High Throughput Binding Analysis Determines the Binding Specificity of ASF/SF2 on Alternatively Spliced Human Pre-mRNAs

B. Chang[1,§], J. Levin[2,§], W.A. Thompson[3,4] and W.G. Fairbrother[*,1,4]

[1]*Department of Molecular and Cell Biology and Biochemistry, Brown University, Providence, Rhode Island 02912, USA*

[2]*Yale School of Medicine, 333 Cedar Street, New Haven, CT 06510, USA*

[3]*Division of Applied Math, Brown University, Providence, Rhode, Island 02912, USA*

[4]*Center for Computational Molecular Biology, Brown University, Providence, Rhode, Island 02912, USA*

Abstract: High throughput immunoprecipitation studies of transcription factors and splicing factors have revolutionized the fields of transcription and splicing. Recent location studies on Nova1/2 and Fox2 have identified a set of cellular targets of these splicing factors. One problem with identifying binding sites for splicing factors arises from the transient role of RNA in gene expression. The primary role of most splicing factors is to bind pre-mRNA co-transcriptionally and participate in the extremely rapid process of splice site selection and catalysis. Pre-mRNA is a labile species with a steady state level that is three orders of magnitude less abundant than mRNA. As many splicing factors also bind mRNA to some degree, these substrates tend to dominate the output of location studies. Here we present an *in vitro* method for screening RNA protein interactions that circumvents these problems. We screen approximately 4000 alternatively spliced exons and the entire hepatitis C genome for binding of ASF/SF2, the only splicing factor demonstrated to function as an oncogene. From the pre-mRNA sequences returned in this screen we discovered physiologically relevant ASF recognition element motifs. ASF binds two motifs: a C-rich and a purine-rich motif. Comparisons with similar data derived from the hnRNP protein PTB reveal little overlap between strong PTB and ASF/SF2 sites. We illustrate how this method could be employed to screen disease alleles with the set of small molecules that have been shown to alter splicing in search for therapies for splicing diseases.

INTRODUCTION

Splice site selection occurs through the recognition of multiple elements: a conserved 5' splice site (GU), a 3' splice site (AG), and a branch point sequence and a polypyrimidine tract that occur close to the 3' splice site [1]. Since the splice sites themselves do not contain enough information to direct all the splicing in a cell, there must be other signals and factors which help to define exon/intron boundaries. These other signals can take the form of intronic splicing enhancers (ISEs) that act to define splice sites from an intronic location. Exonic splicing enhancers (ESEs) are another class of *cis*-acting element that act from an exonic location. ESEs are short motifs that can bind favorably to an SR protein to facilitate interactions with each other and the other snRNPs in the splicosome. These interactions help to promote both constitutive and alternative splicing. The short motifs (7 nucleotides) that define ESEs are somewhat degenerate.

The SR-protein family contains one or two N-terminal RNA recognition motifs (RRMs) that bind RNA and a highly conserved C-terminal arginine/serine-rich (RS) domain that is involved in protein-protein interactions [1]. One such SR protein is Alternative Splicing Factor (ASF/SF2), a well characterized 33-kD splicing activator that promotes tumor formation in mice [2, 3]. ASF/SF2 contains two RRM domains and an RS domain. Each RRM domain can independently bind to RNA; however, binding is optimal when both RRMs are involved. The RS domain is responsible for recruiting other members of the spliceosome to the splice site [4] and also in abrogating the electrostatic repulsive force associated with RNA annealing [5]. In this sense, the RS domain may also help determine binding specificity [5].

Consensus motifs of RNA that bind to ASF/SF2 have been identified. Prior experiments have been carried out using SELEX (Systematic Evolution of Ligands by Exponential Enrichment), an *in vitro* system that iteratively selects sequences from a large, randomized pool in order to derive "winner" sequences based on their binding to the protein of interest. The first model, using a recombinant form of ASF/SF2 that lacked its RS domain (ASFΔRS), produced 49 "winner" sequences after seven rounds of selection with purified protein. These sequences were rich in purine content and fitting the octomer (A/G)GAAGAAC, the decamer AGGACAGAGC, or the decamer AGGACG AAGC. Other sequences rich in purine content did not necessarily bind to ASF/SF2, showing that this binding motif was relatively specific. Additionally, when the same SELEX process was performed on ASF-RBD1 (only the first RRM of ASF/SF2), a very different motif was obtained, suggesting that both RRMs help to determine binding specificity. ASFΔRS was used because the RS domain contains a high arginine content, which would convey a posiitve charge on the protein that could lead to non-specific interactions with negatively charged RNA molecules [6]. A second model sought to improve on the previous by using functional

*Address correspondence to this author at the Department of Molecular and Cell Biology and Biochemistry, Brown University, Providence, Rhode Island 02912, USA; E-mail: fairbrother@brown.edu

[§]These authors contributed equally.

1386-2073/10 $55.00+.00 © 2010 Bentham Science Publishers Ltd.

SELEX in order to directly measure splicing activity rather than binding ability. The selection was carried out for fewer rounds, and the oligos were selected on the basis of their ability to promote splicing in S100 cell extract supplement with recombinant ASF/SF2. This technique conveyed a number of possible advantages. First, motifs determined by binding may be a subset of those determined by function. Additionally, binding is not a necessary correlate to function; in fact, in some cases, high affinity to the substrate may preclude function. The experiment produced a consensus sequence of (C/G)(A/G)(C/G)A(C/G)GA for ASF/SF2 winners [7].

Thus, both a purine-rich consensus motif, based on binding SELEX, and a C-rich consensus motif, based on functional SELEX detecting splicing activity, have been proposed. However, determining binding using SELEX suffers from two drawbacks. First, these previous models used large pools of randomized sequences to perform their assays. These sequences may be relevant, or may never occur in the genome. Thus, our selection procedure begins with a much smaller library of genomic sequences, taken from alternatively spliced exons, genes known to interact with ASF/SF2, and several viral genomes that show significant alternative splicing and the ability to interact with ASF/SF2. Secondly, the SELEX process possesses a potential disadvantage in that as successive rounds of selection are performed in order to "clean" the pool of noise, data from non-winner sequences are lost. Thus, consensus motifs may unjustly reflect only those sequences that were "most highly" enriched rather than all sequences that had been significantly enriched by binding [8].

Here, we present a novel method for identifying ASF binding sites on pre-mRNAs of interest. This approach is an RNA adaption of the MEGAshift protocol described previously [9]. We utilize a T7 tagged ASF/SF2 construct to localize ASF/SF2 binding within the vicinity of approximately 4000 alternatively spliced exons and identify and characterize physiologically relevant ASF binding motifs and sites in the genome.

RESULTS

Design and Synthesis of Oligonucleotide Pool

Our library for selection was comprised mostly of sequences found in the human genome that may participate in some form of alternative splicing. The library included: genes predicted to undergo alternatively splicing based on the ACEScan database [10]; CALCA, which codes for a pre-protein of calcitonin, which was shown to splice differentially in the presence of ASF/SF2 [11]; CASP9, which encodes for caspase-9, whose splicing was shown to be regulated by ASF/SF2 [12]; PLP, a gene coding for a transmembrane lipoprotein that has been shown to be spliced by ASF/SF2 [13]; MAPT, a gene where mutations can lead to defects in alternative splicing resulting in dementia [14]; the genomes of a number of viruses that have been shown to undergo alternative splicing, including SARS [15], HPRT [16], and hepatitis C [17]; three copies of the SFRS1 gene itself (human, dog, mouse) coding ASF/SF2, which has been shown to have alternate splice forms that inactivate the protein, and might occur by autoregulation [18]; and SELEX winner sequences for other proteins, including Eukaryotic Initiation Factor 5A [19], CELF2 (involved in RNA splicing) [20], RNA binding proteins hnrnp K and αCPK-2L (Thisted, Lyakhov *et al.* 2001), the splicing regulator TLS [21], and the tumor suppressor WT1 [22].

The oligonucleotide pool was designed by tiling over these genomic regions of interest, namely pre-mRNA regions around alternatively and constitutively spliced exons, by shifting a window of length 30 nucleotides by 10 nucleotide increments, centered on exons and extending 200 nucleotides toward their intronic flanks (Fig. **1A**). This technique was used in order to ensure full coverage, but gave enough resolution so that binding would be distinct when detected by microarray. Tiling through approximately 4,000 genomic regions resulted in 241,347 oligonucleotides. Universal primer binding sites were appended to each end of the oligonucleotides, and the resulting library of 60-mers was commercially synthesized onto a custom oligonucleotide array.

The oligonucleotide library was commercially synthesized as a custom oligonucleotide array, liberated from the slide, and then amplified *via* the universal primer binding sites at low cycle PCR (Fig. **1A**) [9]. The oligonucleotide pool was transcribed into RNA *via* the T7 RNA Polymerase promoter introduced into the forward primer.

SELECTION OF HIGH AFFINITY ASF LIGANDS

The RNA library was first incubated in whole cell extract with ASF/SF2, and then mixed with the bead/antibody solution as described above. This order of addition was chosen (as opposed to first isolating the protein, and then binding the RNA) in order to allow the protein-RNA binding to occur in whole cell extract. As prior SELEX studies have found that the high affinity ligand of ASF/SF2 was recognized by tra2 protein (not ASF/SF2) in extract, performing the binding reaction in extract will allow us to determine binding specificity in a manner that accounts for the combined action of competitors and cooperative binders that may be present in the extract. In order to prevent non-specific binding of RNA to either the magnetic beads or antibody, sonicated competitor yeast total RNA was added to the binding buffer.

In order to isolate the ASF-bound fraction of the library, we performed an immunoprecipitation of T7-tagged ASF/SF2 from 293 cell extract. Streptavidin magnetic beads were incubated with biotinylated anti-T7 antibody, which was then mixed with whole cell extract from the transfected 293 cells. After three washes with 1X PBS, the protein was eluted from the beads with 0.1 M glycine-HCl. When the elution product for ASF/SF2-transfected extract was visualized by Western blotting using the anti-T7 antibody, a distinct band was seen at 33 kD that was not seen for the elution product of extract from 293 cells that had undergone a "blank" transfection (Fig. **1B**). In order to see if the elution product was in fact pure, a protein gel was SYPRO Ruby stained to visualize all protein in the solution. Relative to the amount of protein in the starting material and flow through, very little remained in the elution material, showing the isolation of ASF/SF2 to be pure (Fig. **1B**).

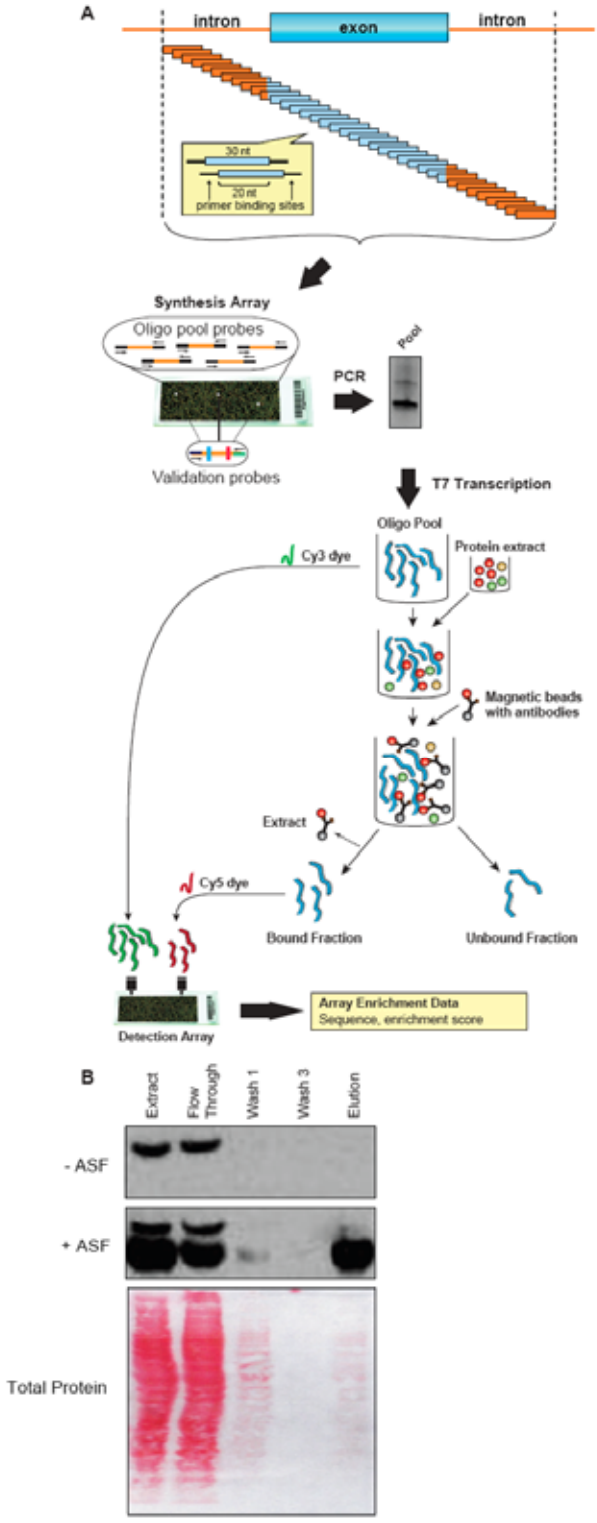

Fig. (1). Experimental scheme. (**A**) The oligo pool was designed by tiling length 30 oligonucleotides in 10 nucleotide increments across approximately 4,000 genomic regions. A total of 241,347 experimental and 90 control sequences were flanked by common primers and ordered as features on a custom microarray. Features were recovered from the array surface by scouring (Materials and Methods) and PCR amplified using the common primers. A T7 promoter was introduced *via* PCR and used to transcribe the pool into RNA. The RNA pool was partitioned into ASF-bound and -unbound fractions *via* co-immunoprecipitation from 293T nuclear extract. The starting pool was then internally labeled with Cy3 dye, and the bound fraction was labeled with Cy5. These two sets of oligos competed for binding on a two-color microarray, resulting in enrichment data. (**B**) T7-tagged ASF protein was immunoprecipitated from 293 cell extract using streptavidin magnetic beads and a biotinylated anti-T7 antibody. After three washes with 1X PBS, ASF was eluted from the beads with 0.1 M glycine-HCl. A distinct band at 33 kD was visualized by Western blotting after elution and was absent in extracts that had undergone a 'blank' transfection. SYPRO Ruby stain was used to visualize all protein in the solution. Little protein remained in the elution material, showing the immunoprecipitation of ASF to be pure.

To rank all 241,347 RNA oligos according to their ability to bind ASF/SF2, a two color array strategy was chosen. After elution and purification, selected RNA was reverse transcribed into cDNA. A longer 20-bp primer that overlapped both the 5' and 3' ends of the RNA was found to function better in this reaction than a standard 15-bp primer overlapping just the 3' end. PCR was performed using the normal 5' sense primer and a 3' antisense primer that includes a T7 promoter. The T7 promoter was used to transcribe antisense RNA that could hybridize to the verification microarray. cDNA was also produced from unbound "starting material" RNA for competitive hybridization on the array. Cy3 (green) was used to label starting material RNA; Cy5 (red) was used to label eluted RNA. After hybridization to the detection array, enrichment was measured as the ratio of oligonucleotide in the bound fraction versus that in the starting pool. Of the 241,347 sequences on the microarray, 50,161 (20.7%) showed a significant green Cy3 signal and were considered as present in the starting pool.

ANNOTATION OF GENOMIC REGIONS WITH ASF/SF2 BINDING SITES

To visualize the binding specificity of ASF/SF2, we first annotated sites on the pre-mRNA where ASF/SF2 bound (Fig. **2A**, **B**). Each nucleotide was associated with the average of the red/green ratio from all overlapping array probes (Fig. **2A**). The base-ten logs of these ratios were then plotted as a function of chromosome position. With the tiling scheme used in this experiment, three probes are expected to overlap each position. Due to incomplete probe coverage, this value was smaller, with an average of 0.61 oligos overlapping each position. In the selected example (the *CDC42BPG* gene), the polypyrimidine tract regions of exons 2 and 7 and also the intronic regions on the 3' end of exons 1, 2, and 7 appeared enriched for ASF binding sites (Fig. **2B**). The direct annotation of ASF on pre-mRNA was written as a custom annotation track for the UCSC genome browser and is available to download.

RNA MAP OF ASF/SF2

The oligos were overlaid onto a generalized model of an intron-exon-intron segment, and the log mean and median were calculated for each position along this model (Fig. **2C**). Each position on the model has an average of 49.6 oligos contributing enrichment data. While the log median RNA map shows a generally continuous enrichment score throughout the map, the log mean RNA map contains three regions of increased enrichment: -100 of the 3' splice site, at the 3' splice site, and -60 of the 5' splice site. These regions represent the top 3 enriched oligos in the data set, such that their scores were high enough to visibly raise the log average plot. These oligos, marked 1 to 3 on the map, are shown in Table **1**. The top oligo in gene PHF14 had a log enrichment score of 3.41, signaling that it was 2597.40 times more represented in the bound fraction than unbound.

DETERMINING THE BINDING SPECIFICITY OF ASF/SF2

To discover the binding specificity of ASF/SF2 on real genomic sequence, motif finding was done on the top 1% of oligos ranked by enrichment, and because of the clear separation between high affinity ligands and the bulk of the oligonucleotide population (apparent in Fig. **2D**), motif finding was also done on the top 0.1% (Fig. **3A**, **B**). From the top 0.1% (n=50) (Fig. **3A**), we discover a purine rich motif when sampling for one motif model, and also a C-rich motif when sampling for two motifs. While the purine rich motif is somewhat consistent with the purine-rich motif found in the original binding SELEX study, the C-rich motif is consistent with Krainer's functional SELEX results [6, 7]. Additionally, sampling for 4 motif models on the top 1% (Fig. **3B**) finds a motif very similar to the purine rich motif found in the top 0.1%. This result demonstrates that the dominant motif can also recovered from a lower enriched fraction of the binding experiment.

To explicity measure the agreement with prior results we scored both prior datasets with the purine and C-rich motifs discovered in this paper. While neither motif was enriched in the original binding SELEX output, both the C-rich and purine-rich motifs were enriched in the functional SELEX dataset (1.8-fold and 1.6-fold above an equally sized random set, respectively (data not shown). This is perhaps expected as the T7 tagged ASF/SF2 protein used in this study is more similar to the full length ASF/SF2 used in the functional SELEX [7] than the RS domain truncated factor used in the original study [6]. Recently, a dataset of *in vivo* ASF/SF2 targets has been published [23]. We tested for the presence of our motifs in the data from an ASF/SF2 cross-linking immunoprecipitation and high throughput sequencing (CLIP-seq) study. Using the pattern search program Patser v3e [24], we annotated the identified 23,632 CLIP tags with the highest Patser score for each of the two motifs. For the purine rich motif, 20,877 (88%) of the CLIP tags have a Patser score greater than threshold – this represents a 2.3-fold enrichment over random nucleotide sequences of the same length (Fig. **4A**). However, the C-rich motif was not significantly enriched (Fig. **4B**). While this result validates the dominant motif discovered in the *in vivo* study, the failure to identify the C-rich motif*ss* could be due to an artifact of the C-terminal tagging or an indirect binding event. The later explanation is especially plausible given prior evidence of ASF/SF2 interactions with proteins such as U2AF65 that bind polypyrimidine tracts [25].

ASF/SF2 COOPERATIVITY WITH PTB

To place ASF/SF2 in the context of other important determinants of splice site selection, we compared the map of ASF/SF2 binding to the map of PTB binding on the 4000 alternatively spliced exons under study. In other words, the binding assay was performed on the same oligo pool, but instead the oligos were selected for enrichment of polypyrimidine tract binding protein (PTB) (Reid *et al.*, submitted). The oligos that had binding enrichment data to both ASF/SF2 and PTB were extracted (n=26350), and their log enrichment scores were scatter plotted against each other. We found that 52.5% of the oligos used in the ASF/SF2 study also had enrichment data for PTB.

From these data, there was a clear lack of oligos that were highly enriched for both PTB and ASF/SF2 binding Fig. (**5**). The scatter plot shows that the sequences that were highly enriched for PTB, in the top 1% of all PTB enriched

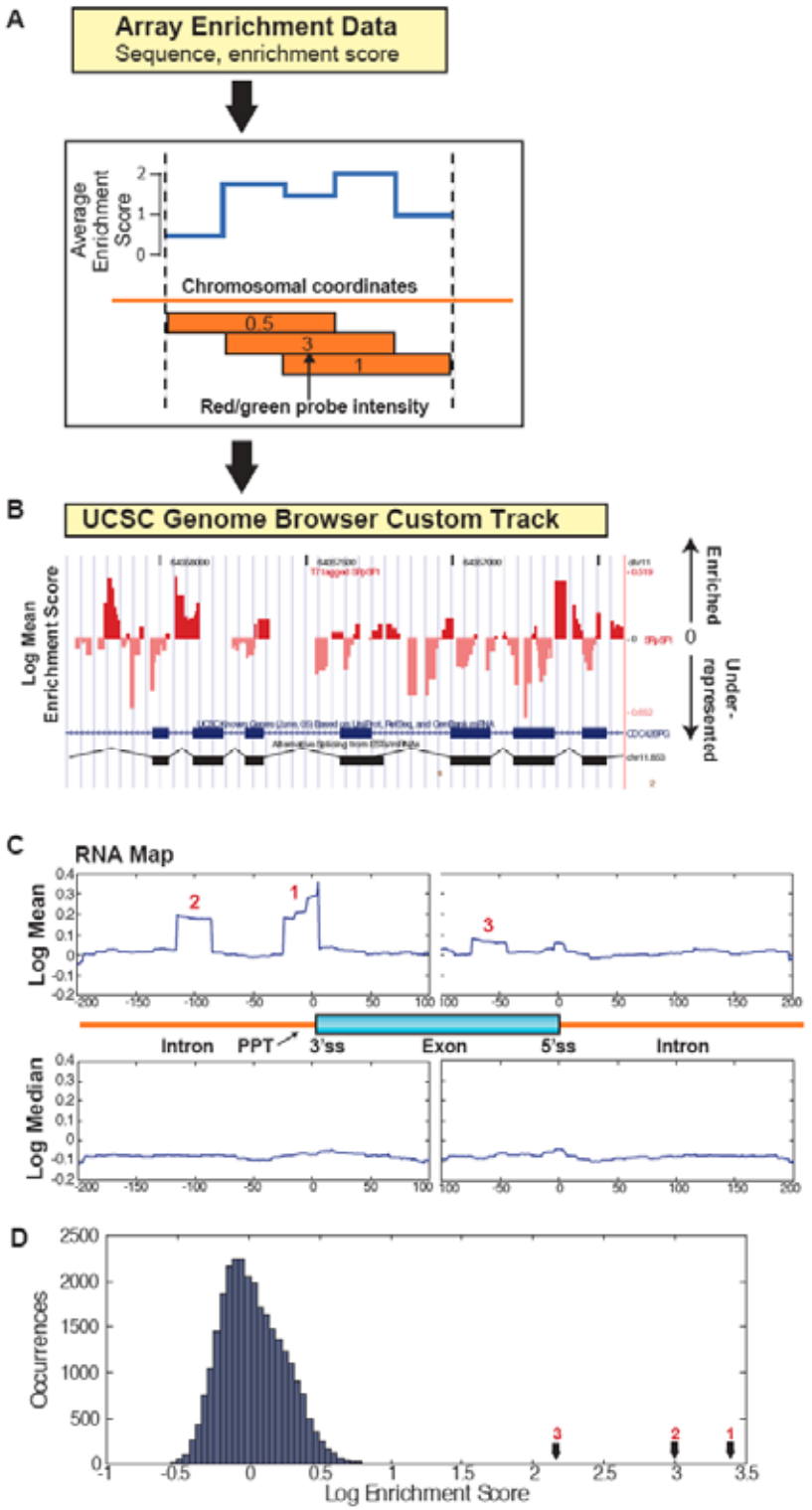

Fig. (2). Annotating the genome with enrichment data. (**A**) The array data were mapped to genomic coordinates and the scores at each location were averaged and converted to base-ten log. An illustration of this averaging step is given for 3 overlapping 30-nt oligos with scores of 0.5, 3, and 1, where the average enrichment score for each 10-nt window is graphed above. (**B**) The approximately 4,000 genomic regions were then visualized individually using a Custom Track in the UCSC Genome Browser. The example browser window given is in the *CDC42BPG* gene, where gene features (exons, introns, alternative splicing, etc.) are given along the bottom, and log average enrichment scores from the bound oligos are represented by red vertical bars. (**C**) A generalized RNA map for ASF enrichment data was made by compiling the information from all ~4,000 regions into one map. The enrichment scores for each oligo were sorted based on their distance from splice sites (x-axis). Both log mean and log median enrichment scores (y-axis) were calculated and plotted for each position 200 bases into the intron and 100 bases into the exon from each splice site. Each of these results in a single map encompassing the entire enrichment data set. The 1, 2, and 3 on the log mean map refer to the three top scoring oligos in Table **1**. (**D**) A histogram was created from the enrichment scores of all the oligos. The top 3 scoring oligos are indicated.

Table 1. Top 3 Oligos Ranked by ASF/SF2 Enrichment Score

Rank	Chromosomal Coordinate	Gene ID	Sequence	Log Enrichment Score
1	chr7:10795214-10795243	PHF14	TTTTTTTTTCTCATATTTTCAACAGATTCT	3.41
2	chr1:10650222-10650251	CASZ1	GGGTTCTTTCTAGGGAGACCTGAGGCCAGC	3.02
3	chr1:159058511-159058540	CAPON	CTCACCACCAGATGCAGCTCCTCCAGCAGC	2.16

sequences, were also slightly enriched for ASF/SF2. While only 12.7% of this segment of oligos have ASF scores in the top 1% of ASF/SF2 bound oligos, these oligos have an average log ASF/SF2 enrichment score of 0.41, indicating a 2.59 times overrepresentation in the bound fraction *vs* unbound for ASF/SF2.

Oligos in the top 1% of ASF/SF2 bound sequences, on the other hand, were not correlated with strong PTB enrichment. This result is consistent with the general models of hnRNP working antagonistically against SR proteins to define splice site selection.

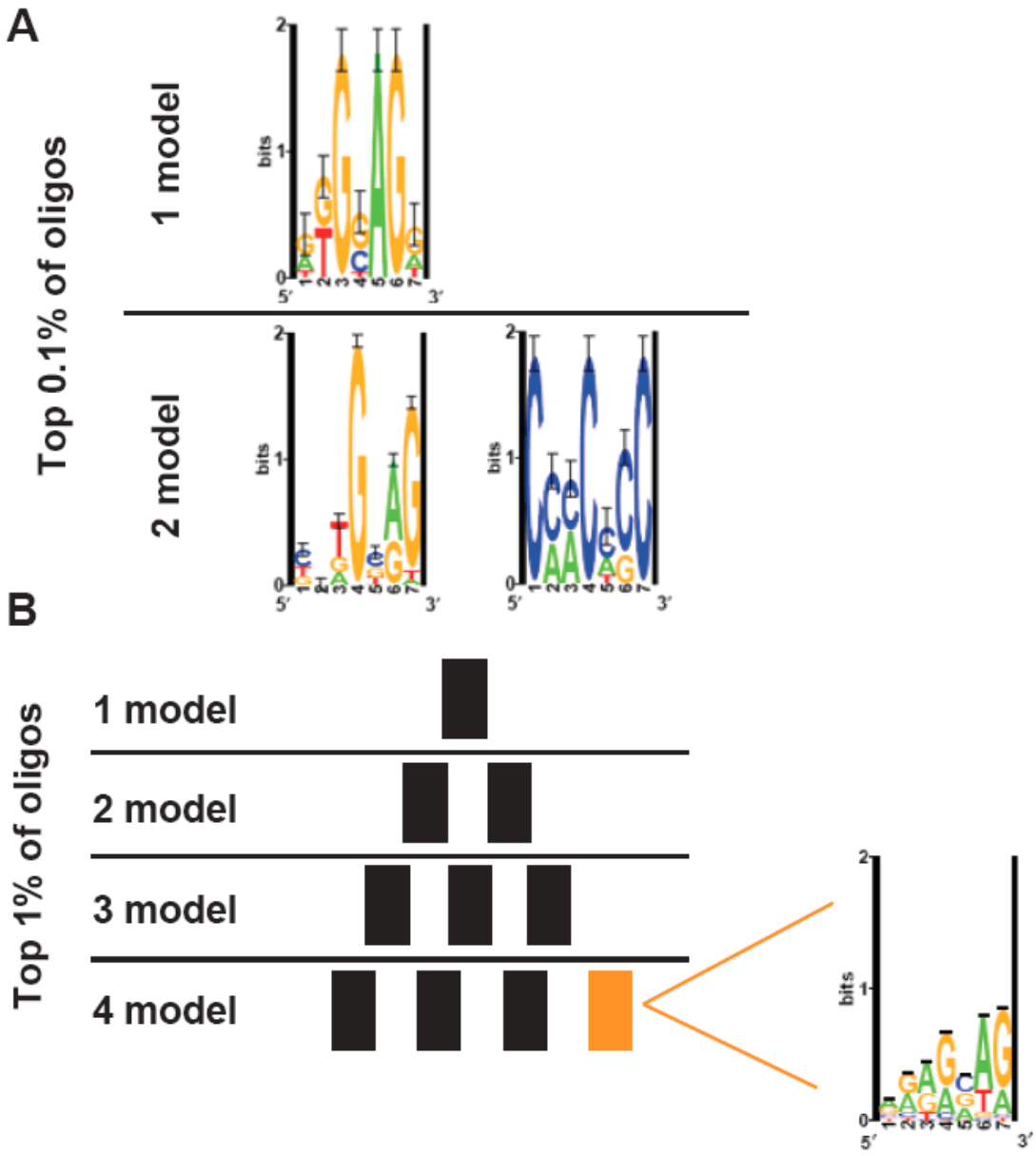

Fig. (3). Discovering and validating ASF/SF2 binding specificity. (**A**) Motif finding was run for the top 0.1% of oligos ranked by enrichment sampling for both 1 and 2 motif models. (**B**) The same process was done for the top 1% of enriched data from 1 to 4 motif models. The shown motif was found when sampling for 4 motif models.

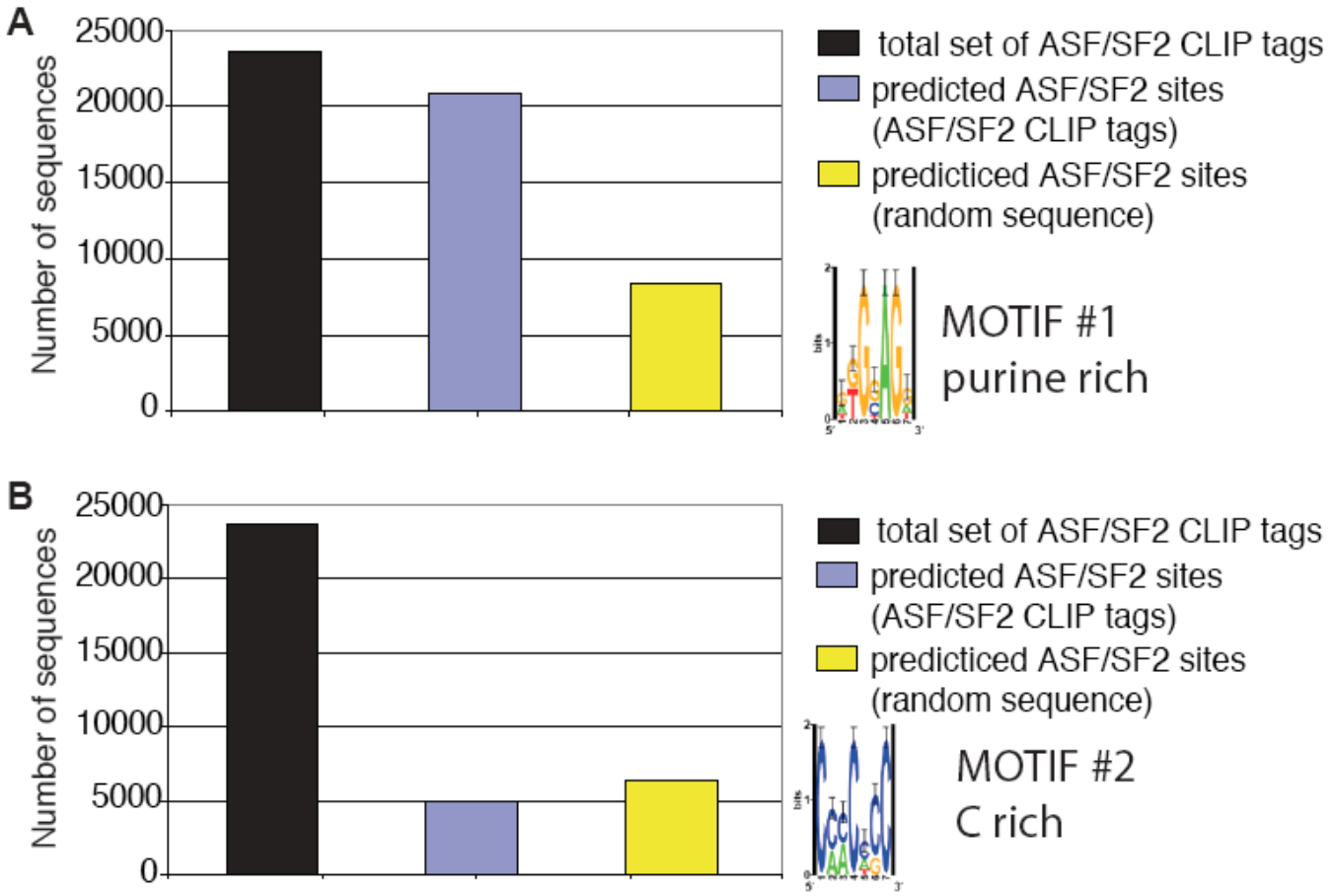

Fig. (4). Presence of motifs in CLIP-seq data. (**A**) The 23,632 identified binding sites by Sanford *et al*. were annotated for presence of the purine-rich motif. For comparison, the presence of the motif was also tested in the shuffled binding site sequences and in random nucleotide sequences of the same length. (**B**) The same procedure was done for the C-rich motif.

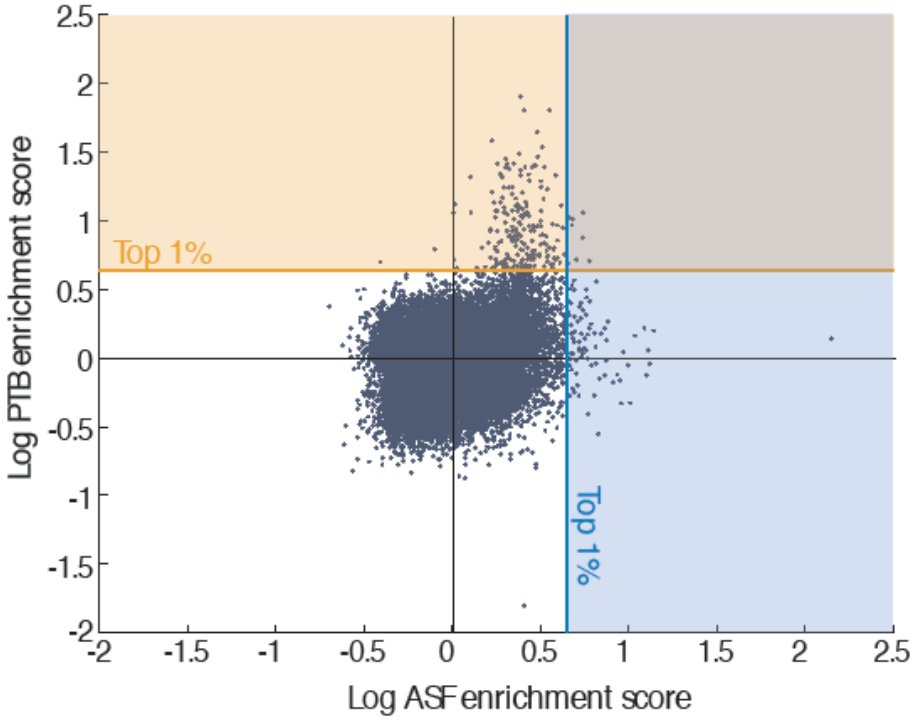

Fig. (5). Comparison of ASF/SF2 and PTB binding. For oligos that had binding enrichment data for both ASF/SF2 and PTB, the log enrichment scores for both proteins are compared on a scatter plot.

Table 2. Small Molecules that Influence Alternative Splice Site Selection

Class	Molecule Name	Mechanism	Regulated Exon
HDAC inhibitor	Sodium butyrate	HDAC inhibitor	*SMN2* exon 7
HDAC inhibitor	Valproic acid	HDAC inhibitor	*SMN2* exon 7
Kinase inhibitor	Aclarucibin	Topo I	*SMN2* exon 7
Phosphatase inhibitor	*N*-(Hexanoyl)sphingosine (C_6-ceramide)	PP1 inhibitor	*BCL-X* and *CASP-9*
Protein-protein interactions	10-Chloro-2,6-dimethyl-2H-pyrido[3',4':4,5]pyrrolo[2,3-g]isoquinoline (IDC16)	SR-protein interaction	HIV-1 mRNA
cAMP pathway	2-(t-Butylamino)-1-(4-hydroxy-3-hydroxymethylphenyl)ethanol sulfate (salbutamol)	Adrenergic antagonist	*SMN2* exon 7
Coupling of transcription and splicing	Dexamethasone	Coupling of transcription and splicing	Insulin receptor mRNA

Table 3. ASF/SF2 Enriched Sequences in Hepatitis C. The Hepatitis C Genome was Tiled Over in the Original Oligo Pool, and the Top 3 Enriched Sequences were Extracted

HepC Coordinate	Sequence	Log Enrichment Score
4760-4789	ACAACCACGCTCCCCCAGGATGCTGTCTCC	0.51
4770-4799	TCCCCCAGGATGCTGTCTCCAGGACTCAAC	0.47
9030-9059	ACTCTCCAGGTGAAATCAATAGGGTGGCCG	0.47

DISCUSSION

Here we report a method of locating the binding sites of the oncogenic splicing factor, ASF/SF2 in alternatively spliced pre-mRNA. This method identifies ligands of the RNA binding protein which can then be subjected to further study. By selectively applying Gibbs Sampling to the bound fraction of the pool, we identified two binding motifs, perhaps correlating with the specificities of ASF/SF2 two RRMs. One motif bears similarity to the results of [6], while the other is similar to the results of [7]. The binding model presented here predicts ASF/SF2 binding in 88% of the recently reported ASF/SF2 binding sites. While these two motifs correspond to the highest affinity sites, we find that these motifs can also be recovered in a less enriched fraction.

In the future, we wish to apply this method towards the identification of disease alleles that alter splicing. Mutations that disrupt splicing play a larger role in hereditary disease than other non-coding lesions. For example, the Human Gene Mutation Database reports that 6,428 hereditary disease alleles disrupt splicing elements, whereas only ~900 alleles are reported to disrupt transcriptional elements [26]. There are numerous examples of splicing mutations that disrupt splicing elements outside the classic 5'ss, 3'ss and branchpoints in cancer related genes like BRCA1, NF1, ATM, and others [27-29].

In addition to mutations that create aberrant isoforms, many studies link RNA processing shifts towards particular spliced isoforms to neoplastic disease. A clear example of this phenomena is CD44, a transmembrane glycoprotein that is heavily alternatively spliced - 10 of its 20 exons are variably included, and certain combinations appear restricted to certain tumors and perhaps even certain cancer prognoses [30]. Along these lines, several genes that affect cell growth, adhesion, migration, invasion, and apoptosis have alternative isoforms that significantly upregulate or abate these functions, can transform cells and are upregulated in tumors. It is important to note that within this set of alternative splicing events, there are targets of the oncogenic splicing factor, ASF/SF2 (e.g., caspase 9, BCL-X, prolactin receptor, Ron, Rac1, fibronectin, FGFR, MDM2, and IIp45). In addition to these results from candidate gene studies, recent genome wide comparisons of alignments of EST from tumor lines and normal tissue have identified an additional 383 cancer specific splicing events by statistically conservative criteria [31].

Recent effort has been put into the identification of small molecules that influence splice site selection [32]. Several examples of these molecules are given in Table **2**. Each of these compounds is a potential therapy for splicing related diseases. Furthermore, SR proteins are often implicated in export and other processes necessary for successful viral replication. Compounds that inhibit particular viral RNA/protein interactions could also serve as an antiviral therapy. Along these lines, we have mapped several ASF/SF2 binding sites in hepatitis C (Table **3**). In the future, we look forward to implementing this binding screen on the set of pre-mRNAs known to be misspliced in disease states. Performing this assay in the presence and absence of small molecules listed in Table **2** should identify particular drugs that may ameliorate or even reverse particular splicing disorders.

METHODS AND MATERIALS

Design of the Array Library and Pool Recovery

We obtained the sequences of genes of interest (including known alternatively spliced exons, computationally-derived

exons from the ACEScan database, constitutively spliced exons, the SF2/ASF gene from human, mouse, and dog, and various viral RNA genomes) from the UCSC Genome Browser (http://www.genome.ucsc.edu), and subjected the sequences to a tiling scheme. Tiling of the 30-mers begins at splice sites, with the 30-mer covering the splice site, and extends in both directions. Resolution is set at 10 base pairs, such that each subsequent 30-mer overlaps 20 base pairs with the previous 30-mer. Tiling extends 100 into the flanking exonic region and 200 into the flanking intronic region. In total, our array library consisted of 241,347 oligos covering approximately 4,000 exons.

DNA was recovered from synthesis arrays by adding 500 μL dH2O to the surface of the array and either thoroughly scouring and resuspending using a sterile 25-guage hypodermic needle or placing in a hybridization chamber and boiling for 1 h. The samples were then sonicated at 50% amplitude for 3 5-second pulses in a Sonic Dismembrator Model 500 (Fisher). Pools were amplified by low cycle PCR (1 min at 94 °C, 20 s at 55 °C, 1:00 at 72 °C first round; 10 s, 20 s, 10 s at each respective temperature for subsequent rounds; final elongation step of 5 min at 72 °C).

ASF PRODUCTION AND IMMUNOPRECIPITATION

The pCGT7-SF2 plasmid contains the gene for the ASF/SF2 protein, under control of the strong CMV enhancer/promoter. The ASF/SF2 also contains an N-terminal epitope tag, MASMTGGQQMG, known as the T7 tag since it corresponds to the first 11 residues of the bacteriophage T7 10 caspid protein [33].

These pCGT7-SF2 plasmids were transfected into 293T cells that had been growing on 150 mm plates in Dulbecco's modified Eagle's medium (DMEM, Gibco) supplemented with 10% fetal bovine serum (FBS, Hyclone) and incubated at 37°C in the presence of 5% CO_2. The transfection procedure begins by shocking the cells for 1 h in pure DMEM (no FBS). Approximately 24 μg of DNA was added to 60 μL Lipofectamine 2000 (Invitrogen) in the presence of DMEM; these cells were allowed to sit in this solution for 5 h, after which the medium was changed to DMEM with 10% FBS.

Forty-eight hours after transfection, the cells were washed with cold PBS and scraped from their plates. The cells were then centrifuged for 3 min at 14K at 4°C. The supernatant was discarded, and the cells were resuspended in 0.5 mL extraction buffer (200 mM KCl, 100 mM Tris, pH 8.0, 0.2 mM EDTA, 0.1 % NP-40, 10% glycerol) with 1 μL PLAC/mL buffer added and sat on ice for 50 min. The cells were then centrifuged for 10 min at 14K at 4°C, after which the supernatant was transferred to a new tube and snap-frozen by liquid nitrogen for storage.

ASF/SF2 could be detected by Western blotting on the cell extract with a biotinylated anti-T7 monoclonal antibody (Novagen), mouse HRP secondary antibody, and Western Lighting Chemiluminescence Reagent (Perkin Elmer). The protein produced a band around 33 kD. To isolate ASF/SF2 from the extract, 1 mL streptavidin magnetic beads (Roche) were washed three times with 1X PBS and then incubated with 1 μg of biotinylated antibody for 1 h at 4°C. After three more washes, the beads were incubated for 30 min in 200 μL binding buffer (200 mM KCl, 100 mM Tris base, pH 8.0, 0.2 mM EDTA, 0.1% NP-40, 10% glycerol, 440 mM yeast total RNA). After several failed attempts, it was found that the yeast total RNA (Sigma), which was included as a "cold" competitor for non-specific interactions between RNA and the antibody or the magnetic bead, had to be sonicated before usage. It is possible that large strands of RNA can take upon conformations around the protein or antibody such that it blocks binding of the ASF/SF2 to the antibody. The beads were then mixed with 50 μL of extract and 150 μL of binding buffer. The mixture was incubated for 1 h at 4°C and then washed three times with 1X PBS to discard anything not bound to the magnetic beads. Two washes with 0.1M glycine-HCl eluted any protein that bound to the beads.

SELECTION AND AMPLIFICATION OF LIGAND BOUND RNA

To the above isolation procedure, 20 μg of RNA was incubated with 50 μL of 293 extract in 130 μL of binding buffer for 30 min at 4°C [6]. This was then added to the beads after 30 min, and the mixture was incubated for 1 h at 4°C. Elution was again performed using 0.1M glycine-HCl. RNA was extracted from this elute by phenol-chloroform and ethanol precipitation.

This RNA was then reverse transcribed using ArrayScript™ M-MLV reverse transcriptase (Ambion). It was found that a longer 20-bp primer that overlapped both the 5' and 3' ends of the RNA worked better in this reaction than a standard 15-bp primer overlapping just the 3' end. Following reverse transcription, 20 cycles of PCR using iProof DNA polymerase (BioRad) was used to make the cDNA. The PCR step was performed using an antisense primer that includes the T7 promoter. This would allow transcription in the following step to produce antisense mRNA, which could bind to the verification array (whose sequences are sense). cDNA was also produced from "starting material" RNA; that is, RNA that had not gone through the selection procedure.

ARRAY HYBRIDIZATION AND UCSC GENOME BROWSER ANALYSIS

RNA that precipitated with ASF/SF2 as well as the pre enrichment starting pool was transcribed from the cDNA using the common flanking priers containing a T7 polymerase promoter. The oligos were labeled with Cy5 and Cy3 dyes, respectively. The MEGAshortscript™ transcription kit was used (Ambion), using 1 μL of 5-(3-aminoallyl)-UTP (Ambion) and no regular UTP. Monoreactive Cy 3 and Cy5 dyes (GE Healthcare) were prepared by mixing them with 45 μL DMSO. To the RNA product, 4.5 μl of coupling buffer (0.1 M Na_2CO_3), 2.5 μL water, and 3 μL of prepared dye were added. The mixture was incubated at room temperature for 1 h; the reaction was terminated by incubation with 6 μL of 4 M hydroxylamine for 15 min. The RNA was then extracted by phenol chloroform and ethanol precipitation.

The following was used as a hybridization solution: 50 μL Blocking Buffer, 30 μL of starting RNA / 45 μL of elution RNA (corresponding to 750 ng of RNA), 10 μl 25 X Fragmentation Buffer, 250 μL 2X Hybridization buffer, and

water up to 500 μL (all buffers by Agilent). This was then injected in the array chamber and incubated at 50°C for 3 h.

The array was scanned on an Axon 4000B scanner and then gridded with Agilent's Feature Extraction software. Only sequences that scored 1 for the "gIsWellAboveBG" value were used for analysis – this corresponds to sequences that showed green (starting) signal at least 2.6 standard deviations higher than the mean calculated background signal. Of 241,347 total sequences, 50,161 sequences fit this criterion, or 20.7% of the array.

The enrichment values were visualized using the UCSC Genome Browser. A wiggle-format custom track was created that performs the enrichment score averaging step across all the genomic regions (Fig. **2**). For each position, the log of the average of enrichment scores is taken for every oligo that overlaps that position, such that each 30-mer has data for 30 positions. A single false scaling data point is added to each continuous genomic region for the purpose of keeping the y-axis constant between different regions.

The data, scripts, and documentation for this project are available for download [34].

Generation of RNA Map

Genomic intron and exon positions were obtained using the Known Genes track from the UCSC Genome Browser [35]. BLAST [36] was used to map each oligo to its nearest splice site. A log mean and log median enrichment score was calculated from overlapping oligos for each 100 bases at the 5' and 3' ends of the exon and 200 intronic bases at the 5' and 3' ends.

Motif Finding and Annotation

Binding motifs were identified using the Gibbs Sampler (V 3.04.006) [37]. Motifs were visualized using SeqLogo [38].

Using Motifs to Score Sequences

Patser v3e was used to score the results from the Sanford *et al* CLIP-seq study [24] and counted the sequences that had a score greater than 0. We shuffled the sequences using an implementation of Altschul's shuffling algorithm [39]. For both the shuffled and random sequence data, we ran 100 trials and averaged the results.

ACKNOWLEDGEMENTS

We would like to thank all members of the Fairbrother lab for help in this performing this work and useful discussions. We would especially like to thank Mathew Gemberling for his early work on developing the oligonucleotide library. We would also like to thank the Brown University UTRA program for summer support (JL) and Martin Maxey and the NSF-UBM for summer support of BC (DUE-0734234). This work was partially supported by a CCMB Scholarship Award (WF) and an NIH COBRE award (WT). The ASF/SF2 plasmid pCGT7-SF2 was generously donated by Javier Cáceres.

REFERENCES

[1] Hastings, M.L.; Krainer, A.R. Pre-mRNA splicing in the new millennium. *Curr. Opin. Cell Biol.,* **2001**, *13*, 302-309.

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

[3] Krainer, A.R.; Conway, G.C.; Kozak, D. The essential pre-mRNA splicing factor SF2 influences 5' splice site selection by activating proximal sites. *Cell,* **1990**, *62*, 35-42.

[4] Zuo, P.; Manley, J.L. Functional domains of the human splicing factor ASF/SF2. *EMBO J.,* **1993**, *12*, 4727-4737.

[5] Shen, H.; Green, M.R. RS domains contact splicing signals and promote splicing by a common mechanism in yeast through humans. *Genes Dev.,* **2006**, *20*, 1755-1765.

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

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

[8] Irvine, D.; Tuerk, C.; Gold, L. SELEXION. Systematic evolution of ligands by exponential enrichment with integrated optimization by non-linear analysis. J. Mol. Biol., 1991, *222*, 739-761.

[9] Tantin, D.; Gemberling, M.; Callister, C.; Fairbrother, W. High throughput biochemical analysis of *in vivo* location data reveals novel distinct classes of POU5F1(Oct4)/DNA complexes. *Genome Res.,* **2008**, *18*, 631-639.

[10] Han, K.; Yeo, G.; An, P.; Burge, C.B.; Grabowski, P.J. A combinatorial code for splicing silencing: UAGG and GGGG motifs. *PLoS Biol.,* **2005**, *3*, e158.

[11] Bai, Y.; Lee, D.; Yu, T.; Chasin, L.A. Control of 3' splice site choice *in vivo* by ASF/SF2 and hnRNP A1. *Nucleic Acids Res.,* **1999**, *27*, 1126-1134.

[12] Massiello, A.; Chalfant, C.E. SRp30a (ASF/SF2) regulates the alternative splicing of caspase-9 pre-mRNA and is required for ceramide-responsiveness. *J. Lipid Res.,* 2006, *47*, 892-897.

[13] Wang, Z.; Xiao, X.; Van Nostrand, E.; Burge, C.B. General and specific functions of exonic splicing silencers in splicing control. *Mol. Cell*, **2006**, *23*, 61-70.

[14] Garcia-Blanco, M.A.; Baraniak, A.P.; Lasda, E.L. Alternative splicing in disease and therapy. *Nat. Biotechnol.,* **2004**, *22*, 535-546.

[15] Chang, C.Y.; Hong, W.W.; Chong, P.; Wu, S.C. Influence of intron and exon splicing enhancers on mammalian cell expression of a truncated spike protein of SARS-CoV and its implication for subunit vaccine development. *Vaccine,* **2006**, *24*, 1132-1141.

[16] Andersson, B.; Hou, S.M.; Lambert, B. Mutations causing defective splicing in the human hprt gene. Environ. Mol. *Mutagen,* **1992**, *20*, 89-95.

[17] Cohen, J. The scientific challenge of hepatitis C. Science, **1999**, *285*, 26-30.

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

[19] Xu, A.; Chen, K.Y. Hypusine is required for a sequence-specific interaction of eukaryotic initiation factor 5A with postsystematic evolution of ligands by exponential enrichment RNA. *J. Biol. Chem.,* **2001**, *276*, 2555-2561.

[20] Marquis, J.; Paillard, L.; Audic, Y.; Cosson, B.; Danos, O.; Le Bec, C.; Osborne, H.B. CUG-BP1/CELF1 requires UGU-rich sequences for high-affinity binding. *Biochem. J.,* **2006**, *400*, 291-301.

[21] Lerga, A.; Hallier, M.; Delva, L.; Orvain, C.; Gallais, I.; Marie, J.; Moreau-Gachelin, F. Identification of an RNA binding specificity for the potential splicing factor TLS. *J. Biol. Chem.,* **2001**, *276*, 6807-6816.

[22] Bardeesy, N.; Pelletier, J. Overlapping RNA and DNA binding domains of the wt1 tumor suppressor gene product. *Nucleic Acids Res.,* **1998**, *26*, 1784-1792.

[23] Sanford, J.R.; Wang, X.; Mort, M.; Vanduyn, N.; Cooper, D.N.; Mooney, S.D.; Edenberg, H.J.; Liu, Y. Splicing factor SFRS1

recognizes a functionally diverse landscape of RNA transcripts. *Genome Res.*, **2009**, *19*, 381-394.

[24] Hertz, G.Z.; Stormo, G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **1999**, *15*, 563-577.

[25] Wu, J.Y.; Maniatis, T. Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell*, **1993**, *75*, 1061-1070.

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

[27] Ars, E.; Serra, E.; Garcia, J.; Kruyer, H.; Gaona, A.; Lazaro, C.; Estivill, X. Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. *Hum. Mol. Genet.*, **2000**, *9*, 237-247.

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

[29] Telatar, M.; Teraoka, S.; Wang, Z.; Chun, H.H.; Liang, T.; Castellvi-Bel, S.; Udar, N.; Borresen-Dale, A.L.; Chessa, L.; Bernatowska-Matuszkiewicz, E.; Porras O, Watanabe, M.; Junker, A.; Concannon, P.; Gatti, RA. Ataxia-telangiectasia: identification and detection of founder-effect mutations in the ATM gene in ethnic populations. *Am. J. Hum. Genet.*, **1998**, *62*, 86-97.

[30] Naor, D.; Nedvetzki, S.; Golan, I.; Melnik, L.; Faitelson, Y. CD44 in cancer. Crit. Rev. Clin. Lab. Sci., **2002**, *39*, 527-579.

[31] Hui, L.; Zhang, X.; Wu, X.; Lin, Z..; Wang, Q.; Li, Y.; Hu, G. Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. *Oncogene*, **2004**, *23*, 3013-3023.

[32] Sumanasekera, C.; Watt, D.S.; Stamm, S. Substances that can change alternative splice-site selection. *Biochem. Soc. Trans.*, **2008**, *36*, 483-490.

[33] Cazalla, D.; Sanford, J.R.; Caceres, J.F. A rapid and efficient protocol to purify biologically active recombinant proteins from mammalian cells. *Protein Expr. Purif.*, **2005**, *42*, 54-58.

[34] http://fairbrother.biomed.brown.edu/data/asf/.

[35] Karolchik, D.; Kuhn, R.M.; Baertsch, R.; Barber, G.P.; Clawson, H.; Diekhans, M.; Giardine, B.; Harte, R.A.; Hinrichs, A.S.; Hsu, F.; Kober, KM.; Miller, W.; Pedersen, J.S.; Pohl, A.; Raney, B.J.; Rhead, B.; Rosenbloom, K.R.; Smith, K.E.; Stanke, M.; Thakkapallayil, A.; Trumbower, H.; Wang, T.; Zweig, A.S.; Haussler, D.; Kent, WJ. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.*, **2008**, *36*, D773-779.

[36] Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **1997**, *25*, 3389-3402.

[37] Thompson, W.; Conlan, S.; McCue, L.A.; Lawrence, C.E. Using the Gibbs Motif Sampler for Phylogenetic Footprinting. In Methods in Molecular Biology, Comparative Genomics, N. Bergman, ed. (Humana Press), **2007**, pp. 403-423.

[38] Schneider, T.D.; Stephens, R.M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **1990**, *18*, 6097-6100.

[39] Altschul, S.F.; Erickson, B.W. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Mol. Biol. Evol.*, **1985**, *2*, 526-538.

Received: May 1, 2009 **Revised: August 8, 2009** **Accepted: August 13, 2009**