Published in final edited form as:
Methods. 2014 February ; 65(3): 350–358. doi:10.1016/j.ymeth.2013.08.019.

Systematical identification of splicing regulatory *cis*-elements and cognate *trans*-factors

Yang Wang and **Zefeng Wang**
Department of Pharmacology, Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599

Abstract

The majority of human genes undergo alternative splicing to generate multiple isoforms with distinct functions. This process is generally controlled by *cis*-acting splicing regulatory elements (SREs) that recruit *trans*-acting factors to promote or inhibit the use of nearby splice sites. The growing interest in understanding the regulatory rules of splicing necessitates the systematic identification of these SREs and their cognate protein factors using experimental and computational approaches. Here we describe a strategy to identify and analyze both *cis*-acting SREs and *trans*-acting splicing factors. This strategy involves a cell-based screen to identify SREs from a random sequences library and a modified RNA affinity purification approach to unbiasedly identify the splicing factors. These methods can be adopted to identify splicing enhancers or silencers in both exons and introns, and can be extended to different cultured cells. The resulting SREs and splicing factors can be further analyzed with a series of computational and experimental approaches. This approach will help us to collect a molecular part-list for splicing regulation, providing a rich data source that enables a better understanding of the "splicing code".

Keywords

Splicing regulatory elements; splicing factors; exonic splicing silencers; intronic splicing enhancers; intronic splicing silencers

1. Introduction

Alternative splicing was found in more than 90% of human genes, serving as a major mechanism to increase protein diversity [1–3]. This process is tightly regulated in different tissues and developmental stages, and dysregulation of splicing is a common cause of various human diseases [4]. The specificity of alternative splicing is mainly determined by the sequence at the exon/intron boundary, known as splice sites and branch point sequence. However, these sequences only contain half of the information required for the accurate exon/intron recognition, suggesting additional information is needed to define splicing specificity. Such information is provided by multiple *cis*-acting splicing-regulatory elements (SREs), which act as either enhancers or silencers of splicing and can exist in either exons or introns. Historically these elements are named as exonic splicing enhancers (ESEs) or

To whom correspondence should be addressed: Zefeng Wang, Phone: 919-966-0131; Fax: 919-966-5640; zefeng@med.unc.edu, Yang Wang, Phone: 919-966-0131; Fax: 919-966-5640; ywang@med.unc.edu.

silencers (ESSs), and intronic splicing silencers (ISSs) or enhancers (ISEs). By recruiting *trans*-acting splicing factors, SREs either activate or inhibit the usage of adjacent splice sites, thus to specifically control alternative splicing [5, 6].

Various computational and experimental approaches have been developed in the past decade to identify the SREs and study their activities. The computational approaches generally use the sequence distribution biases or elevated conservation in different pre-mRNA regions to predictively identify sequence motifs that may control splicing [7–12]. The predicted SREs usually have to be experimentally confirmed by inserting the putative SREs into various splicing reporters. The experimental approaches usually use different screen approaches either *in vitro* or in cultured cells to identify short sequences that are bound by known splicing factors [13, 14] or directly affect splice site selection [15–19]. Historically, the exonic SREs are studied in greater detail compared to intronic SREs, and the splicing enhancers attract more attention than silencers. However, the four types of SREs may play equally important roles in splicing regulation. In addition, most SREs function in a context dependent manner (i.e. act as either splicing enhancer or silencers in different genes), indicating that a systematical identification and study of their functional overlap will provide critical information in understanding splicing regulation.

We have developed a cell-based method, called fluorescence-activated screen (FAS), which can identify short sequences from a random library that control splicing in both exons and introns. The key of this approach is to develop a fluorescence-based splicing reporter that can be spliced into either a non-functional mRNA (by default) or the GFP mRNA depending on the inserted SREs. This reporter is then inserted with a random decamer library and stably transfected into cultured cells. The green cells can be sorted out by flow cytometry to recover the inserted sequences that control splicing. The obtained SREs can then be used as "baits" to identify putative splicing factors that bind to SREs. This approach is very flexible and can be adopted to study splicing regulation in different cell types. In addition, it may be extended to study similar questions in other RNA processing pathways involving regulation by RNA *cis*-elements and *trans*-factors. In this report we describe details of how to design and carry out this approach, and give a brief discussion on the key questions in adopting this method to other scenario.

2. Description of methods

The basic outline of the SREs screen is illustrated in figure 1. The key of this method is to design and construct a splicing reporter that can produce two splicing isoforms, one of which being a functional GFP mRNA. The reporter was constructed so that the default splicing isoform is a non-functional mRNA, therefore when inserted with an SRE this reporter will produce functional GFP that can be used as a marker for cell sorting. The different designs of splicing reporters for ESSs, ESEs, ISSs and ISEs are shown in figure 1A, and the intact GFP gene is divided into two exons in all reporters. For ESS or ISS reporters, we insert a small constitutive exon between two GFP exons. The test exon is included during splicing unless there are ESS or ISS sequences inserted inside or near this exon, which will cause the skipping of the test exon to produce intact GFP mRNA (Fig. 1A) [15, 18]. For ESE reporter, the first GFP fragment is merged with a weak exon that is normally skipped during splicing. When inserted with an ESE, the weak exon will be included to produce intact GFP mRNA. The ISE reporter contains a weak intron between two GFP exons. This intron is normally retained during splicing, while the insertion of an ISE promotes splicing to produce functional GFP [19].

To screen functional SREs, a random sequence library will be inserted into the splicing reporter (Fig. 1A, shown in brown boxes). The resulting plasmid library will be transfected

into the cultured cells and the GFP positive clones will be selected by flow cytometry (Fig. 1B). To determine the randomness of the starting library, we have amplified and sequenced the insertion fragments from total DNA of a pool of stably transfected 293 cells [15, 18]. We found that all four bases were equally represented in each position, suggesting that the starting pool represented an essentially random pool of decamers [15, 18]. This strategy has been successfully used in screening of ESSs, ISSs and ISEs, and we describe the details of construction and screen of library in this paper.

2.1 Construction of splicing reporters for SREs screen

2.1.1 Generating constructs for ESSs screen—To assay for ESS activity, a GFP-based reporter containing three exons is constructed. The first exon of the reporter is amplified by PCR using pEGFP-C1 (Clonetech) as a template with primers P1 and P2 (Table 1) that contains a 5′ splice sites. Exon 3 of the screen reporter is amplified by PCR using pEGFP-C1 as a template with primers P3 and P4 (Table 1). Exon 1 is inserted between *Nhe*I and *Xho*I sites of pEGFP-C1 vector, which replaces the full-length EGFP open reading frame. The resulting vector is cut with *Sac*II and *BamH*I enzymes, and exon 3 is inserted between these two sites, generating a vector, pZW1 that contains a multicloning site between the two EGFP exons. The second exon of the Chinese hamster *DHFR* gene and its flanking introns are amplified by PCR using pB36 vector [20] as a template with primers P5 and P6. The PCR product is digested with *Sal*I and *Pst*I enzymes, and inserted into *Xho*I and *Pst*I digested pZW1 (between the two EGFP exons). The resulting vector is called pZW2, which is a three-exon containing minigene. Exon 1 and exon 3 of pZW2 can form an intact EGFP gene, whereas the eoxn 2 contains an *Xho*I/*Apa*I cloning site. Subsequently the three-exon fragment of pZW2 is cut with *Nhe*I and *BamH*I and ligated to the site-specific integration plasmid pCDNA5/FRT digested with *Nhe*I and *BamH*I. The resulting vector pZW4 is used for further stable transfection.

To insert random decamer sequences into the screen reporter pZW4, the foldback primer P7 (Table 1) is extended with Klenow enzyme and dNTPs for 20 min at room temperature (final concentration of P7 primer is 10μM), subsequently the polymerase is heat inactivated. The resulting product is digested overnight with *Xho*I and *Apa*I in a final concentration of 1μM, and ligated into pZW4 vector (Fig. 1B). No purification steps are needed since the reaction buffer is diluted at each step. The final amount of extended and digested primers in the ligation reaction is around 2.4 pmol in a 15 μl system.

2.1.2 Generating constructs for ISSs screen—The ISS screen minigene reporter is generated from the backbone vector, pZW1, which includes a multicloning site between two EGFP exons as described above. To make the reporter for the ISSs screen, two PCR reactions are used to amplify a constitutive exon – exon 6 of human *SIRT1* gene (Ensembl ID: ENSG00000096717) together with portions of its flanking introns. First, the upstream intron 5, exon 6 and 11bp of the downstream intron 6 of *SIRT1* gene are amplified by PCR with primers P8 and P9 (Table 1). A second PCR is used to amplify the fragment containing position 12 to position 266 of the downstream intron 6 of *SIRT1* gene using primers P10 and P11 (Table 1). The two PCR fragments are cloned into pZW1 to generate the resulting construct, pZW9, which is a three-exon splicing reporter with exon 1 and 3 forming an intact GFP gene and a multicloning site at 11 bp downstream of the 5′ss of the test exon 2 (SirT1 exon 6). The minigene reporter pZW9 is transferred into the site-specific integration plasmid pcDNA5/FRT by *Nhe*I/*BamH*I digestion and ligation, generating the vector pZW11 that can be used for further stable transfection.

To insert the random decamer library into pZW11, the foldback primer P7 (Table 1) is extended with Klenow enzyme and dNTPs for 20 min at room temperature as described

above. The resulting product is digested with *Xho*I and *Apa*I, and ligated into pZW11 vector as described above.

2.1.3 Generating construct for ISEs screen—To make the reporter for ISE screen, a retained intron from intron 4 of *C7orf26* (RefSeq: NM_024067) is fused with GFP exons in pZW1. Briefly a PCR reaction is carried out with primers P1 and P12 to amplify the first GFP exon of pZW1, resulting the first GFP exon linked with the 5′ end of C7orf26 intron that contain *Nhe*I/*Xho*I restriction sites. This PCR product is used to replace the first exon of the pZW1. Subsequently a two-step PCR reaction is conducted: In the first step the intron 4 of *C7orf26* (with primers P13 and 14) is amplified and the second GFP exon is amplified with primers P15 and P4, producing two fragments with a short overlap of the 3′ end of *C7orf26* intron. In the second step the two PCR products are linked together with primers P13 and P4, generating a fragment containing intron 4 of *C7orf26* and the GFP exon2. This DNA fragment is cloned into the pZW1 backbone with *Hind*III/*BamH*I, replacing the original GFP eoxn 2. The resulting minigene contains two GFP eoxns separated by a weak intron (*C7orf26* intron 4) harboring an *Xho*I/*Apa*I cloning site. To increase the sensitivity of ISE screen, additional mutations are introduced in the 3′ splice site to increase its strength [19]. The resulting reporter is transferred into pcDNA5/FRT vector between *Nhe*I/*BamH*I sites, creating the vector pZW15C for further screen. The insertion of a random decamer library is carried out as described earlier.

2.1.4 Constructs with heterologous exon or intron for validation of SRE activities—The FAS method uses a single exon/intron from a particular gene, it is possible that some SREs identified only function in this particular reporter since the activities of SREs are often context-dependent. Therefore it is necessary to confirm the generality of SREs in other exons or introns. Here we describe two commonly used splicing reporters with cloning sites near or inside a test exon to validate the SREs. However other splicing reporters can be used for this purpose, and we often test the SREs with a modular splicing reporter containing multiple cloning sites around an alternative exon [21].

To test ESSs in a heterologous exon, a minigene reporter pZW8 is constructed as following: The exon 6 of human *SIRT1* gene (Ensembl ID: ENSG00000096717) and portions of its flanking introns are amplified by two PCR reactions targeting to the 5′ and 3′ halves of the exon and corresponding flanking intron sequences by using primers P16/P17 and P18/P19. The resulting PCR products are digested with *Xho*I/*Hind*III and *Eco*RI/*Sac*II respectively, and inserted into the multicloning site of pZW1 between *Nhe*I and *Sac*II. A cloning site is generated inside the test exon at 21bp downstream of the 3′ splice site and the test sequences can be inserted into the *SIRT1* exon through *Hind*III/*Kpn*I digestion and ligation.

To examine the activities of ISSs and ISEs in a heterologous intronic context, a reporter, pZW2C, is modified from pZW2 that is used in the FAS-ESS screen. The pZW2 is digested with *Xho*I/*Apa*I and filled in with an oligonucleotide (obtained by annealing primers P20 and P21, Table 1) to destroy the exonic restriction sites. Then a new *Xho*I/*Apa*I restriction site is introduced at 18 nt downstream of the exon 2 by three consecutive PCR reactions: the first PCR uses primers P22 and P23 to amplify the 5′ half of the minigene in pZW2 whose exonic restriction sites is destroyed, the second PCR reaction uses primers P24 and P25 to amplify the 3′ half of the minigene in pZW2, and the third PCR reaction uses the products from PCR 1 and 2 as templates and uses primers P23 and P24 to amplify a fragment that contains intronic restriction sites. The resulting product of PCR3 is inserted into pZW2 digested with *Nhe*I/*Pst*I to obtain the reporter pZW2B. To increase the ISS and ISE detection sensitivity, the pZW2C is further generated by weakening the 3′ ss of exon 2 in pZW2B with site-directed mutagenesis so that exon 2 is included in about 50% of mRNA in the absence of ISS and ISE.

2.2 Generating the library for the FAS screens

To make a library of the random decamer sequences for the SREs screen, 2.5 μl out of 15 μl ligation product is mixed with 100 μl thawed ElectroMAX DH5α-E cells on ice and incubated for 10 min. Subsequently the cell/DNA mixture is moved into a chilled 0.1 cm cuvette. The cuvette should be gently tapped to ensure that the cell/DNA mixture contact all the way across the bottom of the chamber. The mixture is electroporated with BioRad GenePulser II electroporator using the condition 2 kV, 200 , 25 μF. After electroporation, 1.5 ml of SOC medium is immediately added to the cells in the cuvette, and the solution is transferred to a 15 ml snap-cap tube and shaken at 250 rpm (37°C) for 1 hour. The resulting cells are diluted into 5 ml with SOC medium. To determine the efficiency of the electroporation transformation, 200 μl out of 5 ml culture are plated into a 15 cm LB plate, the remainder cells are kept in 4°C for further plating. After 16 hours incubation at 37°C, colonies are counted to examine the transformation efficiency and determine plating conditions. The remainder *E. coli* cells are plated into 15 cm LB plates in a density of 5,000 to 10,000 colonies per plate. To achieve one to two fold coverage of the $4^{10} = \sim 10^6$ possible DNA decamers, 200 plates are needed and about eight electroporation transformations should be done to obtain sufficient transformed *E. coli* cells. The *E. coli* cells from the 200 plates are collected by gently wash with LB medium and the plasmid DNA libraries are purified from the transformed cells using QIAfilter Plasmid Mega kit (QIAGEN). Typically, around 1.5 mg of plasmid DNA of SRE library can be obtained.

2.3 Transfection and screen of the SREs library

For the stable transfection of the SRE library, Flp-In 293 cells are used. These cells include a single integrated Flp Recombination Target (FRT) site at a transcriptionally active genomic locus, which can facilitate integration of the library sequences into the specific site in the genome of mammalian cells.

In each transfection, about 2×10^7 Flp-In 293 cells are co-transfected with pOG44 and the pcDNA5/FRT vector containing SREs library (e.g. pZW4 for ESS screen) at a 9:1 ratio in a 15 cm tissue culture dish. The pOG44 encodes the recombinase Flp to mediate a homologous recombination event between the FRT sites, causing the reporter plasmid to be inserted by site-specific recombination into a single target site of the Flp-In 293 cells. To select stable transfectants, the transfected cells are expanded by a 1 to 4 dilution one day after transfection. The cells are grown for another day, and hygromycin B (Roche) is added into the medium to a final concentration of 100 μg/ml. The clones of stably transfected cells are usually visible after 10–13 days. These clones are trypsin digested and pooled for flow cytometry analysis (Fig. 1B). Typically, about one in 1,000–5,000 cells are GFP positive. The green cells are sorted out by fluorescence-activated cell sorting (FACS) using a Cytomation MoFlo high-speed sorter (or a similar sorter) into 96-well plates with one cell per well. About 10% of wells will have visible single colonies after 10–14 days. These colonies are transferred into duplicated 96-well plates and allowed to grow for 4–6 days into confluence. One plate is applied for flow cytometry analysis to reconfirm the GFP expression, and the other duplicated plate is used for total DNA purification. In rare situation, multiple cells will stick together and be sorted into the same well, and the wells with multiple colonies are checked by eyes and discarded before transferring to duplicated plates.

To purify DNAs from a 96-well plate, the cells are washed twice with PBS and incubated overnight in 50 μl of Lysis Buffer (10 mM Tris-Hcl, pH 7.5; 10 mM EDTA; 10 mM NaCl; 0.5% Sarcosyl; 1 mg/ml Proteinase K) at 60 °C in a humid atmosphere. The next morning, DNAs are precipitated by adding 100 μl of cold NaCl-EtOH mixture (add 750 μl of 5 M NaCl to 50 ml 100% EtOH pre-incubated at −20°C) to each well. The 96-well plate is kept

on the bench for 15–30 minutes until the visible DNA precipitation appearing on the bottom of the plate. The plate is inverted on paper towels covered by Kimwipes to discard the solution, and 150 μl of 70% EtOH is added to wash the wells. The DNA pellet will stick to the plate during the wash. The washing step is repeated for 2 to 3 times, and the plate is left to dry in the air (usually takes 20 min). The DNA is then dissolved by adding 50 μl of distilled water to each well. The total DNAs from GFP-positive clones are used as templates for PCR with primers P26 and P27 (Table 1), and the PCR products are cleaned up and sequenced to recover the SRE sequences. Around 200 transfections in 20 batches are needed to achieve about one-fold coverage of decamer library (1 million positive clones), which typically generate more than 100 SRE decamer sequences. Based on the sequence similarity, the SRE decamer sequences can be clustered into different groups by using CLUSTALW with default setting parameters (http://www.ebi.ac.uk/Tools/msa/clustalw2/).

2.4 Validation of identified SREs in a heterologous context and different cell type

To access the specificity of the FAS screen, the identified SREs are often validated by inserted back to splicing reporters that are transiently transfected to cells to assay for splicing changes. This validation is carried out both in the same splicing reporter and cell type used for the screen and in a different gene context and cell type. The SREs usually control splicing through recruiting specific splicing factors whose expression levels or activities may be different in various cell types. To validate the identified SREs in different cell types, we use HeLa cells since they are readily transfectable and easy to culture. For each transient transfection, 0.2 μg of splicing reporter containing selected SRE decamer is transfected into HeLa cells in a 24-well plate using lipofectamine 2000 (Invitrogen). After 24 hours of transfection, total RNAs from the transfected cells are isolated with TRIzol reagent (Invitrogen) by following standard protocols. The resulting RNAs are dissolved in 50 μl of RNase-free water and treated with 2 μl RNas-free DNase (Promega) at 37 °C for 45 minutes to remove DNA contamination. The reaction is stopped by adding 1 μl of DNase Stop solution and incubating at 65 °C for 15 minutes.

To synthesize the cDNA, 1μg of total RNA is mixed with 1 μl 10 mM dNTP mix, 1 μl primer P27 (10 μM) and RNase free water into a final volume of 13 μl, and the mixture is heated to 65 °C for 5 minutes and incubated at 4 °C for at least 1 minute. The resulting mixture is then incubated with 4 μl 5x First-Strand Buffer, 1 μl 0.1M DTT, 1 μl RNaseOUT and 1 μl of SuperScript III RT (200 units/μl) at 55 °C for 50 minutes and then inactivated at 70 °C for 15 minutes. One tenth of the cDNA product is used as the template for PCR amplification with primers P26 and P27 (25 cycles of amplification, with a trace amount of Cy5-dCTP adding to nonfluorescent dNTPs). RT-PCR products are separated on 10% PAGE gels and scanned with a Typhoon 9400 scanner (Amersham Biosciences). The amount of each splicing isoform is measured with ImageQuant 5.2.

On the other hand, the screens are conducted using a constant intron or exon with its flanking intron from a particular gene, which may lead to the identified SREs only functioning in a sequence context specific to this intron or exon. To directly address this, another splicing reporter with different exon/intron sequences (e.g. pZW9, pZW2C) is inserted with selected SREs. The resulting constructs are transfected into 293T or HeLa cells, and the change of splicing is detected with RT-PCR as described above.

2.5 Identification and validation of core SRE motifs

To identify the short core motifs that likely possess intrinsic SRE activities, statistical over-representation of hexamers in the set of recovered decamer sequences will be used as a criterion. For this analysis, each SRE decamer will be extended into a 14-mer by appending 2 nt of the vector sequence at each end (to allow for cases in which SRE activity derived

from sequence boundary of the vector). The 14-mer sequences will be broken into overlapping hexamers, and the numbers of each hexamer are counted. The repetitive hexamers occurring more than what are expected by chance (usually occurring at least three times) are clustered into several groups by sequence similarity using CLUSTALW, and the consensus motif of each group will be identified by aligning all hexamers of that group. These consensus motifs usually represent common patterns found in SRE decamers, and thus are treated as core motif of SREs.

To determine whether these significantly overrepresented short motifs have intrinsic SRE activity, one hexamer (i.e. exemplar) resembling the consensus of each group will be chosen for further experiments. Briefly, these exemplars will be inserted into different splicing reporters (e.g. pZW4, pZW11 or pZW15C), and their effects on splicing outcome can be analyzed by RT-PCR as described in section 2.4. The similar process can be carried out to identify over-representative pentamers or heptamers, and the consensus motifs can be identified and validated for each SRE group.

2.6 Identification of putative splicing factors for each SRE group

SREs usually control splicing through specifically recruiting *trans*-acting splicing factors that activate or inhibit splicing. The identification of novel SREs enables the identification of cognate splicing factors that specifically bind to different groups of SREs in an unbiased manner. To this end, an RNA affinity purification approach is developed based on a previously described protocol [22]. Briefly a short RNA fragment containing three copies of SRE exemplar is incubated with whole cell extract, and RNAprotein complexes are isolated with streptavidin beads and proteins specifically bound by SREs are identified with mass spectrometry (Fig. 2A).

The short (~21-nt) RNA fragments containing three copies of SRE exemplars are synthesized with 5′ biotin followed by two 18-carbon spacers (Ambion/Invitrogen) as RNA "baits". For each RNA sample, approximately 2.5×10^8 HeLa cells (NCCC, Minneapolis) are harvested at ~95% confluence and resuspended with 2.5 ml ice-cold resuspension buffer (50 mM Tris-HCl, pH 8.0, 150 mM NaCl). Cells are mixed with 2.5 ml 2× lysis buffer (50mM Tris-HCl, pH 8.0, 150mM NaCl, 15 mM NaN3, 1% (v/v) NP-40, 2 mM DTT, 2 mM PMSF, 2× protease inhibitor mix), lysed for 5 min and then centrifuged at 12,000g for 20 min at 4 °C. Subsequently 0.75 nmol biotinylated RNAs are added to the supernatants and incubated for 2 h at 4 °C. Next, 50 μl streptavidin-agarose beads (Sigma) are added into the mixture and incubated for 2 h at 4 °C with slow inverting rotation. The beads are washed 3 times with 4 ml lysis buffer (50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 15 mM NaN3, 0.5% NP-40, 1 mM DTT, 1 mM PMSF, 1× protease inhibitor mix), resuspended in 40 μl final volume and mixed with 10 μl of 5× SDS loading buffer. The proteins are then separated with a 10% SDS-PAGE gel and stained with Coomassie blue. The gels are kept in 3% acetic acid for further mass spectrometry analysis. The bands of interest that contain candidate protein *trans*-factors are cut and analyzed by ESI-MS/MS on a Q-Tof (Micromass) mass spectrometer.

2.6 Validation of splicing factors for SREs

The putative splicing factors are identified according to their specific bindings to the SRE groups, it remains to be tested whether the resulting factors are indeed responsible for the SRE activity. To determine whether these identified putative protein factors regulate splicing through directly binding to the specific SRE group, overexpression and knockdown approaches are used. Figure 2B shows a diagram of such experiments using ISS as an example [18]. A splicing reporter containing a cognate ISS or a control element is co-transfected into cells with the expression vector of the putative splicing factor. If the *trans*-

acting factor indeed controls splicing by binding to ISS, it will increase the level of exon skipping in the reporter inserted with cognate ISS but not in control reporter (Fig. 2B left panel). Conversely, the knockdown of *trans*-factor by RNAi will decrease exon skipping in reporters inserted with cognate ISS but not in control reporter (Fig. 2B right panel).

For knockdown of protein factors, the siRNAs targeting specific splicing factors are purchased from Dharmacon (On-target SMARTpool, with scrambled dsRNA controls). For each transfection, 6 μl of siRNA pool is mixed with 194 μl of RPMI-1640 medium. 2 μl of lipofectamine 2000 is mixed with 63 μl of RPMI-1640 medium. Subsequently these two mixtures are incubated together at room temperature for 20 minutes. At the same time, 125k HEK-293T cells are plated into each well of 24-well plates. The resulting mixture is added to the cells. The second day, the medium will be replaced with 1 ml of fresh medium. After another 24 hours of siRNA transfection, the cells are transfected with 0.2 μg splicing reporters containing cognate SRE exemplars. One day later, the transfected cells are collected for further analyses using RT-PCR and western blot (to measure RNA splicing and protein decrease). For the over-expression experiments, the full-length cDNAs of different *trans*-factors are cloned into pcDNA3 expression vector with a FLAG tag. In each transfection, 0.8 μg of protein factor expression vectors are co-transfected with 0.2 μg splicing reporters containing cognate SRE exemplars. After three days, the transfected cells are harvested to purify total RNAs and proteins, followed by RT-PCR and western blot analyses that determine RNA splicing and protein expression.

3. Results and Data Analysis

3.1 Typical results of a FAS screen

Here we use the FAS-ISE screen as an example to illustrate the typical results from this approach. The detail results and dataset from the FAS-ISE screen are available in our previous paper [19]. A random decamer sequences library was inserted into the screen minigene reporter pZW15C between *Xho*I/*Apa*I sites, and the ligation product was transformed into enough DH-5α *Escherichia coli* cells to generate the ISE decamers sequence library. The resulting library was transfected into FlpIn-HEK-293 cells, which contain a single site-specific recombination site for stable integration.

Over a span of eight months, we have conducted 208 transfections in 15 batches to achieve a one-fold coverage of the entire decamcer sequence library. The clones from 96-well plates were tested for the positive GFP expression using flow cytometry (Fig. 3A), and only the DNAs from the positive clones were used for subsequent PCR and sequencing. In total 117 FAS-ISE decamers have been identified, 109 of which are unique [19]. Based on the sequence similarity, the identified FAS-ISE decamers can be clustered into different groups by CLUSTALW. To validate the screen, we randomly selected six identified decamers and inserted them into a new splicing reporter, pZW2C that contains a cassette exon with its flanking introns, and test if they can promote splicing in a heterologous context. We found that all the tested ISE decamers significantly promoted cassette exon inclusion as judged by increase of PSI (percent-spliced-in) value as compared to the neutral sequence ($p<0.05$) (Fig. 3B). Such validation was conducted in both 293 and HeLa cells (Fig. 3B), and we were able to confirm the ISE activity for all tested sequences in both cell types, indicating that the false positive rate of FAS-ISE screen is low and the identified ISEs have general enhancer activities across different cell types.

We further extracted the core motifs of the identified ISE decamers using statistical overrepresentation approach. According to the sequence similarity, these ISE hexamers were clustered into six groups, and the hexamers in each group were multiply aligned to identify candidate motifs [19]. From each ISE group, we selected a hexamer exemplar that resembles

the consensus motif, and tested their intrinsic ISE activities in pZW15C reporter (Fig. 3C). The resulting reporters were transfected into HEK-293T cells and the splicing change was examined by RT-PCR. We found that all exemplars promoted intron splicing (Fig. 3C). In addition, the ISE activities of these exemplars were also validated in another splicing reporter, pZW2C, which contains a heterologous intron [19].

The splicing factors that specifically bind to different ISE groups are identified in an unbiased manner using an RNA affinity purification approach as described earlier. We have identified 17 known or predicted RNA-binding proteins from 30 bands. Several well-known splicing factors (e.g., hnRNP H1 and hnRNP F) and some novel splicing factors were included in the list of identified proteins. To validate if the identified splicing factors were indeed responsible for the ISE activities, we used hnRNP H1 as an example to test its function using splicing reporters containing cognate ISE. The results showed over-expression of hnRNP H1 promoted the inclusion of the cassette exon. Consistently, the RNAi of hnRNP H1 inhibited the inclusion of the cassette exon (Fig. 3D).

3.2 Data analysis by computational approaches

The newly identified ISE sequences provide a large dataset that can be analyzed with various computational approaches. Here we described two straightforward examples. The first analysis is to examine the distribution biases of the newly identified ISEs in different regions of pre-mRNA, which often provide clues of the *in vivo* function of these elements. The positional distributions of FAS-ISEs were examined in human exons and associated introns. Briefly, the FAS-ISE hexamers in each group were counted near the constitutive exons (CEs), skipped exons (SEs), alternative 3′SS exons (A3Es) and alternative 5′SS exons (A5Es). The data showed that all FAS-ISE groups are enriched in introns *versus* exons and many FAS-ISE groups peak at upstream of the 3′SS or downstream of the 5′SS, which is similar to the distribution pattern of ESSs. Such distribution suggests that ISE may suppress pseudo-exons and help define alternative splice sites, which were further confirmed by experiments [19].

We often measure if the identified elements are conserved across different species in certain region of pre-mRNA, which is another strong indication of their function. The relative conservation patterns of FAS-ISEs in different pre-mRNA regions were analyzed using a previously developed scoring system [23]. Briefly, the exons conserved in human and mouse genomes were obtained from the UCSC Genome Browser, and classified into SEs, CEs and A5Es or A3Es. We obtained 281 human/mouse orthologous A5Es, 301 A3Es, 1649 SEs and 26340 CEs. The degrees of conservation were computed as the *P* values for the Conserved Occurrence Rate (COR) statistic. A smaller *P* value indicates more conserved hexamers in particular regions. Strikingly, the ISE set were more conserved in exons rather than introns and were remarkably conserved in exonic extension regions of A3Es or A5Es (Detail results can be found in ref [19] Fig. 3C).

4. Concluding remarks

Alternative splicing is a key mechanism to expand coding diversity in higher eukaryotes. This process is mainly regulated by SREs and cognate splicing factors. Here we describe a method to unbiasedly identify SREs from a random decamer library and their cognate putative splicing factors. This screen system has a number of technical features. First, GFP is selected as the reporter of different splicing outcomes, thus there is no growth advantage expected between cells with the two splicing isoforms. Second, the exons of the reporter don’t share any homology to each other, minimizing the likelihood of DNA recombination that would complicate the screen. Third, the FlpIn system and a host cell line containing a single FRT integration site are used to generate a library of stably transfected cells, ensuring

that only a single minigene is inserted at the same genomic location of different clones. In addition, insertion into the same locus also ensures consistent expression of the minigene reporter. Forth, one selection step followed by a single round of PCR is used during the screen, avoiding the sequence biases that may result from multiple rounds of selection and PCR in SELEX. Finally, the green cells are recovered with FACS sorting, giving us a high sensitivity (in practice, one positive in 10 thousand cells can be reliably recovered).

Several factors could affect the sensitivity and specificity of this cell-based screen. The SREs identified in any cell-based screen will presumably reflect the set of *trans*-regulators expressed in the cell type used, thus the tissue specific SREs will probably be missed (i.e., false negatives). General splicing factors are ubiquitously expressed and 293 cells are routinely used for splicing assays, use of 293 cells might allow the identification of a large fraction of SREs that have activities in a broad range cell types. Indeed, the ESSs, ISEs and ISSs identified by this method usually functioned in another cell type [15, 18, 19]. However, it is likely that certain SREs were not detected because the corresponding *trans*-factors are not expressed in 293 cells (or expressed at very low levels). Use of random decamer library might also result in missing of long SRE motifs (> 10 bases) or SREs that function only when present in multiple copies. Therefore it will be interesting in the future to conduct the screen using a tissue-specific cell type or a random library with different lengths.

Generally we used a splicing reporter that is constitutively spliced as the non-fluorescence isoform in absence of candidate SREs, thus this screen was designed to detect those SREs whose activity is strong enough to change splicing of constitutive exon. We also used an empirical threshold of FACS sorting to capture the SREs that turn dark cells into green cells. This setup ensured a low positive rate (i.e. high specificity) but might miss some weak SREs. In the future, it may worth adjusting the FACS sorting thresholds to collect cells in different fractions. This will help to recover cells with weak fluorescence (i.e. smaller change of PSI). The tradeoff of using a range of thresholds is that more clones will have to be recovered and sequenced, and more false positive SREs will be identified.

A different strategy based on statistical analysis was commonly used to identify the SREs. These methods discovered putative SRE motifs either based on sequence distribution biases or elevated conservation [7–12], or based on the enriched sequence motifs in the endogenous protein targets [24] or tissue specific exons [25]. The method described here presents a complementary approach that can identity general SREs with strong activities. As described earlier, the tissue specific SREs will not be recovered unless we carry out this screen in tissue-specific cell types. In comparison, analysis of Nova binding site or neuronal exons will identify neuronal specific SREs (e.g. binding site of Nova or nPTB) [24, 25].

We also applied modified RNA affinity purification approach to unbiasedly identify the putative splicing factors that bind to cognate SREs. It has been previously shown that many proteins can bind to RNA with relatively low sequence specificities, leading to high noise in affinity purifications of protein factors associated with SREs. We have modified this method with several improvements to minimize such noise. We use a short RNA (21-23 nt) with three copies of SRE exemplars to increase the binding specificity as the use of a short bait will “force” all proteins to compete for the same region. We also use a long spacer to separate biotin from RNA, making the RNA more accessible by the proteins. Finally the RNA-protein incubation is carried out in a large volume with a lower protein concentration for an extended time (> 4 hours), which helps to reach binding equilibrium.

Acknowledgments

This work is supported by NIH grant R01-CA158283 and the Jefferson Pilot award to Z.W.

Reference

1. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Nature. 2008; 456:470–476. [PubMed: 18978772]
2. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Nat Genet. 2008; 40:1413–1415. [PubMed: 18978789]
3. Castle JC, Zhang C, Shah JK, Kulkarni AV, Kalsotra A, Cooper TA, Johnson JM. Nat Genet. 2008; 40:1416–1425. [PubMed: 18978788]
4. Cooper TA, Wan L, Dreyfuss G. Cell. 2009; 136:777–793. [PubMed: 19239895]
5. Wang Z, Burge CB. RNA. 2008; 14:802–813. [PubMed: 18369186]
6. Matlin AJ, Clark F, Smith CW. Nat Rev Mol Cell Biol. 2005; 6:386–398. [PubMed: 15956978]
7. Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Science. 2002; 297:1007–1013. [PubMed: 12114529]
8. Zhang XH, Chasin LA. Genes Dev. 2004; 18:1241–1250. [PubMed: 15145827]
9. Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G. Mol Cell. 2006; 22:769–781. [PubMed: 16793546]
10. Yeo GW, Nostrand EL, Liang TY. PLoS Genet. 2007; 3:e85. [PubMed: 17530930]
11. Voelker RB, Berglund JA. Genome Res. 2007; 17:1023–1033. [PubMed: 17525134]
12. Aznarez I, Barash Y, Shai O, He D, Zielenski J, Tsui LC, Parkinson J, Frey BJ, Rommens JM, Blencowe BJ. Genome Res. 2008; 18:1247–1258. [PubMed: 18456862]
13. Liu HX, Zhang M, Krainer AR. Genes Dev. 1998; 12:1998–2012. [PubMed: 9649504]
14. Tian H, Kole R. J Biol Chem. 2001; 276:33833–33839. [PubMed: 11454855]
15. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. Cell. 2004; 119:831–845. [PubMed: 15607979]
16. Yu Y, Maroney PA, Denker JA, Zhang XH, Dybkov O, Luhrmann R, Jankowsky E, Chasin LA, Nilsen TW. Cell. 2008; 135:1224–1236. [PubMed: 19109894]
17. Culler SJ, Hoff KG, Voelker RB, Berglund JA, Smolke CD. Nucleic Acids Res. 2010
18. Wang Y, Xiao X, Zhang J, Choudhury R, Robertson A, Li K, Ma M, Burge CB, Wang Z. Nat Struct Mol Biol. 2013; 20:36–45. [PubMed: 23241926]
19. Wang Y, Ma M, Xiao X, Wang Z. Nat Struct Mol Biol. 2012
20. Fairbrother WG, Chasin LA. Mol Cell Biol. 2000; 20:6816–6825. [PubMed: 10958678]
21. Xiao X, Wang Z, Jang M, Burge CB. Proc Natl Acad Sci U S A. 2007; 104:18583–18588. [PubMed: 17998536]
22. Dominski Z, Yang XC, Kaygun H, Dadlez M, Marzluff WF. Mol Cell. 2003; 12:295–305. [PubMed: 14536070]
23. Wang Z, Xiao X, Van Nostrand E, Burge CB. Mol Cell. 2006; 23:61–70. [PubMed: 16797197]
24. Zhang C, Frias MA, Mele A, Ruggiu M, Eom T, Marney CB, Wang H, Licatalosi DD, Fak JJ, Darnell RB. Science. 2010; 329:439–443. [PubMed: 20558669]
25. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ. Nature. 2010; 465:53–59. [PubMed: 20445623]

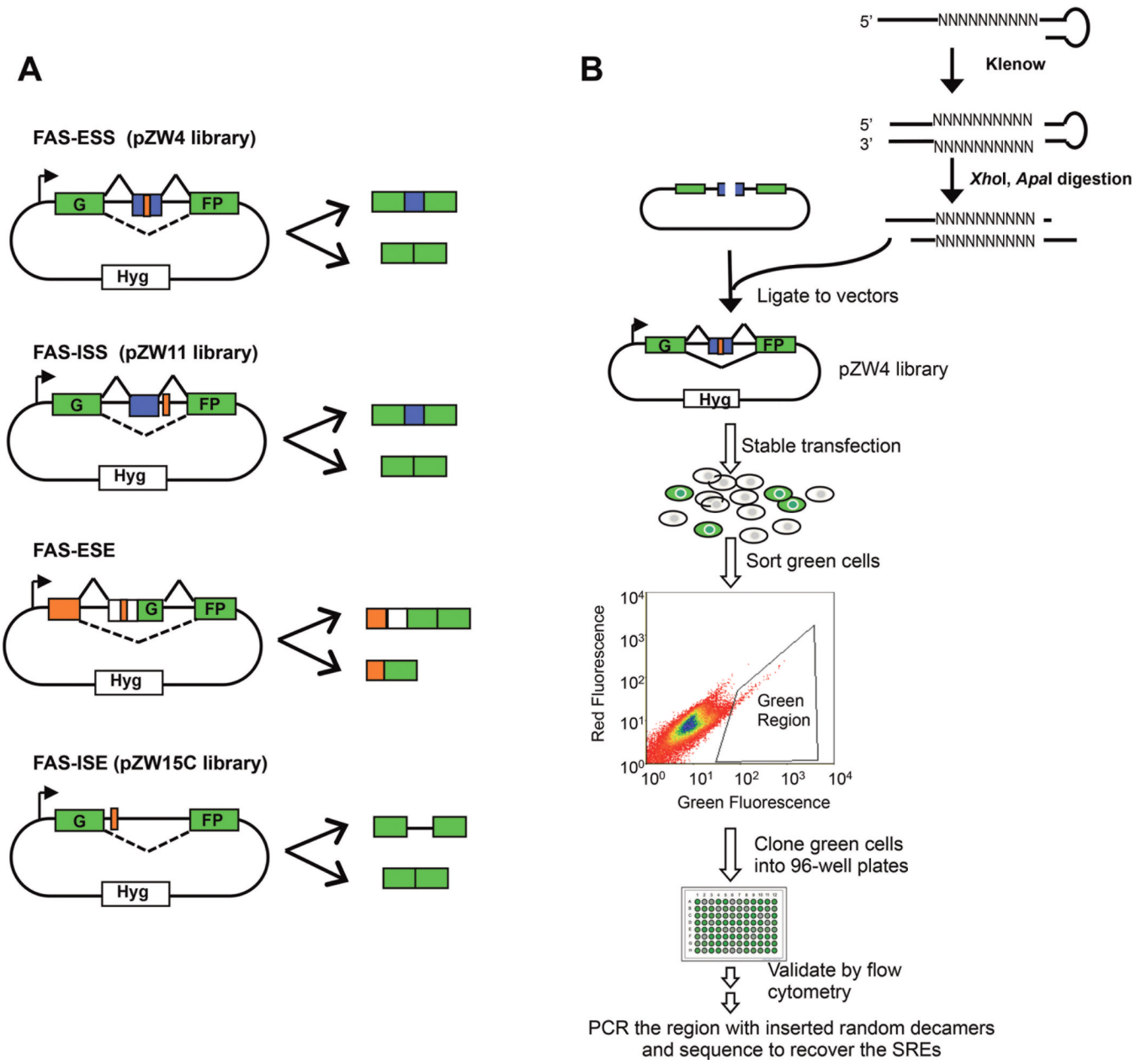

Figure 1. Diagrammatic representation of FAS screen strategy

(A) Design of splicing reporters for screen of ESS, ISS, ESE and ISE. (B) The foldback primer is synthesized with a random decamer nucleotide sequence, extended with Klenow enzyme, digested with XhoI/ApaI enzymes and ligated to screening vectors (e.g. pZW4) to make the SREs library. The resulting library are stably transfected into Flp-In 293 cells. After selection, the green cells, which are generated upon SREs insertion, are sorted into 96-well plates by fluorescence-activated cell sorting (FACS) using a Cytomation MoFlo high-speed sorter. DNAs are purified from these green cells, and the regions with inserted random decamers are amplified. The SRE sequences are recovered by sequencing.

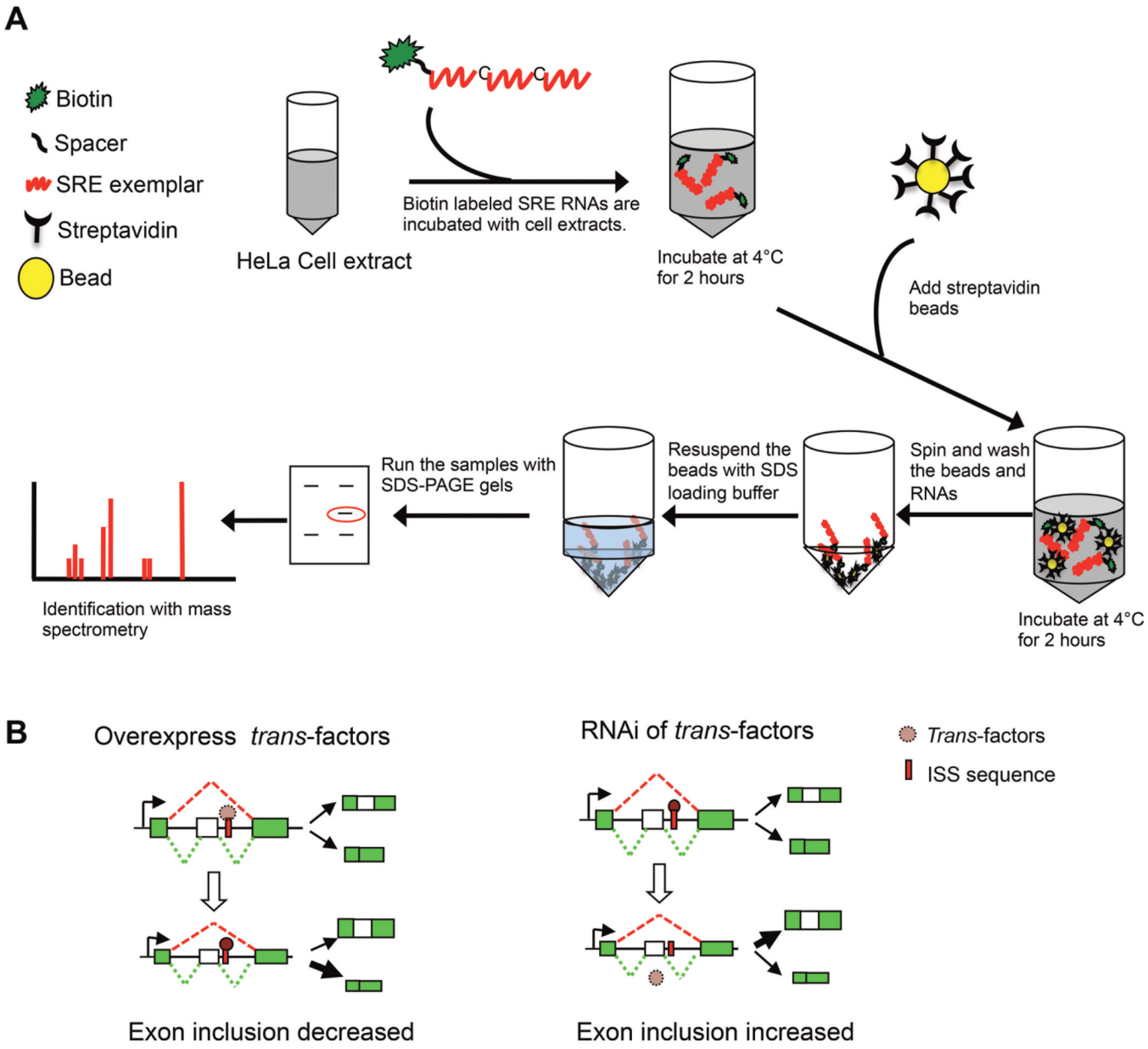

Figure 2. Identification and validation of cognate splicing factors

(A) Schematic diagram of the RNA-affinity purification method. Biotinylated RNAs containing three copies of SRE exemplars are incubated with HeLa cell extract. The RNA-protein complexes are subsequently purified with streptavidin beads, and separated on SDSPAGE gels. The specific bands found in SRE groups but not in controls are cut from the gels for identification with mass spectrometry. (B) Reconfirm the activity of splicing factors with over-expression and RNAi experiments. The factors binding to ISSs are used as examples.

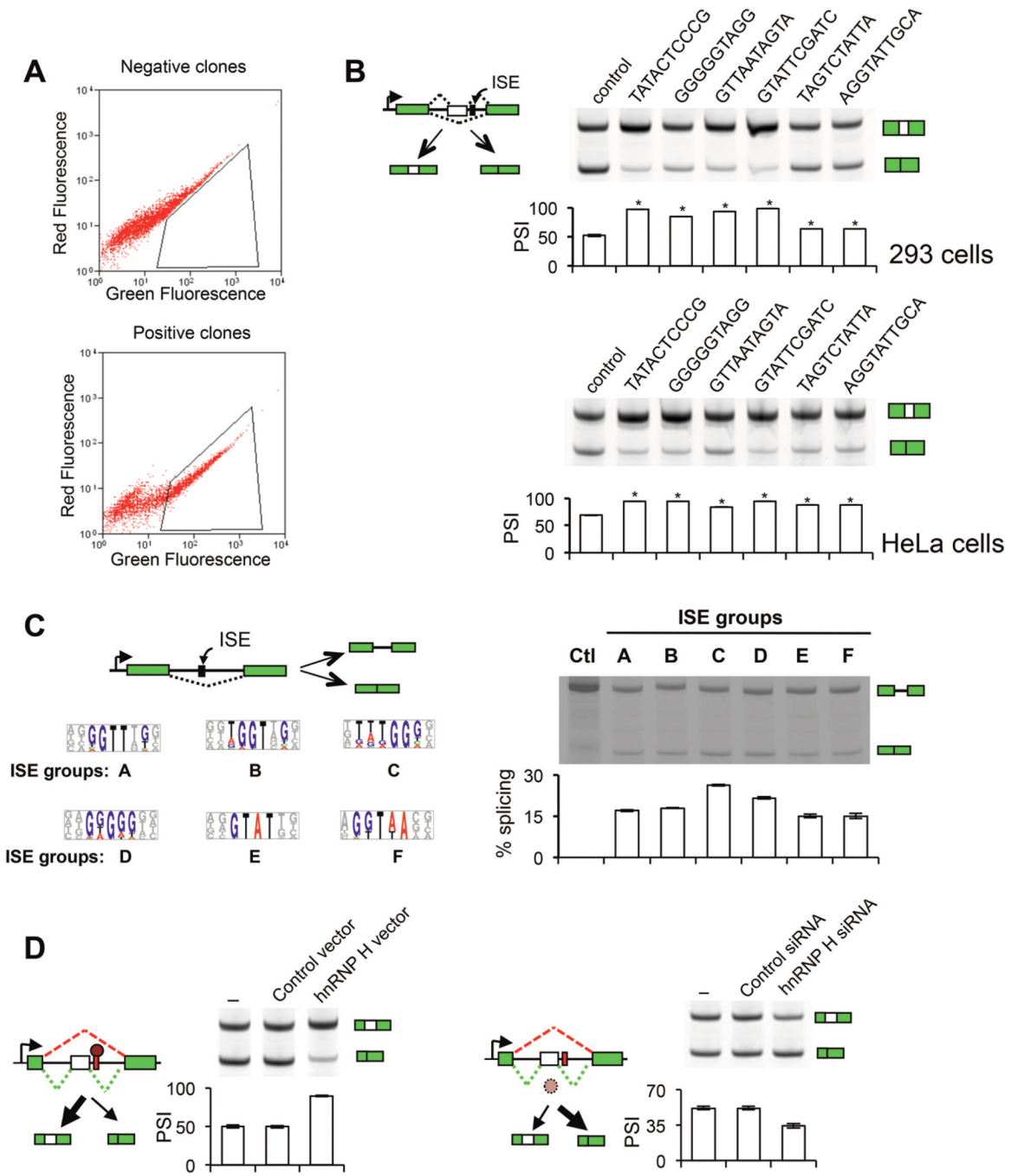

Figure 3. Typical results obtained by FAS-ISE screen

(A) Flow cytometry results from positive and negative clones from 96-well plates. This step was used to reconfirm the sorting result, and total DNAs from positive clones were used for PCR and sequencing. (B) Validation of recovered ISE decamers in a heterologous context and different cell types. All sequences can significantly increase the exon inclusion when inserted at downstream of alternative exon. The means and S.T.D. of PSI were plotted. *P* values were calculated with the paired T test (n=3). "*" indicates significant difference. (C) The consensus motifs of six ISE motifs identified by FAS-ISE screen. We inserted the examplar sequences of each group into the splicing reporter originally used in screen to

confirm the intrinsic ISE activity of each group. (D) Over-expression and RNAi of hnRNP H1 affected splicing of a cassette exon containing the cognate ISE. Left, the splicing reporter was co-transfected with hnRNP H1 and control vector. Right, the splicing reporter was transfected 48 hours after RNAi of hnRNP H1 or control. The test exon contain a putative binding site of hnRNP H1 at the downstream intron (indicated by red box).

Table 1

Primer sequences used in generating the vectors

Primers	Sequences	Note
P1	CACGCTAGCGCTACCGGTCGCCAC	Forward primer for GFP exon 1
P2	CACCTCGAGACTTACCTGGACGTAGCCTTCGG	Reverse primer for GFP exon1
P3	CACCCGCGGTCTCTTTCTTCCAGGAGCGCACCATCTTC	Forward primer for GFP exon 2
P4	CACGGATCCTTACTTGTACAGCTCGTCCATG	Reverse primer for GFP exon 2
P5	CACAGTCGACGCAGCCCTTGCCCATGCCC	Forward primer for DHFR exon 2
P6	AAGGGCTGCAGAAAGGCTGGAAC	Reverse primer for DHFR exon 2
P7	CACCTCGAG(N10)GGGCCCACACGTTTTTTTCGTGTGGGCCC	Foldback primer for generating random decamer sequences
P8	CACGTCGACCTGCAGGATTTTAGCCCTG	Forward primer for intron 5, exon 6 and 11bp of downstream intron 6 of SIRT1
P9	CACAAGCTTCTCGAGCAACAAATTACCTGATTAAAAAT	Reverse primer for intron 5, exon 6 and 11bp of downstream intron 6 of SIRT1
P10	CACGAATTCATGTGGGCCCATATTTTAGGAATTGTTC	Forward primer for position 12 to position 266 of the downstream intron 6 of SIRT
P11	CACCCGCGGACAACTTGCTTATGATCCTGAC	Reverse primer for position 12 to position 266 of the downstream intron 6 of SIRT
P12	GTGCTCGAGTAGTGAGAAGCAACCTGGACGTAGCCTTCGG	Reverse primer linking GFP exon1 with C7orf26 intron 4
P13	CACAAGCTTCGGGCCCTAAATACTCCGATTGCGGC	Forward primer for C7orf26 intron 4
P14	GAAGATGGTGCGCTCCTGAGGTGGAGTTTTG	Reverse primer linking GFP exon2 with C7orf26 intron 4
P15	CAAAACTCCACCTCAGGAGCGCACCATCTTC	Forward primer linking C7orf26 intron 4 with GFP exon2
P16	CACCTCGAGCTGCAGGATTTTAGCCCTG	Forward primer for 5′ half of exon 6 of SIRT1
P17	CACAAGCTTCAAGATGCTGTTGCAAAGG	Reverse primer for 5′ half of exon 6 of SIRT1
P18	CACGAATTCGGTACCTTGTAAATACAAAGTTGAC	Forward primer for 3′ half of exon 6 and corresponding intron of SIRT1
P19	CACCCGCGGACAACTTGCTTATGATCCTGAC	Reverse primer for 3′ half of exon 6 and corresponding intron of SIRT1
P20	TCGACTACGTACATGCGGCC	Forward primer for destroying the exonic enzyme sites of pZW2
P21	GCATGTACGTAG	Reverse primer for destroying the exonic enzyme sites of pZW2
P22	TAGGGCCCCAGTCTCGAGACCCCAAATTACCTTC	Forward primer for amplifying the 5′ half of the minigene in pZW2
P23	CACGCTAGCGCTACCGGTCGCCAC	Reverse primer for amplifying the 5′ half of the minigene in pZW2
P24	TCTCGAGACTGGGGCCCTAAGATGAGGATTCTAGGGG	Forward primer for amplifying the 3′ half of the minigene in pZW2
P25	AAGGGCTGCAGAAAGGCTGGAAC	Reverse primer for amplifying the 3′ half of the minigene in pZW2
P26	AGTGCTTCAGCCGCTACCC	Forward primer for amplifying GFP positive clones
P27	GTTGTACTCCAGCTTGTGCC	Reverse primer for amplifying GFP positive clones