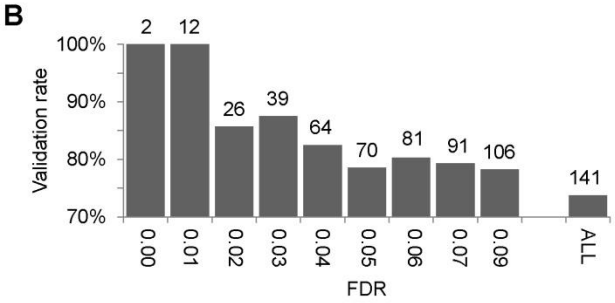
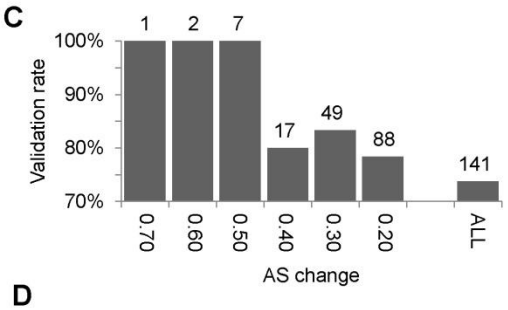
SUPPLEMENTAL FIGURES AND TABLES

A

	A=MCF10A-3D B=MCF10A-2D	A=MCF10A-3D B=HeLa	A=MCF10A-2D B=HeLa
A+B+	60	37	32
A+B-	300	323	284
A-B+	256	132	137
A-B-	60823	46237	46276
***P*-value**	5.5E-72	1.09E-42	1.05E-36





D

Number of AS events tested/ validated			All cell lines combined	MCF-10A 3-D	MCF-10A 2-D	HeLa 2-D
AS event classes	Cassette exon (CA)	Exon inclusion	58/80	30/46	19/22	9/12
		Exon skipping	46/61	26/35	9/12	11/14
	Alternative 3' splice site (AA)		9/9	7/7	2/2	-
	Alternative 5' splice site (AD)		4/4	3/3	1/1	-
	Intron retention (IR)		2/2	1/1	1/1	-
Total AS events			119/155	67/92	32/38	20/25
% total validation			**77%**	**73%**	**84%**	**80%**

Figure S1. Validation statistics of SRSF1-regulated CA events (Related to Figure 2). (A) Statistical significance of overlapping CA exons. The Fisher's exact test was used to determine whether the observed overlap between experiments is due to chance. Column names indicate two experiment instances designated A and B. The top four rows show the contingency tables where A+B+ indicates overlapping AS events, significant in both A and B; A+B- are events significant in A but not B, A-B+ are significant in B but not A and the marginal group A-B- are all the CA-exons detected although not significant in neither sample. The fifth row shows the Fisher's exact P-value **(B-C)** The validation rate for CA events (skipping and inclusion) in MCF10A 3-D and 2-D cells together with HeLa cells was computed using **(B)** different false discovery rates (FDR) or **(C)** different absolute AS change as thresholds. The column labeled "ALL" indicates the validation rate for all tested targets together. The number of AS events in each category is indicated. (**D**) Number of SRSF1-regulated AS events validated by RT-PCR.

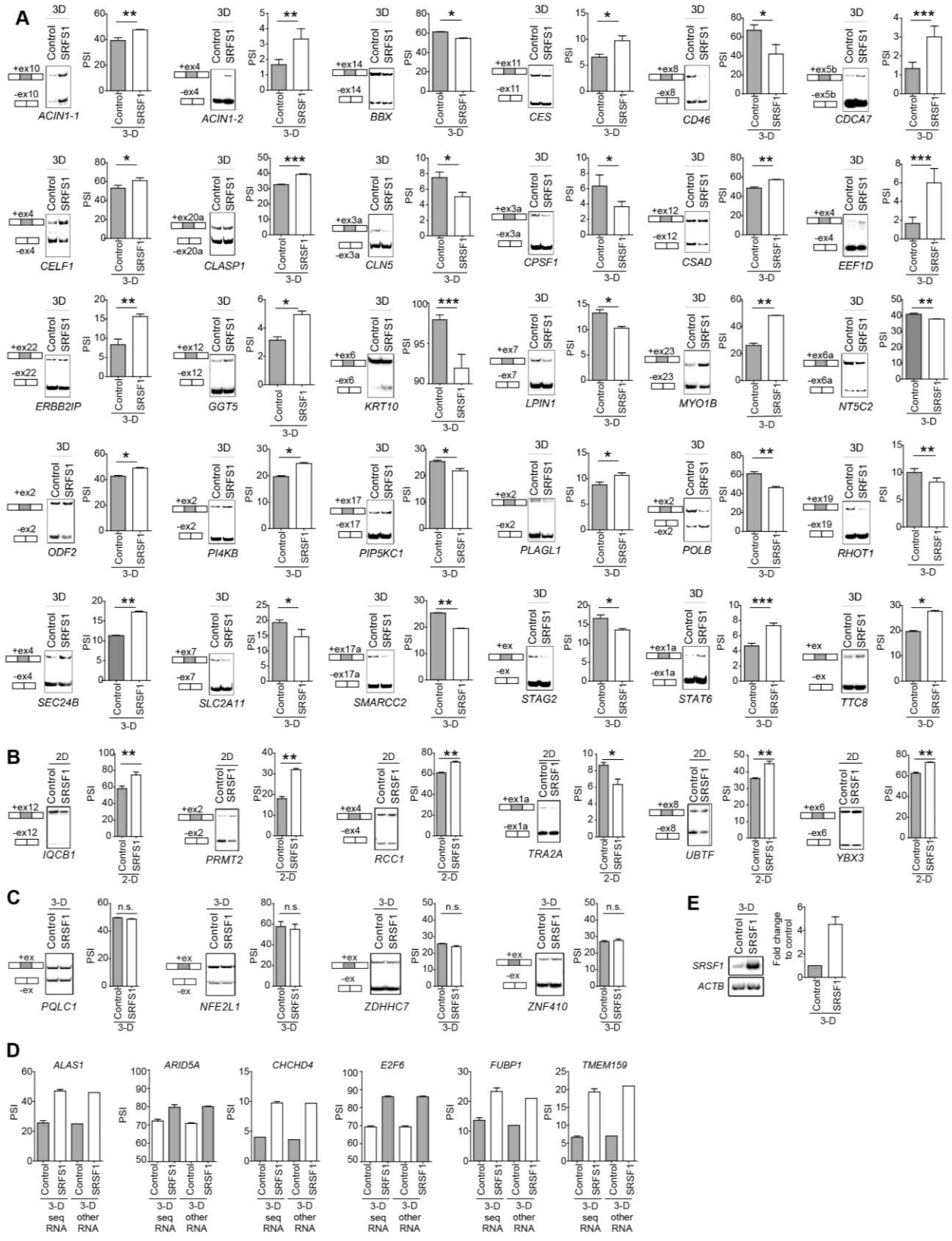


Figure S2. RT-PCR validation of SRSF1-regulated CA events in MCF-10A cells (Related to Figure 2). (A, B) Total RNA from 3-D (A) or 2-D (B) control or SRSF1-OE MCF-10A cells was analyzed by radioactive RT-PCR using primers in the upstream and downstream exons, followed by native PAGE

and autoradiography. The structure of each isoform is indicated (not to scale). CA-exons are colored. The percent spliced in (PSI) was quantified for each condition (n≥3; t-test *** $P<0.0005$, ** $P<0.005$, * $P<0.05$). Error bars, s.e.m. **(C)** CA-exons not regulated by SRSF1 were analyzed as described in (A,B). **(D)** The reproducibility of the RT-PCR validations was assessed by using RNA different ('other RNA') from the sample used to generate the RNA-seq libraries ('seq RNA'). **(E)** *SRSF1* transcript levels were analyzed by radioactive RT-PCR in control and SRSF1-overexpressing cells, normalized to *ACTB* levels (n≥3).

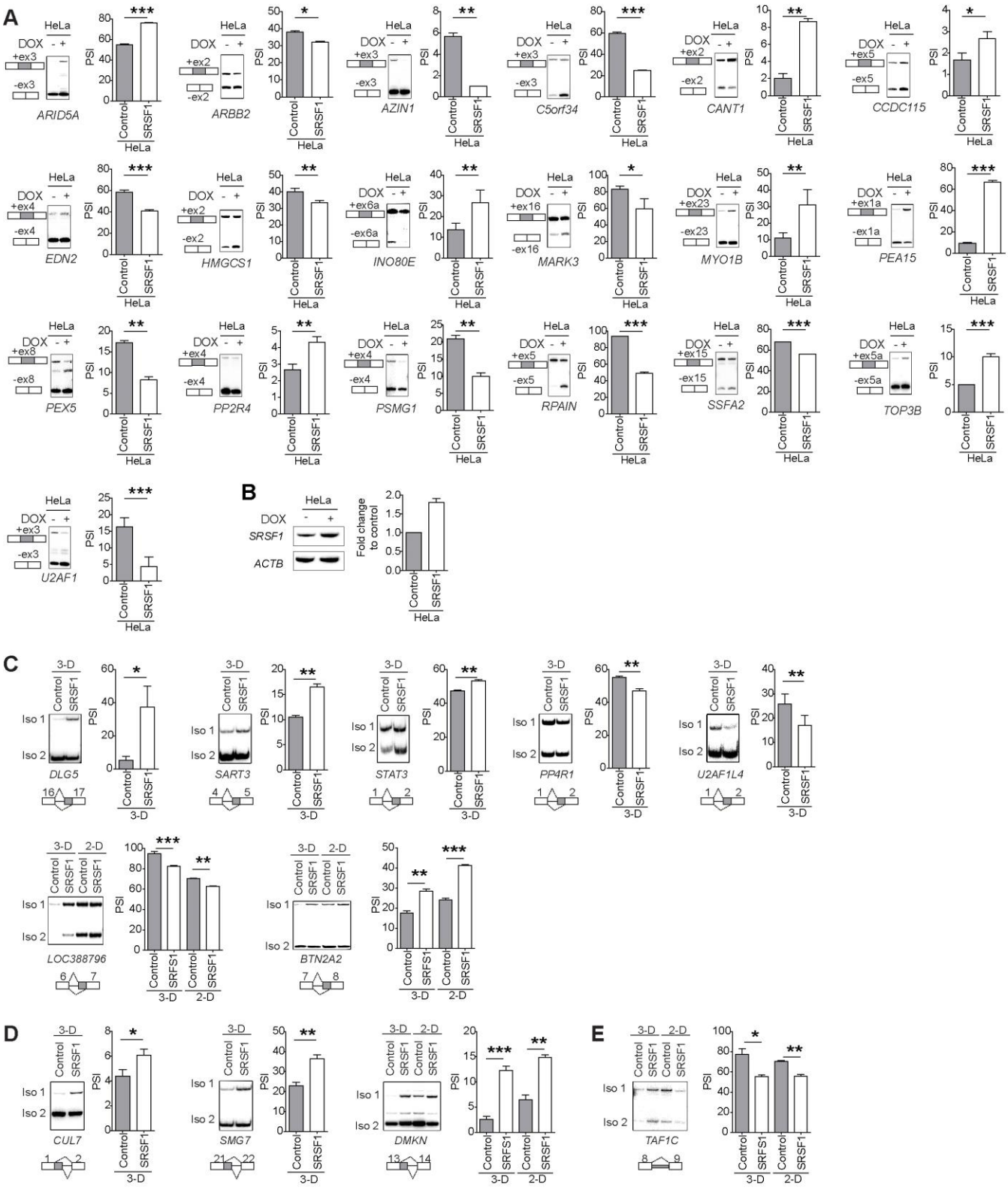
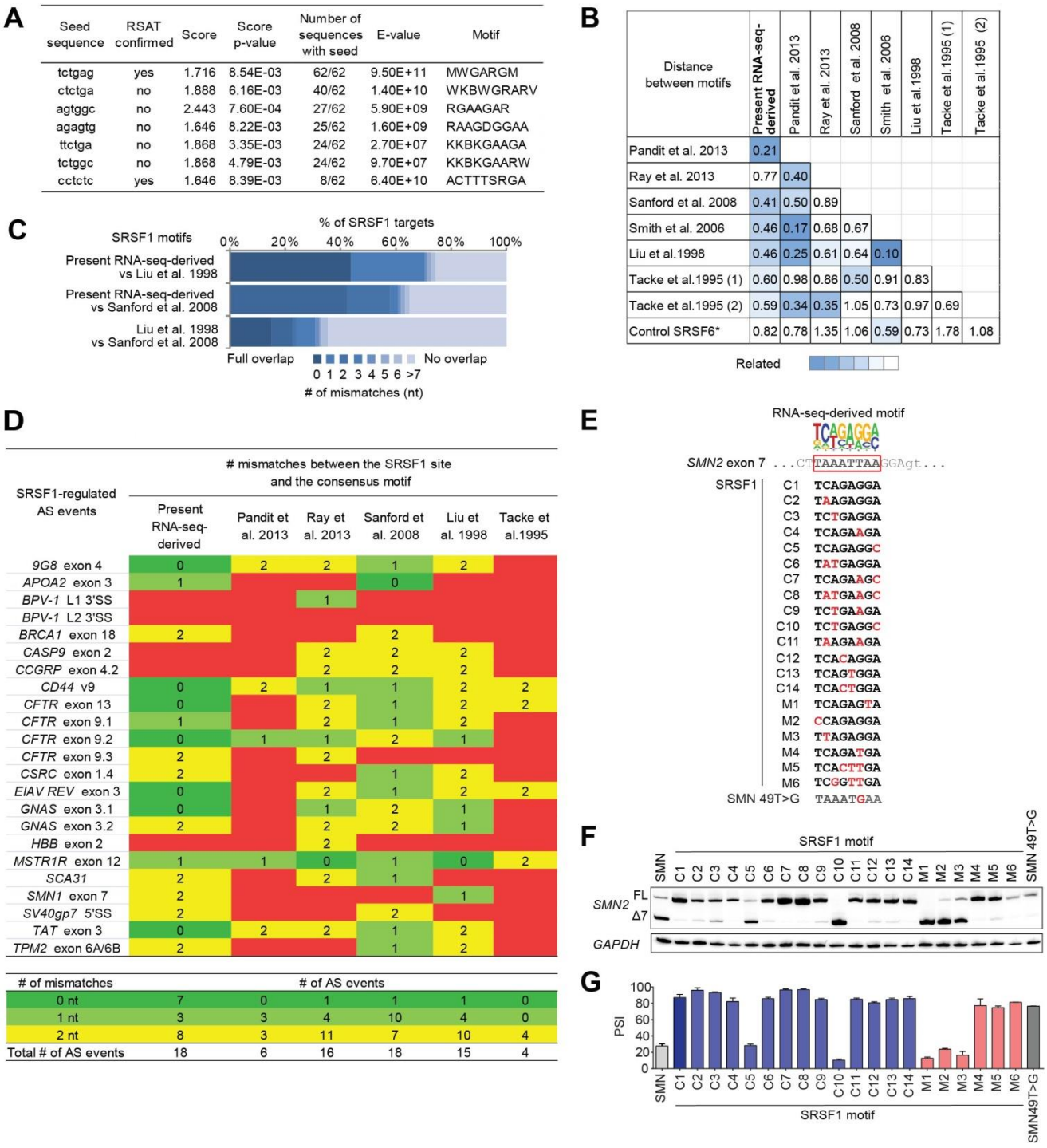


Figure S3. RT-PCR validation of SRSF1-regulated splicing events in HeLa (A-B) and in MCF-10A (C-E) cells (Related to Figure 2). (A) Inducible HeLa cells were treated with doxycyline (DOX) for 48 h to induce the expression of T7-SRSF1. Total RNA from SRSF1-OE DOX treated (+DOX) or control mock-treated (-DOX) cells was analyzed by radioactive RT-PCR using primers in the upstream and

downstream exons, followed by native PAGE and autoradiography. The structure of each isoform is indicated (not to scale). CA-exons are colored. The percent spliced in (PSI) was quantified for each condition (n≥3; t-test *** $P<0.0005$, ** $P<0.005$, * $P<0.05$). Error bars, s.e.m. **(B)** *SRSF1* transcript levels were analyzed by radioactive RT-PCR in control and SRSF1-overexpressing HeLa cells, normalized to *ACTB* levels (n≥3). **(C-E)** RT-PCR validation of SRSF1-regulated AA (C), AD (D) or IR (E) events**.** Total RNA from 3-D or 2-D control or SRSF1-OE MCF-10A cells was analyzed by radioactive RT-PCR using primers in the upstream and downstream exons, followed by native PAGE and autoradiography. The structure of each isoform is indicated (not to scale). Alternatively spliced sequences are colored. The percent spliced in (PSI) was quantified for each condition (n≥3; t-test *** $P<0.0005$, ** $P<0.005$, * $P<0.05$). Error bars, s.e.m.



Seed sequence	RSAT confirmed	Score	Score p-value	Number of sequences with seed	E-value	Motif
tctgag	yes	1.716	8.54E-03	62/62	9.50E+11	MWGARGM
ctctga	no	1.888	6.16E-03	40/62	1.40E+10	WKBWGRARV
agtggc	no	2.443	7.60E-04	27/62	5.90E+09	RGAAGAR
agagtg	no	1.646	8.22E-03	25/62	1.60E+09	RAAGDGGAA
ttctga	no	1.868	3.35E-03	24/62	2.70E+07	KKBKGAAGA
tctggc	no	1.868	4.79E-03	24/62	9.70E+07	KKBKGAARW
cctctc	yes	1.646	8.39E-03	8/62	6.40E+10	ACTTTSRGA

Distance between motifs	Present RNA-seq-derived	Pandit et al. 2013	Ray et al. 2013	Sanford et al. 2008	Smith et al. 2006	Liu et al.1998	Tacke et al.1995 (1)	Tacke et al.1995 (2)
Pandit et al. 2013	0.21							
Ray et al. 2013	0.77	0.40						
Sanford et al. 2008	0.41	0.50	0.89					
Smith et al. 2006	0.46	0.17	0.68	0.67				
Liu et al.1998	0.46	0.25	0.61	0.64	0.10			
Tacke et al.1995 (1)	0.60	0.98	0.86	0.50	0.91	0.83		
Tacke et al.1995 (2)	0.59	0.34	0.35	1.05	0.73	0.97	0.69	
Control SRSF6*	0.82	0.78	1.35	1.06	0.59	0.73	1.78	1.08

SRSF1-regulated AS events	# mismatches between the SRSF1 site and the consensus motif: Present RNA-seq-derived	Pandit et al. 2013	Ray et al. 2013	Sanford et al. 2008	Liu et al. 1998	Tacke et al.1995
9G8 exon 4	0	2	2	1	2	
APOA2 exon 3	1			0		
BPV-1 L1 3'SS			1			
BPV-1 L2 3'SS						
BRCA1 exon 18	2			2		
CASP9 exon 2			2	2	2	
CCGRP exon 4.2			2	2	2	
CD44 v9	0	2	1	1	2	2
CFTR exon 13	0		2	1	2	2
CFTR exon 9.1	1		2	1	2	
CFTR exon 9.2	0	1	1	2	1	
CFTR exon 9.3	2		2			
CSRC exon 1.4	2			1	2	
EIAV REV exon 3	0		2	1	2	2
GNAS exon 3.1	0		1	2	1	
GNAS exon 3.2	2		2	2	1	
HBB exon 2			2			
MSTR1R exon 12	1	1	0	1	0	2
SCA31	2		2	1		
SMN1 exon 7	2				1	
SV40gp7 5'SS	2			2		
TAT exon 3	0	2	2	1	2	
TPM2 exon 6A/6B	2			1	2	

# of mismatches	# of AS events					
0 nt	7	0	1	1	1	0
1 nt	3	3	4	10	4	0
2 nt	8	3	11	7	10	4
Total # of AS events	18	6	16	18	15	4

Figure S4. SRSF1 motif discovery, comparison and validation (Related to Figure 3). (A) Seed sequences used to derive the seeded psp-MEME SRSF1 motif in Figure 3A. **(B)** Divergence between the present RNA-seq-derived SRSF1 motif and previous motifs from the literature, as calculated using the Kullback-Leiber distance, indicated by the shading (see Experimental Procedures). *SRSF6 motif from (Liu et al., 1998) was used as a negative control. **(C)** Number of nucleotides overlapping between the present RNA-seq-derived SRSF1 motif and previous motifs in SRSF1-regulated targets identified by RNA-seq. **(D)** Predictive power of different SRSF1 motifs in previously reported SRSF1-regulated alternative splicing events. The number of mismatches between each actual SRSF1 site and the SRSF1

consensus motif is indicated and color-coded. The bottom part summarizes the number of SRSF1-regulated targets predicted using each of the motifs and classified into the following categories: 0/1/2 nt mismatches. **(E-G)** Mutational analysis of the RNA-seq-derived SRSF1 motif in a *SMN2* minigene reporter. (**E**) Variants of the SRSF1 sequences were inserted into *SMN2* exon 7 at the indicated positions. Variants predicted to match the consensus motif (C1-C14), or to abolish binding (M1-M6) are colored in red. The 49T>G mutation previously shown to promote exon inclusion (Singh et al., 2004) was used as a positive control. (**H**) *SMN2* exon 7 inclusion was analyzed by radioactive RT-PCR using primers in the upstream and downstream exons, followed by native PAGE and autoradiography. (**G**) The percent spliced in (PSI) was quantified for each condition (n=4). Error bars, s.d.

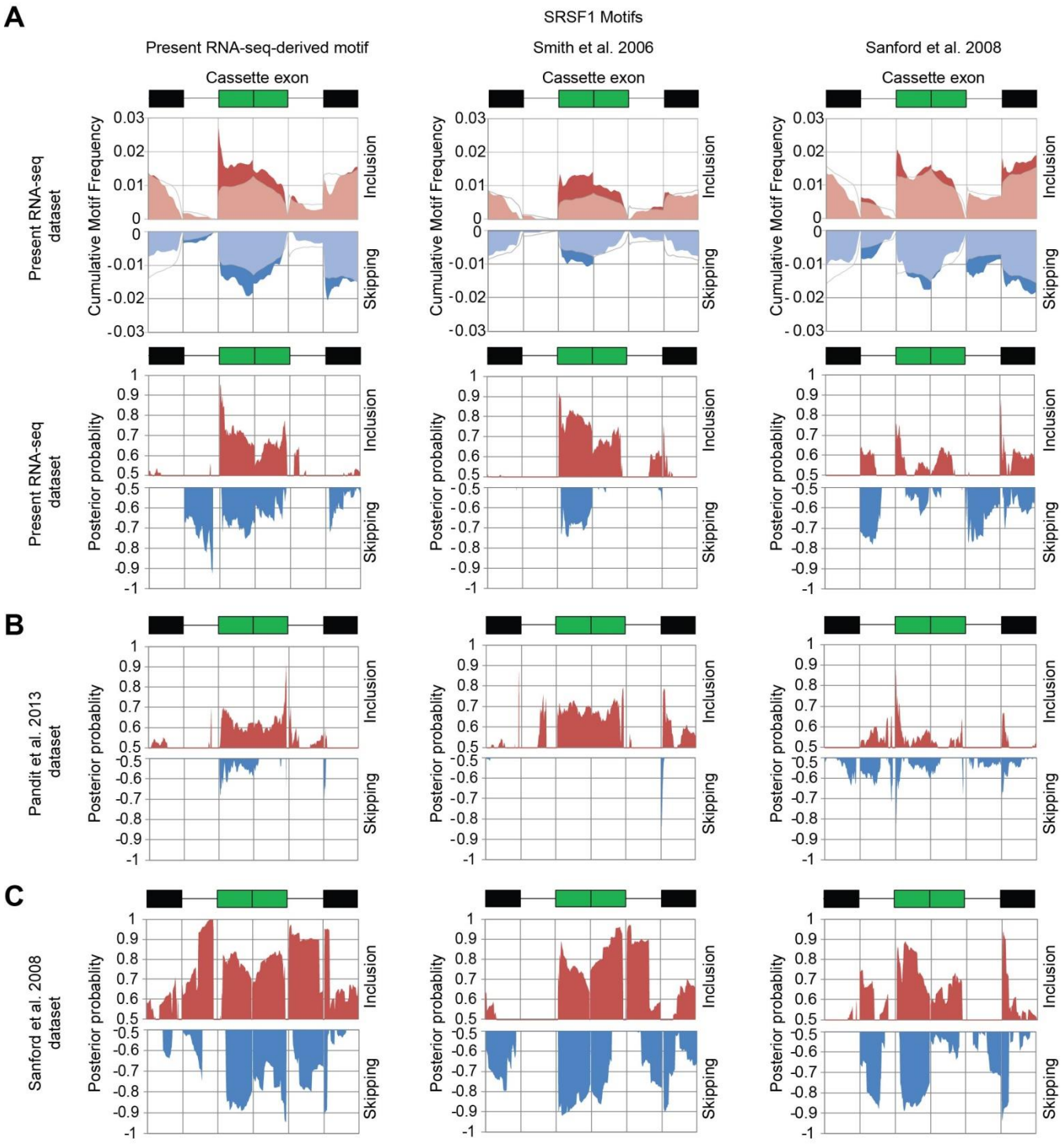
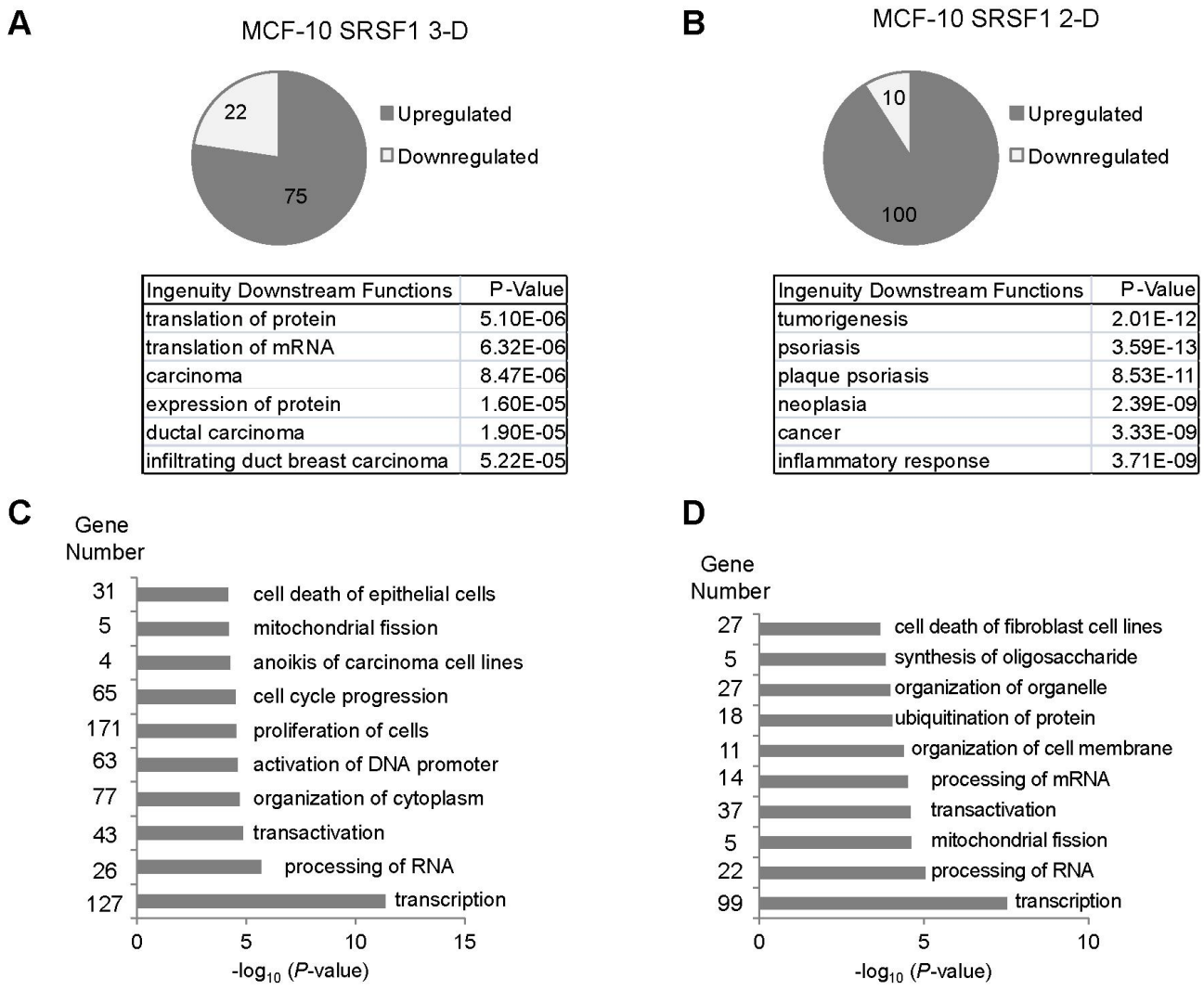


Figure S5. Comparison of SRSF1-motif regulatory maps (Related to Figure 4). (A) SRSF1-motif frequency (top panel) and probability (lower panel) maps were derived using the RNA-seq data and the various SRSF1 motifs, as described in Figure 3. **(B,C)** SRSF1-motif probability maps were derived using the data from Pandit et al. 2013 **(B)** or Sanford et al. 2008 **(C)** using the indicated motifs.



Ingenuity Downstream Functions	P-Value
translation of protein	5.10E-06
translation of mRNA	6.32E-06
carcinoma	8.47E-06
expression of protein	1.60E-05
ductal carcinoma	1.90E-05
infiltrating duct breast carcinoma	5.22E-05

Ingenuity Downstream Functions	P-Value
tumorigenesis	2.01E-12
psoriasis	3.59E-13
plaque psoriasis	8.53E-11
neoplasia	2.39E-09
cancer	3.33E-09
inflammatory response	3.71E-09

Figure S6. SRSF1-regulated expression and splicing changes are relevant to the phenotype (Related to Figure 5). (A-B) Expression changes in control or SRSF1-overexpressing MCF-10A cells grown in 3-D **(A)** or 2-D **(B)** cultures were determined by RNA-seq (see details in Experimental Procedures). The number of significantly upregulated or downregulated genes in SRSF1-overexpressing cells compared to control cells is plotted. Pathway enrichment using Ingenuity (see details in Experimental Procedures) suggests that changes in 3-D grown cells are associated with biological functions connected to breast cancer. **(C-D)** Pathway and functions enriched in SRSF1-splicing targets in 3-D **(C)** and 2-D MCF-10A **(D).**

Supplemental Tables (see excel file)

Table S1. List of all SRSF1-regulated splicing events detected by RNA-seq (Related to Figure 1)

Table S2. Full annotation of SRSF1-regulated CA exons (Related to Figure 2)

Table S3. Validations of SRSF1-regulated AA, AD and IR events (Related to Figure 2)

Table S4. Splicing-factor motif enrichment in SRSF1-regulated splicing events (Related to Figure 3 and Figure 4)

Table S5. Association between splice-site strength and SRSF1-binding motifs (Related to Figure 3 and Figure 4)

Table S6. SRSF1-induced expression changes in MCF-10A cells (Related to Figure 5)

Table S7. AS detected in SRSF1-overexpressing human breast tumors (Related to Figure 5 and 6)

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Three-dimensional assays

MCF-10A stable cell lines were seeded on chamber slides coated with Matrigel Growth Factor Reduced (BD Biosciences) as described (Debnath et al., 2003). At least 100 acini were imaged at the indicated time points using a Zeiss Axiovert 200M microscope and AxioVision 4.5 software (Zeiss).

Immunofluorescence procedures were performed as described (Anczukow et al., 2012). Microscopy was performed on a Zeiss Observer instrument (Zeiss). Cleaved caspase-3 (Cell Signaling) and ki67 (Zymed) primary antibodies, and Alexa Fluor 568 anti-mouse and 488 anti-rabbit secondary antibodies (Invitrogen), were used at 1/100 and 1/500 dilution, respectively. Acini were scored positive for ki67 when at least five cells within the acini were stained, and positive for cleaved caspase-3 when at least one cell in the lumen was stained.

Western blot analysis

For 3-D cultured acini, cells were harvested as described above. For protein extraction, cells were washed with PBS and lysed in RIPA buffer (50 mM Tris pH 8, 150 mM NaCl, 1% (v/v) NP40, 0.5% (w/v) deoxycholate, 0.1% (w/v) SDS) supplemented with Complete Protease Inhibitor Cocktail tablets (Roche) and 20 mM β-glycerophosphate. Equal amounts of total protein, as measured by Bradford's assay, were loaded on a 12% SDS-polyacrylamide gel, transferred onto a nitrocellulose membrane (Millipore) and blocked in 5% (w/v) milk in Tween 20-TBST (50 mM Tris pH 7.5, 150 mM NaCl, 0.05% (v/v) Tween 20). Blots were incubated with SRSF1 (Zhang and Krainer, 2004), β-catenin (BD), or Tubulin (Sigma) primary antibodies. IR-Dye 680 anti-mouse or IR-Dye 800 anti-rabbit immunoglobulin G (IgG) secondary antibodies (Licor) were used for infrared detection and quantification with an Odyssey Imaging System (Licor).

RNA-sequencing

For RNA extraction from 3-D cultured acini, cells were washed with PBS, and Matrigel was dissolved by incubating slides at 4 °C in Cell Recovery Solution (BD). Total RNA from 3-D or 2-D grown cells was extracted by ultracentrifugation through a cesium-chloride cushion, treated with DNAse I (Promega), followed by phenol-chloroform extraction and ethanol precipitation, and was then used to prepare libraries for RNA-seq. Briefly, 10 µg of total RNA was hybridized to Dynal oligo(dT) beads (Invitrogen) in two sequential rounds, according to the manufacturer's instructions. RNA was fragmented by incubation with fragmentation buffer (Ambion) at 70 °C for 4 min. Fragmented RNA was purified by ethanol precipitation. First-strand synthesis was performed using random hexamer primers (Invitrogen) and Superscript II reverse transcriptase (Invitrogen). Double-stranded cDNA was synthesized by first incubating the RNA with second-strand buffer, RNase Out, and dNTPs (Illumina), on ice for 5 minutes. The reaction mix was then treated with DNA Pol I and RNase H at 16 °C for 2.5 h. To create blunt-end DNA, the double-stranded cDNA was then incubated with T4 DNA polymerase,

Klenow DNA polymerase, and T4 polynucleotide kinase at 20 °C for 30 min. A single 'A' base was then added using Klenow (3' to 5' exo) and dATP. Paired-end adaptors (Illumina) were ligated using a Rapid DNA Ligation kit (Roche). Size selection (200 base pairs) of DNA was performed by cutting the target fragment out of a 2% agarose gel. The amplified DNA library was prepared using the Expand High Fidelity PLUS PCR System (Roche). Each library was sequenced in duplicate on a Genome Analyzer IIx (Illumina) to produce paired-end 36-nt reads.

Analysis of splicing changes

AS events (CA, AD, AA and IR) were identified and quantified using SpliceTrap (Wu et al., 2011). This tool combines RNA-seq data with prebuilt transcript models to quantify the level of inclusion of every exon in a transcript. The transcript models are exon trios, composed of alternative-exon candidates with their annotated flanking exons. They were derived from the hg18 TXdb database, which accounts for a total of 299,718 exon-trio models (Wu et al., 2011). SpliceTrap utilizes the Bowtie read aligner to align reads against TXdb, which was used as a reference.

SpliceDuo

Introduction: One challenge in RNA-seq data analysis is the estimation of significant variation across samples. Gene expression is usually measured from RNA-seq data in RPKM units, which often fluctuate from ~zero to tens of thousands. This large dynamic range makes it possible to identify variation in fold-change units and assign statistical significance through methods that exploit data variance (Anders and Huber, 2010). However, exon-inclusion profiles have a smaller dynamic range, consisting of probabilistic measures (from 0 to 1) with a beta-like ("U"-shaped) distribution (Wu et al., 2011). A common practice in the field is to compute splicing variance by estimating the difference between PSI values (percent spliced in) across conditions, and using a cutoff such as ΔPSI=|0.1| or |0.2| (Han et al., 2014). However, this threshold-based approach does not address the problem of assigning statistical significance to splicing variation. To this end, we developed the SpliceDuo algorithm.

Given a paired list of PSI values $\forall n \in E \rightarrow \{P_{e1}, P_{e2}\}$ describing two measurable instances of exon *e* in an experiment *E*, SpliceDuo implements the Thin Plate Spline (TPS) transformation (Wahba, 1990) to convert the two-dimensional (2D) field (P_{e1}, P_{e2}) of irregularly spaced data, into three-dimensional (3D) space (P_{e1}, P_{e2}, D), where D is the density of data points on the coordinates (P_{e1}, P_{e2}). We then use the distribution of points D to assign FDR values.

We assume that for most instances of e, P_{e1} is approximately equal to $P_{e2.}$ such that $P_{e1} = P_{e2} \pm \varepsilon$ where ε is the methodological error. However, there is a quantity *m*, a subset of *E* that meets $(m > 0 \subset n) \in E: \{P_{e1} \ncong P_{e2}\}$ accounting for noticeable variation in (P_{e1}, P_{e2}) pairs. Under these assumptions, in a 2D field, the density of (P_{e1}, P_{e2}) points will be high around $P_{e1} = P_{e2} \pm \varepsilon$ and low at $P_{e1} \ncong P_{e2}$. Therefore, the inverse of the points' density can be used to estimate the data deviation from the probability density distribution.

Thin Plate Spline: TPS is a non-parametric regression algorithm often applied to image alignment (Joshia et al., 2007) and shape matching (Belongie et al., 2002). Essentially, it interpolates a three-

dimensional surface onto irregularly spaced data defined by two components (x_i, y_i). The solution of the model consists in choosing a function f that minimizes the physical bending energy of the interpolated surface (Bookstein, 1989). For this reason, the name TPS refers to an analogy involving the bending of a thin sheet of metal or a chocolate wrapper.

The energy function is formulated as follows:

$$E[f] = \int_{\mathbb{R}^n} |M^2 f|^2 dX$$

In which $M^2 f$ is the matrix of second-order partial derivatives of f and $|M^2 f|^2$ is the sum of squares of the matrix entries. The infinitesimal elements of hypervolume are dX = dx1…dxn where xi are the components of X.

$$D^2 f = \left(\frac{\partial^2 xy}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 xy}{\partial x\, \partial y}\right)^2 + \left(\frac{\partial^2 xy}{\partial y^2}\right)^2 dx\, dy$$

M is used to fit a TPS surface using *E[f]* to minimize the residual sum of squares subject to the constraint that the function have a certain level of smoothness, λ, implemented by the generalized Krig model. The Krig model assumes that the unknown function is a realization of a Gaussian random spatial process. The assumed model is additive, $Y = O^a(x) + Z(x) + \varepsilon$, in which O^a is a low-order polynomial and Z is a zero-mean, Gaussian stochastic process with a covariance that is unknown up to a scale constant.

It is practical to formulate the problem with a smoothing parameter for regularization. A function f(xi, yi) is chosen that does not necessarily interpolate exactly all the data points, but that minimizes construction of the TPS:

$$E[f] = \sum_{i=1}^{n} |f(x_i) - y_i|^2 + \lambda \int_{\mathbb{R}^n} |M^2 f|^2 dX$$

Here n is the number of data points defined by the components (X_I, Y_I), and λ is the smoothing parameter. Smoothness is quantified by the integral of squared λ-th order derivatives. For one dimension and λ=2, the smoothness is the integrated square of the second derivative of the function. λ=0 corresponds to no smoothness constraints, and the data are interpolated. λ=∞ corresponds to just fitting the polynomial base model by ordinary least squares.

The TPS regression model is explained in detail in (Bookstein, 1989; Wahba, 1990). We implemented the TPS model using the 'tps' and 'predict.surface' functions of the 'fields' R package http://cran.r-project.org/web/packages/fields/fields.pdf.

Algorithm: Below we describe the SpliceDuo algorithm that assigns FDR values to splicing changes.

Given $\forall n \in E \rightarrow \{P_{e1}, P_{e2}\}$, the values of P_{e1}, P_{e2} are discretized to a matrix $M = [P_{e1}, P_{e2}]_{i,j,b}$, (Figure S10B) in which the size of M is given by the binning parameter b.

1. M is used to fit a TPS surface using the energy function E[f] implemented by the Krig model. The main advantage of this implementation is that once a TPS/Krig object is created, the estimator can be rapidly found for other data and smoothing parameters, provided that the locations remain unchanged. This allows a faster implementation of the 'tps' R package.

   The output of this step is a TPS radial-basis function represented by the following parameters:

   - Maximum likelihood estimation of error (φ),
   - Maximum likelihood estimation of covariance (ρ)

2. We create a grid G using the maximum likelihood values φ, ρ, the smoothing parameter λ, and the size *n.* The output can be adjusted by the parameters *x, y* and *S*: *x*y* is the number of cells in the grid, and *S* is the extrapolation Boolean. When S=1, extrapolation is used.

3. We perform the G→G' conversion, in which G' is a grid of *p-values*, computed from the Z-scores of $N(G(\mu^c, \sigma^c))$. N is a normal distribution derived from μ^c and σ^c, which are the mean and standard deviation, respectively, of the *c* confidence interval of the data. The p-values are then corrected by the FDR procedure.

4. A maximum *p-value* cutoff is set on G' and used to select significant variation in P_{e1}, P_{e2} pairs. The superposition of the original with the generated noise model can easily highlight significant changes in the data.

Feature optimization: To achieve maximal performance, the following parameters were optimized

- The binning parameter b of the frequency matrix M.
- The smoothing parameter λ.
- The grid size *x*y* of the grid G.
- The confidence interval *c* of the distribution $N(G(\mu^c, \sigma^c))$.
- The extrapolation Boolean S.

To optimize the set of parameters, we compared two technically identical replicates of RNA-seq data through a series of iterations in which we tested all possible combinations of parameters (a total of 1000 combinations).

List of optimized parameters and their values

Parameters	**Possible Values**
Binning Parameter *b*	0.1, 0.5, 0.01, 0.05
Smoothing parameter λ	0, 0.0001, 0.001, 0.01,1,2
Grid Size *x*y*	10*10, 20*20, 50*50, 100*100
Confidence interval *c*	0.7, 0.8, 0.9, 0.95, 0.99

Extrapolation S	0,1

The metrics that were used to determine the optimal parameters were the AS discovery rate, true positive rate, total number of exons, and minmax values.

The CA exon discovery rate (ASD) is the ratio of predicted AS events in the data compared to the total number of known events (TS):

$$ASD = \frac{AS}{AS + TS}$$

Whether an event is a known AS event is determined through a comparison to the TXdb database, which contains annotations for tens of thousands of known AS events (Wu et al., 2011).

The true positive rate (TPR) compares the number of AS events detected when merging the two NGS data replicates into a single file (true positives) to the same number of AS events detected in only one replicate, but not the other (FP):

$$TPR = \frac{TP}{TP + FP}$$

The total number of detected AS events (N), corresponds to significant changes with *p-value* <0.05.

ASD, TPR and N are finally combined through the minmax approach to produce a single score:

$$minmax = \frac{ASD_{max} - ASD}{ASD_{max}} + \frac{TPR_{max} - TPR}{TPR_{max}} + \frac{\log_{10} N_{max} - \log_{10} N}{\log_{10} N_{max}}$$

Looking at the minmax value from each possible combination, the parameter set with the lowest minmax value is chosen as the optimal set of parameters. Based on our simulation, we chose the following parameters for SpliceDuo analysis: b=0.05, λ=0.001, x*y=10*10, *c*=0.9, S=1.

TCGA RNA-seq data analysis

A total of 57 SRSF1-overexpressing cell lines (≥2-fold change) and 46 controls (≥-0.01 and ≤0.01 fold change) were analyzed with SpliceTrap.

List of TCGA breast tumor samples with or without SRSF1 overexpression

Tumor Group ID	TCGA Tumor	SRSF1 overexpression level	TCGA RNA-seq ID	Does not overexpress (≥2x) any SR proteins besides SRSF1
SRSF1/22	TCGA-C8-A1HL-01	6.7	51464022-7dc8-42a4-8214-be47ab87a555	TRUE
SRSF1/5	TCGA-AO-A1KS-01	4.3	15aba40e-f01a-459e-aa7d-874ad768eb39	TRUE

SRSF1/6	TCGA-AR-A0TY-01	3.7	16c15728-1027-4b99-9399-a13b819c168d	TRUE
SRSF1/40	TCGA-E2-A14V-01	2.6	b554c4a9-b18d-4d2a-9adc-95781f5117c3	TRUE
SRSF1/37	TCGA-E2-A15H-01	2.4	a9fde415-1397-4a49-9c8b-e7fed32bf9c1	TRUE
SRSF1/46	TCGA-C8-A130-01	2.3	c7c70aca-370d-44ed-9610-3641b76fa5c2	TRUE
SRSF1/25	TCGA-AN-A0XR-01	2.2	61580f5a-5938-437d-910b-04541720bc31	TRUE
SRSF1/18	TCGA-GM-A2DA-01	4.5	41b88db2-9151-46d6-a8d0-c0a17e622a44	FALSE
SRSF1/55	TCGA-C8-A12U-01	2.2	fd65c23f-a026-4b5d-9ffb-1e8f8019b225	FALSE
SRSF1/56	TCGA-A2-A25B-01	5.4	fe7cbe2d-2e22-4ca4-aa8b-e1ae81d90ffe	FALSE
SRSF1/29	TCGA-A1-A0SK-01	4.2	79bb03e0-9bf3-4c81-a44e-4907542e33d5	FALSE
SRSF1/42	TCGA-BH-A1FM-01	4.1	b9316f28-2335-4055-95a7-e48d86f72dad	FALSE
SRSF1/19	TCGA-BH-A1F2-01	4.0	43e98f55-664d-48f0-8eb9-dd7d06152c86	FALSE
SRSF1/38	TCGA-AN-A04D-01	3.7	aaff09c1-7c1f-4421-af06-694d615ac72f	FALSE
SRSF1/24	TCGA-BH-A1F5-01	3.7	57de5390-e100-4d48-8e8c-46980b0c7b9f	FALSE
SRSF1/43	TCGA-AN-A04C-01	3.6	c0512cba-bfe1-4d51-8fb8-262b5a38e3cf	FALSE
SRSF1/45	TCGA-E2-A14P-01	3.4	c6f65c79-8d8b-4784-be80-64a69b4cc000	FALSE
SRSF1/32	TCGA-A2-A0T3-01	3.3	8d2ad50a-29b0-4b9f-906b-6722c69f0927	FALSE
SRSF1/28	TCGA-AR-A24H-01	3.3	705eb8a5-afa0-47ba-afef-9d2639eb22da	FALSE
SRSF1/16	TCGA-E2-A1L7-01	3.2	3fb0f2a9-8c50-4df6-9159-995951a42f3b	FALSE
SRSF1/34	TCGA-AO-A03O-01	3.2	98b113af-a275-4665-8480-0ebe27367f65	FALSE
SRSF1/13	TCGA-D8-A143-01	3.2	330f7b75-6840-4af6-9c12-278e1d75e080	FALSE
SRSF1/41	TCGA-A8-A07R-01	3.2	b7708ee9-c206-4caf-88f6-63eca6400506	FALSE
SRSF1/4	TCGA-C8-A1HM-01	3.2	11f2ae41-f230-4685-b19f-19e14542577f	FALSE
SRSF1/47	TCGA-C8-A1HF-01	3.2	c80ff812-5259-4978-9928-1e7d138bbcb0	FALSE
SRSF1/14	TCGA-A2-A1FW-01	3.1	36389ee8-7301-4896-a060-d6b594d495e0	FALSE
SRSF1/52	TCGA-AN-A0AT-01	3.0	f888fc09-259f-42a1-b4fe-04403fcef6f8	FALSE
SRSF1/33	TCGA-AO-A1KP-01	3.0	8ffc0191-7697-4f2d-97c3-b324e284300b	FALSE
SRSF1/8	TCGA-E2-A14N-01	2.8	1cacb904-523c-4042-a384-cc24ce643c2b	FALSE
SRSF1/3	TCGA-A8-A09W-01	2.8	0ed34724-ddc4-4d5b-ae25-d798d42c47ba	FALSE
SRSF1/20	TCGA-B6-A402-01	2.7	46b26380-be4c-4412-9dc2-8a1b3ca46005	FALSE
SRSF1/15	TCGA-AR-A2LK-01	2.7	3f761da6-f76e-4c38-ad9a-2bd36ea80f43	FALSE
SRSF1/35	TCGA-EW-A1IY-01	2.6	9b4f1411-1789-4a09-9f82-e36c00f5cd21	FALSE
SRSF1/23	TCGA-D8-A13Y-01	2.6	535a7ce3-1aee-4efd-8494-2e3e7d9f0b80	FALSE
SRSF1/39	TCGA-AO-A12F-01	2.5	afc1f310-6951-40bd-9961-04bb9037b392	FALSE
SRSF1/9	TCGA-AR-A0U2-01	2.5	206e3d9b-64eb-45b6-9aa9-928340484709	FALSE
SRSF1/50	TCGA-E9-A1RB-01	2.5	dbce421e-d024-4388-8265-bb18c153ac15	FALSE
SRSF1/27	TCGA-A8-A092-01	2.5	66c59bdb-823a-4258-9615-b9b98974bce5	FALSE
SRSF1/31	TCGA-AO-A0J3-01	2.4	885e0d64-c485-44cc-b437-d41a039a92f0	FALSE
SRSF1/17	TCGA-E9-A5FL-01	2.4	41a5b666-fbaf-4080-b6ab-2ea3efcf5341	FALSE
SRSF1/26	TCGA-BH-A202-01	2.3	637160f2-ca4a-40f3-8bba-b25a4676be4b	FALSE
SRSF1/7	TCGA-E9-A248-01	2.3	194fa6f2-499b-453f-95d0-a947596ff454	FALSE
SRSF1/49	TCGA-E9-A1RH-01	2.3	da3d2e3d-5386-4022-88d9-ac33fd4572eb	FALSE
SRSF1/44	TCGA-AN-A0AJ-01	2.3	c5d898e7-3160-4742-b401-0bb6c04e2289	FALSE
SRSF1/1	TCGA-AR-A24S-01	2.2	008886b5-e60a-4dca-9730-ababd8ae1e94	FALSE

SRSF1/12	TCGA-A2-A25C-01	2.2	27f30aab-e356-4e6d-8f3c-d5f731f35bea	FALSE
SRSF1/54	TCGA-AO-A0JB-01	2.2	fcff11f7-001b-4640-be13-30e527e4ddf4	FALSE
SRSF1/11	TCGA-E9-A22H-01	2.2	22bcc46e-3566-4bfe-83c2-eddb8d9a5c1f	FALSE
SRSF1/2	TCGA-A8-A095-01	2.2	05a845e1-278d-4879-b5ab-c4653ba51543	FALSE
SRSF1/57	TCGA-C8-A1HN-01	2.2	ff20324c-a37e-4b9f-88b7-ce5a9e8f06f5	FALSE
SRSF1/10	TCGA-C8-A27B-01	2.2	225a8ab7-48b4-4c67-b310-46346f93cc11	FALSE
SRSF1/51	TCGA-AR-A256-01	2.1	ea0add07-2cd7-4230-a37d-c621bec85b17	FALSE
SRSF1/30	TCGA-E2-A14Y-01	2.1	86eb9b62-abee-4253-9f67-b358ed1748d9	FALSE
SRSF1/36	TCGA-EW-A1J1-01	2.1	9f55d11f-53db-408a-b9f4-3a0acb236aef	FALSE
SRSF1/53	TCGA-E2-A105-01	2.1	fa350888-da92-477d-a79c-7a25fe351925	FALSE
SRSF1/48	TCGA-AR-A2LL-01	2.1	cf3e7395-c3a1-46b5-8ae5-e5588923fb35	FALSE
SRSF1/21	TCGA-A2-A0YM-01	2.1	4a488588-db2c-4db8-a635-632b790e577e	FALSE
CONTROL/25	TCGA-AR-A24W-01	0.1	6086f8a0-ccdd-40a2-aabd-e4c42cbf26cd	n/a
CONTROL/15	TCGA-B6-A0I5-01	0.1	28f684d2-2913-41a9-b432-8abf64c7554a	n/a
CONTROL/21	TCGA-E2-A14T-01	0.1	4273af2b-c823-4912-aeed-45a9f1397fa6	n/a
CONTROL/13	TCGA-E9-A1RA-01	0.1	22aa5f58-35ff-4318-81ec-8ba695ba3f2a	n/a
CONTROL/9	TCGA-BH-A0AV-01	0.1	1763262a-1ceb-4ed5-a589-36968e96b1e1	n/a
CONTROL/8	TCGA-D8-A141-01	0.1	15f05ea4-6a16-49f2-97bd-240be29a96c2	n/a
CONTROL/5	TCGA-E9-A54X-01	0.1	0d13bf66-6f1a-4326-ac9c-24a4e934a37c	n/a
CONTROL/30	TCGA-AO-A129-01	0.1	7359d5ac-69a3-4c8f-8c09-4c24e3545825	n/a
CONTROL/17	TCGA-B6-A0RH-01	0.1	300a51a4-3fef-47e8-98f2-8fcb0a8f0964	n/a
CONTROL/3	TCGA-E2-A1B0-01	0.1	0adb3052-f0e8-4eac-967d-daa7b8ac9cb3	n/a
CONTROL/14	TCGA-D8-A1XC-01	0.1	28debfe8-16e5-44d6-a9a7-28887e61f8c0	n/a
CONTROL/6	TCGA-A2-A3XU-01	0.1	1021bfc4-4f27-4e43-b4f2-04784569906c	n/a
CONTROL/26	TCGA-BH-A0BZ-01	0.1	64c6a628-936f-4a99-93a3-8579a458f5c6	n/a
CONTROL/39	TCGA-A8-A07G-01	0.1	85e776a5-43f5-4e7d-b675-c998791c413c	n/a
CONTROL/18	TCGA-AN-A0FV-01	0.1	36a6c0dc-789e-476c-aa12-367563f25638	n/a
CONTROL/29	TCGA-A8-A07W-01	0.0	6f609739-cc90-42e7-af00-85ac68208429	n/a
CONTROL/2	TCGA-D8-A1XT-01	0.0	05ddeda5-3779-478a-8481-a2f18072e8b6	n/a
CONTROL/4	TCGA-AR-A0U0-01	0.0	0b020e24-040f-436b-984d-69cff92f061f	n/a
CONTROL/38	TCGA-BH-A1EN-01	0.0	80670658-fea6-42e9-82da-7df593b1453a	n/a
CONTROL/27	TCGA-BH-A18J-01	0.0	672b751f-383f-4300-baba-a81fa2c333ad	n/a
CONTROL/1	TCGA-A2-A0YC-01	0.0	04cf2014-5f71-45cf-a449-3af1a9449c61	n/a
CONTROL/41	TCGA-E9-A1NF-01	0.0	8ededc6f-0a8d-4146-bab3-116f17664bd7	n/a
CONTROL/35	TCGA-AN-A0FX-01	0.0	7dcd6e14-a7a0-4bbd-83ce-9d3ffa69ef3c	n/a
CONTROL/24	TCGA-A8-A06U-01	0.0	576a24c2-4da7-41ad-9c37-4cb005976678	n/a
CONTROL/11	TCGA-AR-A5QQ-01	0.0	19f61454-464e-48de-af60-91aa95a66249	n/a
CONTROL/23	TCGA-D8-A1XL-01	0.0	549a3fac-84ce-49a1-9690-c8e7dc5b5d1d	n/a
CONTROL/20	TCGA-LL-A5YP-01	0.0	41fa9e01-db32-42e8-9e60-87594e5d7dab	n/a
CONTROL/45	TCGA-AC-A2B8-01	0.0	e16d6a4a-a82d-4e2f-90b7-efdf40e3961f	n/a
CONTROL/16	TCGA-A7-A13F-01	0.0	2bf63257-413d-40c2-a9f9-1bfb8e7a2ebc	n/a
CONTROL/12	TCGA-A2-A1G4-01	0.0	1f44b1d9-8dc3-4ddb-8b05-0c2d6038d70b	n/a
CONTROL/7	TCGA-A8-A08B-01	0.0	14502ddf-7933-4be1-9a9b-32fd132f328e	n/a
CONTROL/19	TCGA-A8-A079-01	0.0	3faba5b4-bb41-4598-87a4-77dc20447972	n/a
CONTROL/40	TCGA-AO-A128-01	0.0	8e6b0f42-5e26-46ea-a06a-3a7c26d965a9	n/a

CONTROL/22	TCGA-A2-A1FV-01	0.0	4f0afea9-830d-4057-a6ca-8d4183f70b57	n/a
CONTROL/43	TCGA-AR-A1AJ-01	-0.1	974909ad-b9b6-47db-a0f6-da7a2780c025	n/a
CONTROL/37	TCGA-AC-A2FE-01	-0.1	800f6089-3982-408d-8e78-702afa8c83bc	n/a
CONTROL/34	TCGA-E2-A1B5-01	-0.1	7c51b5ad-cecd-41f6-8908-5d1ca13128af	n/a
CONTROL/33	TCGA-AR-A24U-01	-0.1	774437b9-285e-4b82-b0bb-0a13e28ebd3b	n/a
CONTROL/10	TCGA-A1-A0SI-01	-0.1	19563808-839d-410a-a10a-68e693e24bd9	n/a
CONTROL/36	TCGA-EW-A1IZ-01	-0.1	7ec20760-c02a-4688-bd33-ea84dbdc6e7f	n/a
CONTROL/31	TCGA-A2-A4RX-01	-0.1	73865900-ab18-4f04-a2b2-27d17dd9f03b	n/a
CONTROL/46	TCGA-OL-A5D8-01	-0.1	e2cf8377-b97d-4f85-a6f6-5ea6bc2ecde9	n/a
CONTROL/44	TCGA-BH-A1ES-01	-0.1	9b8499c1-1985-4f36-8813-3af5f389488a	n/a
CONTROL/32	TCGA-D8-A27H-01	-0.1	74925915-0801-448a-abb3-7ccc8e8ab278	n/a
CONTROL/28	TCGA-A2-A0ST-01	-0.1	6c3617cd-e7e2-4aa7-81ad-fda0b75f6d63	n/a
CONTROL/42	TCGA-BH-A28Q-01	-0.1	94950f32-36e7-4fb2-b03e-05035e2e5c1c	n/a

We detected a total of 434,654 distinct AS events in the control set, including CA, AA, AD, and IR. From these, we discarded AS events with a mean absolute error >0.1 and reproducibility of <10 samples. The resulting set of 353,015 AS events were averaged to compile a unique control set.

Each SpliceTrap profile generated for the 57 SRSF1-overexpressing samples was compared to the unique control using SpliceDuo. Finally we selected AS events with FDR<0.1, averaged the AS change values, and annotated their reproducibility as the number of samples in which a significant AS was detected (Table S14).

RT-PCR validation

For 3-D cultured acini, cells were harvested as described above. Total RNA was extracted using Trizol (Invitrogen) from 2-D or 3-D cultured cells. Following DNAse I digestion (Promega), phenol-chloroform extraction, and ethanol precipitation, 1 µg of RNA was reverse-transcribed with Improm-II reverse transcriptase (Promega). Semi-quantitative, radioactive touch-down PCR (29 cycles) with [α-^{32}P]dCTP was used to amplify endogenous transcripts with the primers indicated in Table S3. PCR products were separated by 8% native PAGE, and bands were quantified with a phosphorimager (Fuji Image Reader FLA-5100). The ratio of each isoform was normalized to the sum of the different isoforms.

Motif discovery

60 CA-exons shared between 2-D and 3-D samples were used as a training set. A set of 53 CA-exons (shared between HeLa and 2-D or HeLa and 3-D samples) were used as a test set. By "shared" we mean pairs of CA-exons that follow either one of the following criteria: (i) FDR<0.1 in both samples and |AS change|>0.1; or (ii) FDR<0.1 for at least one sample, |AS change|>0.1 and |Δ(*sample1, sample2*)|<0.1 (Table S3). An additional group of 500 CA-exons unaffected by SRSF1 overexpression was used as a negative control.

We implemented *de-novo* motif discovery using three different methods:

1. psp-MEME: This is a variation of the MEME suite (http://meme.nbcr.net/meme/) which allows position-specific priors to assign a probability that a motif starts at each possible location in a sequence. The training set was used as 'positive' and the control as 'negative'. The motif size was set to 4-10, to accommodate the expected binding site of a typical RNA-binding protein.

2. Seeded psp-MEME: we utilized psp-MEME with an external seed. MEME implements the Expectation Maximization algorithm (EM), which begins by creating a function of the expectation (E), then maximizing (M) this function to the expected values of E, and utilizing the output of M as the input of the next E step, iteratively. In the case of MEME, the initial E step is given by a seed motif, representing the best initial guess of the program on the correct motif's position and structure. By using a seeded MEME, we allow an external source to provide the initial seed. The seeds we implemented were derived by two different programs:

    2.a. RSAT: input seeds were derived from sequence analysis with the oligo-diff function of the RSAT program (http://rsat.ulb.ac.be/), to detect enriched motifs in a 'positive'/'negative' comparison, in this case using the training and control sets.

    2.b. 6-mers Enrichment: For each of the 4096 possible RNA 6-mers, we computed their frequency in the training (N_{pos}) and control (N_{neg}) sets, and we then adjusted them by their average frequency in the shuffled sets (N_{pos_shuff} and N_{neg_suff}). These were obtained by randomly shuffling within each sequence, computing the frequency, and taking the average of 1000 iterations. Finally, the enrichment was represented as follows:

$$E_{6mer} = log_2(\frac{N_{pos} - N_{pos\_shuf}}{N_{neg} - N_{neg\_shuff}})$$

The resulting motifs were tested for test/control enrichment using the Fischer's exact test. The FDR procedure was used to adjust the p-values.

Motif comparison

For position weight matrix comparisons, we computed the dissimilarity score Kullback–Leibler (KL) as previously described (Roepcke et al., 2005). Briefly, the matrix representing the longest motif (length W) is shifted along the second matrix and KL scores are computed using the formula below. We imposed two constrains to eliminate uninformative shifts: (i) At least half of the shortest matrix must overlap with the longest matrix; and (ii) such overlap must be at least four nucleotides long. Finally, the lowest KL score among all informative shits (T) is reported as a measure of the dissimilarity between both matrices.

$$KL = \frac{1}{2}\frac{1}{W}\left[\sum_{j=1.}^{W}\sum_{b=A}^{T}\left[PWM1(j,b) \cdot \log(\frac{PWM1(j,b)}{PWM2(j,b)}) + PWM2(j,b) \cdot \log(\frac{PWM2(j,b)}{PWM1(j,b)})\right]\right]$$

PMW1(j, b) is the probability of finding base b at position j in Motif 1, W is the length of the longer motif, T is the number of informative shifts, and A is the set of all possible alignments for a valid shift.

Preparation of RNA and protein samples for NMR

SRSF1 RRM1+2 ORF corresponding to amino acids 1 to 196 was cloned in the pET24 expression vector. A GB1 tag was fused at the N-terminus of the protein to increase its solubility and stability. The protein was overexpressed at 37 °C in *E. coli* BL21 (DE3) codon plus cells in minimal M9 medium containing 1 g per liter $^{15}NH_4Cl$ and 4 g per liter glucose. Protein was purified by two successive nickel affinity chromatography (QIAGEN) steps using an N-terminal 6×His tag cloned between GB1 and the N-terminus of the protein, dialyzed against NMR buffer (150 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 50 mM L-Glu, 50 mM L-Arg, 0.05% β-mercaptoethanol, and 20 mM Na_2HPO_4 at pH=7) and concentrated to 0.16 mM with a 10-kDa molecular mass cutoff Centricon device (Vivascience). The GB1 tag was kept for all NMR and ITC titrations, as it does not influence the protein's interaction with RNA (Clery et al., 2013).

WT and mutant RNA oligonucleotides were purchased from Dharmacon, deprotected according to the manufacturer's instructions, lyophilized and resuspended in NMR buffer. NMR titrations were performed in the NMR buffer at 40 °C.

Isothermal titration calorimetry

ITC experiments were performed on a VP-ITC instrument (Microcal), calibrated according to the manufacturer's instructions. Protein and RNA samples were dialyzed against NMR buffer. The concentrations of proteins and RNAs were determined using optical-density absorbance at 280 and 260 nm, respectively. 20 mM of all the tested RNAs was titrated with 160 mM of recombinant protein by 40 injections of 6 ml every 5 min at 40 °C. Raw data were integrated, normalized for the molar concentration, and analyzed using Origin 7.0 software, according to a 1:1 RNA:protein ratio binding model. Due to RNA degradation during ITC titration, RNA concentrations were corrected from 20 to 10 mM to allow fitting of the ITC curves with an N value of 1.

Minigene reporter assays

The wild-type *SMN2* minigene construct was constructed by inserting an additional 1290 bp from intron 6 (corresponding to chr5:69370918-69372208) into the previously described pCI-neo-SMN2 (Hua et al., 2008). The SRSF1 wild-type and mutant motifs were introduced by site-directed mutagenesis using Phusion High Fidelity polymerase (Life Technologies) and primers containing the corresponding mutations (Sigma Aldrich). 0.5 µg of each plasmid was transfected in HEK293 cells using Lipofectamine 2000 (Invitrogen). Cells were collected after 48 hrs and total RNA was extracted using Trizol (Invitrogen). Following DNAse I digestion (Promega), phenol-chloroform extraction, and ethanol precipitation, 1 µg of RNA was reverse-transcribed with Improm-II reverse transcriptase (Promega). Semi-quantitative, radioactive PCR (26 cycles) with [α-^{32}P]dCTP was used to amplify endogenous transcripts with the following primers: SMN-T7-F2 5'-

TACTTAATACGACTCACTATAGGCTAGCCTCG-3' and SMN-EX8-75+5-R 5'-AAGTACTTACCTGTAACGCTTCACATTCCAGATCTGTC-3'. PCR products were separated by 5% native PAGE, and bands were quantified with a phosphorimager (Fuji Image Reader FLA-5100). The ratio of each isoform was normalized to the sum of the different isoforms.

Regulatory maps

Regulatory maps are binding-site density profiles used to uncover positional biases in splicing regulators in both exons and introns. It is arguably harder to find positional biases in exons than in introns, mainly because exons are relatively short and variable in size. If exonic sequences are indented towards the splice sites (SS), the sequence search space will decrease as we move away from the SS, and so will the probability of detecting informative signals. In contrast, introns are long enough, such that an arbitrary sequence search space can be imposed on them (e.g., 100 nucleotides next to the 3'SS). Exon size normalization is a possible solution to these irregularities, although it is debatable whether this procedure is biologically justified: i.e., does splicing regulation "sense" exon size in any way? Can cis-acting element scores be averaged? To overcome these limitations, we derived a Bayesian probability model based on the following guidelines and assumptions:

1. A splicing factor "F" binds and regulates the sequence target "S" through the motif "M" located in the substring (*p*, *q*). Here, *p* is the starting nucleotide of the sequence search space (i.e., the 3' or 5' SS) and *q* is any other nucleotide, either upstream or downstream from *p*. In this way, we implement an "expanding window" rather than a "sliding window," which drops the requirement for signals to converge within the boundaries of an arbitrary window size and implicitly assumes that regulatory motifs are positioned relative to the SS.

2. $P(F \rightarrow S_{(p,q)}|M_{(p,q)})$ is the observed likelihood of M (e.g., SRSF1 binding sites) in a set of putative targets S of F (i.e., targets of SRSF1). Instances of M were predicted by mapping regulatory motifs using the program SFmap (Akerman et al., 2009; Paz et al., 2010). SRSF1 targets were selected based on our RNA-seq data analysis. $P(F \rightarrow S_{(p,q)}|M_{(p,q)})$ is comparable to motif density units used in previous studies.

3. Most splicing regulatory maps reported so far, show a "control" curve, alongside the $P(F \rightarrow S_{(p,q)}|M_{(p,q)})$ curve. This control is included to report nucleotide content biases in the sequence search space. However, it can be visually challenging to interpret discrepancies between two highly irregular curves, such as $P(F \rightarrow S_{(p,q)}|M_{(p,q)})$ and the control. To generate a single informative curve, we integrated the control in the model as a marginal probability function $P(F \rightarrow S_{(p,q)}|\neg M_{(p,q)})$ which accounts for the instances of M detected with SF-map in 500 randomly selected cassette exons not predicted as targets of F by RNA-seq data.

4. We modeled the prior probability of M assuming that P(M) at the interval (*p, q*) is similar to the posterior probability of $P(M_{(p,[q-1])}|F \rightarrow S_{(p,[q-1])})$. This is plausible, given that splicing-factor binding sites are 5-9 nt long, so the binding probability should vary little between *q-1* and *q*. In addition, this procedure waves the necessity of curve smoothing.

5. Finally, $P(\neg M_{(p,[q-1])})$ is the complementary of the prior probability $P(\neg M_{(p,[q-1])}) = 1 - P(M_{(p,[q-1])})$.

The complete model is defines as follows:

$$P(M_{(p.q)}|F \rightarrow S_{(p.q)}) = \frac{P(F \rightarrow S_{(p,q)}|M_{(p,q)})\, P(M_{(p,[q-1])})}{P(F \rightarrow S_{(p,q)})}$$

$$P(F \rightarrow S_{(p,q)}) = (P(F \rightarrow S_{(p,q)}|M_{(p,q)})P(M_{(p,[q-1])})) + (P(M_{(p,q)}|\neg M_{(p,q)})P(\neg M_{(p,[q-1])}))$$

We analyzed three different datasets: (1) SRSF1 targets predicted by more than one RNA-seq experiment in this study; (2) AS changes from microarray data (Pandit et al., 2013); and (3) the intersection between RNA-seq (this study) and CLIP-seq targets (Sanford et al., 2008; Sanford et al., 2009) (Table S3). We queried these datasets with three SRSF1 motifs: (1) the present RNA-seq-derived motif; (2) a functional-SELEX-derived motif (Smith et al., 2006); and (3) CLIP-derived motifs (Sanford et al., 2008). We also constructed regulatory maps for SRSF2, SRSF3, SRSF5, SRSF6 and SRSF7 based on binding motifs from SFmap, using dataset 1. The following sequence blocks were analyzed: (1) up to 100 nt exonic sequences downstream of the 3' SS; (2) up to 100 nt exonic sequences upstream of the 5' SS; (3) 100 nt intronic sequences upstream of the 3' SS; (4) 100 nt intronic sequences downstream of the 5' SS; and (5) exonic sequences from the flanking upstream and downstream exons, as in (1) and (2). In all cases, the SS dinucleotides were excluded.

SMN2 and *ADAR2* mutational analysis

SMN2 exon7 and *ADAR2* exon 8 mutants (*m*) and wild-type (*wt*) sequences were screened for the present RNA-seq-derived SRSF1 binding motif using the Weighted Rank scoring function from the SFmap program (Akerman et al., 2009). For every mutant and at every nucleotide position, we calculated creation/loss scores of SRSF1 motif as the aggregated difference:

$$\text{creation/loss score} = \sum_{\forall m \in M} (S_m - S_{wt})$$

Where m is every mutant form the set of mutant set M and S is the SRSF1 motif score calculated for SFmap. Positive numbers indicate creation and negative numbers indicate loss of SRSF1 motifs at each position. The mutant sequences are shown in below.

SMN2 mutant sequences

Group	PMID	Mutation	PSI
SMN2	15272122	6C>T; 51A>C	100%

Rescue	16385450	6C>T; 7A>C	100%
	14766219	6C>T; 14A>C; 15U>G; 16C>G	97%
	15272122	6C>T; 25 G>U; 26G>U; 54A>G	97%
	16385450	6C>T; 11A>G	96%
	15272122	6C>T; 1G>U; 54A>G	95%
	14766219	6C>T; 4U>G; 5U>G; 11A>C; 12A>C; 12A>C	95%
	15272122	6C>T; 43del3	94%
	16385450	6C>T; 10C>G; 11A>G	93%
	15272122	6C>T; 45A>G	93%
	15272122	6C>T; 45A>C	92%
	14766219	6C>T; 12A>C	91%
	15272122	6C>T; 49del3	91%
	15272122	6C>T; 49U>G	88%
	15272122	6C>T; 46del3	87%
	15272122	6C>T; 51A>U	87%
	14766219	6C>T; 11A>C;12A>C;13A>C;14A>C; 15U>G, 16C>G	81%
	14766219	6C>T; 14A>C;15U>C	78%
	19716110	6C>T; 29G>C	70%
	10931943	6C>T; mDM1-T	70%
	14766219	6C>T; 35G>U	68%
	14766219	6C>T; 16C>G	63%
SMN2	14766219	6C>T; 11A>C;12A>C;13A>C	52%
	12604607	6C>T; MUTa	47%
	14766219	6C>T; 5U>G	47%
	12604607	6C>T; MUTf	39%
	14766219	6C>T; 3U>G;5U>G	31%
	12604607	6C>T; MUTe	30%
	12604607	6C>T; MUTc	29%
	10931943	6C>T; 8G>U;9A>U; 11A>U;12A>U	28%
	10931943	6C>T; 36A>U;38A>U; 45A.U;47A>U;48A>U	28%
	14766219	6C>T; 11A>C	28%
	15272122	6C>T; 40del3	25%
	14766219	6C>T; 4U>G	21%
	15272122	6C>T; 1G>U;25G>;26G>U;54A>G	19%
	15272122	6C>T; 47A>C	18%
	12604607	6C>T; MUTd	17%
	16385450	6C>T; 9A>U; 10C>G	10%
	14766219	6C>T; 3U>G;4U>G	8%

	14766219	6C>T; 3U>G	6%
	14766219	6C>T; 3U>G;4U>G;5U>G	5%
	16385450	6C>T; 9A>U; 10C>G	4%
	16385450	6C>T; 7A>U;8G>U;9A>U; 10C>U;11A>U	0%
	10931943	6C>T; 19A>U;20A>U;22G>U;23A>U;25G>U;26G>U	0%

ADAR2 mutant sequences

Group	PMID	Mutation	PSI
ADAR2 a	16793546	68A>C;69A>T;70A>C;71C>A; 72T>C;73C>G;74G>A	100%
	16793546	68A>C;70A>C;71C>G; 72T>A	100%
	16793546	63G>C;64C>T;66C>A;67T>C; 68A>G	100%
	16793546	57G>C;58C>T;59T>C;60G>A;61A>C;62A>G;63G>A	100%
ADAR2 b	16793546	67T>C;68A>T;69A>C;72T>G;73C>A	75%
	16793546	60G>C;61A>T;62A>C;63G>A;65C>G;66C>A	80%
	16793546	51T>C;52G>T;53G>C;54C>A;55T>C;56C>G;57G>A	80%
	16793546	72T>C;73C>T;74G>C;76G>C;77C>G;78T>A	85%
ADAR2 c	16793546	55T>C;56C>T;57G>C;58T>C	63%
	16793546	39G>C;41G>C;42G>A;43T>C;45C>A	63%
	16793546	74G>C;75A>T;76G>C;77C>A;78T>C;80G>A	70%
	16793546	69A>C;79A>T;72T>A	70%
ADAR2 d	16793546	42G>C;44G>C;45C>A;46A>C;47A>G;48T>A	33%
	16793546	48T>C;49C>T;50A>C;51T>A;521>C;54C>A	45%
	16793546	46A>T;47A>C;48T>A;50A>G;51T>a	45%
	16793546	WT	52%

SUPPLEMENTAL REFERENCES

Akerman, M., David-Eden, H., Pinter, R.Y., and Mandel-Gutfreund, Y. (2009). A computational approach for genome-wide mapping of splicing factor binding sites. Genome Biol *10*, R30.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol *11*, R106.

Belongie, S., Malik, J., and Puzicha, j. (2002). Shape matching and object recognition using shape contexts. IEEE Trans Pattern Anal Mach Intell.

Bookstein, F.L. (1989). Principal wraps: Thin-plate splines and the decomposition of deformation. IEEE Trans Pattern Anal Mach Intell.

Debnath, J., Muthuswamy, S.K., and Brugge, J.S. (2003). Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. Methods *30*, 256-268.

Han, A., Stoilov, P., Linares, A.J., Zhou, Y., Fu, X.D., and Black, D.L. (2014). De novo prediction of PTBP1 binding and splicing targets reveals unexpected features of its RNA recognition and function. PLoS computational biology *10*, e1003442.

Hua, Y., Vickers, T.A., Okunola, H.L., Bennett, C.F., and Krainer, A.R. (2008). Antisense masking of an hnRNP A1/A2 intronic splicing silencer corrects SMN2 splicing in transgenic mice. American journal of human genetics *82*, 834-848.

Joshia, A.A., Shattuckb, D.W., Thompsonb, P.M., and Leahya, R.M. (2007). Registration of cortical surfaces using sulcal landmarks for group analysis of MEG data. Int Congr Ser.

Liu, H.X., Zhang, M., and Krainer, A.R. (1998). Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev *12*, 1998-2012.

Pandit, S., Zhou, Y., Shiue, L., Coutinho-Mansfield, G., Li, H., Qiu, J., Huang, J., Yeo, G.W., Ares, M., Jr., and Fu, X.D. (2013). Genome-wide Analysis Reveals SR Protein Cooperation and Competition in Regulated Splicing. Molecular cell *50*, 223-235.

Paz, I., Akerman, M., Dror, I., Kosti, I., and Mandel-Gutfreund, Y. (2010). SFmap: a web server for motif analysis and prediction of splicing factor binding sites. Nucleic Acids Res *38*, W281-285.

Roepcke, S., Grossmann, S., Rahmann, S., and Vingron, M. (2005). T-Reg Comparator: an analysis tool for the comparison of position weight matrices. Nucleic Acids Res *33*, W438-441.

Sanford, J.R., Coutinho, P., Hackett, J.A., Wang, X., Ranahan, W., and Caceres, J.F. (2008). Identification of nuclear and cytoplasmic mRNA targets for the shuttling protein SF2/ASF. PLoS One *3*, e3369.

Sanford, J.R., Wang, X., Mort, M., Vanduyn, N., Cooper, D.N., Mooney, S.D., Edenberg, H.J., and Liu, Y. (2009). Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. Genome Res *19*, 381-394.

Singh, N.N., Androphy, E.J., and Singh, R.N. (2004). In vivo selection reveals combinatorial controls that define a critical exon in the spinal muscular atrophy genes. Rna *10*, 1291-1305.

Smith, P.J., Zhang, C., Wang, J., Chew, S.L., Zhang, M.Q., and Krainer, A.R. (2006). An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. Hum Mol Genet *15*, 2490-2508.

Wahba, G. (1990). Spline Models for Observational Data. SIAM.

Wu, J., Akerman, M., Sun, S., McCombie, W.R., Krainer, A.R., and Zhang, M.Q. (2011). SpliceTrap: a method to quantify alternative splicing under single cellular conditions. Bioinformatics *27*, 3010-3016.