SUPPLEMENTARY DATA

Splicing of many human genes involves sites embedded within introns

Steven Kelly[1,†], Theodore Georgomanolis[2,†], Anne Zirkel[2,†], Sarah Diermeier[3,§], Dawn O'Reilly[4], Shona Murphy[4], Gernot Längst[3], Peter R. Cook[4], and Argyris Papantonis[2,*]

[1] Department of Plant Sciences, University of Oxford, Oxford, OX1 3RB, United Kingdom
[2] Centre for Molecular Medicine, University of Cologne, Cologne, D-50931, Germany
[2] Institut für Biochemie III, University of Regensburg, Regensburg, D-93053, Germany
[4] Sir William Dunn School of Pathology, University of Oxford, Oxford, OX1 3RE, United Kingdom

* To whom correspondence should be addressed. Tel: +49-221-478-96987; Fax: +49-221-478-4833; Email: argyris.papantonis@uni-koeln.de

[†]The authors wish it to be known that these authors contributed equally to this study.

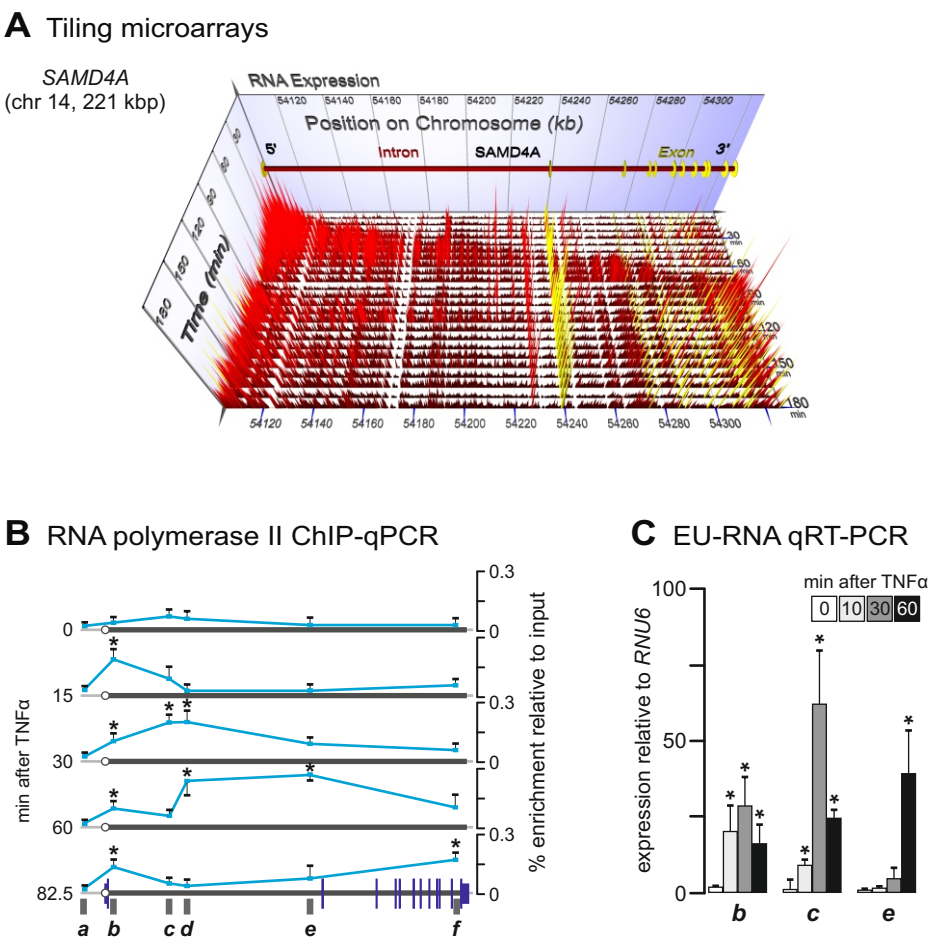
[§]Present Address: Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York, 11724, U.S.A.

Supplementary Data include:

– **Supplementary Figures S1-S7**

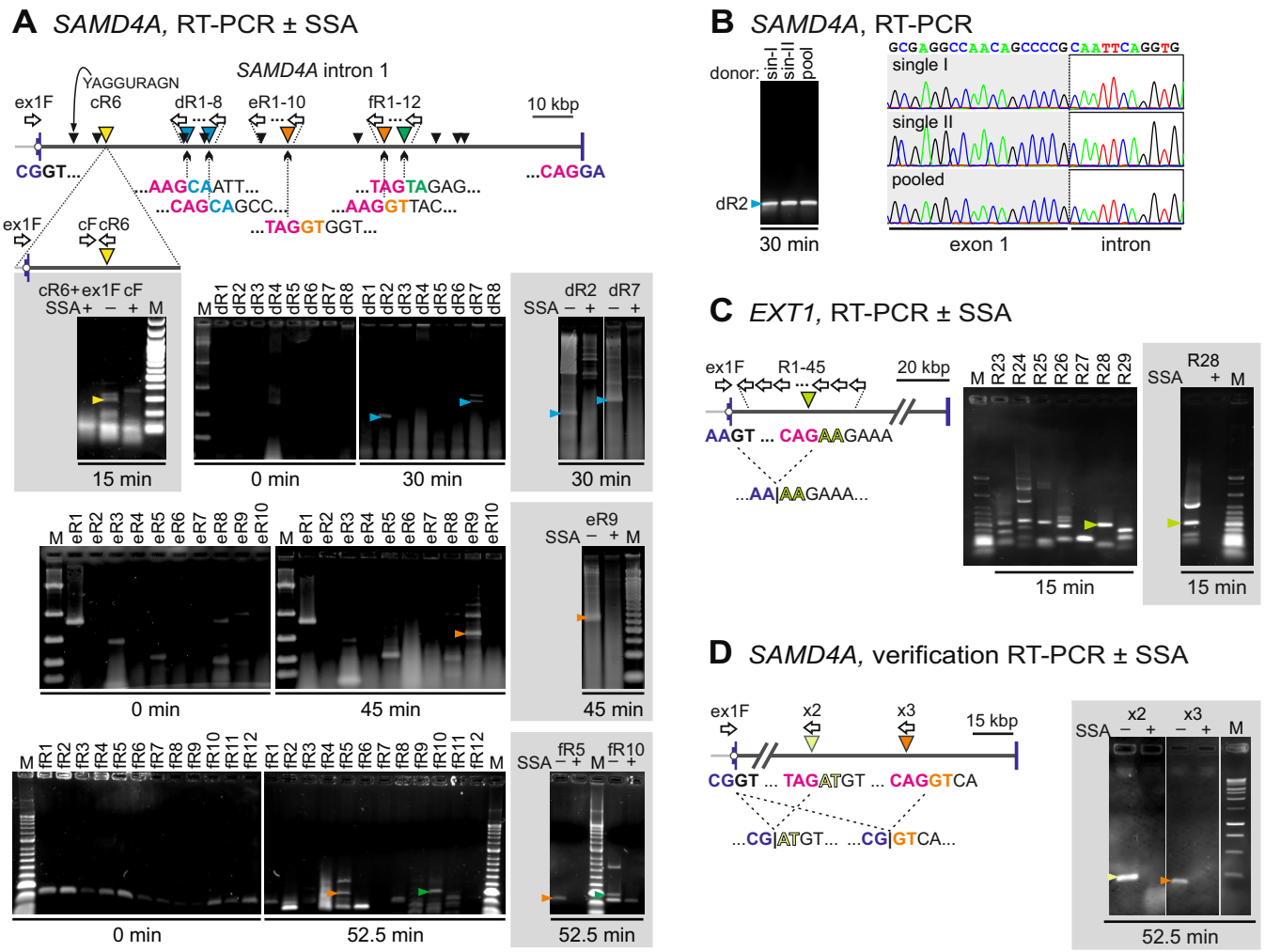
– **Supplementary Table S1**

Supplementary Figure S1



Supplementary Figure S1. Waves of nascent transcription in *SAMD4A* intron 1. HUVECs were treated with TNFα for 0-82.5 min. (**A**) A study, with 7.5-minute temporal resolution, of the "transcriptional wave" along TNFα-activated *SAMD4A* gene, using total RNA applied to a tiling microarray. Red signal peaks represent intronic (nascent), and yellow ones exonic normalized RNA levels (this panel is reproduced from ref. 7). (**B**) Binding of RNA polymerase II, determined by ChIP coupled to qPCR (using an antibody targeting phospho-Ser2 in the C-terminal domain of the largest subunit – and so the elongating form of the polymerase). The % enrichment (± SD; *n*=3) for segments *a*-*f* of *SAMD4A* (*bottom*) at different times post-stimulation is given relative to input. Results reflect those seen in panel A and in Figure 2A. *: significantly different from 0-min levels ($P<0.01$; two-tailed unpaired Student's *t*-test). (**C**) Nascent RNAs copied from segments *b*, *c*, and *e* in intron 1, detected by qRT-PCR. HUVECs were grown in 5-ethynyl-uridine (EU) for 5 min, biotin "clicked" on to the resulting EU-RNA, and now-biotinylated RNAs selected using streptavidin-coated magnetic beads; after DNase-treatment, specific regions of (nascent) EU-RNA (± SD; *n*=3) were detected by qRT-PCR, and levels normalized relative to those of *RNU6* EU-RNA. Again, results reflect those seen with microarrays (panel A) and ChIP-qPCR (panel B). *: significantly different from 0-min levels ($P<0.01$; two-tailed unpaired Student's *t*-test).

Supplementary Figure S2

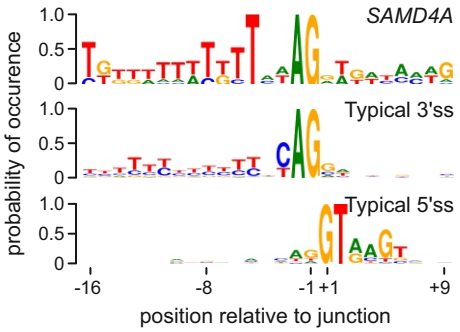


Supplementary Figure S2. Detection of recursive splicing in long human introns. HUVECs were treated with TNFα for 0-52.5 min before total RNA was isolated. (**A**) Intronic (nascent) RNA was detected by RT-PCR, amplimers resolved by electrophoresis, gels stained and imaged, and bands detected only after stimulation sequenced. The map (*top*) of intron 1 in *SAMD4A* (*blue* lines: exons 1/2) shows the primers used (*white arrows*). Forward primer "ex1F" targeting exon 1 is used successively with each reverse primer targeting indicated regions at 1-kbp intervals ("dR1-8", "eR1-10", and "fR1-12"). Coloured arrowheads mark RS sites, and hybrid sequences are indicated (exonic sequences – *blue*; donor GT at the 5′ end of intron 1 – *black*; splice junctions – *vertical lines*; RS acceptors – *red*; bases resulting from recursive splicing that would go on to be used subsequently as donors – *blue*, *orange*, or *green*; the RS site from Figure 2A is also indicated – *yellow*). Typical gel images are shown below (M: size markers), and arrowheads (coloured as above) indicate bands with hybrid exon-intron sequences. *Grey boxes*: pre-treatment with spliceostatin A (SSA) abolishes indicated bands. (**B**) The RS hybrid resulting from RT-PCR using "exF1+dR2" was verified in total RNA preparations from HUVECs derived from two single donors (*single I* and *II*), and from a cell pool different from that analyzed in Figure 2 (*pooled*). Sanger sequencing chromatographs of the three amplimers (*blue arrowhead*) are shown; they all encode the expected hybrid sequence (*top*). (**C**) Mapping (as in panel A) of a spliceostatin-sensitive hybrid from a site (*light-green* arrowhead) in the 273 kbp-long intron 1 of *EXT1*. The map (*left*) shows the 5′ end of the intron, and forward (exonic) primer ex1F is used successively with 45 reverse primers targeting points at ~1-kbp intervals along intron 1; images of typical gels are shown (*right*). *Grey box*: pre-treatment with spliceostatin A (SSA) abolishes the indicated band. (**D**) RNA-seq analysis uncovered seven RS sites in *SAMD4A* intron 1 (Figure 3C), 4 of which had not been seen using RT-PCR; therefore, we verified that two of these were hybrid products using RT-PCR (pairing ex1F with reverse primers x2 and x3). Images of typical gels are shown. *Yellow* and *orange* arrowheads mark (SSA-sensitive) amplimers encoding the expected hybrids (*left*). (**E**) Effect of serum starvation on recursive splicing. Total RNA from HUVECs in full growth medium (-), or from HUVECs serum-starved overnight (+) was isolated, DNase-treated, and used in RT-PCR to amplify the three most frequently detected hybrids in constitutively-expressed genes (according to RNA-seq data). The *HDLBP*, *UBR5*, and *KALRN* iS hybrids are detected under both growth conditions, as shown by gel electrophoresis. *Bottom*: parts of the sequences that are joined to form these exon-intron junctions are shown. (***F***) Effect of exosome inactivation on recursive splicing. A gene encoding a key subunit of the exosome, *RRP40*, was knocked down using siRNAs in

HUVECs. Total RNA was harvested from cells at the appropriate times post-stimulation, DNase-treated, and used in qRT-PCR. No significant effect was detected on the levels of two *SAMD4A* hybrids (dR2 and xR3), and one from *HDLBP*; mRNA levels of constitutively-expressed genes *CDC42* and *RCOR1* remained unaffected and serve as negative controls, while PROMPT levels from the *TMEM97* promoter rise as expected (21) and serve as a positive control. *: significantly different (P<0.01; two-tailed unpaired Student's t-test).

Supplementary Figure S3

A RS consensus



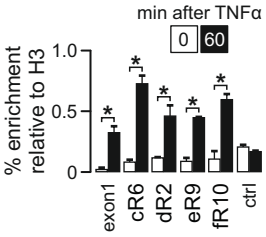
B *SAMD4A* RS properties

#	Position (kbp from TSS)	iS site (acc|donor)	Branch point (bp from junction)	pY tract (bp from BP)	5′ ss (ME score)	3′ ss (ME score)
1	14.2	AA**AG**|**ACT**T	TTA**A**T (-22)*	+4	-11.96	**5.87**
2	36.3	GA**AG**|**CA**AT	TTA**A**G (-22)	+2	-6.38	**4.58**
3	41.4	AC**AG**|**CA**GC	CTA**A**A (-22)*	+3	-7.48	**5.62**
4	60.2	CT**AG**|**GT**GG	GTG**A**C (-26)*	+6	**4.88**	**4.34**
5	80.7	CT**AG**|**AT**GT	GTA**A**A (-21)	+4	-0.6	**7.11**
6	84.6	AA**AG**|**GT**TA	CTC**A**C (-25)*	+8	**6.58**	**8.59**
7	94.5	TT**AG**|**TA**GA	GTA**A**C (-21)	+2	-6.11	**6.98**
8	100.9	GC**AG**|**GT**CA	CTG**A**G (-38)	+9	**9.28**	**8.17**
exon 1	0.19	CCCG|**GT**AA	–	–	**17.08**	–
exon 2	134.1	TC**AG**|GAAT	CTT**A**T (-27)	+14	–	**12.18**

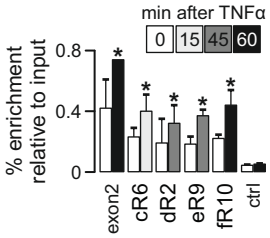
*verified by RT-PCR (see **Figure 3**)

C ChIP

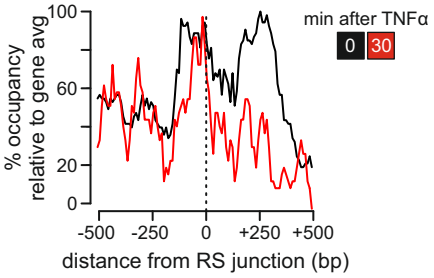
(i) H3K36me3



(ii) U2AF65



(ii) nucleosome occupancy

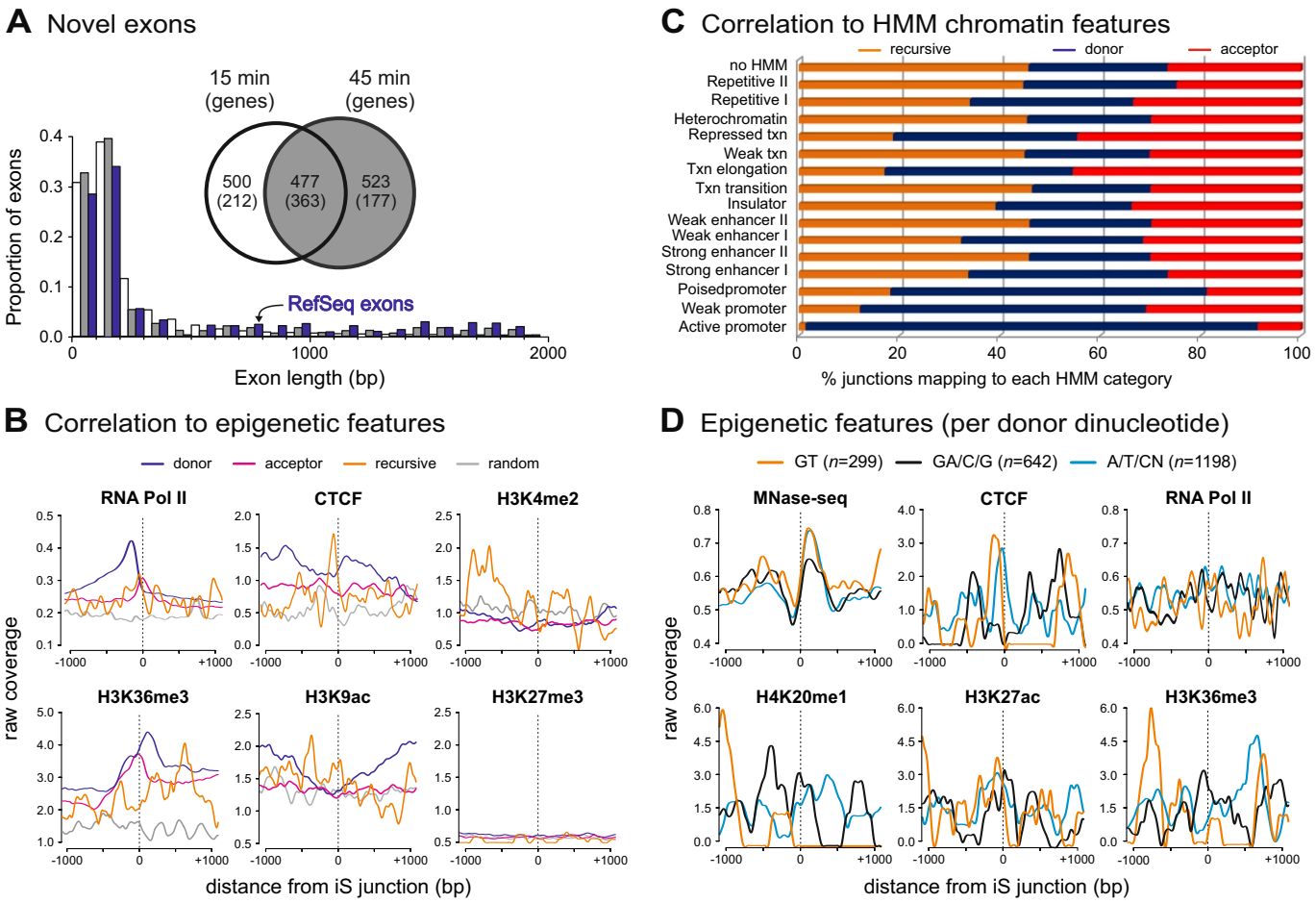


D *SAMD4A* RNA-seq overview

Feature mapped	Events	Reads
read-through	18	472
end-adjusted	*14*	*165*
exon-exon junctions	19	1871
inferred	*11*	*251*
recursive splicing	8	10
inferred	*5*	*9*

Supplementary Figure S3. Features of the *SAMD4A* recursive splicing sites. (**A**) The consensus motif (derived using WebLogo 3; ref. 23) for eight RS sites in *SAMD4A* intron 1 resembles that of canonical acceptors, but not donors (motifs at canonical sites from **Figure 4A**). (**B**) Predicted splicing potential of the eight *SAMD4A* sites compared to that of the canonical donor and acceptor sites at each end of intron 1 (*bottom*). Also indicated are the position relative to the TSS (in kbp), the hybrid sequence (*vertical line* separates the 3′ acceptor dinucleotide – in *red* – from the dinucleotide – in *bold black* – that could become a new donor in a subsequent RS event), the branch point (BP) identified using the yUnAy consensus (branch 'A' in *bold*), the start position (relative to the branch-point) of a poly-pyrimidine (pY) tract of 12 nucleotides, and maximum entropy (ME; ref. 31) scores (positive values in *bold*) calculated assuming each motif serves as a 5′ or a 3′ splice site. (**C**) Four RS sites in *SAMD4A* intron 1 (selected as each carries a different donor dinucleotide) carry H3K36me3 marks, and associate with U2AF65. HUVECs were treated with TNFα for 0-60 min, chromatin isolated, and ChIP performed using antibodies targeting (**i**) histone H3 carrying tri-methyl marks on lysine 36 (H3K36me3) or (**ii**) splicing factor U2AF65, and the primers indicated in the map (*left*). After stimulation, enrichments (± SD; n=3) are normalized relative to those obtained with H3 or "input" chromatin. Profiles are similar to those given by canonical sites at exon/intron boundaries, but unlike a control region from within the intron that is not spliced ("ctrl", encodes the *Drosophila* recursive-splicing motif). *: significantly different from 0-min levels (P<0.01; two-tailed unpaired Student's t-test). (**iii**) Mean nucleosome occupancy ±500 bp around the *SAMD4A* RS sites at 0 (*black line*) and 30 min (*red line*) after TNFα stimulation, as given by MNase-seq data (25). (**D**) Splicing events detected by RNA-seq in *SAMD4A*. "Read-through" refers to reads mapping across exon/intron boundaries; "end-adjusted" refers to read-through after removing reads mapping to the first and last *SAMD4A* exons (these skew results as the background of pre-existing mRNAs yield 5′- and 3′-UTR reads); "exon-exon junctions" refers to conventional splicing junctions seen in RefSeq genes; "inferred" refers to read pairs in which one maps solely to within an exon and the other to another exon (so neither read runs across the potential exon-exon junction); "recursive splicing" refers to hybrids 1-8 shown on the left, and "inferred" again to reads supporting these which flank, but do not contain, the actual junction sequence.

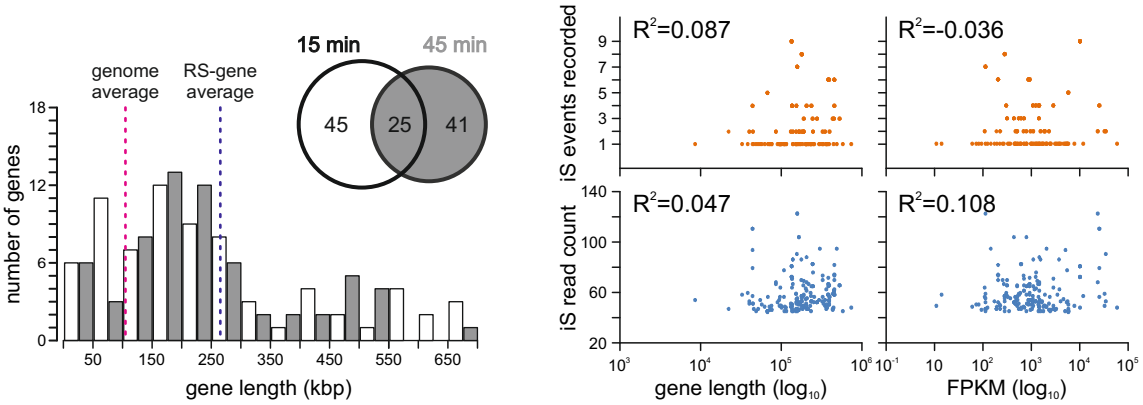
Supplementary Figure S4



Supplementary Figure S4. Novel exons and chromatin modifications associated with recursive splicing sites. (**A**) Novel exons. An exon was categorized as "novel" if (**i**) it is not found in any *Ensembl* gene model, and (**ii**) hybrid sequences linking its 5′ and/or 3′ ends to a known exon are also seen in our poly(A)$^+$-selected RNA-seq libraries. The Venn diagram shows the number of unique versus shared novel exons seen in the 15- (*black*) and 45-min (*grey*) libraries. The plot shows the size distributions of novel and RefSeq exons (*blue*). (**B**) Some epigenetic features of RS sites (*orange*) compared to canonical donors (*blue*) and acceptors (*red*); randomly-selected (but actively-transcribed) regions from the same introns provide controls (*grey dotted lines*). Data sources: ChIP-seq for histone marks and CTCF (ENCODE data, unstimulated HUVECs), and RNA polymerase II (HUVECs, 30 min post-TNFα). A "silent" mark, H3K27me3 – not enriched at any of our sites – provides another control. (**C**) Correlation of RS, donor, and acceptor sites (color-coded as above) to Hidden Markov Models (HMM) of chromatin features derived using an integrative annotation of ChIP-seq data on chromatin modifications generated by the ENCODE consortium on HUVECs (26). (**D**) Epigenetic features around RS sites categorized according to dinucleotides at positions +1/+2 (GT: *orange line*, GA/C/G: *black line*, or A/T/CN *blue line*). Data sources: as above. All three dinucleotides tend to be associated with a nucleosome positioned immediately 3′ of the site, but GT and A/T/CN groups are more similar to one another and distinct from GN ones (except for H4K20me1).

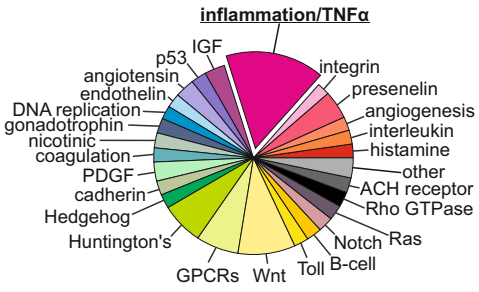
Supplementary Figure S5

A RS detection and correlation to gene features (top 100 events/library)

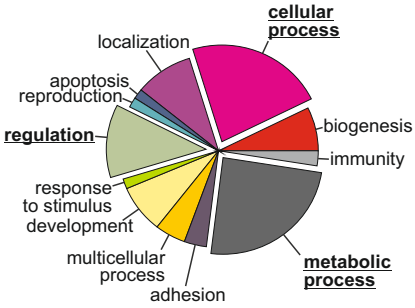


B GO term enrichment analysis (top 100 events/library)

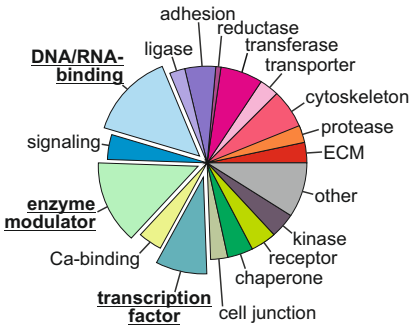
Signaling pathway:



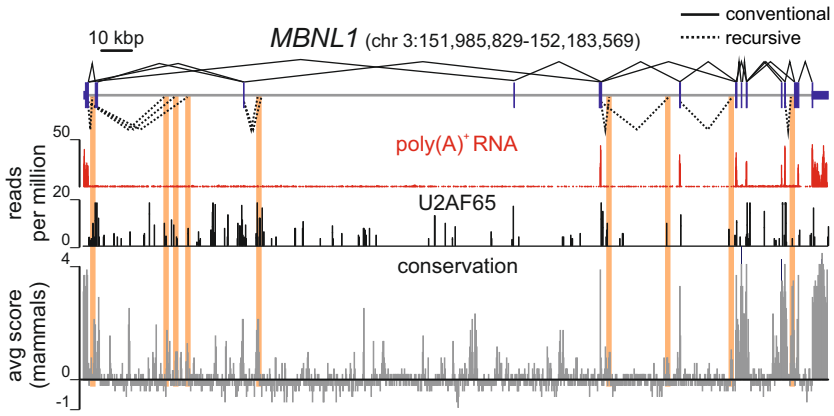
Biological process:



Protein function:



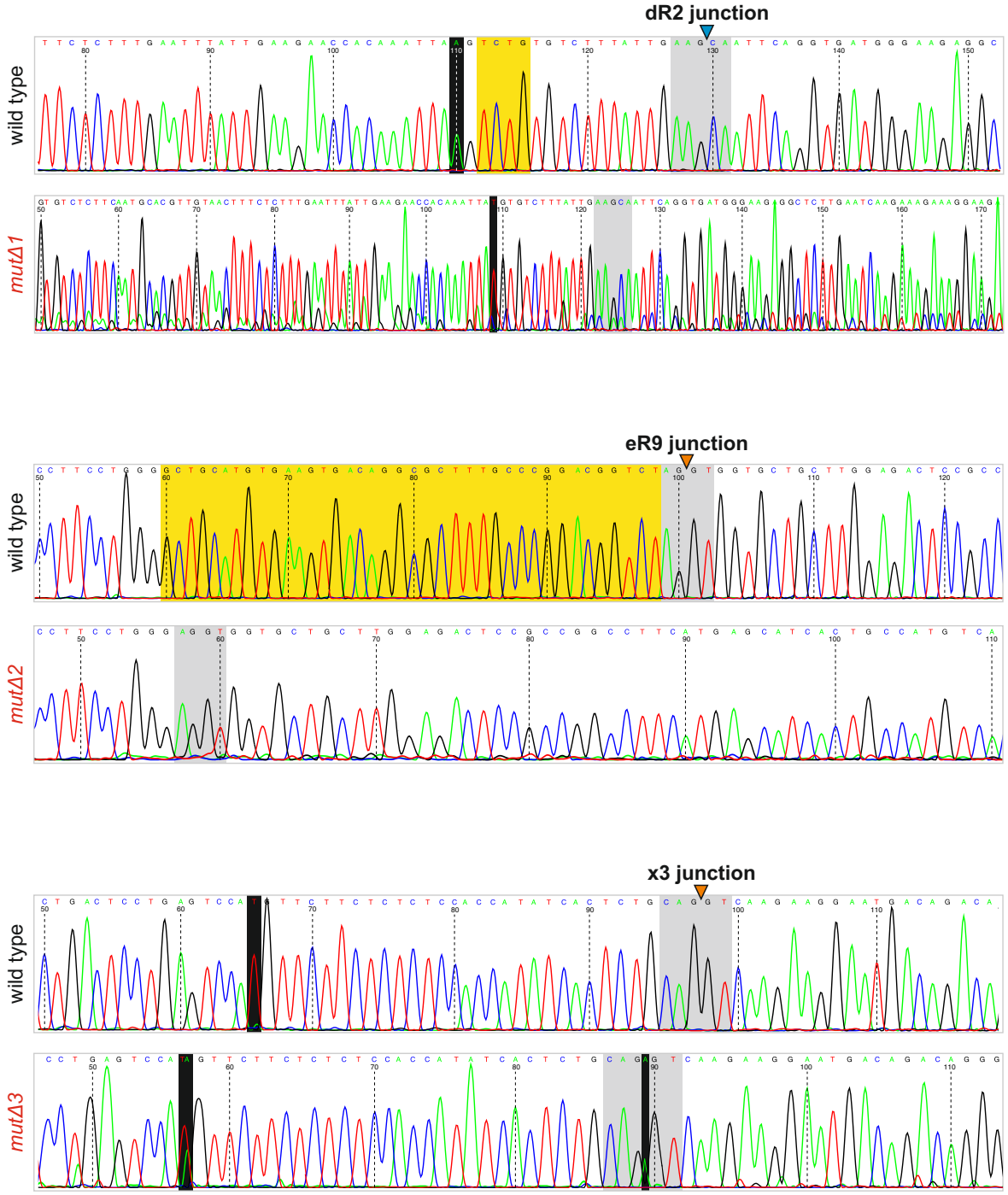
C Conservation at RS sites (genome browser view)



Supplementary Figure S5. Features and conservation of recursive splicing sites (from RNA-seq data). (**A**) Features of the top 100 most-frequently observed RS events. The Venn diagram shows the number of genes that host the top 100 RS events in the 15- (*white*) and 45-min (*grey*) library, respectively; 25 are shared. The histogram shows the distribution of genes in each library according to gene length (same colour code). Genes with hybrids (average ~244 kbp) are longer than the average gene (~63 kbp). The plots (*right*) show that neither the number of different iS hybrids seen per gene, nor their total number of read counts (Fragments Per Kilobase of transcript per Million mapped reads; FPKM), correlate well with increasing gene length (R^2<0.2; Pearson's correlation coefficient). (**B**) Gene Ontology (GO) terms for the genes hosting the top 100 RS events recorded by RNA-seq. The analysis was performed using PANTHER (http://www.pantherdb.org/) on a list of 111 genes (from panel A). Results for "signalling pathways", "biological process", and "protein function" are displayed as pie charts, and overrepresented categories are indicated (*bold, underlined*; must include >10% of dataset). (**C**) Conservation at recursive splicing (RS) sites. The gene model of *MBNL1* depicts conventional (*solid lines*) and recursive splicing (*dotted lines*) alongside tracks for poly(A)$^+$-selected RNA (60 min after TNFα; *red*), U2AF65 binding (HeLa, from ref. 27; *black*), and sequence conservation between placental mammals (from the UCSC browser; *grey*). Segments involved in RS (unfiltered; *highlighted orange*) tend to have intermediate levels of conservation (between those of intron and exons).

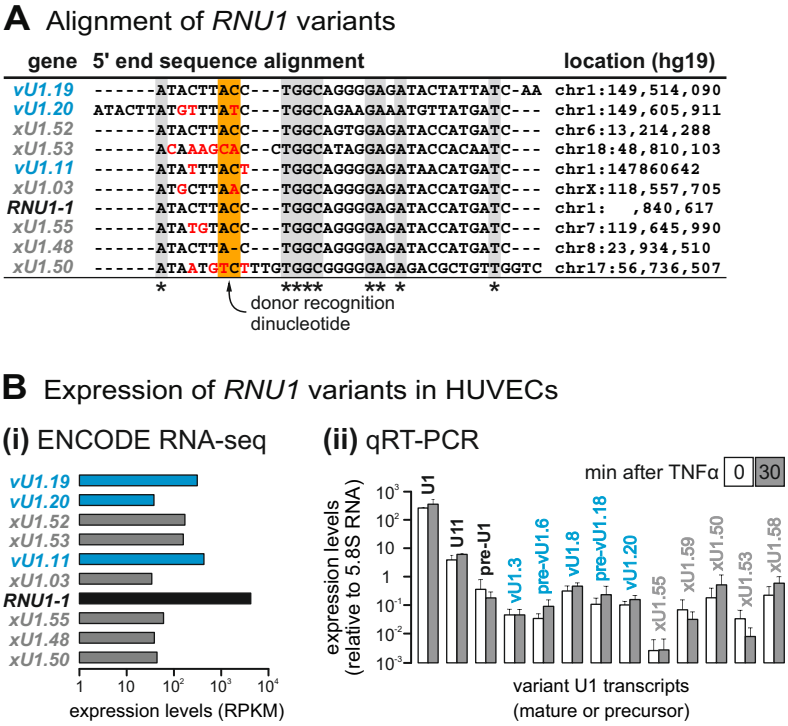
Supplementary Figure S6

Alignments of clone sequences (chromatographs)



Supplementary Figure S6. Identification of clone mutations in RS sites after CRISPR-Cas9n manipulations. The chromatographs resulting from Sanger sequencing of the three mutated HEK293A clones, *mutΔ1-3*, compared to the wild type clone, wt, which was transfected with vectors carrying no sgRNA sequences. In all three cases, RS junctions are denoted by an arrowhead and RS-acceptors/-donors highlighted grey, stretches wild-type sequences deleted in the mutants are highlighted yellow, and nucleotide positionss changed or inserted (as in the x3 junction) highlighted black.

Supplementary Figure S7



Supplementary Figure S7. Identification of robustly expressed *RNU1* variants in HUVECs. (**A**) Alignment of some *RNU1* variants that might be involved in successive RS events. The sequence of the *RNU1-1* gene on chromosome 1 (*black bold*) was used as a probe in a BLAT search against the human genome (http://genome.ucsc.edu/cgi-bin/hgBlat); a set of >60 similar sequences (xU1.01-58; *not shown*) were returned – most were unannotated. Nine are presented here (with 5′ ends aligned, conserved bases marked by *grey* shading and stars). Their sequences are similar to that of *RNU1-1* and carry "donor recognition" dinucleotides (*orange highlight*) that could pair with non-canonical (without a GT) or atypical (with a GT, but without other typical bases encompassing it) splicing donors. (**B**) Expression of *RNU1* variants in HUVECs stimulated with TNFα for 0 or 30 min. (**i**) A plot showing the expression levels of the nine most highly expressed *RNU1* variants (previously unidentified – *grey*; from ref. 39 – *blue*) alongside the typical *U1* transcript (*black*). Data, presented in "reads per million", from ENCODE data (http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeCshlLongRnaSeq). (**ii**) A plot showing expression levels of mature or precursor ("pre-") transcripts for variant U1 genes that produce unique amplicons in qRT-PCR. 0- (*white*) and 30-min (*grey*) levels are expressed relative to those of 5.8S RNA (±SD; *n*=3); colour-coding as in panel B,i.

Supplementary Table S1. A full catalogue of recursive splicing sites in HUVECs (after filtering). A .BED file containing the 2,389 unique recursive splicing sites identified in HUVECs at 15 and 45 min post-stimulation. The base indicated (by chromosomal and strand location; genome reference build: hg19) corresponds to the +2 position after the RS junction.