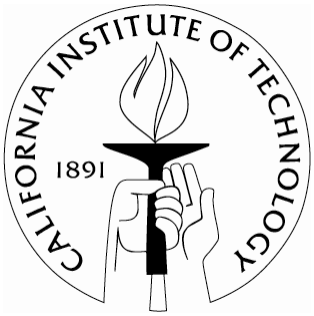
Reprogramming Alternative Splicing Using *Cis*-Acting Intronic Control Elements

Thesis by

Stephanie June Culler

In Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy



California Institute of Technology

Pasadena, California

2009

(Defended May 27, 2009)

Dedicated to Manigeh, Ron, and Bradley

Acknowledgements

The work described in this thesis represents a true ‘trial by fire’ and there are many people that deserve recognition. First and foremost, I would like to thank my advisor Prof. Christina Smolke for accepting me to be in her first group of graduate students. Christina has taught me the fundamentals of molecular biology and how to be critical at every step of a project. I feel very fortunate to have been afforded the opportunity to help start this lab from scratch, witness the evolution of our initial projects and to set an example for new students joining the lab. More importantly, I thank her for always being there to answer questions and to offer guidance. I thank her for giving me the independence to develop all of my projects and for her constant support. Christina has truly been a wonderful mentor and has set a very high bar of how to write, communicate and manage every aspect of science.

I would very much like to thank Prof. Dave Tirrell, Prof. Peter Dervan and Prof. Ellen Rothenberg for serving as my thesis committee. I appreciate their constant encouragement and wisdom that they have graciously offered throughout this journey. Prof. Tirrell, I thank you for the guidance and mentorship you have provided. I would like to also acknowledge Prof. Asthagiri for allowing us to utilize his TC facility and Prof. Niles Pierce for sharing his radioactive room with us after the lab moved. Their generosity has been much appreciated. Special recognition goes to Shelly Diamond and Diana Perez of the flow cytometry facility for helping with over 200+ hours of sorting. Their expertise in flow cytometry and guidance really played a key role in finishing the ISRE project. I appreciate their willingness to continue sorts into the late hours of the night.

I have had the fortunate experience of collaborating with a truly outstanding bioinformaticist, Rodger Voelker. He has played a significant role in all of the bioinformatic analyses performed for the ISRE project and has provided much guidance. My bioinformatic knowledge has been significantly increased because of his wisdom. I would also like to thank Prof. Gene Yeo as well, for assistance in answering bioinformatic related questions and for providing a critical review of the ISRE manuscript.

Being in the first group of graduate students in Prof. Smolke's lab has been a very special experience and I have many of my co-workers to thank! I would like to first thank Dr. Kristy Hawkins for being my rock throughout this process and for being there for me during the best and the most difficult points of this journey. She is an amazing scientist and her work ethic is both inspiring and unparalleled. Thank you for always inspiring me to be a better scientist and for of course, sharing your body building world with me! To Dr. Kevin G. Hoff, thank you for being my second mentor and for always lending an extremely critical eye on my work as well as for everyone else's in the lab. Your mentorship skills and amazing knowledge of Biology will give you tons of success when you have your own lab. Dr. Travis Bayer, thank you for making science fun and for Mexico City and the Rado! Dr. Chase Beisel, I thank you for being generous with your knowledge and your willingness to always provide a helping hand with experiments. I will miss you and Amy! Special recognition goes out to other members of the Smolke lab; Yvonne Chen, Joe Liang, Mike Siddiqui, Drew Kennedy, Josh Michener, Katie Galloway, Jay Vowles and Midori Greenwood-Goodwin. Dr. Mike D and Dr. Cara, thanks for sharing your lab with us and for being really great people to be around.

I would like to thank my friends for always supporting me and allowing me to share my personal triumphs and failures with them. I would like to thank Dr. Seema Patel, Dr. Sophia Chen and Delfa Mejia, for always being there for me and for their friendship. I don't know if I would have been able to finish without their support! Thanks to Dr. Ash Swaminathan as well. To Jenny and Dr. Tristan (my third roomie), I have enjoyed going through this experience with both of you and will miss our romantic comedy nights. Jenny your work ethic provides daily inspiration for me and I thank for you being a wonderful roommate. I would also like to thank the PCC orchestra for their support of my scientific endeavors, for the opportunity to share my music with the community and for having the opportunity to work with professional musicians from the LA Phil!

Finally, I am very grateful to my parents, my brother and my family for their love and constant support. My parents have always encouraged me throughout every endeavor and they always know how to cheer me up when times are tough.

Abstract

Alternative splicing is a process by which multiple protein isoforms are generated from a single coding region by altering the ways in which exons are joined together. This pathway is used by cellular systems to both increase proteomic diversity from a limited number of genes and to precisely control gene expression. Bioinformatics and comparative genomics approaches have provided significant sequence and functional insight into the regulatory sequences that occur within exonic regions of a transcript. *In vitro* and *in vivo* strategies have also been developed to screen for exonic splicing enhancers and silencers (ESEs and ESSs) from small, randomized libraries. Much less is known about intronic splicing enhancers and silencers (ISEs and ISSs), although recent bioinformatics approaches are beginning to shed some light on these regulatory sequences. A thorough understanding of both exonic and intronic regulators is necessary to enable the programming of alternative splicing patterns, which will provide a powerful tool for interrogating and manipulating cellular function.

We developed a generalizable *in vivo* screening strategy for generating intronic splicing regulatory elements (ISREs). Our high-throughput approach employed a systematic screening strategy with extensive genome-wide bioinformatic analyses and experimental characterization which included a small-scale RNAi screen. Using this approach, we identified ISRE consensus motifs, characterized the splicing regulatory networks (SRNs) associated with these regulatory elements and generated a model for ISRE regulatory function. Highlighting the complexity of SRNs, we found that *cis*-acting intronic regulatory sequences function through combinatorial effects from multiple elements and trans-acting factors, and that the immediate transcript context has a

dominant effect on ISRE function. Overall, this screening strategy provides a general method for generating regulatory sequences of alternative splicing events, which provide powerful tools for gene expression control.

We next extended from our studies on cellular screening strategies for generating splicing regulatory elements, to build novel platforms that support the construction of protein-responsive alternative splicing control elements. Protein-binding RNA aptamers were inserted into key intronic locations of an alternatively spliced transcript to enable the detection of intracellular protein concentrations and to translate detection events to the regulation of alternative splicing patterns and thus gene expression. We demonstrate that these RNA elements can serve as autonomous control devices by linking endogenous nuclear protein levels to gene expression events and external stimuli to complex cellular phenotypes. These synthetic alternative splicing regulators can be implemented combinatorially to regulate alternative splicing patterns in response to multiple inputs. In addition, we applied these synthetic regulators to the rewiring of endogenous signal transduction pathways and building of novel regulatory networks for user-defined phenotypes. Our work provides an early example of a novel class of RNA-based "intelligent" therapeutics by directing increased signaling through pathways associated with disease to the triggering of apoptosis. These programmable sensing-actuation molecules will be broadly applied in health and medicine towards the early diagnosis and treatment of disease.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Chapter I. Introduction

1.1. Post-transcriptional processing plays a significant role in regulating gene expression

Proper control of gene expression is essential for normal cellular function, development and reproduction. Eukaryotic gene expression is regulated at both the transcriptional and post-transcriptional level. Post-transcriptional regulation arises through cellular control of RNA processing, splicing, export, localization, and turn-over[1], and allows cells greater control over gene expression patterns[2]. mRNA splicing, export, and turn-over are tightly coupled processes. Alternative splicing creates proteomic diversity and controls protein isoform levels by regulating the patterns in which exons are assembled. Cells can also control gene expression by varying mRNA turnover rates for specific transcripts through nonsense-mediated decay (NMD), a cellular surveillance system in which mRNAs containing premature termination codons are selectively degraded. Recent studies indicate that alternative splicing and NMD are often coupled such that gene expression can be controlled in both a spatial and temporal manner[3].

1.2. Pre-mRNA splicing

Mature eukaryotic mRNAs are formed from precursor mRNAs (pre-mRNAs) that contain introns and exons. The intron sequences are removed and the exon sequences are ligated together to form a mature mRNA in a process called splicing (Figure 1.1a). A typical human gene contains relatively short exons (50–250 base pairs (bp)) which are

separated by significantly longer introns (hundreds to thousands of bp)[4]. Splicing requires exon recognition followed by accurate cleavage at exon-intron boundaries which are determined by the nearly invariant GU and AG intronic dinucleotides at the 5' and 3' exon-intron junctions, the polypyrimidine tract $(Y)_n$ and the A residue (adenosine) that serves as the branch point[5,6]. Trans-acting factors are recruited to assemble across splice sites, forming a catalytically active complex known as the spliceosome which is responsible for the excision of introns and recombination of exons[7]. The spliceosome is composed of five small nuclear ribonucleoprotein particles (snRNPs) and over 100 protein factors[8,9] and utilizes RNA–RNA, RNA–protein, and protein–protein interactions to correctly excise introns and to splice exons[10].

The components of the spliceosome coordinate the stepwise associations, dissociations and conformational changes of the pre-mRNA, snRNPs, and protein complexes necessary for splicing and transcript release to occur[9]. As seen in Figure 1.1b, intron removal takes place in two S_N2-type transesterification reactions. In the first step, the 5' splice site is attacked by the 2'-hydroxyl of the branch site adenosine, generating the exon 1 intermediate and a branched intron 'lariat' attached to exon 2. In the second step, the 3' splice site is attacked by the 3'-hydroxyl of the free exon 1 intermediate. The final products are the spliced mRNA and the excised intron in lariat form [11].

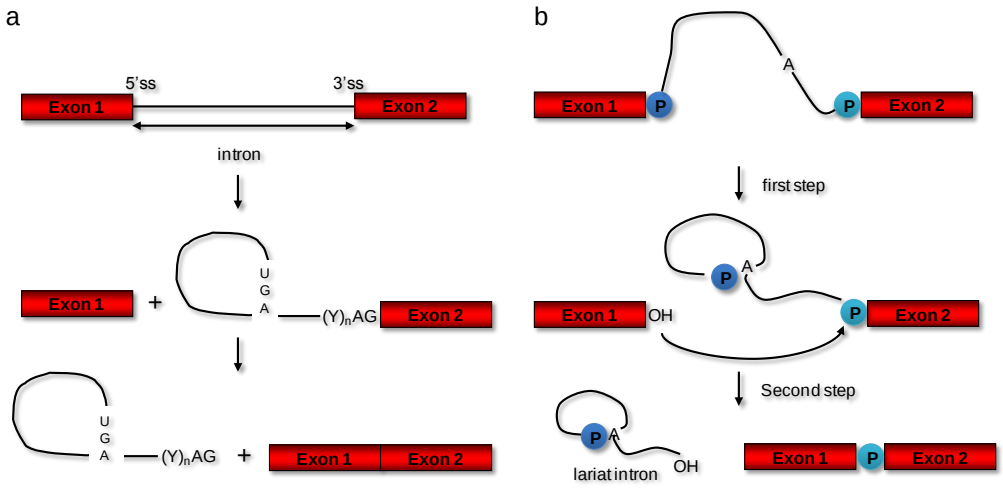


Figure 1.1. Mechanism of splicing. (a) Consensus mammalian 5' splice site (5' ss), branch point (A), polypyrimidine tract (Y_n), and 3' splice site (3' ss) sequences are shown. Splicing takes place in two transesterification steps. The first step results in two reaction intermediates: the detached 5' exon and an intron 3'– exon fragment in a lariat structure. The second step ligates the two reactions and releases the intron lariat. (b) Two-step transesterification pathway of pre-mRNA splicing. First, the phosphodiester bond at the 5' ss is attacked by the 2'-hydroxyl of an adenosine at the branch point, which generates a free 5' exon and an intron lariat-3' exon. Subsequently, the 3'-hydroxyl of the 5' exon attacks the phosphodiester bond at the 3' ss, leading to exon ligation and an intron lariat-3' exon.

1.3. Basic principles of alternative splicing

Alternative splicing is a process by which multiple protein isoforms are generated from a single coding region by altering the ways in which the exons are joined together,

or defined, during the splicing process[12]. Alternative splicing allows for the enrichment of the transcriptome and proteome of higher organisms without the need for genome expansion[9]. The potential of alternative splicing to generate more protein isoforms from a single gene than the number of genes in an entire organism helps to explain the discrepancy between the low number of human protein-coding genes (~26,000) and the number of human proteins, which is estimated to be more than 90,000[13,14]. Recent studies using high-throughput deep sequencing suggest that the extent of alternative splicing is significantly greater in humans than was previously estimated with approximately 92–94% of human multi-exon genes being alternatively spliced[15,16]. While these studies have provided catalogues of splicing events, further characterization is needed to determine if these spliced isoforms have a particular function[17,18]. A current challenge in the field of splicing is the development of high-throughput cell-based assays to evaluate the function of these spliced variants[17].

Alternative splicing occurs through various modes including exon skipping/inclusion, alternative 3' ss, alternative 5' ss, mutually exclusive exons, intron retention, and alternative initial/terminal exons (Figure 1.2)[5,11,19]. Consensus motifs contain only about half of the information required for accurate recognition of exons and introns in human transcripts[20]. It has recently been discovered that human introns contain many sequences that resemble consensus splice site sequences, or 'decoy' splice sites, that are rarely recognized by the splicing machinery[21]. At the level of exon definition, pseudoexons, which are composed of these decoy signals, outnumber real exons by an order of magnitude[22,23]. Despite the potential for errors, the splicing process occurs with high fidelity, implying that there are additional sequences besides canonical elements that

play a role in the splicing of most transcripts. These elements include specific *cis*-acting elements in exons and introns, which aid in splice site recognition and act as enhancers or repressors of splicing. Within the spliceosome, splicing is carried out by the interplay of these *cis*-acting sequences and trans-acting factors that modulate them, leading to a "splicing code"[24].

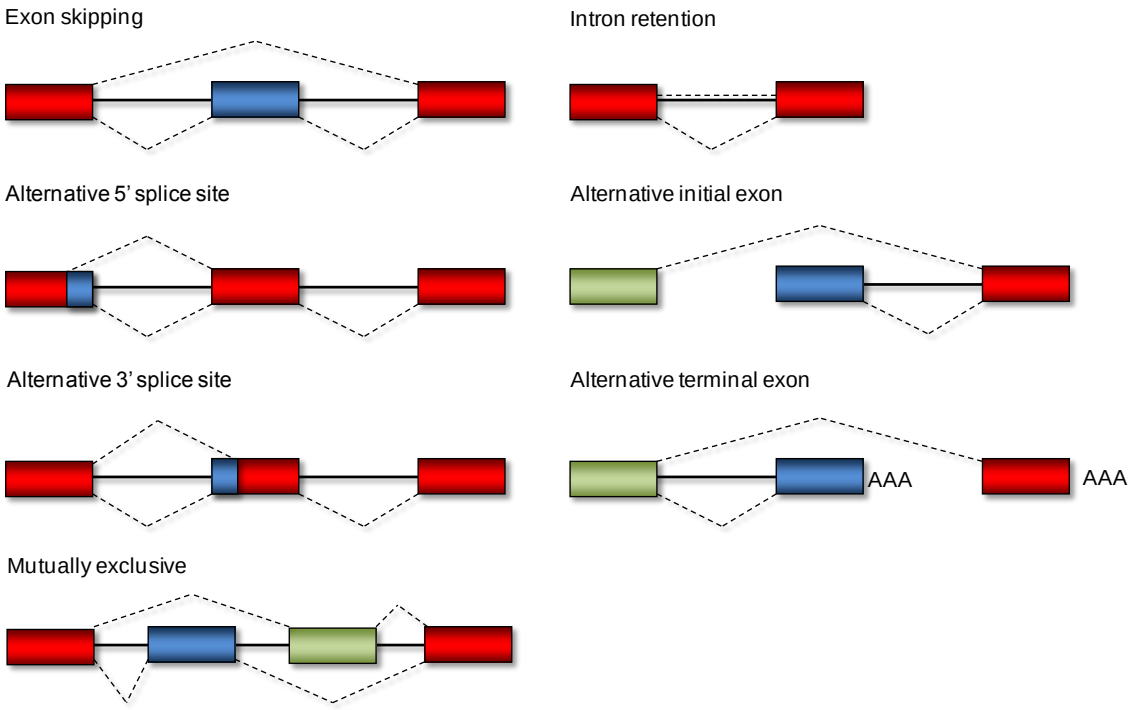


Figure 1.2. Major forms of alternative splicing. These modes are responsible for the generation of functionally distinct transcripts. Labels: introns are represented by solid lines and dashed lines indicate splicing activities.

1.4. Elements of a "splicing code"

The auxiliary elements that make up the splicing code include exonic and intronic splicing enhancers (ESEs and ISEs, respectively), which aid in exon recognition, and exonic and intronic splicing silencers (ESSs and ISSs, respectively), which suppress exon inclusion (Figure 1.3)[11,19]. These splicing regulatory elements (SREs) generally function by recruiting trans-acting splicing factors that activate or suppress splice site recognition or spliceosome assembly[19]. Mutating, removing, and/or shifting the location of these sequences will affect the overall splicing pattern of a transcript[25]. In contrast to the canonical splice sites whose sequence and position are well characterized, the complex code formed by these auxiliary SREs is only partially understood. In addition, the regulatory proteins that interact with these specific sequences to either stimulate or repress exon recognition have been only partially elucidated. Therefore the generation of a complete "splicing code" will require the elucidation of all types of SREs and the corresponding trans-acting factors that regulate them.

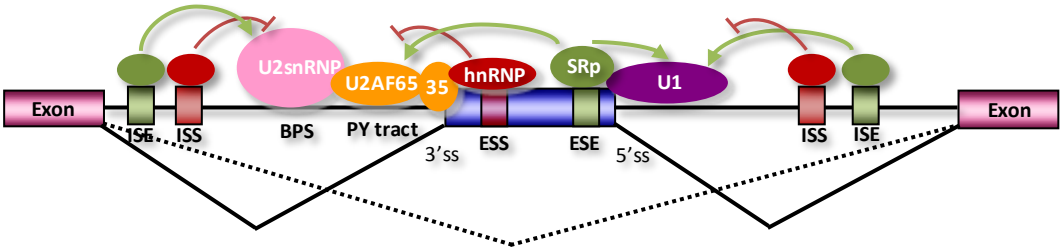


Figure 1.3. Elements of a "splicing code". The diagram illustrates regulated splicing. Green arrows illustrate the positive activity of splicing enhancers (green bars) on the selection of adjacent splice sites in the alternative exon (blue bar). Red arrows with flat bars indicate the negative activity of splicing silencers (red bars) on the regulation of

adjacent splice sites. Splicing is regulated by *cis*-elements (ESEs, ESSs, ISSs, and ISEs) and the trans-acting factors that bind them (SR proteins, hnRNP, and unknown factors).

The most well characterized family of regulatory proteins that bind both intronic and exonic *cis*-elements are the serine/arginine-rich (SR) and SR-like proteins. The human SR family contains ten identified members[26] that are thought to mediate interactions between splicing factors bound to the 5' and 3' splice sites. All SR proteins have a modular structure and contain either one or two copies of an RNA-recognition motif (RRM) and a C-terminal end highly enriched in arginine and serine dipeptides (RS domain)[27]. The RRMs serve to mediate sequence-specific binding to the RNA, which determines substrate specificity, and the RS domains are involved in protein-protein interactions that are essential for the recruitment of the splicing machinery[5]. Most exonic splicing enhancers (ESEs) are purine rich and different ESEs are recognized by various subsets of SR proteins[8]. Additionally, SR protein functions have been extended to mRNA export[28], mRNA stability[29], protein translation[30], and nuclear export[31].

The heterogeneous nuclear ribonucleoproteins (hnRNPs) are a diverse class of proteins that bind to both exonic splicing silencers (ESSs) and intronic splicing silencers (ISSs) [12]. Similar to SR proteins, hnRNPs have a modular structure consisting of one or more RRMs that may be involved with RNA binding and domains that are presumed to mediate interactions with other proteins. Amongst the best characterized members of this family are hnRNP A1[32] and PTB (hnRNP I)[33,34]. This family of proteins function by a variety of mechanisms and often serve as antagonists to SR proteins through competitive

binding to the transcript[35]. These factors can often block essential interactions between spliceosome components to inhibit splicing[36,37].

Overexpression of SR proteins and hnRNP proteins has been shown to affect the splicing patterns of alternatively spliced pre-mRNAs *in vivo*[38]. Most hnRNP and SR proteins shuttle continuously between the nucleus and cytoplasm and consequently their subcellular distribution can shift in response to stress signals[9]. Recent studies have also demonstrated that the relative amounts of SR and hnRNP A/B proteins are important in regulating patterns of alternative splicing in a tissue-specific and developmentally-regulated manner[39] . The expression of these proteins is unique to each cell type and thus the regulation of expression and activity of these proteins is critical for normal alternative splicing and cellular function. Moreover, a diverse set of diseases are associated with changes in expression of trans-acting splicing factors[40]. Disease-related changes in splicing factors are potentially useful biomarkers for disease diagnosis and classification. It has been suggested that the modulation of the relative stoichiometries of splicing factors can be used to regulate the alternative splicing of disease-relevant mRNAs[9]. As such, high-throughput small molecule screens are currently being developed to select for agents that modulate splicing factor ratios towards targeted therapies[41].

1.5. Nonsense-mediated decay is a surveillance pathway in eukaryotes

Studies have indicated that approximately a third of alternative mRNA isoforms are targets for NMD, an mRNA surveillance system that targets aberrant transcripts[3,42].

The NMD pathway targets aberrant transcripts and removes potentially harmful truncated versions of proteins[43]. A recent investigation of the human genome sequence and databases of expressed sequence tags (ESTs) identified a coupling between NMD and alternative splicing[3]. NMD substrates from aberrant splicing processes include transcripts with retained introns, skipped exons and extended 5' or 3' UTRs[44,45]. NMD specifically targets transcripts containing premature termination codons (PTCs) that can be introduced into transcripts as a result of DNA rearrangements, frameshifts, nonsense mutations or errors during transcription and splicing[45,46]. In higher eukaryotes, PTCs are recognized when they occur 50–55 nucleotides (nt) upstream of an exon-exon junction (Figure 1.4)[47]. The current model of the NMD mechanism is that during pre-mRNA processing the spliceosome deposits exon junction complexes (EJCs) at sites of intron removal. During the first round of translation the ribosome displaces the EJCs in its path and then dissociates from the mRNA at the stop codon[45]. However, if a PTC is present the ribosome will stop and will fail to remove any EJCs downstream from the PTC. Interactions between EJC proteins and several release factors[48] trigger mRNA decay through a deadenylation-dependent pathway[49].

NMD plays a significant role in human diseases and inherited cancers[9]. It has been suggested that 25% of all mutations causing genetic disorders and cancers target corresponding mRNAs to NMD[50]. The ability of nonsense transcripts to be targeted or to evade NMD has an effect on the genotype-phenotype results of these mutations[43,51]. Gain-of-function mutations can occur when nonsense transcripts evade NMD and result in the production of truncated proteins that are potentially harmful to cells. Whereas, loss-of function mutations creating PTCs targeted to NMD can modify the severity of the

disease phenotypes. Therefore, novel therapeutic approaches targeting NMD may be able to modify diseases with similar genetic phenotypes[6]. Therapeutic strategies targeting the coupling of NMD with alternative splicing may be also used to fine tune levels of specific trans-factors including splicing factors and other RNA binding proteins[24].

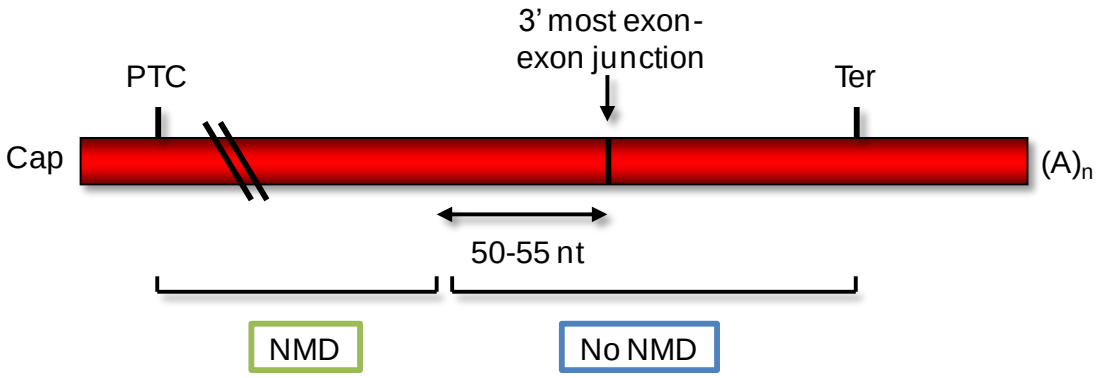


Figure 1.4. NMD and the position of the exon-exon junction. Only the 3'-most exon-exon junction within a generic mammalian mRNA is shown. A PTC that is located in the region indicated in green, which is followed by an exon-exon junction more than 50–55 nt downstream, elicits NMD, whereas a PTC that is located in the region indicated in blue fails to elicit NMD. The normal termination codon (Ter) usually resides in the 3'-most exon.

1.6. Alternative splicing and human disease

The inaccurate recognition of exon-intron boundaries or the failure to remove an intron produces aberrant mRNAs that are either degraded or encode for defective protein isoforms. Studies have suggested that approximately 50% of disease-causing mutations

are found to affect splicing[43]. Mutations affect splicing by disrupting canonical splice sites and auxiliary elements, by creating cryptic splice sites or by altering RNA secondary structure[52]. Mutations that alter trans-acting factors may result in global splicing defects which can have very specific phenotypic outputs[53]. Therefore, understanding the potential effects of single-nucleotide mutations that alter pre-mRNA splicing will enable researchers to develop new therapeutics and treatments that target certain genetic diseases and a variety of cancers[46,54]. Moreover, further elucidation of the *cis*-acting elements that regulate alternative splicing is needed to determine the extent of which single nucleotide polymorphisms (SNPs) modulate splicing through these elements.

1.7. Trans- and *cis*-acting technologies that alter pre-mRNA splicing

Several therapeutic approaches have been developed to alter the splicing patterns of target genes or specific splice variants. Antisense oligonucleotides (AOs) have been used extensively to target alternative spliced transcripts in order to correct disease causing splicing defects. AOs are designed to hybridize and block one or more sequences in the target pre-mRNA that are essential to the splicing event. Several systems based on this general strategy have been used to alter the splice site selection of transcripts involved in disease including breast cancer 1 (*BRCA1*), *SMN*, *β-globin*, the *CFTR* gene in cystic fibrosis, and apoptosis regulator (Bcl-x)[9,29,55]. An antisense technology was recently developed that mimics the functions of SR proteins to restore wild-type splicing in *BRCA1* and *SMN2* pre-mRNA transcripts[46]. This approach, termed ESSENCE (Exon-Specific Splicing ENhancement by small Chimeric Effectors), utilized a peptide-nucleic

acid (PNA) hybrid containing RS dipeptide repeats coupled with an antisense oligonucleotide targeting a mutated ESE in the *SMN2* gene to correct splice site selection [56]. ESSENCE was also demonstrated on the *BRCA1* gene and was successful at suppressing the effects of a mutation on exon skipping *in vitro*.

SMN2 pre-mRNA transcripts have also been targeted by another AO approach which uses a tailed bi-functional antisense oligonucleotide[57]. In this approach, the antisense portion of the oligonucleotide targets the molecule to exon 7 in *SMN2* and the tailed portion contains an ESE element such that splicing factors will be recruited to influence splice site selection. The latter technology was tested successfully *in vitro* and *in vivo*[57,58]. RNAi approaches may also be used to eliminate aberrantly spliced mRNAs by targeting specific isoforms[43]. However, the applications of these and other antisense technologies are limited because of off-target effects, toxicity, efficiency, and issues with delivery.

Alternative splicing is a viable target for pharmacological modulation with small molecules[9]. High-throughput screening strategies have been developed and employed to select for small molecule inhibitors of SR proteins[59], SR protein kinases (SRPKs)[60] and Cdc2-like kinases (CLKs)[61]. These compounds were shown to modulate splicing, although for only a few splicing events. Further large-scale screens are needed to find more potent and specific modulators of alternative splicing which may be used as a general therapeutic approach for treating diseases[9]. In addition, small molecule based therapies have had success as pharmacological agents because they circumvent some of the major issues of delivery that has been encountered by nucleic acids[9]. While recent

studies have employed cell-based selection strategies, these screens were employed on a modest scale and have only examined the splicing patterns of a limited set of genes. The potential of small molecule based strategies targeting alternative splicing can be determined with the development of more robust, quantitative and specific cell-based splicing assays[9].

In addition to trans-acting technologies, *cis*-acting RNA-based regulatory systems have been developed towards the regulation of pre-mRNA splicing in yeast and mammalian cells[62]. Regulation of splicing in these systems is exerted by small molecule responsive aptamers which were inserted into the intronic regions of spliced transcripts. The aptamer sensing components used in these platforms were shown to not be modular and consequently these systems have not been extended to sense other biomolecules. The adaptation of these platforms towards therapeutic applications have also been hampered by the lack of existing aptamers with suitable pharmacological properties, such as limited cell toxicity[62]. The extension of these designs towards applications in health and medicine may be realized with the identification of new aptamers and with the modification of these platforms to be responsive to protein biomarkers.

1.8. *Cis*-acting regulators of alternative splicing

Much effort has been directed toward the characterization of *cis*-acting exonic regulatory sequences. Specifically, *in vitro* and *in vivo* strategies have been implemented to screen for ESEs and ESSs from small randomized libraries[63–67] and within genome

sequence data[68–70]. Identified ESSs are able to control the selection of alternative 5' and 3' ss when placed between competing sites[71]. Results from these *in vivo* studies demonstrate that not all RNA sequences that have been selected against SR proteins are splicing enhancers, underlining the importance of functional screens. Some of the selected exonic splicing regulators displayed significant similarities to naturally occurring regulators, whereas others were novel. From these and other results, it has been suggested that most exons are likely to have multiple regulators (ESEs and ESSs) that act as weak splicing signals which have an additive effect on splicing[26,64]. ESEs and ESSs also play critical roles in directing splicing to consensus splice sites rather than decoy sites[72-74]. However, few exonic regulators may be strong enough on their own to regulate splicing individually[65].

In contrast, fewer ISEs and ISSs have been characterized and little is known regarding the auxiliary factors by which they are bound. These elements are generally short, variable in sequence, individually weak and present in multiple copies[38]. Several common intronic regulator motifs have been observed, including the GGG triplet[75,76], purine-rich motifs[10] or polypyrimidine tracts present in the 3' intronic regions[10]; however, most newly found auxiliary elements tend to be quite degenerate[77]. ISSs and ISEs have been identified near alternatively spliced exons and their mechanistic actions appear to be antagonistic[78]. ISSs may inhibit exon inclusion by recruiting splicing repressors, which directly antagonize splicing factor binding, or by recruiting repressors to multiple binding sites resulting in a 'zone of silencing'[19]. Identified ISSs are variable in sequence and recruit members of the serine/arginine-rich (SR) or the heterogeneous nuclear ribonucleoprotein (hnRNP) protein families[11]. Similar to exonic regulatory elements,

ISSs have been shown to inhibit the inclusion of pseudoexons into mature mRNAs[79]. Several ISEs have been characterized[80], however the nuclear factors that regulate alternative splicing through these sequences have not been elucidated[81]. Further adding to these complexities, are observations that some intronic splicing regulatory elements (ISREs) act as inhibitors upstream of a specific splice site and as enhancers downstream of that splice site[81]. These results support a role for intronic elements in regulating splicing patterns in a combinatorial manner [8]. Despite the widespread importance of ISREs, a systematic experimental characterization or iterative functional screen has yet to be applied towards developing a functional definition of these elements.

1.9. Engineering *cis*-acting intronic regulators of alternative splicing

Cis-acting regulators of alternative splicing play key roles in regulating the form and function of protein isoforms produced from a given gene in response to various signals received by the cell. Therefore, the ability to program alternative splicing patterns will provide a powerful tool to interrogate and manipulate cellular function. Despite the critical role of alternative splicing in creating phenotypic complexity and regulating gene expression, the sequence composition and function of ISREs have not been well elucidated. As a result of this lack of knowledge around how alternative splicing events are encoded at a genetic level, researchers very rarely incorporate control of alternative splicing events in synthetic genetic networks.

My thesis work has centered around the development of strategies for generating synthetic regulators of alternative splicing and for implementing genetic circuits based on the control of alternative splicing events. Chapter II describes the development of a high-

throughput *in vivo* screen for ISRE function. This high-throughput approach combined a systematic screening strategy with extensive genome-wide bioinformatic analyses and experimental characterization. The implementation of this strategy yielded insight into the sequence composition of ISREs, the splicing regulatory networks (SRNs) associated with these sequences and the mechanisms in which they achieve regulation. Chapter III describes an extension of this platform to support the construction of protein-responsive alternative splicing regulatory elements based on the integration of protein-binding RNA aptamers into key intronic locations of a target alternatively spliced transcript. This protein-responsive platform was adapted to detect disease biomarkers, reprogram natural signaling pathways, and control biologically-relevant processes, such as apoptosis, in response to increased signaling through pathways associated with disease. The molecular platforms described in Chapters 2 and 3 represent powerful tools to regulate alternative splicing events and thus gene expression. In addition, the ability to reprogram biological function in response to endogenous protein levels has broad applications in health and medicine, where such molecular tools can provide the basis for the design of targeted "intelligent" therapeutics. Chapter 4 provides a perspective on the general applications of such genetically encoded technologies and future work needed to further characterize these synthetic regulatory systems.

References

1. Maquat, L.E. & Carmichael, G.G. Quality control of mRNA function. *Cell* **104,** 173–176 (2001).

2. Maniatis, T. & Reed, R. An extensive network of coupling among gene expression machines. *Nature* **416,** 499–506 (2002).

3. Lewis, B.P., Green, R.E. & Brenner, S.E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A* **100,** 189–192 (2003).

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

5. Cartegni, L., Chew, S.L. & Krainer, A.R. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* **3,** 285–298 (2002).

6. Nissim-Rafinia, M. & Kerem, B. Splicing regulation as a potential genetic modifier. *Trends Genet* **18,** 123–127 (2002).

7. Goldstrohm, A.C., Greenleaf, A.L. & Garcia-Blanco, M.A. Co-transcriptional splicing of pre-messenger RNAs: considerations for the mechanism of alternative splicing. *Gene* **277,** 31–47 (2001).

8. Smith, C.W. & Valcarcel, J. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem Sci* **25,** 381–388 (2000).

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

10. Hastings, M.L. & Krainer, A.R. Pre-mRNA splicing in the new millennium. *Curr Opin Cell Biol* **13,** 302–309 (2001).

11. Black, D.L. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72,** 291–336 (2003).

12. Maniatis, T. & Tasic, B. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418,** 236–243 (2002).

13. Modrek, B., Resch, A., Grasso, C. & Lee, C. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res* **29,** 2850–2859 (2001).

14. Woodley, L. & Valcarcel, J. Regulation of alternative pre-mRNA splicing. *Brief Funct Genomic Proteomic* **1,** 266–277 (2002).

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

16. Wang, E.T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476 (2008).

17. Ledford, H. Human genes are multitaskers. *Nature* **456,** 9 (2008).

18. Multitudes of messages. *Nat Genet* **40,** 1385 (2008).

19. Matlin, A.J., Clark, F. & Smith, C.W. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol* **6,** 386–398 (2005).

20. Lim, L.P. & Burge, C.B. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A* **98,** 11193–11198 (2001).

21. Cote, J., Dupuis, S. & Wu, J.Y. Polypyrimidine track-binding protein binding downstream of caspase-2 alternative exon 9 represses its inclusion. *J Biol Chem* **276,** 8535–8543 (2001).

22. Zhang, X.H., Heller, K.A., Hefter, I., Leslie, C.S. & Chasin, L.A. Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Res* **13,** 2637–2650 (2003).

23. Sun, H. & Chasin, L.A. Multiple splicing defects in an intronic false exon. *Mol Cell Biol* **20,** 6414–6425 (2000).

24. Blencowe, B.J. Alternative splicing: new insights from global analyses. *Cell* **126,** 37–47 (2006).

25. Graveley, B.R., Hertel, K.J. & Maniatis, T. A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. *Embo J* **17,** 6747–6756 (1998).

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

27. Chandler, S.D., Mayeda, A., Yeakley, J.M., Krainer, A.R. & Fu, X.D. RNA splicing specificity determined by the coordinated action of RNA recognition motifs in SR proteins. *Proc Natl Acad Sci U S A* **94,** 3596–3601 (1997).

28. Maatta, A.M. *et al.* Non-small cell lung cancer as a target disease for herpes simplex type 1 thymidine kinase-ganciclovir gene therapy. *Int J Oncol* **24,** 943–949 (2004).

29. Villemaire, J., Dion, I., Elela, S.A. & Chabot, B. Reprogramming alternative pre-messenger RNA splicing through the use of protein-binding antisense oligonucleotides. *J Biol Chem* **278,** 50031–50039 (2003).

30. Sanford, J.R., Gray, N.K., Beckmann, K. & Caceres, J.F. A novel role for shuttling SR proteins in mRNA translation. *Genes Dev* **18,** 755–768 (2004).

31. Caceres, J.F., Misteli, T., Screaton, G.R., Spector, D.L. & Krainer, A.R. Role of the modular domains of SR proteins in subnuclear localization and alternative splicing specificity. *J Cell Biol* **138,** 225–238 (1997).

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

33. Valcarcel, J. & Gebauer, F. Post-transcriptional regulation: the dawn of PTB. *Curr Biol* **7,** R705–708 (1997).

34. Wagner, E.J. & Garcia-Blanco, M.A. Polypyrimidine tract binding protein antagonizes exon definition. *Mol Cell Biol* **21,** 3281–3288 (2001).

35. Caceres, J.F. & Kornblihtt, A.R. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet* **18,** 186–193 (2002).

36. Izquierdo, J.M. *et al.* Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Mol Cell* **19,** 475–484 (2005).

37. Sharma, S., Falick, A.M. & Black, D.L. Polypyrimidine Tract Binding Protein Blocks the 5' Splice Site-Dependent Assembly of U2AF and the Prespliceosomal E Complex. *Mol Cell* **19,** 485–496 (2005).

38. Ladd, A.N. & Cooper, T.A. Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol* **3,** reviews0008 (2002).

39. Hanamura, A., Caceres, J.F., Mayeda, A., Franza, B.R., Jr. & Krainer, A.R. Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors. *Rna* **4,** 430–444 (1998).

40. Ule, J. Ribonucleoprotein complexes in neurologic diseases. *Current opinion in neurobiology* **18,** 516–523 (2008).

41. Stoilov, P., Lin, C.H., Damoiseaux, R., Nikolic, J. & Black, D.L. A high-throughput screening strategy identifies cardiotonic steroids as alternative splicing modulators. *Proc Natl Acad Sci U S A* **105,** 11218–11223 (2008).

42. Green, R.E. *et al.* Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. *Bioinformatics* **19 Suppl 1,** i118–121 (2003).

43. Wang, G.S. & Cooper, T.A. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8,** 749–761 (2007).

44. Hilleren, P. & Parker, R. mRNA surveillance in eukaryotes: kinetic proofreading of proper translation termination as assessed by mRNP domain organization? *Rna* **5,** 711–719 (1999).

45. Chang, Y.F., Imam, J.S. & Wilkinson, M.F. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* **76,** 51–74 (2007).

46. Cartegni, L. & Krainer, A.R. Correction of disease-associated exon skipping by synthetic exon-specific activators. *Nat Struct Biol* **10,** 120–125 (2003).

47. Maquat, L.E. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat Rev Mol Cell Biol* **5,** 89–99 (2004).

48. Czaplinski, K. *et al.* The surveillance complex interacts with the translation release factors to enhance termination and degrade aberrant mRNAs. *Genes Dev* **12,** 1665–1677 (1998).

49. Lejeune, F., Li, X. & Maquat, L.E. Nonsense-mediated mRNA decay in mammalian cells involves decapping, deadenylating, and exonucleolytic activities. *Mol Cell* **12,** 675–687 (2003).

50. Culbertson, M.R. RNA surveillance. Unforeseen consequences for gene expression, inherited genetic disorders and cancer. *Trends Genet* **15,** 74–80 (1999).

51. Holbrook, J.A., Neu-Yilik, G., Hentze, M.W. & Kulozik, A.E. Nonsense-mediated decay approaches the clinic. *Nat Genet* **36,** 801–808 (2004).

52. Pagani, F. & Baralle, F.E. Genomic variants in exons and introns: identifying the splicing spoilers. *Nat Rev Genet* **5,** 389–396 (2004).

53. Garcia-Blanco, M.A., Baraniak, A.P. & Lasda, E.L. Alternative splicing in disease and therapy. *Nat Biotechnol* **22,** 535–546 (2004).

54. Venables, J.P. *et al.* Identification of alternative splicing markers for breast cancer. *Cancer Res* **68,** 9525–9531 (2008).

55. Singh, R.N. Evolving concepts on human SMN pre-mRNA splicing. *RNA biology* **4,** 7–10 (2007).

56. Khoo, B., Akker, S.A. & Chew, S.L. Putting some spine into alternative splicing. *Trends Biotechnol* **21,** 328–330 (2003).

57. Skordis, L.A., Dunckley, M.G., Yue, B., Eperon, I.C. & Muntoni, F. Bifunctional antisense oligonucleotides provide a trans-acting splicing enhancer that stimulates SMN2 gene expression in patient fibroblasts. *Proc Natl Acad Sci U S A* **100,** 4114–4119 (2003).

58. Baughan, T.D., Dickson, A., Osman, E.Y. & Lorson, C.L. Delivery of bifunctional RNAs that target an intronic repressor and increase SMN levels in an animal model of spinal muscular atrophy. *Hum Mol Genet* **18,** 1600–1611 (2009).

59. Soret, J. *et al.* Selective modification of alternative splicing by indole derivatives that target serine-arginine-rich protein splicing factors. *Proc Natl Acad Sci U S A* **102,** 8764–8769 (2005).

60. Fukuhara, T. *et al.* Utilization of host SR protein kinases and RNA-splicing machinery during viral replication. *Proc Natl Acad Sci U S A* **103,** 11329–11333 (2006).

61. Muraki, M. *et al.* Manipulation of alternative splicing by a newly developed inhibitor of Clks. *J Biol Chem* **279,** 24246–24254 (2004).

62. Suess, B. & Weigand, J.E. Engineered riboswitches: overview, problems and trends. *RNA Biol* **5,** 24–29 (2008).

63. Coulter, L.R., Landree, M.A. & Cooper, T.A. Identification of a new class of exonic splicing enhancers by in vivo selection. *Mol Cell Biol* **17,** 2143–2150 (1997).

64. Liu, H.X., Zhang, M. & Krainer, A.R. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev* **12,** 1998–2012 (1998).

65. Schaal, T.D. & Maniatis, T. Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. *Mol Cell Biol* **19,** 261–273 (1999).

66. Tian, H. & Kole, R. Strong RNA splicing enhancers identified by a modified method of cycled selection interact with SR protein. *J Biol Chem* **276,** 33833–33839 (2001).

67. Wang, Z. *et al.* Systematic identification and analysis of exonic splicing silencers. *Cell* **119,** 831–845 (2004).

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

69. Fairbrother, W.G. *et al.* RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res* **32,** W187–190 (2004).

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

71. Wang, Z., Xiao, X., Van Nostrand, E. & Burge, C.B. General and specific functions of exonic splicing silencers in splicing control. *Mol Cell* **23,** 61–70 (2006).

72. Pozzoli, U. & Sironi, M. Silencers regulate both constitutive and alternative splicing events in mammals. *Cell Mol Life Sci* **62,** 1579–1604 (2005).

73. Zhang, X.H., Kangsamaksin, T., Chao, M.S., Banerjee, J.K. & Chasin, L.A. Exon inclusion is dependent on predictable exonic splicing enhancers. *Mol Cell Biol* **25,** 7323–7332 (2005).

74. Zhang, X.H., Leslie, C.S. & Chasin, L.A. Computational searches for splicing signals. *Methods* **37,** 292–305 (2005).

75. McCullough, A.J. & Berget, S.M. An intronic splicing enhancer binds U1 snRNPs to enhance splicing and select 5' splice sites. *Mol Cell Biol* **20,** 9225–9235 (2000).

76. Brudno, M. *et al.* Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. *Nucleic Acids Res* **29,** 2338–2348 (2001).

77. Blencowe, B.J. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci* **25,** 106–110 (2000).

78. Charlet, B.N., Logan, P., Singh, G. & Cooper, T.A. Dynamic antagonism between ETR-3 and PTB regulates cell type-specific alternative splicing. *Mol Cell* **9,** 649–658 (2002).

79. Fairbrother, W.G. & Chasin, L.A. Human genomic sequences that inhibit splicing. *Mol Cell Biol* **20,** 6816–6825 (2000).

80. Yeo, G., Hoon, S., Venkatesh, B. & Burge, C.B. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci U S A* **101,** 15700–15705 (2004).

81. Venables, J.P. Downstream intronic splicing enhancers. *FEBS Lett* **581,** 4127–4131 (2007).

Chapter II. Functional selection of intronic splicing elements provides insight into their regulatory mechanism

Abstract

Despite the critical role of alternative splicing in generating proteomic diversity and regulating gene expression, the sequence composition and function of intronic splicing regulatory elements (ISREs) have not been well elucidated. Here, we employed a high-throughput *in vivo* Screening PLatform for Intronic Control Elements (SPLICE) to identify 125 unique ISRE sequences from a random nucleotide library. Bioinformatic analyses reveal consensus motifs that resemble splicing regulatory elements and binding sites for characterized splicing factors and that are enriched in the introns of naturally-occurring spliced genes, supporting their biological relevance. *In vivo* studies, including an RNAi silencing study, demonstrate that ISRE sequences can exhibit combinatorial regulatory activity and that complex splicing regulatory networks (SRNs) are involved in the regulatory effect of a single ISRE. From our results, we propose three mechanisms through which ISREs interact with splicing factors to achieve regulatory function: direct binding / competition, recruitment, and agonist interaction.

2.1. Introduction

Post-transcriptional gene regulatory mechanisms play central roles in programming the complexity of biological systems. One such process is alternative splicing, a dynamic mechanism that produces multiple protein isoforms from a single gene by altering the ways in which exons are joined from a single pre-mRNA[1]. Splicing patterns are regulated by the interplay between auxiliary *cis*-acting elements that include exonic and intronic splicing enhancers (ESEs and ISEs, respectively) and exonic and intronic splicing silencers (ESSs and ISSs, respectively) and the trans-acting factors that modulate them, leading to a 'splicing code'[2]. The lack of high-throughput *in vivo* methods for analyzing the function of spliced variants and the *cis*-acting elements involved in the regulation of these transcripts has hindered the functional validation of spliced transcripts discovered through recent genome-wide mRNA sequencing studies[3-5]. Bioinformatic and experimental analyses have identified several RNA motifs that regulate splicing, where much of this effort has been directed toward the functional characterization of *cis*-acting exonic regulatory sequences[6-9]. Despite the widespread importance of intronic splicing regulatory elements (ISREs), knowledge regarding their sequence composition, the mechanisms through which they regulate splicing and the regulatory networks of trans-acting splicing factors by which they are bound, or splicing regulatory networks (SRNs), is limited. The development of a functional definition of ISREs and the elucidation of corresponding SRNs is of great interest given that > 90% of human genes are alternatively spliced[10] and that up to 50% of disease-causing mutations affect splicing[11].

Several properties of ISREs have complicated their functional characterization. ISSs and ISEs have been identified near alternatively spliced exons; however, their actions appear to be antagonistic[12] suggesting that they behave in a combinatorial manner[13]. In addition, the activities of some sequences are context dependent[10,14]. ISSs may inhibit exon inclusion by recruiting splicing repressors that directly antagonize splicing factor binding or by recruiting repressors to multiple binding sites resulting in a 'zone of silencing'[15]. While several ISEs have been characterized[16], the trans-acting factors that bind these sequences remain unknown[17].

To begin to generate a functional definition of ISREs, we have developed a generalizable *in vivo* screening strategy for ISREs, which we call SPLICE (Screening PLatform for Intronic Control Elements). SPLICE was used to identify intronic sequences that regulate the inclusion of an alternatively spliced exon that triggers rapid transcript decay through nonsense-mediated decay (NMD). Our high-throughput approach combines a systematic screening strategy, extensive genome-wide bioinformatic analyses and experimental characterization, including an RNAi silencing study, to identify ISRE consensus motifs, characterize the SRNs associated with these global regulatory elements and generate a model for ISRE regulatory function. Our results indicate that *cis*-acting intronic regulatory sequences function through combinatorial effects from multiple elements and trans-acting factors, and that the immediate transcript context has a dominant effect on ISRE function. In addition, our results support three mechanisms for ISRE regulatory function: direct binding / competition, recruitment, and agonist interaction.

2.2. Results

2.2.1. SPLICE: a Screening PLatform for Intronic Control Elements

SPLICE is a high-throughput *in vivo* screen for ISRE function based on a reporter construct encoding the green fluorescent protein (GFP) fused 5' of a three-exon, two-intron mini-gene. The alternatively-spliced middle exon harbors a premature termination codon (PTC) that triggers mRNA degradation through the NMD pathway[18]. Auxiliary elements that regulate alternative splicing are normally positioned in proximity to splice sites[16,19,20] and have been shown to vary in length between 10- to 30-nt[19]. We implemented SPLICE with the SMN1 mini-gene containing a random 15-nucleotide (nt) library positioned 45-nt upstream of the 3' ss in the first intron (Figure 2.1a). Therefore, cells with a high level of exon 7 inclusion display lower GFP fluorescence than cells in which this exon is excluded. By coupling NMD to splicing efficiency, ISREs with a range of activities can be selected using fluorescence activated cell sorting (FACS).

To test the utility of NMD as the basis of SPLICE we examined the difference in fluorescence between a NMD-based reporter construct (NMD control), containing a 15-nt control insert and a PTC in exon 7, relative to a construct lacking a PTC (GFP-SMN1 control). All constructs were stably transfected into HEK-293 FLP-In cells to generate isogenic cell lines. Flow cytometry (Figure 2.1a and Figure S2.1a) and fluorescence microscopy analyses (Figure S2.1b) reveal that the fluorescence difference between the GFP-SMN1 and NMD controls is ~22-fold. Transcript isoform analysis through quantitative real time-PCR (qRT-PCR) indicates that the level of exon 7 inclusion in the NMD control is ~60-fold less than the GFP-SMN1 control (Figure S2.1c,d), supporting that differences in fluorescence are due to exon 7 inclusion.

A library of synthetic DNA oligonucleotides containing a random 15-nt region (~$1x10^9$ sequences) was ligated into the NMD control construct and transformed into *Escherichia coli*. Library constructs were purified from ~$1x10^6$ pooled transformants, representing ~0.1% of possible sequences. The pooled library was stably transfected into HEK-293 FLP-In cells and ~450,000 stable transformants were generated (Methods). Sequencing of the library before and after transfection demonstrated minimal sequence bias at each position (Figure S2.1e). FACS analysis indicated that ~0.05%–0.1% of the cell population exhibits fluorescence levels greater than the NMD control, corresponding to putative ISSs. Positive cells were bulk sorted, grown 2-3 weeks and re-analyzed by flow cytometry. The round-one pool exhibits an approximate six fold increase in mean fluorescence compared to the NMD control (Figure 2.1a) and was re-sorted into different groups based on fluorescence ranges (A, B, and C) to further enrich the population and select for sequences varying in splicing regulatory activity (Figure 2.1a and Figure S2.2). The enriched populations were analyzed by flow cytometry, and the mean fluorescence levels correlated well with their sorted sections (Figure 2.1b).

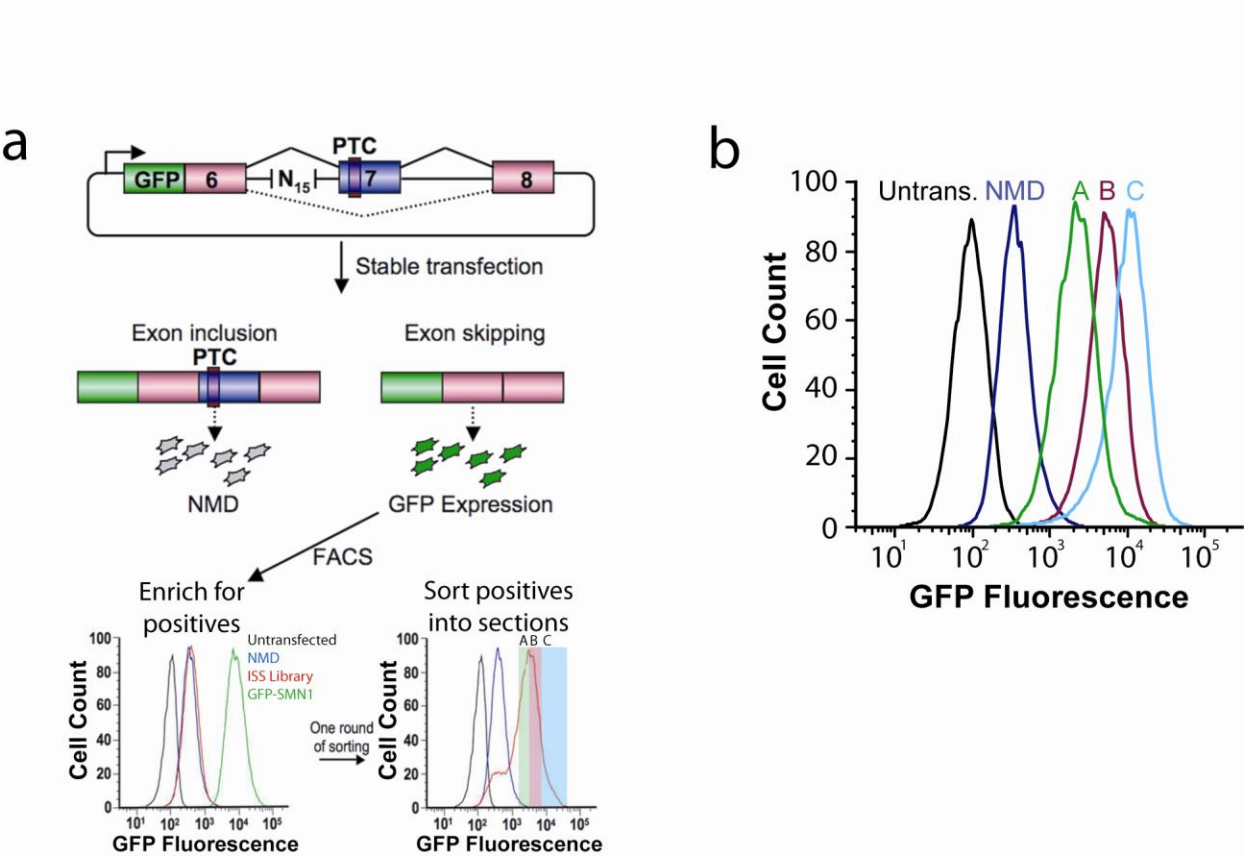


Figure 2.1. A Screening PLatform for Intronic Control Elements (SPLICE) provides a generalizable *in vivo* screening strategy for ISREs. **(a)** The application of SPLICE to the screening of ISRE sequences. SPLICE couples an exon inclusion event in a mini-gene (SMN1) to the expression level of a fluorescent reporter protein (GFP) through a NMD-based reporter system. A random nucleotide library cloned into unique restriction sites in intron 6 is screened for ISSs by sorting cells exhibiting fluorescence levels higher than the negative control (NMD). The enriched cells are expanded and later sorted into sections (A, B, C) based on user-designated fluorescence levels in a second screening round. **(b)** The enriched cell populations maintain the fluorescence levels of the sorted sections (A, B, C). Following the second round of sorting, the fluorescence levels of expanded populations were re-analyzed through flow cytometry to confirm maintenance of expression levels.

2.2.2. Recovered ISRE sequence composition correlates with sorted sections

We identified 125 unique sequences with enhanced fluorescence from 480 sequenced isolates (Table S2.1). Three of the recovered ISRE sequences exhibit significant (12 of 15-nt) similarity to portions of the SMN1 mini-gene, suggesting that these sequences may be involved in the cooperative assembly of repressor elements on the SMN1 transcript (Figure S2.3). The sequences have a higher level of G (35.8%) and reduced levels of T (22%), C (18.6%), and A (23.6%) (Figure S2.4a). The dinucleotide CC is overrepresented in the ISRE dataset, while others, such as AC, AG, CA, GT, TA, TC, and TG, are only slightly enriched (Figure S2.4b).

Recovered 15-mers were subjected to hierarchical clustering to determine the overall sequence similarity between elements (Figure 2.2 and Figure S2.5)[6,9]. SPLICE-generated sequences are generally diverse (>95% of sequences differ by more than 1-nt), indicating that the majority of the recovered sequences arose from independent selection. We evaluated the association between clusters of sequences (using a dissimilarity score cutoff of 1.1) and the fluorescent section from which they were sorted. In particular, clusters 11 and 13 show a significant association with the sorted sections, while clusters 10 and 18 do not. The resulting clusters generally contain sequences from identically sorted sections suggesting that sequence composition correlates with cellular fluorescence (Figure 2.2).

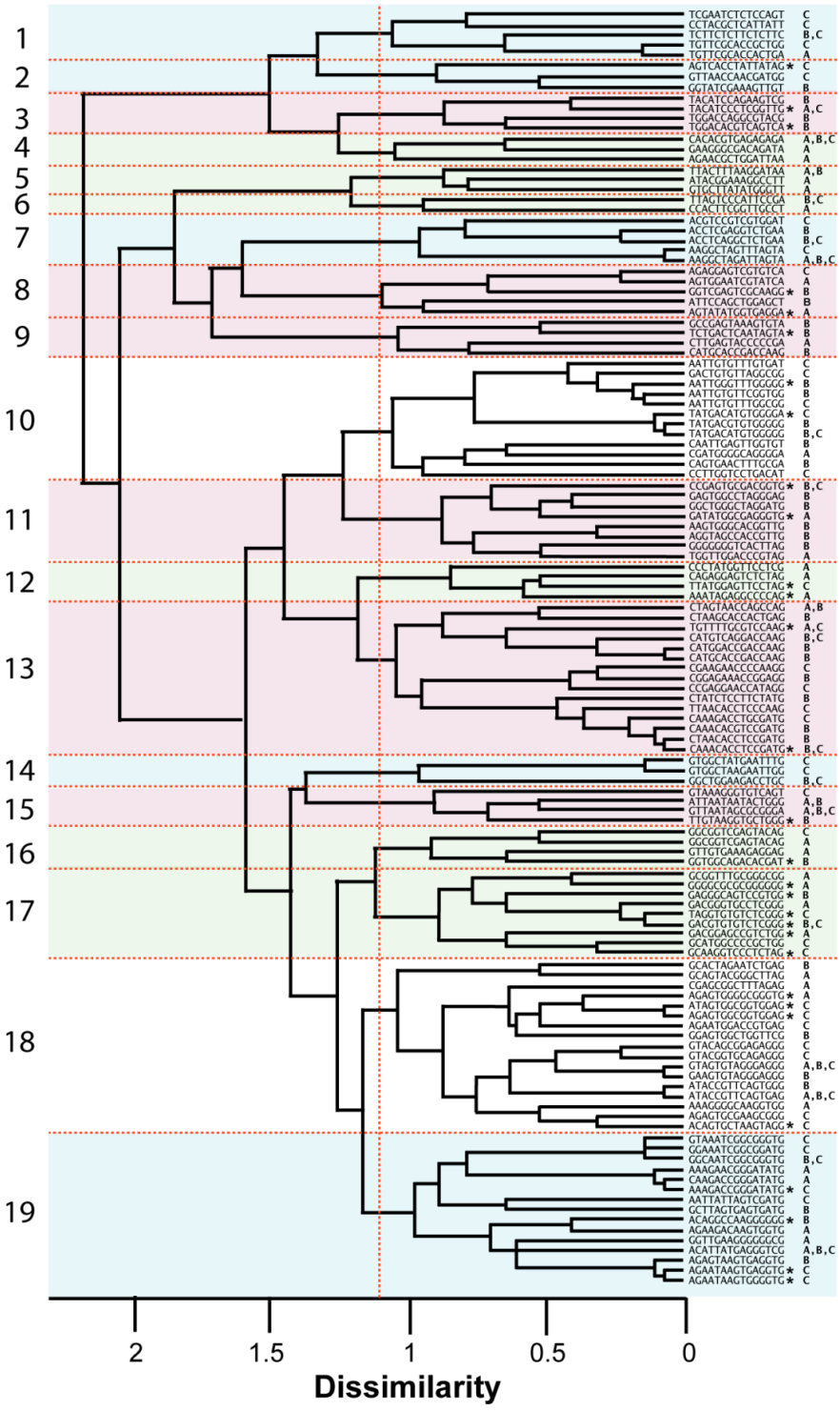


Figure 2.2. Hierarchical clustering of recovered ISREs indicates sequence composition correlates with sorted sections. Hierarchical clustering applied to 125 recovered ISRE sequences identified 19 clusters using a dissimilarity cutoff of 1.1. Clusters that have over

40% sequence representation from one sorted section are indicated (A, green; B, red; C, blue). Sorted sections for each sequence are denoted. Starred sequences were subjected to additional studies to examine regulatory activity.

2.2.3. GCCS clustering of recovered ISREs identifies motifs similar to known splicing factor binding sites

The sequence alignment results indicate that multiple motifs of varying lengths occur within the 15-nt dataset (Figure S2.5). For analyzing datasets of this nature, Graph Clustering by Common Substrings (GCCS)[21] is better suited than hierarchical clustering. We analyzed a 19-nt region including the 15-mer ISRE sequence and 2-nt of the flanking regions for sequence enrichment. Since RNA binding proteins typically recognize short sequence motifs, we restricted our analysis to n-mers ranging from 4–6-nt. We determined the enrichment of n-mers in a sample of 125 sequences using a confidence interval for the binomial distribution based on probabilities expected for 19-nt oligonucleotides containing 15-nt of uniformly random bases flanked by the 2 constant bases present in the experimental system. In the ISRE dataset, 241 n-mers consisting of 39 4-mers, 93 5-mers, and 109 6-mers were significantly enriched ($\alpha_{1\text{-tailed}} = 0.1$; Figure 2.3a and Table S2.2). The GCCS analysis grouped 80.1% of the statistically enriched 4-6-nt n-mers into 30 consensus motif clusters (Figure 2.3b and Tables S2.3 and S2.4, Methods).

Many of the consensus motifs identified by the GCCS analysis resemble known binding sites for trans-acting splicing factors (Figure 2.3b and Table S2.5). The largest number of motifs resembles binding sites of the hnRNP family of proteins (class 1). In

particular, class 1 motifs resemble binding sites for several known repressors of splicing: hnRNP A1 (TAGGG)[22], hnRNP F/H (GGGGG)[23], the polypyrimidine tract binding protein PTB (hnRNP I, CT-rich)[24] and hnRNP L (CA-rich)[25]. The significant similarity between binding sites for the hnRNP family of proteins and the enriched motifs supports the possible functional role of selected ISREs.

Several identified motifs resemble known binding sites for the SR protein family (class 2) whose members act as general splicing factors[1]. Enriched ISREs within class 2 resemble binding sites for SF2/ASF (GAAGAA)[26], SRp40 (ACAAG)[27], SRp30c (CTGGATT)[14], SC35 (AGGAGAT)[28], 9G8 (GACC)[28], and Tra2β $(GAA)_n$[29]. While the examples of SR proteins involvement in splicing repression are limited, the enrichment of motifs similar to binding sites for members of this family suggests that their role in intronic regulation may be more widespread than previously thought.

Several of the enriched motifs identified in our dataset resemble the major 5' splice site (ss) consensus sequence GT[A/G]AGT (class 3)[30]. All four motifs in class 3 contain an AGT core element, and the enriched motif TAAGTG is almost identical to the canonical 5' ss sequence and the hnRNP G binding motif AAGT[31]. The occurrence of 5' ss motifs within intronic regulatory elements has been noted[32] and computational analyses have identified conserved elements that are similar to the consensus 5' ss within mammalian intronic regions[21,33]. In addition, the enrichment of 5' ss motifs was previously observed in an *in vivo* screen for ESSs[9]. Taken together, these results add support to the role of cryptic 5' ss in regulating alternative splicing.

Other enriched motifs in our dataset (GTGT, GGTGG, TTGTGT, and GGTT) resemble known binding sites for the CELF/Bruno-like family (class 4). This family of

proteins regulates alternative splicing patterns by binding sequences that contain CTG repeats and exhibit a higher affinity for GT repeats[34]. The motifs GTGT and TGTG resemble binding sites to a well-characterized member of this family, CUG-BP1, which has been shown to bind TGT-containing sequences[34]. The GTGT motif may also serve as a binding site for hnRNP M[35].

GCCS identified 5 motifs that represent either novel regulatory elements or weak binding sites for characterized splicing factors (class 5). The [A/G]TGGC motif is similar to a degenerate CELF protein binding site and the motif TCGG[G/C] shares up to 80% sequence identity to a hnRNP A1 binding site. Strikingly, the GCTGG, CGA[T/G] and TATG motifs have not been previously identified. Therefore, in addition to identifying elements resembling binding sites for characterized trans-acting splicing factors, SPLICE generated novel regulatory elements.

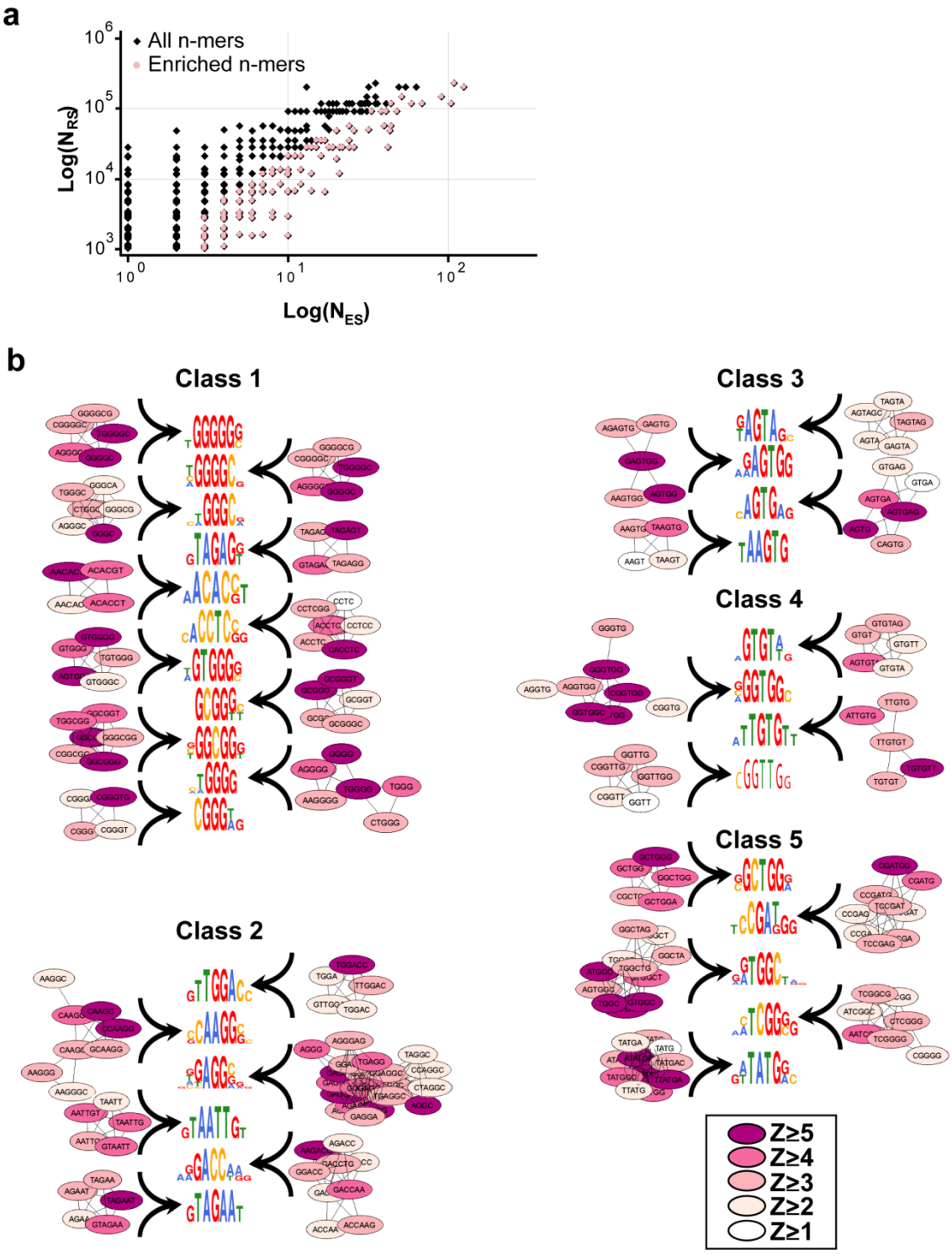


Figure 2.3. Enriched motifs and GCCS clusters derived from recovered ISRE sequences map to known and unknown splicing factors. **(a)** Scatter-plot for the occurrence

frequency of all 4–6-nt n-mers in the enriched sample set (N_{ES}) vs. a corresponding random sample set (N_{RS}) (black). A similar scatter-plot based on n-mers determined to be significantly enriched in the recovered ISREs is overlaid (pink). **(b)** Consensus motif groupings according to resemblance to binding sites for trans-acting splicing factors. Motif classes include enriched ISREs that are similar to the binding sites of the hnRNP, SR and CELF families of proteins and the 5' ss (classes 1–4, respectively). Class 5 consists of previously unidentified elements and may represent novel regulatory sequences. The graph clusters representing the enriched n-mers used to construct each consensus motif are shown. Vertices are colored according to the enrichment Z-scores.

2.2.4. Enriched ISRE n-mers resemble known splicing regulatory elements

To investigate the potential functional role and general significance of the identified ISRE motifs, we examined the number of pentamer motifs identified in our enriched n-mer dataset (Table S2.2) that are identical to pentamers in published sets of splicing regulatory elements (SREs). We analyzed data corresponding to four SRE classes: ESEs[6,7], ESSs[7,9], ISEs[16], and computationally identified conserved intronic elements. The latter class includes conserved intronic sequences (CISs)[21], ISREs[33], pentamers enriched in intronic regions of excluded exons in neural progenitor (NP) cells[36] and motifs enriched upstream of weak polypyrimidine (PY) tracts in AT- and GC-rich introns[37]. Significant overlap exists between the enriched pentamers and ESSs, ISEs, donor intronic (DI) elements in NP cells and motifs enriched upstream of weak PY tracts ($P < 0.05$ for ESSs, $P << 0.0001$ for ISEs, NP DI elements and weak PY tracts) (Figure 2.4a). The dominant motifs that overlap between SPLICE-generated pentamers and ISEs

and weak PY elements are G-rich elements, similar to hnRNP A/B and hnRNP F/H binding sites and the canonical 5' ss. The results suggest that the selected elements may function as general splicing silencers and intronic modulators of splicing (as both silencers and enhancers) depending on their context across various cell types and are likely regulated by general splicing factors. The comparison between enriched pentamers and conserved acceptor intronic elements (AI) for CIS and ISRE datasets demonstrate some overlap ($P < 0.05$). However, the observed overlap is far less than expected, suggesting that SPLICE selected against these elements.

2.2.5. Genome-wide analysis demonstrates that enriched ISREs associate with spliced exons

The biological relevance of selected motifs was examined by assessing the association of enriched motifs with naturally occurring alternative and constitutive splicing events. The occurrence of enriched motifs in the region 80-nt upstream of the AI regions flanking skipped exons was determined using a database of alternatively spliced junctions throughout the human genome[21]. A portion of SPLICE-generated ISREs significantly associate with alternative splicing (2 of 30; $P_{\text{t-test}} < 0.01$; class 4 only) and constitutive splicing (10 of 30; $P_{\text{t-test}} < 0.05$; all classes except 3) (Figure 2.4b). The entire population of consensus n-mers significantly associates with constitutive splicing ($P_{\text{t-test}} = 1.8e^{-8}$). This association is unexpected since selected ISREs are located within an alternatively spliced gene. However, the alternative exon 7 of the SMN1 mini-gene strongly favors inclusion, such that it may display regulatory signals similar to those involved in constitutive splicing, potentially biasing the sequence composition of selected

ISREs towards the association with constitutive splicing. In addition, our ISREs are likely to be more enriched in ISSs, which have been shown to be enriched in the intronic flanks of constitutively spliced exons[2]. Results from our genome-wide association analysis suggest that selected ISREs serve an important role in defining constitutive and alternative splice sites.

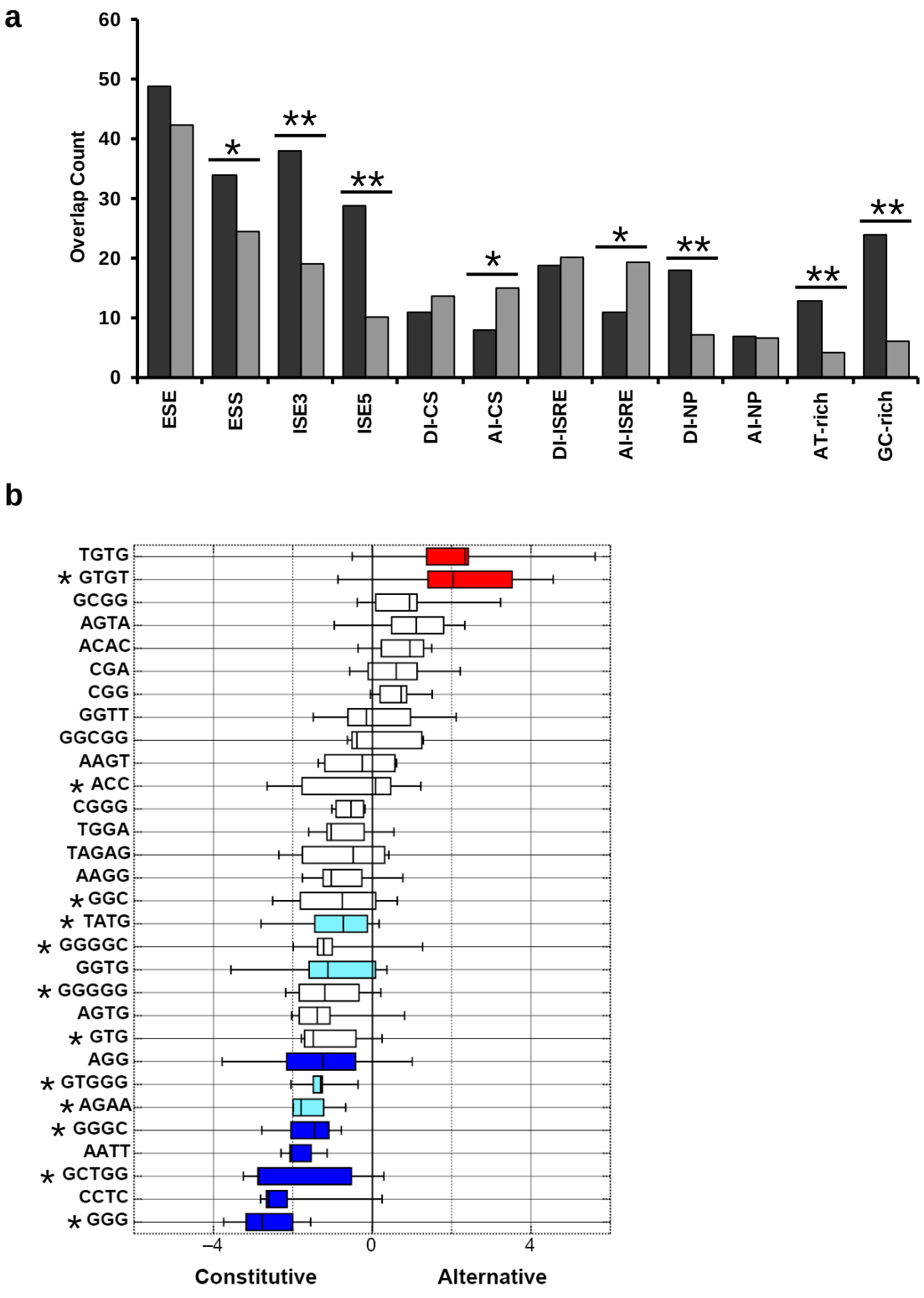


Figure 2.4. Enriched n-mers overlap with both experimentally and computationally derived SREs and associate with constitutive and alternative splicing. **(a)** Overlap of

enriched n-mers from recovered ISRE sequences with known classes of SREs. Observed (black) and expected overlap (gray) between datasets is shown. *P*-values derived from the chi-squared test of association are as follows: * $P < 0.05$ and ** $P << 0.0001$. **(b)** Box-plots revealing the distribution of TA-scores for GCCS derived ISREs. The GCCS consensus motifs that are significantly associated with alternative splicing are shown in red ($P_{t\text{-}test} < 0.01$) and those that are significantly associated with constitutive splicing are shown in shades of blue (dark blue, $P_{t\text{-}test} < 0.01$; light blue, $P_{t\text{-}test} < 0.05$). In total, 9 consensus motifs are biased toward alternative splicing and 21 consensus motifs display a bias towards constitutive splicing. Elements exhibiting no significant association with either category are not shaded. Starred motifs are present in hexamers subjected to RNAi silencing studies to examine regulated splicing.

2.2.6. Recovered ISRE sequences enable tuning of alternative splicing

The silencer activity of recovered 15-mers was validated by characterizing individual sequences that were selected randomly from 13 of the 19 hierarchical clusters (Figure 2.2). We analyzed an initial set of 18 15-mers (Figure 2.5a) and 4 known ISS sequences: an hnRNP H[38], 2 PTB[39,40] and a U2AF65 binding sites[39]. Individual sequences were cloned into our NMD-based reporter, stably transfected into HEK-293 FLP-In cells and analyzed by flow cytometry. Of the known ISSs tested, only the U2AF65 element demonstrates significant silencing activity relative to the NMD control, exhibiting an ~1.5-fold higher fluorescence level (Figure 2.5b). This result is in line with studies demonstrating that the silencing mechanisms of several characterized ISREs are context dependent[41]. In contrast, 16 of the selected sequences display significant silencer activity

($P << 0.001$) and 2 exhibit enhancer activity relative to the NMD control ($P < 0.05$) (Figure 2.5b), and over half exhibit silencing activities equal to or greater than the U2AF65 element. Similar trends were seen upon examination of an additional 12 recovered sequences (Figure S2.6b). In addition, we arranged the sequences into groups representing low (ISS1-5), medium (ISS6-10), and high (ISS11-16) silencing activities and determined the section from which each sequence was recovered (Figure 2.2). The activities of the majority of tested sequences correlated with sectioned populations, where a subset of enriched n-mers GGGGC, GGGC, and GGG correlated significantly with their sorted section ($P << 0.01$) and those sequences that did not correlate were shown to cluster with the appropriate group by sequence. These results support that functional regulatory activity is related to sequence.

To directly examine changes in splicing patterns, we analyzed the transcript isoforms of 12 of the recovered sequences and the ISS controls by qRT-PCR. The total transcript levels and the levels of intron retention for the examined ISS and control sequences were similar to the NMD control, while these levels for the selected ISEs differed from the NMD control ($P < .05$) (Figure S2.7a–d). The GFP-SMN1 control exhibits a low level of the skipped exon isoform compared to the NMD control ($P < 0.05$) (Figure 2.5c). As expected, for most of the recovered and control ISS sequences the levels of the skipped exon isoform are significantly higher than the NMD control ($P < 0.05$), with the exception of ISS15 ($P = 0.51$) and ISS8 ($P = 0.40$). In addition, the ISE sequences exhibited lower levels of the skipped exon isoform relative to the NMD control ($P < 0.05$). Therefore, expression levels of the skipped exon isoform generally confirm the activity of the sequences observed by fluorescence measurements.

All constructs except for the GFP-SMN1 control are expected to exhibit low levels of the exon 7 included isoform, as this isoform should be rapidly degraded through NMD. The GFP-SMN1 control exhibits a high level of exon 7 inclusion (99.7%), ~60-fold more than the NMD control. Exon inclusion levels for the ISS controls do not differ from the NMD control ($P > 0.35$), with the exception of PTB(2), which had a higher level of exon inclusion ($P < 0.005$). Exon inclusion levels for 8 of the 10 recovered ISS sequences (ISS5, ISS8-13) and the ISE sequences range from 2 to 20-fold less than the NMD control ($P < 0.05$), whereas ISS15 and ISS16 exhibited increased levels of the exon included isoform relative to the NMD control ($P < 0.05$, Figure S2.7e). The elevated levels of exon inclusion observed from several ISSs is not a result of cryptic splice sites as determined by analyzing the sizes of the RT-PCR amplification products (data not shown). Overall, the majority of sequences that display increased fluorescence have decreased levels of exon 7 inclusion compared to the NMD control, supporting their silencer function.

2.2.7. ISRE sequences function in a different cell type

The relative levels[42] and activities[43] of trans-acting splicing factors vary widely across different cell types, which may result in *cis*-acting sequences exhibiting different regulatory activities. To determine whether the selected ISREs are cell type specific, we examined their regulatory function in a second cell line. We first examined the fluorescence of ISS1-16, ISE1, and the NMD and GFP-SMN1 controls in a transient transfection assay in the HEK-293 cell line to verify that regulatory activity was observed under these conditions. Flow cytometry analysis reveals that transiently transfected cells

display increased expression levels and population distributions relative to stable cell line assays (Figure S2.8). As such, the relative expression of the GFP-SMN1 control is only ~4.1-fold that of the NMD control (Figure 2.5d). Despite the decreased sensitivity of the transient transfection assay, the qualitative activity of 15 of the recovered ISREs was maintained and 11 sequences exhibited significantly increased expression ($P < 0.05$).

We next investigated whether the recovered sequences function in HeLa cells. The GFP-SMN1 control displays a approximate six fold higher level of expression than the NMD construct in HeLa cells in the transient transfection assay (Figure 2.5e). The ISRE sequences display a range of expression levels, but all are significantly different than the NMD control ($P < 0.05$). The majority of examined sequences (12 of 16) maintain the same trend in activity in HeLa cells as was observed in HEK-293 cells and ANOVA analysis of the activities in both cell lines shows a strong correlation ($P < 0.0005$). In contrast, four of the tested sequences (ISS3, 5, 6, and 13) exhibit enhancer activity relative to the NMD control in HeLa cells, which may be due to differences in levels of trans-acting factors between the cell lines. The results support that most sequences recovered from SPLICE retain function in a cell line different from which they were selected and may represent global splicing regulators.

2.2.8. Analysis of recovered ISRE sequences in a different transcript supports context dependent function

The context dependence of *cis*-regulatory elements on splice site choice has been shown[44] and we have observed little activity from known silencers in the context of the SMN1-NMD reporter system (Figure 2.5b). A subset of the selected ISREs was tested for

context dependent function by examining their activity in a second NMD-based reporter, based on the BRCA1 gene consisting of exons 17, 18, and 19[45], via transient transfection in HEK-293 cells. Selected ISRE sequences were inserted 50-nt upstream from the 3' ss of exon 18. Analysis of the reporter constructs by flow cytometry reveals a approximate two fold difference between the positive and negative controls ($P < 0.05$, Figure 2.5f). Only 3 of the tested sequences (ISS14, 17 and 18) exhibit significant silencer activity ($P < 0.05$) in the context of the BRCA1 mini-gene, while 1 sequence (ISS15) exhibits enhancer activity ($P < 0.05$). Transcript isoform analysis indicates that the level of exon 18 inclusion in the NMD control is ~12.5 fold less than the GFP-BRCA1 control (Figure S2.7f) and that splicing patterns for a subset of tested ISRE sequences were similar to the NMD control (Figure S2.7g), validating the lack of ISRE activity observed by fluorescence measurements. A predicted secondary structure analysis of the intronic regions shows that individual ISREs change the overall structure of each intron very little (Figure S2.9). However, the predicted secondary structure of the SMN1 intron is significantly different than that for the BRCA1 intron. The results suggest that the regulatory activity of SPLICE-generated ISREs is likely dependent on specific properties of the mini-gene in which they are selected.

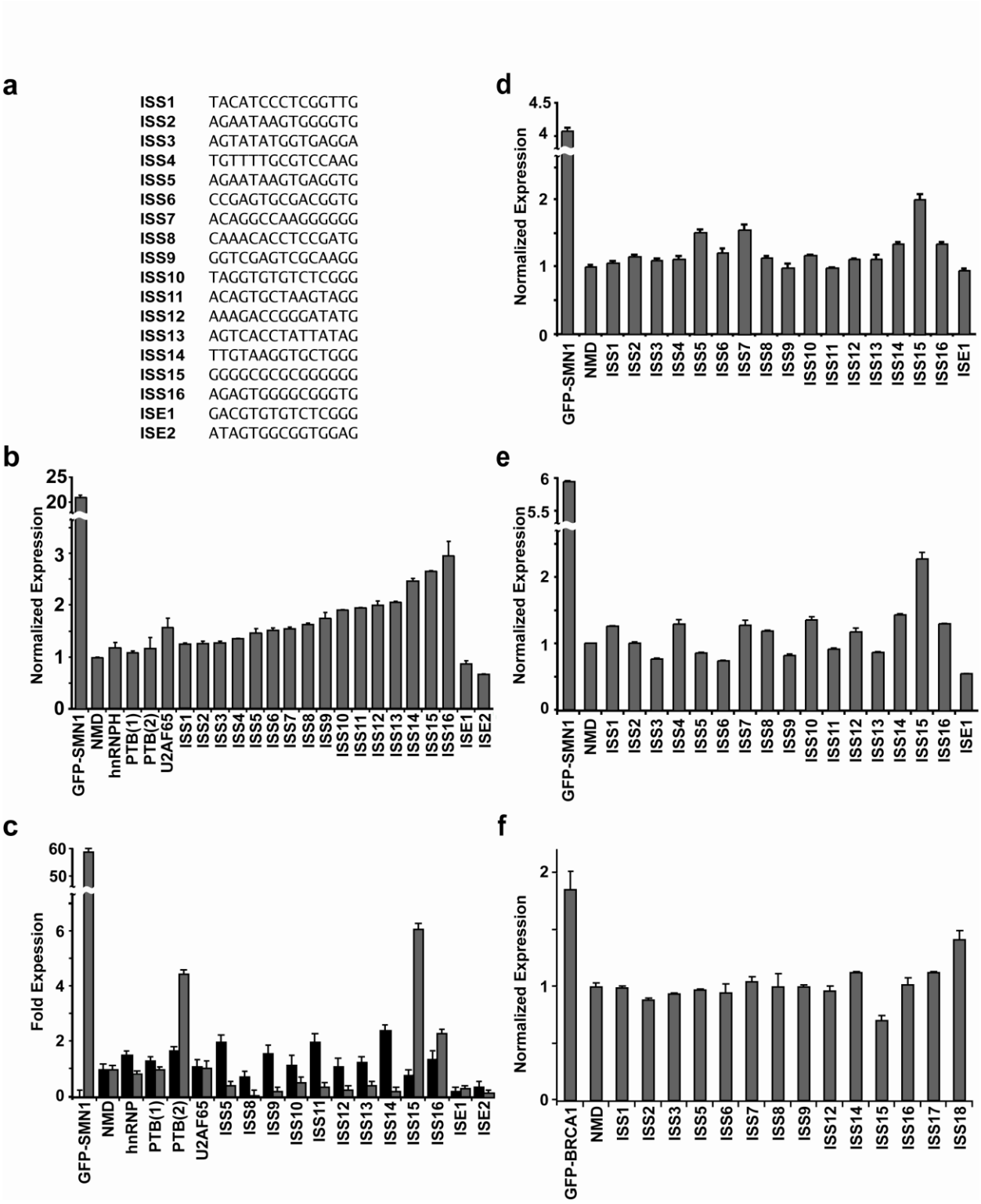


Figure 2.5. Functional analysis of recovered ISRE sequences. **(a)** Recovered ISRE sequences examined for regulatory activity. **(b)** Flow cytometry analysis of HEK-293

FLP-In stable cell lines generated for recovered ISRE sequences and control constructs. For all reported activities, the mean GFP levels from two independent experiments were determined and normalized to the NMD control. Normalized expression and average error are reported. ISRE sequences are labeled according to function. **(c)** qRT-PCR analysis of the ISS control sequences and 12 selected sequences with primer sets specific for exon 7 included (black bars) and excluded (gray bars) products. Expression levels of duplicate PCR samples were normalized to the levels of *HPRT*. Fold expression data is reported as the mean expression for each sample divided by the mean NMD expression value $\pm$ the average error. **(d)** Flow cytometry analysis of recovered ISRE sequences and control constructs transiently transfected in HEK-293 cells. **(e)** Flow cytometry analysis of recovered ISRE sequences and control constructs transiently transfected in HeLa cells. **(f)** Flow cytometry analysis of recovered ISRE sequences and control constructs in the BRCA1 mini-gene transiently transfected in HEK-293 cells.

2.2.9. Analysis of enriched hexamers confirms independent and combinatorial function

We examined the silencer activities of representative hexamers from consensus motifs within the GCCS clusters through transient assays in HEK-293 cells to confirm the activity of individual motifs. Hexamers resembling the PTB, hnRNP H, SF2/ASF, Sam and the CELF protein binding sites, the 5' ss and an unknown motif were examined (classes 1–5; Figure 2.3b). Silencing activity was investigated by comparing expression levels of the hexamer alone to the hexamer with double point mutations (loss of function) and to the hexamer in duplicate (Figure 2.6a). A majority of the mutated hexamers (PTB, hnRNP H, Sam and unknown motifs) exhibits significant loss of function ($P < 0.05$;

classes 1, 2, and 5). However, only one of the hexamers (class 4) displays an increase in silencer activity when present in duplicate. The results indicate that while individual hexamers exhibit silencing activity and likely represent core ISREs, they do not necessarily behave in an additive manner likely due to context and spacing requirements. For example, the duplicate hnRNP H hexamer does not exhibit increased silencing, whereas ISS15, which differs from the duplicate hexamer by 3 cytosine residues, exhibits strong silencer activity. The additional residues may provide spacing between the G-rich hexamers important for functional activity.

To test the possibility of combinatorial control within the context of a selected 15-mer, we examined two ISS sequences that contained multiple enriched hexamers. Most of the recovered ISRE sequences contain several enriched n-mers, where 88% of all extended 15-mers (plus 2-nt flanking region) contain at least one enriched hexamer (Table S2.6). ISS5 contains 7 enriched hexamers resembling the 5' ss and binding sites for the CELF and SF2/ASF proteins, which cluster into 3 main regions within the sequence (Figure 2.6b). We introduced 2 point mutations within each region and in combination and assessed their activity through transient transfection assays in HEK-293 cells. Mutations within each region of ISS5 decrease expression to levels comparable to the NMD control, indicating that each silencing zone has regulatory activity. Region 1 contains two overlapping hexamers resembling the SF2/ASF binding motif, where one of these is the SF2/ASF representative hexamer that did not demonstrate silencing activity in the hexamer analysis studies, suggesting that the regulatory function of this hexamer is context dependent (Figure 2.6a). Simultaneous mutations to regions 2 and 3 resulted in expression levels comparable to or slightly higher than the individual mutations ($P <$

0.05), indicating that the regulatory function of ISS5 is likely not due to combinatorial recognition of motifs.

In contrast, analysis of a two-zone ISS sequence, ISS8, demonstrated that enriched hexamers can exhibit combinatorial control over ISS activity (Figure 2.6c). The extended ISS8 sequence contains 8 enriched hexamers resembling preferred binding sites for the PTB and hnRNP L proteins and a novel element that overlaps regions 1 and 2. A similar analysis of ISS8 shows that the individual mutations within each zone disrupt silencer activity to levels below the NMD control ($P < 0.05$), resulting in an ~18% decrease in activity (Figure 2.6c). Mutations to both regions in combination result in an ~25% decrease in silencer activity, suggesting that the hexamer regions work together to effect silencer activity ($P < 0.005$). Therefore, the 'zones of silencing' in our recovered ISRE sequences consisting of enriched hexamers exhibit regulatory function independently and in combination with other zones, but the effects are context dependent and may depend on the specific trans-acting factors involved.

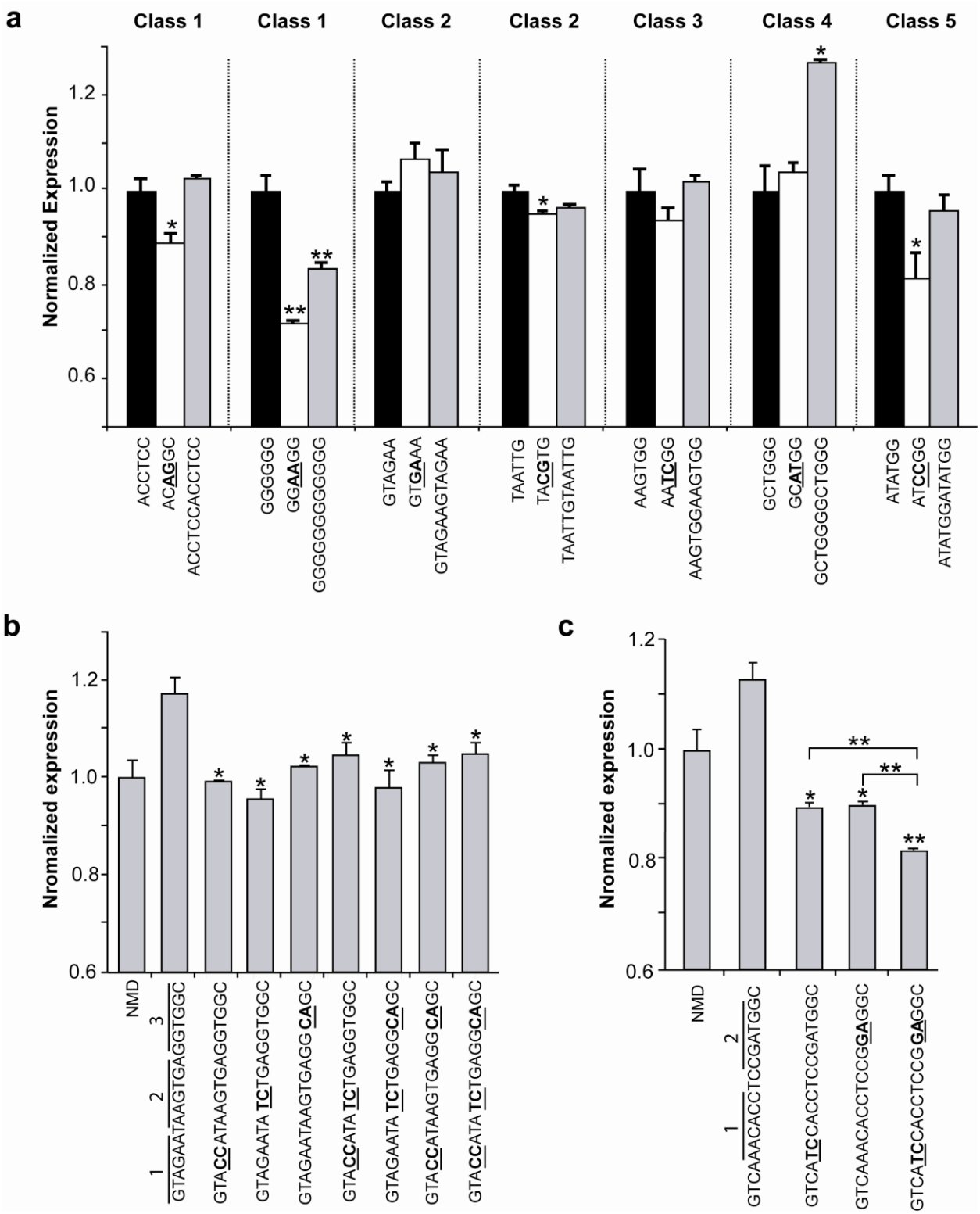


Figure 2.6. Enriched ISRE hexamers demonstrate silencer activity. **(a)** Individual hexamer analysis supports the silencing activity of enriched hexamer sequences.

Representative hexamers from each class of GCCS clusters and corresponding mutant and duplicate sequences were characterized in transient transfection assays in HEK-293 cells. For all reported data, silencing activity was assessed by flow cytometry analysis, where the mean GFP levels from two independent experiments were normalized to the wild-type hexamer construct. Normalized expression and average error are reported. *P* -values derived from the Student's t-test are as follows: * $P < 0.05$ and ** $P < 0.01$. **(b)** Mutational analysis of an ISS sequence supports the silencing activity of individual hexamer regions. The combined and individual activity of hexamer regions within the context of an ISS sequence was examined by introducing 2 point mutations into 3 regions, in combination and separately into all 3 hexamer regions of ISS5. **(c)** Mutational analysis of an ISS sequence supports the silencing activity of combined hexamer regions. The combined and individual activity of hexamer regions within the context of an ISS sequence was examined by introducing 2 point mutations separately and in combination into 2 hexamer regions of ISS8.

2.2.10. Splicing factor depletion influences ISRE regulated splicing in vivo

Many of the sequence classes identified by GCCS analysis resemble known or predicted binding sites for trans-acting splicing regulators. To validate the functional significance of the GCCS-identified sequence classes and uncover the associated trans-acting factors or SRNs, we screened a panel of siRNAs targeting known splicing regulators (hnRNP H, hnRNP A1, PTB, CUG-BP1, and SF2/ASF) for resulting effects on the splicing patterns of selected hexamers in stable cell lines. RNAi-mediated silencing of each gene resulted in a substantial reduction ($\geq$ 70%) of the targeted protein (Figure 2.7a)

and displayed minimal effects on the other splicing factors examined (Figure S2.10a). Hexamers from classes 1 (ACCTCC, GGGGGG), 2 (GTAGAA), 4 (GCTGGG) and 5 (ATATGG), which harbor potential binding sites for the selected trans-acting factors, and a random insert control were subjected to the RNAi-based screen using a mini-gene lacking a PTC to avoid any siRNA-mediated effects on the NMD pathway.

We analyzed the splicing patterns of the hexamer and control constructs through qRT-PCR analysis. In the presence of the mock siRNA, four of the hexamers exhibit silencing activity, a higher ratio of exon exclusion to inclusion, relative to the GFP-SMN1 control (Figure 2.7b and Figure S2.10b). In contrast, one hexamer (GGGGGG) exhibits enhancer activity in the presence of the mock siRNA. Splicing of constructs containing the GCCS hexamers were significantly affected by the depletion of at least one, and in some cases multiple, trans-acting factors (Figure 2.7c and Figure S2.10c). In contrast, siRNA-mediated depletion of the selected trans-acting factors had statistically insignificant effects ($P >> 0.05$) on the splicing pattern of the GFP-SMN1 control.

The splicing pattern of three hexamer constructs exhibited significant changes in response to the depletion of one of the trans-acting factors. The GGGGGG enhancer motif matches the hnRNP F/H binding site[23]. Depletion of hnRNP H leads to a 2.3-fold increase in exon exclusion, demonstrating that this factor enhances the recognition of the 3' ss, most likely through direct binding to the hexamer. The ACCTCC motif is similar to the CT-rich PTB binding site[24], but depletion of PTB leads to only a marginal decrease in exon exclusion of the construct. However, depletion of CUG-BP1 leads to a significant two fold decrease in exon exclusion. Although PTB has been shown to act antagonistically to CELF proteins[12], it is unlikely that CUG-BP1, which binds CTG and

GT-rich motifs, directly binds to the ACCTCC hexamer, suggesting it may be recruited through interactions with other regulatory proteins. Analysis of the splicing of a novel motif, ATATGG, reveals that depletion of hnRNP A1 leads to an increase in exon exclusion levels (~2.3-fold). The hexamer and flanking regions contain a GGG motif that may be a weak binding site for hnRNP A1. However, any direct binding of hnRNP A1 likely competes with other regulatory factors since its depletion leads to an increase in exon exclusion. Alternatively, modulation of hnRNP A1 levels may affect the levels of other trans-acting factors that play a role in the splicing regulatory effect of the hexamer.

Two hexamers constructs exhibit significant changes in their splicing pattern in response to depletion of multiple factors. The GTAGAA motif closely resembles the SF2/ASF SELEX-derived binding site (GAAGAA)[26], although the hexamer and flanking regions contain two GT repeats, which may serve as binding sites for CUG-BP1, and a TAGA motif, which may be a weak binding site for hnRNP A1. Depletion of hnRNP H, hnRNP A1, CUG-BP1 and SF2/ASF led to a 2.5-fold or greater reduction in exon exclusion levels for the construct. One possible mechanism is that SF2/ASF, CUG-BP1 and hnRNP A1 directly compete for binding to the GTAGAA hexamer and that hnRNP H acts positively in the recruitment of these factors. Both hnRNP H and CUG-BP1 have been shown to form an RNA-dependent suppressor splicing complex[46], suggesting that many of these factors may be involved in an inhibitory splicing complex that aids in the recruitment of a factor that directly binds to the transcript. The GCTGGG motif and flanking regions contain GT and TG dinucleotides and a CTG element that resemble CUG-BP1 binding sites. Depletion of CUG-BP1 results in a four fold decrease in exon exclusion of the construct. Depletion of PTB and SF2/ASF also cause significant

decreases in exon exclusion, although the preferred binding sites of these factors don't resemble any motifs within the GCTGGG hexamer and flanking regions. The results suggest that CUG-BP1 may be directly involved in binding to the GCTGGG hexamer, while PTB or SF2/ASF may be recruited by CUG-BP1 or other trans-acting factors.

2.2.11. Splicing factor depletion alters splicing of endogenous genes containing ISREs

Our genome-wide analysis revealed that selected ISREs are enriched in the intronic regions flanking constitutively and alternatively spliced endogenous genes (Figure 2.4b). To determine the biological significance of these associations, we analyzed the splicing patterns of 10 alternatively spliced endogenous genes containing an intronic hexamer under depletion of trans-acting splicing factors (Figure 2.7d and Table S2.10). Each target gene was analyzed through qRT-PCR in the presence of a mock siRNA and a siRNA targeting the splicing factor that showed the most significant effect on the splicing pattern of each hexamer in the SMN1 mini-gene depletion studies (Figure 2.7c). We observed significant changes in the alternative splicing patterns of all targeted genes upon splicing factor depletion ($P < .05$), where 7 of the 10 genes showed increased exon inclusion supporting the ISS activity of the selected hexamers. In contrast, 3 genes (*RREB1*, *CAMK2G*, and *HNRNPA2B1*) displayed higher levels of exon exclusion, indicating that hexamers GTAGAA, GCTGGG, and ATATGG can function as ISEs within endogenous genes. The significant changes in splicing of the synthetic SMN1 mini-gene and endogenous genes containing selected hexamers upon splicing factor depletion support the functional role of SPLICE identified ISREs through known trans-acting factors. These studies further highlight the context dependent nature of ISRE

function, where a given sequence can display enhancer and silencer functions in different transcripts.

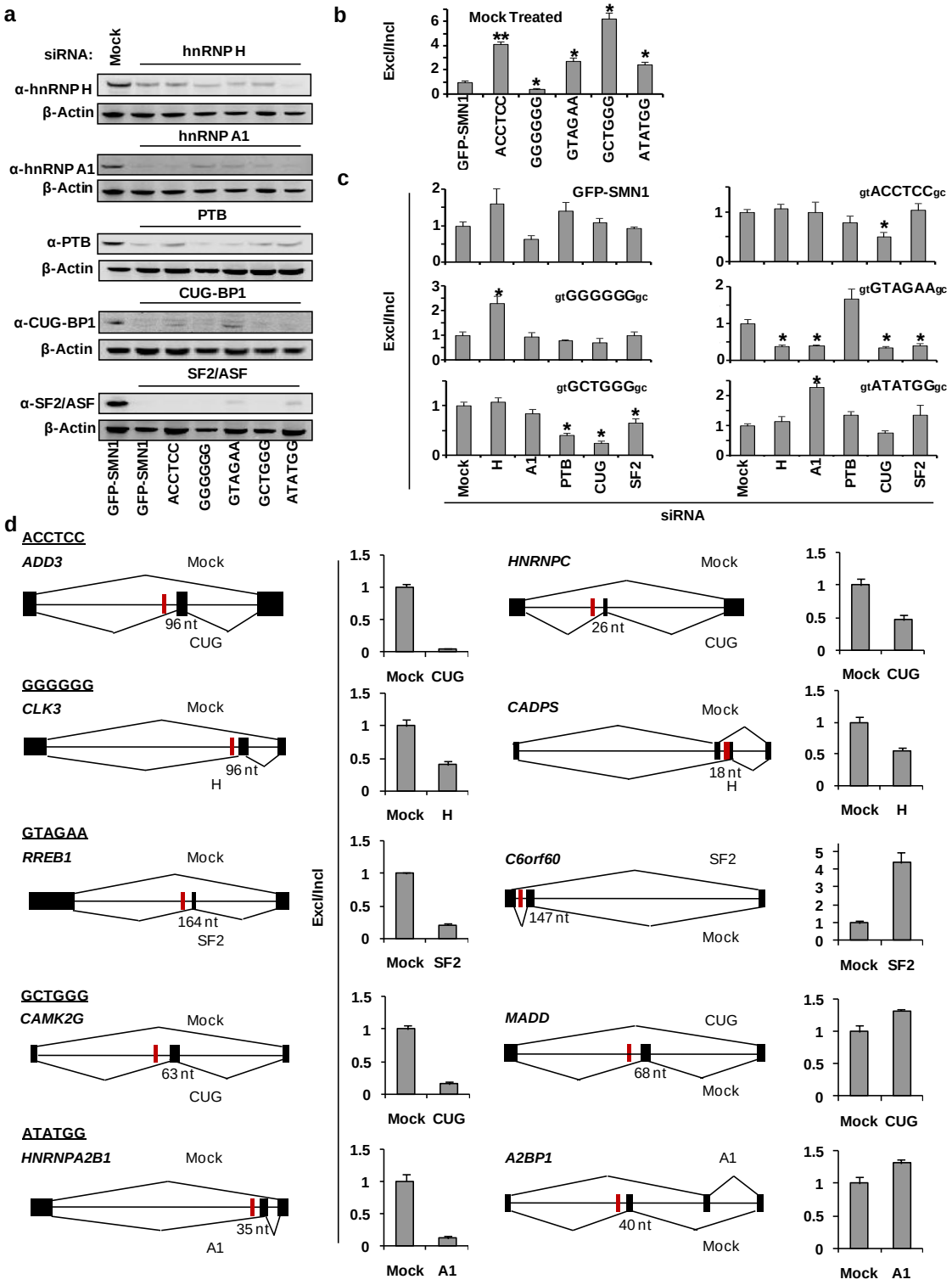


Figure 2.7. The effects of *in vivo* depletion of splicing factors on ISRE regulated splicing patterns of synthetic and endogenous genes. **(a)** Western blot analysis of total cell lysates prepared from the ISRE hexamer and GFP-SMN1 control cell lines treated with siRNAs targeted to trans-acting splicing factors and a mock siRNA negative control. β-Actin was used as a loading control for all blots. The results of the GFP-SMN1 mock treated lysate is representative of all mock treated cell lines. **(b)** qRT-PCR analysis of the mock treated ISRE hexamer and GFP-SMN1 control cell lines with primer sets specific for exon 7 included and excluded products. Expression levels of duplicate PCR samples were normalized to the levels of *HPRT*. Data is reported as the ratio of the mean expression of the exon excluded isoform to the exon included isoform normalized to the ratio for the GFP-SMN1 control ± the average error. **(c)** qRT-PCR analysis of the siRNA treated ISRE hexamer and GFP-SMN1 control cell lines with primer sets specific for exon 7 included and excluded products. Data is reported as the ratio of the mean expression of the exon excluded isoform to the exon included isoform normalized to the ratio for the mock siRNA treated cell line control ± the average error. *P* -values derived from the Student's t-test are as follows: * $P < 0.05$ and ** $P < 0.01$. **(d)** qRT-PCR analysis of the siRNA treated GFP-SMN1 control cell lines with primer sets specific for exon included and excluded products of 10 endogenous genes. The splicing patterns of each gene are diagrammed where black bars represent exons and red bars represent the location of conserved ISRE hexamer motifs. Data is reported as the ratio of the mean expression of the exon excluded isoform to the exon included isoform normalized to the ratio for the mock siRNA treated cell line control ± the average error.

2.2.12. Models of ISRE mediated regulation of alternative splicing

Based on the location of our ISREs in the mini-gene construct, these elements likely exhibit their regulatory activity through interacting with trans-acting factors that enhance or inhibit the binding of general splicing factors, such as U2AF65 or the U2 snRNP complex, at the 3' ss. Based on our studies, we propose three models for ISRE regulation of alternative splicing based on direct or indirect interactions with trans-acting factors in the SRN. The first model suggests that direct binding of a specific factor or the competitive binding of multiple factors to the ISRE sequence plays a role in splicing regulation and is supported by results from the class 1 (GGGGGG), 2 (GTAGAA) and 5 (ATATGG) ISREs (Figure 2.8a). This model is further supported by a recent study describing the juxtaposition of an ESS and ESE that results in hnRNP H and F competing for binding with SF2/ASF[47]. A second model is based on results from the class 1 (ACCTCC), 2 (GTAGAA), and 4 (GCTGGG) ISREs and proposes that direct binding of a specific factor is involved in the extensive recruitment of or is itself recruited by several other regulatory factors, thereby resulting in a recruitment pathway for ISRE regulation (Figure 2.8b). Previous work has suggested that splicing factors may be components of larger regulatory complexes in which binding selectivity is dictated by protein-protein interactions[47,48]. The third model is supported by results from the class 5 (ATATGG) ISRE and is based on an interaction with an agonist factor, where the level of that factor may affect the levels of other splicing factors that play a role in the regulation of the ISRE (Figure 2.8c).

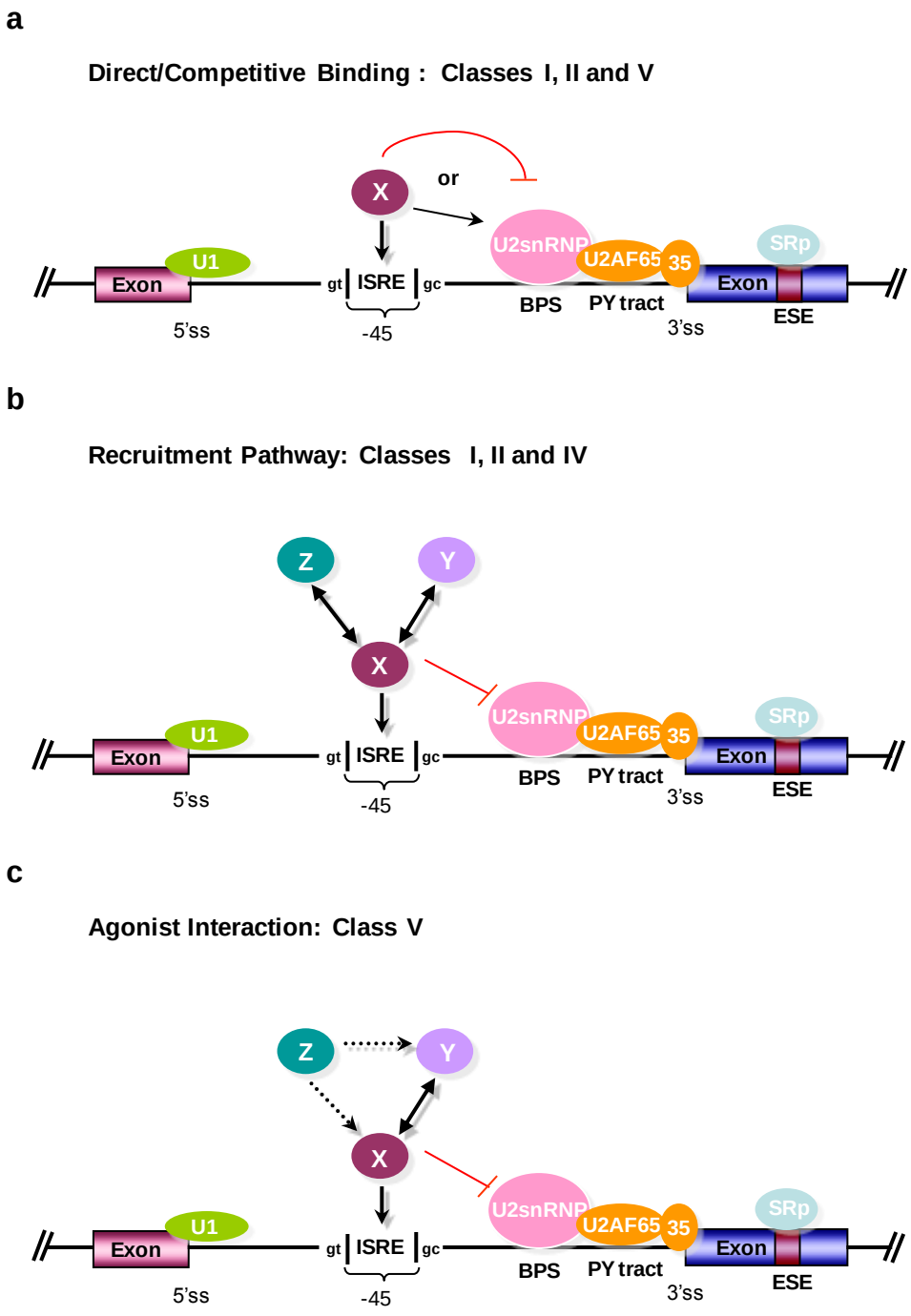


Figure 2.8. Three models for ISRE regulation. **(a)** A direct binding / competition model for ISRE regulation. Factor X, represents a protein that directly binds to the ISREs, factors Y and Z represent proteins that do not directly bind to the ISRE sequence. We present our ISRE mechanistic models based on a simplified model of ESE-dependent 3'

ss activation by SR proteins involving the recruitment of U2AF65 to the PY tract and subsequent recruitment of the U2 snRNP complex. Selected ISEs stabilize the interactions of general splicing factors to the 3' ss, whereas ISSs destabilize this process. In the direct binding / competition model, factor X binds to the selected ISRE and either enhances (black) or suppresses (red) recognition of the 3' ss. **(b)** A recruitment pathway model for ISRE regulation. This model is based on the direct binding of factor X to the ISRE leading to the recruitment of additional regulatory factors Y and Z. **(c)** An agonist interaction model for ISRE regulation. This model is based on an interaction between the splicing factors and an agonist factor Z, where the level of factor Z may affect the levels of the splicing factors (X and Y) that play a role in the regulation of the ISRE.

2.3. Discussion

Significant advances in our understanding of the mechanisms that guide splice site selection and the distributions of regulatory elements has aided the formulation of an early version of a 'splicing code'[9]. However, currently missing from this draft is a thorough understanding of the sequence characteristics and function of ISREs. The ISREs obtained from our *in vivo* selection provide a diverse composition of sequences that correlate to an array of splicing activities, which enable tuning of alternative splicing, and reveal motifs that resemble binding sites for many known and novel trans-activating factors that are enriched in introns of naturally-occurring spliced genes throughout the human genome. The identified motifs offer a rich dataset to expand the splicing code, determine the extent of single nucleotide polymorphisms (SNPs) that modulate splicing

through ISREs and refine bioinformatic search algorithms for genome-wide identification of intronic regulators, which will facilitate the diagnosis and treatment of disease.

The majority of the tested ISREs retained function when tested in a second cell type, but not a second transcript. These results support that although differences in cellular constituents can lead to differential splicing patterns (i.e., tissue specific splicing), a dominant factor guiding function is likely the immediate transcript context in which these cellular factors bind. Therefore, refinements of the splicing code will likely benefit from identification of co-regulatory sequences that may be identified in functional screens examining pair-wise or combinatorial motifs. Such combinatorial motifs were identified by SPLICE and experimental characterization supports the properties of context dependent and combinatorial regulation. The results from our RNAi silencing study highlight the complexity of the SRNs associated with ISRE function, suggesting a role for multiple splicing factors influencing regulation at a single motif and supporting models for ISRE function in which direct binding, recruitment or agonist interactions with upstream factors interfere or enhance the recruitment of the basal splicing complex. The splicing factor depletion studies also provided experimental validation of ISRE regulation of endogenous alternatively spliced transcripts further supporting their biological significance and context dependent function. Our work sets the stage for larger-scale characterization studies of the identified ISREs and associated trans-acting factors, which will further elucidate ISRE regulatory activity and mechanism, including the role of combinatorial control and sequence context in the function of these elements. Our results provide the first large scale analysis of ISREs *in vivo* and highlight that an

understanding of the complex interplay between multiple factors at a single binding site is necessary to further define the splicing code.

2.4. Materials and Methods

2.4.1. Base SPLICE constructs

Plasmids were constructed using standard molecular biology techniques[49]. All enzymes, including restriction enzymes and ligases, were obtained through New England Biolabs unless otherwise noted. DNA synthesis was performed by Integrated DNA Technologies, Inc. Ligation products were electroporated into *E. coli* DH10B (Invitrogen) using a GenePulser XP system (BioRAD), and clones verified through colony PCR and restriction mapping. All cloned constructs were sequence verified through Laragen. Primer sequences and plasmid descriptions are available in Tables S2.7 and S2.8, respectively.

The GFP-SMN1 mini-gene fusion construct was constructed through a PCR assembly and site-directed mutagenesis strategy. A region encompassing exons 6 through 8 of the *SMN1* mini-gene was amplified through PCR from template pCISMNxΔ6-wt[50] with primers Ex6 and Ex8 and PfuUltra high-fidelity DNA polymerase (Stratagene). The *GFP* gene was amplified from the template pKW430[51] with primers GFP1 and GFP2. The GFP-SMN1 gene fusion was constructed by performing PCR assembly on the resulting purified products (Qiagen) as templates and flanking primers GFP1 and Ex8. The resulting gene fusion product was digested with Xho I and Kpn I and ligated into the corresponding restriction sites of the mammalian expression vector pcDNA5/FRT (Invitrogen), resulting in the positive control vector pCS238. A PTC (TAA at position +1

in exon 7) and ISRE insertion sites Eco RV/Cla I in intron 6 (positions -62 and -51 from 3' ss of exon 7, respectively) and Bam HI/Pml I in intron 7 (positions +43 and +59 from 5' ss of exon 7, respectively) were introduced by site-directed mutagenesis with primers ECmutF1/ECmutR1, PmlImutF/PmlImutR, and BamHImutF/BamHImutR using a Quickchange II Kit (Stratagene) according to manufacturer's instructions, resulting in the base NMD reporter construct pCS516. ISRE sequences were digested and ligated into the Eco RV and Cla I restriction sites within intron 6 of the base NMD construct.

The GFP-BRCA1 mini-gene fusion was constructed through a PCR assembly and site-directed mutagenesis strategy. A portion of the wild-type human *BRCA1* gene was amplified from HEK-293 genomic DNA as previously described[45] using reported primers P2, P3, P4, and P5 with the exception of the forward primer for exon 17 (Ex17) and the reverse primer for exon 19 (Ex19). The resulting wild-type BRCA1 mini-gene contains shortened introns and wild-type exons 17, 18, and 19. The *GFP* gene was amplified from template pKW430[51] with primers GFP1 and GFP3. The GFP-BRCA1 gene fusion was PCR assembled using primers GFP1 and Ex19, digested with Xho I and Kpn I and ligated into the corresponding restriction sites of pcDNA5/FRT, resulting in the GFP-BRCA1 positive control construct pCS990. A PTC (TAA at position +3 in exon 18) and ISRE insertion sites Eco RV/Cla I in intron 17 (positions -61 and -50 from 3' ss of exon 18, respectively) were introduced as described above with primers ECmutF2 and ECmutR2, resulting in the base BRCA1-NMD reporter construct pCS1008. ISS sequences were cloned into intron 17 of the base NMD construct as described above.

2.4.2. Cell culture, transfections, stable cell lines and FACS

HEK293 FLP-In cells (Invitrogen) were cultured in D-MEM supplemented with 10% fetal bovine serum (FBS) and 100 μg/ml Zeocin at 37ºC in 5% CO_2. HeLa cells were cultured in MEM media supplemented with 10% FBS. Transfections for all cell lines were carried out with Fugene (Roche) according to the manufacturer's instructions. All cell culture media was obtained from Invitrogen.

HEK-293 FLP-In stable cell lines were generated by co-transfection of the appropriate SMN1 mini-gene construct with a plasmid encoding the Flp recombinase (pOG44) in growth medium without Zeocin according to the manufacturer's instructions (Invitrogen). The library stable selections were carried out in 225 cm^2 flasks containing ~$4x10^7$ HEK-293 FLP-In cells where 37 μg of pOG44 and 3.7 μg of the SMN1 ISRE plasmid library (10:1 ratio) were co-transfected. Fresh medium was added to the cells 24 h after transfection. The cells were expanded by a 1:4 dilution and Hygromycin B was added to a final concentration of 200 μg/ml 48 h after transfection. In total, ~450,000 stable transformants were pooled from 60 transfections. Clones were harvested by trypsinization, pooled and analyzed on a FACS Aria (Becton Dickinson Immunocytometry Systems) 10–14 days after transfection. GFP fluorescence was excited at 488 nm and emission was measured with a FITC filter. Detailed sorting procedures are presented in Figure S2.2. In the first screening round, positive cells were bulk sorted into 96-well plates, where no more than 25,000 cells were collected into a single well. After ~1–2 weeks of growth, positives were re-sorted into 3 fractions (A, B, and C) based on varying fluorescence levels (Figure S2.2b). Positive cells were bulk sorted in the second screening round as described for the first round. Total genomic DNA from bulk sorted

cells was purified using the DNeasy Blood & Tissue total DNA purification kit (Qiagen) according to the manufacturer's instructions and used as a template for amplification of recovered ISRE sequences with primers Lib3 and Lib4. The recovered ISRE fragments were then digested, ligated into the corresponding sites of pCS516 and sequenced verified by Functional Biosciences, Inc.

For transient transfection studies, HEK293 and HeLa cells were seeded in 12-well plates at ~$5x10^4$ cells per well 16 to 24 h prior to transfection. Cell lines were transfected with 625 ng of the appropriate GFP-SMN1 or GFP-BRCA1 mini-gene constructs. The cells were harvested by trypsinization, pooled and analyzed on a FACS Aria 48 h after transfection. Experiments were carried out on different days and transfections were completed in duplicate, where the mean GFP fluorescence of the transfected population and the average error between samples is reported. A comparison of FACS gating procedures used in transient and stable assays is presented in Figure S2.8. Cell lines harboring PTC-containing transcripts tend to increase in fluorescence at higher passage numbers (>10), whereas the GFP-SMN1 cell line does not. As such, the fluorescence levels of the enriched cell populations at the time of sorting (Figure 2.1b) do not directly match the expression levels for individual recreated clones (Figure 2.5b). To minimize differences in expression due to such instabilities, the analysis of all stable cell lines was performed at an identical, early passage.

2.4.3. qRT-PCR analysis

Total cellular RNA was purified from stably transfected HEK-293 Flp-In cells using GenElute mammalian total RNA purification kit (Sigma) according to the manufacturer's instructions, followed by DNase treatment (Invitrogen). cDNA was synthesized using Superscript III reverse transcriptase (Invitrogen) according to the manufacturer's instructions. qRT-PCR analysis was performed using isoform-specific primers (Tables S2.9 and S2.10). Expression levels of duplicate PCR samples were normalized to the levels of *HPRT* (Hypoxanthine-guanine phosphoribosyltransferase**)**. Fold expression data is reported as the mean expression for each sample divided by the mean NMD expression value $\pm$ the average error.

2.4.4. siRNA mediated silencing of trans-acting splicing factors

siRNAs targeting hnRNP H, hnRNP A1, PTB, CUG-BP1, and SF2/ASF and a mock control siRNA were purchased from Dharmacon and are listed in Table S2.11. All duplexes were resuspended in 1X PBS to a concentration of 20 µM. Briefly, HEK-293 FLP-In cells were plated at ~2 X 10^5 cells per well in 6-well plates. After 24 h, the cells were transfected with individual siRNA duplexes to a final concentration of 50 nM using Lipofectamine RNAiMAX (Invitrogen) according to the manufacturer's instructions. Cells were collected for RNA isolation and western blotting 48 h after transfection.

2.4.5. Western blot analysis

Whole-cell extracts were prepared from harvested cells using M-PER mammalian protein extraction reagent (Pierce) and equal amounts of protein (50 µg) were resolved on

4-12 % SDS-PAGE gels (Invitrogen) and transferred onto Protran nitrocellulose membranes (Whatman) using the Trans-Blot SD semi-dry transfer cell (BioRad). After blocking with 5% BSA in TBST, the membranes were incubated with the specified antibodies overnight at 4ºC. After incubation, the membranes were washed with TBST and then incubated with the corresponding secondary antibody conjugated with HRP. Signals were detected using the ECL western blotting substrate (Thermo Scientific) according to the manufacturer's protocol. The primary antibody dilutions were 1:500 for goat anti-hnRNP H (N-16), 1:1000 for goat anti-Actin (I-19), 1:200 for goat anti-hnRNP A1 (Y-15), 1:200 for mouse anti-PTB (SH54), 1:200 for mouse anti-SF2/ASF (96) and 1:200 for mouse anti-CUG-BP1 (3B1). The secondary antibody dilutions were 1:10,000 for donkey anti-goat IgG-HRP (sc-2020) and 1:10,000 for goat anti-mouse IgG-HRP (sc-2005). All of the antibodies were purchased from Santa Cruz Biotechnology Inc. The relative band intensities were measured by densitometry analyses using Quantity One (BioRad).

2.4.6. Discovery of sequence motifs enriched in ISRE sequences

A sliding-window count of all n-mers (4–6-nt) within the nonredundant sample set of 125 sequences was performed. Two nucleotides flanking the 5' and 3' ends of the random region were included to account for bias due to the constant sequences. A similar sliding-window count on a set of 450,000 computer generated sequences containing a uniformly random 15-nt region flanked by the same constant nucleotides was performed to calculate the maximum likelihood probabilities for expected occurrences (see

Methods). For both data sets the counts were transformed into probabilities and the enrichment was determined according to the binomial confidence interval method[21].

2.4.7. Overlap of ISRE sequences with known splicing regulatory elements

The set of pentamers enriched in the ISRE sequences were compared to previously compiled lists of ESEs[6,7], ESSs[7,9], and ISEs[16] (http://www.snl.salk.edu/~geneyeo/stuff/ papers/supplementary/ISRE/). These datasets were originally reported as hexamers, such that pentameric equivalents were created by extracting all pentamers that occurred at least one time within the original datasets. The ISRE enriched pentamers were also compared to ISREs[33], CISs[21], and motifs enriched upstream of weak PY tracts[37]. Both of these datasets were composed of various length n-mers and were adjusted to pentameric equivalents to achieve independent sampling by extracting all pentamers that occurred at least once. Lastly, the ISRE pentamers were compared to conserved pentamers enriched in intronic regions of exons excluded in NP cells[36]. Since these were reported as pentamers no adjustments were necessary.

The significance of overlap between datasets was determined using a 2x2 Chi-test of association. Each pentamer was classified according to which of the four ways it could be distributed: (1) in both sets, (2) in set A but not set B, (3) not in set A but in set B, (4) in neither set. The counts for each distribution were then used to calculate the likelihood that this arrangement could have occurred randomly (according to the Chi-distribution with 1-degree of freedom).

2.4.8. Statistical Analysis

Data are expressed as normalized or fold expression ± average error where applicable. Student's *t*-test and Anova analyses were performed using Microsoft Excel. $P < .05$ were taken to be significant.

2.4.9. ISRE library and ISS controls construction

A random 15-nt ISRE library was generated through PCR using a 47-nt template (ISStemp) with primers Lib1 and Lib2. The library PCR was conducted for 12 cycles in a 100 µl reaction containing 20 pmol DNA template, 300 pmol each Lib1 and Lib2, 200 µM each dNTPs, 1.6 mM $MgCl_2$, and 10 U *Taq* DNA polymerase (Roche). ISS and negative controls were constructed by replacing the random 15-nt region in the above template with previously characterized ISS sequences and scrambled sequences, respectively. The resulting ISRE library, ISS control, and negative control fragments were digested with Eco RV and Cla I and ligated into the corresponding restriction sites within intron 6 of pCS516. Control ISS sequences correspond to previously characterized binding sites for U2AF65 (TTTTTTTTTCCTTTTTTTTTCCTTTT; pCS668)[39]; hnRNP H (TAAATGTGGGACCTAGA; pCS669)[38], PTB(1) (TAGCATCAGCCTGG TGCCTACCTTCGGCCCC; pCS670)[39]; PTB(2) (TCTTCTCTTCTCTTCTCTTC; pCS667)[40]. In addition, 15 scrambled sequences were examined in place of the 15-nt random region as negative control constructs. The base random 15-nt sequence ACCTCAGGCTCTGAA (pCS517) was subsequently used as the negative control for all FACS experiments.

2.4.10. Quantitative RT-PCR analysis

Total cellular RNA was purified from stably transfected HEK-293 Flp-In cells using GenElute mammalian total RNA purification kit (Sigma) according to the manufacturer's instructions, followed by DNase treatment (Invitrogen). cDNA was synthesized using a gene-specific primer for the pcDNA5/FRT vector (SMN1cDNA) and Superscript III reverse transcriptase (Invitrogen) according to the manufacturer's instructions. qRT-PCR analysis was performed using isoform-specific primers (Tables S2.9 and S2.10) where each reaction contained 1 µL template cDNA, 10 pmol of each primer and 1X iQ SYBR green supermix (BioRAD) to a final volume of 25 µL. Reactions were carried out using a iCycler iQ system (BioRAD) for 30 cycles (95ºC for 15 s, 72ºC for 30 s). The purity of the PCR products was determined by melt curve analysis. Data analysis was completed using the iCycler IQ system software v.3.1.7050 (BioRAD). Isoform-specific relative expression was calculated using the ΔCt (change in cycling threshold) method[52]. Expression levels were normalized to the levels of *HPRT* (Hypoxanthine-guanine phosphoribosyltransferase**)**. Fold expression data is reported as the mean expression for each sample divided by the mean NMD expression value $\pm$ the average error.

2.4.11. Discovery of sequence motifs enriched in ISRE sequences

Sequence motifs were constructed from the significantly enriched ($P < 0.1$) n-mers using the graph clustering method and software (GCCS)[21] with the following parameters: minimum cluster size = 4, rounds of clustering = 5, minimum substring

length = 5 (rounds 1–3) and 4 (rounds 4 and 5). GCCS uses the MCL algorithm[53,54] to find clusters. Parameters were set as follows: MCL inflation = 3 and MCL scheme = 4. The other MCL parameters were set to default values. To validate the enrichment of ISRE motifs, the GCCS analysis was repeated using 5 sets of 125 random 15-mers with the same constant flanking bases. The average number of significantly enriched n-mers observed in the random samples (RS) was only 91 and each yielded an average of 11 clusters.

2.4.12. Hierarchical clustering

A distance matrix for ISRE sequences recovered from SPLICE was produced using the Jukes-Cantor method[55] in which the distance is defined by the maximum likelihood estimate of the number of nucleotide substitutions between two sequences (Matlab default method) (http://mathworks.com). The distance matrix was then used to cluster sequences using the standard average linkage hierarchical clustering implemented in Matlab. 15-nt clusters were defined by using a dissimilarity cutoff of 1.1 in the dendrogram. Sequences within each cluster were then aligned with ClustalX using default parameters[56].

2.4.13. RNA structural analysis

RNA secondary structure predictions were performed using RNAfold[57].

Acknowledgments

We thank R. Diamond and D. Perez (Caltech Cell Sorting Facility) for FACS assistance and expert technical advice, A. Krainer for providing the pCISMNxΔ6-wt construct.

References

1. Black, D.L. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72**, 291–336 (2003).

2. Blencowe, B.J. Alternative splicing: new insights from global analyses. *Cell* **126**, 37–47 (2006).

3. Wang, E.T. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–6 (2008).

4. Castle, J.C. et al. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nat Genet* **40**, 1416–25 (2008).

5. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**, 1413–5 (2008).

6. Fairbrother, W.G., Yeh, R.F., Sharp, P.A. & Burge, C.B. Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007–13 (2002).

7. Zhang, X.H. & Chasin, L.A. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev* **18**, 1241–50 (2004).

8. Coulter, L.R., Landree, M.A. & Cooper, T.A. Identification of a new class of exonic splicing enhancers by in vivo selection. *Mol Cell Biol* **17**, 2143–50 (1997).

9. Wang, Z. et al. Systematic identification and analysis of exonic splicing silencers. *Cell* **119**, 831–45 (2004).

10. Wang, Z. & Burge, C.B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *Rna* **14**, 802–13 (2008).

11. Wang, G.S. & Cooper, T.A. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8**, 749–61 (2007).

12. Charlet, B.N., Logan, P., Singh, G. & Cooper, T.A. Dynamic antagonism between ETR-3 and PTB regulates cell type-specific alternative splicing. *Mol Cell* **9**, 649–58 (2002).

13. Smith, C.W. & Valcarcel, J. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem Sci* **25**, 381–8 (2000).

14. Simard, M.J. & Chabot, B. SRp30c is a repressor of 3' splice site utilization. *Mol Cell Biol* **22**, 4001–10 (2002).

15. Matlin, A.J., Clark, F. & Smith, C.W. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol* **6**, 386–98 (2005).

16. Yeo, G., Hoon, S., Venkatesh, B. & Burge, C.B. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci U S A* **101**, 15700–5 (2004).

17. Venables, J.P. Downstream intronic splicing enhancers. *FEBS Lett* **581**, 4127–31 (2007).

18. Green, R.E. et al. Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. *Bioinformatics* **19 Suppl 1**, i118-21 (2003).

19. Ladd, A.N. & Cooper, T.A. Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol* **3**, (2002).

20. McCullough, A.J. & Berget, S.M. G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol Cell Biol* **17**, 4562–71 (1997).

21. Voelker, R.B. & Berglund, J.A. A comprehensive computational characterization of conserved mammalian intronic sequences reveals conserved motifs associated with constitutive and alternative splicing. *Genome Res* **17**, 1023–33 (2007).

22. Burd, C.G. & Dreyfuss, G. RNA binding specificity of hnRNP A1: significance of hnRNP A1 high-affinity binding sites in pre-mRNA splicing. *Embo J* **13**, 1197–204 (1994).

23. Markovtsov, V. et al. Cooperative assembly of an hnRNP complex induced by a tissue-specific homolog of polypyrimidine tract binding protein. *Mol Cell Biol* **20**, 7463–79 (2000).

24. Chan, R.C. & Black, D.L. Conserved intron elements repress splicing of a neuron-specific c-src exon in vitro. *Mol Cell Biol* **15**, 6377–85 (1995).

25. Hui, J., Stangl, K., Lane, W.S. & Bindereif, A. HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats. *Nat Struct Biol* **10**, 33–7 (2003).

26. Tacke, R. & Manley, J.L. The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. *Embo J* **14**, 3540–51 (1995).

27. Liu, H.X., Zhang, M. & Krainer, A.R. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev* **12**, 1998–2012 (1998).

28. Cavaloc, Y., Bourgeois, C.F., Kister, L. & Stevenin, J. The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. *Rna* **5**, 468–83 (1999).

29. Tacke, R., Tohyama, M., Ogawa, S. & Manley, J.L. Human Tra2 proteins are sequence-specific activators of pre-mRNA splicing. *Cell* **93**, 139–48 (1998).

30. Lim, L.P. & Burge, C.B. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A* **98**, 11193–8 (2001).

31. Nasim, M.T., Chernova, T.K., Chowdhury, H.M., Yue, B.G. & Eperon, I.C. HnRNP G and Tra2beta: opposite effects on splicing matched by antagonism in RNA binding. *Hum Mol Genet* **12**, 1337–48 (2003).

32. Pagani, F. et al. A new type of mutation causes a splicing defect in ATM. *Nat Genet* **30**, 426–9 (2002).

33. Yeo, G.W., Nostrand, E.L. & Liang, T.Y. Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS Genet* **3**, e85 (2007).

34. Marquis, J. et al. CUG-BP1/CELF1 requires UGU-rich sequences for high-affinity binding. *Biochem J* **400**, 291–301 (2006).

35. Hovhannisyan, R.H. & Carstens, R.P. Heterogeneous ribonucleoprotein m is a splicing regulatory protein that can enhance or silence splicing of alternatively spliced exons. *J Biol Chem* **282**, 36265–74 (2007).

36. Yeo, G.W. et al. Alternative splicing events identified in human embryonic stem cells and neural progenitors. *PLoS Comput Biol* **3**, 1951–67 (2007).

37. Murray, J.I., Voelker, R.B., Henscheid, K.L., Warf, M.B. & Berglund, J.A. Identification of motifs that function in the splicing of non-canonical introns. *Genome Biol* **9**, R97 (2008).

38. Chen, C.D., Kobayashi, R. & Helfman, D.M. Binding of hnRNP H to an exonic splicing silencer is involved in the regulation of alternative splicing of the rat beta-tropomyosin gene. *Genes Dev* **13**, 593–606 (1999).

39. Singh, R., Valcarcel, J. & Green, M.R. Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. *Science* **268**, 1173–6 (1995).

40. Perez, I., Lin, C.H., McAfee, J.G. & Patton, J.G. Mutation of PTB binding sites causes misregulation of alternative 3' splice site selection in vivo. *Rna* **3**, 764-78 (1997).

41. Buratti, E., Stuani, C., De Prato, G. & Baralle, F.E. SR protein-mediated inhibition of CFTR exon 9 inclusion: molecular characterization of the intronic splicing silencer. *Nucleic Acids Res* **35**, 4359–68 (2007).

42. Hanamura, A., Caceres, J.F., Mayeda, A., Franza, B.R., Jr. & Krainer, A.R. Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors. *Rna* **4**, 430–44 (1998).

43. Wollerton, M.C. et al. Differential alternative splicing activity of isoforms of polypyrimidine tract binding protein (PTB). *Rna* **7**, 819–32 (2001).

44. Pozzoli, U. & Sironi, M. Silencers regulate both constitutive and alternative splicing events in mammals. *Cell Mol Life Sci* **62**, 1579–604 (2005).

45. Liu, H.X., Cartegni, L., Zhang, M.Q. & Krainer, A.R. A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. *Nat Genet* **27**, 55–8 (2001).

46. Paul, S. et al. Interaction of muscleblind, CUG-BP1 and hnRNP H proteins in DM1-associated aberrant IR splicing. *Embo J* **25**, 4271–83 (2006).

47. Mauger, D.M., Lin, C. & Garcia-Blanco, M.A. hnRNP H and hnRNP F complex with Fox2 to silence fibroblast growth factor receptor 2 Exon IIIc. *Mol Cell Biol* **28**, 5403–19 (2008).

48. Venables, J.P. et al. Multiple and Specific mRNA Processing Targets for the Major Human hnRNP Proteins. *Mol Cell Biol* (2008).

49. Sambrook, J. & Russell, D.W. *Molecular Cloning: a laboratory manual*, (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2001).

50. Cartegni, L. & Krainer, A.R. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nat Genet* **30**, 377–84 (2002).

51. Stade, K., Ford, C.S., Guthrie, C. & Weis, K. Exportin 1 (Crm1p) is an essential nuclear export factor. *Cell* **90**, 1041–50 (1997).

52. Livak, K.J. & Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402–8 (2001).

53. Dongen, S. University of Utrecht (2000).

54. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575–84 (2002).

55. Jukes, T.a.C., CR. Evolution of protein molecules. in *Mammalian protein metabolism* 21–123 (Academic Press, New York, 1969).

56. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. & Higgins, D.G. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876–82 (1997).

57. Hofacker, I.L. et al. Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte f. Chemie* **125**, 167–188 (1994).

58. Cartegni, L., Hastings, M.L., Calarco, J.A., de Stanchina, E. & Krainer, A.R. Determinants of exon 7 splicing in the spinal muscular atrophy genes, SMN1 and SMN2. *Am J Hum Genet* **78**, 63–77 (2006).

59. Hentze, M.W. & Kulozik, A.E. A perfect message: RNA surveillance and nonsense-mediated decay. *Cell* **96**, 307–10 (1999).

60. Singh, R. & Valcarcel, J. Building specificity with nonspecific RNA-binding proteins. *Nat Struct Mol Biol* **12**, 645–53 (2005).

61. Buratti, E. & Baralle, F.E. Influence of RNA secondary structure on the pre-mRNA splicing process. *Mol Cell Biol* **24**, 10505–14 (2004).

62. Hiller, M., Zhang, Z., Backofen, R. & Stamm, S. Pre-mRNA Secondary Structures Influence Exon Recognition. *PLoS Genet* **3**, e204 (2007).

63. Chou, M.Y., Rooke, N., Turck, C.W. & Black, D.L. hnRNP H is a component of a splicing enhancer complex that activates a c-src alternative exon in neuronal cells. *Mol Cell Biol* **19**, 69–77 (1999).

64. Hutchison, S., LeBel, C., Blanchette, M. & Chabot, B. Distinct sets of adjacent heterogeneous nuclear ribonucleoprotein (hnRNP) A1/A2 binding sites control 5' splice site selection in the hnRNP A1 mRNA precursor. *J Biol Chem* **277**, 29745–52 (2002).

65. Miriami, E., Margalit, H. & Sperling, R. Conserved sequence elements associated with exon skipping. *Nucleic Acids Res* **31**, 1974–83 (2003).

66. Chou, M.Y., Underwood, J.G., Nikolic, J., Luu, M.H. & Black, D.L. Multisite RNA binding and release of polypyrimidine tract binding protein during the regulation of c-src neural-specific splicing. *Mol Cell* **5**, 949–57 (2000).

67. Itoh, H., Washio, T. & Tomita, M. Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes. *Rna* **10**, 1005–18 (2004).

68. Paronetto, M.P., Achsel, T., Massiello, A., Chalfant, C.E. & Sette, C. The RNA-binding protein Sam68 modulates the alternative splicing of Bcl-x. *J Cell Biol* **176**, 929–39 (2007).

69. Lim, L.P. & Burge, C.B. A computational analysis of sequence features involved in recognition of short introns. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 11193–11198 (2001).

Supplementary Information

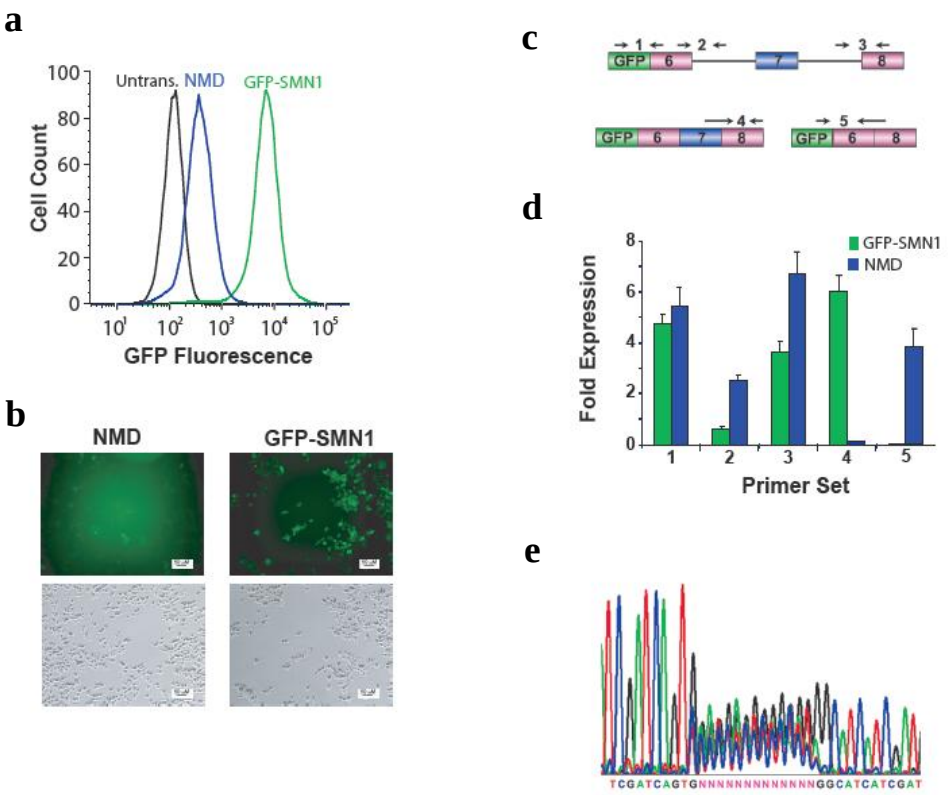


Figure S2.1. Fluorescence expression/analysis of SPLICE control constructs and library sequence bias. **(a)** Flow cytometry histograms of the stable cell lines expressing the control constructs. An untransfected HEK-293 FLP-In cell population (Untrans.) was also analyzed for reference. **(b)** Microscope images of stable cell lines expressing the negative (NMD) and positive (GFP-SMN1) control constructs. Upper panels: GFP fluorescence, lower panels: phase contrast images. **(c)** Schematic representing the relative locations of primer set binding for transcript isoform analysis by qRT-PCR. **(d)** qRT-PCR analysis of the NMD and GFP-SMN1 control cell lines supports decay of the PTC harboring isoform. The observed high level of exon 7 inclusion for the GFP-SMN1 control are in line with previous observations for the splicing of the SMN1 mini-gene[58]. Our transcript isoform analysis also reveals that levels of exon 7 exclusion are elevated in the NMD control compared to the GFP-SMN1 control, suggesting that the PTC may have a

secondary effect of increasing the levels of the exon excluded transcript. Such observations have been previously observed and may be the result of nonsense-associated altered splicing[59]. Expression levels were normalized to the levels of *HPRT* (Hypoxanthine-guanine phosphoribosyltransferase**)**. Data presented is the mean expression of duplicate PCR samples $\pm$ the average error. **(e)** DNA sequencing analysis of purified genomic DNA from HEK-293 cell lines harboring the library constructs. The transfected library exhibits a slight sequence bias at positions 1 and 15, but all other positions are free of bias. A comparison of sequencing results from two independent transfections supports that the sequence bias at these positions is minimal, indicating that the ISRE library represents an essentially random pool (data not shown).

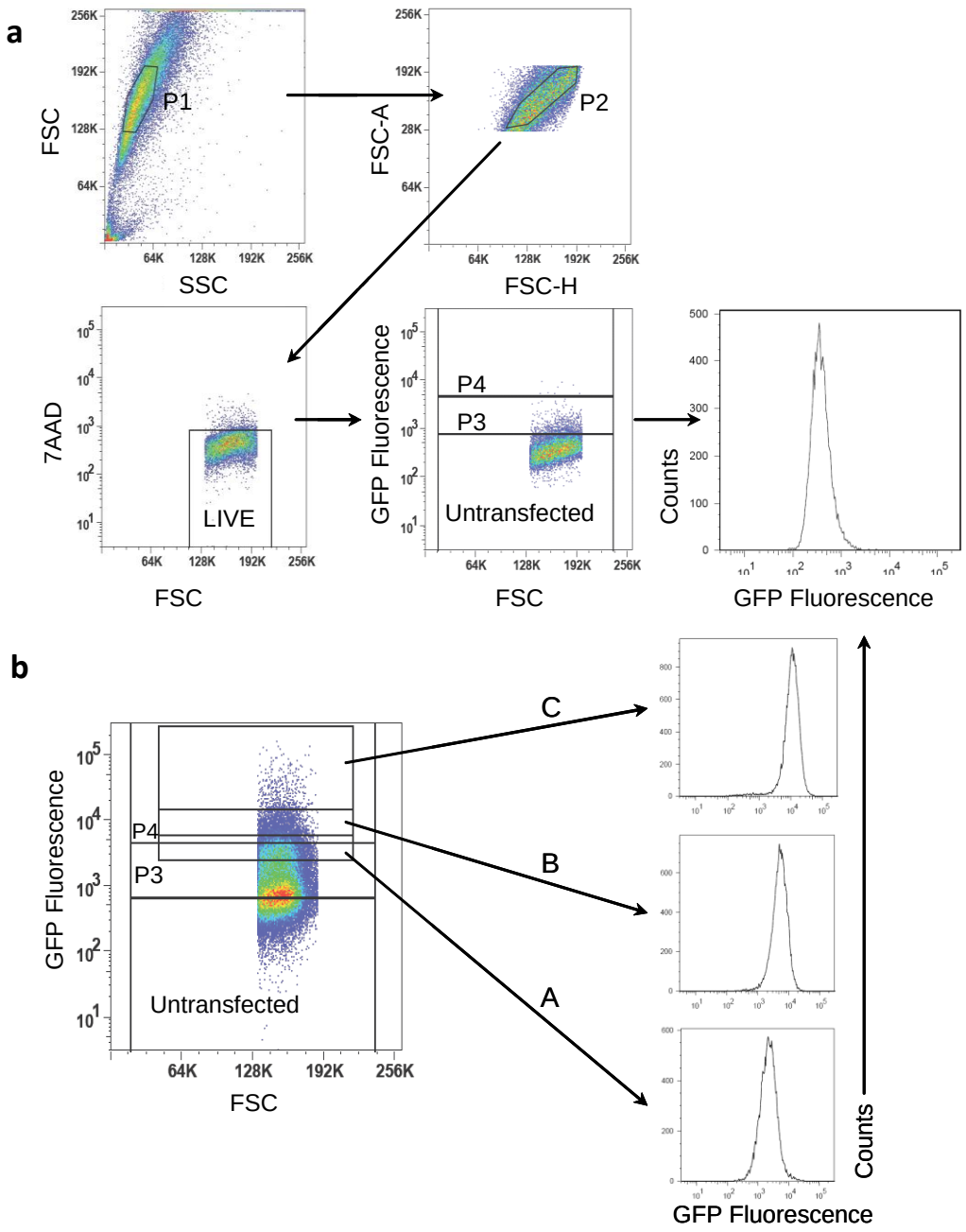


Figure S2.2. FACS analysis and ISRE library sorting scheme. **(a)** FACS analysis and gating procedure for all HEK-293 FLP-In cells. As an example, flow cytometry data from the stable NMD control is presented. Dot plots show initial gating of stable cells (P1), followed by P2 gating for cell uniformity (i.e., to remove cell aggregates) and finally the selection of live cells using 7-Amino-Actinomycin D (7AAD) staining. The P3 gate

reflects the GFP positive cells and the P4 gate is drawn to indicate the upper GFP fluorescence limit of the NMD control population. P4 was used as the gate for the selection of ISS positive cells. The histogram reports the intensity of GFP fluorescence in the NMD control population. **(b)** FACS analysis of ISS positive stable cells after one round of sorting. Cells from gates A, B, and C were sorted and the resulting histograms indicate the intensity of GFP fluorescence after 1 week in culture.

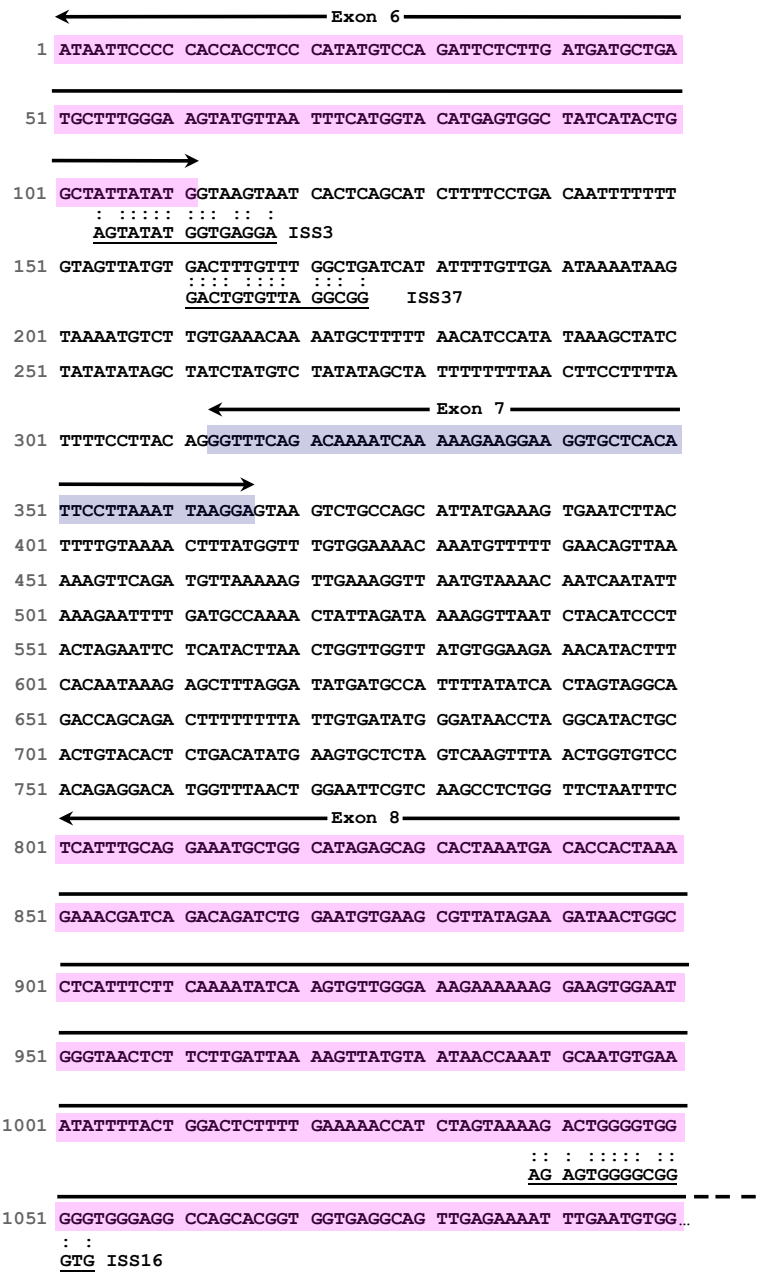


Figure S2.3. SPLICE-generated ISREs contain elements similar to sequences within SMN1. The sequence of the wild-type SMN1 mini-gene used in this study is shown. SPLICE-generated sequences similar to portions of the SMN1 mini-gene are underlined and labeled accordingly. Sequences ISS3, ISS16, and ISS37 are similar to sections of intron 6 and exon 8 and regions spanning the exonic and intronic portion of the 5' ss in

exon 6. Splicing repression may be a result of cooperative repressor binding to multiple silencer elements creating a ‘zone of silencing’ (at one site or at overlapping sites) between splice sites or nucleation that causes the looping out of RNA between repressor elements[15,60]. The sequence composition of ISS3, 16 and 37 supports a model of cooperative assembly of repressor elements to the SMN1 transcript in regulating splicing repression. The silencer activity of ISS3 and ISS16 has been confirmed (see Figure 2.5b).

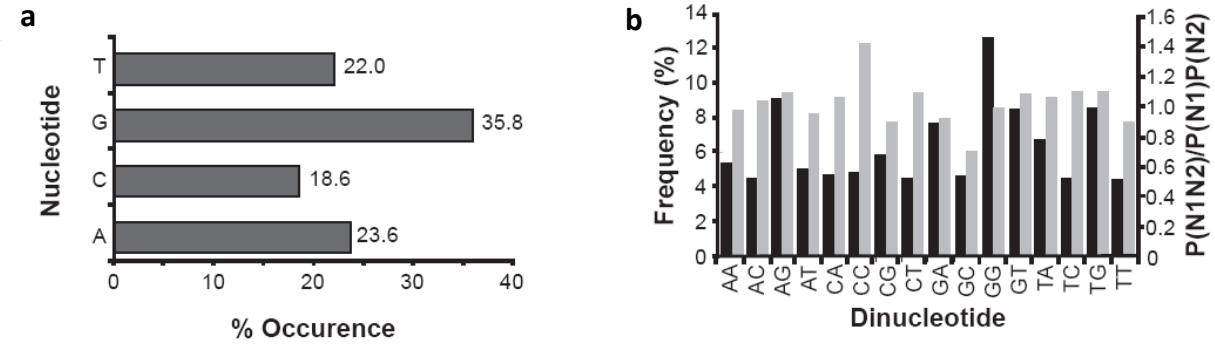


Figure S2.4. Overall compositional features of recovered ISRE sequences. **(a)** Overall nucleotide composition of the recovered ISREs. The occurrence value of each nucleotide prior to enrichment for ISRE activity was 25%. SPLICE-generated sequences have a higher level of G (35.8%) and reduced levels of T (22%), C (18.6%), and A (23.6%). **(b)** Dinucleotide frequency (black) and odds ratios (gray) in recovered ISRE. The occurrence of dinucleotides within all 125 ISRE sequences was calculated (black). The odds ratio for each dinucleotide was determined by dividing the probably of a dinucleotide occurrence within the 125 SPLICE selected sequences by the probability of an individual nucleotide occurrence within selected sequences (P(N1N2)/P(N1)P(N2)) (gray). The dinucleotide CC is overrepresented in the ISRE dataset, while others, such as AC, AG, CA, GT, TA, TC, and TG, are only slightly enriched.

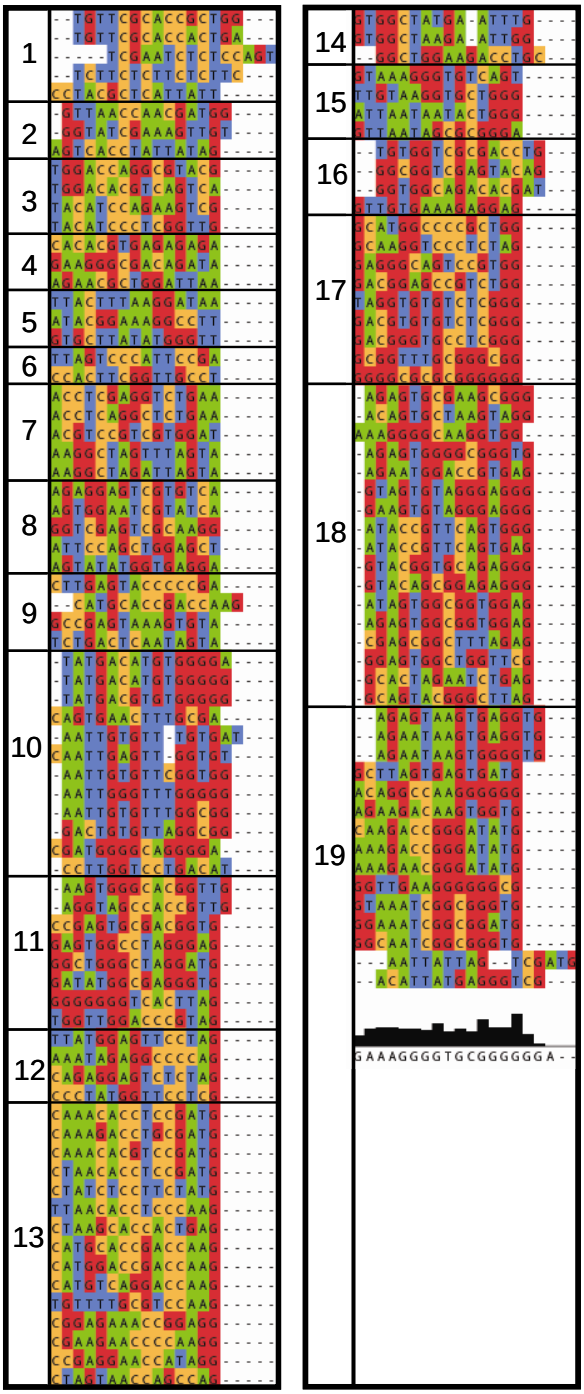


Figure S2.5. ISRE hierarchical clusters and sequence alignment. ISRE hierarchical clusters and sequence alignment of individual clusters. Sequences were aligned using ClustalX[56].

a

ISS17	GCAAGGTCCCTCTAG
ISS18	GACGGAGCCGTCTGG
ISS19	AGAGTGGCGGTGGAG
ISS20	GATATGGCGAGGGTG
ISS21	GGTGGCAGACACGAT
ISS22	AAATAGAGGCCCCAG
ISS23	TTATGGAGTTCCTAG
ISS24	GAGGGCAGTCCGTGG
ISS25	TGGACACGTCAGTCA
ISS26	TCTGACTCAATAGTA
ISS27	AATTGGGTTTGGGGG
ISS28	TATGACATGTGGGGA

b

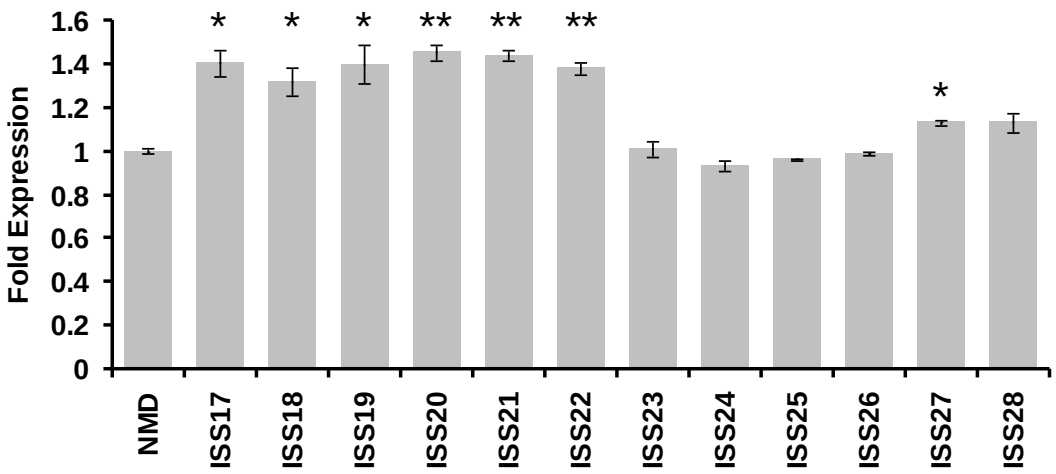


Figure S2.6. The activity of additional recovered ISRE sequences is validated by stable cell line assays. **(a)** Additional recovered ISRE sequences examined for regulatory activity. **(b)** Flow cytometry analysis of HEK-293 FLP-In stable cell lines generated for each recovered ISRE sequence and control construct. Mean GFP levels from two independent experiments were determined and normalized to the NMD control. The fold

expression of each sample relative to NMD and average error are reported. Resulting *P*-values in comparison to the NMD control: * $P < 0.03$ and ** $P < 0.01$.

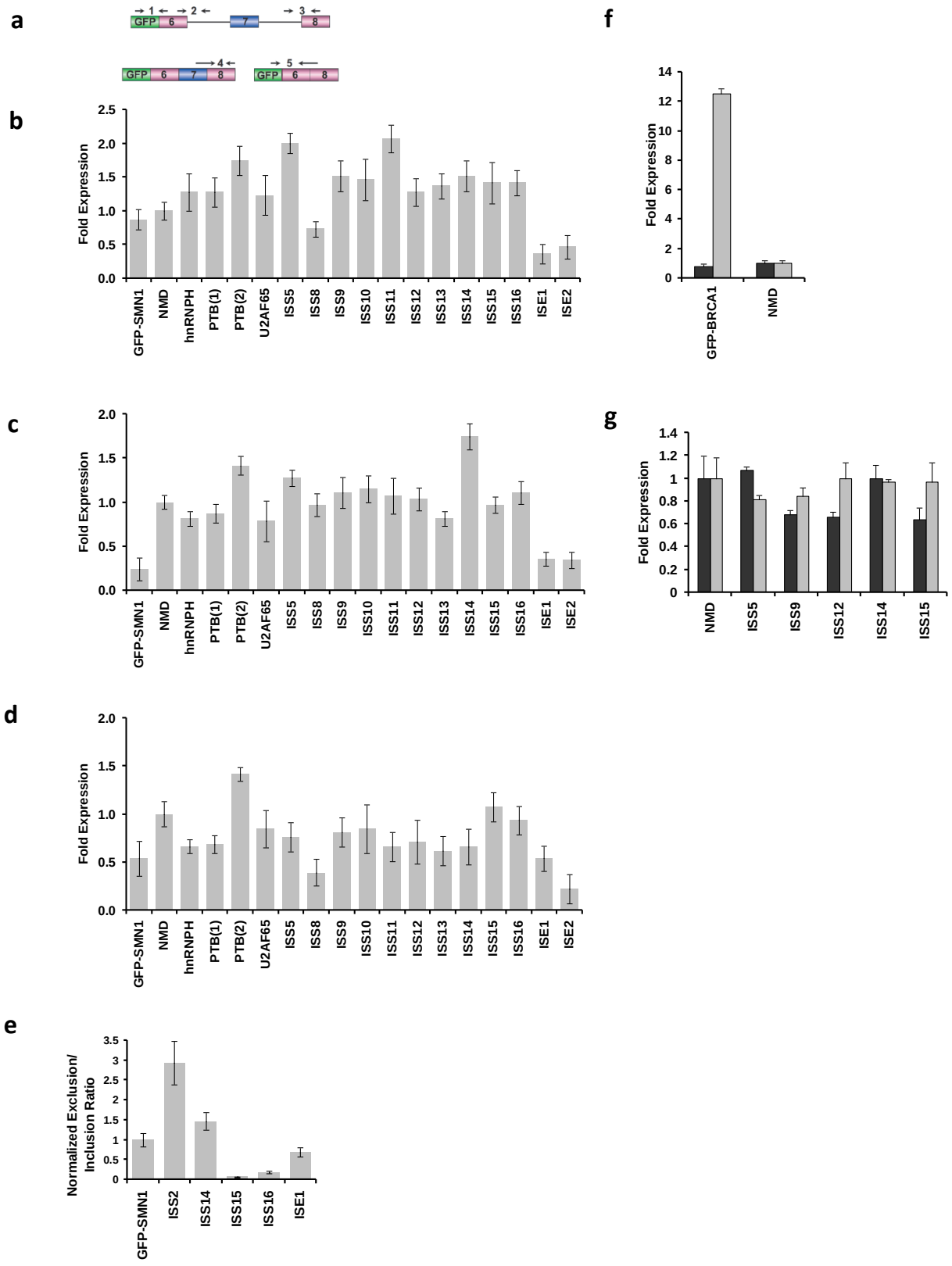

a
GFP
6
7
8
1
2
3
4
5
b
Fold Expression
GFP-SMN1
NMD
hnRNPH
PTB(1)
PTB(2)
U2AF65
ISS5
ISS8
ISS9
ISS10
ISS11
ISS12
ISS13
ISS14
ISS15
ISS16
ISE1
ISE2
c
d
e
Normalized Exclusion/ Inclusion Ratio
ISS2
f
GFP-BRCA1
g

Figure S2.7. Additional qRT-PCR isoform analysis of recovered ISREs and control constructs. **(a)** Schematic representing the relative locations of primer set binding on the reporter system for transcript isoform analysis. **(b)** qRT-PCR analysis with primer set 1. Results demonstrate that overall transcript levels for the GFP-SMN1, ISS controls, ISSs and ISEs did not significantly differ from the NMD control ($P = 0.20$). For all subsequent analyses, expression levels of duplicate PCR samples were normalized to the levels of *HPRT*. Fold expression data is reported as the mean expression for each sample divided by the mean NMD expression value $\pm$ the average error. **(c)** qRT-PCR analysis with primer set 2. The levels of intron 6 retained in transcripts containing the selected and control ISS sequences are similar to the NMD control ($P = 0.48$). In contrast, intron 6 retention in ISE transcripts are similar to the GFP-SMN1 control ($P = 0.74$) and different from the NMD control ($P < 0.05$), suggesting that intron 6 in the GFP-SMN1 control and ISEs are processed similarly by the general splicing machinery. The retention level of intron 6 for the GFP-SMN1 control is statistically different from the NMD control ($P < 0.05$). **(d)** qRT-PCR analysis with primer set 3. The levels of intron 7 retention for the recovered and control ISS sequences and the GFP-SMN1 are similar to the NMD control ($P = 0.23$). The intron 7 retention levels in ISE transcripts are significantly different from the NMD control ($P < .05$). **(e)** qRT-PCR analysis with primer sets 4 and 5 on selected ISREs inserted in the non-NMD-based GFP-SMN1 control construct. Stable cell lines containing ISS2, ISS14 and ISE1 maintained selected ISRE function; however, ISS15 and ISS16 displayed significant enhancer activity ($P < 0.05$). The results suggest that ISS15 and ISS16 may exhibit enhanced fluorescence levels in the context of the NMD reporter due to the evasion of the NMD process. Data is reported as the expression ratio

of the mean expression of the exon excluded isoform to the exon included isoform normalized to the ratio for the GFP-SMN1 control $\pm$ the average error. **(f)** qRT-PCR analysis of the NMD and GFP-BRACA1 constructs with primer sets specific for exon 18 excluded (black bars) and included (gray bars) products, supports decay of the PTC harboring isoform. **(g)** qRT-PCR analysis of selected sequences inserted into the BRCA1-NMD construct with primer sets specific for exon 18 excluded (black bars) and included

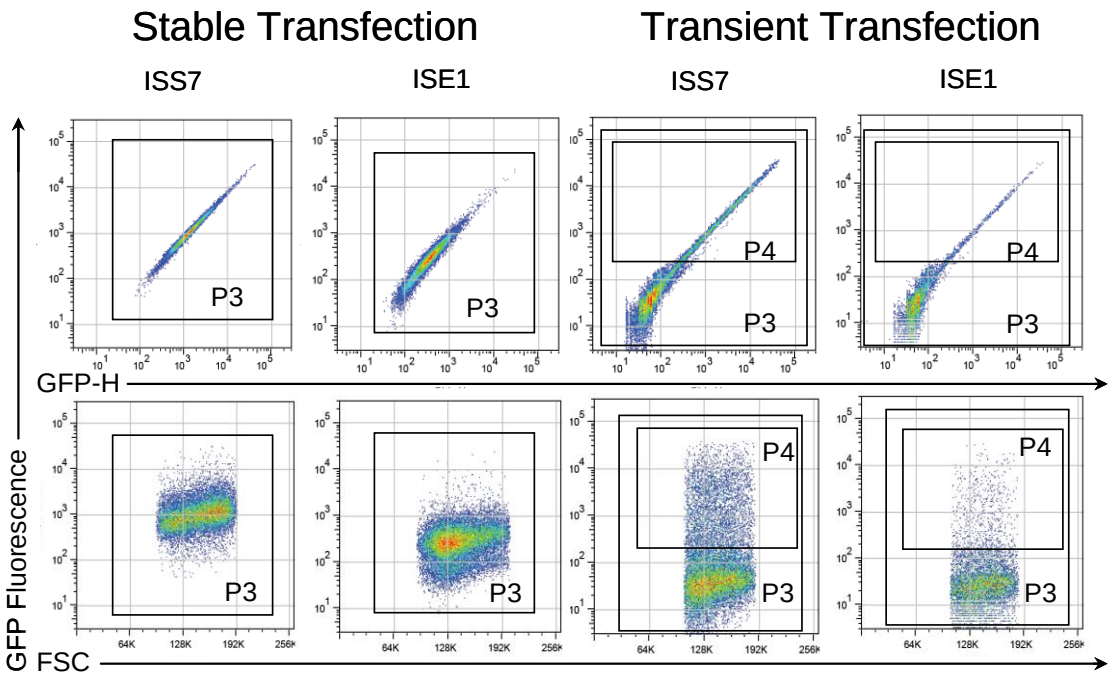


Figure S2.8. Assessment of splicing regulatory activity through stable and transient transfection assays. Sixteen recovered ISS sequences (ISS1-ISS16) and 1 recovered ISE sequence (ISE1) were examined for regulatory activity in both transient and stable transfection assays. Examples of assay results for two recovered sequences (ISS7, ISE1) are shown. For the stable cell line assays, mean GFP fluorescence levels were determined using gate P3. For the transient transfection assays, the P3 gate represents the untransfected cell population and the P4 gate represents the GFP-positive cells. The results of an ANOVA analysis applied to data from the transient and stable assays indicate that the two methods are not statistically similar (P = 0.27).

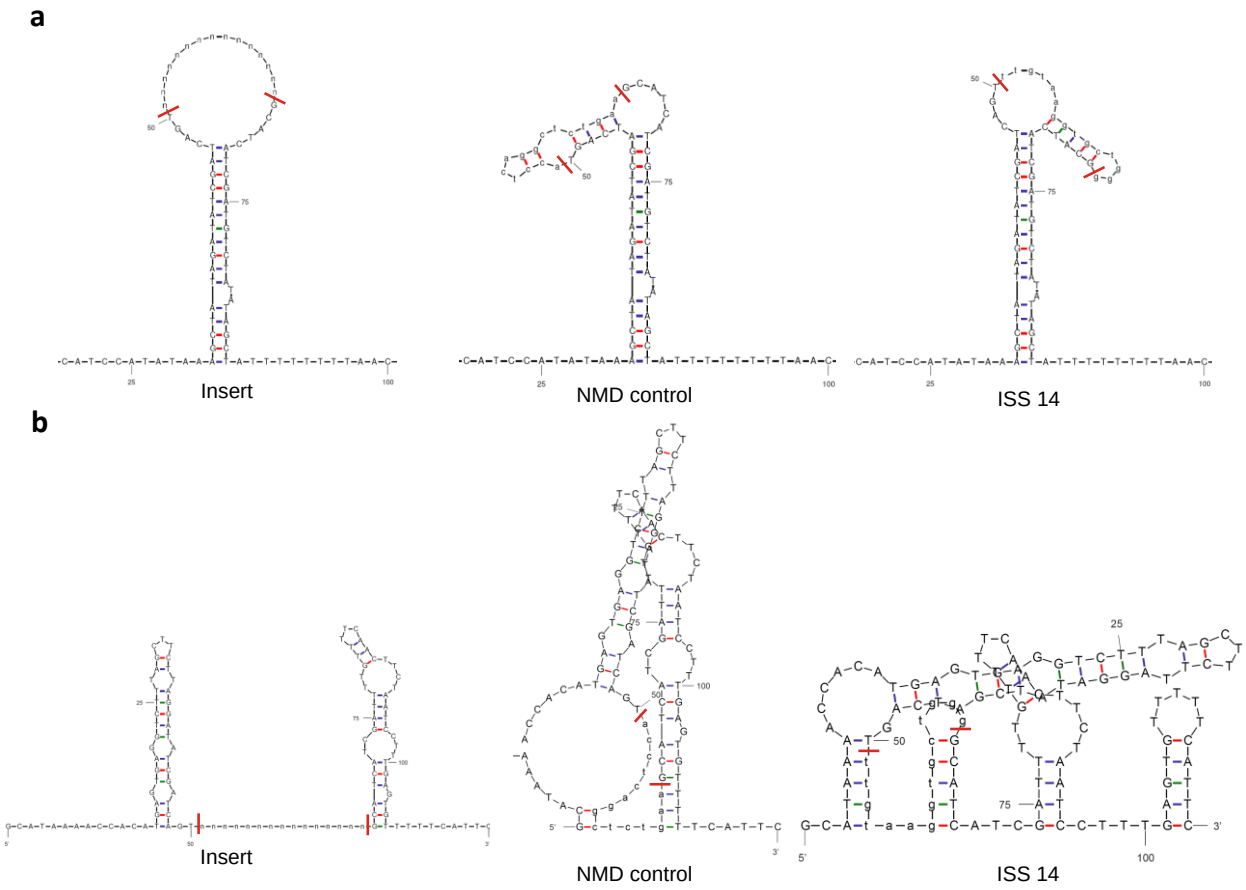


Figure S2.9. Predicted secondary structure for ISREs in the SMN1 and BRCA1 mini-genes. RNA secondary structures have been proposed to play a role in alternative splicing[61]. It is possible that the insertion of sequences within intron 6 of the SMN1 mini-gene may have resulted in a secondary structure that disrupts the binding of trans-acting factors. We examined structures of a region -50- to + 50-nt from the 15-mers within the SMN1 mini-gene using RNAfold[57]. **(a)** Predicted secondary structure for the intronic regions +/- 50-nt of any ISRE sequence, the random 15-nt (NMD control) and ISS14 in SMN1 mini-gene. Inserted sequences are located in position 51 to 65 in lower case letters and denoted by a red dash. The inserted sequences are predicted to be generally located

within a looped region. The local secondary structure around 15-mers within the SMN1 mini-gene is dominated by a hairpin structure. **(b)** Predicted secondary structure for the intronic regions +/- 50-nt of any ISS sequence, the random 15-mer (NMD control) and ISS14 in the BRCA1 mini-gene. Inserted sequences are located in position 51 to 65 in lower case letters and denoted by a red dash. The inserted sequences are predicted to be generally located within single stranded regions. The major structure within the BRCA1 mini-gene is a double hairpin. Splicing motifs are preferentially found in single-stranded contexts[62]. Taken together, the results suggest that the lack of selected ISRE function within the BRCA1 mini-gene may be due to differences in local secondary structure around the ISREs.

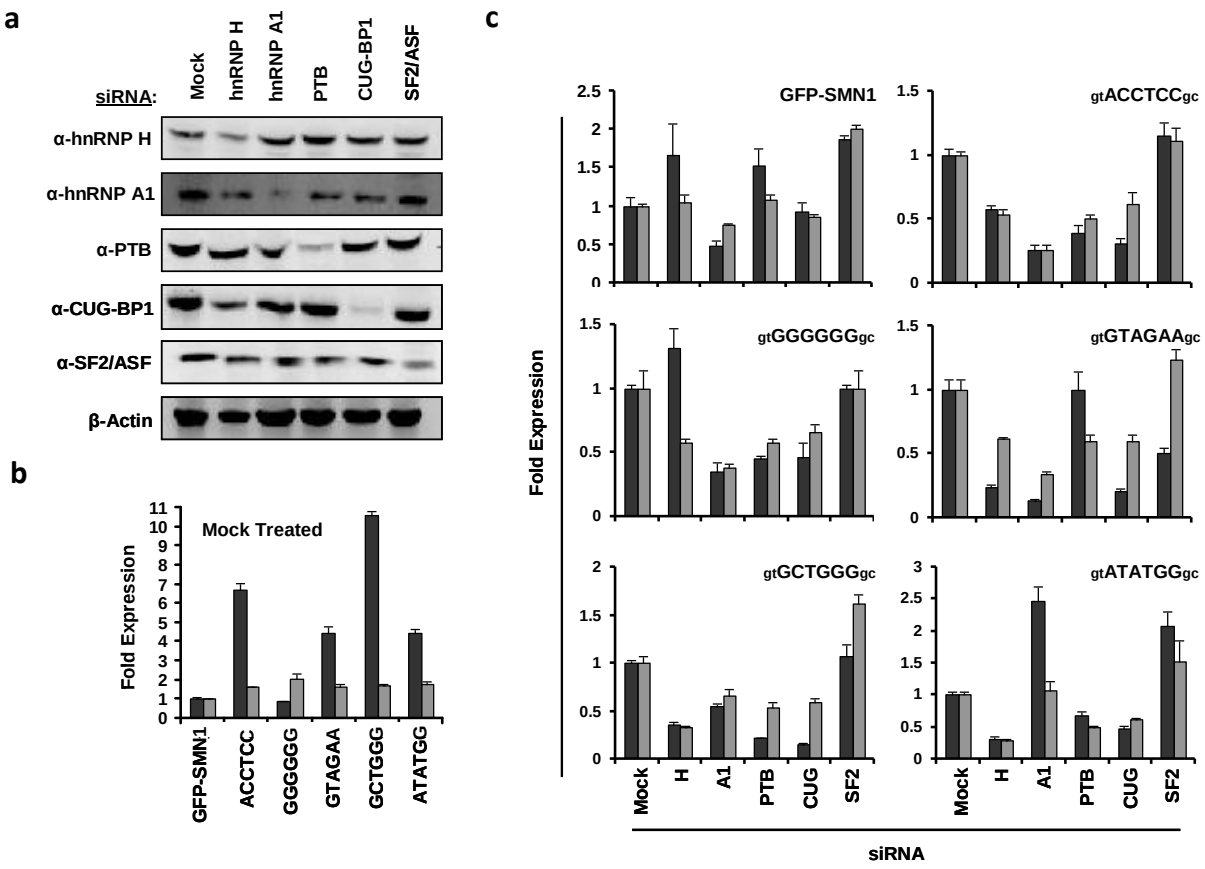


Figure S2.10. The effects of *in vivo* depletion of splicing factors on ISRE regulated splicing patterns. **(a)** Western blot analysis of total cell lysates prepared from the GFP-SMN1 control cell line treated with siRNAs targeted to trans-acting splicing factors and a mock siRNA negative control. β-Actin was used as a loading control for all blots. The results demonstrate that individual siRNAs have minimal to no off-target affects. **(b)** qRT-PCR analysis of the mock treated ISRE hexamer and GFP-SMN1 control cell lines with primer sets specific for exon 7 excluded (black bars) and included (gray bars) products. Expression levels of duplicate PCR samples were normalized to the levels of *HPRT*. Fold expression data is reported as the mean expression for each sample divided by the mean GFP-SMN1 control expression value ± the average error. **(c)** qRT-PCR analysis of the siRNA treated ISRE hexamer and GFP-SMN1 control cell lines with primer sets specific for exon 7 excluded (black bars) and included (gray bars) products.

Fold expression data is reported as the mean expression for each sample divided by the mean mock siRNA treated cell line control expression value $\pm$ the average error.

Table S2.1. Identified ISRE regulatory sequences

ISS sequences	Name	Tested stably	Tested transiently	Heterologous transcript
GACGTGTGTCTCGGG	ISE1	Y	Y	Y
ATAGTGGCGGTGGAG	ISE2	Y	N	N
TACATCCCTCGGTTG	ISS1	Y	Y	Y
AGAATAAGTGGGGTG	ISS2	Y	Y	Y
AGTATATGGTGAGGA	ISS3	Y	Y	Y
TGTTTTGCGTCCAAG	ISS4	Y	Y	Y
AGAATAAGTGAGGTG	ISS5	Y	Y	Y
CCGAGTGCGACGGTG	ISS6	Y	Y	Y
ACAGGCCAAGGGGGG	ISS7	Y	Y	Y
CAAACACCTCCGATG	ISS8	Y	Y	Y
GGTCGAGTCGCAAGG	ISS9	Y	Y	Y
TAGGTGTGTCTCGGG	ISS10	Y	Y	Y
ACAGTGCTAAGTAGG	ISS11	Y	Y	Y
AAAGACCGGGATATG	ISS12	Y	Y	Y
AGTCACCTATTATAG	ISS13	Y	Y	Y
TTGTAAGGTGCTGGG	ISS14	Y	Y	Y
GGGGCGCGCGGGGGG	ISS15	Y	Y	Y
AGAGTGGGGCGGGTG	ISS16	Y	Y	Y
GCAAGGTCCCTCTAG	ISS17	Y	N	N
GACGGAGCCGTCTGG	ISS18	Y	N	N
AGAGTGGCGGTGGAG	ISS19	Y	N	N
GATATGGCGAGGGTG	ISS20	Y	N	N
GGTGGCAGACACGAT	ISS21	Y	N	N
CCGAGGAACCATAGG	ISS23	Y	N	N
AAATAGAGGCCCCAG	ISS24	Y	N	N
TTATGGAGTTCCTAG	ISS25	Y	N	N
GAGGGCAGTCCGTGG	ISS26	Y	N	N
TGGACACGTCAGTCA	ISS27	Y	N	N
TCTGACTCAATAGTA	ISS28	Y	N	N
AATTGGGTTTGGGGG	ISS29	Y	N	N
TATGACATGTGGGGA	ISS30	Y	N	N
CCCTATGGTTCCTCG	ISS31			
GACGGGTGCCTCGGG	ISS32			
GGCTGGAAGACCTGC	ISS33			
GGAGTGGCTGGTTCG	ISS34			
GGCTGGGCTAGGATG	ISS35			
ACCTCAGGCTCTGAA	ISS36			
GACTGTGTTAGGCGG	ISS37			
AAAGAACGGGATATG				
TCGAATCTCTCCAGT				
CCTACGCTCATTATT				
TCTTCTCTTCTCTTC				
TGTTCGCACCGCTGG				
TGTTCGCACCACTGA				
GTTAACCAACGATGG				
GGTATCGAAAGTTGT				

TACATCCAGAAGTCG
TGGACCAGGCGTACG
CACACGTGAGAGAGA
GAAGGGCGACAGATA
AGAACGCTGGATTAA
TTACTTTAAGGATAA
ATACGGAAAGGCCTT
GTGCTTATATGGGTT
TTAGTCCCATTCCGA
CCACTTCGGTTGCCT
ACGTCCGTCGTGGAT
ACCTCGAGGTCTGAA
AAGGCTAGTTTAGTA
AAGGCTAGATTAGTA
AGAGGAGTCGTGTCA
AGTGGAATCGTATCA
ATTCCAGCTGGAGCT
GCCGAGTAAAGTGTA
CTTGAGTACCCCCGA
CATGCACCGACCAAG
AATTGTGTTTGTGAT
GACTGTGTTAGGCGG
AATTGTGTTTGGCGG
TATGACGTGTGGGGG
TATGACATGTGGGGG
CAATTGAGTTGGTGT
CGATGGGGCAGGGGA
CAGTGAACTTTGCGA
CCTTGGTCCTGACAT
GAGTGGCCTAGGGAG
GGCTGGGCTAGGATG
AAGTGGGCACGGTTG
AGGTAGCCACCGTTG
GGGGGGGTCACTTAG
TGGTTGGACCCGTAG
CCCTATGGTTCCTCG
CTAGTAACCAGCCAG
CTAAGCACCACTGAG
CATGTCAGGACCAAG
CATGGACCGACCAAG
TATGCCTCCCCGATA
CGAAGAACCCCAAGG
CGGAGAAACCGGAGG
CTATCTCCTTCTATG
TTAACACCTCCCAAG
CAAAGACCTGCGATG
CAAACACGTCCGATG
CTAACACCTCCGATG
GTGGCTATGAATTTG
GTGGCTAAGAATTGG
GGCTGGAAGACCTGC
GTAAAGGGTGTCAGT
ATTAATAATACTGGG
GTTAATAGCGCGGGA
TGTGGTCGCGACCTG
GGCGGTCGAGTACAG

GTTGTGAAAGAGGAG
GCGGTTTGCGGGCGG
GACGGGTGCCTCGGG
GCATGGCCCCGCTGG
GCACTAGAATCTGAG
GCAGTACGGGCTTAG
CGAGCGGCTTTAGAG
AGAATGGACCGTGAG
GGAGTGGCTGGTTCG
GTACAGCGGAGAGGG
GTACGGTGCAGAGGG
GTAGTGTAGGGAGGG
GAAGTGTAGGGAGGG
ATACCGTTCAGTGGG
ATACCGTTCAGTGAG
AAAGGGGCAAGGTGG
AGAGTGCGAAGCGGG
GTAAATCGGCGGGTG
GGAAATCGGCGGATG
GGCAATCGGCGGGTG
AAAGAACGGGATATG
CAAGACCGGGATATG
AATTATTAGTCGATG
GCTTAGTGAGTGATG
AGAAGACAAGTGGTG
GGTTGAAGGGGGGCG
ACATTATGAGGGTCG
AGAGTAAGTGAGGTG

The splicing activity of the first 30 sequences was assessed by stable transfection assays. Additional sequences validated through transiently transfection assays are indicated (Y= Tested and N=Not tested) in addition to those tested in a second transcript.

Table S2.2. Significantly enriched ISRE n-mers

Field	Description
n-mer	The n-mer
Length	N-mer length
Count(ISS)	Counts observed in ISS sample
Count(RS)	Counts observed in RS sample
N(ISS)	Total counts performed in ISS sample
N(RS)	Total counts performed in RS sample
P(ISS)	Probability of n-mer in ISS sample
P(RS)	Probability of n-mer in RS sample
CI(low)	"Lower cutoff for confidence interval (alpha= 0.02, two tailed)"
CI(high)	"Upper cutoff for confidence interval (alpha= 0.02, two tailed)"
Z	Z-score
P(Z)	P-value based on Z-score

n-mer	length	count(ISS)	count(RS)	N(ISS)	N(RS)	P(ISS)	P(RS)	CI(low)	CI(high)	Z	P(Z)
AAGG	4	19	28158	500	1799620	0.038	0.015647	0.009956	0.024508	4.02621	5.67E-05
AAGT	4	10	20912	500	1799620	0.02	0.01162	0.006882	0.019556	1.74801	0.080462
AGAA	4	11	20909	750	2699430	0.014667	0.007746	0.004585	0.013058	2.16143	0.030662
AGAG	4	14	28151	1000	3599240	0.014	0.007821	0.004971	0.012286	2.21742	0.026594
AGGC	4	32	49223	500	1799620	0.064	0.027352	0.019439	0.03836	5.02257	5.10E-07
AGGG	4	21	28200	500	1799620	0.042	0.01567	0.009975	0.024537	4.73884	2.15E-06
AGTA	4	11	20932	750	2699430	0.014667	0.007754	0.004591	0.013069	2.15759	0.03096
AGTG	4	24	28201	500	1799620	0.048	0.015671	0.009975	0.024537	5.8182	5.95E-09
ATGG	4	22	28223	500	1799620	0.044	0.015683	0.009985	0.024552	5.09436	3.50E-07
CCGA	4	11	21079	500	1799620	0.022	0.011713	0.006951	0.019672	2.13739	0.032566
CCTC	4	10	21415	750	2699430	0.013333	0.007933	0.004724	0.013292	1.66665	0.095584
CGGG	4	18	28018	500	1799620	0.036	0.015569	0.009896	0.024414	3.68909	0.000225
CGGT	4	10	21095	500	1799620	0.02	0.011722	0.006958	0.019683	1.71939	0.085544
GACC	4	12	21062	500	1799620	0.024	0.011704	0.006944	0.01966	2.55586	0.010593
GAGG	4	26	28015	750	2699430	0.034667	0.010378	0.006591	0.016305	6.56052	5.36E-11
GAGT	4	16	21335	500	1799620	0.032	0.011855	0.007058	0.019849	4.16024	3.18E-05
GATG	4	13	28209	750	2699430	0.017333	0.01045	0.006647	0.016392	1.85334	0.063834
GCGG	4	18	28223	750	2699430	0.024	0.010455	0.006651	0.016398	3.64572	0.000267
GGAG	4	19	28339	750	2699430	0.025333	0.010498	0.006685	0.016451	3.98487	6.75E-05
GGCG	4	15	28442	750	2699430	0.02	0.010536	0.006715	0.016497	2.53766	0.01116
GGCT	4	11	21106	500	1799620	0.022	0.011728	0.006962	0.019691	2.13292	0.032931
GGGC	4	43	49067	500	1799620	0.086	0.027265	0.019367	0.038259	8.06106	7.56E-16
GGGG	4	42	28105	2000	7198480	0.021	0.003904	0.002479	0.006143	12.2506	1.67E-34
GGGT	4	11	21111	500	1799620	0.022	0.011731	0.006965	0.019694	2.1321	0.032999
GGTG	4	21	27984	750	2699430	0.028	0.010367	0.006583	0.016291	4.76593	1.88E-06
GGTT	4	10	21126	500	1799620	0.02	0.011739	0.006971	0.019705	1.71456	0.086425
GTAG	4	24	55917	750	2699430	0.032	0.020714	0.015022	0.028501	2.16957	0.03004
GTGA	4	20	49127	500	1799620	0.04	0.027299	0.019395	0.038298	1.74258	0.081408
GTGG	4	44	56695	750	2699430	0.058667	0.021003	0.015265	0.028833	7.19059	6.45E-13
GTGT	4	26	49154	1000	3599240	0.026	0.013657	0.009687	0.019221	3.36222	0.000773
TAGA	4	14	27883	500	1799620	0.028	0.015494	0.009837	0.024323	2.26367	0.023595
TAGG	4	16	35047	500	1799620	0.032	0.019475	0.012986	0.02911	2.02634	0.04273
TAGT	4	13	27981	750	2699430	0.017333	0.010366	0.006582	0.016289	1.88362	0.059616
TATG	4	15	35246	500	1799620	0.03	0.019585	0.013075	0.029241	1.68025	0.09291
TGAG	4	15	35211	500	1799620	0.03	0.019566	0.013059	0.029218	1.6842	0.092142
TGGA	4	15	28342	500	1799620	0.03	0.015749	0.010036	0.024633	2.55883	0.010502
TGGC	4	40	56418	500	1799620	0.08	0.03135	0.022796	0.042973	6.24044	4.36E-10
TGGG	4	24	35134	500	1799620	0.048	0.019523	0.013025	0.029167	4.60088	4.21E-06
TGTG	4	17	34998	1000	3599240	0.017	0.009724	0.006474	0.014582	2.34428	0.019064
AAAGA	5	4	4874	500	1799620	0.008	0.002708	0.000946	0.007727	2.2758	0.022858
AACAC	5	4	4855	375	1349720	0.010667	0.003597	0.001255	0.010269	2.28582	0.022265
AAGAC	5	5	4829	375	1349720	0.013333	0.003578	0.001245	0.01024	3.16238	0.001565
AAGGC	5	8	11863	375	1349720	0.021333	0.008789	0.004406	0.017456	2.60165	0.009278
AAGGG	5	7	6614	375	1349720	0.018667	0.0049	0.001968	0.012151	3.81559	0.000136
AAGTG	5	7	6506	375	1349720	0.018667	0.00482	0.001922	0.012037	3.8693	0.000109

AATTG	5	6	6471	375	1349720	0.016	0.004794	0.001907	0.012	3.14001	0.001689
ACCAA	5	4	4776	500	1799620	0.008	0.002654	0.000918	0.007646	2.32261	0.0202
ACCGT	5	4	4779	375	1349720	0.010667	0.003541	0.001225	0.010186	2.32219	0.020223
ACCTC	5	5	4900	375	1349720	0.013333	0.00363	0.001272	0.010318	3.12256	0.001793
AGAAT	5	5	4745	375	1349720	0.013333	0.003516	0.001213	0.010149	3.21047	0.001325
AGACC	5	4	4807	375	1349720	0.010667	0.003561	0.001236	0.010216	2.30871	0.02096
AGAGG	5	7	6553	375	1349720	0.018667	0.004855	0.001942	0.012087	3.84579	0.00012
AGAGT	5	4	4903	375	1349720	0.010667	0.003633	0.001273	0.010321	2.26321	0.023623
AGGGC	5	7	11795	375	1349720	0.018667	0.008739	0.004372	0.01739	2.06499	0.038924
AGGGG	5	8	6582	375	1349720	0.021333	0.004877	0.001954	0.012117	4.57194	4.83E-06
AGGTG	5	5	6563	375	1349720	0.013333	0.004863	0.001946	0.012097	2.35725	0.018411
AGTGA	5	6	4826	500	1799620	0.012	0.002682	0.000932	0.007687	4.02655	5.66E-05
AGTGG	5	10	6699	375	1349720	0.026667	0.004963	0.002004	0.01224	5.97609	2.29E-09
ATATG	5	6	6591	375	1349720	0.016	0.004883	0.001958	0.012127	3.08677	0.002023
ATGGC	5	14	11824	375	1349720	0.037333	0.00876	0.004387	0.017418	5.93424	2.95E-09
ATTAT	5	4	4997	625	2249520	0.0064	0.002221	0.000785	0.006267	2.21807	0.02655
CAAGG	5	10	6657	375	1349720	0.026667	0.004932	0.001986	0.012196	6.00339	1.93E-09
CACCT	5	4	4692	375	1349720	0.010667	0.003476	0.001192	0.01009	2.36473	0.018043
CAGTG	5	6	6656	375	1349720	0.016	0.004931	0.001985	0.012195	3.05845	0.002225
CCAAG	5	7	6510	375	1349720	0.018667	0.004823	0.001924	0.012041	3.86729	0.00011
CCGAG	5	5	6559	375	1349720	0.013333	0.00486	0.001944	0.012093	2.35878	0.018335
CCGAT	5	4	4808	375	1349720	0.010667	0.003562	0.001237	0.010217	2.30823	0.020987
CCTCC	5	4	4871	625	2249520	0.0064	0.002165	0.000756	0.006184	2.2766	0.02281
CGAGT	5	4	4886	375	1349720	0.010667	0.00362	0.001266	0.010303	2.27118	0.023136
CGATG	5	8	6496	375	1349720	0.021333	0.004813	0.001918	0.012027	4.61974	3.84E-06
CGGGA	5	4	4850	375	1349720	0.010667	0.003593	0.001253	0.010263	2.28819	0.022127
CGGGG	5	5	6625	375	1349720	0.013333	0.004908	0.001972	0.012162	2.33353	0.01962
CGGGT	5	4	4712	375	1349720	0.010667	0.003491	0.0012	0.010112	2.35486	0.01853
CGGTG	5	5	6503	375	1349720	0.013333	0.004818	0.001921	0.012034	2.38046	0.017291
CGGTT	5	4	4834	375	1349720	0.010667	0.003582	0.001246	0.010246	2.2958	0.021688
CTAGG	5	5	6483	375	1349720	0.013333	0.004803	0.001912	0.012013	2.38826	0.016928
CTCGG	5	5	6722	375	1349720	0.013333	0.00498	0.002014	0.012264	2.29696	0.021621
CTGGG	5	6	6289	375	1349720	0.016	0.00466	0.001831	0.011808	3.22319	0.001268
GACCA	5	4	4763	375	1349720	0.010667	0.003529	0.001219	0.010168	2.32994	0.019809
GAGGA	5	5	4770	625	2249520	0.008	0.00212	0.000733	0.006118	3.19377	0.001404
GAGGC	5	10	11892	375	1349720	0.026667	0.008811	0.00442	0.017485	3.69855	0.000217
GAGGG	5	8	6569	500	1799620	0.016	0.00365	0.001461	0.009089	4.57631	4.73E-06
GAGTA	5	4	4835	375	1349720	0.010667	0.003582	0.001247	0.010247	2.29532	0.021715
GAGTG	5	7	6725	500	1799620	0.014	0.003737	0.001511	0.009212	3.7592	0.00017
GATAT	5	4	4827	375	1349720	0.010667	0.003576	0.001244	0.010238	2.29914	0.021497
GATGG	5	10	6628	500	1799620	0.02	0.003683	0.00148	0.009136	6.01865	1.76E-09
GCACC	5	4	4837	375	1349720	0.010667	0.003584	0.001248	0.010249	2.29437	0.021769
GCGGG	5	10	6456	500	1799620	0.02	0.003587	0.001425	0.009	6.1336	8.59E-10
GCGGT	5	4	4843	375	1349720	0.010667	0.003588	0.00125	0.010256	2.29151	0.021934
GCTGG	5	8	6494	500	1799620	0.016	0.003609	0.001437	0.00903	4.61807	3.87E-06
GGACC	5	5	4794	375	1349720	0.013333	0.003552	0.001231	0.010202	3.18228	0.001461
GGAGG	5	7	6609	625	2249520	0.0112	0.002938	0.001179	0.007303	3.8143	0.000137
GGAGT	5	4	4919	375	1349720	0.010667	0.003644	0.001279	0.010339	2.25573	0.024087
GGATA	5	4	4790	375	1349720	0.010667	0.003549	0.00123	0.010198	2.31688	0.02051
GGCGG	5	11	6706	625	2249520	0.0176	0.002981	0.001204	0.007364	6.69827	2.11E-11
GGCTA	5	5	4800	375	1349720	0.013333	0.003556	0.001233	0.010209	3.17886	0.001479
GGGAG	5	6	6735	500	1799620	0.012	0.003742	0.001514	0.00922	3.02259	0.002506
GGGCA	5	4	4761	375	1349720	0.010667	0.003527	0.001219	0.010166	2.33091	0.019758
GGGCG	5	5	6655	500	1799620	0.01	0.003698	0.001488	0.009157	2.32071	0.020302
GGGGC	5	21	11858	375	1349720	0.056	0.008786	0.004404	0.017451	9.78905	1.25E-22
GGGGG	5	17	6589	1875	6748580	0.009067	0.000976	0.000391	0.002435	11.2025	3.96E-29
GGGTG	5	7	6537	500	1799620	0.014	0.003632	0.001451	0.009064	3.85142	0.000117
GGTGG	5	14	6731	625	2249520	0.0224	0.002992	0.00121	0.00738	8.87407	7.05E-19
GGTTG	5	6	6641	500	1799620	0.012	0.00369	0.001484	0.009146	3.06306	0.002191
GTAAA	5	7	11813	375	1349720	0.018667	0.008752	0.004381	0.017407	2.06066	0.039336
GTAGA	5	10	11703	375	1349720	0.026667	0.008671	0.004327	0.0173	3.75724	0.000172
GTGAG	5	9	13633	500	1799620	0.018	0.007575	0.003971	0.014404	2.68747	0.0072
GTGGC	5	20	19065	375	1349720	0.053333	0.014125	0.008168	0.024319	6.4307	1.27E-10
GTGGG	5	12	13696	500	1799620	0.024	0.007611	0.003995	0.01445	4.21516	2.50E-05
GTGTA	5	8	11859	375	1349720	0.021333	0.008786	0.004404	0.017452	2.6027	0.009249
GTGTT	5	7	11954	375	1349720	0.018667	0.008857	0.004451	0.017546	2.027	0.042663
GTTGG	5	8	13697	500	1799620	0.016	0.007611	0.003996	0.014451	2.15776	0.030947
TAAGT	5	5	6561	500	1799620	0.01	0.003646	0.001459	0.009083	2.35658	0.018444
TAATT	5	5	6619	500	1799620	0.01	0.003678	0.001477	0.009129	2.33437	0.019576
TAGAA	5	6	6486	375	1349720	0.016	0.004805	0.001913	0.012016	3.13329	0.001729

TAGAG	5	7	8315	375	1349720	0.018667	0.006161	0.002716	0.013912	3.09377	0.001976
TAGGC	5	8	13605	375	1349720	0.021333	0.01008	0.005283	0.01915	2.18095	0.029187
TAGTA	5	5	6516	625	2249520	0.008	0.002897	0.001155	0.007244	2.37311	0.017639
TATGA	5	5	6628	375	1349720	0.013333	0.004911	0.001974	0.012166	2.33239	0.01968
TATGG	5	9	8411	375	1349720	0.024	0.006232	0.00276	0.01401	4.37004	1.24E-05
TCAGT	5	5	6748	500	1799620	0.01	0.00375	0.001518	0.009231	2.28584	0.022264
TCCGA	5	6	6654	375	1349720	0.016	0.00493	0.001985	0.012193	3.05931	0.002218
TGAGG	5	9	8379	375	1349720	0.024	0.006208	0.002745	0.013977	4.38416	1.16E-05
TGGAC	5	5	6631	375	1349720	0.013333	0.004913	0.001975	0.012169	2.33125	0.01974
TGGCT	5	5	6581	500	1799620	0.01	0.003657	0.001465	0.009099	2.34889	0.018829
TGGGC	5	10	13583	375	1349720	0.026667	0.010064	0.005271	0.019129	3.22006	0.001282
TGGGG	5	12	8309	375	1349720	0.032	0.006156	0.002713	0.013906	6.39366	1.62E-10
TGTGG	5	6	8242	375	1349720	0.016	0.006106	0.002683	0.013837	2.45835	0.013958
TGTGT	5	6	6606	1000	3599240	0.006	0.001835	0.000736	0.004568	3.07548	0.002102
TTAGT	5	5	6596	500	1799620	0.01	0.003665	0.00147	0.009111	2.34315	0.019122
TTATG	5	6	8434	375	1349720	0.016	0.006249	0.00277	0.014033	2.39546	0.0166
TTGTG	5	7	8242	375	1349720	0.018667	0.006106	0.002683	0.013837	3.12078	0.001804
AACACC	6	3	1051	375	1349720	0.008	0.000779	0.000104	0.005807	5.00613	5.55E-07
AAGACC	6	4	1099	375	1349720	0.010667	0.000814	0.000112	0.005869	6.67681	2.44E-11
AAGGGC	6	3	2802	375	1349720	0.008	0.002076	0.000539	0.007955	2.51905	0.011767
AAGGGG	6	3	1567	375	1349720	0.008	0.001161	0.000208	0.006465	3.88538	0.000102
AAGTGG	6	3	1570	375	1349720	0.008	0.001163	0.000208	0.006469	3.88041	0.000104
AATCGG	6	3	1493	375	1349720	0.008	0.001106	0.000191	0.006372	4.01211	6.02E-05
AATTGT	6	3	1098	375	1349720	0.008	0.000814	0.000112	0.005868	4.87458	1.09E-06
ACACCT	6	3	1070	375	1349720	0.008	0.000793	0.000107	0.005832	4.95198	7.35E-07
ACACGT	6	3	1118	375	1349720	0.008	0.000828	0.000116	0.005894	4.82098	1.43E-06
ACCAAG	6	3	1566	375	1349720	0.008	0.00116	0.000207	0.006464	3.88703	0.000101
ACCGTT	6	3	1118	375	1349720	0.008	0.000828	0.000116	0.005894	4.82098	1.43E-06
ACCTCC	6	3	1103	375	1349720	0.008	0.000817	0.000113	0.005875	4.86105	1.17E-06
AGAGGA	6	3	1082	375	1349720	0.008	0.000802	0.000109	0.005847	4.91846	8.72E-07
AGAGTG	6	3	1607	375	1349720	0.008	0.001191	0.000217	0.006515	3.82024	0.000133
AGGGAG	6	3	1551	500	1799620	0.006	0.000862	0.000153	0.004839	3.91151	9.17E-05
AGGGGC	6	5	2965	375	1349720	0.013333	0.002197	0.00059	0.008145	4.60244	4.18E-06
AGGTGG	6	3	1554	375	1349720	0.008	0.001151	0.000205	0.006449	3.90704	9.34E-05
AGTAGC	6	3	2827	375	1349720	0.008	0.002095	0.000547	0.007984	2.50008	0.012416
AGTGAG	6	4	1587	500	1799620	0.008	0.000882	0.000159	0.004873	5.35549	8.53E-08
AGTGGC	6	4	2920	375	1349720	0.010667	0.002163	0.000576	0.008093	3.54163	0.000398
AGTGGG	6	4	1585	375	1349720	0.010667	0.001174	0.000212	0.006487	5.36048	8.30E-08
AGTGTA	6	3	1106	375	1349720	0.008	0.000819	0.000114	0.005879	4.85298	1.22E-06
ATATGG	6	6	1546	375	1349720	0.016	0.001145	0.000203	0.006439	8.48791	2.10E-17
ATCGGC	6	3	2780	375	1349720	0.008	0.00206	0.000533	0.007929	2.53593	0.011215
ATTGTG	6	3	1470	375	1349720	0.008	0.001089	0.000186	0.006343	4.05327	5.05E-05
CAAGGC	6	5	2932	375	1349720	0.013333	0.002172	0.00058	0.008107	4.63833	3.51E-06
CAAGGG	6	3	1539	375	1349720	0.008	0.00114	0.000201	0.00643	3.93234	8.41E-05
CACCTC	6	3	1030	375	1349720	0.008	0.000763	0.0001	0.00578	5.06762	4.03E-07
CCAAGG	6	7	1610	375	1349720	0.018667	0.001193	0.000217	0.006519	9.78205	1.34E-22
CCAGGC	6	3	2816	375	1349720	0.008	0.002086	0.000544	0.007971	2.5084	0.012128
CCGATG	6	3	1597	375	1349720	0.008	0.001183	0.000214	0.006502	3.83631	0.000125
CCTCGG	6	3	1594	375	1349720	0.008	0.001181	0.000214	0.006499	3.84116	0.000122
CGATGG	6	7	1576	375	1349720	0.018667	0.001168	0.00021	0.006476	9.90067	4.14E-23
CGCTGG	6	3	1512	375	1349720	0.008	0.00112	0.000195	0.006396	3.97876	6.93E-05
CGGGAT	6	3	1101	375	1349720	0.008	0.000816	0.000113	0.005872	4.86645	1.14E-06
CGGGGC	6	4	2873	375	1349720	0.010667	0.002129	0.000561	0.008038	3.585	0.000337
CGGGTG	6	4	1488	375	1349720	0.010667	0.001102	0.00019	0.006366	5.57367	2.49E-08
CGGTGG	6	4	1557	375	1349720	0.010667	0.001154	0.000205	0.006452	5.42011	5.96E-08
CGGTTG	6	3	1558	375	1349720	0.008	0.001154	0.000206	0.006454	3.90035	9.61E-05
CTAGGC	6	3	2821	375	1349720	0.008	0.00209	0.000545	0.007977	2.50462	0.012258
CTCGGG	6	3	1591	375	1349720	0.008	0.001179	0.000213	0.006495	3.84602	0.00012
CTGGGC	6	4	2641	375	1349720	0.010667	0.001957	0.000491	0.007766	3.81387	0.000137
GACATG	6	3	1553	375	1349720	0.008	0.001151	0.000204	0.006447	3.90871	9.28E-05
GACCAA	6	3	1126	375	1349720	0.008	0.000834	0.000117	0.005904	4.79991	1.59E-06
GACCTG	6	3	1547	375	1349720	0.008	0.001146	0.000203	0.00644	3.91881	8.90E-05
GAGTGG	6	4	1582	375	1349720	0.010667	0.001172	0.000211	0.006484	5.3668	8.01E-08
GATATG	6	4	1550	375	1349720	0.010667	0.001148	0.000204	0.006444	5.43525	5.47E-08
GATGGC	6	8	2823	375	1349720	0.021333	0.002092	0.000546	0.00798	8.14456	3.81E-16
GCAAGG	6	3	1535	375	1349720	0.008	0.001137	0.0002	0.006425	3.93915	8.18E-05
GCTGGA	6	3	1087	375	1349720	0.008	0.000805	0.00011	0.005854	4.90465	9.36E-07
GCTGGG	6	4	1480	375	1349720	0.010667	0.001097	0.000188	0.006356	5.59212	2.24E-08
GGAGGC	6	4	2891	375	1349720	0.010667	0.002142	0.000567	0.008059	3.56828	0.000359
GGAGGG	6	3	1564	500	1799620	0.006	0.000869	0.000155	0.004851	3.88979	0.0001

GGATAT	6	3	1087	375	1349720	0.008	0.000805	0.00011	0.005854	4.90465	9.36E-07
GGCGGG	6	6	1527	500	1799620	0.012	0.000849	0.000149	0.004816	8.54717	1.26E-17
GGCGGT	6	3	1116	375	1349720	0.008	0.000827	0.000116	0.005892	4.82628	1.39E-06
GGCTAG	6	3	1543	375	1349720	0.008	0.001143	0.000202	0.006435	3.92556	8.65E-05
GGCTGG	6	3	1475	500	1799620	0.006	0.00082	0.00014	0.004767	4.04369	5.26E-05
GGGAGG	6	3	1494	500	1799620	0.006	0.00083	0.000144	0.004785	4.00978	6.08E-05
GGGATA	6	3	1099	375	1349720	0.008	0.000814	0.000112	0.005869	4.87187	1.11E-06
GGGCGG	6	3	1606	500	1799620	0.006	0.000892	0.000162	0.004891	3.82127	0.000133
GGGGCG	6	3	1516	375	1349720	0.008	0.001123	0.000196	0.006401	3.97181	7.13E-05
GGGGGC	6	6	2780	375	1349720	0.016	0.00206	0.000533	0.007929	5.94795	2.72E-09
GGGGGG	6	10	1546	1750	6298670	0.005714	0.000245	4.35E-05	0.001385	14.5575	5.23E-48
GGGTGG	6	5	1591	500	1799620	0.01	0.000884	0.00016	0.004877	6.84782	7.50E-12
GGTGGC	6	10	2938	375	1349720	0.026667	0.002177	0.000581	0.008114	10.1586	3.03E-24
GGTTGG	6	3	1544	500	1799620	0.006	0.000858	0.000152	0.004832	3.92331	8.73E-05
GTAAAG	6	5	3287	375	1349720	0.013333	0.002435	0.000693	0.008517	4.27842	1.88E-05
GTAATT	6	5	2886	375	1349720	0.013333	0.002138	0.000565	0.008053	4.68928	2.74E-06
GTACAG	6	4	3303	375	1349720	0.010667	0.002447	0.000699	0.008535	3.21956	0.001284
GTAGAA	6	5	2839	375	1349720	0.013333	0.002103	0.000551	0.007998	4.74249	2.11E-06
GTAGAG	6	5	3255	375	1349720	0.013333	0.002412	0.000683	0.00848	4.30867	1.64E-05
GTATAC	6	3	2834	375	1349720	0.008	0.0021	0.000549	0.007992	2.49481	0.012602
GTGAGG	6	6	3343	375	1349720	0.016	0.002477	0.000712	0.008581	5.26377	1.41E-07
GTGGCT	6	5	2866	375	1349720	0.013333	0.002123	0.000559	0.00803	4.71178	2.46E-06
GTGGGC	6	4	4664	375	1349720	0.010667	0.003456	0.001182	0.01006	2.37863	0.017377
GTGGGG	6	8	3311	375	1349720	0.021333	0.002453	0.000701	0.008544	7.38201	1.56E-13
GTGTAG	6	4	3242	375	1349720	0.010667	0.002402	0.000679	0.008465	3.26747	0.001085
GTTATG	6	4	3288	375	1349720	0.010667	0.002436	0.000694	0.008518	3.23123	0.001233
GTTGGA	6	3	2833	375	1349720	0.008	0.002099	0.000549	0.007991	2.49556	0.012576
TAAGTG	6	4	1947	375	1349720	0.010667	0.001443	0.000299	0.006934	4.70162	2.58E-06
TAATTG	6	4	1923	375	1349720	0.010667	0.001425	0.000293	0.006905	4.73989	2.14E-06
TAGAAT	6	4	1532	375	1349720	0.010667	0.001135	0.0002	0.006421	5.47463	4.38E-08
TAGAGG	6	3	1937	375	1349720	0.008	0.001435	0.000296	0.006922	3.35564	0.000792
TAGAGT	6	4	1550	375	1349720	0.010667	0.001148	0.000204	0.006444	5.43525	5.47E-08
TAGGGA	6	3	1529	375	1349720	0.008	0.001133	0.000199	0.006417	3.9494	7.83E-05
TAGTAG	6	3	1945	625	2249520	0.0048	0.000865	0.000179	0.004167	3.34476	0.000824
TATGAC	6	3	1566	375	1349720	0.008	0.00116	0.000207	0.006464	3.88703	0.000101
TATGGC	6	5	3316	375	1349720	0.013333	0.002457	0.000703	0.00855	4.25135	2.12E-05
TCAGTG	6	4	2000	375	1349720	0.010667	0.001482	0.000312	0.006999	4.61938	3.85E-06
TCCGAG	6	3	1961	375	1349720	0.008	0.001453	0.000302	0.006951	3.32607	0.000881
TCCGAT	6	3	1595	375	1349720	0.008	0.001182	0.000214	0.0065	3.83954	0.000123
TCGGCG	6	3	1999	375	1349720	0.008	0.001481	0.000312	0.006998	3.28023	0.001037
TCGGGG	6	3	1959	375	1349720	0.008	0.001451	0.000302	0.006949	3.32851	0.000873
TGACAT	6	3	1546	375	1349720	0.008	0.001145	0.000203	0.006439	3.92049	8.84E-05
TGAGGC	6	4	3245	375	1349720	0.010667	0.002404	0.00068	0.008468	3.26508	0.001094
TGGACC	6	4	1539	375	1349720	0.010667	0.00114	0.000201	0.00643	5.45924	4.78E-08
TGGCGG	6	4	1988	375	1349720	0.010667	0.001473	0.000309	0.006984	4.63774	3.52E-06
TGGCTG	6	3	1996	500	1799620	0.006	0.001109	0.000233	0.005252	3.2832	0.001026
TGGGGC	6	6	3240	375	1349720	0.016	0.002401	0.000678	0.008463	5.3766	7.59E-08
TGGGGG	6	4	2007	375	1349720	0.010667	0.001487	0.000314	0.007007	4.60874	4.05E-06
TGTCAG	6	3	1914	375	1349720	0.008	0.001418	0.00029	0.006894	3.38444	0.000713
TGTGGG	6	3	1953	375	1349720	0.008	0.001447	0.0003	0.006942	3.33587	0.00085
TGTGTT	6	4	1581	375	1349720	0.010667	0.001171	0.000211	0.006482	5.36891	7.92E-08
TGTTCG	6	3	2005	375	1349720	0.008	0.001486	0.000314	0.007005	3.2731	0.001064
TTATGA	6	4	1553	375	1349720	0.010667	0.001151	0.000204	0.006447	5.42875	5.68E-08
TTGGAC	6	3	1535	375	1349720	0.008	0.001137	0.0002	0.006425	3.93915	8.18E-05
TTGTGT	6	3	1550	375	1349720	0.008	0.001148	0.000204	0.006444	3.91375	9.09E-05
TTTGCG	6	3	2005	375	1349720	0.008	0.001486	0.000314	0.007005	3.2731	0.001064

Table S2.3. GCCS clusters derived from the ISRE enriched n-mers

Field	Description
n-mer	The n-mer (4-6mers)
clustID	ClusterID
GCS	Greatest Common Substring
Len	n-mer length
aligned	Aligned n-mers
wWeight	Edge weight
count	Count of n-mer in ISS dataset
Zscore	Z-score for n-mer
round	Clustering round in which produced cluster
vDegree	Vertex degree (number of other vertices attached)
TA	Association score

n-mer	clustID	GCS	len	aligned	wWeight	count	Zscore	round	vDegree	TA
AGGTG	1	GGTG	5	'AGGTG--	2.35725	5	2.35725	1	1	0
AGGTGG	1	GGTG	6	'AGGTGG-	3.90704	3	3.90704	1	5	0.6
CGGTG	1	GGTG	5	'CGGTG--	2.38046	5	2.38046	1	1	0
CGGTGG	1	GGTG	6	'CGGTGG-	5.42011	4	5.42011	1	5	0.6
GGGTG	1	GGTG	5	'GGGTG--	3.85142	7	3.85142	1	1	0
GGGTGG	1	GGTG	6	'GGGTGG-	6.84782	5	6.84782	1	5	0.6
GGTGG	1	GGTG	5	'-GGTGG-	8.87407	14	8.87407	1	4	1
GGTGGC	1	GGTG	6	'-GGTGGC	10.1586	10	10.1586	1	4	1
AAGGC	2	AAGG	5	'--AAGGC-	2.60165	8	2.60165	1	1	0
AAGGG	2	AAGG	5	'--AAGGG-	3.81559	7	3.81559	1	2	1
AAGGGC	2	AAGG	6	'--AAGGGC	2.51905	3	2.51905	1	2	1
CAAGG	2	AAGG	5	'-CAAGG--	6.00339	10	6.00339	1	4	1
CAAGGC	2	AAGG	6	'-CAAGGC-	4.63833	5	4.63833	1	5	0.6
CAAGGG	2	AAGG	6	'-CAAGGG-	3.93234	3	3.93234	1	6	0.47
CCAAGG	2	AAGG	6	'CCAAGG--	9.78205	7	9.78205	1	4	1
GCAAGG	2	AAGG	6	'GCAAGG--	3.93915	3	3.93915	1	4	1
CGGCGG	3	GGCGG	6	'CGGCGG-	3.80021	3	3.80021	1	5	1
GGCGG	3	GGCGG	5	'-GGCGG-	6.69827	11	6.69827	1	5	1
GGCGGG	3	GGCGG	6	'-GGCGGG	8.54717	6	8.54717	1	5	1
GGCGGT	3	GGCGG	6	'-GGCGGT	4.82628	3	4.82628	1	5	1
GGGCGG	3	GGCGG	6	'GGGCGG-	3.82127	3	3.82127	1	5	1
TGGCGG	3	GGCGG	6	'TGGCGG-	4.63774	4	4.63774	1	5	1
AGGGGC	4	GGGGC	6	'AGGGGC-	4.60244	5	4.60244	1	4	1
CGGGGC	4	GGGGC	6	'CGGGGC-	3.585	4	3.585	1	4	1
GGGGC	4	GGGGC	5	'-GGGGC-	9.78905	21	9.78905	1	4	1
GGGGCG	4	GGGGC	6	'-GGGGCG	3.97181	3	3.97181	1	4	1
TGGGGC	4	GGGGC	6	'TGGGGC-	5.3766	6	5.3766	1	4	1
CGCTGG	5	GCTGG	6	'CGCTGG-	3.97876	3	3.97876	1	4	1
GCTGG	5	GCTGG	5	'-GCTGG-	4.61807	8	4.61807	1	4	1
GCTGGA	5	GCTGG	6	'-GCTGGA	4.90465	3	4.90465	1	4	1
GCTGGG	5	GCTGG	6	'-GCTGGG	5.59212	4	5.59212	1	4	1
GGCTGG	5	GCTGG	6	'GGCTGG-	4.04369	3	4.04369	1	4	1
AGTGGG	6	GTGGG	6	'AGTGGG-	5.36048	4	5.36048	1	4	1
GTGGG	6	GTGGG	5	'-GTGGG-	4.21516	12	4.21516	1	4	1
GTGGGC	6	GTGGG	6	'-GTGGGC	2.37863	4	2.37863	1	4	1
GTGGGG	6	GTGGG	6	'-GTGGGG	7.38201	8	7.38201	1	4	1
TGTGGG	6	GTGGG	6	'TGTGGG-	3.33587	3	3.33587	1	4	1
GTAGAG	7	TAGAG	6	'GTAGAG-	4.30867	5	4.30867	1	3	1
TAGAG	7	TAGAG	5	'-TAGAG-	3.09377	7	3.09377	1	3	1
TAGAGG	7	TAGAG	6	'-TAGAGG	3.35564	3	3.35564	1	3	1
TAGAGT	7	TAGAG	6	'-TAGAGT	5.43525	4	5.43525	1	3	1
ATTGTG	8	TGTG	6	'ATTGTG--	4.05327	3	4.05327	1	2	1
TGTGT	8	TGTG	5	'--TGTGT-	3.07548	6	3.07548	1	2	1

TGTGTT	8	TGTG	6	'--TGTGTT	5.36891	4	5.36891	1	2	1
TTGTG	8	TGTG	5	'-TTGTG--	3.12078	7	3.12078	1	2	1
TTGTGT	8	TGTG	6	'-TTGTGT-	3.91375	3	3.91375	1	4	0.33
GGGGG	9	GGGGG	5	'-GGGGG-	11.2025	17	11.2025	1	3	1
GGGGGC	9	GGGGG	6	'-GGGGGC	5.94795	6	5.94795	1	3	1
GGGGGG	9	GGGGG	6	'-GGGGGG	14.5575	10	14.5575	1	3	1
TGGGGG	9	GGGGG	6	'TGGGGG-	4.60874	4	4.60874	1	3	1
AAGTGG	10	AGTG	6	'-AAGTGG	3.88041	3	3.88041	2	2	1
AGAGTG	10	AGTG	6	'AGAGTG-	3.82024	3	3.82024	2	2	1
AGTGG	10	AGTG	5	'--AGTGG	5.97609	10	5.97609	2	2	1
GAGTG	10	AGTG	5	'-GAGTG-	3.7592	7	3.7592	2	2	1
GAGTGG	10	AGTG	6	'-GAGTGG	5.3668	4	5.3668	2	4	0.33
AGAGG	11	AGG	5	'-AGAGG---	3.84579	7	3.84579	4	13	1
AGAGGA	11	AGG	6	'-AGAGGA--	4.91846	3	4.91846	4	13	1
AGGC	11	AGG	4	'---AGGC--	5.02257	32	5.02257	4	6	1
AGGG	11	AGG	4	'---AGGG--	4.73884	21	4.73884	4	4	1
AGGGAG	11	AGG	6	'---AGGGAG	3.91151	3	3.91151	4	8	0.86
CCAGGC	11	AGG	6	'-CCAGGC--	2.5084	3	2.5084	4	6	1
CTAGGC	11	AGG	6	'-CTAGGC--	2.50462	3	2.50462	4	6	1
GAGG	11	AGG	4	'--GAGG---	6.56052	26	6.56052	4	13	1
GAGGA	11	AGG	5	'--GAGGA--	3.19377	5	3.19377	4	13	1
GAGGAG	11	AGG	6	'--GAGGAG-	5.56098	4	5.56098	4	14	0.92
GAGGC	11	AGG	5	'--GAGGC--	3.69855	10	3.69855	4	17	0.68
GAGGG	11	AGG	5	'--GAGGG--	4.57631	8	4.57631	4	15	0.83
GAGGGG	11	AGG	6	'--GAGGGG-	5.4483	4	5.4483	4	15	0.83
GGAGG	11	AGG	5	'-GGAGG---	3.8143	7	3.8143	4	14	0.92
GGAGGC	11	AGG	6	'-GGAGGC--	3.56828	4	3.56828	4	18	0.64
GGAGGG	11	AGG	6	'-GGAGGG--	3.88979	3	3.88979	4	15	0.83
GGGAGG	11	AGG	6	'GGGAGG---	4.00978	3	4.00978	4	14	0.92
TAGGC	11	AGG	5	'--TAGGC--	2.18095	8	2.18095	4	6	1
TGAGG	11	AGG	5	'-TGAGG---	4.38416	9	4.38416	4	13	1
TGAGGC	11	AGG	6	'-TGAGGC--	3.26508	4	3.26508	4	17	0.68
ATATG	12	TATG	5	'-ATATG--	3.08677	6	3.08677	4	10	1
ATATGG	12	TATG	6	'-ATATGG-	8.48791	6	8.48791	4	10	1
GATATG	12	TATG	6	'GATATG--	5.43525	4	5.43525	4	10	1
GTTATG	12	TATG	6	'GTTATG--	3.23123	4	3.23123	4	10	1
TATG	12	TATG	4	'--TATG--	1.68025	15	1.68025	4	10	1
TATGA	12	TATG	5	'--TATGA-	2.33239	5	2.33239	4	10	1
TATGAC	12	TATG	6	'--TATGAC	3.88703	3	3.88703	4	10	1
TATGG	12	TATG	5	'--TATGG-	4.37004	9	4.37004	4	10	1
TATGGC	12	TATG	6	'--TATGGC	4.25135	5	4.25135	4	10	1
TTATG	12	TATG	5	'-TTATG--	2.39546	6	2.39546	4	10	1
TTATGA	12	TATG	6	'-TTATGA-	5.42875	4	5.42875	4	10	1
AAGACC	13	ACC	6	'AAGACC---	6.67681	4	6.67681	4	6	1
ACCAA	13	ACC	5	'---ACCAA-	2.32261	4	2.32261	4	3	1
ACCAAG	13	ACC	6	'---ACCAAG	3.88703	3	3.88703	4	3	1
AGACC	13	ACC	5	'-AGACC---	2.30871	4	2.30871	4	6	1
GACC	13	ACC	4	'--GACC---	2.55586	12	2.55586	4	6	1
GACCA	13	ACC	5	'--GACCA--	2.32994	4	2.32994	4	8	0.64
GACCAA	13	ACC	6	'--GACCAA-	4.79991	3	4.79991	4	8	0.64
GACCTG	13	ACC	6	'--GACCTG-	3.91881	3	3.91881	4	6	1
GGACC	13	ACC	5	'-GGACC---	3.18228	5	3.18228	4	6	1
AATCGG	14	CGG	6	'AATCGG--	4.01211	3	4.01211	4	5	1
ATCGGC	14	CGG	6	'-ATCGGC-	2.53593	3	2.53593	4	5	1
CGGGG	14	CGG	5	'---CGGGG	2.33353	5	2.33353	4	2	1
CTCGG	14	CGG	5	'-CTCGG--	2.29696	5	2.29696	4	5	1
CTCGGG	14	CGG	6	'-CTCGGG-	3.84602	3	3.84602	4	6	0.73
TCGGCG	14	CGG	6	'--TCGGCG	3.28023	3	3.28023	4	5	1
TCGGGG	14	CGG	6	'--TCGGGG	3.32851	3	3.32851	4	6	0.73
AGAA	15	AGAA	4	'--AGAA-	2.16143	11	2.16143	4	4	1
AGAAT	15	AGAA	5	'--AGAAT	3.21047	5	3.21047	4	4	1
GTAGAA	15	AGAA	6	'GTAGAA-	4.74249	5	4.74249	4	4	1
TAGAA	15	AGAA	5	'-TAGAA-	3.13329	6	3.13329	4	4	1
TAGAAT	15	AGAA	6	'-TAGAAT	5.47463	4	5.47463	4	4	1
AATTG	16	AATT	5	'--AATTG-	3.14001	6	3.14001	4	4	1
AATTGT	16	AATT	6	'--AATTGT	4.87458	3	4.87458	4	4	1
GTAATT	16	AATT	6	'GTAATT--	4.68928	5	4.68928	4	4	1
TAATT	16	AATT	5	'-TAATT--	2.33437	5	2.33437	4	4	1
TAATTG	16	AATT	6	'-TAATTG-	4.73989	4	4.73989	4	4	1
ACCTC	17	CCTC	5	'-ACCTC--	3.12256	5	3.12256	4	5	1

ACCTCC	17	CCTC	6	'-ACCTCC-	4.86105	3	4.86105	4	5	1
CACCTC	17	CCTC	6	'CACCTC--	5.06762	3	5.06762	4	5	1
CCTC	17	CCTC	4	'--CCTC--	1.66665	10	1.66665	4	5	1
CCTCC	17	CCTC	5	'--CCTCC-	2.2766	4	2.2766	4	5	1
CCTCGG	17	CCTC	6	'--CCTCGG	3.84116	3	3.84116	4	5	1
AGGGC	18	GGGC	5	'-AGGGC-	2.06499	7	2.06499	4	5	1
CTGGGC	18	GGGC	6	'CTGGGC-	3.81387	4	3.81387	4	5	1
GGGC	18	GGGC	4	'--GGGC-	8.06106	43	8.06106	4	5	1
GGGCA	18	GGGC	5	'--GGGCA	2.33091	4	2.33091	4	5	1
GGGCG	18	GGGC	5	'--GGGCG	2.32071	5	2.32071	4	5	1
TGGGC	18	GGGC	5	'-TGGGC-	3.22006	10	3.22006	4	5	1
AGTA	19	AGTA	4	'-AGTA--	2.15759	11	2.15759	4	4	1
AGTAGC	19	AGTA	6	'-AGTAGC	2.50008	3	2.50008	4	4	1
GAGTA	19	AGTA	5	'GAGTA--	2.29532	4	2.29532	4	4	1
TAGTA	19	AGTA	5	'TAGTA--	2.37311	5	2.37311	4	4	1
TAGTAG	19	AGTA	6	'TAGTAG-	3.34476	3	3.34476	4	4	1
AGTG	20	GTG	4	'-AGTG--	5.8182	24	5.8182	4	3	1
AGTGA	20	GTG	5	'-AGTGA-	4.02655	6	4.02655	4	5	0.6
AGTGAG	20	GTG	6	'-AGTGAG	5.35549	4	5.35549	4	5	0.6
CAGTG	20	GTG	5	'CAGTG--	3.05845	6	3.05845	4	3	1
GTGA	20	GTG	4	'--GTGA-	1.74258	20	1.74258	4	3	1
GTGAG	20	GTG	5	'--GTGAG	2.68747	9	2.68747	4	3	1
CGGTT	21	GGTT	5	'CGGTT--	2.2958	4	2.2958	4	4	1
CGGTTG	21	GGTT	6	'CGGTTG-	3.90035	3	3.90035	4	4	1
GGTT	21	GGTT	4	'-GGTT--	1.71456	10	1.71456	4	4	1
GGTTG	21	GGTT	5	'-GGTTG-	3.06306	6	3.06306	4	4	1
GGTTGG	21	GGTT	6	'-GGTTGG	3.92331	3	3.92331	4	4	1
AGTGTA	22	GTGT	6	'AGTGTA-	4.85298	3	4.85298	4	4	1
GTGT	22	GTGT	4	'-GTGT--	3.36222	26	3.36222	4	4	1
GTGTA	22	GTGT	5	'-GTGTA-	2.6027	8	2.6027	4	4	1
GTGTAG	22	GTGT	6	'-GTGTAG	3.26747	4	3.26747	4	4	1
GTGTT	22	GTGT	5	'-GTGTT-	2.027	7	2.027	4	4	1
CGGG	23	CGGG	4	'CGGG--	3.68909	18	3.68909	4	3	1
CGGGA	23	CGGG	5	'CGGGA-	2.28819	4	2.28819	4	3	1
CGGGT	23	CGGG	5	'CGGGT-	2.35486	4	2.35486	4	3	1
CGGGTG	23	CGGG	6	'CGGGTG	5.57367	4	5.57367	4	3	1
GCGG	24	GCGG	4	'GCGG--	3.64572	18	3.64572	4	4	1
GCGGG	24	GCGG	5	'GCGGG-	6.1336	10	6.1336	4	4	1
GCGGGC	24	GCGG	6	'GCGGGC	3.67273	4	3.67273	4	4	1
GCGGGT	24	GCGG	6	'GCGGGT	5.00325	3	5.00325	4	4	1
GCGGT	24	GCGG	5	'GCGGT-	2.29151	4	2.29151	4	4	1
AACAC	25	ACAC	5	'AACAC--	2.28582	4	2.28582	4	3	1
AACACC	25	ACAC	6	'AACACC-	5.00613	3	5.00613	4	3	1
ACACCT	25	ACAC	6	'-ACACCT	4.95198	3	4.95198	4	3	1
ACACGT	25	ACAC	6	'-ACACGT	4.82098	3	4.82098	4	3	1
AAGT	26	AAGT	4	'-AAGT-	1.74801	10	1.74801	4	3	1
AAGTG	26	AAGT	5	'-AAGTG	3.8693	7	3.8693	4	3	1
TAAGT	26	AAGT	5	'TAAGT-	2.35658	5	2.35658	4	3	1
TAAGTG	26	AAGT	6	'TAAGTG	4.70162	4	4.70162	4	3	1
AGTGGC	27	GGC	6	'AGTGGC---	3.54163	4	3.54163	5	7	1
ATGGC	27	GGC	5	'-ATGGC---	5.93424	14	5.93424	5	7	1
GATGGC	27	GGC	6	'GATGGC---	8.14456	8	8.14456	5	7	1
GGCT	27	GGC	4	'---GGCT--	2.13292	11	2.13292	5	5	1
GGCTA	27	GGC	5	'---GGCTA-	3.17886	5	3.17886	5	5	1
GGCTAG	27	GGC	6	'---GGCTAG	3.92556	3	3.92556	5	5	1
GTGGC	27	GGC	5	'-GTGGC---	6.4307	20	6.4307	5	7	1
GTGGCT	27	GGC	6	'-GTGGCT--	4.71178	5	4.71178	5	10	0.67
TGGC	27	GGC	4	'--TGGC---	6.24044	40	6.24044	5	7	1
TGGCT	27	GGC	5	'--TGGCT--	2.34889	5	2.34889	5	10	0.67
TGGCTG	27	GGC	6	'--TGGCTG-	3.2832	3	3.2832	5	10	0.67
CCGA	28	CGA	4	'-CCGA---	2.13739	11	2.13739	5	6	1
CCGAG	28	CGA	5	'-CCGAG--	2.35878	5	2.35878	5	6	1
CCGAT	28	CGA	5	'-CCGAT--	2.30823	4	2.30823	5	8	0.71
CCGATG	28	CGA	6	'-CCGATG-	3.83631	3	3.83631	5	8	0.71
CGATG	28	CGA	5	'--CGATG-	4.61974	8	4.61974	5	4	1
CGATGG	28	CGA	6	'--CGATGG	9.90067	7	9.90067	5	4	1
TCCGA	28	CGA	5	'TCCGA---	3.05931	6	3.05931	5	6	1
TCCGAG	28	CGA	6	'TCCGAG--	3.32607	3	3.32607	5	6	1
TCCGAT	28	CGA	6	'TCCGAT--	3.83954	3	3.83954	5	8	0.71
AAGGGG	29	GGG	6	'AAGGGG	3.88538	3	3.88538	5	3	1

AGGGG	29	GGG	5	'-AGGGG	4.57194	8	4.57194	5	3	1
CTGGG	29	GGG	5	'CTGGG-	3.22319	6	3.22319	5	2	1
GGGG	29	GGG	4	'--GGGG	12.2506	42	12.2506	5	3	1
TGGG	29	GGG	4	'-TGGG-	4.60088	24	4.60088	5	2	1
TGGGG	29	GGG	5	'-TGGGG	6.39366	12	6.39366	5	5	0.4
GTTGGA	30	TGGA	6	'GTTGGA--	2.49556	3	2.49556	5	4	1
TGGA	30	TGGA	4	'--TGGA--	2.55883	15	2.55883	5	4	1
TGGAC	30	TGGA	5	'--TGGAC-	2.33125	5	2.33125	5	4	1
TGGACC	30	TGGA	6	'--TGGACC	5.45924	4	5.45924	5	4	1
TTGGAC	30	TGGA	6	'-TTGGAC-	3.93915	3	3.93915	5	4	1

Table S2.4. Summary of the enriched ISRE n-mers and GCCS clustering performance

	ISRE sequences	**Random Sample**
Total n-mers	5376	5376
Probability $>Cl_{high}$ [21]	241	91
Clustered	193	64
% clustered	80.1%	70.33%
Number of Clusters	30	11

Table S2.5. Detailed comparison of GCCS clusters consensus motifs to known trans-acting factor binding sites

Class	Pictogram	Similar To
1		hnRNP F/H consensus binding site (GGGGG)[23], which functions as either a splicing enhancer or silencer[63]. Contains a G-triplet, a known ISE sequence[20] that is abundant in mammalian introns[16].
1		High affinity hnRNPA1 binding site (TAGGG) indentified by SELEX[22]. Contains a G-triplet, a known ISE sequence[20] that is abundant in mammalian introns[16].
1		Exhibits similarity to the hnRNP F/H and hnRNPA1 binding sites and may represent a weak binding site for these factors. Contains a G-triplet, a known ISE sequence[20] that is abundant in mammalian introns[16].
1		hnRNP A1 binding site (TAGAGT)[64]
1		High affinity hnRNP L binding site (CA-rich) indentified by SELEX and an ISE element comprised of variable-length CA repeats[25]. A/C-rich ESSs[8].
1		CTCC and CCTCCC repeats indentified by computational analysis of introns flanking skipped exons[65]. CT-rich intronic sequences that act as PTB binding sites[24,66].
1		Exhibits similarity to the hnRNP F/H and hnRNPA1 binding sites and may represent a weak binding site for these factors. Contains a G-triplet, a known ISE sequence[20] that is abundant in mammalian introns[16].
1		Exhibits similarity to the hnRNP F/H and hnRNPA1 binding sites and may represent a weak binding site for these factors.
1		Exhibits similarity to the hnRNP F/H and hnRNPA1 binding sites and may represent a weak binding site for these factors.
1		High affinity hnRNP A1 binding site (TAGGG) indentified by SELEX[22]. Contains a G-triplet, a known ISE sequence[20] that is abundant in mammalian introns[16].
1		Exhibits similarity to the hnRNP F/H and hnRNPA1 binding sites and may represent a weak binding site for these factors. Contains a G-triplet, a known ISE sequence[20] that is abundant in mammalian introns[16].
2		SRp30c recognition sequence (CTGGATT) that is critical for binding[14].
2		SRp40 binding site (ACAAG)[27]
2		SC35 binding site (AGGAGAT)[28]. A purine-rich element (AGGG) identified in introns flanking skipped exons[65].
2		Sam68 binding site (TAAA)[67,68].
2		9G8 high-affinity binding site (GAC) identified by SELEX[28].

2		SF2/ASF high-affinity binding site (GAAGAA) identified by SELEX[26]. Tra2β high-affinity binding site $(GAA)_n$ identified by SELEX[29].
3		Major 5'ss consensus sequence (GT[A/G]AGT)[69].
3		Major 5'ss consensus sequence (GT[A/G]AGT)[69].
3		Major 5'ss consensus sequence (GT[A/G]AGT)[69].
3		Major 5'ss consensus sequence (GT[A/G]AGT)[69]. hnRNP G binding motif (AAGT)[31].
4		CELF/Bruno-like family of proteins that bind GT repeats with high affinity[34]. CUG-BP1 binding sites consisting of TGT-repeats[34]. hnRNP M binding sites consisting of poly(G) and poly(T) homopolymers[35].
4		CELF/Bruno-like family of proteins that bind GT repeats with high affinity[34].
4		CELF/Bruno-like family of proteins that bind GT repeats with high affinity[34]. CUG-BP1 binding sites consisting of TGT-repeats[34].
4		CELF/Bruno-like family of proteins that bind GT repeats with high affinity[34].
5		N/A. A novel regulatory element.
5		N/A. A novel regulatory element.
5		CELF binding site (GT repeats)[34].
5		hnRNPA1 binding site (TAGGG)[22].
5		N/A. A novel regulatory element.

Table S2.6. Overlap of enriched hexamers with extended recovered ISRE sequences

Extended ISS sequence	Enriched Hexamers			
GTTCGAATCTCTCCAGTGC				
GTCCTACGCTCATTATTGC				
GTTCTTCTCTTCTCTTCGC				
GTTGTTCGCACCGCTGGGC	CGCTGG	CTGGGC	GCTGGG	TGTTCG
GTTGTTCGCACCACTGAGC	TGTTCG			
GTAGTCACCTATTATAGGC				
GTGTTAACCAACGATGGGC	CGATGG			
GTGGTATCGAAAGTTGTGC				
GTTACATCCAGAAGTCGGC				
GTTACATCCCTCGGTTGGC	CCTCGG	CGGTTG	GGTTGG	
GTTGGACCAGGCGTACGGC	CCAGGC	GTTGGA	TGGACC	TTGGAC
GTTGGACACGTCAGTCAGC	ACACGT	GTTGGA	TTGGAC	
GTCACACGTGAGAGAGAGC	ACACGT			
GTGAAGGGCGACAGATAGC	AAGGGC			
GTAGAACGCTGGATTAAGC	CGCTGG	GCTGGA	GTAGAA	
GTTTACTTTAAGGATAAGC				
GTATACGGAAAGGCCTTGC	GTATAC			
GTGTGCTTATATGGGTTGC	ATATGG			
GTTTAGTCCCATTCCGAGC	TCCGAG			
GTCCACTTCGGTTGCCTGC	CGGTTG			
GTACGTCCGTCGTGGATGC				
GTACCTCGAGGTCTGAAGC				
GTACCTCAGGCTCTGAAGC				
GTAAGGCTAGTTTAGTAGC	AGTAGC	GGCTAG		
GTAAGGCTAGATTAGTAGC	AGTAGC	GGCTAG		
GTAGAGGAGTCGTGTCAGC	AGAGGA	GTAGAG	TAGAGG	TGTCAG
GTAGTGGAATCGTATCAGC				

GTGGTCGAGTCGCAAGGGC	AAGGGC	CAAGGG	GCAAGG			
GTATTCCAGCTGGAGCTGC	GCTGGA					
GTAGTATATGGTGAGGAGC	ATATGG	GTGAGG				
GTGCCGAGTAAAGTGTAGC	AGTGTA	GTAAAG	GTGTAG			
GTTCTGACTCAATAGTAGC	AGTAGC					
GTCTTGAGTACCCCCGAGC						
GTCATGCACCGACCAAGGC	ACCAAG	CAAGGC	CCAAGG	GACCAA		
GTAATTGTGTTTGTGATGC	AATTGT	ATTGTG	GTAATT	TAATTG	TGTGTT	TTGTGT
GTGACTGTGTTAGGCGGGC	GGCGGG	TGTGTT				
GTAATTGGGTTTGGGGGGC	GGGGGC	GGGGGG	GTAATT	TAATTG	TGGGGG	
GTAATTGTGTTCGGTGGGC	AATTGT	ATTGTG	CGGTGG	GTAATT	GTGGGC	TAATTG
	TGTGTT	TGTTCG	TTGTGT			
GTAATTGTGTTTGGCGGGC	AATTGT	ATTGTG	GGCGGG	GTAATT	TAATTG	TGGCGG
	TGTGTT	TTGTGT				
GTTATGACATGTGGGGAGC	GACATG	GTGGGG	GTTATG	TATGAC	TGACAT	TGTGGG
	TTATGA					
GTTATGACGTGTGGGGGGC	GGGGGC	GGGGGG	GTGGGG	GTTATG	TATGAC	TGGGGG
	TGTGGG	TTATGA				
GTTATGACATGTGGGGGGC	GACATG	GGGGGC	GGGGGG	GTGGGG	GTTATG	TATGAC
	TGACAT	TGGGGG	TGTGGG	TTATGA		
GTCAATTGAGTTGGTGTGC						
GTCGATGGGGCAGGGGAGC	CGATGG	TGGGGC				
GTCAGTGAACTTTGCGAGC	TCAGTG	TTTGCG				
GTCCTTGGTCCTGACATGC	GACATG	TGACAT				
GTCCGAGTGCGACGGTGGC	CGGTGG	GGTGGC	TCCGAG			
GTGAGTGGCCTAGGGAGGC	AGGGAG	AGTGGC	GAGTGG	GGAGGC	GGGAGG	TAGGGA
	TAGTAG					
GTGGCTGGGCTAGGATGGC	CTGGGC	GATGGC	GCTGGG	GGCTAG	GGCTGG	GTGGCT
	TGGCTG					
GTGATATGGCGAGGGTGGC	ATATGG	GATATG	GGGTGG	GGTGGC	TATGGC	
GTAAGTGGGCACGGTTGGC	AAGTGG	AGTGGG	CGGTTG	GGTTGG	GTGGGC	TAAGTG
GTAGGTAGCCACCGTTGGC	ACCGTT					
GTGGGGGGGTCACTTAGGC	GGGGGG	GTGGGG	TGGGGG			
GTTGGTTGGACCCGTAGGC	GGTTGG	GTTGGA	TGGACC	TTGGAC		
GTCCCTATGGTTCCTCGGC	CCTCGG					
GTCAGAGGAGTCTCTAGGC	AGAGGA	CTAGGC				

GTTTATGGAGTTCCTAGGC	CTAGGC					
GTAAATAGAGGCCCCAGGC	CCAGGC	TAGAGG				
GTCTAGTAACCAGCCAGGC	CCAGGC					
GTCTAAGCACCACTGAGGC	TGAGGC					
GTTGTTTTGCGTCCAAGGC	CAAGGC	CCAAGG	TTTGCG			
GTCATGTCAGGACCAAGGC	ACCAAG	CAAGGC	CCAAGG	GACCAA	TGTCAG	
GTCATGGACCGACCAAGGC	ACCAAG	CAAGGC	CCAAGG	GACCAA	TGGACC	
GTTATGCCTCCCCGATAGC	GTTATG					
GTCGAAGAACCCCAAGGGC	AAGGGC	CAAGGG	CCAAGG			
GTCGGAGAAACCGGAGGGC	GGAGGG					
GTCCGAGGAACCATAGGGC	TCCGAG					
GTCTATCTCCTTCTATGGC	TATGGC					
GTTTAACACCTCCCAAGGC	AACACC	ACACCT	ACCTCC	CAAGGC	CACCTC	CCAAGG
GTCAAAGACCTGCGATGGC	AAGACC	CGATGG	GACCTG	GATGGC		
GTCAAACACGTCCGATGGC	ACACGT	CCGATG	CGATGG	GATGGC	TCCGAT	
GTCTAACACCTCCGATGGC	AACACC	ACACCT	ACCTCC	CACCTC	CCGATG	CGATGG
	GATGGC	TCCGAT				
GTCAAACACCTCCGATGGC	AACACC	ACACCT	ACCTCC	CACCTC	CCGATG	CGATGG
	GATGGC	TCCGAT				
GTGTGGCTATGAATTTGGC	GTGGCT					
GTGTGGCTAAGAATTGGGC	GTGGCT					
GTGGCTGGAAGACCTGCGC	AAGACC	GACCTG	GCTGGA	GGCTGG	GTGGCT	TGGCTG
GTGTAAAGGGTGTCAGTGC	GTAAAG	TCAGTG	TGTCAG			
GTATTAATAATACTGGGGC	TGGGGC					
GTGTTAATAGCGCGGGAGC						
GTTTGTAAGGTGCTGGGGC	GCTGGG	TGGGGC				
GTTGTGGTCGCGACCTGGC	GACCTG					
GTGGCGGTCGAGTACAGGC	GGCGGT	GTACAG	TGGCGG			
GTGTTGTGAAAGAGGAGGC	AGAGGA	GGAGGC				
GTGGTGGCAGACACGATGC	GGTGGC					

GTGCGGTTTGCGGGCGGGC	GGCGGG	GGGCGG	TTTGCG			
GTGGGGCGCGCGGGGGGGC	GGGGCG	GGGGGC	GGGGGG	GTGGGG	TGGGGC	
GTGAGGGCAGTCCGTGGGC	GTGAGG	GTGGGC				
GTGACGGGTGCCTCGGGGC	CCTCGG	CGGGGC	CGGGTG	CTCGGG	TCGGGG	
GTTAGGTGTGTCTCGGGGC	CGGGGC	CTCGGG	TCGGGG			
GTGACGTGTGTCTCGGGGC	CGGGGC	CTCGGG	TCGGGG			
GTGACGGAGCCGTCTGGGC	CTGGGC					
GTGCATGGCCCCGCTGGGC	CGCTGG	CTGGGC	GCTGGG			
GTGCAAGGTCCCTCTAGGC	CTAGGC	GCAAGG				
GTGCACTAGAATCTGAGGC	TAGAAT	TGAGGC				
GTGCAGTACGGGCTTAGGC						
GTCGAGCGGCTTTAGAGGC	TAGAGG					
GTAGAGTGGGGCGGGTGGC	AGAGTG	AGTGGG	CGGGTG	GAGTGG	GGCGGG	GGGCGG
	GGGGCG	GGGTGG	GGTGGC	GTAGAG	GTGGGG	TAGAGT
	TGGGGC					
GTATAGTGGCGGTGGAGGC	AGTGGC	CGGTGG	GGAGGC	GGCGGT	TGGCGG	
GTAGAGTGGCGGTGGAGGC	AGAGTG	AGTGGC	CGGTGG	GAGTGG	GGAGGC	GGCGGT
	TAGAGT	TGGCGG	GTAGAG			
GTAGAATGGACCGTGAGGC	GTAGAA	GTGAGG	TAGAAT	TGAGGC	TGGACC	
GTGGAGTGGCTGGTTCGGC	AGTGGC	GAGTGG	GAGTGG	GGCTGG	GTGGCT	TGGCTG
GTGTACAGCGGAGAGGGGC	AGGGGC	GTACAG				
GTGTACGGTGCAGAGGGGC	AGGGGC					
GTGTAGTGTAGGGAGGGGC	AGGGAG	AGGGGC	AGTGTA	GGAGGG	GGGAGG	GTGTAG
	TAGTAG	TAGGGA				
GTGAAGTGTAGGGAGGGGC	AGGGAG	AGGGGC	AGTGTA	GGAGGG	GGGAGG	GTGTAG
	TAGTAG	TAGGGA				
GTATACCGTTCAGTGGGGC	ACCGTT	AGTGGG	GTATAC	GTGGGG	TCAGTG	TGGGGC
GTATACCGTTCAGTGAGGC	ACCGTT	AGTGAG	GTATAC	GTGAGG	TCAGTG	TGAGGC
GTAAAGGGGCAAGGTGGGC	AAGGGG	AGGGGC	AGGTGG	GCAAGG	GTAAAG	GTGGGC
GTAGAGTGCGAAGCGGGGC	AGAGTG	CGGGGC	GTAGAG	TAGAGT		
GTACAGTGCTAAGTAGGGC	GTACAG					
GTGTAAATCGGCGGGTGGC	AATCGG	ATCGGC	CGGGTG	GGCGGG	GGGTGG	GGTGGC
	TCGGCG	GGTGGC				
GTGGAAATCGGCGGATGGC	AATCGG	ATCGGC	GATGGC	TCGGCG		

GTGGCAATCGGCGGGTGGC	AATCGG TCGGCG	ATCGGC	CGGGTG	GGCGGG	GGGTGG	GGTGGC
GTAAAGAACGGGATATGGC	ATATGG TATGGC	CGGGAT	GATATG	GGATAT	GGGATA	GTAAAG
GTCAAGACCGGGATATGGC	AAGACC TATGGC	ATATGG	CGGGAT	GATATG	GGATAT	GGGATA
GTAAAGACCGGGATATGGC	AAGACC TATGGC	ATATGG GTAAAG	CGGGAT	GATATG	GGATAT	GGGATA
GTAATTATTAGTCGATGGC	CGATGG	GATGGC	GTAATT			
GTGCTTAGTGAGTGATGGC	AGTGAG	GATGGC				
GTACAGGCCAAGGGGGGGC	AAGGGG	CAAGGG	CCAAGG	GGGGGC	GGGGGG	GTACAG
GTAGAAGACAAGTGGTGGC	AAGTGG	GGTGGC	GTAGAA			
GTGGTTGAAGGGGGGCGGC	AAGGGG	GGGCGG	GGGGCG	GGGGGC	GGGGGG	
GTACATTATGAGGGTCGGC	TTATGA					
GTAGAGTAAGTGAGGTGGC	AGGTGG TAGAGT	AGTGAG	GGTGGC	GTAGAG	GTGAGG	TAAGTG
GTAGAATAAGTGAGGTGGC	AGGTGG TAGAAT	AGTGAG	GGTGGC	GTAGAA	GTGAGG	TAAGTG
GTAGAATAAGTGGGGTGGC	AAGTGG TAGAAT	AGTGGG TAAGTG	GGGTGG	GGTGGC	GTAGAA	GTGGGG

Table S2.7. Primer and oligonucleotide sequences

Name	**Primer Sequence (5 ’- 3’)**
Ex6	CATGGACGAGCTGTACGTTAACATAATTCCCCCACCACCTC
Ex8	CGCTCG AGCACATACGCCTCACATACATTTTG
GFP1	GCGGTACCATGGTGAGCAAGGGCG
GFP2	GGTGGTGGGGGAATTATGTTAACGTACAGCTCGTCCATGCC
ECmutF	CTTTTTAACATCCATATAAAGCTATCGATATCTAGCTATCGAT GTCTATATAGCTATTTTTTTTAACT
ECmutR	AGTTAAAAAAAAATAGCTATATAGACATCGATAGCTAGATATCG ATAGCTTTATATGGATGTTAAAAAG
PmlImutF	CATTATGAAAGTGAATCTTACTTTTGTAACACGTGATGGTTTG TGGAAAACAAATGTTTTTGAA
PmlImutR	TTCAAAAACATTTGTTTTCCACAAACCATCACGTGTTACAAAAG TAAGATTCACTTTCATAATG
BamHImutF	CTTTTGTAACACGTGATGGTTTGTGGGATCCAAATGTTTTTGAA CAGTTAAAAAGTTC
BamHImutR	GAACTTTTTAACTGTTCAAAAACATTTGGATCCCACAAACCATC ACGTGTTACAAAAG
P2	TAAGAAGCTAAAGAGCCTCACTCATGTGGTTTTATGCAGC
P3	TGAGGCTCTTTAGCTTCTTA
P4	AGATAGAGAGGTCAGCGATTTGCAATTCTGAGGTGTTAAA
P5	AATCGCTGACCTCTCTATCT
Ex17	CATGGACGAGCTGTACGTTAACATGCTCGTGTACAAGTTTGCC
Ex19	CGCTCGAGAAGTACTTACCTCATTCAGCATTTTTC
GFP3	GCAAACTTGTACACGAGCATGTTAACGTACAGCTCGTCCATGCC
ECmutF2	TTTAGCTTCTTAGGATATCACTTATCGATTTTGTTTTCAAC
ECmutR2	GTTGAAAACAAAATCGATAAGTGATATCCTAAGAAGCTAAA
ISStemp	GCGCGATATCGATCAGT (N_{15}) GCATCATCGATGCGC
Lib1	GCGCGATATCGATCAGT
Lib2	GCGCATCGATGATGC
Lib3	GAAACAAAATGCTTTTTAACATCCATA
Lib4	GGAAAATAAAAGGAAGTTAAAAAAAATAGC
SMN1cDNA	TAGAAGGCACAGTCGAGG

Table S2.8. Plasmid constructs used in this work

Name	**Description**
pCS238	GFP-SMN1. Contains the wild-type SMN1 mini-gene fused to the N-terminus of GFP. Positive control used for all flow cytometry analysis.
pCS516	SMN1 NMD-based reporter construct. Contains the SMN1 mini-gene with a PTC in exon 7 fused to the N-terminus of GFP. Recovered ISREs as well as control ISS were inserted into this construct.
pCS517	SMN1 NMD-based containing random 15-mer. Negative control used for all flow cytometry analysis.
pCS990	GFP-BRCA1. Contains the wild-type BRCA1 mini-gene fused to the N-terminus of GFP. Positive control used for flow cytometry analysis.
pCS1008	BRCA1 NMD-based reporter construct. Contains the BRCA1 mini-gene with a PTC in exon 17 fused to the N-terminus of GFP. Recovered ISREs were inserted into this construct.
pCS668	U2AF65 binding site inserted into pCS516.
pCS669	hnRNP H binding site inserted into pCS516.
pCS670	PTB (1) binding site inserted into pCS516.
pCS667	PTB (2) binding site inserted into pCS516.

Table S2.9. Primer sequences for SMN1 transcript isoform analysis through qRT-PCR

Name	**Forward Primer (5' - 3')**	**Reverse Primer (5' - 3')**	**Isoform**
Pair 1	CTCCCATATGTCCAGATCT	AGCATTTTGTTTCACAAGACA	Ex 6 and Int 6
Pair 2	CACTAGTAGGCAGACCAG	CAGTTATCTTCTATAACGCTTCAC	Int 7 and Ex 8
Pair 3	TAAATTAAGGAGAAATGCT	GGTTTTTCAAAAGAGTCCAGTAA	Ex 7/8 and Ex 8
Pair 4	TGAGCAAAGACCCCAA	CCAGCATTTCCATATAATAG	GFP and Ex 6/8
Pair 5	TGAGCAAAGACCCCAA	TGATAGCCACTCATGTACC	GFP and Ex 6
Pair 6	CAAAGATGGTCAAGGTCGCAAG	GGCGATGTCAATAGGACTCC	HPRT

Table S2.10. Primer sequences for endogenous transcript isoform analysis through qRT-PCR

Name	Gene	Hexamer	Sequence (5' – 3')	Isoform	Type of Alternative splicing
Fw.ADD3ex15_16	ADD3	ACCTCC	TGAAAAATTAGAAGAA AACCATGAGC	Exon15/16	cassette
Fw.ADD3ex14_16	ADD3	ACCTCC	GGCC TAG AAGAAA ACCATG AGC	Exon14/16	cassette
Rv.ADD3ex16	ADD3	ACCTCC	CTTCGATTTTCTCTGGA GACT	ADD3 cDNA, Exon15/16, Exon 14/16	cassette
Fw.hnRNPCex1_3	HNRNPC	ACCTCC	CCC CTT CTT GTT TTC GGC TTT	Exon1/3	cassette
Fw.hnRNPCex2_3	HNRNPC	ACCTCC	CTT CAGCTACATTTT C GGCTTT	Exon2/3	cassette
Rv.hnRNPCex3	HNRNPC	ACCTCC	CGAAAAGATTGCCTCC ACAT	hnRNPC cDNA and Exon1/3, Exon 2/3	cassette
Fw.CLK3ex4	CLK3	GGGGGG	CCGTGACAGCGATACA TAC	Exon 4/5, Exon4/6	cassette
Rv.CLK3ex4_5	CLK3	GGGGGG	GTTGGCTTCTCGAGGAG G	Exon 4/5	cassette
Rv.CLK3ex4_6	CLK3	GGGGGG	CCACAATCTCATCGAG GAGG	Exon 4/6	cassette
Rv.CLK3cDNA	CLK3	GGGGGG	CAAGCACTCCACCACCT	CLK3 cDNA	cassette
Fw.CADPSex16	CADPS	GGGGGG	GAAAGATATTGTTACCC CAGT	Exon 16/19, Exon16/18, Exon16/17	mutually exclusive
Rv.CADPSex16_18	CADPS	GGGGGG	CCTTTTGATTCTCTTCG ATTTTG	Exon16/18,	mutually exclusive
Rv.CADPSex16_19	CADPS	GGGGGG	GGCCTACATTTTCTTCG ATTTTG	Exon 16/19	mutually exclusive
Rv.CADPSex16_17	CADPS	GGGGGG	CTCTCTTTTTCCCTTCG ATTTTG	Exon16/17	mutually exclusive
Rv.CADPScDNA	CADPS	GGGGGG	AAG CTT TTT GGC AGG AGT GA	CADPS cDNA	mutually exclusive
Fw.c6orf60ex15_16	C6orf60	GTAGAA	CTTTACAAGTGTCATTA GAAGAAATG	Exon 15/16	cassette
Fw.c6orf60ex14_16	C6orf60	GTAGAA	CCA ACA GAT AAG ATT AGA AGA AAT GG	Exon 14/16	cassette
Rv.c6orf60ex16	C6orf60	GTAGAA	GATCTGGTCTCTTTCTG TAAGC	C6orf60 cDNA, Exon 15/16, Exon 14/16	cassette
Fw.RREB1ex11_12	RREB1	GTAGAA	GATAGCACAGACAGTC AGTCG	Exon11/12	cassette
Fw.RREB1ex10_12	RREB1	GTAGAA	ACA CAC ACT GAC AGT CAG TCG	Exon10/12	cassette
Rv.RREB1ex12	RREB1	GTAGAA	CTCCTCCTCCGGCTCAT	RREB1 cDNA, Exon11/12,	cassette

				Exon10/12	
Fw.MADDex35	MADD	GCTGGG	AGTTCCCTGTGCGAC	Exon35/36, Exon35/37	cassette
Rv.MADDex35_36	MADD	GCTGGG	TCTATGAAAACCTGATTGTGCA	Exon35/36	cassette
Rv.MADDex35_37	MADD	GCTGGG	TAATTTCAGGAACTGATTGTGCA	Exon35/37	cassette
Rv.MADDcDNA	MADD	GCTGGG	TAGTACAGCTCCCGACACTT	MADD cDNA	cassette
Fw.CAMK2Gex13_14	CAMK2G	GCTGGG	CGGGCAAGCTGCCAAAAG	Exon13/14	cassette
Fw.CAMK2Gex12_14	CAMK2G	GCTGGG	GAA CTT CTC AGC TGC CAA AAG	Exon12/14	cassette
Rv.CAMK2Gex14	CAMK2G	GCTGGG	TTGACACCGCCATCCG	CAMK2G cDNA, Exon13/14, Exon12/14	cassette
Fw.A2BP1ex15_17	A2BP1	ATATGG	GCAGACATTTATGGTGGTTATG	Exon15/17	mutually exclusive
Fw.A2BP1ex16_17	A2BP1	ATATGG	TAA ATT GCT GCA GGG TGG TTA TG	Exon16/17	mutually exclusive
Rv.A2BP1ex17	A2BP1	ATATGG	CTGTCACTGTAGGCAGCG	A2BP1 cDNA, Exon15/17, Exon16/17	mutually exclusive
Fw.HNRNPA2B1ex1	HNRNPA2B1	ATATGG	CTCTAGCGGCAGTAGCA	Exon1/2, Exon1/3	cassette
Rv. HNRNPA2B1ex1_2	HNRNPA2B1	ATATGG	GTTTCTAAAGTTTTCTCCATCGCG	Exon1/2	cassette
Rv. HNRNPA2B1ex1_3	HNRNPA2B1	ATATGG	GTTCCTTTTCTCTCTCCATCGC	Exon1/3	cassette
Rv. HNRNPA2B1cDNA	HNRNPA2B1	ATATGG	CCTCAAACTTTCTTCTGTGG	HNRNPA2B1 cDNA, Exon1/2, Exon1/3	cassette

Table S2.11. siRNA duplex sequences

Targeted mRNA	**Target Sequence (5'- 3')**
hnRNP H	GAUCCACCACGAAAGCUUA
hnRNP A1	CAACUUCGGUCGUGGAGGA
PTB	CGUCAAAGGAUUCAAGUUC
CUG-BP1	GAGCCAACCUGUUCAUCUA
SF2/ASF	CGUGGAGUUUGUACGGAAA

Chapter III. Engineering complex phenotypes by reprogramming alternative splicing

Abstract

Alternative splicing affects more than 90% of human genes, increases proteomic diversity, controls gene expression and contributes significantly to the high level of phenotypic complexity in mammals[1]. The ability to program alternative splicing patterns will provide a powerful tool to interrogate and manipulate cellular function. Toward this aim, we have developed a novel framework for the construction of single input-single output RNA control devices based on the direct coupling of a sensor component, composed of a protein-binding RNA aptamer and an actuator component, composed of an alternatively spliced transcript. These devices function to sense changes in intracellular protein concentrations to regulate alternative splicing and the expression of any gene. This framework can be extended to process multiple inputs and to rewire endogenous signal transduction pathways to create novel regulatory networks for user defined phenotypes. Our devices can therefore integrate cellular information by linking external stimuli to complex phenotypes, thus controlling cellular function. This platform provides a novel class of RNA-based "smart" therapeutics towards the treatment and diagnosis of disease.

3.1. Introduction

Normal cellular functions depend on the tight regulation of post-transcriptional gene regulatory mechanisms. One such process is alternative splicing, which produces multiple protein isoforms from a single gene by altering the ways in which exons are joined from a single pre-mRNA[2]. The regulation of alternative splicing is critical for the temporal and spatial expression of cellular factors[3]. Splicing patterns are tightly regulated by the interplay between auxiliary *cis*-acting elements and the trans-acting factors that modulate them, leading to a 'splicing code' that is susceptible to mutations and misregulation[4]. Up to 50% of disease-causing mutations affect splicing yielding a variety of therapeutic targets, shedding light on the need for the development of tools that reprogram splicing decisions[3]. Recently, synthetic RNA-based regulatory systems that process small molecule inputs have been developed towards the regulation of pre-mRNA splicing in yeast and mammalian cells [5]. The extension of these designs as cellular therapeutics has been hampered by the limited number of aptamer domains that can be used in the engineering of these platforms and the utility of existing ligands which are restricted by the lack of cell permeability and toxicity to cells [5]. Moreover, there are only a few examples of synthetic and naturally occurring protein-dependent RNA-based regulatory systems [6–8]. Since much of cellular physiology is regulated by protein expression and activity, the ability to interrogate and regulate these processes is an important goal in biological and medical research. Given the unique properties of synthetic RNA-based regulatory systems that are suitable for therapeutic applications, including tunable regulation and rapid response times to input availability[8], the creation of such a system towards the regulation of alternative splicing based on sensing protein biomarkers would have broad applications in biotechnology and medicine.

3.2. Results

Here, we describe a novel framework for the construction of single input-single output RNA control devices based on the direct coupling of a sensor component, composed of a protein-binding RNA aptamer and an actuator component, composed of an alternatively spliced transcript. This framework consists of a reporter construct that expresses the gene encoding the green fluorescent protein (GFP) fused 3'of a three exon, two-intron mini-gene. The alternatively-spliced transcript contains a middle exon that harbors a stop codon and a protein binding aptamer positioned in either of two introns (Figure 3.1a). Regulation of this device is exerted by RNA aptamers that sense intracellular changes in protein concentration and alter splice site selection through steric hindrance of spliceosomal components either at the 3'or 5' splice sites (ss). Therefore, cells with a high level of exon inclusion should display lower GFP fluorescence than cells in which this exon is excluded.

We implemented the platform with the SMN1 mini-gene as an actuator and the MS2 coat protein binding aptamer as the sensor component[9,10]. To determine the utility of this device, we preformed a systematic experimental analysis of aptamer mediated regulation of splicing by inserting the MS2 coat protein aptamer at various positions upstream the 3'ss and downstream the 5'ss (Figure 3.1b). Starting at both consensus splice sites the MS2 aptamer was inserted into 6 separate positions spaced by 15-nucleotides (nts). The resulting constructs were stably integrated into HEK-293 FLP-In cells to generate isogenic cell lines. To assess the effect of the MS2 coat protein on reporter gene splicing, we transiently expressed an MS2 coat protein fusion with monomeric DsRed and an SV40 NLS in these stable cell lines (Figure 3.1b). The nuclear

localization of the MS2-DsRed fusion was confirmed by confocal microscopy (data not shown). In the presence of the MS2-DsRed fusion we find that 3 positions exhibit significant silencer activity ($P < 0.05$) near both splice sites, while 2 near the 3'ss and 1 near the 5'ss exhibit enhancer activity as compared to the expression of DsRed alone ($P < 0.05$) (Figure 3.1c). These results demonstrate that aptamer-regulated control is position dependent and likely occurs through steric hindrance of spliceosome assembly.

To examine changes in splicing patterns, we analyzed the transcript isoforms of the devices and controls in the presence and absence of the MS2-DsRed fusion by qRT-PCR. The exon exclusion to inclusion ratios for each device (Figure 3.1d) significantly correlates with the observed relative device expression in Figure 3.1c ($P << 0.01$, ANOVA). To examine the regulatory effect of the insertion of the MS2 aptamer into the SMN1 transcript we compared the fluorescence of our devices to the wildtype mini-gene in the absence of the MS2-DsRed fusion. The aptamer secondary structure alone has a significant effect on the regulation of the SMN1 transcript, where the majority of positions display significant silencer activity (Figure 3.1e). Interestingly, the effect of RNA secondary structure was insignificant at position 3 (MS2-3) which displayed significant silencer activity in our expression studies; whereas the presence of the MS2 aptamer at position 10 (MS2-10) had the opposite effect of what was observed in the expression studies. Therefore, the modulation of RNA secondary structure within introns of alternatively spliced transcripts may be a component of future design efforts for these frameworks. The exon exclusion to inclusion expression ratios confirm the regulatory activity of our synthetic devices observed by fluorescence measurements and that secondary structure alone can modulate splicing patterns.

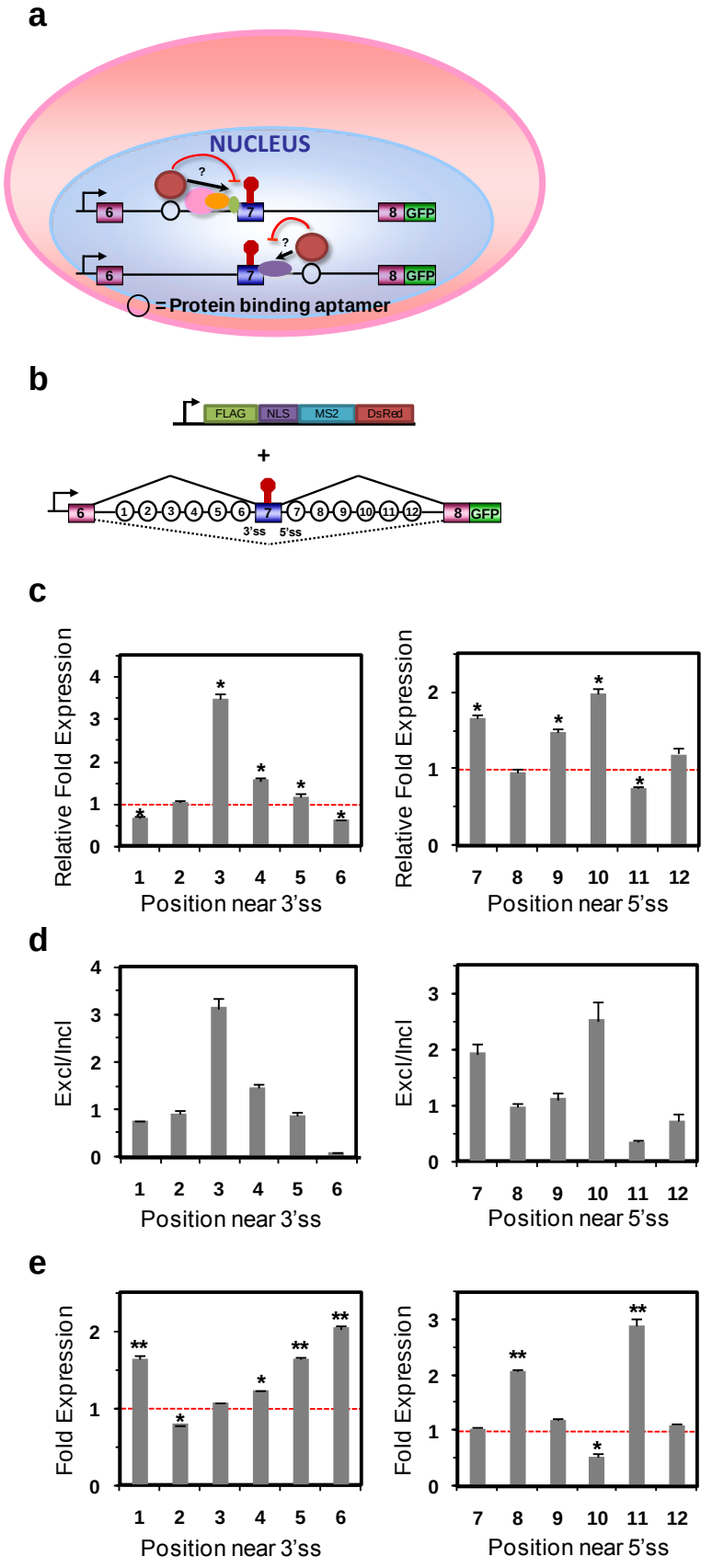

a
NUCLEUS
6
7
8 GFP
= Protein binding aptamer
b
FLAG
NLS
MS2
DsRed
+
3'ss
5'ss
c
Relative Fold Expression
Position near 3'ss
Position near 5'ss
d
Excl/Incl
Position near 3'ss
Position near 5'ss
e
Fold Expression
Position near 3'ss
Position near 5'ss

Figure 3.1. RNA device framework and analysis. **(a)** Aptamer-based splicing regulatory platform. The framework consists of a fluorescent reporter protein (GFP) fused 3' of the SMN1 mini-gene. The SMN1 mini-gene contains an alternative exon (exon 7) that harbors a stop codon and a protein binding aptamer in either of two introns. **(b)** Platform used for the systematic analysis of aptamer mediated regulation of alternative splicing and heterologous MS2-DsRed fusion construct. Starting at both consensus splice sites, the MS2 aptamer was inserted into 6 separate positions (1-6 near 3'ss and 7-12 near 5'ss) spaced by 15-nucleotides. The FLAG-NLS-MS2-DsRed fusion construct was expressed in the HEK-293 stable cell lines contacting these devices. **(c)** Flow cytometry analysis of HEK-293 FLP-In stable cell lines containing the MS2 regulatory devices. For all reported activities, the mean GFP levels from two independent experiments were determined and normalized to the mutant devices in the presence and absence of ligand. Relative fold expression and average error are reported. *P*-values derived from the Student's t-test are as follows: $*P < 0.05$ and $**P < 0.01$. **(d)** qRT-PCR analysis of the MS2 regulatory devices with primer sets specific for exon 7 included and excluded products. Expression levels of duplicate PCR samples were normalized to the levels of *HPRT*. Data is reported as the ratio of the mean expression of the exon excluded isoform to the exon included isoform normalized to the ratio for the mutant control device ± the average error. **(e)** Flow cytometry analysis of HEK-293 FLP-In stable cell lines containing the MS2 regulatory devices normalized to the wild-type SMN1 mini-gene in the absence of the MS2-DsRed fusion.

To demonstrate the functional modularity of our platform and to investigate the ability to rationally reprogram natural signal transduction pathways using our devices, we replaced the MS2 sensor component with RNA aptamers towards β-catenin, NF-κB p50 and NF-κB p65. Initially to examine the modularity of our device we replaced the MS2 aptamer with an RNA aptamer towards β-catenin as well as mutant aptamer that is the reverse compliment of the β-catenin aptamer at the positions exhibited the most significant suppression and enhancement of splicing by MS2; 3 and 6 within our framework (Figure 3.1b)[11]. β-catenin is a central component of the well-characterized Wnt pathway that regulates cell growth and differentiation during embryonic development and tumorigenesis [11]and during tumorigenesis β-catenin is nuclear localized to regulate the transcription of oncogenic target genes [11]. The β-catenin devices were stably integrated into HEK-293 FLP-In cells and leukotriene D_4 (LTD_4) was added to induce β-catenin signaling and translocation to the nucleus (Figure 3.2a). The regulatory activity of these devices was assessed over a 48 h time period in the presence of LTD_4 compared to control constructs which contain a mutant β-catenin aptamer (B-catΔ) (Figure 3.2b). The B-cat-6 device response demonstrates increasing gene expression over time ($P < 0.05$), corresponding to an increase in exon skipping (Figure 3.2b). These results demonstrate component modularity of our platform and the ability to sense and rewire endogenous signal transductions pathways. In contrast, the B-cat-3 device does not respond to LTD_4 stimulated β-catenin signaling. The MS2 device studies demonstrated exon enhancement at position 6, where as the B-cat-6 device shows increased skipping. As compared to the MS2 studies, the β-catenin device responses suggest that different proteins ligands likely have varying positional effects on splicing as

result of protein-protein interactions, steric hindrance of spliceosome machinery and specific RNA structural modification resulting from ligand binding.

To further examine component modularity and to reprogram additional signaling pathways, we inserted aptamers towards NF-κB p50 and NF-κB p65 into position 3 of our platform (Figure 3.2c). The NF-κB p50 and p65 dimers play a significant role in inflammation and disease by binding to κB sites in promoters or enhancers of target genes[12]. Two aptamers towards NF-κB p50 (p50(1) and p50(2)) and one towards p65 were individually inserted into our platform as were mutant versions of these aptamers [12-14]. These devices were stably integrated into HEK-293 FLP-In cells and tumor necrosis factor alpha (TNF-α) was supplemented to induce NF- κB signaling and subsequent translocation of NF- κB to the nucleus (Figure 3.2c)[15]. The regulatory activity of these devices was assessed over a 48 h time period (Figure 3.2d). The p65-3 device displays increasing gene expression over time ($P < 0.01$), corresponding to an increase in exon skipping (Figure 3.2d). These results further demonstrate the component modularity of our platform and regulatory functionalities in which different ligands can achieve significant suppression of splicing at a single site.

Interestingly, both p50-3 devices show a decrease in gene expression ($P < 0.05$), correlating with a decrease in exon skipping. To determine whether the NF- κB devices response to TNF-α is directly mediated by NF- κB, we developed a p50-DsRed fusion construct (Figure 3.2e). The p50-DsRed fusion construct was transiently expressed in the stable cell lines containing the NF- κB p50 devices and regulatory activity of these devices was assessed in the presence of the fusion and absence of the fusion compared to mutant aptamer containing control constructs (Figure 3.2f). Both device output responses

(p50(1) and p50(2)) demonstrate decreasing gene expression ($P < 0.05$) and correlate well with the observed changes in gene expression seen with TNF-α stimulated NF- κB regulation (Figure 3.2d). Furthermore, these results suggest that our signal transduction device responses are due to specific effector-mediated gene regulatory effects imparted by highly specific target recognition abilities and that different protein binding at a given site will impart differential regulation. As such, aptamer position and the target protein will provide key tuning capabilities for our devices.

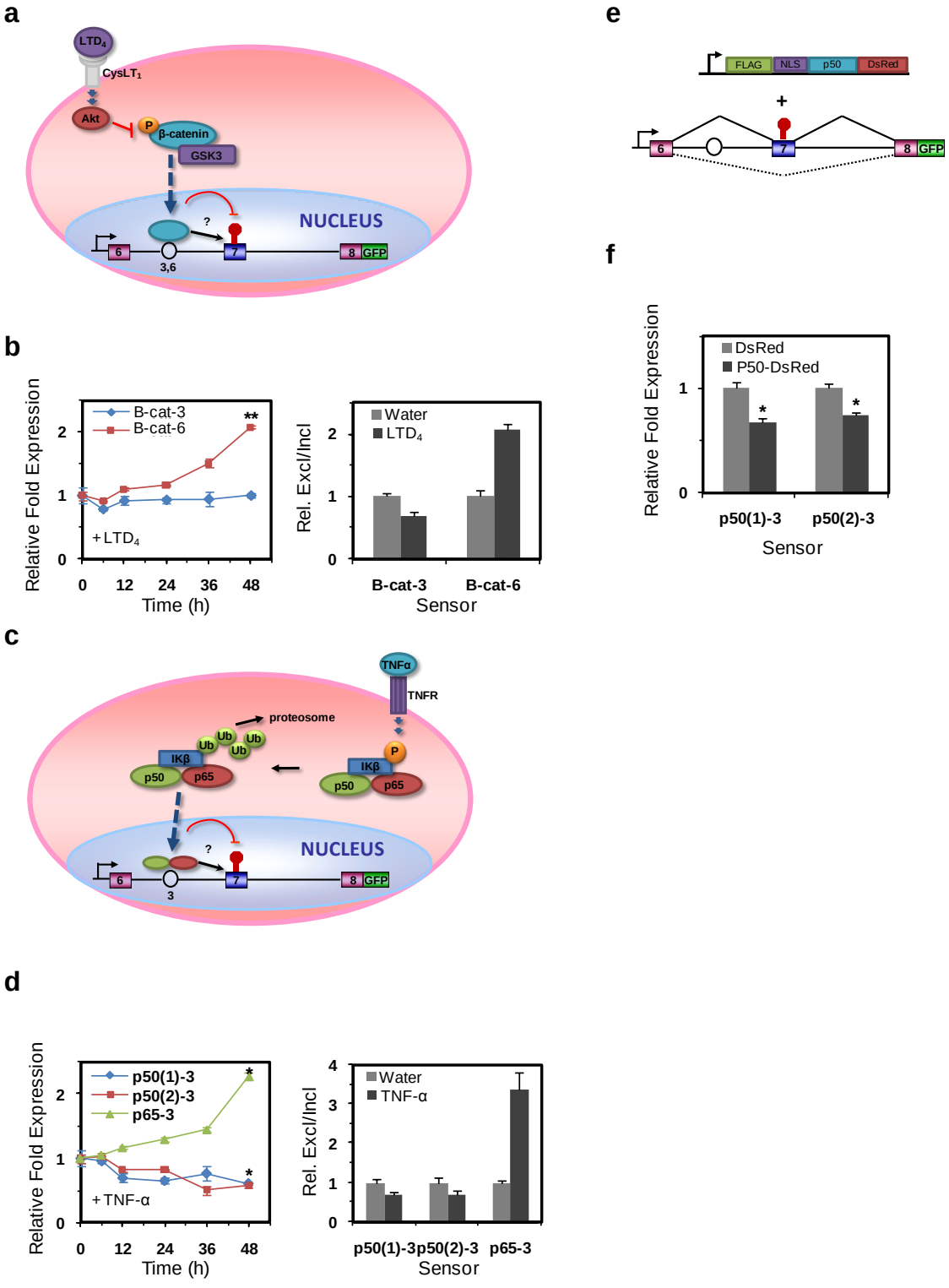


Figure 3.2. RNA device component modularity and reprogramming of endogenous signaling pathways. **(a)** Schematic of the translocation of β-catenin into the nucleus upon

stimulation by LTD_4 where binding of the RNA aptamer alters splice site choice. **(b)** Flow cytometry and qRT-PCR analysis of HEK-293 FLP-In stable cell lines containing the β-catenin regulatory devices (B-cat-3 and B-cat-6). Cells were stimulated with 80 nM LTD_4 for 48 h. **(c)** Schematic of the translocation of NF-κB dimers p50 and p65 into the nucleus upon stimulation by TNF-α where binding of the RNA aptamer alters splice site choice. **(d)** Flow cytometry and qRT-PCR analysis of HEK-293 FLP-In stable cell lines containing the NF-κB regulatory devices (p50(1)-3, p50(2)-3 and p65-3). Cells were stimulated with 20ng/ml TNF-α for 48 h. **(e)** Platform used for the analysis of NF-κB p50 mediated regulation of alternative splicing and heterologous p50-DsRed fusion construct. **(f)** Flow cytometry analysis of HEK-293 FLP-In stable cell lines containing the NF-κB p50 regulatory devices. Regulatory activity was assessed in the presence of the p50-DsRed fusion and absence of the fusion (DsRed only) compared to mutant aptamer containing control constructs.

To enable the design of complex gene circuits, there is a need for modular genetic components that can be used to sense different stimuli and generate stimulus specific phenotypes. To extend our regulatory platform towards multiple input processing we integrated both wildtype and mutant MS2 sensor components into positions 3 and 10 of the framework (Figure 3.3a). These devices were stably integrated into HEK-293 and the output response of the devices was examined in the presence and absence of the MS2-DsRed fusion. The regulatory response of the device is low in the absence of the MS2-DsRed fusion (Figure 3.3b). In contrast, the responses at both positions are equivalent with a significant increase in gene expression ($P < 0.01$). The combination of both inputs

yields a slightly higher response (~.5 fold) than the individual inputs ($P < 0.01$). The exon exclusion to inclusion ratios in the absence the MS2-DsRed fusion are undetectable, while each MS2-DsRed input and in combination show a significant increase in exon exclusion validating the platforms regulatory function (Figure 3.3c).

We adapted our multi-input platform to process both the heterologous MS2-DsRed and the endogenous NF-κB p50 inputs, by replacing the MS2 aptamer in position 3 with the aptamer towards NF-κB p50 (Figure 3.3d). Likewise, mutants of both aptamers were inserted into the framework. These devices were stably integrated into HEK-293 and the output response of the devices was examined in absence of both inputs (p50 and MS2) and in the presence of either input or both (Figure 3.3e). We find that the output response is low in the absence of both ligands. The device response at both positions is equivalent to the regulatory activity observed with our single input devices demonstrating that platform function is programmable (Figures. 3.1c and 3.2d), where the platform response to p50 is that decreased gene expression, where as the response to MS2-DsRed is of increased gene expression ($P < 0.05$). Interestingly, the device response in the presence of both ligands is greater than the sum of the individual components, suggesting the combined inputs have a synergistic or combinatorial effect on splicing regulation. The transcript isoform ratios for the devices correlate well with the observed gene expression responses and therefore validate their regulatory function (Figure 3.3f).

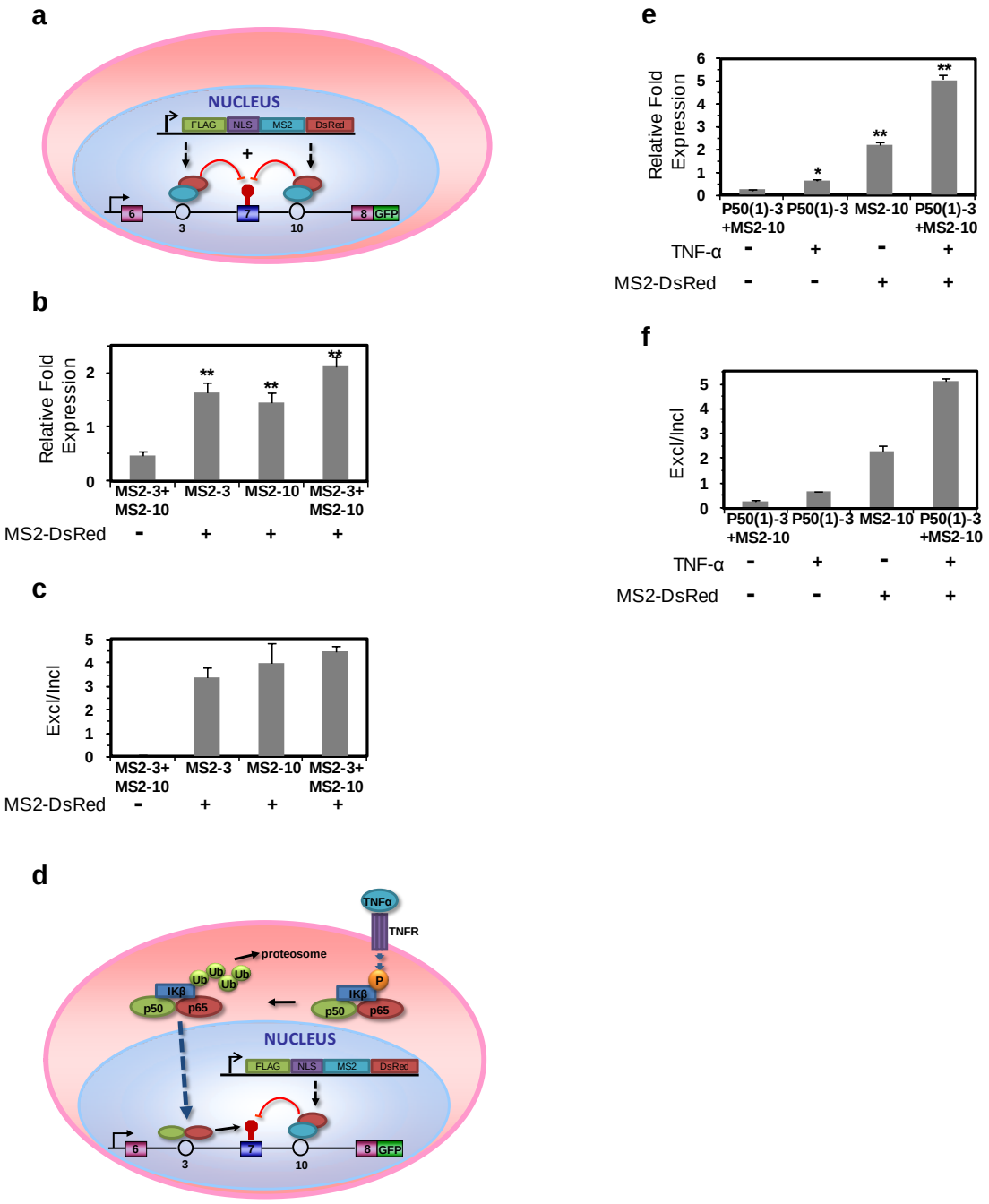


Figure 3.3. Multi-input processing platform and analysis. **(a)** MS2 multi-input splicing regulatory platform. The MS2 sensor components were inserted into positions 3 and 10 of the framework. **(b)** Flow cytometry analysis of HEK-293 FLP-In stable cell lines containing the MS2 multi-input regulatory devices. Regulatory activity was assessed in

the presence of the MS2-DsRed fusion and absence of the fusion (DsRed only) compared to mutant aptamer containing control constructs. **(c)** qRT-PCR analysis of HEK-293 FLP-In stable cell lines containing the MS2 multi-input regulatory devices. **(d)** MS2 and NF-κB p50 multi-input splicing regulatory platform. The NF-κB p50 and MS2 sensor components were inserted into positions 3 and 10 of the framework respectively. Cells were stimulated with 20ng/ml TNF-α to induce the NF-κB pathway and the MS2-DsRed fusion construct was transiently expressed. **(e)** Flow cytometry analysis of HEK-293 FLP-In stable cell lines containing the MS2 and NF-κB p50 multi-input regulatory devices. Regulatory activity was assessed in the absence of the both inputs (MS2-DsRed fusion and TNF- α) and in the presence of either input or both. **(f)** qRT-PCR analysis of HEK-293 FLP-In stable cell lines containing the MS2 and NF-κB p50 multi-input regulatory devices.

To investigate whether our platform can be used to regulate biological processes, we replaced the GFP reporter in our MS2-3 mutant and wildtype devices with *Puma*, a proapoptotic gene (Figure 3.4a). It has been demonstrated that the overexpression of Puma induces rapid apoptosis in mammalian cells through a Bax- and mitochondria-dependent pathway[16]. Apoptosis was assessed by flow cytometry 48h after transfection of the MS2-DsRed fusion. We find that apoptosis was significantly induced in cells containing the wildtype MS2-3 device as compared to cells containing the mutant device MS2Δ-3 which displayed an insignificant change in apoptosis (Figure 3.4b). To examine the ability to rationally reprogram signal transduction pathways to regulate apoptosis, we replaced the MS2 aptamer with the wildtype and mutant aptamers towards β-catenin (position 6) and NF-κB p65 (position 3) (Figure 3.4c). These aptamers were chosen

because they demonstrated increased exon exclusion activities which should upregulate the expression of Puma. LTD_4 or TNF-α was added to HEK-293 stable cell lines containing these devices to induce either the β-catenin or NF-κB pathway. Apoptosis was significantly induced in cells containing the B-cat-6 ($P < 0.01$) and p65-3 devices ($P < 0.05$), while cells containing the mutant devices (B-catΔ-6 and p65Δ-3) did not display a change in apoptosis (Figure 3.4d). The device mediated regulation of Puma and resulting apoptosis is in line with percentages observed in Puma overexpression studies[17]. These results show that both heterologous as well as endogenous inputs can be processed by our platform to regulate apoptosis.

To demonstrate the scalability of our platform in regulating apoptosis, we replaced Puma with the suicide gene therapy system, herpes simplex thymidine kinase (HSV-TK) in devices regulated by β-catenin and NF-κB p65 (Figure 3.4e). The HSV-TK system confers sensitivity to the prodrug ganciclovir (GCV) and has been shown to induce apoptosis in target cells[18]. We examined HSV-TK/GCV-induced apoptosis in HEK-293 cell lines stably expressing these devices in the presence of varying amounts (10 or 100 μM) and in the absence of GCV. To induce either the β-catenin or NF-κB pathway, LTD_4 or TNF-α was added to these cell lines. The B-cat-6 wildtype device in the presence of LTD_4 demonstrates significant sensitivity to GCV with an average cell survival of ~21% at 100 μM GCV ($P < 0.01$) (Figure 3.4f). In contrast, B-catΔ-6 in the presence and B-cat-6 in the absence of LTD_4, displayed similar survival rates of ~60%. The observed reduced cell survival may due to having sufficient expression of the basal spliced HSV-TK isoform to induce apoptosis upon addition of GCV as well as high dosages of GCV alone have been shown to induce cell death in the ranges of 5–30% in

cell culture [19]. In the presence of TNF-α, the wildtype p65-3 device demonstrates significant sensitivity to GCV with an average survival of ~23% at 100 μM GCV ($P < 0.01$), while the p65Δ-3 device shows little sensitivity with average survival of ~93% at 100 μM GCV. Similar to the B-cat-6 devices, the wildtype p65-3 device in the absence of TNF-α shows some sensitivity to GCV with an average survival of ~70% at 100 μM GCV. The observed sensitivity to GCV as seen with both devices in the presence of inducer is in line with studies overexpressing HSV-TK [19]. These results demonstrate that rewired endogenous signaling pathways can regulate apoptosis using either a suicide gene therapy system or overexpression of a proapoptotic gene.

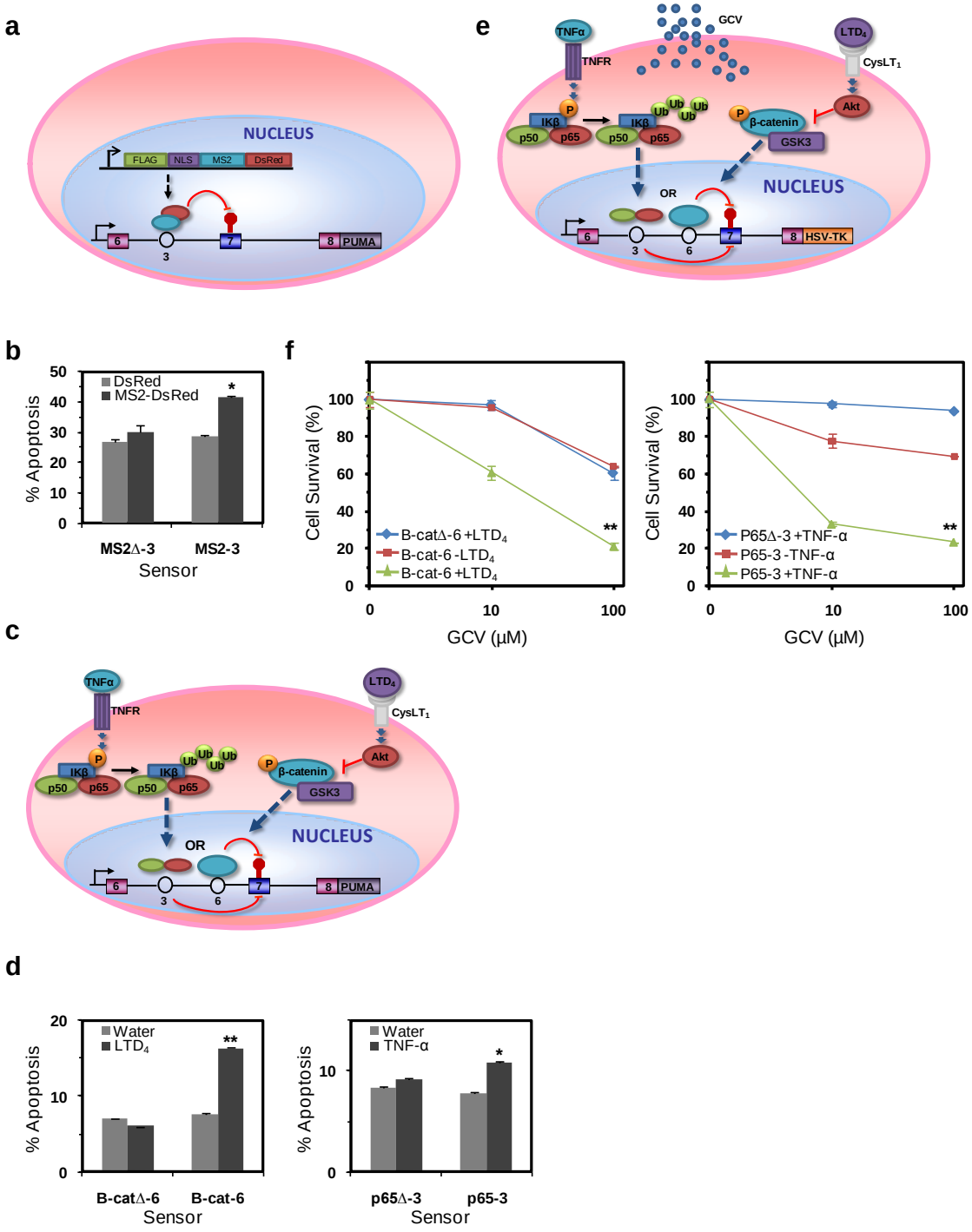


Figure 3.4. Platform mediated regulation of apoptosis. **(a)** MS2 splicing regulatory platform to induce apoptosis. The MS2 sensor component was inserted into position 3 of a framework consisting of Puma fused 3' of the SMN1 mini-gene. **(b)** Flow cytometry

analysis of apoptotic HEK-293 FLP-In stable cell containing the MS2 regulatory devices that mediate Puma induced apoptosis. Regulatory activity was assessed in the presence of the MS2-DsRed fusion and absence of the fusion (DsRed only) compared to mutant aptamer containing control constructs. **(c)** NF-κB p65 and β-catenin regulatory platforms to induce apoptosis. The NF-κB p65 and β-catenin sensor components were inserted into positions 3 and 6 of the framework respectively. The platform consists of Puma fused 3' of the SMN1 mini-gene. Cells were stimulated with either 20ng/ml TNF-α or 80 nM LTD_4 to induce the NF-κB or β-catenin pathways respectively. **(d)** Flow cytometry analysis of apoptotic HEK-293 FLP-In stable cell containing the NF-κB p65 and β-catenin regulatory devices that mediate Puma induced apoptosis. Regulatory activity was assessed in the presence of either inputs (TNF-α or LTD_4) and absence of both compared to mutant aptamer containing control constructs. **(e)** NF-κB p65 and β-catenin regulatory platforms to induce apoptosis using the HSV-TK suicide gene system. The NF-κB p65 and β-catenin sensor components were inserted into positions 3 and 6 of the framework respectively. The platform consists of HSV-TK fused 3' of the SMN1 mini-gene. Ganciclovir (GCV) was added to the cells to induce cell death. **(f)** Flow cytometry analysis of apoptotic HEK-293 FLP-In stable cell containing the NF-κB p65 and β-catenin regulatory devices that mediate HSV-TK induced cell death. Regulatory activity was assessed in the presence of either inputs (TNF-α or LTD_4) and absence of both compared to mutant aptamer containing control constructs.

3.3. Discussion

We have developed a novel framework for construction of both single and multiple input-single output RNA regulatory devices. We demonstrated that component modularity enabled the sensing of both heterologous as well as endogenous proteins and can be used to regulate the expression of any gene. Our platform can be interfaced with natural gene regulatory networks and signaling pathways to create programmable cells or cellular biosensors. The extension of our framework towards the processing of multiple inputs can be used in the creation of RNA-based platforms with sophisticated functionalities and to create complex gene regulatory networks to interrogate and program cellular function. Our platform can utilize both heterologous as well as reprogrammed endogenous signaling pathways to regulate apoptosis demonstrating the integration of cellular function to create complex user defined phenotypes. The platforms described here represent powerful tools to regulate gene expression and have broad applications in health and medicine where they can be used as 'intelligent' therapeutics towards the treatment and diagnosis of disease.

3.4. Materials and Methods

3.4.1. Base RNA device constructs

Plasmids were constructed using standard molecular biology techniques[20]. All enzymes, including restriction enzymes and ligases, were obtained through New England Biolabs unless otherwise noted. DNA synthesis was performed by Integrated DNA Technologies, Inc and DNA 2.0. Ligation products were electroporated into *E. coli* DH10B (Invitrogen) using a GenePulser XP system (BioRAD), and clones verified

through colony PCR and restriction mapping. All cloned constructs were sequence verified through Laragen. Primer sequences and plasmid descriptions are available in Tables S3.1 and S3.2, respectively.

The SMN1-GFP mini-gene fusion construct (pCS3001) was constructed through PCR amplification, digestion and ligation in the appropriate expression vector. A region encompassing the last nine nucleotides of exons 6 through the first 21 nucleotides of exon 8 of the *SMN1* mini-gene was amplified through PCR from template pCS3030 with primers Ex6 and Ex8 and PfuUltra II fusion high-fidelity DNA polymerase (Stratagene) and the resulting PCR product was digested with Nhe I. The *SMN1* mini-gene DNA synthesis was performed by DNA 2.0 to contain restriction sites Kpn I, Eco RV, Cla I (positions -87, -61 and -50 from 3'ss of exon7, respectively) and Xho I, Hind III, Bam HI, and Xba I (positions +10, +50, +70 and +97 from 5'ss of exon 7, respectively). The *GFP* gene was amplified from the template pKW430[21] with primers GFP1 and GFP2. The resulting PCR product was digested with Apa I and Nhe I and ligated into the corresponding restriction sites of the mammalian expression vector pcDNA5/FRT (Invitrogen). The SMN1-minigene PCR product was then ligated into the Nhe I restriction site of the resulting construct, creating the vector SMN1-GFP. The aptamer sequences used in this study are CGTACACCATCAGGGTACG[22], for MS2; CGTACCCATCAGGGTACG,[22] for MS2Δ; AGGCCGATCTATGGACGCTA AGGCACACCGGATACTTTAACGATTGGCT[23], for β-catenin; TCGGTTAGC AATTTCATAGGCCACACGGATATCGCAGGTATCTAGCCGGA (reverse compliment), for β-cateninΔ; GCATCCTGAAACTGTTTTAAGGTTGGCCGATGC[14], for NF-κB p50(1); CGTAGCCGGTTGGAAT TTTGTCAAAGTCCTACG (reverse

compliment), for NF-κB p50(1)Δ; GATCTTGAAACTGTTTTAAGGTTGGCCGATC[13], for NF-κB p50(2); GAAGCTTACAAGAAGGACAGCACGAATAAAACC TGCGTAAATCCGCCCCATTTGTGTAAGGGTAGTGGGTCGAATTCCGCTCA[12], for NF-κB p65; and ACTCGCCTTAAGCTGGGTGATGGGAATGTGT TTACCCCGCCTAAATGCGTCCAAAATAAGCACGACAGGAAGAACATTCGAAG (reverse compliment), for NF-κB p65Δ.

Specific aptamer and mutant cassette sequences containing portions of the SMN1-minigene were digested and ligated into the appropriate restriction sites within the SMN1-GFP construct (Table S3.3). Briefly, cassettes containing the wildtype and mutant MS2 coat protein aptamers were annealed, digested with the appropriate restriction enzymes and ligated into SMN1-GFP. β-catenin aptamer constructs were generated through PCR using templates β-cat-3, β-catΔ-3, β-cat-6, and β-catΔ-6 with forward primers Bcat3, BcatΔ3, Bcat6, and BcatΔ6 respectively with reverse primer AptRv. Similarly, for the NF-κB p50 and p65 aptamer constructs, the templates; NF-κBp50(1)-3, NF-κBp50(1)Δ-3, NF-κBp50(2)-3, NF-κBp65-3, and NF-κBp65Δ-3 were PCR amplified with forward primers p50(1), p50(1)Δ, p50(2), and p65, respectively, with reverse primer AptRv. The resulting β-catenin, NF-κB p50 and p65 PCR products were digested with the appropriate restriction enzymes and ligated into the corresponding restriction sites of SMN1-GFP. To construct RNA devices containing the wildtype and mutant MS2 coat protein aptamers in positions 3 and 10, the MS2-3 and MS2Δ-3 annealed cassettes were digested with Eco RV and Xho I and ligated into the corresponding restriction sites of SMN1-GFP containing the wildtype and mutant MS2 coat protein aptamers in position 10 (pCS3018 and pCS3019, respectively). To construct RNA devices containing the

wildtype and mutant NF-κB p65 aptamers in position 3 and the wildtype and mutant MS2 coat protein aptamers, the NF-κB p65 PCR products from above were digested with Eco RV and Xho I and were ligated into the corresponding restriction sites of SMN1-GFP containing the wildtype and mutant MS2 coat protein aptamers in position 10.

The SMN1-Puma and SMN1-TK fusions were constructed in two steps. The human *Puma* gene was amplified from template pORF5-hPUMA (Invivogen) with primers Puma1 and Puma2 to contain a flexible Gly-Ser linker (GGSGGS) at the 5' end. The resulting PCR product was digested with Nhe I and Pme I and ligated into the corresponding restriction sites of pcDNA5/FRT. The SMN1-GFP constructs containing the wildtype and mutant MS2 coat protein, β-catenin and NF-κB p65 aptamers in positions 3 and 6 of *SMN1* intron 6 (Figure 3.1a) were digested with Nhe I and the resulting SMN1-minigenes were ligated into the corresponding restriction site in the above vector containing the *Puma* gene. The resulting constructs are MS2-3-Puma, MS2Δ-3-Puma, β-cat-6-Puma, β-catΔ-6-Puma, p65-3-Puma and p65Δ-3-Puma (pCS3004-pCS3009, Table S3.2). Similarly, to create the SMN1-TK constructs the *HSV-TK* gene was amplified from the template CD19t-Tk-T2A-IL15op_epHIV7 using primers TK1 and TK2. The resulting PCR product was digested with Nhe I and Pme I and ligated into corresponding restriction sites of pcDNA5/FRT. The Nhe I digested SMN1-minigenes containing the wildtype and mutant β-catenin and NF-κB p65 aptamers in positions 3 and 6 of *SMN1* intron 6 were ligated into the corresponding restriction site of the above vector containing the *HSV-TK* gene. This created constructs β-cat-6-TK, β-catΔ-6-TK, p65-3-TK, and p65Δ-3-TK (pCS3010-–pCS3013, Table S3.2).

To create the MS2-DsRed control construct (pCS3014), the *DsRed* monomer gene was amplified from template pDsRed-monomer (Clontech) with primers DsRed1 and DsRed2, digested with Bam HI and Not I and ligated into the corresponding restriction sites of pcDNA5/FRT. The *MS2 coat protein* gene was amplified in two steps from template pHis-BIVT-MS2-RSp55[24] using primers MS2-1 and MS2-2 for the first round and MS2-3 and MS2-2 for the second to include a FLAG epitope (DYKDDDDK)[25] followed by an SV40 NLS (PKKKRKV)[26], 5' of the *MS2* gene. The final PCR product was digested with Kpn I and Bam HI and ligated into the corresponding restriction sites of the above vector containing DsRed monomer. The NF-κBp50-DsRed expression construct (pCS3015) was assembled similarly where the *NF-κB p50* gene was amplified in two steps from template pGAD424[27] using primers p50-1 and p50-2 for the first round and p50-3 and p50-2 to include a FLAG epitope followed by a SV40 NLS, 5' of the *NF-κB p50* gene. The final PCR product was digested with Nhe I and Kpn I and ligated into the corresponding restriction sites of the above vector containing DsRed monomer.

3.4.2. Cell culture, transfections, stable cell lines and flow cytometry

HEK-293 FLP-In cells (Invitrogen) were cultured in D-MEM supplemented with 10% fetal bovine serum (FBS) and 100 μg/ml Zeocin at 37ºC in 5% CO_2. Transfections were carried out with Fugene (Roche) according to the manufacturer's instructions. All cell culture media was obtained from Invitrogen.

HEK-293 FLP-In stable cell lines were generated by co-transfection of the appropriate SMN1 mini-gene construct with a plasmid encoding the Flp recombinase (pOG44) in growth medium without Zeocin according to the manufacturer's instructions (Invitrogen). Stable selections were carried out in 6-well plates seeded with ~$2x10^5$ HEK-293 FLP-In cells per well where 1.8 µg of pOG44 and .2 µg of the SMN1 mini-gene construct (10:1 ratio) were co-transfected. Fresh medium was added to the cells 24 h after transfection. The cells were expanded by a 1:4 dilution and Hygromycin B was added to a final concentration of 200 µg/ml 48 h after transfection. Clones were harvested by trypsinization, pooled and analyzed using a Quanta Cell Lab Cytometer (Beckman Coulter; Fullerton, CA) 10-14 days after transfection. GFP and DsRed fluorescence was excited at 488 nm and emission was measured through a 525-nm filter and a 610-nm band-pass filter respectively. For the NF-κB induction studies, cells were treated with 20ng/mL TNF-α (Sigma) for 48 h where indicated[28]. For the β-catenin induction studies, cells were serum starved 2h before stimulation with 80nM leukotriene D_4 (LTD_4) (Sigma) for 48 h in the absence of serum[29].

For transient transfection studies, HEK-293 stable cell lines containing SMN1 mini-genes were seeded in 24-well plates at ~$5x10^4$ cells per well 16 to 24 h prior to transfection. Cell lines were transfected with 250 ng of the appropriate MS2-DsRed or NF-κBp50-DsRed expression constructs. The cells were harvested by trypsinization, pooled and analyzed by flow cytometry 48 h after transfection. Experiments were carried out on different days and transfections were completed in duplicate, where the mean GFP fluorescence of the DsRed transfected population and the average error between samples is reported. For the induction of β-catenin and NF-κB pathways, HEK-293 stable cell

lines containing SMN1 mini-genes were seeded in a 24-well plate at ~$5x10^4$ cells per well 16 to 24 h prior to induction with LTD_4 or TNF-α. Stimulation of both pathways was carried out for 48h and the cells were harvested by trypsinization, pooled, and analyzed by flow cytometry. For the ganciclovir (GCV) sensitivity assays, HEK-293 stables cell lines containing SMN1 mini-genes containing either the β-catenin and NF-κB p65 aptamers were seeded in a 24-well plate at ~$5x10^4$ cells per well 16 to 24 h prior to induction with LTD_4 or TNF-α. At the time of induction cells were either left untreated or incubated with increasing concentrations of GCV (10 or 100μM). After 96h the cells were harvested by trypsinization, pooled, and analyzed by flow cytometry.

3.4.3. Apoptosis assays

Stable cell lines were harvested by trypsinization as described above. Pooled cells were washed in cold phosphate-buffered saline (PBS). Cells were stained with Pacific Blue annexin V and 7-aminoactinomycin D (7-AAD) using the Vybrant Apoptosis Assay Kit (Invitrogen) according to the manufacturer's instructions. The fluorescence of the stained cells was measured using a Quanta Cell Lab Cytometer where the Pacific Blue dye was excited using a UV light source and measured through a 465/430 band-pass filter (FL1). GFP and 7-AAD were excited with a 488-nm laser and measured through a 535-nm band-pass (FL2) and 670-nm long pass filter (FL3) respectively.

3.4.4. qRT-PCR analysis

Total cellular RNA was purified from stably transfected HEK-293 Flp-In cells using GenElute mammalian total RNA purification kit (Sigma) according to the manufacturer's instructions, followed by DNase treatment (Invitrogen). cDNA was synthesized using a gene-specific primer for the pcDNA5/FRT vector (SMN1cDNA) and Superscript III reverse transcriptase (Invitrogen) according to the manufacturer's instructions. qRT-PCR analysis was performed using isoform-specific primers (Table S3.4) where each reaction contained 1 µL template cDNA, 10 pmol of each primer and 1X iQ SYBR green supermix (BioRAD) to a final volume of 25 µL. Reactions were carried out using a iCycler iQ system (BioRAD) for 30 cycles (95ºC for 15 s, 72ºC for 30 s). The purity of the PCR products was determined by melt curve analysis. Data analysis was completed using the iCycler IQ system software v.3.1.7050 (BioRAD). Isoform-specific relative expression was calculated using the ΔCt (change in cycling threshold) method4. Expression levels of duplicate PCR samples were normalized to the levels of *HPRT* (Hypoxanthine-guanine phosphoribosyltransferase**)**. Fold expression data is reported as the mean expression for each sample divided by either the mean untreated expression value or the expression of the mutant aptamer cell line $\pm$ the average error.

3.4.5. Statistical analysis

Data are expressed as normalized or fold expression $\pm$ average error where applicable. Student's *t*-test and Anova analyses were performed using Microsoft Excel. $P < .05$ were taken to be significant.

Acknowledgments

We thank K. Hertel for providing the pHis-BIVT-MS2-RSp55 construct, M. Jensen for providing CD19t-Tk-T2A-IL15op_epHIV7 and L. J Maher for providing pGAD24.

References

1. Wang, E.T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476 (2008).

2. Black, D.L. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72,** 291–336 (2003).

3. Wang, G.S. & Cooper, T.A. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8,** 749–761 (2007).

4. Blencowe, B.J. Alternative splicing: new insights from global analyses. *Cell* **126,** 37–47 (2006).

5. Suess, B. & Weigand, J.E. Engineered riboswitches: overview, problems and trends. *RNA biology* **5,** 24–29 (2008).

6. Vaish, N.K. *et al.* Monitoring post-translational modification of proteins with allosteric ribozymes. *Nat Biotechnol* **20,** 810–815 (2002).

7. Ray, P.S. *et al.* A stress-responsive RNA switch regulates VEGFA expression. *Nature* **457,** 915–919 (2009).

8. Win, M.N., Liang, J.C. & Smolke, C.D. Frameworks for programming biological function through RNA parts and devices. *Chem Biol* **16,** 298–310 (2009).

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

10. Schneider, D., Tuerk, C. & Gold, L. Selection of high affinity RNA ligands to the bacteriophage R17 coat protein. *J Mol Biol* **228,** 862–869 (1992).

11. Lee, H.K. *et al.* beta-catenin regulates multiple steps of RNA metabolism as revealed by the RNA aptamer in colon cancer cells. *Cancer Res* **67,** 9315–9321 (2007).

12. Wurster, S.E. & Maher, L.J., 3rd Selection and characterization of anti-NF-kappaB p65 RNA aptamers. *Rna* **14,** 1037–1047 (2008).

13. Mi, J. *et al.* H1 RNA polymerase III promoter-driven expression of an RNA aptamer leads to high-level inhibition of intracellular protein activity. *Nucleic Acids Res* **34,** 3577–3584 (2006).

14. Chan, R. *et al.* Co-expression of anti-NFkappaB RNA aptamers and siRNAs leads to maximal suppression of NFkappaB activity in mammalian cells. *Nucleic Acids Res* **34,** e36 (2006).

15. Hoffmann, A., Levchenko, A., Scott, M.L. & Baltimore, D. The IkappaB-NF-kappaB signaling module: temporal control and selective gene activation. *Science* **298,** 1241–1245 (2002).

16. Buskirk, A.R. & Liu, D.R. Creating small-molecule-dependent switches to modulate biological functions. *Chem Biol* **12,** 151–161 (2005).

17. Christofk, H.R. *et al.* The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature* **452,** 230–233 (2008).

18. Beltinger, C. *et al.* Herpes simplex virus thymidine kinase/ganciclovir-induced apoptosis involves ligand-independent death receptor aggregation and activation of caspases. *Proc Natl Acad Sci U S A* **96,** 8699–8704 (1999).

19. Song, Y., Kong, B., Ma, D., Qu, X. & Jiang, S. Procaspase-3 enhances the in vitro effect of cytosine deaminase-thymidine kinase disuicide gene therapy on human ovarian cancer. *Int J Gynecol Cancer* **16,** 156–164 (2006).

20. Sambrook, J. & Russell, D.W., *Molecular Cloning: a laboratory manual*, 3 ed. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2001).

21. Stade, K., Ford, C.S., Guthrie, C. & Weis, K. Exportin 1 (Crm1p) is an essential nuclear export factor. *Cell* **90,** 1041–1050 (1997).

22. Villemaire, J., Dion, I., Elela, S.A. & Chabot, B. Reprogramming alternative pre-messenger RNA splicing through the use of protein-binding antisense oligonucleotides. *J Biol Chem* **278,** 50031–50039 (2003).

23. Lee, H.K., Choi, Y.S., Park, Y.A. & Jeong, S. Modulation of oncogenic transcription and alternative splicing by beta-catenin and an RNA aptamer in colon cancer cells. *Cancer Res* **66,** 10560–10566 (2006).

24. Graveley, B.R., Hertel, K.J. & Maniatis, T. A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. *Embo J* **17,** 6747–6756 (1998).

25. Chubet, R.G. & Brizzard, B.L. Vectors for expression and secretion of FLAG epitope-tagged proteins in mammalian cells. *Biotechniques* **20,** 136–141 (1996).

26. Kalderon, D., Roberts, B.L., Richardson, W.D. & Smith, A.E. A short amino acid sequence able to specify nuclear location. *Cell* **39,** 499-509 (1984).

27. Cassiday, L.A. & Maher, L.J., 3rd Yeast genetic selections to optimize RNA decoys for transcription factor NF-kappa B. *Proc Natl Acad Sci U S A* **100,** 3930–3935 (2003).

28. Hellweg, C.E., Arenz, A., Bogner, S., Schmitz, C. & Baumstark-Khan, C. Activation of nuclear factor kappa B by different agents: influence of culture conditions in a cell-based assay. *Ann N Y Acad Sci* **1091,** 191–204 (2006).

29. Mezhybovska, M., Wikstrom, K., Ohd, J.F. & Sjolander, A. The inflammatory mediator leukotriene D4 induces beta-catenin signaling and its association with antiapoptotic Bcl-2 in intestinal epithelial cells. *J Biol Chem* **281,** 6776–6784 (2006).

Supplementary Information

Table S3.1. Primer sequences

Name	**Primer sequence (5' to 3')**
Ex6	GCGCGCTAGCATGTATTATATGGTAAGTAAT CACTCAGC
Ex8	ATAGCTAGCGCTGCT ACCTGC CAGC
GFP1	GCGCGCTAGCGTG AGCAAGGGCGAG
GFP2	GCGCGGGCCCTTAGTACAGCTCGTCCATGCC
Puma1	ATAGTTTAAACGGTGGTTCT GGTGGTTCTGCCCGCGCACGCCAGGAG
Puma2	GCGC GTTTAAACTTA ATTGGGCTCCATCTCGGG
TK1	GCGCGCTAGCGTGACAGGGGGAATG GC
TK2	GCGCGTTTAAACTTAGTTAGCCTCCCCCATCTC
DsRed1	ATAGGATCCGACAACACCG AGGACGTCAT
DsRed2	ATAGCGGCCGCCTACTGGGAGCCGGAG
MS2-1	AGCCAAAAAAAAAACGCAAAGTGGCTTCTAACTTTACTCAGTTCGTTC
MS2-2	ATAGGATCCACCACCACCACCGTAGATGCCG
MS2-3	ATAGGTACCATGGATTACAAGGATGACGATGACAAGCCAAAAAAAA AACGCAAAGTG GCTTCTAACTTTAC
P50-1	AGCCAAAAAAAAAACGCAAAGTGGCAGAAGATGATCCATATTTGGGAAG
P50-2	ATAGGTACCGTCATCACTTTTGTCACAACCTTC
P50-3	ATAGCTAGCATGGATTACAAGGATGACGATGACAAGCCAAAAAAAA AACGCAAAGTGGCAGAAGATGATCC
SMN1cDNA	TAGAAGGCACAGTCGAGG
Bcat3	ATATGATACTAGCTATCAGGCCGA
Bcat3Δ	ATATGATATCTAGCTATCTCGGTTAG

Bcat6	ATA TATCGATGTCTATAT AGCTATTTTTTT TAA CTT
Bcat6Δ	ATATATCGATGTCTATAT AGCTATTTTTTT TAA CTT
AptRv	ATACTCGAGCAGACTTACTCCTTAATTTAAGGAATG
p50(1)	ATATGA TATCTA GCT ATCCGCGC
p50(1)Δ	ATATGA TATCTA GCT ATCCGTAGC C
p50(2)	ATATGA TATCTA GCT ATCCGCGC
p65	ATAT GATATCTAGCTATCGAAGC TACAAG AAGGACAGCAC

Table S3.2. Plasmid constructs used in this work

Name	Description
pCS3001	SM1-GFP. Contains the wild-type SMN1 mini-gene fused to the C-terminus of GFP.
pCS3002	SMN1-Puma. Contains the wild-type SMN1 mini-gene fused to the C-terminus of Puma.
pCS3003	SMN1-TK. Contains the wild-type SMN1 mini-gene fused to the C-terminus of HSV-TK.
pCS3004	MS2-3-Puma. Wildtype SMN1 mini-gene containing the MS2 aptamer in position 3 of intron 6 fused to the C-terminus of Puma.
pCS3005	MS2Δ-3-Puma. Wildtype SMN1 mini-gene containing the mutant MS2 aptamer in position 3 of intron 6 fused to the C-terminus of Puma.
pCS3006	β-cat-6-Puma. Wildtype SMN1 mini-gene containing the β-catenin aptamer in position 6 of intron 6 fused to the C-terminus of Puma.
pCS3007	β-catΔ-6-Puma. Wildtype SMN1 mini-gene containing the mutant β-catenin aptamer in position 6 of intron 6 fused to the C-terminus of Puma.
pCS3008	p65-3-Puma. Wildtype SMN1 mini-gene containing the NF-κB p65 aptamer in position 3 of intron 6 fused to the C-terminus of Puma.
pCS3009	p65Δ-3-Puma. Wildtype SMN1 mini-gene containing the mutant NF-κB p65 aptamer in position 3 of intron 6 fused to the C-terminus of Puma.
pCS3010	β-cat-6-TK. Wildtype SMN1 mini-gene containing the β-catenin aptamer in position 6 of intron 6 fused to the C-terminus of HSV-TK.
pCS3011	β-catΔ-6-TK. Wildtype SMN1 mini-gene containing the mutant β-catenin aptamer in position 6 of intron 6 fused to the C-terminus of HSV-TK.
pCS3012	p65-3-TK. Wildtype SMN1 mini-gene containing the NF-κB p65 aptamer in position 3 of intron 6 fused to the C-terminus of HSV-TK.
pCS3013	p65Δ-3-TK. Wildtype SMN1 mini-gene containing the mutant NF-κB p65 aptamer in position 3 of intron 6 fused to the C-terminus of HSV-TK.
pCS3014	MS2-DsRed. Contains the FLAG-NLS-MS2 gene fused to the C-terminus of DsRed.

pCS3015	NF-κBp50-DsRed. Contains the FLAG-NLS-NF-κBp50 gene fused to the C-terminus of DsRed.
pCS3016	MS2-10. Wildtype SMN1 mini-gene containing the MS2 aptamer in position 10 of intron 7 fused to the C-terminus of GFP.
pCS3017	MS2Δ-10. Wildtype SMN1 mini-gene containing the mutant MS2 aptamer in position 10 of intron 7 fused to the C-terminus of GFP.
pCS3018	MS2-3. Wildtype SMN1 mini-gene containing the MS2 aptamer in position 10 of intron 7 fused to the C-terminus of GFP.
pCS3019	MS2Δ-3. Wildtype SMN1 mini-gene containing the mutant MS2 aptamer in position 10 of intron 7 fused to the C-terminus of GFP.

Table S3.3. Aptamer cassette sequences used in the construction of the RNA devices. Aptamer sequences are italicized and nucleotides added for strengthening aptamer secondary structure are in Red.

Name	**Position**	**Restriction sites**	**Cassette (5'-3')**
MS2-1	1	Kpn I/Cla I	ATATGGTACCAACA*CGTACACCATCAGGGT ACG*TCCATATAAAGCTATAGATATCTAGCT ATCGATATAT
MS2Δ-1	1	Kpn I/Cla I	ATATGGTACCAACA*CGTACCCATCAGGGTA CG*TCCATAT AAAGCTATAGATATCTAGCTATCGAT ATAT
MS2-2	2	Kpn I/Cla I	ATATGGTACCAACATCCATATAAAGCTAT*C GTACACCA TCAGGGTACG*AGATATCTAGCTATCGATTA T
MS2Δ-2	2	Kpn I/Cla I	ATATGGTACCAACATCCATATAAAGCTAT*C GTACCCATCAGGGTACG*AGATATCTAGCTA TCGATTAT
MS2-3	3	Eco RV/Xho I	ATATGATATCTAGCTATC*CGTACACCATCAG GGTACG*GA TGTCTATATAGCTATTTTTTTTAACTTCCTT TATTTTCCT TACAGGGTTTCAGACAAAATCAAAAAGAA GGAAGGTGCTCACATTCCTTAAATTAAGG AGTAAGTCTGCTCGAGA TAT
MS2Δ-3	3	Eco RV/Xho I	ATATGATATCTAGCTATC*CGTACCCATCAG GGTACG*GATGTCTATATAGCTATTTTTTTTA ACTTCCTTTATTTTCCTTACAGGGTTTCAGA CAAAATCAAAAAGAAGGAAGGTGCTCACA TTCCTTAAATTAAGGAGTAAGTCTGCTCGA GATAT
MS2-4	4	Cla I/Xho I	ATATATCGATGTCTATATAGCT*CGTACACC ATCAGGTACG*ATTTTTTTTAACTTCCTTTAT

			TTTCCTTACAGGGTTTCAGACAAAATCAAA AAGAAGGAAGGTGCTCACATTCCTTAAAT TAAGGAGTAAGTCTGCTCGAGATAT
MS2Δ-4	4	Cla I/Xho I	ATATATCGATGTCTATATAGCT*CGTACCCAT CAGGGTACG*ATTTTTTTTAACTTCCTTTATT TTCCTTACAGGGTTTCAGACAAAATCAAAA AGAAGGAAGGTGCTCACATTCCTTAAATT AAGGAGTAAGTCTGCTCGAGATAT
MS2-5	5	Cla I/Xho I	ATATATCGATGTCTATATAGCTATTTTTTTT AACTTC*CGTACACCATCAGGGTACG*CTTTAT TTTCCTTACAGGGTTTCAGACAAAATCAAA AAGAAGGAAGGTGCTCACATTCCTTAAAT TAAGGAGTAAGTCTGCTCGAGATAT
MS2Δ-5	5	Cla I/Xho I	ATATATCGATGTCTATATAGCTATTTTTTTT AACTTC*CGTACCCATCAGGGTACG*CTTTATT TTCCTTACAGGGTTTCAGACAAAATCAAAA AGAAGGAAGGTGCTCACATTCCTTAAATT AAGGAGTAAGTCTGCTCGAG ATAT
MS2-6	6	Cla I/Xho I	ATATATCGATGTCTATATAGCTATTTTTTTT AACTTCCTTTATTTTCCTTAC*CGTACACCAT CAGGGTACG*AGGGTTTCAGACAAAATCAA AAAGAAGGAAGGTGCTCACATTCCTTAAA TTAAGGAGTAAGTCTGCTCGAGATAT
MS2Δ-6	6	Cla I/Xho I	ATATATCGATGTCTATATAGCTATTTTTTTT AACTTCCTTTATTTTCCTTAC*CGTACCCATC AGGGTACG*AGGGTTTCAGACAAAATCAAA AAGAAGGAAGGTGCTCACATTCCTTAAAT TAAGGAGTAAGTCTG CTCGAGATAT
MS2-7	7	Cla I/Xho I	ATATATCGATGTCTATATAGCTATTTTTTTT AACTTCCTTTATTTTCCTTACAGGGTTTCAG ACAAAATCAAAAAGAAGGAAGGTGCTCAC ATTCCTTAAATTAAGGAGTAAGTCTG*CGTA CACCATCAGGGTACG*CTCGAGATAT
MS2Δ-7	7	Cla I/Xho I	ATATATCGATGTCTATATAGCTATTTTTTTT AACTTCCTTTATTTTCCTTACAGGGTTTCAG ACAAAATCAAAAAGAAGGAAGGTGCTCAC

			ATTCCTTAAATTAAGGAGTAAGTCTG*CGTA CCCATCAGGGTACG*CTCGAGATAT
MS2-8	8	Xho I/HindIII	ATATCTCGAGCCAGCATTA*CGTACACCATC AGGGTACG*TGAAAGTGAATCTTACTTTTGT AAAAAAGCTTATAT
MS2Δ-8	8	Xho I/HindIII	ATATCTCGAGCCAGCATTA*CGTACCCATCA GGGTACG*TGAAAGTGAATCTTACTTTTGTA AAAAAGCTTATAT
MS2-9	9	Xho I/HindIII	ATATCTCGAGCCAGCATTATGAAAGTGAA TCTTA*CGTACACCATCAGGGTACG*CTTTTGT AAAAAAGCTTATAT
MS2Δ-9	9	Xho I/HindIII	ATATCTCGAGCCAGCATTATGAAAGTGAA TCTTA*CGTACCCATCAGGGTACG*CTTTTGTA AAAAAGCTTATAT
MS2-10	10	Xho I/Bam HI	ATATCTCGAGCCAGCATTATGAAAGTGAA TCTTACTTTTGTAAAAAAGC*CGTACACCATC AGGGTACG*TTCTTTATGGTTTGTGGGATCC ATAT
MS2Δ-10	10	Xho I/Bam HI	ATATCTCGAGCCAGCATTATGAAAGTGAA TCTTACTTTTGTAAAAAAGC*CGTACCCATCA GGGTACG*TTCTTTATGGTTTGTGGGATCCA TAT
MS2-11	11	HindIII/Bam HI	ATATAAGCTTCTTTATGGTTTGT*CGTACACC ATCAGGGTACG*GGGATCCATAT
MS2Δ-11	11	HindIII/Bam HI	ATATAAGCTTCTTTATGGTTTGT*CGTACCCA TCAGGGTACGG*GGGATCCATAT
MS2-12	12	Bam HI/Xba I	ATATGGATCCAAATGTTT*CGTACACCATCAG GGTACG*TTGAACAGTTAATCTAGAATAT
MS2Δ-12	12	Bam HI/Xba I	ATATGGATCCAAATGTTT*CGTACCCATCAG GGTACG*TTGAACAGTTAATCTAGAATAT
β-cat-3	3	Eco RV/Xho I	ATATGATATCTAGCTATC*AGGCCGATCTATG GACGCTATAGGCACACCGGATACTTTAACGAT TGGCT*GATGTCTATATAGCTATTTTTTTTAA

			CTTCCTTTATTTTCCTTACAGGGTTTCAGAC AAAATCAAAAAGAAGGAAGGTGCTCACAT TCCTTAAATTAAGGAGTAAGTCTGCTCGAG ATAT
β-catΔ-3	3	Eco RV/Xho I	ATATGATATCTAGCTATC*TCGGTTAGCAATT TCATAGGCCACACGGATATCGCAGGTATCTA GCCGGA*GATGTCTATATAGCTATTTTTTTT AACTTCCTTTATTTTCCTTACAGGGTTTCAG ACAAAATCAAAAAGAAGGAAGGTGCTCAC ATTCCTTAAATTAAGGAGTAAGTCTGCTCG AGATAT
β-cat-6	6	Cla I/Xho I	ATATATCGATGTCTATATAGCTATTTTTTTT AACTTCCTTTATTTTCCTTAC*AGGCCGATCT ATGGACGCTATAGGCACACCGGATACTTTAAC GATTGGCT*AGGGTTTCAGACAAAATCAAAA AGAAGGAAGGTGCTCACATTCCTTAAATT AAGG AGTAAGTCTG CTCGAG ATAT
β-catΔ-6	6	Cla I/Xho I	ATATATCGATGTCTATATAGCTATTTTTTTT AACTTCCTTTATTTTCCTTAC*TCGGTTAGCA ATTTCATAGGCCACACGGATATCGCAGGTATC TAGCCGGA*AGGGTTTCAGACAAAATCAAA AAGAAGGAAGGTGCTCACATTCCTTAAAT TAAGG AGTAAGTCTG CTCGAG ATAT
NF-κBp50(1)-3	3	Eco RV/Xho I	ATATGATATCTAGCTATC*GCATCCTGAAACT GTTTTAAGGTTGGCCGATGC*GATGTCTATAT AGCTATTTTTTTTAACTTCCTTTATTTTCCT TACAGGGTTTCAGACAAAATCAAAAAGAA GGAAGGTGCTCACATTCCTTAAATTAAGG AGTAAGTCTG CTCGAG ATAT
NF-κBp50(1)Δ-3	3	Eco RV/Xho I	ATATGATATCTAGCTATC*CGTAGCCGGTTG GAATTTTGTCAAAGTCCTACG*GATGTCTATA TAGCTATTTTTTTTAACTTCCTTTATTTTCC TTACAGGGTTTCAGACAAAATCAAAAAGA AGGAAGGTGCTCACATTCCTTAAATTAAG

			GAGTAAGTCTG CTCGAG ATAT
NF-κBp50(2)-3	3	Eco RV/Xho I	ATATGATATCTAGCTATCCGCGC*GATCTTG AAACTGTTTTAAGGTTGGCCGATC*GCGCGGA TGTCTATATAGCTATTTTTTTTAACTTCCTT TATTTTCCTTACAGGGTTTCAGACAAAATC AAAAAGAAGGAAGGTGCTCACATTCCTTA AATTAAGGAGTAAGTCTG CTCGAG ATAT
NF-κBp65-3	3	Eco RV/Xho I	*GAAGCTTACAAGAAGGACAGCACGAATAAAA CCTGCGTAAATCCGCCCCATTTGTGTAAGGG TAGTGGGTCGAATTCCGCTCA*GATGTCTATA TAGCTATTTTTTTTAACTTCCTTTATTTTCC TTACAGGGTTTCAGACAAAATCAAAAAGA AGGAAGGTGCTCACATTCCTTAAATTAAG G AGTAAG
NF-κBp65Δ-3	3	Eco RV/Xho I	*ACTCGCCTTAAGCTGGGTGATGGGAATGTGT TTACCCCGCCTAAATGCGTCCAAAATAAGCAC GACAGGAAGAACATTCGAAG*GATGTCTATAT AGCTATTTTTTTTAACTTCCTTTATTTTCCT TACAGGGTTTCAGACAAAATCAAAAAGAA GGAAGGTGCTCACATTCCTTAAATTAAGG AGTAAG

Table S3.4. Primer sequences for transcript isoform analysis through qRT-PCR

Name	**Forward Primer (5' - 3')**	**Reverse Primer (5' - 3')**	**Isoform**
Pair 1	GTATTATATGGAAATGCTGG	GAA GGTGGTCACGAGGG	Ex6/8 and GFP
Pair 2	TAAATTAAGGAGAAATGCT	GAA GGTGGTCACGAGGG	Ex7/8 and GFP
Pair 3	CAAAGATGGTCAAGGTCGCAAG	GGCGATGTCAATAGGACTCC	HPRT

Chapter IV. Conclusions and future prospects

4.1. Applications of SPLICE and further experimental characterizations

Elucidating the complex regulatory networks that control splicing represents a major challenge in the post-genomic era. Significant advances in our understanding of the mechanisms that guide splice site selection and the distribution of regulatory elements has aided the formulation of an early version of a 'splicing code'. Chapter 2 describes the development of a novel high-throughput cell-based screening strategy for ISRE function. We propose that SPLICE is a general screening platform that can be easily modified to select for a variety of *cis*-acting regulatory elements within any transcript. Although the initial 'proof-of-principle' of this screen was demonstrated with an alternative splicing pattern of exon skipping, the platform can be adapted towards other modes of alternative splicing including alternative 3' / 5' ss selection and mutually exclusive exon splicing. Using this platform, one can also screen for *cis*-acting regulators in different exonic and intronic locations.

One future application of this technology will be to select for libraries of cell-specific intronic or exonic regulatory elements that can be used to target gene expression in a tissue-specific and temporal manner. Toward this goal, our platform can be extended to develop a cell-specific screening strategy by incorporating a counter-selection step in a second cell line. This modified strategy can be utilized as well for the generation of cancer-specific and disease-specific regulators of alternative splicing. However, since we observed that the majority of ISREs generated through this screening strategy function in a second cell type, the diversity of the initial pool of oligonucleotides utilized for such

cell-specific screens may have to be significantly larger than what was examined in the study presented here. It is also likely that mutagenesis approaches will have to be applied at every round of these selections in combination with sequencing and experimental validation of selected elements to assess population enrichment and regulatory activity.

Given the role of alternative splicing in disease, the use of small molecules to target alternative splicing events has been investigated[1,2]. Robust and quantitative cell-based assays are needed to identify small molecule-based therapies targeted to specific splicing events[3]. Two screens have been recently developed, although on a modest scale, to identify small-molecule inhibitors of alternative splicing[4,5]. We have demonstrated that SPLICE is a robust cell-based screen and therefore can be used to screen for small molecule inhibitors of therapeutically-relevant spliced transcripts. The same strategy could likewise be applied to transcripts containing SPLICE selected ISREs to identify small molecules that affect their activities by interfering with the activity of corresponding trans-acting factors. Small molecules that regulate the activities of *cis*-acting regulatory elements will likely provide a means to modify alternative splicing of specific disease-relevant mRNAs.

Larger-scale experimental studies are needed to further characterize the SRNs involved in ISRE function as well as determining how large a role combinatorial control plays in the function of these elements. Several studies have applied high-throughput RNAi-based screening to probe gene expression in diseases such as cancer, where they have been used to gain insights into specific disease-associated pathways and their resulting phenotypes. A genome-wide RNAi-based screening strategy could be used to identify specific RNA-binding proteins that modulate the alternative splicing patterns of

transcripts containing selected ISRE elements. Results from these studies will aid in the elucidation of the trans-acting networks or SRNs involved in the regulation of these elements. In addition, high-throughput sequencing methods can be applied in combination with genome-wide RNAi-based screens to investigate transcriptome wide changes in alternative splicing. Bioinformatic approaches employed in our study can be applied to determine the spectrum of *cis*-elements that are conserved and likely involved in RNAi induced changes in global splicing patterns. Results from these studies will aid in the determination of a general set of rules that define the regulation and context dependence of intronic control elements.

Further experimental characterization is needed to identify how SPLICE selected ISRE sequences utilize complex SRNs to regulate splicing patterns in targeted cell types. Experimental analysis of selected ISREs demonstrated that their function is determined by combinatorial effects of multiple elements. To further characterize the combinatorial nature of these elements, SPLICE can be adapted to contain a synthetic oligonucleotide library comprising a combination of enriched ISRE elements spaced by neutral elements of similar length. This screen can be used to select for synthetic regulatory elements of enhanced function in a specific transcript or multiple transcripts towards the selection of universal or modular regulatory sequences. The experimental characterization of the SRNs involved in the regulation of these selected sequences as well as mutational studies that determine which elements play a role in regulation will provide significant insight into how combinatorial regulation is achieved.

Our studies also revealed that selected ISREs retain function when tested in a second cell type, but not in a second transcript. ISRE function was examined in a second

NMD-based reporter, based on the BRCA1 gene that contained an intron with significantly different predicted RNA secondary structure than the examined SMN1 intron. Given these differences, the context dependence of selected ISREs should be examined in several alternatively spliced mini-gene contexts. Also, the context dependence of these elements should be further explored within the introns of the endogenous genes containing these elements that were examined in our RNAi-based depletion study. Mini-genes containing the alternative spliced portions of these genes could be constructed to determine if the splicing of these transcripts depend on the presence of SPLICE generated ISREs. Overall, this work provides the first large-scale analysis of ISREs *in vivo* and our results highlight than an understanding of the complex interplay between multiple factors at a single binding site is necessary to further define the splicing code.

4.2. Applications of engineered ligand control of alternative splicing

Toward the goal of engineered regulation of alternative splicing we have created a platform to support the construction of protein-responsive alternative splicing regulatory elements based on the integration of protein-binding RNA aptamers into key intronic locations of a target alternatively spliced transcript. This protein-responsive platform was adapted to detect disease biomarkers, reprogram natural signaling pathways, and control biologically-relevant processes, such as apoptosis, in response to increased signaling through pathways associated with disease. *In vitro* splicing studies are needed to validate that the observed gene regulatory responses are directly modulating

alternative splicing patterns in response to changing levels of the input protein. Since our qRT-PCR measurements of the spliced isoforms correlate well with observed gene expression data for these RNA control systems, we anticipate that the results from *in vitro* splicing studies will be in line with the *in vivo* measurements.

This protein-responsive alternative splicing control platform can be interfaced with natural gene regulatory networks and signaling pathways to build programmable cells or cellular biosensors. As biosensors, these genetic elements can be used to interface with, investigate and perturb natural biological systems to report on the functional properties of regulatory networks. By replacing the aptamer component of the system with various protein-binding aptamers, this modular platform can be applied to early disease detection for a variety of different diseases. Therefore, as protein aptamers are selected towards desired targets, such as disease, cancer and viral biomarkers, they can be readily integrated into our ligand-regulated splicing platform. Furthermore, the ability to reprogram biological function in response to endogenous protein levels has broad applications in health and medicine, where such molecular tools can provide the basis for the design of targeted “intelligent” therapeutics.

References

1. Soret, J. *et al.* Selective modification of alternative splicing by indole derivatives that target serine-arginine-rich protein splicing factors. *Proc Natl Acad Sci U S A* 102, 8764–8769 (2005).

2. Fukuhara, T. *et al.* Utilization of host SR protein kinases and RNA-splicing machinery during viral replication. *Proc Natl Acad Sci U S A* 103, 11329–11333 (2006).

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

4. Stoilov, P., Lin, C.H., Damoiseaux, R., Nikolic, J. & Black, D.L. A high-throughput screening strategy identifies cardiotonic steroids as alternative splicing modulators. *Proc Natl Acad Sci U S A* 105, 11218–11223 (2008).

5. O'Brien, K., Matlin, A.J., Lowell, A.M. & Moore, M.J. The biflavonoid isoginkgetin is a general inhibitor of Pre-mRNA splicing. *J Biol Chem* 283, 33147–33154 (2008).