

Molecular Cell, Volume 50

Supplemental Information

Genome-wide Analysis Reveals SR Protein

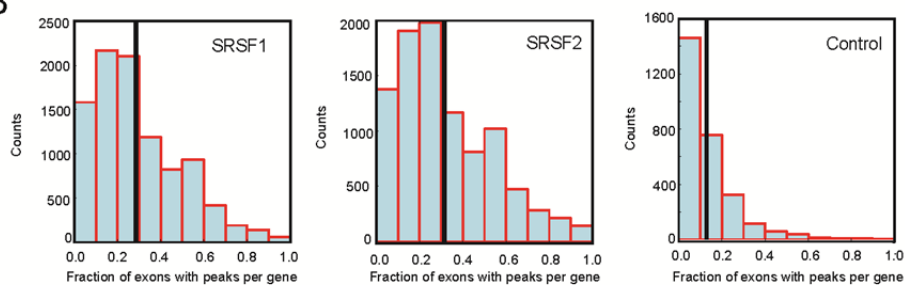
Cooperation and Competition in Regulated Splicing

Shatakshi Pandit, Yu Zhou, Lily Shiue, Gabriela Coutinho-Mansfield, Hairi Li, Jinsong Qiu, Jie Huang, Gene W. Yeo, Manuel Ares, Jr., and Xiang-Dong Fu

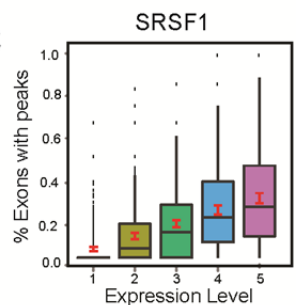
A

	Total Annotated	SRSF1		SRSF2	
		Peaks in	% bound	Peaks in	% bound
Coding with intron	18338	7385	40.27	7226	39.40
Coding intronless	2282	63	2.76	84	3.68
Non-coding	9253	1116	12.06	1075	11.62
miRNA	736	15	2.04	11	1.49
lincRNA	3345	172	5.14	177	5.29

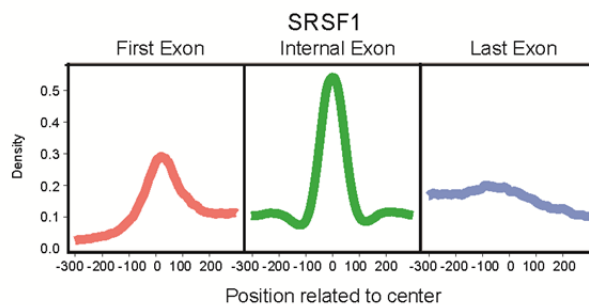
B



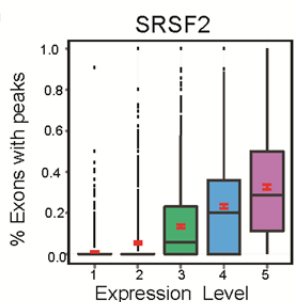
C



E



D



F

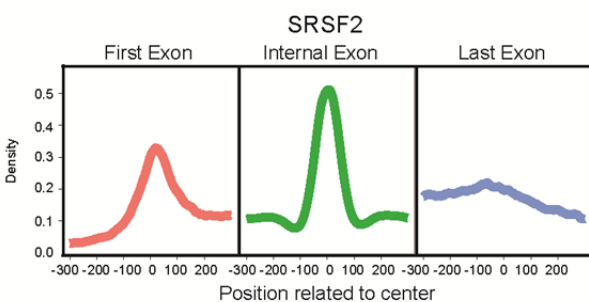


Figure S1. SR Protein Binding Landscape in the Mouse Genome, Related to Figure 1

(A) Statistics of SRSF1 and SRSF2 binding in different classes of genes. miRNA and lincRNA represent annotated subsets of non-coding RNA.

(B) Fraction of exons bound by SRSF1 or SRSF2. The transcripts evaluated had at least 3 exons and at least one SR protein binding peak. The vertical black line represents the average of exon coverage at the current tag density.

(C and D) Correlation between binding and gene expression for SRSF1 (C) and SRSF2 (D). Genes with different expression levels determined using array (built in exon junction arrays) are binned into five groups.

(E and F) SR protein binding on first exon (red), internal exons (green) and last exon (blue) for SRSF1 (E) and SRSF2 (F).

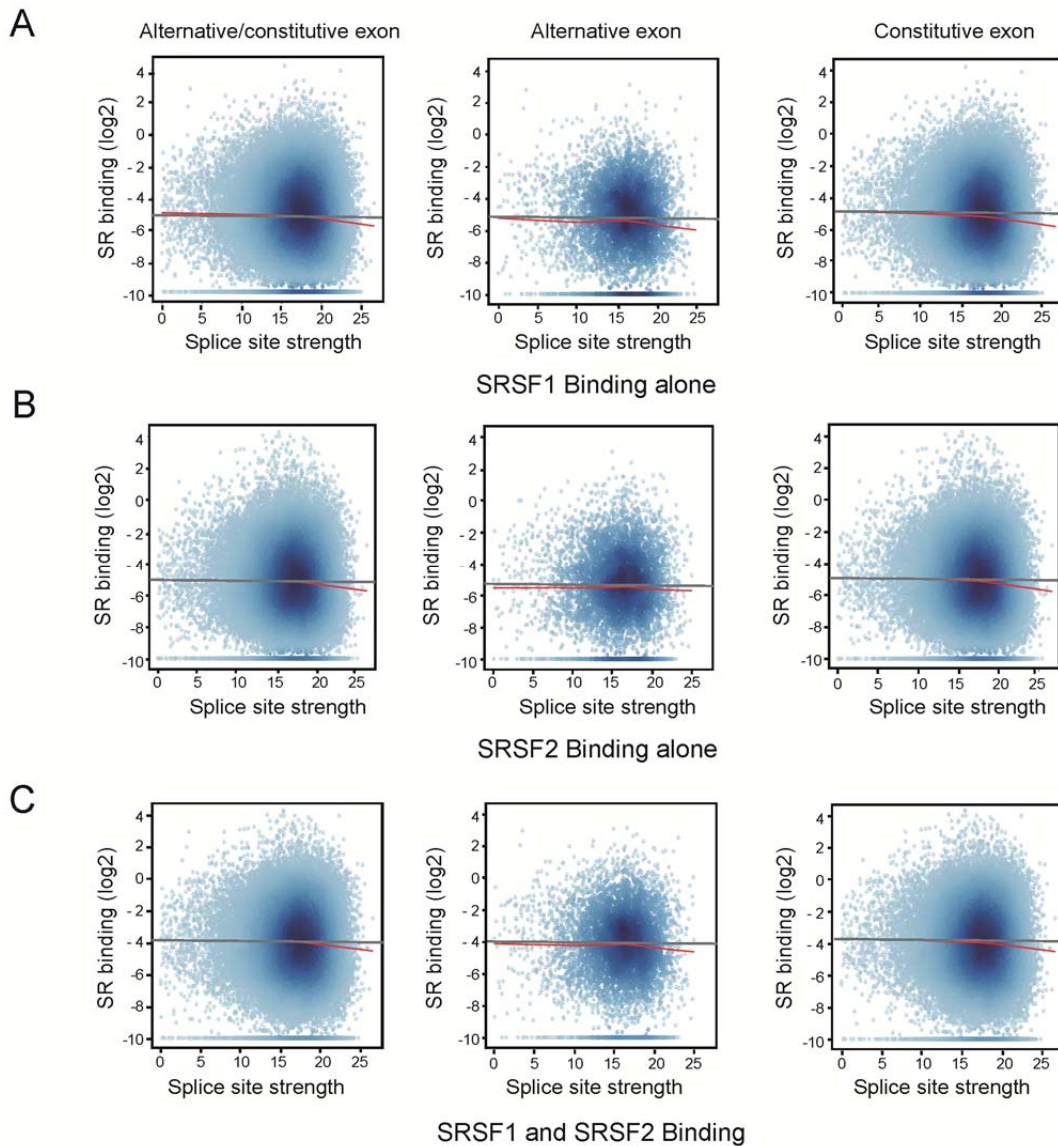


Figure S2. The Reverse Correlation of SR Binding with Splice Site Strength, Related to Figure 3

(A-C) Binding of SRSR1 (A), SRSF2 (B), and SRSF1plus SRSF2 (C) on all internal exons (left), alternative exons (middle), and constitutive exons (right) in relationship to splice site strength. Red line indicates a slight trend of decreased binding on exons with stronger splice sites, which was determined by the local polynomial regression fitting according to R function loess with default parameters. Grey line shows a hypothetical line without such trend.

Supplemental Experimental Procedures

Cell Culture, Immunoprecipitation, RNA Extraction, and Splicing Profiling

Previously generated SRSF1 and SRSF2 tet-repressible MEFs were maintained in DMEM plus 10% FBS without tetracycline (Clontech 631106). SRSF1 or SRSF2 protein was depleted by growing the cells for five days in the presence of 10 ug/ml of Doxycycline. For immunoprecipitation, cells were lysed according to the CLIP-seq protocol and the lysate was incubated with Dynabeads (Protein A + G) coupled to anti-HA antibody (Abcam ab91110). RNA for analysis on splicing arrays was isolated using the RNeasy kit (Qiagen). Reverse transcription was carried out with 1 µg total RNA per 25 µl reaction using the MMLV reverse transcriptase and random primers (Promega). 1 µl cDNA was used for semi-quantitative PCR or qPCR.

For surveying SR protein binding in the absence of another SR protein, we treated SRSF1-HA MEFs with siRNA against SRSF2. SRSF2 siRNA (Dharmacon onTargetplus J-044306-05) was transfected using siPORT Amine (Ambion) as transfecting agent. SRSF1-HA MEFs grown to 30-40% confluency were trypsinized and resuspended in cell culture medium and kept warm. 15 µl of siPORT was diluted in 500 µl Optimem media and incubated for 10 min, to which 30 µM siRNA diluted in 500 µl Optimem media was added and further incubated for another 10 min. The nucleic acid/siPORT amine complexes were transferred to a 10 cm culture plate and overlaid with the cell suspension. After 48 hrs, an aliquot of cells was used to harvest RNA and knockdown was confirmed by qPCR and Western blotting. The rest of the cells were crosslinked and used for CLIP-seq, as described (Yeo et al., 2009).

Analysis of splicing changes on exon junction arrays was performed as previously described (Du et al., 2010). RASL-seq analysis of splicing switches was carried out as detailed (Li et al., 2012; Zhou et al., 2012). A pool of oligonucleotides was designed to detect 3726 alternative splicing events conserved between humans and mice. RASL-seq primers were annealed to 1 µg of total RNA isolated from mock-depleted and SR protein depleted MEFs. After ligation, 5 µl eluted ligated oligos was used for PCR amplification and the bar-coded PCR products were sequenced on HiSeq2000. Data analysis was as described (Zhou et al., 2012).

Plasmid Construction and Transfection

CDC45I exon 3–exon 5 was amplified using 5'-GGATCCCCAGCTGTGAGCAGTTTTCA-3' and 5'-GGCTTTCACACACCTGTCCTAAGCTT-3'. The PCR product was purified and cloned into pcDNA 3.1 (+) at the BamH1 and HindIII sites. The mutant plasmids were generated using deletion primers adjacent to the binding site. *CCNL* exon 3–exon 5 was amplified using 5'-ATCGGTACCCAAAATAAGGGTTGTTAGACT-3' and 5'-GCCGATATCTCCGGTAACTTAACTCTCC -3'. The PCR product was purified and cloned into pcDNA 3.1 (+) at the BamH1 and HindIII sites.

SRSF2 MEF cells were transfected with the indicated plasmids using Lipofectamine 2000 (Life Technology). Cells were first grown to ~30% confluency in 24 well plates. Lipofectamine-DNA complexes were generated following manufacturer's protocol with 1µg plasmid and 2µl Lipofectamine in Optimem media. The complexes were overlaid on the cells in the media with sera for 6 hrs. Media was then changed to

serum containing with or without Dox (to induce SRSF2 depletion). After 48 hrs, the cells were harvested for RNA and protein analyses.

Bioinformatics Analysis

All sequenced tags were mapped to the mouse genome (mm9) by using Bowtie (version 0.12.7) (Langmead B, 2012). Two mismatches were allowed in the first 25nt region and only those tags with one unique hit in the mouse genome were kept. Multiple tags mapped to the same genomic position were collapsed into one or up to 4 tags to remove potential PCR amplicons. The UCSC genome browser refGene and knownGene were merged to generate gene references to calculate the tag distribution. Binding clusters (or peaks) were identified on annotated genes under a p-value cutoff $1e-5$ (Benjamini and Hochberg corrected) as described (Xue et al., 2009). To count tags mapped on exon-exon or exon-intron junctions, we extracted non-redundant 70nt junction sequences (35nt exonic and 35nt intronic flanking 5' or 3' splice site) from annotated genes and then mapped CLIP-seq tags to the sequence pools. All analysis of CLIP-seq data was done using custom Python and R scripts, BEDTools (Quinlan and Hall, 2010) and Kent source package (Kent et al., 2002). To detect SR protein binding in various gene categories, we used the merged reference genes from the UCSC genome browser refGene and knownGene (mm9) to extract intron-containing and intronless coding and non-coding genes. Primary miRNA annotation was based on miRBase version 18 (Kozomara and Griffiths-Jones, 2011) and lincRNA was according to Ulitsky and colleagues (Ulitsky et al., 2011). Overlapping isoforms of one gene were merged to one cluster before computation. To determine whether SR proteins show any preference

for first, last or internal exons, we calculated the binding for each SR protein on unique exons from genes containing at least three exons and having at least one SR protein binding peak.

Calculation of the Percentage of Exons per Gene Bound by SR Proteins

We selected genes with at least 3 exons and at least one SR protein CLIP-seq peak across the whole gene body to determine the correlation between SR protein binding and gene expression. Genes were first partitioned into 5 equal bins depending on the level of expression determined with microarray. SR protein binding was plotted as the fraction of genes bound in each bin.

To determine the overlap between SRSF1 and SRSF2 binding, the number of overlaps between CLIP-seq peaks was calculated for the two SR proteins in exons and introns separately. If a genomic interval had both SRSF1 and SRSF2 peaks with at least 50% overlap, it was calculated as an overlap. In each bin of gene expression, the ratio of overlap was calculated as the number of intervals with overlap to the total number of intervals bound by either SRSF1 or SRSF2.

Correlation between SR Binding and Splicing Signals

We first extracted genes having minimum 3 exons with at least one SR protein binding peak. Based on the number of CLIP-seq tags in the exons of each gene, we removed genes whose RPKM are either too high (top 5%) or too low (bottom 5%). For each gene, we calculated the maximum exonic tag density and normalized the tag density of each exon to that number. We then made a composite map using the exon center as

pivot on exons grouped by exon splice site strength. The 3' and 5' splice site (ss) scores for each exon were calculated using MatEnt (Yeo and Burge, 2004). The exons were equally partitioned to 3 bins (strong, middle and weak) based on the sum of 5'ss and 3'ss scores.

To further investigate the relationship between exon splice site strength and SR binding, we compared the gene expression normalized counts of SR CLIP-tags on an exon with the exon's splice site scores. Unique internal exons were extracted from refGene transcripts and the splice scores were calculated by using MaxEnt (Yeo and Burge, 2004). For each exon, SRSF1 and SRSF2 CLIP-seq tags mapped to it were counted and then normalized to exon length and total number of tags of the CLIP-seq to get a SR binding signal (reads per 100nt per million tags). To eliminate the effect of RNA abundance, the binding signal of each exon was normalized to its gene's expression from 3'-end RNA-seq (data not shown). To avoid possible bias brought by very high or very low expressed genes, 5532 genes whose RPKM ranged from 5 to 500 were used and this gave us a total of 59565 exons. To check possible different patterns between alternative and constitutive exons, the exons were partitioned into two groups, constitutive or alternative, according to whether they overlapped with the annotated alternative splicing events from the UCSC genome browser knownAlt track. Each point represents the normalized signal of SRSF1, SRSF2 or both SRSF1 and SRSF2 binding together along with the splice site score on the scatter plot, the darkness of the color represents overlapping of points. The trend was revealed by local polynomial regression fitting using R function loess with default parameters.

To detect correlation between SR protein binding and flanking intron length, three consecutive non-redundant exon blocks from genes with at least one SR peak were extracted and equally divided into 5 groups according to upstream or downstream intron length from smallest to longest. The fraction of blocks having SR protein binding on the middle exon was calculated for each.

To investigate SR protein binding in intronic regions, pseudo exons or decoy splice sites were identified in the mouse genome, as described (Tilgner et al., 2009). Briefly, gene clusters were built by masking all known exons including flanking 70nt intronic sequences in annotated genes, mRNAs and spliced ESTs. Repeat elements were also masked. For the remaining sequences, intervals with AG-GT pairs separated by 50 and 250nts were searched. Pseudo exons were defined as those having sum of 5' and 3' splice site scores larger than 17.05 (median of real exons); 5' decoy site as those having donor splice site score larger than 8.73 (median of real 5'ss); 3' decoy site as those having acceptor splice site score larger than 8.65 (median of real 3'ss). SR protein binding peaks were aligned by their center; the average number of pseudo exons, 5' and 3' decoy sites in the flanking 250nt regions within a 25nt window were counted.

Motif Search

Two different methods were used to identify potential SRSF1 and SRSF2 binding motifs. The first was to detect significantly enriched hexamers in SR CLIP-seq peaks by comparing the counts in the peak sequences and in control sequences randomly generated from similar background. We ran the program, as described (Xue et al., 2009), which calculated the Z-score for each hexamer and clustered significantly enriched hexamers

into different groups according to the similarity of sequences. The same analysis was carried out for SR protein binding peaks after filtering out those having more than 25% overlap with peaks in control CLIP sample. The second method used a similar pipeline, as recently outlined (Zhang and Darnell, 2011). Significantly frequent deletion sites were identified for SRSF1 and SRSF2 respectively based on permutation test. Program Homer was used to find significant motifs based on 10nt sequences on both sides of individual CIMS sites.

Correlation between SR Binding and Splicing Responses

RNA functional maps were generated for SR protein depletion-induced skipping and inclusion splicing events in 4 regions: 100nt exonic and 300nt intronic intervals flanking the upstream 5'ss and 3'ss, and downstream 5'ss and 3'ss. For the control set, similar information was extracted from three consecutive constitutive exons from the UCSC genome browser knownGene. For each map, only events that had at least one SR protein binding peak in any of the 4 regions were used. The splice sites were used as pivot to align the events together to calculate the average tag density at each position for each event.

Change in Binding Profiles after Knockdown

Each gene's expression was estimated using the total number of CLIP-seq tags located in it, and only those genes that showed no statistically significant change (FDR >0.1; fold-change <1.5) were kept. Genes that contain small transcripts and those with fewer than 3 exons were excluded, leaving 13,293 genes for further analysis.

To determine significant changes in SRSF1 binding upon SRSF2 knockdown, only those peaks that were reproducibly detected (60% overlap) in the two biological repeats for both control and knockdown were considered. In order to account for gene expression variations, we normalized the tags to the mean of the tag counts associated for each gene. We next performed *t*-test to evaluate the change in each binding site (peak) and p-value <0.05 and fold-change ≥ 2 were considered statistically significant. For the categories (increased or decreased binding), we made composite peak maps by aligning the center of each SRSF1 binding event and counting the number of SRSF2 peaks that overlap at every position in a 400nt window, and then calculated the ratio to the total number of SRSF1 locations.

Supplemental References

Du, H., Cline, M.S., Osborne, R.J., Tuttle, D.L., Clark, T.A., Donohue, J.P., Hall, M.P., Shiue, L., Swanson, M.S., Thornton, C.A., *et al.* (2010). Aberrant alternative splicing and extracellular matrix gene expression in mouse models of myotonic dystrophy. *Nature structural & molecular biology* *17*, 187-193.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome research* *12*, 996-1006.

Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* *39*, D152-157.

Langmead B, T.C., Mihai Pop, Steven L. Salzberg (2012). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* *10*.

Li, H., Qiu, J., and Fu, X.D. (2012). RASL-seq for massively parallel and quantitative analysis of gene expression. *Current protocols in molecular biology* / edited by Frederick M. Ausubel ... [et al.] *Chapter 4*, Unit 4 13 11-19.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* *26*, 841-842.

Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcarcel, J., and Guigo, R. (2009). Nucleosome positioning as a determinant of exon recognition. *Nature structural & molecular biology* *16*, 996-1001.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* *147*, 1537-1550.

Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.S., Zhang, C., Yeo, G., Black, D.L., Sun, H., *et al.* (2009). Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol Cell* 36, 996-1006.

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

Yeo, G.W., Coufal, N.G., Liang, T.Y., Peng, G.E., Fu, X.D., and Gage, F.H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature structural & molecular biology* 16, 130-137.

Zhang, C., and Darnell, R.B. (2011). Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nature biotechnology* 29, 607-614.

Zhou, Z., Qiu, J., Liu, W., Zhou, Y., Plocinik, R.M., Li, H., Hu, Q., Ghosh, G., Adams, J.A., Rosenfeld, M.G., *et al.* (2012). The Akt-SRPK-SR Axis Constitutes a Major Pathway in Transducing EGF Signaling to Regulate Alternative Splicing in the Nucleus. *Mol Cell* 47, 422-433.