Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts

Jeremy R Sanford, Xin Wang, Matthew Mort, et al.

Genome Res. published online December 30, 2008
Access the most recent version at doi:10.1101/gr.082503.108

Splicing Factor SFRS1 Recognizes a Functionally Diverse Landscape of RNA Transcripts

Jeremy R. Sanford[1†], Xin Wang[2], Mathew Mort[3], Natalia VanDuyn[4], David N. Cooper[3], Sean D. Mooney[5], Howard J. Edenberg[4,5], and Yunlong Liu[2]

1. Department of Molecular, Cellular and Developmental Biology, University of California Santa Cruz, 1156 High Street, Santa Cruz, California 95064

2. Department of Medicine, Division of Biostatistics, Indiana University School of Medicine, 635 Barnhill Drive, Indianapolis, Indiana 46202

3. Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff CF15 4XN, United Kingdom,

4. Department of Biochemistry and Molecular Biology,

5. Department of Medical and Molecular Genetics, Indiana University School of Medicine, 635 Barnhill Drive, Indianapolis, Indiana 46202

†To whom all correspondence should be addressed (email): sanford@biology.ucsc.edu

Running title: RNA Map for Splicing Factor SFRS1.

Key words: RNA binding protein; SR protein; alternative splicing; high-throughput sequencing; pre-mRNA splicing; mutation; bioinformatics

Abstract

Metazoan genes are encrypted with at least two superimposed codes: the genetic code to specify the primary structure of proteins and the splicing code to expand their proteomic output via alternative splicing. Here, we define the specificity of a central regulator of pre-mRNA splicing, the conserved, essential splicing factor SFRS1. Cross-linking immunoprecipitation and high-throughput sequencing (CLIP-Seq) identified 23,632 binding sites for SFRS1 in the transcriptome of cultured human embryonic kidney cells. SFRS1 was found to engage many different classes of functionally distinct transcripts including mRNA, miRNA, snoRNAs, ncRNAs and conserved intergenic transcripts of unknown function. The majority of these diverse transcripts share a purine-rich consensus motif corresponding to the canonical SFRS1 binding site. The consensus site was not only enriched in exons cross-linked to SFRS1 *in vivo*, but was also enriched in close proximity to splice sites. mRNAs encoding RNA processing factors were significantly over-represented, suggesting that SFRS1 may broadly influence the post-transcriptional control of gene expression *in vivo*. Finally, a search for the SFRS1 consensus motif within the *Human Gene Mutation Database* identified 181 mutations in 82 different genes that disrupt predicted SFRS1 binding sites. This comprehensive analysis substantially expands the known roles of human SR proteins in the regulation of a diverse array of RNA transcripts.

This manuscript is accompanied by four Supplementary Figures, one Supplementary Table and four database files. The database files can be loaded as custom tracks in the UCSC Genome Browser.

INTRODUCTION

Metazoan genomes are encoded with multiple overlapping layers of information required for the precise control of gene expression. The splicing code has co-evolved with the genetic code and regulates the post-transcriptional expression of protein coding genes (for review see Wang & Cooper 2007). In the nucleus, splicing is required to remove intervening sequences (introns) from precursor messenger RNAs (pre-mRNAs) and to correctly join protein-encoding regions (exons) together. Inclusion of an exon into the mature mRNA is regulated by *cis*-acting RNA elements known as exonic or intronic splicing enhancers and silencers (ESE, ISE and ESS, ISS, respectively) that function to recruit *trans*-acting RNA binding proteins. In the cytoplasm, these same RNA elements are decoded by tRNAs and the ribosome in order to template protein synthesis. Alternative splicing allows a single gene to express many different protein isoforms by including all, some or none of a specific exon sequence in the mRNA (for review see Maniatis and Tasic 2002). Current estimates suggest that at least 70% of protein coding genes undergo alternative splicing (Wang & Cooper 2007). However, understanding how these events are regulated and coordinated represents a major challenge.

Classification of functional *cis*-acting RNA elements on a global scale is required to begin the arduous task of defining the specific outputs from every human gene (For review, see Wang and Burge 2008). Combinations of elegant computational studies and biochemical assays are now beginning to address this important problem in gene regulation. A variety of bioinformatics strategies have identified hundreds of putative *cis*-acting sequences that are enriched with regulated exons or introns. These studies have included sequences that function as independent, non-redundant regulatory units and those that work together and have therefore co-evolved (Fairbrother et al. 2002; Zhang and Chasin 2004; Zhang et al. 2005; Wang et al. 2006; Xiao et al. 2007; Friedman et al. 2008). One *caveat* is that the cognate RNA binding proteins often escape identification. By contrast, biochemical studies such as the selected

evolution of ligands by exponential enrichment (SELEX) can reveal the nucleotide sequences recognized by specific RNA binding proteins (Tuerk and Gold 1990; Tacke and Manley 1995; Liu et al. 1998; Cavaloc et al. 1999; Smith et al. 2006). However, these sequences are often degenerate and lack sufficient specificity to reveal the global organization of protein-RNA interactions. One of the most powerful methodologies used to examine interactions of RNA binding proteins with their cognate targets is the RNA immunoprecipitation-microarray experiment (RIP-Chip). RIP-Chip is similar to chromatin IP-microarray analysis (ChIP-Chip) with the exception that it is RNA-protein rather than DNA-protein interactions that are assayed (Tenenbaum et al. 2002; Sanchez-Diaz and Penalva. 2006). RIP-Chip is a robust method that has been applied to both yeast and metazoan systems and can reveal relationships between transcripts and regulatory RNA binding proteins. Interpretation of RIP-Chip experiments however requires caution owing to several technical considerations. In the absence of cross-linking reagents, RNA binding proteins are free to dissociate from their endogenous RNA targets and re-associate with higher affinity binding sites, thereby giving rise to a risk of false discovery (Mili and Steitz 2004). Formaldehyde cross-linking of intact cells can preserve *in situ* protein-RNA interactions but can also induce protein-protein cross-links, thereby increasing the likelihood that RNA targets associated with other RNA binding proteins may be inadvertently co-purified. Despite these issues, numerous studies employing RIP-Chip have suggested that RNA binding proteins exhibit distinct binding specificities (Brown et al. 2001; Hieronymus and Silver 2003; Gerber et al. 2004; Gama-Carvalho et al. 2006; Olson et al. 2007). These observations have led to the formulation of the hypothesis that the coordinated post-transcriptional control of functionally related transcripts is organized by specific RNA binding proteins (Keene 2007).

The RIP-related cross-linking immunoprecipitation (CLIP) method sidesteps several of the pitfalls described above (Ule et al. 2003; Ule et al. 2005a). The primary advantages of this method over the other standard RNA IP methods include the following: (1) photo cross-linking of intact cells preserves the *in situ* RNA binding specificity, (2) partial RNase digestion liberates the

protein binding site from the full length transcript, (3) stringent purification conditions decrease contamination thereby enhancing the specificity of the assay, and (4) cloning and sequencing of purified RNA fragments directly identifies both the genomic locus from which the RNA transcript was derived, and the region recognized by individual RNA binding proteins. CLIP analysis has been successfully performed in a study of the murine RNA binding protein NOVA, a neural gene-specific alternative splicing factor involved in the human neurological disorder, paraneoplastic opsoclonus myoclonus ataxia (POMA; Ule and Darnell 2006). CLIP analysis of NOVA identified a network of pre-mRNAs encoding proteins involved in post-synaptic functions that are regulated by NOVA at the level of alternative splicing. These data allowed the creation of a 'genomic map' capable of predicting alternative splicing events based upon the NOVA binding position (Ule et al. 2006).

Splicing Factor Arginine/Serine-Rich 1 (SFRS1, also known as SF2, ASF, ASF/SF2 and SF2/ASF) is a highly conserved, essential pre-mRNA splicing factor with dual functions in constitutive and alternative splicing (Ge et al. 1991; Krainer et al. 1991). SFRS1 is a member of a large protein family known as the Serine and Arginine-rich proteins (SR proteins). SR proteins have a modular domain structure comprising one or two amino-terminal RNA Recognition Motifs (RRMs) and a carboxyl-terminal domain composed almost exclusively of alternating serine and arginine repeats. SR proteins function at early stages of spliceosome assembly and play an important role in specifying splice site selection (for review see Lin and Fu 2007). In cultured cells, SFRS1 shuttles between the nucleus and the cytoplasm and also participates in post-splicing RNA processing reactions including mRNA export, stability, nonsense-mediated decay and translation (Caceres et al. 1998; Huang and Steitz 2001; Lemaire et al. 2002; Sanford et al. 2004; Zhang and Krainer 2004). Although many of the biochemical roles of SR proteins in pre-mRNA splicing can be carried out by other family members, SFRS1 is absolutely required during both murine and nematode embryogenesis and for the maintenance of genome stability in mammalian cell culture models (Wang et al. 1996; Longman et al. 2000; Li and Manley 2005;

Xu et al. 2005). *SFRS1* is a proto-oncogene, located at 17q21.3-q22, which is amplified in many types of human tumor (Karni et al. 2007). The comprehensive identification of SFRS1 mRNA targets promises to improve our understanding of the diverse biological roles of SFRS1.

Despite numerous *in vitro* and *in vivo* studies, the RNA binding specificity of SFRS1 is not fully understood. Here we combine CLIP with high-throughput sequencing (CLIP-Seq). CLIP-Seq proffers a comprehensive and cost-effective system for the identification of biologically relevant *cis*-acting RNA elements recognized by specific RNA binding proteins in the context of their cellular environment. Our data provide an unprecedented evaluation of the RNA binding specificity of the essential splicing factor SFRS1. CLIP-Seq identified a purine-rich consensus motif, which is present in a majority of exonic, intronic, ncRNA and intergenic transcripts co-precipitated by SFRS1. Analysis of exonic RNA fragments bound by SFRS1 revealed an enrichment of binding sites between 21-40 nt from the 5' or 3' splice sites. This sequence and its positional specificity were then used to predict SFRS1 binding sites that have been disrupted by mutations causing human inherited disease. This analysis identified 181 mutations (in 82 different genes) associated with genetic disease that abolished putative binding sites for SFRS1. These results suggest that defective protein-RNA interactions may play a rather broader role in human inherited disease than has previously been anticipated.

RESULTS

Cross-linking Immunoprecipitation of SFRS1

We used CLIP to identify *cis*-acting RNA elements recognized by splicing factor SFRS1. Three independent cultures of human embryonic kidney cells (HEK293T) exposed to UV and one control culture without UV-irradiation were used to prepare whole cell extracts as previously described (Sanford et al. 2005). Following DNase and RNase treatment of the extracts, SFRS1 was precipitated using mAb 96 (Hanamura et al. 1998). Immunoprecipitation of SFRS1 was confirmed by western blot analysis. Precipitation of SFRS1 is dependent upon antibody and independent of UV irradiation (Fig 1A; compare lanes 2, 3, 5, 6, 7 with lane 4). Radiolabeled SFRS1-RNA complexes were visualized by autoradiography in parallel with the western blot analysis. Precipitation of the ^{32}P-labeled complexes by the SFRS1 antibody was dependent upon UV cross-linking (Fig 1B, compare lanes 1,2 with 4-6). Surprisingly, SFRS1 can be radiolabeled in the absence of UV irradiation. However, incorporation of ^{32}P is not dependent upon T4PNK, indicating that this signal is unrelated to SFRS1-RNA complexes (Fig. 1B, lanes 1,2). The bracket in Figure 1B designates the region purified from nitrocellulose membranes and corresponding to a 10-15 kDa increase in the apparent molecular weight of free SFRS1 (Fig. 1B, compare bracket to arrowhead marking the expected migration of SFRS1). RNA extracted from the nitrocellulose membrane was then amplified by reverse transcription polymerase chain reaction (RT-PCR). For comparison, non-selected input RNA was purified and amplified from RNase-treated HEK293T whole cell extracts. We directly sequenced CLIP-derived and input amplicons using the 454 Genome Sequencer FLX system (Margulies et al. 2005). As expected from the mobility of the SFRS1-RNA complexes, the majority of purified RNA fragments were between 45-65 nt in length (Fig. 1C).

High-Throughput Sequencing

SFRS1-bound RNAs were analyzed in four independent CLIP-Seq experiments. In total, 932,152 reads were obtained from SFRS1-bound RNA. In addition, 670,448 reads from non-selected input RNA were generated from three independent experiments for comparison (Table 1). 99.8% and 91.6% of sequences derived from CLIP and input amplicon libraries contained information from both 5' and 3' RNA linkers, indicating that the vast majority of RNA fragments were sequenced completely. The sequences corresponding to RNA fragments were then filtered, based upon a minimum allowable size of 30 bp and no ambiguous base calls. After this filtering step, all redundant amplicon sequences were removed leaving a pool of 135,318 and 218,108 CLIP- or input-derived RNA fragments, respectively. The unique RNA fragments were aligned to the human genome using the Blast-like Alignment Tool (BLAT; Kent 2002). RNA fragments derived from CLIP or input RNA samples mapped to 58,953 and 3,374 loci respectively, indicating that the CLIP sequences were far more diverse than those derived from the input samples. Overlapping clusters of unique amplicons were used to define the contiguous sequence blocks (henceforth referred to as "blocks"). Blocks derived from non-selected total RNA served as a reference sample. We assume that these unselected blocks correspond to the most abundant transcripts and may be a potential source of contamination in the CLIP-Seq dataset. Therefore, blocks common to both the input and CLIP-Seq were excluded from further analysis, despite the possibility that some very abundant RNAs might actually bind to SFRS1. By applying this stringent filter, the abundant class may have been lost. However, this conservative approach ensured that only extensively enriched sequences would have been captured. Only 5% of the CLIP-derived blocks were present in the reference set.

SFRS1 consensus motif

Identification of the consensus site for SFRS1 is important in order to understand mechanisms of splice site selection. We used the motif finding algorithm *Multiple EM for Motif Elicitation* (MEME; Bailey and Elkan 1995) to identify RNA-binding sequences shared by sequenced fragments that cross-link to SFRS1 *in vivo*. There were 681 unique sequence blocks

present in at least three of the four CLIP-Seq experiments and absent from the input sequences. We randomly split this set in half, using 340 to train the MEME algorithm, and holding the remaining 341 blocks in reserve as part of a gold standard dataset. After picking a single representative amplicon at random from each sequence block, MEME identified a purine-rich octamer containing a GAAGAA core (Fig. 2A). This motif is similar to several SFRS1 recognition sites previously identified by binding SELEX experiments and by mutational analysis of splicing enhancers in the fibronectin extra-domain A and cardiac troponin T alternative cassette exons (Caputi et al. 1994; Ramchatesingh et al. 1995; Tacke and Manley 1995). Computational methods searching for ESEs have also identified and validated the GAAGAA sequence as a functional splicing enhancer (Fairbrother et al, 2002). We then calculated a positional weight matrix (PWM) from the SFRS1 consensus motif. The predictive power of the PWM was evaluated using two different statistical plots: the Accuracy curve, which calculates the accuracy of the binding site prediction as a function of matching score cutoff thresholds and the Receiver Operating Characteristic curves (ROC), which evaluate sensitivity and specificity of the binding site model (Fig. 2B,C). For each plot, we used the PWM to scan a gold standard dataset, consisting of amplicons from the remaining 341 blocks as a positive component and an equal number of sequences picked at random from intergenic deserts as a negative component. These measurements established the maximum accuracy, sensitivity and specificity of PWM as 78%, 81% and 77%, respectively. We therefore consider that the consensus binding site model presented in Figure 2 has a high probability of correctly identifying SFRS1 binding sites *in silico*.

Classification of SFRS1 binding sites

The UCSC Known Gene and Rfam databases were used to annotate each sequence block. Figure 3A depicts the strategy used for annotation. A total of 23,699 blocks were identified in the SFRS1 CLIP-Seq experiment (Fig. 3B). The majority (73%) of these blocks mapped to loci annotated as protein coding genes. Of the 17,365 blocks present within protein coding genes, 83% were associated with exonic sequences. We sub-classified these exon-

associated blocks into those that were contained exclusively within a single exon (10,532 blocks; 60%), those that spanned an exon-exon junction (2,245; 13%), those that spanned an exon-intron boundary (1,065; 6%) and those contained within an intronless gene (681; 4%). The remaining 17% (2,911) of blocks mapping to protein coding genes were present within introns. Three files containing all of the genomic coordinates of blocks targeted by SFRS1, blocks present in the input sample and the positions of SFRS1 consensus sites can be found in the Supplementary Materials online.

SFRS1 *binding sites associated with alternative splicing*

SFRS1 is a well-characterized splicing factor with roles in the regulation of alternative splicing (for review see Lin and Fu 2007). We extracted a non-redundant version of the AltEvents database from the UCSC Genome Browser and used it to classify SFRS1 target exons based upon their relationship to constitutive or alternative splicing (Table 2). Some 88.8% (12,304) of the exonic binding sites of SFRS1 were localized to constitutive exons whereas only 11.2% (1,538) had evidence of alternative splicing in this database. Alternative cassette exons were the single most abundant classification followed by exons containing alternative 3' or 5' splice sites. Retained introns were the least common. The observed levels of these exons within the SFRS1 CLIP-Seq dataset differed significantly from the expected levels of each classification based upon their ratios to all exons annotated by the UCSC Known Gene database (Table 2). Both cassette exons and retained introns were under-represented in the pool of SFRS1 targets (Fisher's Exact test $P < 1.6 \times 10^{-5}$ and 6.5×10^{-4}, respectively). By contrast, exons with alternative 5' and 3' splice sites were enriched in the CLIP-Seq data relative to the genome (Fisher's Exact test $P < 9.5 \times 10^{-14}$ and 6.4×10^{-9}, respectively). Recent work from Biamonti and co-workers (Ghigna et al. 2005) suggests that binding sites for SFRS1 in constitutive exons may regulate the inclusion or exclusion of adjacent cassette exons. We found that SFRS1 bound in equal proportions to exons immediately upstream or downstream of cassette exons (5' or 3' adjacent exons, respectively). We observed a significant enrichment in

the CLIP-Seq dataset relative to the human genome for binding sites located within 5' and 3' adjacent exons (Fisher's Exact test $P < 1.3x10^{-10}$ and $5.4x10^{-11}$, for upstream and downstream exons, respectively).

Intronless Genes

Not all intragenic blocks fell within intron-containing genes; 681 binding sites in 332 intronless mRNAs were identified in this experiment. 100 blocks mapped to 30 different histone genes. Several other interesting intronless genes identified as SFRS1 targets include those encoding transcription factors, such as *JUN*, *JUND*, *SOX4*, *SOX12*, *FOXC1,* and those encoding post-translational regulators, such as *UBC*, *SUMO1* and *SUMO2*. These findings are consistent with roles for shuttling SR proteins in the nuclear export of histone 2a mRNA (Huang and Steitz 2001). However, our experiment suggests that SFRS1 regulates the post-transcriptional expression of many functionally diverse intronless genes. One subtle difference between intronless and intron-containing targets is a modest enrichment of SFRS1 binding sites within the UTRs of intronless genes relative to spliced transcripts (Supplementary Fig. 1). Some 43% of blocks mapping to intronless genes were present in UTRs as compared to 25% of blocks within spliced transcripts, respectively (Fisher's Exact test $P<0.004$; Supplementary Fig. 1). However, the biological significance of these data is unclear.

Intergenic and noncoding RNA targets

A significant proportion (26%) of all blocks mapped to regions not annotated as protein coding genes. According to the Rfam database (Griffiths-Jones et al. 2003), only 80 of the 6,218 intergenic blocks have been annotated as non-coding RNA genes. Among the ncRNAs represented in the SFRS1 CLIP-Seq data are 23 snoRNAs, 3 microRNA precursors, and the 5.8S rRNA, *XIST* and *MALAT1* RNAs. It is possible that the remainder may correspond to as yet undiscovered ncRNAs. Phylogenetic conservation was used to evaluate the significance of SFRS1 binding sites in RNA transcribed from intergenic loci. The conservation scores, or

PhastCons scores, were downloaded from the UCSC Genome Browser and reflect the overall conservation among seventeen vertebrate species (Felsenstein and Churchill 1996; Siepel et al. 2005). We determined the mean conservation score for each nucleotide within 1,200 bp of the center of each intergenic SFRS1 binding site. These data were then compared to intergenic regions randomly selected from the human genome (Supplementary Fig. 2). The majority of SFRS1-bound RNA transcripts derived from intergenic regions were found to be highly conserved across multiple vertebrate lineages, suggesting a high degree of negative selective pressure at these sites.

SFRS1 binds to a subset of functionally related mRNAs

Keene (2007) has proposed that RNA binding proteins may coordinately regulate the post-transcriptional expression of functionally related genes. To determine if particular classes of gene may be influenced by SFRS1, gene ontology (GO) annotations were assigned to SFRS1 mRNA targets present in three out of four CLIP-Seq experiments using EASE (Expression Analysis Systematic Explore; Hosack et al. 2003). Annotation enrichment was ascertained by computing the EASE score for each GO classification. The top 10 (based upon Holm-corrected EASE scores; Hosack et al. 2003) most enriched classes of mRNA bound by SFRS1 are shown in Figure 4A. Based upon biological process annotations, involving several redundant layers, it is clear that the most enriched SFRS1 RNA targets encode proteins involved in gene expression in general and RNA processing specifically. In order to avoid potential bias due to highly abundant transcripts, we also defined the GO of the non-selected, RNase-digested input RNA. The enrichment of each annotation term in both the CLIP-Seq and non-selected input RNA relative to the human genome was calculated and ranked using EASE Scores. Relative to both the genome and non-selected input RNA, SFRS1 target mRNAs were found to be highly enriched for genes involved in biological processes related to pre-mRNA splicing, RNA processing and ribosome biogenesis (Fig. 4B). By contrast, mRNAs encoding proteins functioning in processes such as chromatin and nucleosome assembly did not differ

significantly between SFRS1 RNA targets and non-selected RNase-treated Input RNA samples, despite being enriched relative to the genome as a whole. As expected, the non-selected RNase-treated Input RNA samples contained a much wider array of GO-terms than the CLIP-Seq targets (not shown).

Validation of SFRS1-RNA interactions

CLIP-Seq analysis of SFRS1 identified thousands of potential RNA targets. As with any high throughput method, it is necessary to gauge the accuracy of the CLIP-Seq targets using a secondary assay. In order to validate the SFRS1-RNA interactions, we used the RNA-immunoprecipitation (RIP) assay to determine if 78 RNA transcripts selected at random from the CLIP-Seq dataset interact with SFRS1 under native conditions. SFRS1 was found to be efficiently and specifically immunoprecipitated from whole cell extracts prepared from HEK293T cells (Fig. 5A, compare lanes 3 and 4). The increased mobility of SFRS1 observed in lane 4 is due to partial dephosphorylation of SFRS1 and can be blocked by inclusion of phosphatase inhibitors during the IP incubation (data not shown). Co-precipitated RNA was then isolated from input, control IP and anti-SFRS1 IP samples and analyzed by RT-PCR (Fig. 5B and Supplementary Fig. 3). Of the 78 randomly selected target transcripts, 58 were detectable in both the input and anti-SFRS1 IP samples but not in RNA isolated from the control IP. These validated targets include many intergenic transcripts, demonstrating that these unannotated RNA transcripts are associated with SFRS1 *in vivo*. Nine interactions were classified as false positive on the basis that the transcript was either detected in both the control and SFRS1 IP (see Supplementary Fig. 3, ZNF66) or present only in the input sample but absent from the IP. It is possible that some non-validated interactions between SFRS1 and target RNAs were undetectable under non-cross-linking conditions. Eleven RNA targets were classified as technical failures since no detectable signal was observed in any RNA sample. In total, approximately 73% of randomly selected SFRS1 target transcripts could be validated by the RIP

assays (Fig. 5C). These data were in good agreement with the statistical evaluation of the SFRS1 consensus motif which correctly recognized ~78% of target RNA fragments.

The SFRS1 consensus motif is enriched near the boundaries of exons identified by CLIP-Seq

Our data suggest that the robust consensus motif presented in Figure 2 represents the canonical binding site for SFRS1. However, only a small proportion of blocks were used to generate the model. We next asked if the consensus motif was enriched in blocks derived from exon sequences bound by SFRS1 relative to randomly selected exonic sequences. The binding site density (number of binding sites exceeding the matching score cutoff threshold per nucleotide) for 8,693 blocks present in constitutive exons and 426 blocks within alternative cassette exons was calculated. The average binding site density was compared to an equal number of 55 nt sequence blocks selected at random from non-targeted exons. Figure 6 indicates that the consensus site is highly enriched within exonic sequence blocks captured by CLIP-Seq as compared to 55 nt blocks selected at random from exons across the genome ($P<2.2 \times 10^{-22}$, Wilcoxon test). Additionally, these calculations indicate that each sequence block contains on average ~1.7 consensus binding site. The number of binding sites for SFRS1 within exonic sequence blocks ranges from as few as 0 matches to the consensus motif to as many as 16 sites. We also observed a modest enrichment in consensus site density in alternative cassette exons versus constitutive exons cross-linked to SFRS1, suggesting that on average there are slightly more binding sites for SFRS1 in alternative exons ($P<0.005$, Wilcoxon test).

The spatial relationship between binding site positions and splice sites can provide important mechanistic insights into molecular functions of RNA-binding proteins (Ule et al. 2006). Biochemical studies suggest that SR proteins function at early stages of spliceosome assembly and promote recognition of splice sites (Fu 1993; Kohtz et al. 1994; Graveley et al. 2001; Shen and Green 2004). Indeed, the proximity of SR protein binding sites to splice sites is

positively correlated with the *in vitro* splicing efficiency of reporter pre-mRNAs (Graveley and Maniatis 1998). To determine if the distribution of SFRS1 binding sites is restricted to specific positions within exons, we scanned experimentally observed blocks, full-length exons targeted by CLIP-Seq and randomly selected exons from the genome with PWMs corresponding to the SFRS1 consensus site and the reverse complement of the motif. The distance (in base-pairs, bp) from each consensus site was measured to the nearest 5' or 3' splice site. The frequency of binding sites at specific nucleotide positions was determined and adjusted for the uneven length distribution of human exons (Majewski and Ott 2002). The highest frequency of SFRS1 consensus sites were located in blocks near exon-intron boundaries, specifically between 20-41 bp from 5' and 3' splice sites (Fig. 7A, blue lines). By contrast, the anti-sense PWM failed to detect a significant number of matching sites within the CLIP-Seq blocks (Fig. 7A, red lines). We performed two additional calculations to ensure that the observed positions of SFRS1 consensus sites were not biased by our analytical methods. First, we compared the distribution of sites identified by the sense and anti-sense PWMs in full-length exons identified by CLIP-Seq and randomly selected exons from the human genome (Fig. 7B, blue and red lines, respectively). Again, SFRS1 consensus sites identified by the sense PWM exhibited a clear bias towards the exon boundaries. By contrast, sites identified by the anti-sense PWM are evenly distributed across CLIP-Seq exons. Both PWMs identified matching sequences in randomly selected exons. However, only a weak bias towards exon boundaries was observed in the case of the sense PWM, whereas the antisense motif showed no apparent bias (Fig. 7C, blue and red lines, respectively). These data demonstrate that although consensus SFRS1 binding sites are present throughout exon sequences, experimentally observed blocks, presumably containing engaged binding sites for SFRS1, are enriched at fixed positions relative to splice sites. The distribution of SFRS1 consensus sites in alternative cassette exons was found to be virtually identical to those in constitutive exons (not shown). Finally, we directly examined the positional distribution of the amplicons themselves, relative to splice sites. For each amplicon sequence that was derived from an exon or exon-intron boundary, we measured the distance in

base-pairs (bp) from the midpoint of the amplicon sequence to the nearest 5' or 3' splice site. As with previous calculations, we determined the frequency of amplicon midpoints in 10 bp bins extending away from the splice sites. Each amplicon midpoint was counted only once with respect to either a 5' or 3' splice site, thereby ensuring that the genomic coordinates of each midpoint measurement contributed only to the nearest splice site. The amplicon midpoint frequencies were directly compared to an equal number of randomly selected points from exon sequences extracted from the human genome and from amplicon sequences derived from input RNA samples. The average frequencies for each bin were calculated over 40 replicate samplings. Figure 7D demonstrates that the midpoints of CLIP-Seq amplicons were restricted to specific positions relative to 5' and 3' splice sites (Fig. 7D, blue lines). Amplicons picked from the input samples also showed a slight positional bias relative to 5' and 3' splice sites (Fig. 7D, orange lines). However, CLIP-Seq amplicon midpoints were clearly enriched relative to the input samples. By contrast, randomly selected control "midpoints" displayed no positional bias (Fig. 7D, red lines). This analysis is independent of the positional weight matrix and therefore directly confirms that SFRS1 binding sites are enriched at specific distances (approximately 20-41 nt) relative to splice sites. Given that the median lengths of internal constitutive and cassette exons identified by CLIP-Seq were found to be 142 nt and 158 nt, respectively, which are somewhat longer than their counterparts in the UCSC Known Gene database (125 nt and 108 nt, respectively), the data presented above reflect a clear positional bias of SFRS1 binding sites.

Many human disease mutations disrupt SFRS1 consensus sites

Single nucleotide substitutions or point mutations often alter the genetic code by producing aberrant protein products. However, although nonsense mutations introduce premature termination codons into the open reading frames of disease genes, it is often much more difficult to rationalize the pathogenic basis of missense and synonymous mutations. One explanation is that point mutations can manifest their detrimental effects through RNA processing. It is now well established that defects in pre-mRNA splicing and the regulation of

alternative splicing can induce heritable disease in humans (for review see Wang and Cooper, 2007). Studies of the *BRCA1*, *SMN*, *CFTR, GH1 and ATM* genes (among others) have demonstrated that all classes of point mutations, including nonsense mutations, can disrupt exonic splicing regulatory elements and induce aberrant alternative splicing (Teraoka et al. 1999; Liu et al. 2001; Cartegni and Krainer 2002; Moseley et al. 2002; Kashima and Manley 2003; Pagani et al. 2003). Based upon these results and others, a considerable effort to identify splicing-relevant mutations using PWM generated by both binding and functional SELEX is now underway (Smith et al. 2006). However, as stated above, different approaches for identifying the binding specificity of SFRS1 yield results that do not always concur. These differences serve to confound our understanding of the pathology of human inherited disease.

To investigate the potential impact of human disease-causing mutations on RNA processing involving SFRS1, exons from the *Human Gene Mutation Database* (HGMD; http://www.hgmd.org) were scanned with the PWM generated by CLIP-Seq. This dataset comprised 21,700 single nucleotide substitutions giving rise to either missense, synonymous or nonsense mutations either causing or associated with human inherited disease (Stenson et al. 2003). We scored mutations that abolished predicted SFRS1 binding sites relative to the wild-type allele, based on the thresholds established in Figure 2. As a control, exons from the Seattle SNPs database (http://pga.gs.washington.edu), containing 1,436 validated human polymorphisms, were also scanned with the PWM. The high allele frequencies of these polymorphisms are broadly indicative of their functional neutrality. In total, we identified 181 disease-causing single nucleotide substitutions (0.83%) in 83 different genes that ablate potential binding sites for SFRS1. Missense mutations accounted for the largest percentage (57%) of lost SFRS1 binding sites, whereas nonsense mutations made up the remaining 43% of mutations. By contrast, the Seattle SNPs database contained only 3 different polymorphic sites (0.21%) that were predicted to give rise to the loss of an SFRS1 binding site. We therefore found that substitutions resulting in the loss of SFRS1 binding sites were enriched

approximately four-fold in the HGMD mutation dataset relative to the control dataset (Fig. 8A, Fisher's Exact test, *P*-value < 10^{-5}). These data are consistent with previous studies showing that purifying selection reduces single nucleotide substitutions in exonic positions harboring splicing regulatory sequences (Majewski and Ott, 2002; Fairbrother et al. 2004; Parmley et al. 2006).

We next posed the question of where the mutations causing the loss of SFRS1 binding sites were located relative to splice sites (Fig. 8B). First, we determined the positions of all potential SFRS1 binding sites within exons represented in the HGMD. Potential SFRS1 binding sites were present throughout the disease gene exons and, as expected, these predicted binding sites showed very little positional bias (Fig. 8B, blue lines). By contrast, the majority of mutated SFRS1 binding sites were enriched in positions within 50 bp of the nearest 5' or 3' splice site (Fig. 8B, red lines). These data suggest that human disease mutations that disrupt potential SFRS1 binding sites are located in positions which are wholly compatible with their being physiological binding sites for SFRS1. In support of this conclusion, several of the mutations identified in this computational screen are already known to induce aberrant alternative splicing of the endogenous pre-mRNA in patients. These include three nonsense mutations in *MLH1* (K461X, Liu et al. 2001), *ATM* (E1978X, Teraoka et al. 1999) and *GH1* (E84X, Moseley et al. 2002) as well as two missense mutations in *GH1* (E85G, Moseley et al. 2002) and *NPHP1* (G342R, Betz et al. 2000).

DISCUSSION

High throughput DNA sequencing is rapidly changing the landscape of genomic research (Wold and Myers 2008). Our study is perhaps the first to utilize high-throughput sequencing to analyze protein-RNA interactions. We used the 454 FLX system (Roche) for amplicon sequencing based upon several considerations including read length and the well-validated platform. The read length of the 454 FLX system, using a short read kit, allows for 120-150 bp reads and is ideal for completely sequencing the RNA fragment and linkers produced by CLIP, ensuring that all sequence information is preserved. The longer read length provided by the 454 platform facilitated mapping of sequences to the genome and allowed for detection of many exon-intron and exon-exon junctions in the amplicon library. The primary consideration for future CLIP-Seq experiments is clearly an increased sequencing throughput. Issues arising from the natural abundance of different RNA transcripts and the preferential PCR amplification during library preparation have the potential to introduce many redundant reads. The work presented here cannot be viewed as comprehensive because significant new sequences were discovered in each of the four CLIP-Seq experiments analyzed. However, the majority of the novel sequences share at least one statistically significant match to the consensus SFRS1 binding site. Therefore, for the specific application of CLIP-Seq, data generated by the 454 platform are akin to a large-scale sampling or snapshot. Systems such as the SOLiD platform from ABI, which promise significantly increased throughput, have the potential to deliver truly comprehensive CLIP-Seq analyses.

The work presented above represents a significant step towards elucidating the roles of SFRS1 in post-transcriptional gene expression. Perhaps more importantly, these experiments demonstrate the potential of CLIP-Seq to illuminate the recognition code of RNA binding proteins and their *in situ* binding sites. A comprehensive evaluation of protein-RNA interactions is critical for understanding how RNA binding proteins positively or negatively regulate post-transcriptional processes such as alternative splicing. As future CLIP-Seq experiments increase

the catalogue of known protein-RNA interactions, efforts to integrate binding site data with functional genomic approaches have the potential to reveal the global organization of post-transcriptional regulatory networks in mammalian cells (Ule et al. 2006; Wang and Burge. 2008).

A significant advantage of CLIP-Seq is the large amount of raw data generated by the high throughput sequencing of amplicons. These data facilitate the elucidation of consensus sites using motif-finding algorithms such as MEME. The motif presented in Figure 2 was the only statistically significant sequence identified by MEME. The robust nature of the binding site model allowed for high resolution mapping of SFRS1 binding sites within the amplicon data. This is important because future *in silico* analyses should focus on these positionally restricted windows for identification of SFRS1-regulated exons. Such an approach is exemplified by the search for SFRS1 binding sites abolished by inherited mutations causing human genetic disease. We identified 181 exonic mutations in 82 different disease genes that abolish putative SFRS1 binding sites. Nearly 87% of these mutations were located within 50 bp of the nearest splice site, a region already demonstrated by CLIP-Seq to be enriched in SFRS1 binding sites. It is quite possible that mutations falling outside the preferred zone of SFRS1 binding will have little impact on RNA processing. However, at least five of the mutations identified here have already been correlated with changes in alternative splicing. Given that none of the mutations we identified were apparent in previously published reports identifying large numbers of splicing-relevant disease mutations, the pathological impact of exonic mutations upon splicing could turn out to be quite significant. Our findings argue that defective RNA processing, typically considered unusual in cases of non-synonymous disease mutations, could actually be the rule rather than the exception.

The CLIP method, developed in the Darnell laboratory at Rockefeller University, was first used to identify RNA targets of the splicing regulator NOVA. NOVA and SFRS1 are very different types of splicing factors and these differences are clearly reflected in their *in situ* RNA

binding specificities elucidated by CLIP. NOVA and SFRS1 engage RNA through structurally distinct RNA binding domains. The K-homolog RNA binding domain of NOVA recognizes a pyrimidine-rich YCAY motif that is nearly three-fold more abundant in RNA fragments bound by NOVA relative to non-targeted sequences. By contrast, our study shows that SFRS1 binds a purine-rich octamer with a GAA GAA core. This motif is highly enriched in exons bound by SFRS1 relative to randomly selected exon sequences (Fig. 6). The binding sites for both proteins within pre-mRNA are restricted to specific positions. Intronic and 3'UTR binding sites are most prevalent in the NOVA targets whereas internal exonic binding sites are strongly preferred by SFRS1 (Fig. 7). The positional binding specificities of both proteins can provide insight into their mechanisms of action. By comparing alternative splicing patterns of target transcripts in NOVA knock-out and wild-type mice, Ule and coworkers were able to deduce how different positions of NOVA binding sites influenced splice site selection (Ule et al. 2006). Although we have not yet established the functionality of each binding site for SFRS1 identified by CLIP-Seq, based upon the well established roles of SFRS1 in pre-mRNA splicing, we speculate that exonic binding sites are likely to function as splicing enhancers. However, because the majority of blocks identified by CLIP-Seq are classified as exonic, it is not possible to determine if these binding events occur during spliceosome assembly or instead at a later stage of mRNA processing. By contrast, blocks spanning exon-intron boundaries clearly represent interactions with pre-mRNA whereas those spanning exon junctions are derived from spliced mRNA. Fewer than 5% of all blocks (1,065) mapped to exon-intron boundaries and only 2,245 mapped to exon junctions. Given the proximity of these blocks to splice sites, it is possible that the same RNA elements recognized by SFRS1 influence pre-mRNA splicing and subsequent cytoplasmic steps of post-transcriptional gene expression. Recent work from our laboratory identified binding sites in several mRNAs that are engaged by SFRS1 in both nuclear and cytoplasmic/polysomal mRNA fractions of the cell (Sanford et al. 2008). Clearly, further functional studies are required to elucidate the functions of SFRS1 binding sites identified by CLIP-Seq.

Another interesting finding from our study is that binding sites for SFRS1 are enriched in exons that are adjacent to alternative cassette exons (Table 2). A previous study demonstrated that SFRS1 binding sites in a constitutive exon regulated the skipping of an upstream alternative cassette exon in the receptor tyrosine kinase *RON* gene (Ghigna et al. 2005). We propose that SFRS1 may play a prominent role in regulating this mode of competitive exon skipping by activating downstream splice sites. Finally, there are also significant differences between the functions of proteins encoded by mRNAs targeted by NOVA and SFRS1. NOVA mRNA targets tend to encode proteins involved in pre- and post-synaptic function as well as neuronal inhibition (Ule et al. 2003; Ule et al. 2005b). By contrast, mRNAs encoding other RNA binding proteins are over-represented in the collection of SFRS1 targets. These include a statistically significant enrichment of other splicing factors (Fig. 4). We are confident that the enrichment of RNA binding protein messages is biologically significant for several reasons. First, comparisons of SFRS1 targets with mRNAs present in the non-selected input RNA samples demonstrate that transcript abundance alone does not account for the enrichment of RBP mRNAs in the CLIP-Seq dataset. Secondly, recent experiments describe auto- and *trans*-regulatory post-transcriptional networks involved in homeostatic control of RNA binding protein expression (Lareau et al. 2007; Ni et al. 2007; Barberan-Soler and Zahler 2008; Saltzman et al. 2008). Two hallmarks of this mechanism include alternative splicing-coupled nonsense-mediated decay (AS-NMD) and ultraconserved *cis*-acting regulatory elements within coding exons of RBP mRNAs (Bejerano et al. 2004). In many cases, the ultraconserved regions of RBP genes overlap alternative exons with the potential to induce NMD (Bejerano et al. 2004; Lareau et al. 2007; Ni et al. 2007). In total, we identified 8 out of 111 known ultraconserved regions within exonic sequences. Included in these ultraconserved binding sites are several other genes encoding RNA binding proteins such as *SFRS1* itself, *SFRS6*, *HNRPM*, *PBCP2* and *CLK4* encoding SR protein kinase (Supplementary Fig. 4). We suggest that SFRS1 may be involved in controlling RBP homeostasis.

METHODS

Cell culture

Human embryonic kidney (HEK293T) cells were cultured in DMEM (Sigma), supplemented with 10% fetal calf serum and incubated at 37ºC in the presence of 5% CO_2. For each CLIP experiment, cells were grown to 75% confluence in 15 cm plates.

Cross-Linking Immunoprecipitation (CLIP)

CLIP analysis of SFRS1 was performed as described (Ule et al. 2003) with the following modifications relating to extract preparation and RNase treatment. Whole cell lysates were prepared from UV-treated or control cells as previously described (Sanford et al. 2005). The soluble extract was treated with 30U RQ DNase 1 for 20 min at 37ºC. The reactions were terminated by the addition of 20 mM EDTA. Subsequently, ribosomal subunits were cleared by centrifugation of the extract at 100,000 x g using an Optima Max ultracentrifuge (Beckman Coulter, USA) in a TLA120.2 rotor for 20 min. Cleared extracts were then treated with a dilute cocktail of RNase A/T1 (Ambion, USA) at a final dilution of 1:10,000 for 20 min at 37ºC. 200 U RNaseOut (Invitrogen, USA) was then added to the extract. Proteins were then partially denatured by addition of an equal volume of buffer A (2X PBS, 0.2% SDS, 1% NP-40). An aliquot of each UV-treated extract was used to prepare input RNA fragments. The remainder of the extract was used for immunoprecipitation with anti-SFRS1 monoclonal antibody. Extracts were treated with proteinase K (Ambion, USA) at a final concentration of 2 mg/mL, phenol extracted twice and ethanol precipitated. The trimmed input RNA was then ligated to the 3' RNA linker, followed by the 5' RNA linker, and used as templates for RT-PCR as previously described (Ule et al. 2005a). Gel-purified amplicons from the primary RT-PCR were re-amplified for 15 cycles using HPLC-purified primers that were complementary to the RNA linkers but also contained the 454 capture sequences (Margulies et al. 2005) as described in the original CLIP

protocol (Ule et al. 2005a). Amplicons were gel purified from 2% NuSieve Agarose gels using the QIAX II Gel Extraction kit (Qiagen).

High-throughput sequencing of amplicons

Prior to sequencing, the quality and quantity of gel-purified amplicons was assessed using a DNA LabChip1000 on an Agilent 2100 BioAnalyzer. High-throughput sequencing was performed using the Genome Sequencer FLX system (Roche Diagnostics) following standard protocols (Margulies et al. 2005). Titration runs were performed for all samples.

Primers

All DNA oligonucleotides (sequences available on request) were synthesized by IDT Inc. (Coralville, IA, USA).

454 capture primers:

P5454A (HPLC-purified): 5’ GCC TCC CTC GCG CCA TCA GAG GGA GGA CGA TGC GG 3’

P3454B (HPLC-purified): 5’ GCC TTG CCA GCC CGC TCA GCC GCT GGA AGT GAC TGA CAC 3’

Mapping of high-throughput sequencing data to the human genome

Several QC steps were implemented prior to mapping amplicon sequences to the human genome. We removed any amplicon sequences that did not include a recognizable match to 5’ and 3’ RNA linkers used for amplifying the RNA library. Once amplicons with both linker sequences had been identified, sequences <30 bp were removed, as were sequences containing ambiguous base calls. Finally, to avoid complications from preferential amplification during PCR, redundant identical amplicon sequences were filtered out from each experiment and only representative amplicons were retained. In order to study the binding of SFRS1 on a

genome-wide scale, the filtered amplicons were aligned using BLAT (Kent 2002) against human genome assembly hg18, March 2006 (NCBI build 36, accessed Oct. 18, 2007) (Karolchik et al. 2008). BLAT sequences containing more than 80% repetitive sequence were removed, and only 1 mismatch or 1 gap was allowed so that SNPs and splicing junctions could be accommodated. The annotation strategy focused upon loci containing overlapping unique amplicons. We refer to these regions as sequence blocks (blocks). Blocks from each CLIP-Seq experiment were annotated using the UCSC Known Gene database (http://genome.ucsc.edu/ ; accessed on Oct. 18, 2007) (Karolchik et al. 2008) and the Rfam database. The annotation data was then sub-classified in order to determine the number of blocks targeting specific genomic structures (exon, intron, exon-intron boundary, intergenic etc) and to determine the number of unique genomic structures identified in each experiment. To identify alternatively spliced exons bound by SFRS1, all binding sites were mapped against the Alternative Event Database (derived from AltEvent track in the UCSC Genome Browser (Karolchik et al. 2008)). The binding sites that were not located in alternatively spliced exons were by default designated as constitutive exons (the set of unique exons in the UCSC Known Gene database excluding alternative spliced exons related to AltEvent track).

Modeling and statistical evaluation of the SFRS1 consensus-binding motif

In order to establish precisely where SFRS1 binds, the Multiple Em for Motif Elicitation (MEME) algorithm (version 3.5.7; http://meme.sdsc.edu/meme/intro.html) (Bailey et al. 2006) was used to determine the consensus motif of amplicons in CLIP-hit blocks that did not overlap input blocks. We focused on the 681 blocks detected by 3 out of 4 CLIP samples. A single amplicon sequence was randomly selected from each block. 300 of the randomly selected sequences were used to perform MEME analysis and 300 sequences were used as the positive component of ‘gold standard’ sequences to evaluate the predictive power of the derived consensus motif. This procedure was repeated 20 times. The ROC curve was selected which had the maximum area under the curve (AUC) and its corresponding PWM (Positional Weight Matrix) was taken

as the final prediction of the SFRS1 consensus motif (Fig. 2). The PWM can be found in Supplementary Table 1 in the supporting on-line materials. During each ROC analysis, 40 groups of background sequences were selected to compare with the gold standard sequences. The background sequences were identical in length to the gold standard data set but were selected at random from intergenic desert regions (defined for practical purposes as having no genes within 100,000 bp upstream or downstream) from the chromosomes contributing to each gold standard sequence; any blocks from the CLIP experiments were deleted. After scanning each gold standard (true positive) and background sequence (false positive) using PWM derived from MEME, we computed the binding scores for each octamer, based upon which TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative) rates were calculated at different score cut-off thresholds. Averaging 40 groups of these data, we plotted a final averaged ROC curve, precision-recall curve and accuracy curve using the ROCR package (Sing et al. 2005). As mentioned above, we selected the ROC curve that gave the maximum AUC as our final result; its predictive power is illustrated in Fig. 2b. We adopted the cut-off value 5.2, corresponding to the maximum accuracy of prediction, as the threshold to ascertain whether or not a given octamer was likely to be a *bona fide* binding site. To evaluate the relationship of the SFRS1 PWM to the majority of CLIP-Seq data, the average number of binding sites per nucleotide was calculated for blocks from alternative cassette or constitutive exons identified by CLIP-Seq or an equal number of blocks selected at random from protein coding genes. Wilcoxon's Signed-Rank test was used to evaluate the statistical significance of the data.

Gene ontology analysis

We identified all genes targeted by 3 out of 4 CLIP samples, and then excluded those genes which were targeted by any INPUT sample. The gene list containing these genes was then input into EASE (Hosack et al. 2003) (Expression Analysis Systematic Explorer, version 2.0; http://david.abcc.ncifcrf.gov/ease) to compute the overrepresented functional categories in

"Biological Process", "Cellular Component" and "Molecular Function" systems (Hosack et al. 2003). EASE scores (modified Fisher's Exact test probabilities by penalizing the count of positive agreement by 1) were computed for all categories in each system. Holm's correction method was applied to the data in order to identify the most significantly overrepresented gene categories, because the genes contained in each GO category were not mutually exclusive. We also performed a comparison between the ratios of genes hit by CLIP and the background gene lists (CLIP-hit ratio and Expected ratio). These data provided evidence for annotation enrichment relative to the entire genome. Because transcript abundance may also influence the protein-RNA interactions identified by CLIP-Seq, we also used analyzed mRNAs identified in the non-selected input RNA samples. Comparison of the annotation enrichment between the CLIP-Seq and Input RNA samples was important as a means to identify specific targets that could have originated from highly transcribed genes.

SFRS1 binding site frequency relative to splice sites

Blocks identified by CLIP-Seq or randomly selected exon sequences from the human genome were scanned using the SFRS1 consensus PWM to identify statistically significant binding sites (those with binding scores above the matching score threshold of 5.2). As an additional control, we generated a PWM corresponding to the reverse complementary sequence of the SFRS1 binding site (see Supplementary Table 1). The frequency of binding sites (N) at each nucleotide position (i) relative to a splice site was determined by recording the genomic coordinates for each exonic binding site at the fifth position of the consensus motif. Binding site quantity was adjusted for the uneven distribution of exon lengths by dividing the binding site frequency at a given position (N_i) by the number of exons of at least 2i in length (Majewski and Ott 2002). We favored this approach over normalization of exon length because we felt that the absolute position, rather than the proportional position, of each *cis*-acting element was more likely to be biologically meaningful. The adjusted frequency (N_i') of binding sites was plotted in 10 nt bins relative to the 5' or 3' splice site. Amplicon midpoints were mapped by selecting sequences at

random from blocks identified by CLIP-Seq. Because SFRS1 CLIP-Seq identified many more exons that the input amplicons, we compared the density of amplicon midpoints in each 10 bp bin. Midpoint density is calculated by dividing the number of midpoints at each position (i) by the number of exons at least 2i in length. The distribution of midpoints from targeted and input amplicons was compared to randomly selected positions within exon sequences. This negative control calculation was repeated 40 times.

Identification of SFRS1 binding sites predicted to be disrupted by human disease mutations

Two variation datasets were used to identify the possible role of SFRS1 binding site mutation in the contexts of human inherited disease and inter-individual variation: (A) germline pathogenic substitutions from the *Human Gene Mutation Database,* (B) single nucleotide polymorphisms from the SeattleSNPs resequencing project. All nucleotide variants were mapped to the human genome using the Blast-like alignment tool (BLAT; Kent 2002). If genomic variants mapped to multiple exons (overlapping), then both exons were considered in the analysis. A total of 40,397 coding region pathogenic nucleotide substitutions were retrieved from the *Human Gene Mutation Database* (HGMD; http://www.hgmd.org) (Stenson et al. 2003). However, we only examined those 21,700 single nucleotide substitutions from 440 genes (53.7% of all substitutions) that were located within internal coding exons. The 1436 SNPs derived from the SeattleSNPs resequencing project (http://pga.mbt.washington.edu/), selected for their high allele frequency, were assumed to be functionally neutral.

Mapping mutations predicted to disrupt SFRS1 binding sites

For each nucleotide substitution, a dataset comprising the wild-type and corresponding mutant exons was compiled. Using these data and a sliding window of 8 bp to evaluate the SFRS1 position weight matrix (PWM) using a threshold of 5.2, SFRS1 target sites were determined within the wild-type and mutant exons. Mutations reducing the matching score to levels below

the score cut-off threshold of 5.2 were classified as loss of binding and *vice versa* for gain of binding. The net loss or gain for the set of 8-mers was then used to determine whether the result was an overall loss or gain of SFRS1 at any given position. The frequency and positional distribution of mutations predicted to give rise to a loss of SFRS1 binding sites was determined as described above.

Validation of SFRS1-RNA interactions

Anti-SFRS1 monoclonal antibody (60 uL cell culture supernatant/IP) was bound to 60 µL (packed bead volume) recombinant Protein A Sepharose CL-4B beads (Invitrogen, USA) in 1.0 mL 0.1 M sodium phosphate buffer, pH 8.1. For negative control precipitations, 40 µL beads were equilibrated with 0.1 M sodium phosphate buffer, pH 8.1. Antibody-bound or control beads were then washed 3x in 1.0 mL lysis buffer (20 mM Tris•HCl pH 7.5, 100 mM NaCl, 10 mM $MgCl_2$, 0.5% NP40, 0.5% Triton X100, 1 mini complete protease inhibitor tablet (Roche Diagnostics). 20 µL antibody-bound beads were held in reserve for a western blotting negative control sample (beads, antibody, no extract). Whole cells were prepared by extracting HEK293T cells in lysis buffer as described above. $1/50^{th}$ of the soluble extract was retained as an input reference sample for western blotting and RT-PCR analysis. Equal amounts of soluble extract were incubated with 40 µL anti-SFRS1 beads or control beads at 4ºC for one hr on a rotating mixer. Beads were then washed 4x with lysis buffer. One third of the beads were used for western blot analysis of precipitated protein; the remaining beads were used for RNA extraction with Tri Reagent LS (Sigma) following the manufacturer's protocol. Equal amounts of RNA (approximately 500-600 ng) were treated with 1 U RQ DNase (Promega Corp., USA) for 20 min at 37ºC. RQ DNase was inactivated by the addition of EGTA and incubation at 65ºC for 10 min. Equal amounts of input RNA, anti-SFRS1 IP or control IP (typically 400 ng) were used for cDNA synthesis with oligo dT cellulose and SuperScript III reverse transcriptase (Invitrogen, USA) following the manufacturer's specifications. 25 ng cDNA (or ddH_2O for no template controls) was used as a template for all PCR assays. PCR reactions were performed in 96-well format, 25 µL

per well (1X R-Taq mix (MidWest Scientific) and 200 µM primers) in an Eppendorf EP Mastercycler (Germany) using the following cycle conditions: 95ºC 5 min; 35 cycles of: 95ºC 30 sec, 59ºC 30 sec, 72ºC 60 sec; 72ºC 5 min. PCR products were analyzed on 2% agarose gels and visualized by staining with ethidium bromide.

Acknowledgments

We thank A. Krainer for generously providing anti-SF2 (SFRS1) monoclonal antibody. We thank M. Ares and A. Zahler for comments on the manuscript and J. Bruzik, J. Caceres and D. Unni for helpful discussion. Finally, we thank the reviewers for their efforts in providing a thorough critique of our manuscript. Amplicon sequencing was performed at the Indiana University Center for Genomics and Bioinformatics (Bloomington, Indiana). This work was supported by an American Heart Association Scientist Development Grant to J.R.S. (0830206N), grants from the U.S. National Institutes of Health to J.R.S (1R01GM085121), H.J.E and S.D.M (K22LM009135 and R01LM009722) and financial support from BIOBASE GmbH to D.N.C. and M.M.

Figure Legends

Figure 1. CLIP of SFRS1 from HEK293T cells and amplicon sequencing. (*A*) Western blot analysis of SFRS1 immunopreciptation (IP) from control and UV cross-linked cells. CLIP was performed in from three independent cultures of HEK293T cells (Lanes 5-7). SFRS1 was detected with the same antibody used for IP. The blot was visualized with chemiluminescence (Pierce Super Signal). This panel demonstrates the specificity of the IP; no signal was detected in control IP samples (lane 4), but SFRS1 was efficiently precipitated from both UV-irradiated and control cell extracts. (*B*) Autoradiograph of ^{32}P labeled SFRS1-RNA complexes immobilized on nitrocellulose membrane. Following extensive washes of the immunoprecipitated complex, the beads were incubated with T4PNK (NEB) and [γ^{32}P]-ATP on order to phosphorylate the 5' end of RNA fragments. In the absence of UV-irradiation, SFRS1 can become phosphorylated. However, this does not require T4PNK (compare lanes 1 and 2). These data suggest that a subset of SFRS1 is co-purified with an SR protein kinase even under these stringent conditions. Importantly, in the absence of UV-irradiation, no slower migrating protein-RNA complexes are observed (compare lanes 1-2 with 4-6). The arrow indicates the position of "free" SFRS1 and the bracket defines the region of nitrocellulose excised from the blot. (*C*) Histogram comparing the amplicon length distribution with numbers of reads. Amplicons prepared from both CLIP (blue line) and non-selected input RNA samples (red line) were sequenced using the 454 FLX platform and short read reagents. As expected from panel B, the majority of amplicons (after removal of both linker sequences) are between 40-60 bp in length.

Figure 2. Modeling the *in situ* SFRS1 consensus-binding motif. (*A*) The MEME algorithm was used to identify a consensus motif from 300 amplicons selected at random from a total of 641 blocks common to 3 out of 4 CLIP-Seq experiments. This calculation was repeated 20 times. The motif with the highest sensitivity and specificity (see panel C) is depicted here. The likelihood of finding this motif at random is $<1x10^{-107}$. (*B*) The averaged Accuracy plot for each

positional weight matrix calculated from the MEME results. The prediction accuracy of the SFRS1 site was plotted as a function of matching score cutoff threshold. Maximum accuracy (78%) was achieved at a cutoff score of 5.2. This calculation was repeated 40 times using the gold standard data set. Error bars correspond to the standard deviation from the mean accuracy. (*C*) The averaged Receiver Operator Characteristic curve of the SFRS1. This plot evaluates the sensitivity (true positive rate of discovery) and specificity (false positive discovery rate) as functions of matching score cutoff threshold. The ROC evaluates the ability of the PWM to discriminate between positive (CLIP-Seq derived) and negative components (55 bp fragments selected from intergenic deserts) of the gold standard dataset. This calculation was repeated 40 times. Error bars correspond to standard deviation from the mean sensitivity and specificity.

Figure 3. Classification of *cis*-acting RNA elements bound by SFRS1. (A) Annotation strategy for classifying sequence blocks identified by CLIP-Seq. Following alignment of amplicon sequences to the human genome, blocks of overlapping sequences were defined and subsequently annotated using the UCSC Known Gene and Rfam databases. Blocks were classified at the gene level (“In Gene”, “Out of Gene” and “In noncoding RNA”) and at the transcript level (“In Exon”, “In Intron”, “Exon-Intron Boundary” and “Exon-Junction”). None of the blocks presented here overlapped amplicon sequences obtained from input RNA samples. (*B*) Quantification of block annotations. The majority of blocks were classified as “in gene” and were predominantly associated with exon sequences. Slightly more than 25% of blocks were defined as intergenic, but binding sites for SFRS1 within these transcripts were found to be highly conserved over evolutionary time (See Supplementary Fig. 2).

Figure 4. SFRS1 targets are enriched in mRNAs encoding RNA binding proteins. (*A*) The top ten classes of gene ontology terms enriched in the CLIP data set relative to the expected ratios in the DAVID database (Hosack et al. 2003). The top 10 classes of targets are all related to

gene expression. The numbers of genes observed in each category are indicated in the pie chart. Holm-corrected EASE scores are given for each category (Hosack et al. 2003). (*B*) Comparison of the top 10 CLIP-enriched gene ontology terms with non-selected input mRNA samples. The annotation enrichment relative to the genome is plotted for both CLIP (gray bars) and input (black bars) derived mRNAs. mRNAs encoding splicing factors are most highly enriched in CLIP.

Figure 5. Validation of SFRS1-RNA interactions by RNA-IP RT-PCR. (A) Western blot analysis of proteins precipitated by the anti-SFRS1 monoclonal antibody. SFRS1 was detected in both the input extract (lane 1) and the material immunoprecipitated with anti-SFRS1 (lane 4) but not the controls beads (lanes 2 and 3). The blot was visualized as described in Figure 1. (B) Examples of RT-PCR analysis of endogenous SFRS1-mRNA complexes. RNA extracted from the control IP, input extract and SFRS1 IP, was reverse transcribed using oligo dT and Superscript III (Invitrogen). 78 different primer sets were used to amplify specific transcripts from cDNA. (C) Summary of RT-PCR validation of 78 randomly selected sequence blocks identified by CLIP-Seq. Validated interactions correspond to detectable PCR product in both the input and SFRS1 IP samples. False positive transcripts correspond to PCR products present in the input and the control IP. Technical failures yielded no PCR products from any cDNA sample indicating that the transcript could not be directly validated under these conditions.

Figure 6. The SFRS1 consensus motif was enriched in blocks identified by CLIP relative to randomly selected blocks from exon sequences. The average number of SFRS1 consensus sites per nucleotide was determined and plotted for sequence blocks in 8,693 constitutive (black bars) and 426 alternative cassette exons (gray bars) identified by the CLIP-Seq experiment. For the control group, 55 bp regions were selected at random from equal numbers of constitutive or alternative cassette exons not contained in the pool of SFRS1 CLIP-Seq data. The Wilcoxon test confirmed the mean binding sites per nt were significantly different for CLIP-Seq and control

exons ($P<10^{-22}$). Binding sites for SFRS1 in alternative cassette exons were found to be modestly enriched relative to constitutive exons ($P<0.005$).

Figure 7. SFRS1 binding sites are enriched at fixed positions relative to splice sites. The adjusted frequency of SFRS1 consensus sites within 10 bp bins (N') at a specific position relative to splice sites (i) was calculated by multiplying the number of consensus sites observed by the total number of exons divided by exon ≥ 2i in length. In panels A-C, the blue and red lines represent the sense or anti-sense PWMs, respectively. (A) Positions of SFRS1 binding sites within sequence blocks identified by CLIP-Seq. (B) Positions of SFRS1 binding sites across full length exons targeted by SFRS1. (C) Positions of SFRS1 binding sites across randomly selected exons from the human genome. (D) The distance from splice sites to the midpoints of CLIP-Seq amplicons was calculated as described above. CLIP-Seq and Input Amplicon midpoints (blue and orange lines, respectively) were compared to randomly selected "points" picked from exons selected at random from the genome (red lines). This comparison demonstrated that the amplicons identified by CLIP were enriched relative to the input samples at the boundaries of exons. Likewise, randomly selected 'points' differ dramatically with respect to the experimentally observed amplicon midpoints.

Figure 8. Disruption of SFRS1 binding sites can cause human inherited disease. (A) Single nucleotide substitutions causing loss of predicted SFRS1 binding sites in the *Human Gene Mutation Database* (http://www.hgmd.org) and the Seattle SNPs database (http://pga.gs.washington.edu) were identified by scanning reference and mutated exon sequences with the SFRS1 PWM. The proportion of entries in each database giving rise to a loss of SFRS1 sites was then plotted and a statistically significant difference between the HGMD and Seattle SNPs datasets observed (P-value < 10^{-5}; Fisher's Exact test). (B) Disease mutations resulting in the loss of SFRS1 binding sites were found to be largely confined to exon boundaries. The left and right panels plot mutated sites relative to the 5' and 3' splice sites,

respectively. The blue line in each plot represents the distribution of SFRS1 binding sites throughout the HGMD exons. The red lines correspond to the distribution of sites ablated by disease causing mutations. These data demonstrate that although binding sites for SFRS1 can be predicted across HGMD exons, disease mutations tend only to disrupt those SFRS1 binding sites that are in close proximity to splice sites.

Table 1. Summary of 454 FLX sequencing data for CLIP-Seq and Input derived amplicon libraries.

	CLIP-Seq	Input
Total Reads	932152	670448
Reads containing sequences from 5' and 3' RNA linker	931138	614467
Filtered ampicons (N=0; Length>30 bp)	794226	353283
Unique filtered amplicons	135318	218108
Stringent BLAT matches to HG18	58953	3374

Table 2. SFRS1 RNA targets with annotated alternative splicing events (Alt Events database). The table documents the expected and experimentally observed number of exons from each classification. The expected value is based upon the ratio of each classification divided by the total number of exons (231,385) in the UCSC Known Gene database. The Significance of differences between expected and observed values was calculated using Fishers Exact Test. Odds ratio were calculated to estimate the level enrichment of each classification in the CLIP-Seq dataset. The 95% Confidence intervals (Low and High C.I., respectively) are also shown.

Alternative Event	Expected	Observed	P value	Odds Ratio	Low C.I.
Cassette	840	724	1.57×10^{-5}	0.85	(0.78, 0.91)
Retained Intron	196	151	6.50×10^{-4}	0.76	(0.64, 0.84)
Alternative 5'ss	131	221	9.51×10^{-14}	1.77	(1.54, 2.03)
Alternative 3'ss	203	278	6.36×10^{-9}	1.46	(1.29, 1.65)
Upstream Adjacent	836	1016	1.28×10^{-10}	1.25	(1.17, 1.34)
Downstream Adjacent	836	1020	5.44×10^{-11}	1.26	(1.18, 1.34)

References

Bailey, T.L. and C. Elkan. 1995. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol* **3:** 21-29.

Bailey, T.L., N. Williams, C. Misleh, and W.W. Li. 2006. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* **34:** W369-373.

Barberan-Soler, S. and A.M. Zahler. 2008. Alternative splicing regulation during *C. elegans* development: splicing factors as regulated targets. *PLoS Genet* **4:** e1000001.

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W.J. Kent, J.S. Mattick, and D. Haussler. 2004. Ultraconserved elements in the human genome. *Science* **304:** 1321-1325.

Betz, R., C. Rensing, E. Otto, A. Mincheva, D. Zehnder, P. Lichter, and F. Hildebrandt. 2000. Children with ocular motor apraxia type Cogan carry deletions in the gene (*NPHP1*) for juvenile nephronophthisis. *J Pediatr* **136:** 828-831.

Brown, V., P. Jin, S. Ceman, J.C. Darnell, W.T. O'Donnell, S.A. Tenenbaum, X. Jin, Y. Feng, K.D. Wilkinson, J.D. Keene et al. 2001. Microarray identification of FMRP-associated brain mRNAs and altered mRNA translational profiles in fragile X syndrome. *Cell* **107:** 477-487.

Caceres, J.F., G.R. Screaton, and A.R. Krainer. 1998. A specific subset of SR proteins shuttles continuously between the nucleus and the cytoplasm. *Genes Dev* **12:** 55-66.

Caputi, M., G. Casari, S. Guenzi, R. Tagliabue, A. Sidoli, C.A. Melo, and F.E. Baralle. 1994. A novel bipartite splicing enhancer modulates the differential processing of the human fibronectin EDA exon. *Nucleic Acids Res* **22:** 1018-1022.

Cartegni, L. and A.R. Krainer. 2002. Disruption of an SFRS1-dependent exonic splicing enhancer in *SMN2* causes spinal muscular atrophy in the absence of *SMN1*. *Nat Genet* **30:** 377-384.

Cavaloc, Y., C.F. Bourgeois, L. Kister, and J. Stevenin. 1999. The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. *RNA* **5:** 468-483.

Fairbrother, W.G., R.F. Yeh, P.A. Sharp, and C.B. Burge. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* **297:** 1007-1013.

Fairbrother, W.G., D.. Holste, C.B. Burge, and P.A. Sharp. 2004. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biology* **2.**

Felsenstein, J. and G.A. Churchill. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol Biol Evol* **13:** 93-104.

Friedman, B.A., M.B. Stadler, N. Shomron, Y. Ding, and C.B. Burge. 2008. *Ab initio* identification of functionally interacting pairs of *cis*-regulatory elements. *Genome Res* **18:** 1643-1651.

Fu, X.D. 1993. Specific commitment of different pre-mRNAs to splicing by single SR proteins. *Nature* **365:** 82-85.

Gama-Carvalho, M., N.L. Barbosa-Morais, A.S. Brodsky, P.A. Silver, and M. Carmo-Fonseca. 2006. Genome-wide identification of functionally distinct subsets of cellular mRNAs associated with two nucleocytoplasmic-shuttling mammalian splicing factors. *Genome Biol* **7:** R113.

Ge, H., P. Zuo, and J.L. Manley. 1991. Primary structure of the human splicing factor ASF reveals similarities with *Drosophila* regulators. *Cell* **66:** 373-382.

Gerber, A.P., D. Herschlag, and P.O. Brown. 2004. Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol* **2:** E79.

Ghigna, C., S. Giordano, H. Shen, F. Benvenuto, F. Castiglioni, P.M. Comoglio, M.R. Green, S. Riva, and G. Biamonti. 2005. Cell motility is controlled by SF2/ASF through alternative splicing of the Ron protooncogene. *Mol Cell* **20:** 881-890.

Graveley, B.R., K.J. Hertel, and T. Maniatis. 2001. The role of U2AF35 and U2AF65 in enhancer-dependent splicing. *RNA* **7:** 806-818.

Graveley, B.R. and T. Maniatis. 1998. Arginine/serine-rich domains of SR proteins can function as activators of pre-mRNA splicing. *Mol Cell* **1:** 765-771.

Griffiths-Jones, S., A. Bateman, M. Marshall, A. Khanna, and S.R. Eddy. 2003. Rfam: an RNA family database. *Nucleic Acids Res* **31:** 439-441.

Hanamura, A., J.F. Caceres, A. Mayeda, B.R. Franza, Jr., and A.R. Krainer. 1998. Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors. *RNA* **4:** 430-444.

Hieronymus, H. and P.A. Silver. 2003. Genome-wide analysis of RNA-protein interactions illustrates specificity of the mRNA export machinery. *Nat Genet* **33:** 155-161.

Hosack, D.A., G. Dennis, Jr., B.T. Sherman, H.C. Lane, and R.A. Lempicki. 2003. Identifying biological themes within lists of genes with EASE. *Genome Biol* **4:** R70.

Huang, Y. and J.A. Steitz. 2001. Splicing factors SRp20 and 9G8 promote the nucleocytoplasmic export of mRNA. *Mol Cell* **7:** 899-905.

Karni, R., E. de Stanchina, S.W. Lowe, R. Sinha, D. Mu, and A.R. Krainer. 2007. The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nat Struct Mol Biol* **14:** 185-193.

Karolchik, D., R.M. Kuhn, R. Baertsch, G.P. Barber, H. Clawson, M. Diekhans, B. Giardine, R.A. Harte, A.S. Hinrichs, F. Hsu et al. 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* **36:** D773-779.

Kashima, T. and J.L. Manley. 2003. A negative element in *SMN2* exon 7 inhibits splicing in spinal muscular atrophy. *Nat Genet* **34:** 460-463.

Keene, J.D. 2007. RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet* **8:** 533-543.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664.

Kohtz, J.D., S.F. Jamison, C.L. Will, P. Zuo, R. Luhrmann, M.A. Garcia-Blanco, and J.L. Manley. 1994. Protein-protein interactions and 5'-splice-site recognition in mammalian mRNA precursors. *Nature* **368:** 119-124.

Krainer, A.R., A. Mayeda, D. Kozak, and G. Binns. 1991. Functional expression of cloned human splicing factor SF2: homology to RNA-binding proteins, U1 70K, and *Drosophila* splicing regulators. *Cell* **66:** 383-394.

Lareau, L.F., M. Inada, R.E. Green, J.C. Wengrod, and S.E. Brenner. 2007. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* **446:** 926-929.

Lemaire, R., J. Prasad, T. Kashima, J. Gustafson, J.L. Manley, and R. Lafyatis. 2002. Stability of a PKCI-1-related mRNA is controlled by the splicing factor ASF/SF2: a novel function for SR proteins. *Genes Dev* **16:** 594-607.

Li, X. and J.L. Manley. 2005. Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. *Cell* **122:** 365-378.

Lin, S. and X.D. Fu. 2007. SR proteins and related factors in alternative splicing. *Adv Exp Med Biol* **623:** 107-122.

Liu, H.X., L. Cartegni, M.Q. Zhang, and A.R. Krainer. 2001. A mechanism for exon skipping caused by nonsense or missense mutations in *BRCA1* and other genes. *Nat Genet* **27:** 55-58.

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

Longman, D., I.L. Johnstone, and J.F. Caceres. 2000. Functional characterization of SR and SR-related genes in *Caenorhabditis elegans*. *Embo J* **19:** 1625-1637.

Majewski, J. and J. Ott. 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Res* **12:** 1827-1836.

Maniatis, T. and B. Tasic. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418:** 236-243.

Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-380.

Mili, S. and J.A. Steitz. 2004. Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. *Rna* **10:** 1692-1694.

Moseley, C.T., P.E. Mullis, M.A. Prince, and J.A. Phillips, 3rd. 2002. An exon splice enhancer mutation causes autosomal dominant GH deficiency. *J Clin Endocrinol Metab* **87:** 847-852.

Ni, J.Z., L. Grate, J.P. Donohue, C. Preston, N. Nobida, G. O'Brien, L. Shiue, T.A. Clark, J.E. Blume, and M. Ares, Jr. 2007. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev* **21:** 708-718.

Olson, S., M. Blanchette, J. Park, Y. Savva, G.W. Yeo, J.M. Yeakley, D.C. Rio, and B.R. Graveley. 2007. A regulator of Dscam mutually exclusive splicing fidelity. *Nat Struct Mol Biol*.

Pagani, F., E. Buratti, C. Stuani, and F.E. Baralle. 2003. Missense, nonsense, and neutral mutations define juxtaposed regulatory elements of splicing in cystic fibrosis transmembrane regulator exon 9. *J Biol Chem* **278:** 26580-26588.

Parmley, J.L., J.V. Chamary, L.D. Hurst. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol* **23:** 301-9.

Ramchatesingh, J., A.M. Zahler, K.M. Neugebauer, M.B. Roth, and T.A. Cooper. 1995. A subset of SR proteins activates splicing of the cardiac troponin T alternative exon by direct interactions with an exonic enhancer. *Mol Cell Biol* **15:** 4898-4907.

Saltzman, A.L., Y.K. Kim, Q. Pan, M.M. Fagnani, L.E. Maquat, and B.J. Blencowe. 2008. Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay. *Mol Cell Biol* **28:** 4320-4330.

Sanchez-Diaz, P. and L.O. Penalva. 2006. Post-transcription meets post-genomic: the saga of RNA binding proteins in a new era. *RNA Biol* **3:** 101-109.

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

Sanford, J.R., J.D. Ellis, D. Cazalla, and J.F. Caceres. 2005. Reversible phosphorylation differentially affects nuclear and cytoplasmic functions of splicing factor 2/alternative splicing factor. *Proc Natl Acad Sci USA* **102:** 15042-15047.

Sanford, J.R., N.K. Gray, K. Beckmann, and J.F. Caceres. 2004. A novel role for shuttling SR proteins in mRNA translation. *Genes Dev* **18:** 755-768.

Shen, H. and M.R. Green. 2004. A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly. *Mol Cell* **16:** 363-373.

Siepel, A., G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.W. Hillier, S. Richards et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034-1050.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics* **21:** 3940-3941.

Smith, P.J., C. Zhang, J. Wang, S.L. Chew, M.Q. Zhang, and A.R. Krainer. 2006. An increased specificity score matrix for the prediction of SFRS1-specific exonic splicing enhancers. *Hum Mol Genet* **15:** 2490-2508.

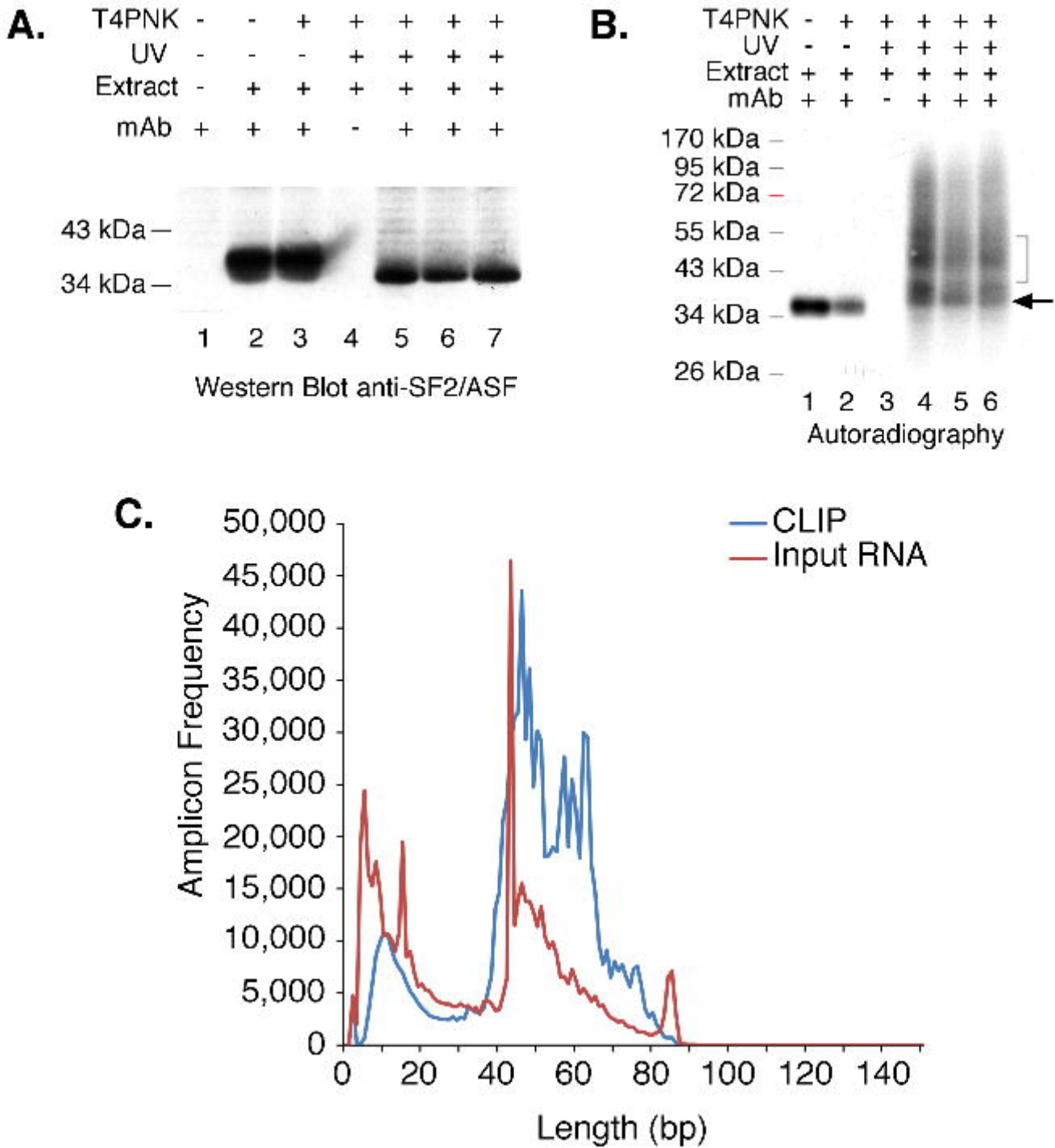
Stenson, P.D., E.V. Ball, M. Mort, A.D. Phillips, J.A. Shiel, N.S. Thomas, S. Abeysinghe, M. Krawczak, and D.N. Cooper. 2003. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* **21:** 577-581.

Tacke, R. and J.L. Manley. 1995. The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. *EMBO J* **14:** 3540-3551.

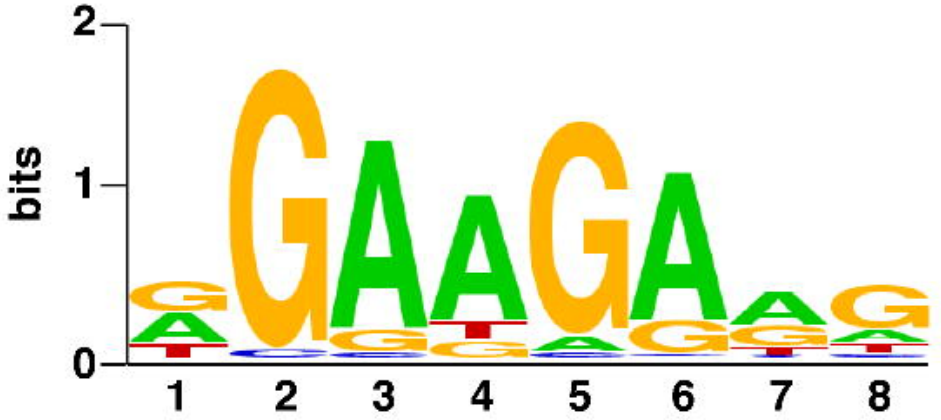
Tenenbaum, S.A., P.J. Lager, C.C. Carson, and J.D. Keene. 2002. Ribonomics: identifying mRNA subsets in mRNP complexes using antibodies to RNA-binding proteins and genomic arrays. *Methods* **26:** 191-198.

Teraoka, S.N., M. Telatar, S. Becker-Catania, T. Liang, S. Onengut, A. Tolun, L. Chessa, O. Sanal, E. Bernatowska, R.A. Gatti et al. 1999. Splicing defects in the ataxia-

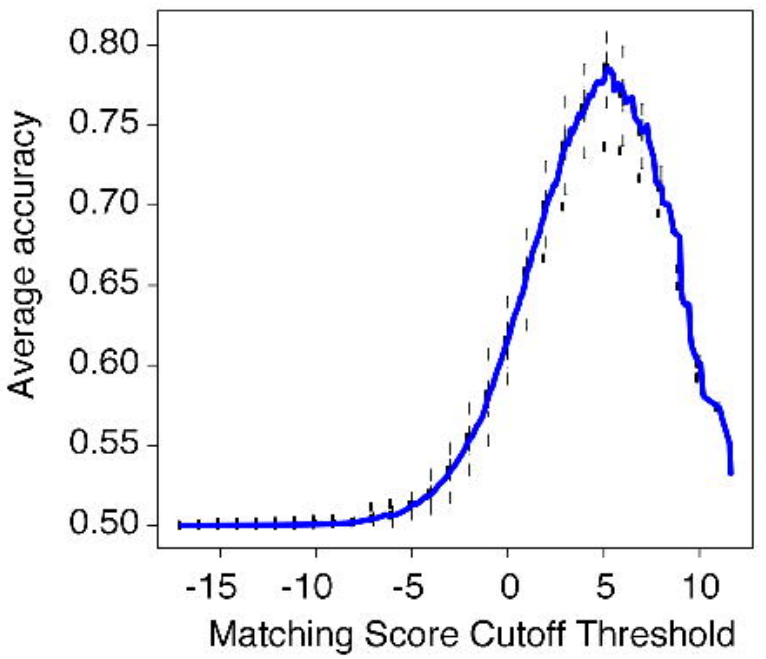
telangiectasia gene, *ATM*: underlying mutations and consequences. *Am J Hum Genet* **64:** 1617-1631.
Tuerk, C. and L. Gold. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249:** 505-510.
Ule, J. and R.B. Darnell. 2006. RNA binding proteins and the regulation of neuronal synaptic plasticity. *Curr Opin Neurobiol* **16:** 102-110.
Ule, J., K. Jensen, A. Mele, and R.B. Darnell. 2005a. CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods* **37:** 376-386.
Ule, J., K.B. Jensen, M. Ruggiu, A. Mele, A. Ule, and R.B. Darnell. 2003. CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302:** 1212-1215.
Ule, J., G. Stefani, A. Mele, M. Ruggiu, X. Wang, B. Taneri, T. Gaasterland, B.J. Blencowe, and R.B. Darnell. 2006. An RNA map predicting Nova-dependent splicing regulation. *Nature* **444**: 580-586.
Ule, J., A. Ule, J. Spencer, A. Williams, J.S. Hu, M. Cline, H. Wang, T. Clark, C. Fraser, M. Ruggiu et al. 2005b. Nova regulates brain-specific splicing to shape the synapse. *Nat Genet* **37:** 844-852.
Wang, G.S. and T.A. Cooper. 2007. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8:** 749-761.
Wang, J., Y. Takagaki, and J.L. Manley. 1996. Targeted disruption of an essential vertebrate gene: ASF/SF2 is required for cell viability. *Genes Dev* **10:** 2588-2599.
Wang, Z. and C.B. Burge. 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14:** 802-813.
Wang, Z., X. Xiao, E. Van Nostrand, and C.B. Burge. 2006. General and specific functions of exonic splicing silencers in splicing control. *Mol Cell* **23:** 61-70.
Wold, B. and R.M. Myers. 2008. Sequence census methods for functional genomics. *Nat Methods* **5:** 19-21.
Xiao, X., Z. Wang, M. Jang, and C.B. Burge. 2007. Coevolutionary networks of splicing cis-regulatory elements. *Proc Natl Acad Sci USA* **104:** 18583-18588.
Xu, X., D. Yang, J.H. Ding, W. Wang, P.H. Chu, N.D. Dalton, H.Y. Wang, J.R. Bermingham, Jr., Z. Ye, F. Liu et al. 2005. ASF/SF2-regulated CaMKIIdelta alternative splicing temporally reprograms excitation-contraction coupling in cardiac muscle. *Cell* **120:** 59-72.
Zhang, X.H. and L.A. Chasin. 2004. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev* **18:** 1241-1250.
Zhang, X.H., C.S. Leslie, and L.A. Chasin. 2005. Dichotomous splicing signals in exon flanks. *Genome Res* **15:** 768-779.
Zhang, Z. and A.R. Krainer. 2004. Involvement of SR proteins in mRNA surveillance. *Mol Cell* **16:** 597-607.


A.
T4PNK - - + + + + +
UV - - - + + + +
Extract - + + + + + +
mAb + + + - + + +
43 kDa
34 kDa
1 2 3 4 5 6 7
Western Blot anti-SF2/ASF
B.
T4PNK - + + + + +
UV - - + + + +
Extract + + + + + +
mAb + + - + + +
170 kDa
95 kDa
72 kDa
55 kDa
43 kDa
34 kDa
26 kDa
1 2 3 4 5 6
Autoradiography
C.
CLIP
Input RNA
Amplicon Frequency
50,000
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
0 20 40 60 80 100 120 140
Length (bp)

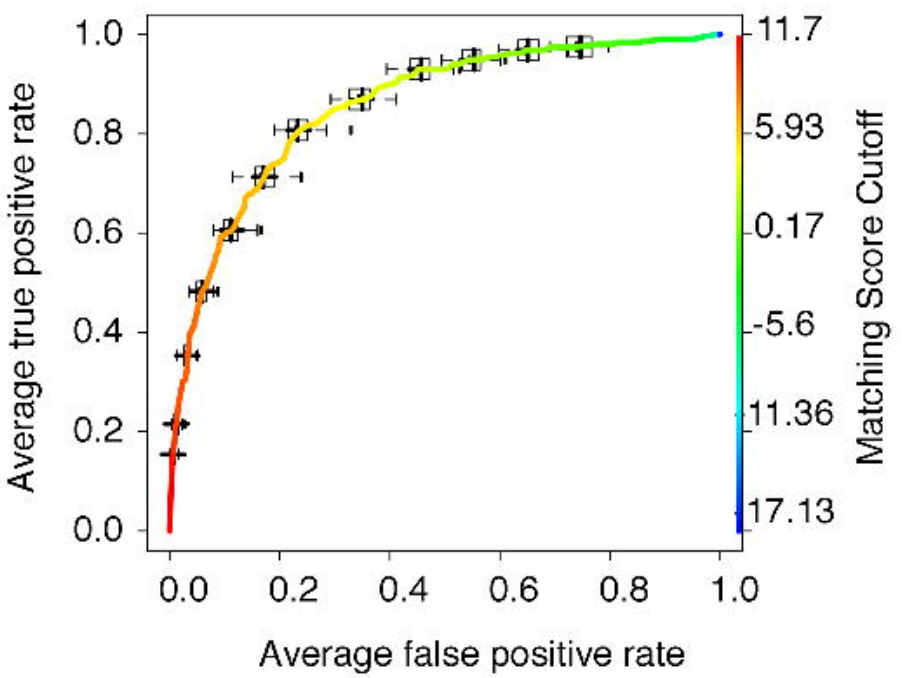
A.

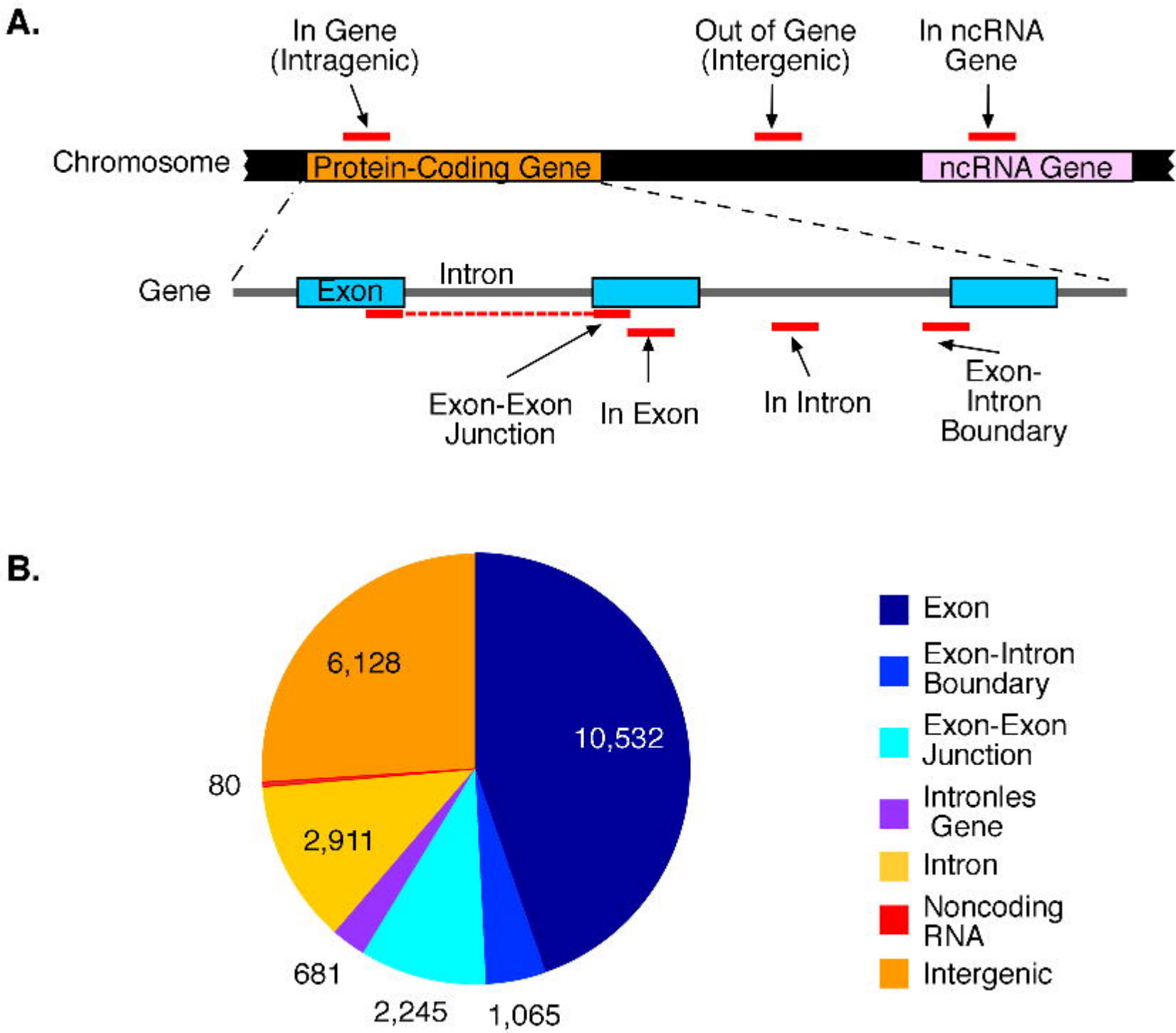

bits
2
1
0
1
2
3
4
5
6
7
8

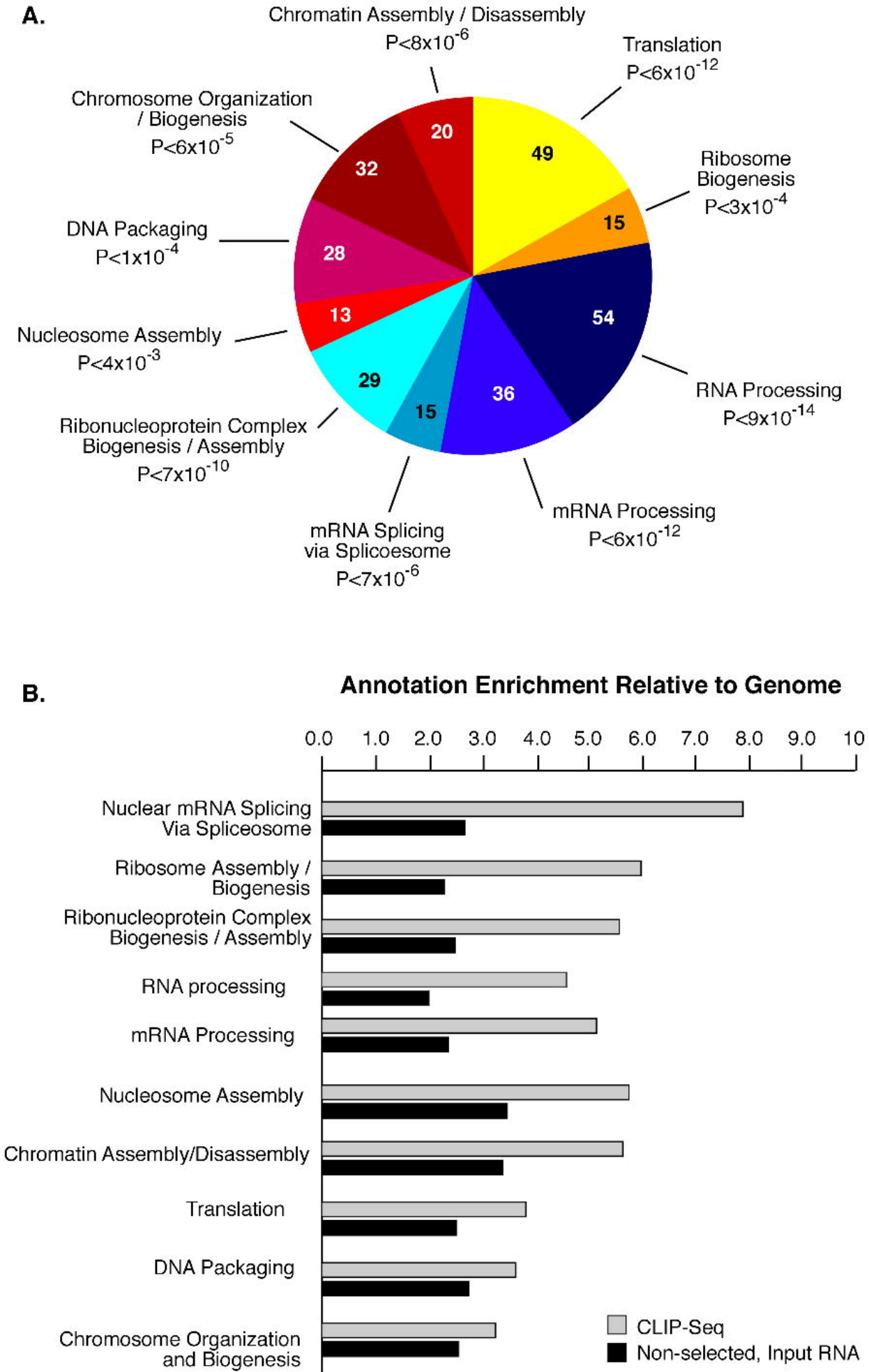
B.

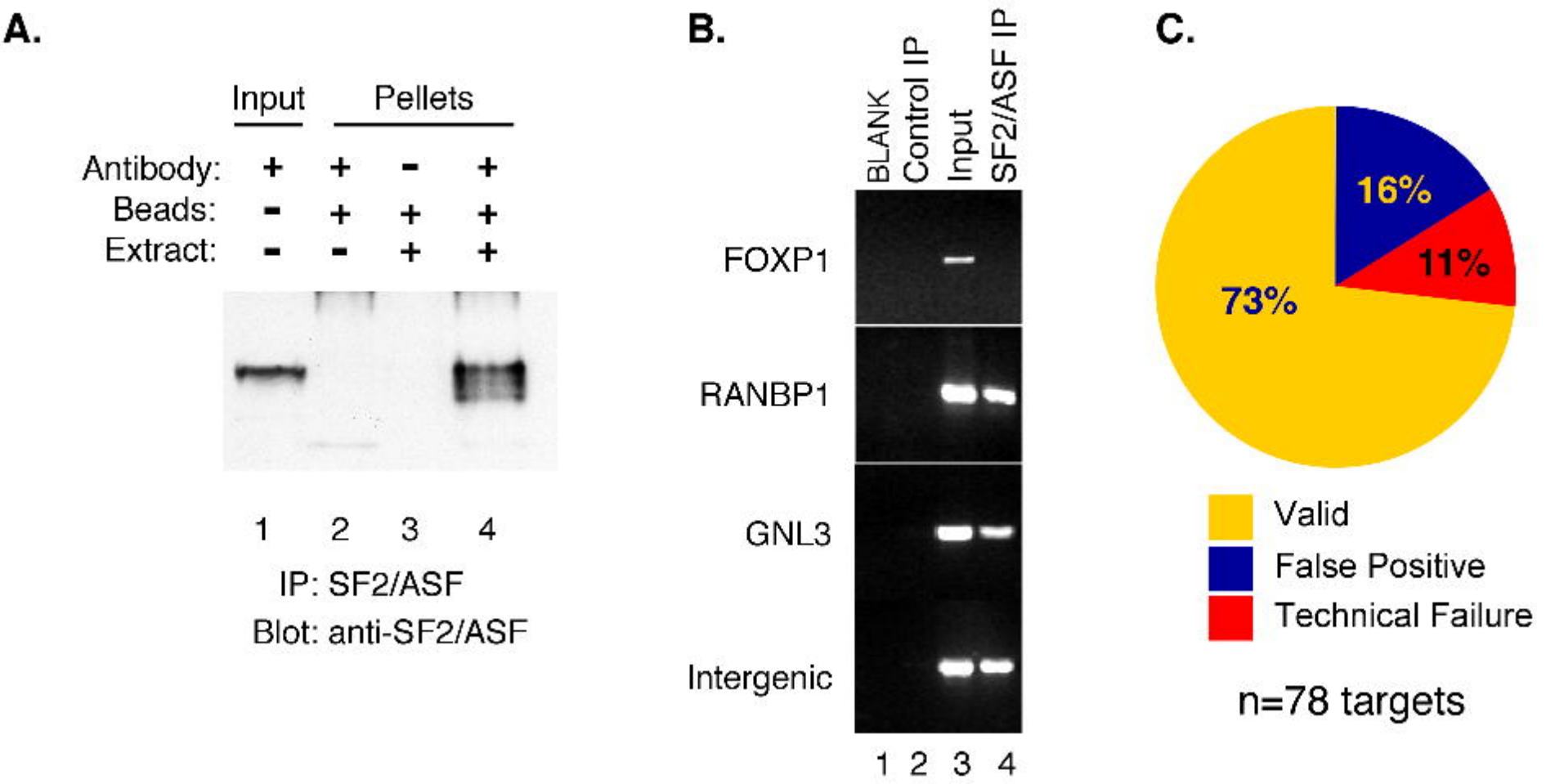

Average accuracy
0.80
0.75
0.70
0.65
0.60
0.55
0.50
-15
-10
-5
0
5
10
Matching Score Cutoff Threshold

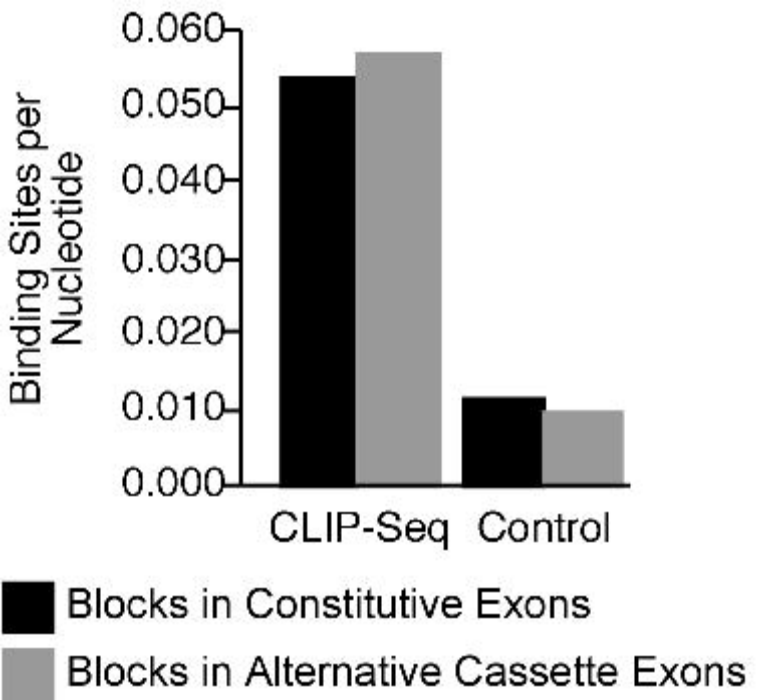
C.


Average true positive rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Average false positive rate
11.7
5.93
0.17
-5.6
11.36
17.13
Matching Score Cutoff


A.
In Gene
(Intragenic)
Out of Gene
(Intergenic)
In ncRNA
Gene
Chromosome
Protein-Coding Gene
ncRNA Gene
Gene
Exon
Intron
Exon-Exon
Junction
In Exon
In Intron
Exon-
Intron
Boundary
B.
6,128
10,532
80
2,911
681
2,245
1,065
Exon
Exon-Intron
Boundary
Exon-Exon
Junction
Intronles
Gene
Intron
Noncoding
RNA
Intergenic


A.
Chromatin Assembly / Disassembly
P<8x10-6
Translation
P<6x10-12
Chromosome Organization
/ Biogenesis
P<6x10-5
Ribosome
Biogenesis
P<3x10-4
DNA Packaging
P<1x10-4
Nucleosome Assembly
P<4x10-3
RNA Processing
P<9x10-14
Ribonucleoprotein Complex
Biogenesis / Assembly
P<7x10-10
mRNA Splicing
via Splicoesome
P<7x10-6
mRNA Processing
P<6x10-12
20
49
32
15
28
13
54
29
36
15
B.
Annotation Enrichment Relative to Genome
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10
Nuclear mRNA Splicing
Via Spliceosome
Ribosome Assembly /
Biogenesis
Ribonucleoprotein Complex
Biogenesis / Assembly
RNA processing
mRNA Processing
Nucleosome Assembly
Chromatin Assembly/Disassembly
Translation
DNA Packaging
Chromosome Organization
and Biogenesis
CLIP-Seq
Non-selected, Input RNA


A.
Input
Pellets
Antibody: + + - +
Beads: - + + +
Extract: - - + +
1 2 3 4
IP: SF2/ASF
Blot: anti-SF2/ASF
B.
BLANK
Control IP
Input
SF2/ASF IP
FOXP1
RANBP1
GNL3
Intergenic
1 2 3 4
C.
16%
11%
73%
Valid
False Positive
Technical Failure
n=78 targets


Binding Sites per Nucleotide
0.060
0.050
0.040
0.030
0.020
0.010
0.000
CLIP-Seq
Control
Blocks in Constitutive Exons
Blocks in Alternative Cassette Exons

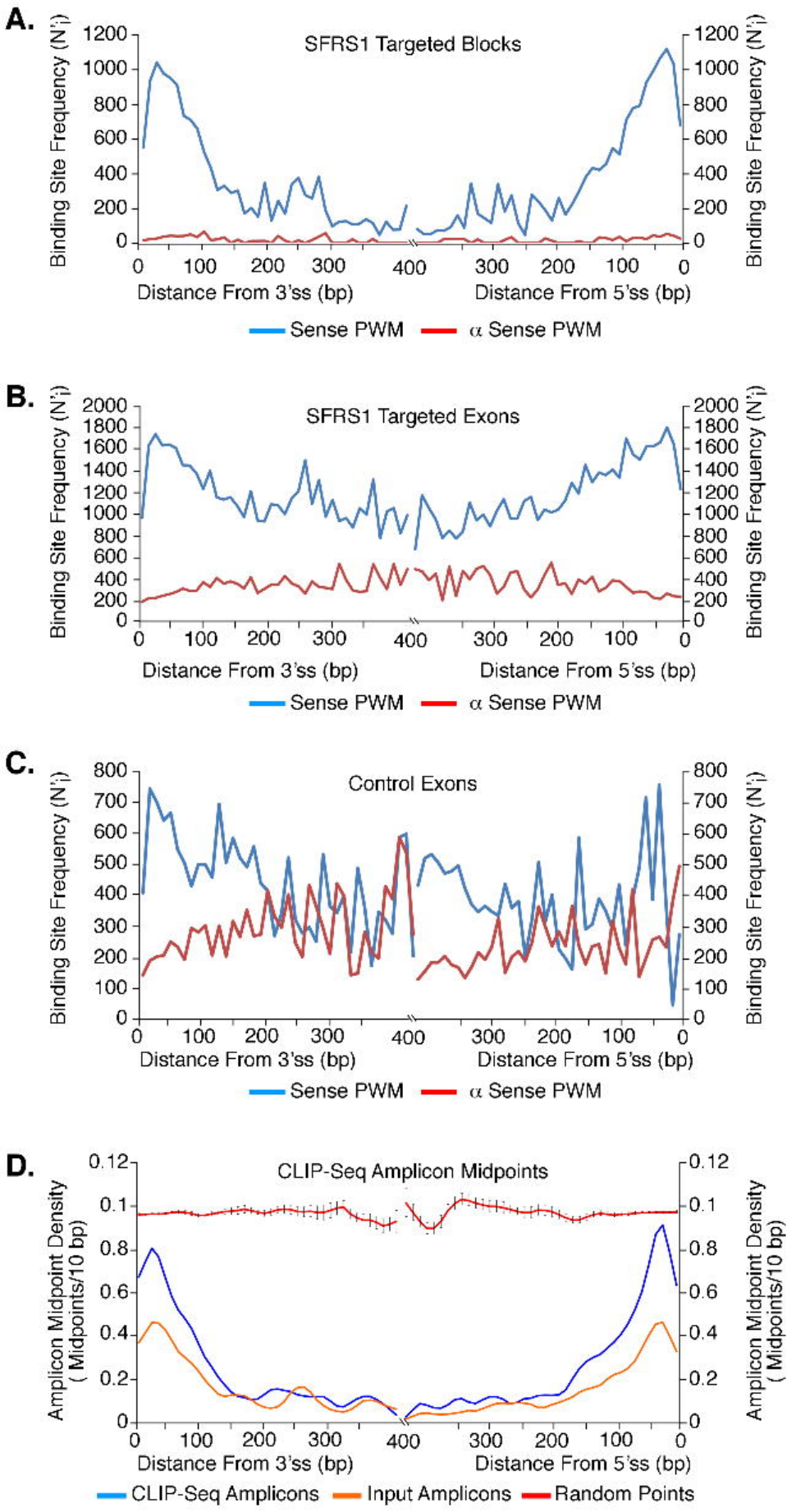
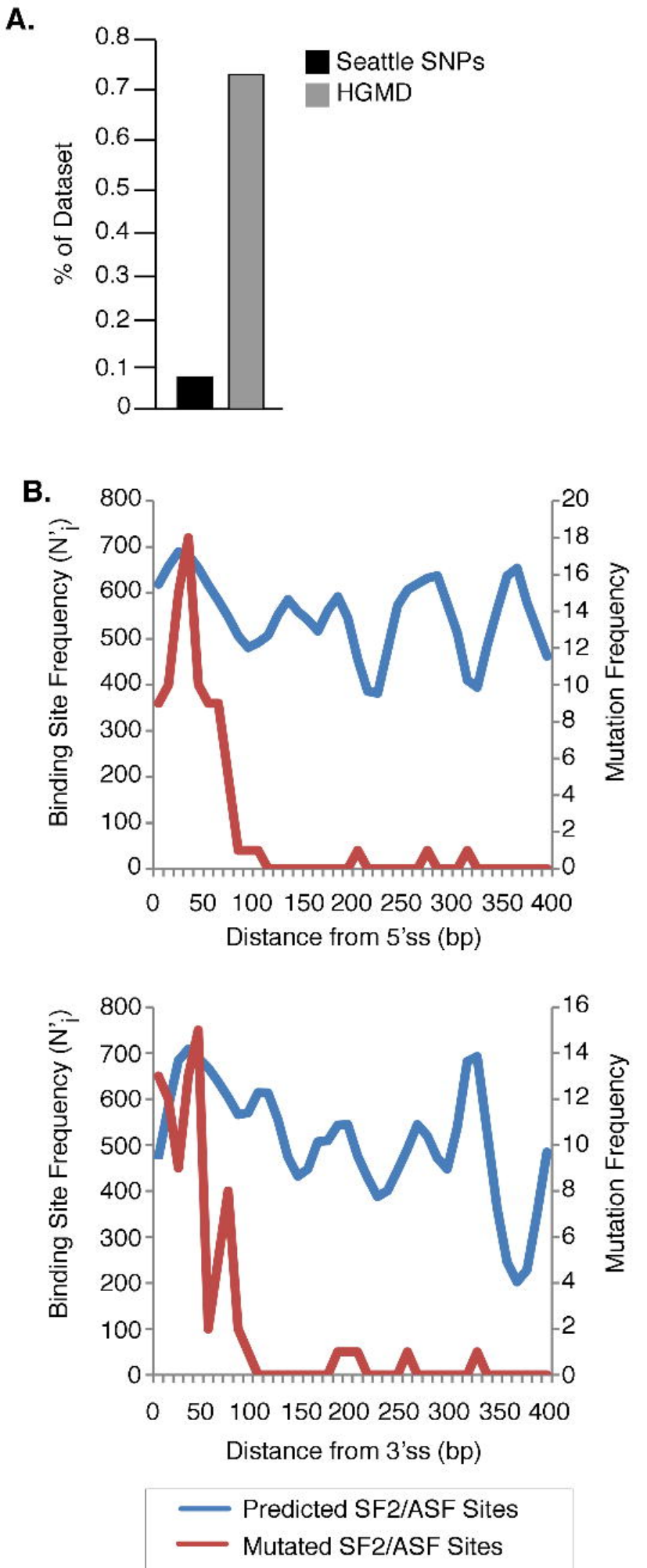

A.
SFRS1 Targeted Blocks
Binding Site Frequency (N'i)
Distance From 3'ss (bp)
Distance From 5'ss (bp)
Sense PWM
α Sense PWM
B.
SFRS1 Targeted Exons
Binding Site Frequency (N'i)
Distance From 3'ss (bp)
Distance From 5'ss (bp)
Sense PWM
α Sense PWM
C.
Control Exons
Binding Site Frequency (N'i)
Distance From 3'ss (bp)
Distance From 5'ss (bp)
Sense PWM
α Sense PWM
D.
CLIP-Seq Amplicon Midpoints
Amplicon Midpoint Density (Midpoints/10 bp)
Distance From 3'ss (bp)
Distance From 5'ss (bp)
CLIP-Seq Amplicons
Input Amplicons
Random Points

Figure 7


A.
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
% of Dataset
Seattle SNPs
HGMD
B.
Binding Site Frequency (N'i)
Mutation Frequency
800
700
600
500
400
300
200
100
0
20
18
16
14
12
10
8
6
4
2
0
0 50 100 150 200 250 300 350 400
Distance from 5'ss (bp)
Binding Site Frequency (N'i)
Mutation Frequency
800
700
600
500
400
300
200
100
0
16
14
12
10
8
6
4
2
0
0 50 100 150 200 250 300 350 400
Distance from 3'ss (bp)
Predicted SF2/ASF Sites
Mutated SF2/ASF Sites