-P conservation track indicates sequence conservation across 46 levels of mammalian evolution.
Extended Data Figure 4 | Inferred recursive splicing patterns in the **OPCML** gene across four separate brains. a, RNA-seq read density patterns for the **OPCML** gene across 12 different regions of four separate brains. Gene isoform displayed is that which included the long first intron within which a significant novel junction was identified. RNA-seq reads are grouped in 5-kb windows. Dotted arrows indicate location of experimentally derived RS-site.
Extended Data Figure 5 | RT–PCR confirmation of RS-sites in human and zebrafish samples, and prediction of mouse RS-exons. **a**, Schematic of primer design used for RT–PCR validation of novel junctions. **b–g**, RT–PCR analysis of CADM2 (**b**), HS6ST3 (**c**), ROBO2 (**d**), PDE4D_1_1 (**e**), PDE4D_1_2 (**f**) and PDE4D_2_2 (**g**) genes around RS-sites using indicated primers. For PDE4D sites, first number after gene name indicates RS-site studied, second number indicates the upstream exon used. See Extended Data Fig. 3g for junctions detected. **h**, RT–PCR analysis of cadm2a RS-site junction in adult male and female zebrafish embryos, together with an alignment of zebrafish (ZF) cadm2a RS-site to human (HS) CADM2 RS-site. **i**, Map of consensus splice site location and in-frame termination codons following RS-sites in indicated mouse genes. Strong consensus splice sites are GTAAG, GTGAG, GTAGG and GTATG. Weak consensus splice sites are GTAAA, GTAAT, GTGGG, GTAAC, GTCAG and GTACG.
Extended Data Figure 6 | Conservation of inferred recursive splicing patterns in the mouse brain. a–h, Normalized FUS iCLIP read density patterns for the Opcml (a), Robo2 (b), Hs6st3 (c), Ank3 (d), Cadm1 (e), Ncam1 (f), Cadm2 (g) and Pde4d (h) genes within the mouse brain. Normalized FUS iCLIP cross-link sites are grouped in 5-kb windows, and the displayed linear regression lines were computed on resulting histograms. Zoomed area at deep intronic loci represents RS-site sequences conserved from humans to mouse.
a. | Exon Class         | Total | Exons starting | GURAG | Percentage |
|---------------------|-------|----------------|-------|------------|
| Constitutive Exons  | 146565| 540            | 0.37  |
| Alternative Exons   | 29289 | 99             | 0.34  |

b. | CADM2: |
|-----|
| RS-site: 1 |
| RS-exon: |
| Expected: 253 |
| Products from dominant isoform promoter P1: |
| 150nt |
| Products from minor isoform promoter P2: |
| 133nt |
| 118nt |
| 103nt |

| Expected mRNA products: |
|--------------------------|
| 1st RS-exon - containing a premature stop codon |
| 2nd RS-exon |
| Other exons |

| Reverse primer located after RS-site 2, Fig 4a |
|-----------------------------------------------|
| C = 253 |
| G = 133 |
| A = 103 |

| Products from dominant isoform promoter P1: |
| 80nt |
| Products from minor isoform promoter P2: |
| 208nt |
| 2nd RS-exon |
| Other exons |

| Forward primer: |
|-----------------|
| Dominant promoter P1 |
| Minor promoter P2 |

| Forward primer from 1st RS-exon: |
| 125nt |
| 98nt |

| PCR cycles: |
|-------------|
| 26 |
| 36 |

| Expected pre-mRNA product: |
|----------------------------|
| 81nt |

| PCR cycles: |
|-------------|
| 35 |

| NTM gene: |
|-----------|
| 200 kb |

| RNA-seq reads/5kb |
|--------------------|
| Annotated RS-exon |
| Dominant promoter P1 |
| Minor promoter P2 |

| Cryptic minor promoter P2: |
|-----------------------------|
| RS-site |
| Annotated RS-exon |

| Products from dominant isoform promoter P1: |
| 208nt |
| Products from minor isoform promoter P2: |
| 2nd RS-exon |
| Other exons |

| PCR cycles: |
|-------------|
| 123 |
| 150 |
| 100 |
| 75 |

| Expected: |
| 253 |
| 133 |
| 103 |

| Forward primer from 1st RS-exon: |
| 80nt |
| 70nt |

| PCR cycles: |
|-------------|
| 26 |
| 36 |

©2015 Macmillan Publishers Limited. All rights reserved
Extended Data Figure 7 | Promoter-dependent inclusion of RS-exons in CADM2 and NTM genes. a, Number of cassette and constitutive exons starting with motif GURAG. b–d, RT–PCR of CADM2 gene in the frontal cortex using primers indicated in b or Fig. 4a. RT–PCR was carried out on one (b) or four (c, d) human brains. In c, the inclusion of the second RS-exon occurs together with the minor promoter. Two bands are present for both PCR reactions due to the presence of an alternatively spliced exon following the RS-exon. This can result in two distinct long or short isoforms. In d, the inclusion of the second RS-exon occurs when the first RS-exon is included. Schematics in c and d represent examined splicing products together with expected length of products. e, RNA-seq read density patterns for the NTM gene and expected human isoforms. RNA-seq reads are grouped in 5-kb windows and linear regression performed on resulting histograms. A cryptic minor promoter/exon detected by RNA-seq is indicated by vertical red line. The annotated RS-exon is indicated by the vertical blue line. Zoomed area represents RS-site sequence at start of the annotated RS-exon. Primers to assess the major and minor promoter products associated with the RS-exon are indicated by coloured arrows. f, RT–PCR of NTM gene around RS-exon using indicated primers. g, RT–PCR analysis of NTM products in which the upstream exon is either derived from the major upstream promoter or the cryptic upstream promoter/exon. RT–PCR was performed in the frontal cortex of three human brains using primer sets indicated by coloured arrows in e. Schematics represent possible splicing products together with expected length of products. Top panel assesses RS-exon inclusion, bottom panel assesses RS-site junction detection.
Extended Data Figure 8 | Recursive splicing regulates the alternative splicing of RS-exons. a, Qiaxcel analysis and quantification of the splicing intermediates of indicated CADM2 splicing reporter products following transfection in SH-SY5Y cells. Primers used are indicated by red arrows in schematic, together with expected products and their sizes. b, RT–PCR analysis of the zebrafish cadm2a mRNA after in vivo injection of AON-2. Sequencing reveals RS-exon inclusion results in subsequent splicing to additional downstream cryptic elements before the second exon, explaining why RS-exon included product size is larger than expected. c, qRT–PCR analysis of exon–exon junctions surrounding the RS-site containing introns following AON-A1 mediated inhibition of RS-site use of the human CADM1 and ANK3 genes (n = 3, 1 experiment) or the zebrafish cadm2a gene (n = 7, 3 separate experiments). d, Splice site scores of reconstituted 5’ splice sites following first step of recursive splicing versus the 5’ splice sites of corresponding recursive exons.
Extended Data Figure 9 | Cryptic elements are frequent in long first introns.

a, UCSC annotated isoforms of the OPCML gene together with spliced expressed sequence tags (ESTs) detected across the OPCML locus. Recursive exon is marked in blue, and the preceding exons produced by minor promoter or cryptic splicing of the long first intron are marked in red. b, Lengths of the 9 introns containing the high-confidence RS-sites compared to other introns across vertebrates. Results are an extension of Fig. 4g. c, Boxplot showing the detected number of unannotated alternative start exons that junction to the dominant second exon of brain expressed genes. Only novel junctions that do not match UCSC/GENCODE transcripts are considered for analysis. Genes are separated into bins based on the first intron length of the canonical isoform. Boxplot presents median, first and third quartile boundaries for each bin. *$P < 10^{-10}$ (Mann–Whitney U test). Only tests between the 100 kb+ bin to other bins are shown. Right panel shows cartoon of the implications of boxplot results.