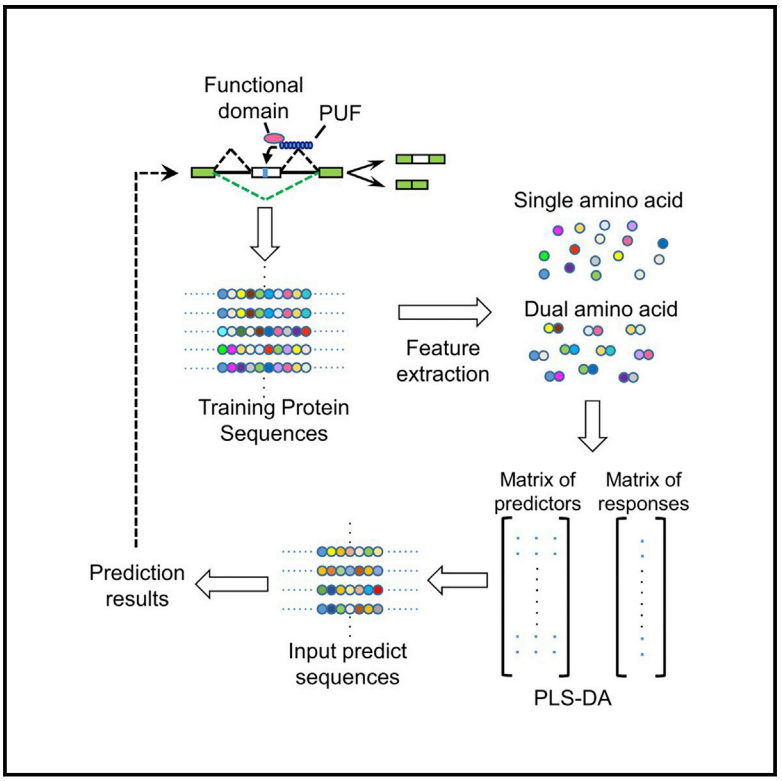
Article

Cell Systems

Modeling and Predicting the Activities of *Trans*-Acting Splicing Factors with Machine Learning

Graphical Abstract



Highlights

- Most low-complexity domains in RBPs can regulate splicing when recruited to pre-mRNAs
- Splicing regulatory activities of RBPs are mainly determined by sequence composition
- Machine learning approach was developed to predict splicing regulatory activity of RBPs
- The predictive model facilitates the design of artificial factors to manipulate splicing

Authors

Miaowei Mao, Yue Hu, Yun Yang, ..., Yi Yang, Xiaoling Li, Zefeng Wang

Correspondence

wangzefeng@picb.an.cn

In Brief

Alternative splicing is mainly regulated by various *trans*-acting splicing factors that specifically bind *cis*-elements. A systematic survey was conducted to study splicing regulatory activities of many RBPs, providing a training set for a machine learning approach to predict splicing regulatory activities of endogenous RBPs and synthetic peptides. This study expanded the repertoire of potential splicing factors and revealed a direct link between the sequence composition and RBP activity.

Mao et al., 2018, Cell Systems *7*, 510–520
November 28, 2018
https://doi.org/10.1016/j.cels.2018.09.002

CellPress

CellPress

Cell Systems
Article

Modeling and Predicting the Activities of *Trans*-Acting Splicing Factors with Machine Learning

Miaowei Mao,[1,2,3] Yue Hu,[1] Yun Yang,[1] Yajie Qian,[2] Huanhuan Wei,[1] Wei Fan,[3] Yi Yang,[2] Xiaoling Li,[3] and Zefeng Wang[1,4,*]
[1]CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, CAS Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai 200031, China
[2]Synthetic Biology and Biotechnology Laboratory, State Key Laboratory of Bioreactor Engineering, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China
[3]Signal Transduction Laboratory, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA
[4]Lead Contact
*Correspondence: wangzefeng@picb.an.cn
https://doi.org/10.1016/j.cels.2018.09.002

SUMMARY

Alternative splicing (AS) is generally regulated by *trans*-splicing factors that specifically bind to *cis*-elements in pre-mRNAs. The human genome encodes ~1,500 RNA binding proteins (RBPs) that potentially regulate AS, yet their functions remain largely unknown. To explore their potential activities, we fused the putative functional domains of RBPs to a sequence-specific RNA-binding domain and systemically analyzed how these engineered factors affect splicing. We discovered that ~80% of low-complexity domains in endogenous RBPs displayed distinct context-dependent activities in regulating splicing, indicating that AS is under more extensive regulation than previously expected. We developed a machine learning approach to classify and predict the activities of RBPs based on their sequence compositions and further validated this model using endogenous RBPs and synthetic polypeptides. These results represent a systematic inspection, modeling, prediction, and validation of how RBP sequences affect their activities in controlling splicing, paving the way for *de novo* engineering of artificial splicing factors.

INTRODUCTION

Genome-wide studies estimated that at least 90% of human genes undergo some degree of alternative splicing (AS), through which a single pre-mRNA can produce multiple isoforms using different combinations of exons (Baralle and Giudice, 2017; Kim et al., 2014; Pan et al., 2008; Wang et al., 2008). The process of AS is tightly regulated in different tissues and developmental stages (Kornblihtt et al., 2013), and disruption of splicing regulation is a common cause of human diseases such as cancer (Chabot and Shkreta, 2016; Cooper et al., 2009; David and Manley, 2010; Song et al., 2018). AS is generally controlled by multiple regulatory *cis*-elements in pre-mRNAs, which specifically recruit *trans*-acting splicing factors to promote or suppress the use of adjacent splice sites (Matera and Wang, 2014; Wang and Burge, 2008). The activities of splicing factors are often context dependent in that the same splicing factor may promote or inhibit splicing reaction by binding to the same *cis*-element at different pre-mRNA locations. For example, the serine/arginine-rich proteins (SR proteins) can specifically bind to their cognate targets within an alternative exon to promote exon inclusion; however, they inhibit exon inclusion when binding to the downstream intron of the alternative exon (Wang et al., 2013).

Typical splicing factors have a modular domain configuration, containing one or several RNA binding domains (RBDs) to specifically recognize cognate *cis*-elements in pre-mRNA targets and the functional domain(s) to affect splicing (Fu and Ares, 2014; Matera and Wang, 2014). For example, SR proteins recognize their target through one or several RNA recognition motifs (RRMs) and affect splicing with RS domains (Graveley, 2000). Therefore, direct tethering of the RS domain can recapitulate the activity of SR proteins (Philipps et al., 2003; Wang et al., 2009). We have previously constructed engineering splicing factors (ESFs) by fusing functional domains of known splicing factors to an RBD with programmable specificity (PUF domain from human PUM1). The resulting ESFs have been successfully used in manipulating AS of endogenous genes in cultured cells and model animals (Li et al., 2016; Wang et al., 2009). Additionally, this strategy also provides a convenient and standard system to assess the function of splicing regulatory domains.

Based on proteomic analyses, the human genome encodes ~1,500 RNA binding proteins (RBPs) with different types of RBDs, including RRMs (~240 RBPs) and HNRNPK-homology domains (KHs, ~60 RBPs) (Gerstberger et al., 2014). A large fraction of these proteins are located in the nucleus and probably play regulatory roles in AS. However, except for several examples, the functional domains and regulatory activities of these potential splicing factors are unclear. To systematically study the activities of potential splicing factors, we engineered a series of ESFs with 63 putative functional domains from 51 human RBPs. Remarkably, we found that many domains in endogenous RBPs have activities to modulate splicing, indicating that AS may be more widely regulated by a variety of RBPs in human cells

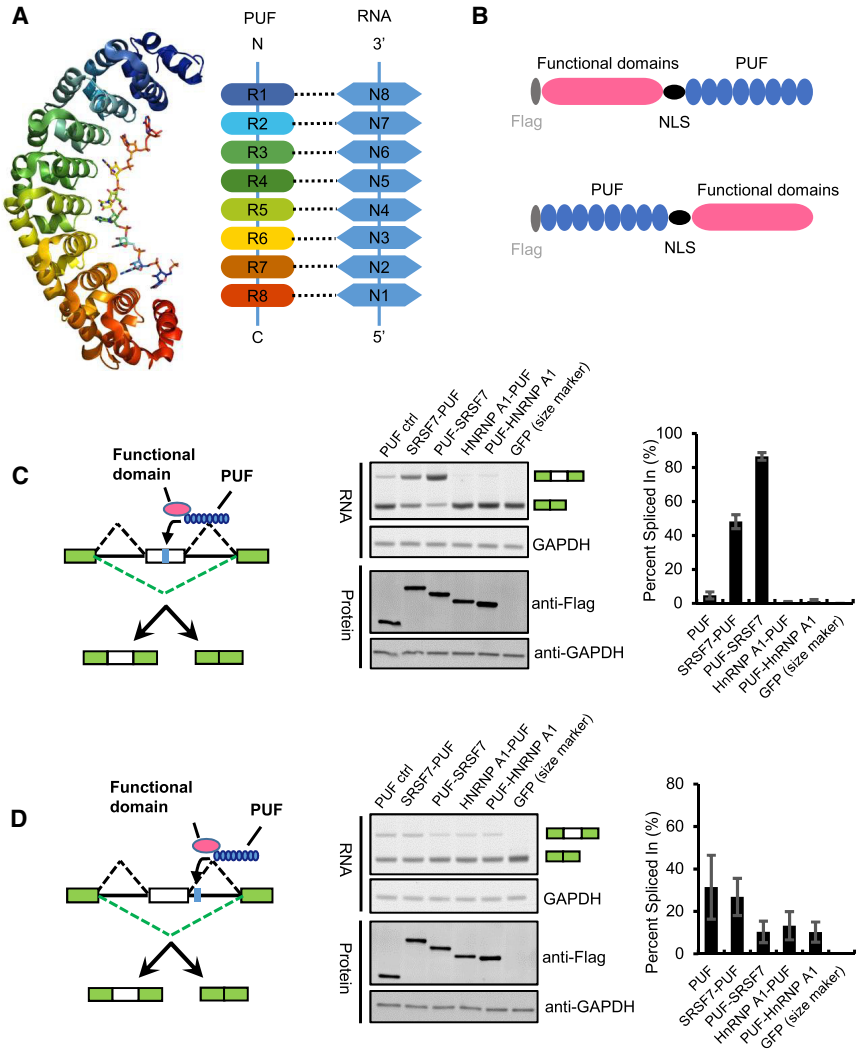


Figure 1. Design and Validation of ESFs That Regulate Alternative Splicing

(A) Crystal structure and the diagram of the PUF domain from human PUM1 (left) bound to its RNA target (PDB: 2YJY). The PUF domain recognizes 8-nt nucleotides in a modular fashion with each repeat recognizing a single base.

(B) Construct of ESFs. PUF domain (indigo) locates at C-terminal and functional domain (magenta) locates at N-terminal (upper); PUF domain locates at N-terminal and functional domain locates at C-terminal (lower); FLAG, epitope tag (dark gray); NLS, nuclear localization signal (black).

(C) Regulation of splicing by ESFs that bind to their cognate site inside the cassette exon of a splicing reporter. The expression vectors of ESFs containing known splicing effector domains and the splicing reporter were co-transfected into HEK293T cells, and exon inclusions were assayed by semi-quantitative RT-PCR (see STAR Methods). Mean values ± SD (n = 3) were plotted.

(D) Regulation of splicing by ESFs that bind to their cognate site at the downstream intron of a cassette exon in a splicing reporter. The experimental details are the same as (C), except that a different splicing reporter was used. Mean values ± SD (n = 3) were plotted.

See also Figures S1 and S2.

than previously expected. This experimentally tested dataset further enabled us to develop a machine learning approach to analyze the association between RBP sequence compositions and their activities in regulating splicing and allowed us to predict and validate splicing regulatory activities of RBPs according to their sequence compositions. Our findings expanded the repertoire of potential splicing factors. To our best knowledge, the framework developed in this study (large survey followed by machine learning) represents the first systematic modeling of RBP functions based on their sequence compositions, which may be useful to investigate other RNA processing pathways beyond splicing regulation.

RESULTS

Design and Validation of ESFs That Regulate Alternative Splicing

Taking advantages of the unique RNA-binding mode of the PUF domain, we have previously designed RBDs with customized specificity to modulate RNA metabolism (Cheong and Hall, 2006; Wei and Wang, 2015). The PUF domain of human PUM1 contains 8 imperfect tandem repeats of ~36 amino acids in three α helices and can recognize 8 nucleotides in an anti-parallel fashion, with each repeat directly interacting with the Watson-Crick edge of the RNA base through hydrogen bonds (Wang et al., 2002) (Figure 1A). Therefore, a PUF scaffold can be modified to recognize any 8-nt RNA sequence with a reprogrammable RNA binding code (Cheong and Hall, 2006; Dong et al., 2011; Filipovska et al., 2011). By combining a customized PUF domain with splicing regulatory domains, we were able to specifically promote or suppress AS of various targets (Qi et al., 2016; Wang et al., 2009; Wang et al., 2012; Wang et al., 2013).

To design an optimal ESF construct for evaluation of different RBPs in controlling splicing, we first test whether the relative position of the PUF and functional domain will affect the activities of ESFs by constructing two types of ESFs with the PUF domain at either the N- or C-terminal of the fusion proteins (Figure 1B). A modified PUF domain, PUF(6-2/7-2), with mutations in repeat 6 and 7 (N1043S/Q1047E and S1079N/E1083Q) to recognize 5′-UUGAUAUA (Cheong and Hall, 2006; Wang et al., 2009), was used to avoid potential interference of endogenous PUM1 in human cells. We chose two classes of functional domains, the arginine/serine–rich domain from SRSF7 and the glycine-rich domain from hnRNP A1, as the representative splicing activator and repressor, respectively. We also included a nuclear localization signal for proper subcellular localization and a FLAG epitope for detection (Figure 1B). The resulting ESFs were then co-expressed with two splicing reporters, of which

the first and third exons are split-GFP exons separated by an alternative exon and its flanking introns. The cognate binding sites of the PUF domain were inserted either inside or downstream of the alternative exon (Figures 1C and 1D, left panels), thus the splicing regulatory activities of ESFs can be measured by RT-PCR using primers corresponding to GFP exons (see STAR Methods).

Consistent with previous studies, ESFs containing the SR domain significantly promoted splicing when binding to exons but inhibited splicing when binding to introns, whereas ESFs containing the Gly-rich domain repressed cassette exon inclusion in both positions (Figures 1C and 1D). Such context-dependent splicing regulatory activities of the functional domains were consistent across different reporters containing distinct exons/introns, as we found that they show similar activities when binding to exons versus introns in another two reporters (Figures S1A and S1B). Importantly, the relative positions of PUF and the functional domain within the fusing proteins did not alter how the ESFs regulate AS in the tested splicing reporters, indicating that the way by which the ESFs were constructed has minimal impact on the splicing regulatory activities of functional domains in our experimental system. In addition, ESFs showed consistent splicing regulatory activities only when the splicing factors were specifically tethered to their target pre-mRNAs through PUF-RNA interaction, whereas the control mCherry-fusion proteins of same factors showed little splicing regulatory activity (or inconsistent with the known activities) (Figure S2). Therefore, ESFs designed using the above strategies can be employed as a reliable tool to test the function of different RBPs in controlling splicing.

ESFs Containing the Low-Complexity Domains from Various RBPs Display Context-Dependent Splicing Regulatory Activities

The majority of the current knowledge on splicing regulation is derived from studies of several typical splicing factors such as SR proteins and hnRNPs. Most of these canonical factors contain one or more RRM domains to bind splicing silencers or enhancers in pre-mRNAs and the low-complexity domains with biased amino acid composition (e.g., RS domain or Gly-rich domain) to affect exon inclusion (Del Gatto-Konczak et al., 1999; Graveley and Maniatis, 1998). Since a significant fraction of RBPs, especially the RRM-containing RBPs, have low-complexity fragments (Castello et al., 2012; Tsai et al., 2014), we speculated that these domains may function to regulate AS when the RBPs bind to pre-mRNAs.

To test this hypothesis, we collected all RRM-containing RBPs in UniProt database and classified them according to the different compositional biases in their low-complexity domains. Based on UniProt annotation, we selected 12 classes of domains from 51 RBPs with the highest frequency of occurrence (Figure 2A; see details in Table S1), including Pro-rich domain, RS domain, Arg-rich domain, G-patch domain, Tyr-rich domain, etc., to generate different ESFs using the strategy outlined in Figure 1. We also constructed ESFs with synthetic polypeptides containing simple repeats of single or dual amino acids enriched in these RBPs (e.g., poly-Pro and poly-RS) (Figure 2A). These ESFs were co-expressed with a splicing reporter containing the cognate binding site in a cassette exon (Figure 2B), and the inclusion of the cassette exon was analyzed using semi-quantitative RT-PCR (Figures 2B and 2C). The activity of each ESF was then represented by the relative change of exon inclusion compared to the control sample in which the same reporter was co-expressed with the PUF domain alone (Figure 2D).

We found that, in addition to the known RS and Gly-rich domains, many other low-complexity domains from RBPs can affect the splicing of alternative exons either positively or negatively (Figure 2D). For example, the domains with charged amino acids (e.g., Asp-rich, Lys-rich, Arg-rich, Arg/Ser-rich domains) generally promoted exon inclusion, while most uncharged repeats (e.g., Tyr-rich, Ala-rich, Pro-rich domains) suppressed exon inclusion. As a control, the PUF fusion proteins containing six randomly selected peptides showed little effect on exon inclusion when co-expressed with the same splicing reporter (Figures S3A and S3B), further confirming the splicing regulatory activities of these low-complexity domains. Similar results were obtained using another splicing reporter with an intronic PUF binding site downstream of the cassette exon (Figure S4). Taken together, these observations indicate that the majority of low-complexity domains in RBPs can regulate AS with consistent activities when binding to pre-mRNA (51 out of 63 domains tested caused more than 30% change in the relative fold of percent-spliced-in [PSI] compared to the PUF-only control), expanding the repertoire of potential splicing factors in the human genome and suggesting that AS is under more extensive regulation by a variety of RBPs in human cells than previously expected.

A thorough survey of putative functional domains in RBPs using two different splicing reporters also produced a general picture for the context-dependent activities of splicing regulators. For each domain, we measured its activity to control exon inclusion upon binding to the cognate sites in two different splicing reporters and plotted their splicing regulatory activity from exonic or intronic sites (Figure 3). We found that the domains enriched with charged amino acid residues (such as SR domains, Arg-rich domains, and poly-His domains) tended to enhance splicing when binding to exons but inhibited splicing from introns. On the other hand, Tyr-rich or Pro-rich domains tended to have an opposite context-dependent activity (i.e., enhanced splicing from introns but inhibited splicing from exons). There were also some domains, such as Gly-rich domains, that inhibited splicing from both exonic and intronic contexts (Figure 3). Intriguingly, none of the tested domains promoted splicing from both exonic and intronic contexts, and there were many more proteins to inhibit splicing from intronic context than to promote splicing. In addition, there were more domains generally functioning as splicing inhibitors than activators (Figure 3; more blue areas than red), implying that the inclusion of cassette exons is more dominantly regulated by negative control than by positive control.

Develop a Machine Learning Approach to Classify Splicing Factors

The correlation between the sequence composition of RBPs and their activities in regulating splicing observed in the above experiments raised the possibility that a computational model may be generated to predict the splicing regulatory activity of any given RBP based on its amino acid composition. To test this possibility, we developed a machine learning approach using the results

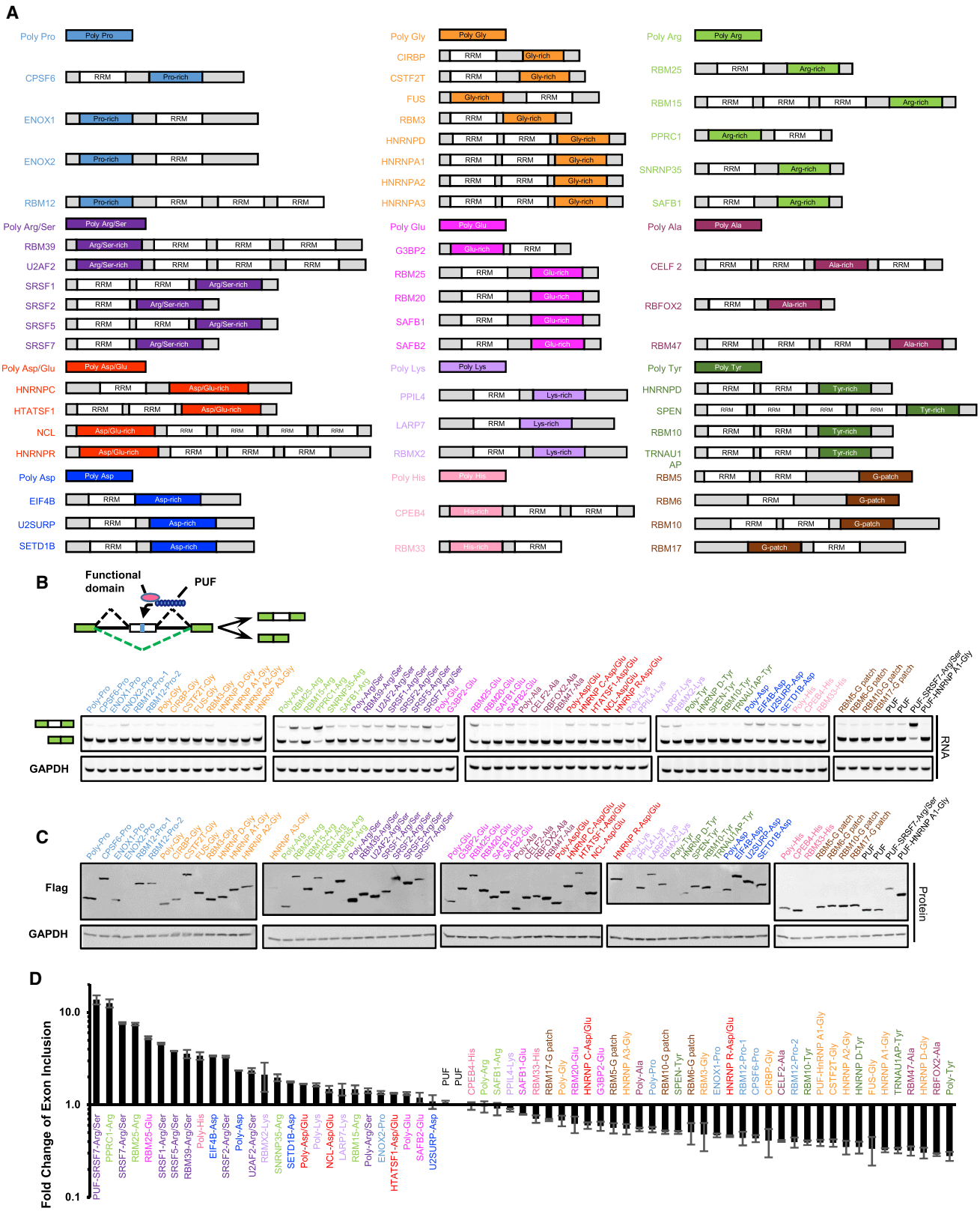


Figure 2. Determine the Splicing Regulatory Activities for Various Functional Domains Binding to Cassette Exon

(A) Classification of potential functional domains of various RBPs based on the UniProt annotation. Low-complexity domains in RRM-containing RBPs were indicated with different colors according to their distinct compositional biases. RRMs were marked with black on a white block. The other regions were plotted by the gray box.

(legend continued on next page)

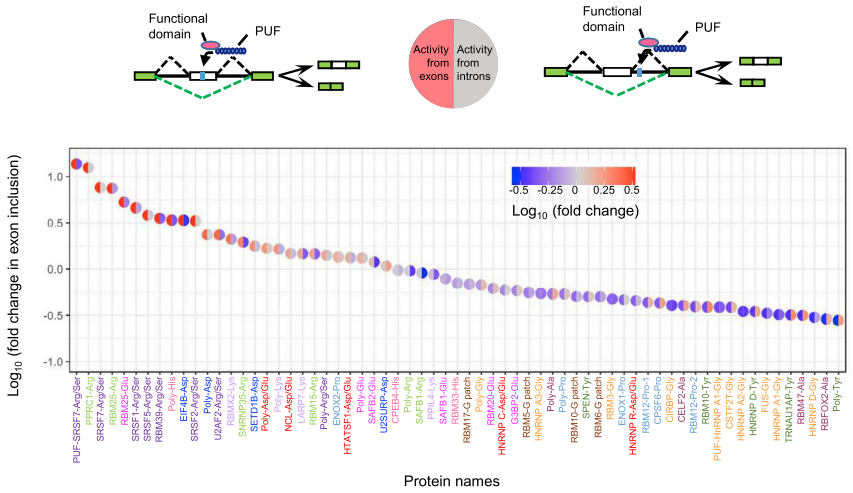


Figure 3. Context-Dependent Activities of Various ESFs with Different Effector Domains

The activities of each ESF in the two splicing models were represented by two semicircles. The left semicircle showed the splicing regulatory activity when binding to the exonic site, whereas the right semicircle showed activity when binding to the intronic site. The relative changes of PSI were color coded with different shades of each semicircle, and all ESFs tested were plotted in the descending order of their activities at exonic context (data from Figures 2 and S4).

See also Figure S4.

from the 63 experiments as the training dataset. Based on the experimental data from the exonic context (Figure 2), we classified all putative functional domains into three groups: splicing activators, splicing inhibitors, and neutral domains, which were defined by the relative fold change in exon inclusion at the threshold of ±0.15 in common logarithm (roughly equivalent to changes of less than 70% for inhibitors or larger than 140% for activators, based on the experimental data from the exonic context [Figure 2]). We further used features of amino acid composition to represent each functional domain (Figure 4A; also see STAR Methods) and assigned an arbitrary value to represent each activity class (−1 for inhibitors, 0 for neutral domains, and 1 for activators). We chose the partial least-squares discriminant analysis (PLS-DA) to solve the link between these features and the functional classification of each domain, as this algorithm can select the predictive features and reduce the dimensions at the same time (Figure 4A; also see STAR Methods).

The basic idea of PLS-DA is to estimate the linear regression between the predictor matrix X(n × p) and the response matrix Y(n × q) with a score matrix T (Figure S5; also see STAR Methods). To compute this model, we assigned the frequency of single and dual amino acids in each domain as features to compose the predictor matrix X(n × p) and used a response matrix Y(n × q) to represent the activity classification of each functional domain. We trained the model with the experiment dataset (Figure 2D) to reduce the dimension of the predictor matrix into two principal components that were subsequently used to generate a two-dimensional scatterplot (Figure 4B). We found that the splicing activators and inhibitors are well separated by the PLS-DA model, indicating that the sequence composition is indeed predictive to the splicing regulatory activity in the exonic context (Figure 4B). As expected, the neutral factors laid between the two classes of splicing regulators; however, the separation was somewhat weaker as judged by overlaps of its 95% confidence ellipse with the other two classes (Figure 4B; overlap of the gray circle with the other two circles).

We further fitted the sequence composition of all the RRM-containing RBPs into this predictive model (excluding splicing factors used to train the model and RRM domains themselves, motif boundaries determined by UniProt annotation) and found that the majority of them are predicted as the splicing inhibitors by our PLS-DA model (Figure 4C) (see more details in Table S3). Intriguingly, only a small fraction of all factors (9%, 44 out of 485 RRM-containing RBPs analyzed) was predicted as neutral factors, probably due to the small number of neutral factors in the training data and poor classification in our model.

In addition, we also used the PLS-DA model to analyze the splicing regulatory activities of the same set of RBPs from the intronic context (Figures S6A and S6B; data from Figure S4) and found that the classification of their activities is less clear, suggesting a weak correlation between RBP sequences and their activities at intronic sites. Besides PLS-DA, several other algorithms were also tested (such as principal-component analysis, Figure S7); however, they failed to produce clear classification in RBPs with different activities.

Validation of the Predictive Model

To quantitatively evaluate the relative strength of any given protein domain in regulating splicing in our predictive model, we defined a simple score system. We first applied the pre-trained PLS-DA model to calculate the coordinates of the given peptide by extracting the sequence features of any given protein fragment (Figure 5A, point G). We then defined a single score using the relative distances from the point G to the center of the confidence ellipses for splicing activators (Figure 5A, point B) and

(B) Promotion or inhibition of exon inclusion by various ESFs that bind to cassette exon. The expression vectors of the ESFs containing various putative effector domains were co-expressed in HEK293T cells with the splicing reporter containing a cognate PUF binding site within the alternative exon. The inclusion of cassette exon in each sample was assayed by semi-quantitative RT-PCR.

(C) Western blot analyses of ESF expression using antibodies recognizing the FLAG tag, with detection of GAPDH as loading controls.

(D) Quantification results of semi-quantitative RT-PCR as shown in Figure 2B. The fold changes of percent-spliced-in (PSI) were normalized to the control sample in which the same splicing reporter was co-expressed with the PUF domain only (mean ± SD, n = 4) and were in descending order.

See also Figures S3 and S12.

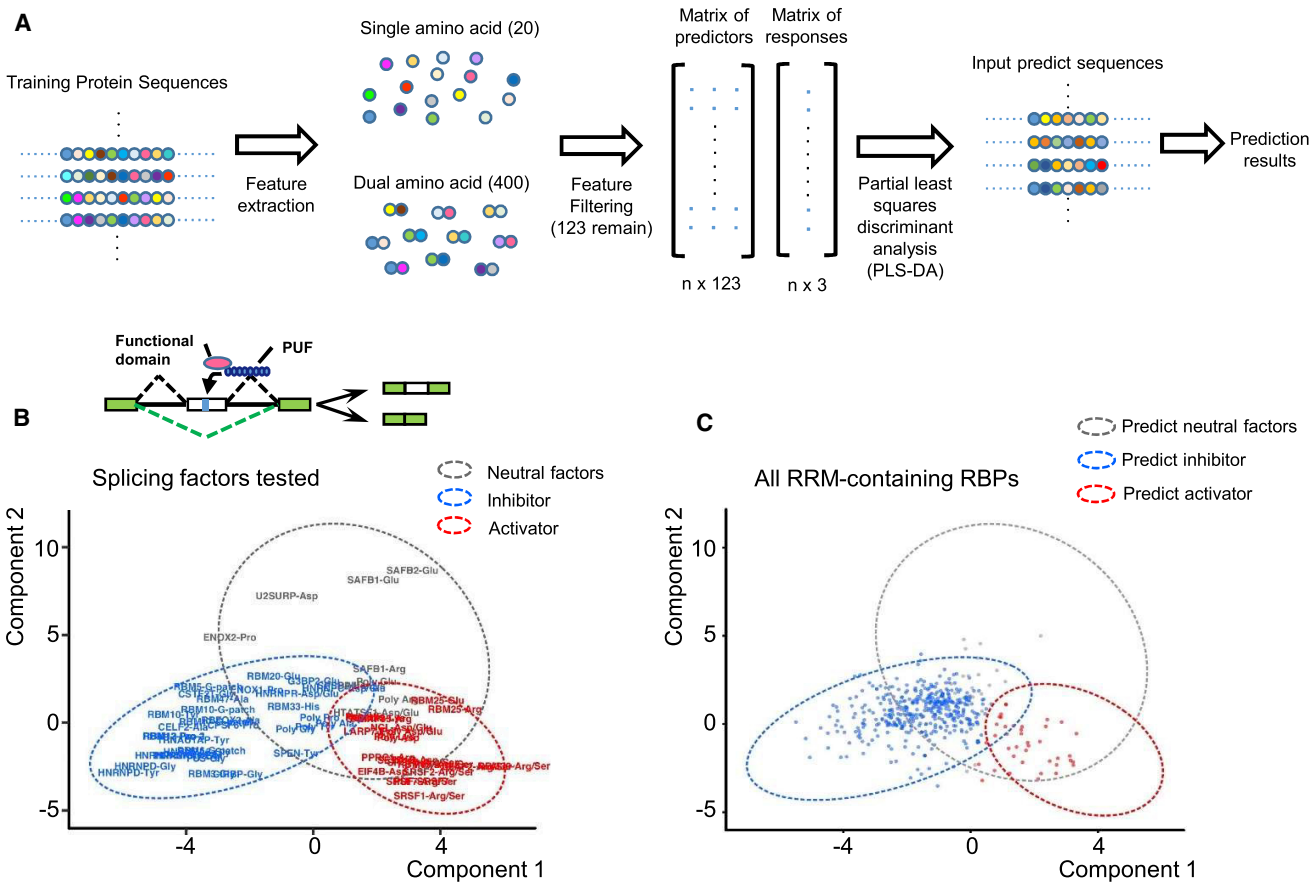


Figure 4. Predict Splicing Regulator Activity from Sequence Composition with a Machine Learning Approach

(A) Workflow of the predictive model constructed with sparse partial least-squares discriminant analysis (sPLS-DA).

(B) Classification of the splicing regulatory activities for all experimentally validated ESFs. The functional domain of each ESF was plotted in a scatterplot of the first two components. The 95% confidence ellipses for the three classes of splicing regulators were calculated from the training dataset.

(C) Prediction of all RRM-containing RBPs for their activities in splicing regulation. The same confidence ellipses in Figure 4B were used here as the threshold for classification of putative splicing regulators. The predicted inhibitors are represented as blue dots, the predicted activators are represented as red dots, and the predicted neutral factors are represented as gray dots.

See also Figures S5–S7.

inhibitors (Figure 5A, point A). The final score was calibrated to the range of [−1, 1] and denoted as score *S*, with a larger value representing a higher possibility to function as a splicing activator.

When applying this scoring system to the training set (Figure 5B, upper panel) and the rest of RBPs (Figure 5B, bottom panel), we again found that many more endogenous RBPs are predicted as splicing inhibitors than splicing activators in both datasets. To experimentally validate the predictions, we constructed ESFs using representative endogenous RBP fragments from different *S* score ranges and determined their activities in regulating the inclusion of a cassette exon (Figure S8). We found a positive correlation between *S* score and the fold changes of splicing relative to the PUF-only control (Figure 5C), suggesting that the predictive score calculated based on our predictive model is indeed correlated with the RBP activity in the regulation of splicing. We noticed that the scores for neutral factors were shifted toward the negative values rather than centered at *S* = 0, which is likely due to the fact that the confidence ellipses for splicing inhibitors and activators are not symmetric, with splicing inhibitors spreading into a much larger area (Figures 4C and 5A). Collectively, these results indicated that the predictive model developed in our study could be used to predict the relative strength of any given protein domain in regulating splicing. It is worth noting that the similar model could not predict well for the regulation from the intronic context (Figure S9), consistent with the poor classification of the training set in such a case (Figure S6A).

Composition of Peptides Determines the Splicing Activities

One advantage to use the PLS-DA model for splicing factor classification is that we can evaluate each sequence feature in the weight predictor matrix for its relative contribution to the final prediction result. To this end, we generated the loading plot for each sequence feature (i.e., the frequency of single and dual amino acids), which gives an estimation of the predictive power for each sequence feature in classifying splicing activators,

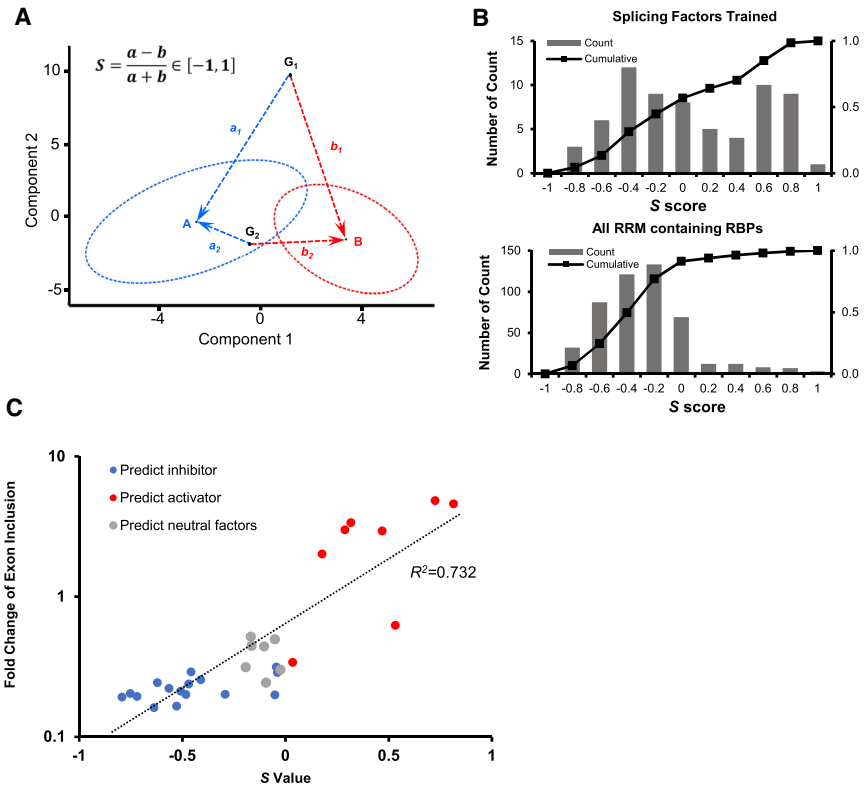


Figure 5. Validation of the Predictive Model

(A) Schematic diagram of the scoring system for prediction of splicing regulatory activity. The confidence ellipses for splicing activators and inhibitors are indicated with red and blue ovals. Point A represents the center of the blue ellipse; point B represents the center of the red ellipse. For any given factors, we calculate their positions (represented with points G_1, G_2, etc.) with our pre-trained sPLS-DA model and determine their distances (*a* and *b*) to the points A and B. The *S* score of each factor is computed from the relative distance from the center of two confidence ellipses.

(B) Frequency distribution of *S* scores for the trained splicing factors (upper panel) and all RRM-containing RBPs (lower panel).

(C) Correlation between relative changes of exon inclusion and the *S* scores of the experimentally verified domains from endogenous RBPs. The predicted inhibitor, activator, and neutral factors are indicated in different colors.

See also Figures S8, S9, and S12.

splicing inhibitors, or neutral factors. As presented in Figure 6A (see more details in Table S4), the loading plot was divided into three parts corresponding to the classification; the red area was highlighted as the predictive features for splicing activator, whereas the blue area indicated the inhibitors and the gray area represented the features that are indicative for neutral factors for splicing regulation. There were only a few features that correlated to the classifications; the distances between the sequence features to the center of the circle represented their contribution weights.

We found that this analysis recapitulated some of the known features that are indicative of the activities of known splicing factors, reassuring the reliability of the loading plot analysis. For example, the SR or RS dipeptides are strong predictors for splicing activators (e.g., SR proteins), while the G and GY are strongly associated with splicing inhibitors (such as members in the hnRNP A1 family). More importantly, this analysis also allowed us to identify new features that may determine the splicing regulatory activities of RBPs. For example, the low-complexity fragments enriched with R, D, and S residues and KR and SD dipeptides may function as splicing activators; the domains enriched with Y, N, NQ, or SG sequences may function as splicing inhibitors; and the predictive features for neutral factors including RE, RD, EQ, etc. (Figure 6A, right panel).

This loading plot analysis also represented the relative contribution for each sequence feature to splicing regulatory activity, which can be directly tested using short repetitive peptides consisting solely of different single or dual amino acids in each feature. Therefore, we synthesized additional ESFs by fusing the PUF domain with different repetitive peptides and tested their activities by co-expressing them with the splicing reporter containing the cognate exonic binding site (Figures 6B and S10). Most predicted features for splicing activators or inhibitors indeed showed expected activities: 7 of 10 predicted activator peptides increased exon inclusion, while all 7 predicted inhibitor peptides decreased exon inclusion (comparing red and blues features in Figure 6B). Moreover, we found that some of the new features even have stronger splicing regulatory activities. For example, the repetitive sequences of KR, K, D, or DE function as stronger splicing activators than the well-known SR repeats (Figure 6B). Given the fact that we made an aggressive prediction based only on simple assumptions, the correlation between the sequence composition and splicing is actually quite impressive. However, we found that many repetitive peptides with predicted neutral activities also regulate the exon inclusion positively or negatively, although on average these features are more "neutral." This high false negative rate for neutral factors may indicate that our model has low sensitivity in extracting sequence features that do not affect splicing, which is probably due to the small number of neutral factors in our training dataset.

Our analysis also enabled *de novo* design and synthesis of splicing regulators using the relative contribution of each sequence feature to the predicted splicing regulatory activity. To this end, we generated novel ESFs with 11 synthetic peptides that are predicted as splicing activators or inhibitors and tested their activities using the same splicing reporter system (Figures 6C and S10C). We found that 10 out of 11 *de novo* designed peptides showed expected activities, with the exception of only one predicted splicing inhibitor (peptide 5). These results demonstrated the feasibility of *de novo* engineering of proteins with customized splicing regulatory activities, paving the way for designing artificial splicing factors as gene manipulation tools.

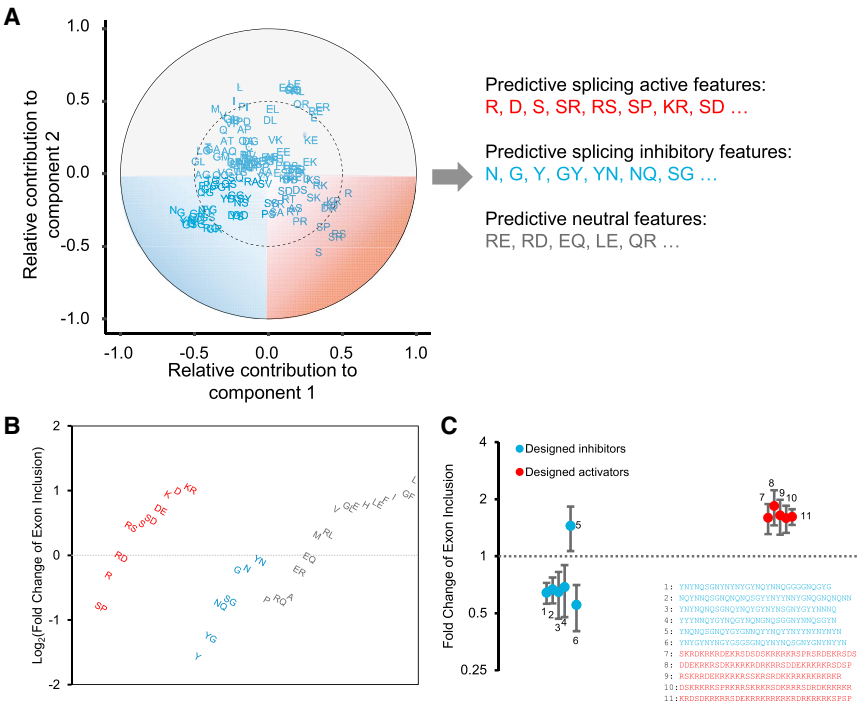


Figure 6. Sequence Compositions of Splicing Factors Determine their Regulatory Activities

(A) Relative contributions of various sequence features in the predictive model, which were calculated by the loading matrix of sPLS-DA model. The red quarter circle covers the features indicative of the predictive features for splicing activators, the blue quarter circle covers the features indicative of the predictive features for splicing inhibitors, and the gray upper semicircle covers the features indicative of the predictive features for neutral factors. The distances from the circle center correspond to the predictive power for each feature.

(B) Validation of different sequence features for the contribution to splicing regulatory activity. The repetitive sequences of different predictive features were used as the effector domains to construct new ESFs, whose activities were determined with a splicing reporter containing cognate binding sites in cassette exon. The experimental details are the same as Figure 2. Red, predictive features for splicing activators; blue, predictive features for splicing inhibitors; gray, predictive features for neutral factors.

(C) Regulation of exon inclusion by ESFs containing *de novo* designed peptides with predicted activities as a splicing inhibitor (blue) and activator (red). The experimental details are the same as Figure 2, and the fold changes of exon inclusion were compared to the PUF-only control (mean ± SD, n = 3).

See also Figures S10–S12.

DISCUSSION

The regulation of AS involved complex functional interactions between *cis*-regulatory elements and *trans*-acting splicing factors. While a number of studies have systematically identified splicing regulatory *cis*-elements by computational prediction and experimental screen (Fairbrother et al., 2002; Ke et al., 2011; Rosenberg et al., 2015; Wang et al., 2012; Wang et al., 2013; Wang et al., 2004; Yu et al., 2008; Zhang and Chasin, 2004), the systematic identification and study of all potential *trans*-acting splicing factors are still missing. Here, we have combined both computational and experimental approaches to systematically examine the RRM-containing RBPs, the largest class of RBPs, for their potential activities in regulating splicing in two different pre-mRNA contexts. We found that most RBPs analyzed can function as either splicing activators or inhibitors through their low-complexity domains, and many function differently in the exonic or intronic contexts (Figure 3). For the regulation at the exonic context, we further developed a machine learning approach to extract key predictive features for splicing activators and inhibitors, which can be used in designing peptides with a customized function in regulating splicing. To our knowledge, this work presents the first large-scale, unbiased survey and predictive identification of potential splicing factors, considerably expanding the current scope of known splicing factors.

This study has been primarily focused on the regulation of skipped exons (a.k.a., cassette exons), which are the most common class of AS events and account for more than a third of all AS events observed in humans (Wang et al., 2008). However, the similar procedures may be easily adopted to the other types of AS events, such as intron retention and alternative 5′ or 3′ splice site usage. In particular, the intron retention has recently been discovered as a key regulatory mechanism in stress response, cell cycle, cancer cell proliferation, and neuronal activity (Boutz et al., 2015; Braun et al., 2017; Dominguez et al., 2016; Mauger et al., 2016), and it will be important to systematically survey all RBPs for their activities in controlling intron retention in future studies.

This study also brings up some unanswered questions. For example, when attempting to classify the splicing regulatory activities of all RBPs, we found that the majority of analyzed RBPs have negative *S* scores (i.e., predicted as splicing inhibitors; Figure 5C), although the training data have a more balanced score distribution. Such bias toward splicing inhibitors was more obvious in the intronic context (Figures 3 and S4). This unbalanced score distribution may be (partially) due to the fact that the two confidence ellipses were asymmetric, so that the *S* score may not classify peptides strictly. However, this unbalanced score distribution could also be caused by the small number of training samples used in our original experiments or, alternatively, because exon inclusion is mostly controlled through negative regulation inside cells (i.e., the long AS isoform is the default). We also found that although the majority of ESFs can be robustly expressed, a few of ESFs expressed poorly even when they showed robust activities in regulating splicing. For example, the ESFs with K-rich or Y-rich domains both showed low protein levels as judged by western blotting but had strong activities in

splicing regulation (Figure 2). With increasing concentrations of poly-Y-PUF expression vectors, we further observed that Y-rich sequence consistently inhibited splicing from exonic contexts in a dose-dependent fashion (Figure S11), confirming the splicing regulation activities observed at low protein levels. Since all ESFs were expressed with the same plasmid backbone and cellular environment, this observation is most likely due to the fast degradation of certain sequences. To reduce potential false positive rates that may be caused by the low baseline exon inclusion level of the original splicing reporter in analyzing splicing inhibitors, we used an additional reporter with higher baseline PSI level to reconfirm a subset of 31 targets, including both endogenous RBPs and synthetic repetitive peptides of predictive features (Figure S12). We found that results were generally consistent across two reporters, with most tested fusion proteins showing inhibitory activities, suggesting that the results in this study are reliable.

Compared to all other proteins, RBPs are significantly enriched with low-complexity fragments containing repetitive sequences (Tsai et al., 2014). While previous studies have demonstrated that some of these low-complexity domains (e.g., RS domain, Gly-rich domain, and Ala-rich domain) can affect AS upon specifically binding to their cognate pre-mRNA targets, this study presents a systematic comparison of the splicing regulatory activity for a comprehensive list of such domains. We demonstrated that most low-complexity domains can positively or negatively affect splicing of alternative exon when recruited to pre-mRNA (Figures 2 and 4B), and the majority of these domains show different splicing regulatory activities between exonic and intronic pre-mRNA contexts (Figure 3). In addition, the low-complexity regions of an RBP generally have the same type of splicing regulatory activity compared to the full-length protein (i.e., both being activators or inhibitors). In rare cases, there are a few RBPs with multiple low-complexity domains of different activities. For example, SREK1 contains a short inhibitory motif and long RS-rich activator motif, and the entire RBP functioned as a splicing activator in our system (Figure S8). The activities of such RBPs are probably determined by functional interplay of these motifs and thus should be studied as exceptional cases.

Our experimental system using PUF fusion proteins has also shown results similar to other tethering systems. For example, the Ala-rich domain of Rbfox-2 (at the C-terminal domain) was found to inhibit splicing from exons and activate splicing from introns (Figure 3), which is consistent with the results using MS2 tethering system (Damianov et al., 2016; Sun et al., 2012; Ying et al., 2017). Additionally, we found that the Tyr-rich domains function as splicing inhibitors when tethered to exons, supporting the recent study using hnRNP A and D families (Gueroussov et al., 2017). To derive a simple and predictive model, our experiments mainly focused on the RBPs with clear RBDs and potential functional domains, thus missing some non-canonical splicing factors. For example, the SR-related protein nSR100 contains only an SR-rich domain but lacks known RBDs; this protein regulates an extensive network of brain-specific alternative exons (Calarco et al., 2009). It will be an exciting challenge to include more complicated and unconventional splicing factors in future modeling.

The detailed mechanisms of how these domains affect splicing may be quite diverse even when they show similar activities. For example, RS domains may enhance splicing by specifically binding to ESEs or branch sites to promote spliceosome assembly (Shen and Green, 2004; Shen et al., 2004), whereas the Pro-rich domain of DAZAP1 can enhance splicing by binding to and neutralizing general splicing inhibitors (Choudhury et al., 2014). In addition, many low-complexity domains in RBPs tend to aggregate to assemble into fibrous structures and hydrogels (Molliex et al., 2015; Murakami et al., 2015), and such self-assembly probably plays a key role in mediating splicing regulation. For example, the self-assembly of the low-complexity C-terminal domain (Tyr-rich sequences) of Rbfox is essential for its activity in regulating splicing (Ying et al., 2017), and the Gly-rich domain in hnRNP A1 may function by self-assembly and propagate along the pre-mRNA targets (Zhu et al., 2001). Our results systematically categorized the low-complexity domains that control splicing through diverse mechanisms and suggested that the sequence compositions of the low-complexity domains determine the splicing regulatory activities of these RBPs. However, it is worth noting that our results were obtained with an artificial system by which the low-complexity domains of RBPs were specifically tethered to pre-mRNAs, and thus the fact that 51 out of 63 domains induce splicing changes in our system does not necessarily prove that they will act as splicing factors in the natural context. For example, the endogenous RBPs localized predominantly outside the nucleus may not bind to pre-mRNAs to control splicing *in vivo*. It is also known that some domains (such as SR proteins) can regulate splicing in a sequence-independent fashion. Such sequence-independent regulatory activity is not the focus of this study because it generally involves the regulation of spliceosome assembly that affects splicing of all exons.

Solving the relationship between the protein sequence and its function is the foundation for *de novo* engineering of proteins with designed activities. However, such a task has only achieved some successes on proteins with defined structures that determine their function (Huang et al., 2016; Moal et al., 2013), such as the massive design of mini-proteins that target influenza (Chevalier et al., 2017). However, the non-structured flexible regions of many proteins often determine their function, and designing non-structured domains is a more difficult problem, especially when they may function through diverse mechanisms. Our study presented a paradigm where the activities of such fragments can be massively assayed in parallel with a cell-based reporter system, generating a large training dataset that enables learning of the relationship between sequence and function without the prior knowledge of structure. Given the fact that predicting the function of unstructured domains from their sequences is an ambitious goal, the true positive rates obtained in this study are encouraging and reasonably reliable despite some mis-classifications. We expect that such a predictive model could be further refined by considering additional parameters (e.g., expression level, turnover rate, and localization of RBPs) or using additional types of reporters. To our knowledge, this machine learning framework represents the first systematic inspection and prediction on the direct link between RBP sequence and its possible function and may be useful to study other RNA processing pathways beyond splicing regulation.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cell Culture
- METHOD DETAILS
  - Cell Transfection
  - Plasmids Construction
  - RNA Extraction and Semi-Quantitative RT-PCR
  - Western Blotting
  - Sparse Partial Least Squares Discriminant Analysis (sPLS-DA)
  - Model Construction
  - Random Peptides Generation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Image and Data Analysis Semi-Quantitative of RT-PCR
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes twelve figures and four tables and can be found with this article online at https://doi.org/10.1016/j.cels.2018.09.002.

ACKNOWLEDGMENTS

We thank Ms. Yun Jiang for help in preparing the paper and members of Wang lab for discussion and comments on this manuscript. This work is supported by the National Natural Science Foundation of China (31570823, 31661143031, and 31730110 to Z.W.), Science and Technology Commission of Shanghai Municipality (17JC1404900 to Z.W. and 18XD1404400 to H.W.), and the National Postdoctoral Program for Innovative Talents (BX20180336 to M.M.). This work is also supported, in part, by the Intramural Research Program of National Institute of Environmental Health Sciences of the National Institutes of Health of USA (Z01 ES102205 to X.L.). M.M. is supported by a Chinese Scholarship Council Scholarship (201406740040) and is a participant in the NIH Graduate Partnership Program. Z.W. is supported by the type A CAS Pioneer 100-Talent program.

AUTHOR CONTRIBUTIONS

Conceptualization, Z.W.; Methodology, M.M., Y.H., and Z.W.; Software, Y.H.; Formal Analysis, Y.H.; Investigation, M.M., Yun.Y., Y.Q., H.W., and W.F.; Writing – Original Draft, M.M., Y.H., and Z.W.; Writing – Review & Editing, M.M., Y.H., Yi.Y., X.L., and Z.W.; Funding Acquisition, M.M., H.W., X.L., and Z.W.

DECLARATION OF INTERESTS

Z.W. has co-founded a company, Enzerna Biosciences, Inc., to commercialize the artificial RNA binding protein using PUF scaffold. The other authors declare no competing interests.

Received: November 24, 2017
Revised: May 10, 2018
Accepted: September 19, 2018
Published: November 7, 2018

REFERENCES

Baralle, F.E., and Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. Nat. Rev. Mol. Cell Biol. *18*, 437–451.

Boutz, P.L., Bhutkar, A., and Sharp, P.A. (2015). Detained introns are a novel, widespread class of post-transcriptionally spliced introns. Genes Dev. *29*, 63–80.

Braun, C.J., Stanciu, M., Boutz, P.L., Patterson, J.C., Calligaris, D., Higuchi, F., Neupane, R., Fenoglio, S., Cahill, D.P., Wakimoto, H., et al. (2017). Coordinated splicing of regulatory detained introns within oncogenic transcripts creates an exploitable vulnerability in malignant glioma. Cancer Cell *32*, 411–426 e411.

Calarco, J.A., Superina, S., O'Hanlon, D., Gabut, M., Raj, B., Pan, Q., Skalska, U., Clarke, L., Gelinas, D., van der Kooy, D., et al. (2009). Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. Cell *138*, 898–910.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M., et al. (2012). Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell *149*, 1393–1406.

Chabot, B., and Shkreta, L. (2016). Defective control of pre-messenger RNA splicing in human disease. J. Cell Biol. *212*, 13–27.

Cheong, C.G., and Hall, T.M. (2006). Engineering RNA sequence specificity of Pumilio repeats. Proc. Natl. Acad. Sci. USA *103*, 13635–13639.

Chevalier, A., Silva, D.A., Rocklin, G.J., Hicks, D.R., Vergara, R., Murapa, P., Bernard, S.M., Zhang, L., Lam, K.H., Yao, G., et al. (2017). Massively parallel de novo protein design for targeted therapeutics. Nature *550*, 74–79.

Choudhury, R., Roy, S.G., Tsai, Y.S., Tripathy, A., Graves, L.M., and Wang, Z. (2014). The splicing activator DAZAP1 integrates splicing control into MEK/Erk-regulated cell proliferation and migration. Nat. Commun. *5*, 3078.

Cooper, T.A., Wan, L., and Dreyfuss, G. (2009). RNA and disease. Cell *136*, 777–793.

Damianov, A., Ying, Y., Lin, C.H., Lee, J.A., Tran, D., Vashisht, A.A., Bahrami-Samani, E., Xing, Y., Martin, K.C., Wohlschlegel, J.A., et al. (2016). Rbfox proteins regulate splicing as part of a large multiprotein complex LASR. Cell *165*, 606–619.

David, C.J., and Manley, J.L. (2010). Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes Dev. *24*, 2343–2364.

Del Gatto-Konczak, F., Olive, M., Gesnel, M.C., and Breathnach, R. (1999). hnRNP A1 recruited to an exon in vivo can function as an exon splicing silencer. Mol. Cell. Biol. *19*, 251–260.

Dominguez, D., Tsai, Y.H., Weatheritt, R., Wang, Y., Blencowe, B.J., and Wang, Z. (2016). An extensive program of periodic alternative splicing linked to cell cycle progression. ELife *5*, https://doi.org/10.7554/eLife.10288.

Dong, S., Wang, Y., Cassidy-Amstutz, C., Lu, G., Bigler, R., Jezyk, M.R., Li, C., Hall, T.M., and Wang, Z. (2011). Specific and modular binding code for cytosine recognition in Pumilio/FBF (PUF) RNA-binding domains. J. Biol. Chem. *286*, 26732–26742.

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

Filipovska, A., Razif, M.F., Nygård, K.K., and Rackham, O. (2011). A universal code for RNA recognition by PUF proteins. Nat. Chem. Biol. *7*, 425–427.

Fu, X.D., and Ares, M., Jr. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. Nat. Rev. Genet. *15*, 689–701.

Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. Nat. Rev. Genet. *15*, 829–845.

Graveley, B.R. (2000). Sorting out the complexity of SR protein functions. RNA *6*, 1197–1211.

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

Gueroussov, S., Weatheritt, R.J., O'Hanlon, D., Lin, Z.Y., Narula, A., Gingras, A.C., and Blencowe, B.J. (2017). Regulatory expansion in mammals of multivalent hnRNP assemblies that globally control alternative splicing. Cell *170*, 324–339 e323.

Huang, P.S., Boyken, S.E., and Baker, D. (2016). The coming of age of de novo protein design. Nature *537*, 320–327.

Ke, S., Shang, S., Kalachikov, S.M., Morozova, I., Yu, L., Russo, J.J., Ju, J., and Chasin, L.A. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. *21*, 1360–1374.

Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature *509*, 575–581.

Kornblihtt, A.R., Schor, I.E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M.J. (2013). Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nat. Rev. Mol. Cell Biol. *14*, 153–165.

Lê Cao, K.A., Boitard, S., and Besse, P. (2011). Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics *12*, 253.

Li, R., Dong, Q., Yuan, X., Zeng, X., Gao, Y., Chiao, C., Li, H., Zhao, X., Keles, S., Wang, Z., et al. (2016). Misregulation of alternative splicing in a mouse model of Rett syndrome. PLoS Genet. *12*, e1006129.

Matera, A.G., and Wang, Z. (2014). A day in the life of the spliceosome. Nat. Rev. Mol. Cell Biol. *15*, 108–121.

Mauger, O., Lemoine, F., and Scheiffele, P. (2016). Targeted intron retention and excision for rapid gene regulation in response to neuronal activity. Neuron *92*, 1266–1278.

Moal, I.H., Moretti, R., Baker, D., and Fernández-Recio, J. (2013). Scoring functions for protein-protein interactions. Curr. Opin. Struct. Biol. *23*, 862–867.

Molliex, A., Temirov, J., Lee, J., Coughlin, M., Kanagaraj, A.P., Kim, H.J., Mittag, T., and Taylor, J.P. (2015). Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. Cell *163*, 123–133.

Murakami, T., Qamar, S., Lin, J.Q., Schierle, G.S., Rees, E., Miyashita, A., Costa, A.R., Dodd, R.B., Chan, F.T., Michel, C.H., et al. (2015). ALS/FTD mutation-induced phase transition of FUS liquid droplets and reversible hydrogels into irreversible hydrogels impairs RNP granule function. Neuron *88*, 678–690.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40*, 1413–1415.

Philipps, D., Celotto, A.M., Wang, Q.Q., Tarng, R.S., and Graveley, B.R. (2003). Arginine/serine repeats are sufficient to constitute a splicing activation domain. Nucleic Acids Res. *31*, 6502–6508.

Qi, Y., Yu, J., Han, W., Fan, X., Qian, H., Wei, H., Tsai, Y.H., Zhao, J., Zhang, W., Liu, Q., et al. (2016). A splicing isoform of TEAD4 attenuates the Hippo-YAP signalling to inhibit tumour proliferation. Nat. Commun. *7*, ncomms11840.

Rosenberg, A.B., Patwardhan, R.P., Shendure, J., and Seelig, G. (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. Cell *163*, 698–711.

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

Shen, H., Kan, J.L., and Green, M.R. (2004). Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. Mol. Cell *13*, 367–376.

Song, X., Zeng, Z., Wei, H., and Wang, Z. (2018). Alternative splicing in cancers: From aberrant regulation to new therapeutics. Semin. Cell Dev. Biol. *75*, 13–22.

Sun, S., Zhang, Z., Fregoso, O., and Krainer, A.R. (2012). Mechanisms of activation and repression by the alternative splicing factors RBFOX1/2. RNA *18*, 274–283.

Tsai, Y.S., Gomez, S.M., and Wang, Z. (2014). Prevalent RNA recognition motif duplication in the human genome. RNA *20*, 702–712.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Wang, X., McLachlan, J., Zamore, P.D., and Hall, T.M. (2002). Modular recognition of RNA by a human pumilio-homology domain. Cell *110*, 501–512.

Wang, Y., Cheong, C.G., Hall, T.M., and Wang, Z. (2009). Engineering splicing factors with designed specificities. Nat. Methods *6*, 825–830.

Wang, Y., Ma, M., Xiao, X., and Wang, Z. (2012). Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. Nat. Struct. Mol. Biol. *19*, 1044–1052.

Wang, Y., and Wang, Z. (2014). Systematical identification of splicing regulatory *cis*-elements and cognate *trans*-factors. Methods *65*, 350–358.

Wang, Y., Xiao, X., Zhang, J., Choudhury, R., Robertson, A., Li, K., Ma, M., Burge, C.B., and Wang, Z. (2013). A complex network of factors with overlapping affinities represses splicing through intronic elements. Nat. Struct. Mol. Biol. *20*, 36–45.

Wang, Z., and Burge, C.B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA *14*, 802–813.

Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., and Burge, C.B. (2004). Systematic identification and analysis of exonic splicing silencers. Cell *119*, 831–845.

Wei, H., and Wang, Z. (2015). Engineering RNA-binding proteins with diverse activities. Wiley Interdiscip. Rev. RNA *6*, 597–613.

Wickham, H. (2009). Ggplot2: Elegant Graphics for Data Analysis (New York: Springer).

Ying, Y., Wang, X.J., Vuong, C.K., Lin, C.H., Damianov, A., and Black, D.L. (2017). Splicing activation by Rbfox requires self-aggregation through its tyrosine-rich domain. Cell *170*, 312–323 e310.

Yu, Y., Maroney, P.A., Denker, J.A., Zhang, X.H., Dybkov, O., Lührmann, R., Jankowsky, E., Chasin, L.A., and Nilsen, T.W. (2008). Dynamic regulation of alternative splicing by silencers that modulate 5' splice site competition. Cell *135*, 1224–1236.

Zhang, X.H., and Chasin, L.A. (2004). Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev. *18*, 1241–1250.

Zhu, J., Mayeda, A., and Krainer, A.R. (2001). Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. Mol. Cell *8*, 1351–1361.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Monoclonal ANTI-FLAG M2 antibody	Sigma-Aldrich	Cat# F1084 RRID: AB_259529
IRDye 800CW Goat anti-Mouse IgG	LI-COR	Cat# 926-32210 RRID: AB_2716640
Anti-mouse IgG antibody	Cell Signaling Technology	Cat# 7076 RRID: AB_330924
GAPDH antibody (FL-335)	Santa Cruz Biotechnology	Cat# sc-25778 RRID: AB_10167668
IRDye 680RD Goat anti-Rabbit IgG	LI-COR	Cat# 926-68071 RRID: AB_10956166
goat anti-rabbit IgG-HRP antibody	Santa Cruz Biotechnology	Cat# sc-2004 RRID: AB_631746
GAPDH (14C10) Rabbit mAb	Cell Signaling Technology	Cat# 2118S RRID: AB_561053
Anti-rabbit IgG, HRP-linked Antibody	Cell Signaling Technology	Cat# 7074S RRID: AB_2099233
HRP-Conjugated GAPDH Antibody	Proteintech Group	Cat# HRP-60004 RRID: AB_2737588
Bacterial and Virus Strains		
Escherichia coli DH5α (Cloning Strains)	New England Biolabs	C2988J
Escherichia coli DH5α (Cloning Strains)	Transgen	CD201-01
Critical Commercial Assays		
High Capacity cDNA Reverse Transcription Kit	Thermo Fisher Scientific	Cat# 4368813
Experimental Models: Cell Lines		
Human Embryonic Kidney cell line 293T (HEK 293T)	ATCC	Cat# CRL-3216 RRID: CVCL_0063
Oligonucleotides		
PCR Primers	GENEWIZ	see Table S1 for details
Recombinant DNA		
pGZ3-exonic site	Wang et al. (2009)	N/A
pZW2C-intronic site	Wang et al. (2012)	N/A
pZW4-exonic site	This paper	N/A
pGZ3-intronic site	This paper	N/A
pGL-None-PUF(6-2/7-2)	Wang et al. (2009)	N/A
pGL-SRSF7-RS-PUF(6-2/7-2)	Wang et al. (2009)	N/A
pGL-HNRNPA1-Gly-PUF(6-2/7-2)	Wang et al. (2009)	N/A
pGL-PUF(6-2/7-2)-SRSF7-RS	This paper	N/A
pGL-PUF(6-2/7-2)-HNRNPA1-Gly	This paper	N/A
pGL-HNRNP F-PUF(6-2/7-2)	This paper	N/A
See other proteins domains used in Table S1		N/A
Software and Algorithms		
R Studio	Open Source	RRID:SCR_000432
R Project for Statistical Computing	Open Source	RRID:SCR_001905
mixOmics (R package)	Lê Cao et al. (2011)	N/A
Python Programming Language	Open Source	RRID:SCR_008394
Biopython	Open Source	RRID:SCR_007173

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Image Lab	Bio-Rad	RRID:SCR_014210
ImageQuantTL	GE Healthcare	RRID:SCR_014246

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and request for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Zefeng Wang (wangzefeng@picb.ac.cn)

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell Culture

The human embryonic kidney cell line 293T (HEK 293T) was grown in Dulbecco's modified Eagle's medium (Thermo Fisher) supplemented with 10% fetal bovine serum (GE Healthcare) at 37°C and 5% CO2 and between 30% and 80% confluency. HEK293T cells used were from ATCC (ATCC Cat# CRL-3216, RRID: CVCL_0063). All human cells have oversight by CAS-MPG Partner Institute for Computational Biology (PICB).

METHOD DETAILS

Cell Transfection

For transfections, HEK 293T cells were seeded onto 24-well plates 1 day before transfection. For each transfection, 1.5 μL Lipofectamine 2000 (Invitrogen) and 0.6 μg plasmids (0.2 μg splicing reporters mixed with 0.4 μg ESF expression vectors) were mixed with 50μL Opti-MEM I Reduced Serum Media (GIBCO) in separate tubes. The solutions were then mixed together gently, incubated for 20 min at room temperature according to the manufacturer's instructions, and added to the wells for 24 hours at 37 °C and then harvested for the subsequent experiments, which are usually carried out 24 hours after transfection.

Plasmids Construction

To generate different ESF expression vectors, we first modified the expression construct for RS-PUF fusing protein by replacing the fragment encoding SR domain with a multi-clonal site that can be used to insert different functional domains (Wang et al., 2009). Specifically, we synthesized and annealed the oligonucleotides containing the *Mlu*I and *Eco*RI sites flanked by *Xho*I and *Bam*HI sites (Table S1), and the resulting dsDNA fragment was ligated into the RS-PUF expression vector digested with *Xho*I and *Bam*HI (Wang et al., 2009).

To generate expression constructs for ESFs with different functional domains, we amplified functional domains from human cDNAs (SuperScript III, Invitrogen) with primers containing a suitable pair of restriction sites (*Xho*I, *Mlu*I, *Eco*RI or *Bam*HI), and digested/ligated into the initial expression vector using the corresponding restriction endonucleases.

To test the effects of ESFs on exon skipping, we used modular reporter systems as described previously to assay for the inclusion of the cassette exon (Wang and Wang, 2014). To insert target sequence of PUF(6-2/7-2) domain into these splicing reporter vectors, pGZ3 and pZW2C, we synthesized and annealed oligonucleotides containing the target sequences flanked by *Xho*I and *Apa*I sites. The resulting double strand fragments were ligated with the digested reporter vectors (digested with *Xho*I and *Apa*I).

To validate the effect of distinct exon/introns, we further constructed two more reporters for the test. We inserted target sequence of PUF(6-2/7-2) domain into these splicing reporter vectors, pZW4 (exonic position, exon and introns were from mouse DHFR) and pGZ3 (intronic position, exon and introns are from and human IGF2BP1) which were used in Figure S1, and modified the splicing sites by PCR approach or primers annealing.

All plasmids were constructed in *E.coli* DH5α strain (#C2988J, New England Biolabs or # CD201-01, Transgen) and confirmed by Sanger sequencing, see Table S1 for more details.

RNA Extraction and Semi-Quantitative RT-PCR

Total RNA was isolated from transfected cells with RNeasy Mini Kit (QIAGEN), followed with treatment by DNase I (RNase-Free DNase Set, QIAGEN) according to the manufacturer's instructions. Total treated RNA was reverse transcribed by High Capacity cDNA Reverse Transcription Kit (Thermo Fisher Scientific) using random primer. A tenth of the RT production was used as the template for PCR amplification (24 cycles, labeled with trace amount of Cy5-dCTP) with primers corresponding to the first and third exon (i.e., GFP exons) of the reporter (Table S1). The resulting PCR productions were separated by 10% TBE PAGE gel, which were further scanned with a Typhoon TRIO Variable Mode Imager (GE Healthcare) and analyzed with ImageQuantTL software (GE Healthcare) as described previously (Wang and Wang, 2014).

Western Blotting

The expression of all ESFs were detected by Western blotting. The total cell pellets were completely lysed in 1×SDS-PAGE loading buffer (mixture of 2×Laemmli Sample Buffer (#1610737, Bio-Rad), RIPA buffer (#89900, Thermo Fisher) and Protease Inhibitor (#04693116001, Roche)) and heated at 99°C for 10 min. The samples were subsequently separated by SDS-PAGE gels and transferred to PVDF membranes. The following antibodies were then used for detection: Monoclonal ANTI-FLAG M2 antibody (F1084, Sigma-Aldrich), IRDye 800CW Goat anti-Mouse IgG (#926-32210, LI-COR), Anti-mouse IgG antibody (#7076, Cell Signaling Technology), GAPDH antibody (FL-335) (sc-25778, Santa Cruz Biotechnology), IRDye 680RD Goat anti-Rabbit IgG (#926-68071, LI-COR) and goat anti-rabbit IgG-HRP antibody (sc-2004, Santa Cruz Biotechnology), GAPDH (14C10) Rabbit mAb (#2118S, Cell Signaling Technology), Anti-rabbit IgG, HRP-linked Antibody (#7074S, Cell Signaling Technology), HRP-Conjugated GAPDH Antibody (#HRP-60004, Proteintech). The fluorescence labeling second antibodies were detected and visualized by Odyssey Imaging Systems (LI-COR) directly. The HRP conjugated primary or second antibodies were detected using an enhanced chemiluminescence detection kit and visualized by ChemiDoc Touch Imaging System (Image Lab, Bio-Rad).

Sparse Partial Least Squares Discriminant Analysis (sPLS-DA)

We tested several machine learning algorithms and finally chose a supervised approach, sparse partial least squares discriminant analysis, to classify the splicing factors according to their activities. The benefit of this method is that it can not only reduce dimensions, but also select important predictive features in the form of loading vectors. Due to the sparsity of the features, the algorithm adds l_1 penalty to the model by applying lasso penalization to select variables.

In sPLS-DA, we denote X as the n×p sample data matrix (i.e., predictor matrix), where n is the number of samples and p is the number of variables (i.e., total features). We used a dummy qualitative matrix as the response matrix Y (n×q) (q means the number of classes, here q=3), where each row in Y could be (1, 0, 0), (0, 1, 0) or (0, 0, 1) to represent inhibitor, neutral factor or activator. Partial least square algorism constructs a set of orthogonal components that maximize the sample covariance between the response and the linear combination of the predictor variables, which can be written as

$$arg \min_{u_h'u_h=1,\ v_h'v_h=1} cov^2\left(u_h'X, v_h'Y\right),$$

where u_h and v_h are the hth left and right singular vector of the singular value decomposition (SVD) of X^TY respectively in each dimension h.

Because of the data sparsity, the method performs variable selection on X data set to select more discriminative features. We set $M_h = X_h^T Y_h$ and applying l_1 penalization to obtain sparse loading vector u_h according to:

$$\min_{u_h,v_h} \|M_h - u_h v_h'\|_F^2 + P_\lambda(u_h),$$

where $P_\lambda(u_h)$ are the soft thresholding functions that approximate Lasso penalty functions.

In the stage of discriminant analysis, the PLS model is formulated as $Y = X\beta + E$, where β is the matrix of the regression coefficients and E is the residual matrix. The β matrix can be calculated by $\beta = W^*V^T$, where V is the matrix consist of loading vectors $(v_1, \ldots, v_H)$ (i.e., the right singular vectors from the SVD). The W^* matrix is given by a transformation: $W^* = W(U^TW)^{-1}$, where W is the matrix containing regression coefficients of the regression X on the latent variable $t_h = v_h'Y$, and U is the matrix containing the loading vectors $(u_1, \ldots, u_H)$.

In the predictive stage, we use the same regression coefficient matrix β to compute $Y_{val} = X_{val}\beta$, where the X_{val} is the predictor matrix of a validated sample and Y_{val} is the predicted class of each validated sample, which is assigned as the column index of the element with the maximal value in this row.

The sPLS-DA applied in our study is adopted from the R language package mixOmics (Lê Cao et al., 2011). We used the PLS-DA function to run the PLS-DA analysis and predict the unknown sequence with the different functions, and use the ten-fold cross validation to test the performance on our data. The score plot and loading plot are drawn with ggplot2 (Wickham, 2009). As a comparison, we also applied the Principal Component Analysis (PCA) algorithm to our experiment data (Figure S5 and Table S2).

Model Construction

Based on the experimental data from the training set, we classified all functional domains into three groups: splicing activators, splicing inhibitors, and neutral domains, which were defined by the relative fold change in exon inclusion at the threshold of ±0.15 in common logarithm (roughly equivalent to changes less than 70% for inhibitors or larger than 140% for activators).

Then we used the features of amino acid composition to represent each functional domain (as shown in Figure 4A), and assigned an arbitrary value to represent each activity class (-1 for inhibitors, 0 for neutral domains and 1 for activators). After extracting the features from tested domains, we excluded the features with extremely low frequency (<0.01) and the features that are almost zero in all samples, and finally used the sPLS-DA algorithm based on the sparsity of the training set data.

We constructed the predictive dataset of all RRM containing-RBPs from UniProt. The Biopython was used to process the fasta file from UniProt, and R Studio provide a statistical computing environment for R Project.

Random Peptides Generation

We generated random peptides (length, 60 amino acids) with equivalent frequency for each amino acid by Python Programming Language (code available upon request).

QUANTIFICATION AND STATISTICAL ANALYSIS

Image and Data Analysis Semi-Quantitative of RT-PCR

The RT-PCR productions were separated by 10% TBE PAGE gel, which were further scanned with a Typhoon TRIO Variable Mode Imager (GE Healthcare) and analyzed with Image Quant software (GE Healthcare) as described previously (Wang and Wang, 2014).

The PSI values were calculated with the intensity of the two bands. The fold changes of percent-spliced-in (PSI) of each sample was normalized to the control sample in which the same splicing reporter was co-expressed with PUF domain only. The n which represented the number of biological replicate was indicated in every figure legend.

DATA AND SOFTWARE AVAILABILITY

All data and software used in this manuscript are available upon request, for contact information see section 'Contact for Reagent and Resource Sharing'.