Designer Exons Inform a Biophysical Model for Exon Definition

Mauricio Arias

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2013

ABSTRACT

Designer Exons Inform a Biophysical Model for Exon Definition

Mauricio Arias

Pre-mRNA molecules in humans contain mostly short internal exons flanked by long introns. To explain the removal of such introns, recognition of the exons instead of recognition of the introns has been proposed. This thesis studies this exon definition mechanism using a bottom-up approach. To reduce the complexity of the system under study, this exon definition mechanism was addressed using designer exons made up of prototype sequence modules of our own design (including an exonic splicing enhancer or silencer). Studies were performed in vitro with a set of DEs obtained from random combinations of the exonic splicing enhancer and the exonic splicing silencer modules. The results showed considerable variability both in terms of the composition and size of the DEs and in terms of their inclusion level. To understand how different DEs generated different inclusion levels, the problem was divided into understanding separately parameters varied between DEs. Subsequent studies focused on each of three parameters: size, ESE composition and ESS composition. The final objective was to be able to combine their effects to predict the inclusion levels obtained for the "random" DEs mentioned previously. To complement this experimental approach an equation was generated in two stages. First a general "framework" equation was obtained modeling a necessary exon definition complex that enabled commitment to inclusion. This equation used rates for the formation and dissociation of this complex without elaborating on the details for how those rates came about. In the second stage, however, formation and dissociation were modeled using novel but intuitive ideas and these models were combined into a final equation. This equation using the single-parameter

perturbation data obtained previously performed well in predicting the inclusion levels of the "random" DEs. Additionally, both the final equation and the mechanisms proposed align well with results published by other groups. In order to make these results more accessible and to open more opportunities to extend them, an initial attempt is presented to identify the proteins involved in the functionality observed for each of the sequences used.

Table of Contents

Chapter 1 **Introduction** 1

References 7

Chapter 2 **Splicing of Designer Exons Reveals Unexpected Complexity in Pre-mRNA Splicing** 11

Abstract 11

Introduction 12

Results 15

Discussion 27

Materials and Methods 32

Authors' Contributions 34

Acknowledgements 34

References 34

Supplemental Material 38

Chapter 3 **A Reductionist Approach to Splicing of Designer Exons: Single-Parameter Perturbations** 45

Introduction 45

Results 47

Discussion 60

Materials and Methods 62

Authors’ Contributions 65

Acknowledgements 65

References 65

Supplemental Material 68

Chapter 4 Derivations of the Equations to Model Exon Definition and Splicing 98

Introduction 98

A Biophysical Model to Explain Splicing Decisions 101

Discussion 116

Authors’ Contributions 118

Acknowledgements 118

References 118

Chapter 5 Splicing of Designer Exons Informs a Biophysical Model for Exon Definition 121

Introduction 121

Results 124

Discussion 137

Materials and Methods 146

Authors’ Contributions 146

Acknowledgements 146

References 147

	Supplemental Material	151
Chapter 6	**Identifying Candidate Effectors**	159
	Introduction	159
	Results	161
	Discussion	173
	Materials and Methods	179
	Authors' Contributions	183
	Acknowledgements	184
	References	184
Chapter 7	**Conclusions**	190

List of Charts, Graphs, and Illustrations

Figures

Figure 2.1 Examples of ambiguity in the identification or testing of splicing regulatory motifs. 14

Figure 2.2 Construction of designer exons. 17

Figure 2.3 Determination of designer exon genotypes and phenotypes. 20

Figure 2.4 Splicing of designer exons. 21

Figure 2.5 Correlations between exon inclusion and splicing regulatory elements. 22

Figure 2.6 Splicing of exons designed without silencers. 25

Figure S2.1 RESCUE-ESEs and PESEs in sequences found by SR protein responsive functional SELEX. 44

Figure 3.1 Construction of designer exons (example). 48

Figure 3.2 Exon inclusion has an optimum size range. 50

Figure 3.3 Addition of a single ESE enhances inclusion level and is position independent. 53

Figure 3.4 Inclusion levels of DEs increase with the number of ESEs present. 56

Figure 3.5 Addition of a single ESS decreases inclusion level and shows some position dependence. 57

Figure 3.6 Inclusion levels of DEs decrease with the number of ESSs present. 59

Figure S3.1 Cartoon of a typical DE-containing minigene. 93

Figure S3.2 Exon inclusion has an optimum size range in a chromosomal context. 93

Figure S3.3 Addition of a single ESE enhances inclusion level and is position independent in a chromosomal context. 94

Figure S3.4 Addition of a single ESS decreases inclusion level and shows some position dependence in a chromosomal context. 95

Figure S3.5 Site-specific recombination reconstitutes the minigene in a specific location of the genome. 96

Figure S3.6 Coupled-standards incorporate two cDNAs into a single molecule. 97

Figure 4.1 Complex kinetics can be described in simpler terms. 102

Figure 4.2 Modeling exon end-to-end contact in exon definition. 112

Figure 5.1 The model accurately predicts the inclusion levels of DEs with single-parameter perturbations. 130

Figure 5.2 The model predicts the size and ESE dependences accurately. 131

Figure 5.3 The model accurately predicts the inclusion levels of DEs for each parameter examined. 132

Figure 5.4 The model accurately predicts the inclusion levels of complex DEs. 134

Figure 5.5 The combinations of ESEs and ESSs are accurately modeled in 110 nt exons. 136

Figure 5.6 The observed psi for complex DEs progressively falls short of prediction as sizes increase above 140 nt. 137

Figure S5.1 Increasing the number of ESEs yields distinctive curves for the stability and the recruitment models for ESE action. 155

Figure S5.2 The recruitment model fails to accurately predict the effect of combining ESEs and ESSs in 110 nt exons. 156

Figure S5.3 Comparison of time course experiments using constructs with multiple ESEs in (Hertel and Maniatis 1998) with time course predictions by the model described in the text. 157

Figure S5.4 The psi of DEs with multiple ESSs as predicted by the model. 158

Figure 6.1 ESS2 has stronger silencer activity than ESS and works at all positions tested. 162

Figure 6.2 Proteins pulled down by RNA baits for the reference sequence, ESE, ESS and ESS2. 165

Figure 6.3 Proteins are enriched differentially for the four bait sequences. 169

Figure 6.4 Western blots confirm the mass spectrometry results. 170

Figure 6.5 Knock down results make hnRNP K an unlikely candidate to explain the effect of ESE. 172

Tables

Table 2.1 Correlation coefficients between inclusion level and the proportion of module bases found in different types of predicted secondary structures. 26

Table 2.2 Designer exon pairs with the same composition but different splicing behavior. 29

Table S2.1 Inclusion level for the set of designer exons used in this chapter. 38

Table 3.1 Effect of splice site sequences on exon inclusion. 51

Table S3.1 Exon inclusion of DEs for size, ESE and ESS perturbation. 81

Table S3.2 Oligomers used. 86

Table 5.1 Best fit for parameters in equations 5.5 and 5.6. 129

Table 5.2 Evaluation of the model. 132

Table S5.1 Values for the parameters in the recruitment model. 154

Table 6.1 Proteins showing differential enrichment with the four RNA molecules. 168

Table 6.2 Short list of candidate proteins for providing the observed functionality of each sequence. 173

Boxes

Box 4.1 Model assumptions and definitions 103

Acknowledgments

To Him who gave me abundant life and in whom all things hold together.

I thank my mother for teaching me to persist when things matter and to my sister for bringing light to my life. I also thank my wife Shannon for her patience, support and advice, and for the family we are creating with my son Gabriel, who has changed everything. To my family for all the support they gave me and to the many professors who pushed me to go farther.

A special thanks to Larry Chasin for his patience, advice and encouragement and to Shengdong Ke, Jonathan Cacciatore, Shulian Shang, Ashira Lubkin, Vincent Anquetil and Dennis Weiss for helpful discussions and advice.

Chapter 1

Introduction

Since the discovery of split genes in the 1970s (Berget et al. 1977), there have been significant advances in our understanding of how these genes function in the cell. Transcription of these genes produces pre-mRNA molecules. These molecules are processed to generate shorter mRNA molecules, in which the intervening sequences have been removed. One crucial aspect in this process is how the cell accurately identifies which regions are to be removed, introns, and which are to be kept and spliced together, exons (Fox-Walsh and Hertel 2009). The importance of this recognition is highlighted by many diseases that occur when proper operation is hindered by mutations (Ward and Cooper 2010).

In the study of this recognition, two models have shaped how splicing is studied (Berget 1995; De Conti et al. 2013). The first model is known as the intron definition model. This model postulates that introns are the units of recognition and that splicing together the sequences that flank them generates mRNA and as a byproduct delineates the exons. The second model is known as the exon definition model. It postulates that exons are the units of recognition and that once two adjacent ones have been recognized they can be joined. Determining adjacency implies that the intron has to be defined and, therefore, exon definition should be followed by intron definition. It has been suggested that both models are valid but are active under different circumstances: intron definition is predominant when introns are small, while exon definition when exons are small and the flanking introns are long (Fox-Walsh et al. 2005).

Intron/exon boundaries play an important role independent of which part is recognized first and their sequence became established functional elements from early on (Mount 1982; Mount and Steitz 1983). However, the information encoded in such sequences seems insufficient to recognize proper exons from a plethora of pseudoexons (Sun and Chasin 2000). Another source of information was postulated to exist in the exons themselves (Reed and Maniatis 1986; Mardon et al. 1987; Cooper and Ordahl 1989; Tsai et al. 1989) in the form of exonic splicing enhancers (ESEs) and exonic splicing silencers (ESSs), which provide positive and negative effects on exon inclusion, respectively. Indeed the number of ESEs and ESSs proposed has made the exons rich and dense in information. So much so that changes to disrupt a target sequence for study are likely to create new functional sequences (Zhang et al. 2009).

Several tools have helped in the study splicing. One that has had great impact is *in vitro* splicing (Krainer et al. 1984). Shortly after its development, use of this tool led to the discovery of a family of proteins that play a fundamental role in splicing: the SR proteins (Ge and Manley 1990; Krainer et al. 1990). Many of the ESEs have been shown to exert their effects through SR proteins (Long and Caceres 2009). Moreover, *in vitro* splicing has allowed the characterization of the effects of another family of proteins, hnRNPs, which in many cases bind ESSs and curtail exon inclusion (Martinez-Contreras et al. 2007). However, *in vitro* splicing has limitations and attempts are being made to address them. One of them is the reliance on shortened versions of the genes due to the slower rate of the *in vitro* splicing reaction compared to *in vivo* experiments (Hicks et al. 2005). This limitation has negatively affected the study of exon definition because the abridged introns usually preclude this mechanism. Another limitation is the uncoupling of transcription and splicing. Evidence has been mounting that transcription rates affect splicing (de la Mata et al. 2003; de la Mata et al. 2011) but this phenomenon of co-transcriptional splicing is not easily studied in traditional *in vitro* experiments (Lazarev and Manley 2007).

The present work builds on several results in the literature, namely the existence of an early irreversible step that determines the inclusion of an exon and the observation of splicing occurring while the transcript is being synthesized, i.e., co-transcriptional splicing. Even though many of the steps in the splicing reactions seem reversible, the existence of early irreversible steps in the splicing process was demonstrated by Lim and Hertel (Lim and Hertel 2004). These researchers characterized an early irreversible step that pairs the splice sites across the intron, implies the inclusion of the exon, occurs after complex E formation and might precede or coincide with the ATP-dependent formation of complex A (Lim and Hertel 2004). This result implies that the splicing outcome can be determined early in the splicing process. As a matter of fact, it is often thought that the splicing outcome is regulated by altering the binding of the initial factors (Black 2003; House and Lynch 2006; Chen and Manley 2009). Subsequent studies have confirmed this but some have shown that other downstream steps in the splicing process can be affected (House and Lynch 2006; Sharma et al. 2008; Chen and Manley 2009). However, even in some of these cases changes in the early steps result in a modificiation in the splicing process that would not enable commitment, effectively poisoning the downstream reactions (House and Lynch 2006). This makes the early steps in splicing a particularly interesting topic to study the decisions involved in exon inclusion.

Another important aspect to consider is the observation that splicing in at least some genes occurs co-transcriptionally (Goldstrohm et al. 2001; Singh and Padgett 2009; Wada et al.

2009; Dujardin et al. 2013). Since splicing is not an event but rather a process involving the formation of several intermediate complexes, co-transcriptional splicing can occur even if the final product is generated after transcription has finalized. In this case, occurrence of the irreversible step before transcription is finalized implies that transcription kinetics could affect the splicing decision (Dujardin et al. 2013).

From studies of natural phenomena particularly in physics we have learned that even complex systems can be understood in terms of simple mechanisms. This simplicity can be hidden by the complex nature of the systems studied, in fact making the underlying mechanisms practically indiscernible. At least two approaches are available for the study of complex systems. A currently favored one attempts to understand the relationships between the parts of a system by studying it as a whole: top-down approach. This approach has been particularly used for the study of emergent properties which are postulated to be unforeseeable from even complete understanding of each individual constituent part (Cohen and Harel 2007). A more traditional approach attempts to understand each part separately, then understand the relationships as the parts are gradually put together and then focus on the system as a whole: a bottom-up approach. Both strategies have different strengths and weaknesses which allow them to be good complements. The studies presented here use a bottom-up approach because of its intrinsic power to make simple relationships apparent, which aligns with our focus on the underlying molecular mechanisms governing splicing.

To reduce the complexity of the system to study, designer exons (DEs) are introduced in Chapter 2 for the study of splicing. These simplified exons are composed of combinations of a small number of naturally occurring modules that include three types of sequences: a prototypical ESE, a prototypical ESS and a “neutral” sequence (also known as the reference sequence). A description of the complete system for the *in vivo* study of DE splicing is included, which satisfies the requirements for exon definition. A report is then given of how a set of DEs composed of random combinations of ESEs and ESSs was made and analyses of exon inclusion are presented. This initial approach sets the stage for more focused experiments but gave little insight into the role of the different parameters varied.

Chapter 3 describes a reductionist approach to the study of exon recognition focusing on three parameters: size, ESE composition and ESS composition. Studying these parameters separately allows an uncomplicated view of their effects on splicing. Experimental assessments of the effects of varying size on inclusion levels are presented, using DEs composed exclusively of reference sequences. The effect in DE inclusion level of a sole ESE is analyzed *in vivo* by placing the ESE in different positions along the exon. After this the effect of adding more copies of ESE to the DE is studied. A similar set of experiments is described for ESS.

Chapter 4 presents the mathematical derivation of a biophysical model to connect inclusion levels to the parameters varied in Chapter 3. This derivation is presented in two parts: a framework model that addresses how commitment progresses based on the states in which the molecules can be, while keeping the details of the molecular mechanisms for the transitions between states general, and a refinement of the model where hypothetical mechanisms to model these transitions are devised and implemented. Clearly delineated assumptions are introduced, which include the existence of an exon definition complex. The full mathematical treatment of a framework model is shown. A general solution is derived for modeling splicing in terms of the rates of formation and dissociation of the exon definition complex and the rate at which exons containing it commit to inclusion. In order to gain an intuitive grasp of the roles of the different components in the resulting complex equation, a simplification is presented based on assumptions regarding the magnitude of the rates involved. A simpler equation is obtained and a biological interpretation of its terms included. This basic equation is then used to explore the effect of different biophysical mechanisms on the inclusion level of DEs. A model is derived for two crucial parameters: size and ESE content. ESSs and reference sequences are modeled as having effects opposed to or similar to ESEs. To account for the effect of size, the collision of exon ends is modeled using results from statistical mechanics of polyelectrolytes. To account for the effect of ESEs, the stability of the exon definition complex is modeled analyzing the frequency of complex disruption by collisions with random molecules.

Chapter 5 expands the reductionist approach to the study of exon recognition introduced in Chapter 3. To explore mechanistic explanations for the observations previously discussed about size, ESE composition and ESS composition, the development of a biophysical model for exon definition of internal exons undergoing co-transcriptional splicing is presented. The mechanisms analyzed are restricted to those that are deemed to affect DEs and natural exons alike. This model features commitment to inclusion before the downstream exon is synthesized and competition between skipping and inclusion fates afterwards. Collision of both exon ends to form an exon definition complex is incorporated to account for the effect of size. Stabilization of the resulting complex is used to model ESE and ESS effects. The performance of this model is evaluated in terms of its ability to reproduce the single-parameter results as well as its ability to predict the inclusion level of the more complex designer exons presented in Chapter 2.

Chapter 6 presents initial attempts at identifying the proteins involved in splicing of DEs. We first introduce a sequence, ESS2, that represents an extension for future DEs and that is used in the process of identifying candidates for the other sequences. The silencer effects of this sequence are studied at different positions along the exon to provide background for the following experiments. For these four sequences, namely ESE, ESS, ESS2 and the reference sequence, a pull down experiment is described. Proteins that bind to each of the molecules are identified through label-free shotgun mass spectrometry. From these experiments, a list of 26 candidates is presented. The candidates are then briefly analyzed in terms of their putative roles in the cell as well as general characteristics and a short list of favored candidates is compiled.

Chapter 7 summarizes the main points presented in the previous chapters. The significance of designer exons is discussed in the context of identifying mechanisms that affect splicing of natural exons. Some conclusions are also presented regarding the significance of unexpected results observed in Chapter 3. Additionally, the usefulness of the equations developed in Chapter 4 is evaluated placing special emphasis on the mechanisms proposed to explain size and ESE effect. The value of identifying the proteins associated with the functionality of each of the sequences studied is discussed briefly. Finally, an exploration is conducted on directions for future research.

References

Berget SM. 1995. Exon recognition in vertebrate splicing. *J Biol Chem* **270**(6): 2411-2414.

Berget SM, Moore C, Sharp PA. 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America* **74**(8): 3171-3175.

Black DL. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72**: 291-336.

Chen M, Manley JL. 2009. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10**(11): 741-754.

Cohen IR, Harel D. 2007. Explaining a complex living system: dynamics, multi-scaling and emergence. *Journal of the Royal Society, Interface / the Royal Society* **4**(13): 175-182.

Cooper TA, Ordahl CP. 1989. Nucleotide substitutions within the cardiac troponin T alternative exon disrupt pre-mRNA alternative splicing. *Nucleic Acids Res* **17**(19): 7905-7921.

De Conti L, Baralle M, Buratti E. 2013. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA* **4**(1): 49-60.

de la Mata M, Alonso CR, Kadener S, Fededa JP, Blaustein M, Pelisch F, Cramer P, Bentley D, Kornblihtt AR. 2003. A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell* **12**(2): 525-532.

de la Mata M, Munoz MJ, Allo M, Fededa JP, Schor IE, Kornblihtt AR. 2011. RNA Polymerase II Elongation at the Crossroads of Transcription and Alternative Splicing. *Genet Res Int* **2011**: 309865.

Dujardin G, Lafaille C, Petrillo E, Buggiano V, Gomez Acuna LI, Fiszbein A, Godoy Herz MA, Nieto Moreno N, Munoz MJ, Allo M et al. 2013. Transcriptional elongation and alternative splicing. *Biochim Biophys Acta* **1829**(1): 134-140.

Fox-Walsh KL, Dou Y, Lam BJ, Hung SP, Baldi PF, Hertel KJ. 2005. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proceedings of the National Academy of Sciences of the United States of America* **102**(45): 16176-16181.

Fox-Walsh KL, Hertel KJ. 2009. Splice-site pairing is an intrinsically high fidelity process. *Proceedings of the National Academy of Sciences of the United States of America* **106**(6): 1766-1771.

Ge H, Manley JL. 1990. A protein factor, ASF, controls cell-specific alternative splicing of SV40 early pre-mRNA in vitro. *Cell* **62**(1): 25-34.

Goldstrohm AC, Greenleaf AL, Garcia-Blanco MA. 2001. Co-transcriptional splicing of pre-messenger RNAs: considerations for the mechanism of alternative splicing. *Gene* **277**(1-2): 31-47.

Hicks MJ, Lam BJ, Hertel KJ. 2005. Analyzing mechanisms of alternative pre-mRNA splicing using in vitro splicing assays. *Methods* **37**(4): 306-313.

House AE, Lynch KW. 2006. An exonic splicing silencer represses spliceosome assembly after ATP-dependent exon recognition. *Nat Struct Mol Biol* **13**(10): 937-944.

Krainer AR, Conway GC, Kozak D. 1990. Purification and characterization of pre-mRNA splicing factor SF2 from HeLa cells. *Genes Dev* **4**(7): 1158-1171.

Krainer AR, Maniatis T, Ruskin B, Green MR. 1984. Normal and mutant human beta-globin pre-mRNAs are faithfully and efficiently spliced in vitro. *Cell* **36**(4): 993-1005.

Lazarev D, Manley JL. 2007. Concurrent splicing and transcription are not sufficient to enhance splicing efficiency. *RNA* **13**(9): 1546-1557.

Lim SR, Hertel KJ. 2004. Commitment to splice site pairing coincides with A complex formation. *Mol Cell* **15**(3): 477-483.

Long JC, Caceres JF. 2009. The SR protein family of splicing factors: master regulators of gene expression. *Biochem J* **417**(1): 15-27.

Mardon HJ, Sebastio G, Baralle FE. 1987. A role for exon sequences in alternative splicing of the human fibronectin gene. *Nucleic Acids Res* **15**(19): 7725-7733.

Martinez-Contreras R, Cloutier P, Shkreta L, Fisette JF, Revil T, Chabot B. 2007. hnRNP proteins and splicing control. *Adv Exp Med Biol* **623**: 123-147.

Mount S, Steitz J. 1983. Lessons from mutant globins. *Nature* **303**(5916): 380-381.

Mount SM. 1982. A catalogue of splice junction sequences. *Nucleic Acids Res* **10**(2): 459-472.

Reed R, Maniatis T. 1986. A role for exon sequences and splice-site proximity in splice-site selection. *Cell* **46**(5): 681-690.

Sharma S, Kohlstaedt LA, Damianov A, Rio DC, Black DL. 2008. Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. *Nat Struct Mol Biol* **15**(2): 183-191.

Singh J, Padgett RA. 2009. Rates of in situ transcription and splicing in large human genes. *Nat Struct Mol Biol* **16**(11): 1128-1133.

Sun H, Chasin LA. 2000. Multiple splicing defects in an intronic false exon. *Mol Cell Biol* **20**(17): 6414-6425.

Tsai AY, Streuli M, Saito H. 1989. Integrity of the exon 6 sequence is essential for tissue-specific alternative splicing of human leukocyte common antigen pre-mRNA. *Mol Cell Biol* **9**(10): 4550-4555.

Wada Y, Ohta Y, Xu M, Tsutsumi S, Minami T, Inoue K, Komura D, Kitakami J, Oshida N, Papantonis A et al. 2009. A wave of nascent transcription on activated human genes. *Proceedings of the National Academy of Sciences of the United States of America* **106**(43): 18357-18361.

Ward AJ, Cooper TA. 2010. The pathobiology of splicing. *J Pathol* **220**(2): 152-163.

Zhang XH, Arias MA, Ke S, Chasin LA. 2009. Splicing of designer exons reveals unexpected complexity in pre-mRNA splicing. *RNA* **15**(3): 367-376.

Chapter 2

Published in RNA 15: 367-376, 2009 (Zhang et al. 2009)

Splicing of Designer Exons Reveals Unexpected Complexity in Pre-mRNA Splicing

Xiang H.-F. Zhang, Mauricio A. Arias, Shengdong Ke, and Lawrence A. Chasin

Department of Biological Sciences, Columbia University, New York, New York 10027, USA

Abstract

Pre-mRNA splicing requires the accurate recognition of splice sites by the cellular RNA processing machinery. In addition to sequences that comprise the branchpoint and the 3’ and 5’ splice sites, the cellular splicing machinery relies on additional information in the form of exonic and intronic splicing enhancer and silencer sequences. The high abundance of these motifs makes it difficult to investigate their effects using standard genetic perturbations, since their disruption often leads to the formation of yet new elements. To lessen this problem, we have designed synthetic exons comprised of multiple copies of a single prototypical exonic enhancer and a single prototypical exonic silencer sequence separated by neutral spacer sequences. The spacer sequences buffer the exon against the formation of new elements as the number and order of the original elements is varied. Over 100 such central designer exons were constructed by random ligation of enhancer, silencer and neutral elements. Each exon was positioned as the central exon in a 3-exon minigene and tested for exon inclusion after transient transfection. The level of inclusion of the test exons was seen to be dependent on the provision of enhancers and

could be decreased by the provision of silencers. In general, there was a good quantitative correlation between the proportion of enhancers and splicing. However, widely varying inclusion levels could be produced by different permutations of the enhancer and silencer elements, indicating that even in this simplified system splicing decisions rest on complex interplays of yet to be determined parameters.

Introduction

In higher organisms, pre-mRNA splicing represents an essential step in the transfer of information from DNA to protein, i.e., the central dogma. Much is known about the chemistry of intron removal catalyzed by the spliceosome, a multi-subunit ribonucleoprotein comparable in size and complexity to the ribosome. Less is known about the recognition of the splice sites, which is the key step in deciphering the information resident in the primary transcript. The splice site sequences themselves – a 9 nt stretch straddling the 5' splice site, a ~ 15 nt region at the 3' splice site (including the polypyrimidine tract) and a 7 nt branch point sequence – do not seem to contain sufficient information for this purpose, since such combinations of sequences occur within large introns at frequencies greater than the actual splice sites (Senapathy et al. 1990; Sun and Chasin 2000; Chasin 2007). Additional information can be provided in the form of splicing enhancer elements located at various positions within the exons (ESEs) or their intronic flanks (ISEs) and in similarly placed splicing silencers (ESSs and ISSs). In general ESE elements are bound by SR-proteins and ESSs and ISSs by hnRNP proteins but proteins outside these exact categories are also often involved (for reviews see (Ladd and Cooper 2002; Black 2003; Bourgeois et al. 2004; Zheng 2004; Pozzoli and Sironi 2005)).

Complete catalogues of these regulatory sequence motifs have been sought by protein binding determinations, functional selections, and validated computational predictions. The results have in a sense been too successful, in that by now at least 75% of the nucleotides in a typical human exon reside in motifs that have been found to influence splicing in one study or another (Chasin 2007). This high density of regulatory information often makes it difficult to make genetic perturbations that cleanly test the role of a particular motif. Three examples of this emergent ambiguity are presented in Fig. 2.1. Example 1 shows 2 typical cases of a SELEX winner from a functional selection for splicing activity (Liu et al. 1998). The fact that many motifs are likely to be found in any random sequence leads to the presence of a high noise level in such experiments. (A complete analysis of all sequences underlying the ESEfinder program is presented in Supplementary Fig. S2.1.) Example 2 shows the sequence resulting from the insertion of a putative exonic splicing silencer (PESS) that we examined for silencing activity in a test exon (Zhang and Chasin 2004). Besides the addition of the ESS, several enhancer motifs were created and a pre-existing silencer of another class was disrupted. Example 3 shows the substitution in a test exon of a predicted exonic splicing regulator (ESR) motif that reduced splicing efficiency (Goren et al. 2006). Concomitant with the substitution, an ESE was disrupted and an additional ESR was created. These examples are typical rather than exceptional. Thus the very act of placing a motif at an exactly specified location often changes the nature of the exon in unintended ways, reminiscent of the Heisenberg uncertainty principle (Heisenberg 1927).

The context of a splicing regulatory motif (Kanopka et al. 1996; Mayeda et al. 1999; Goren et al. 2006) or of a splice site (Lear et al. 1990; Carothers et al. 1993; Hwang and Cohen

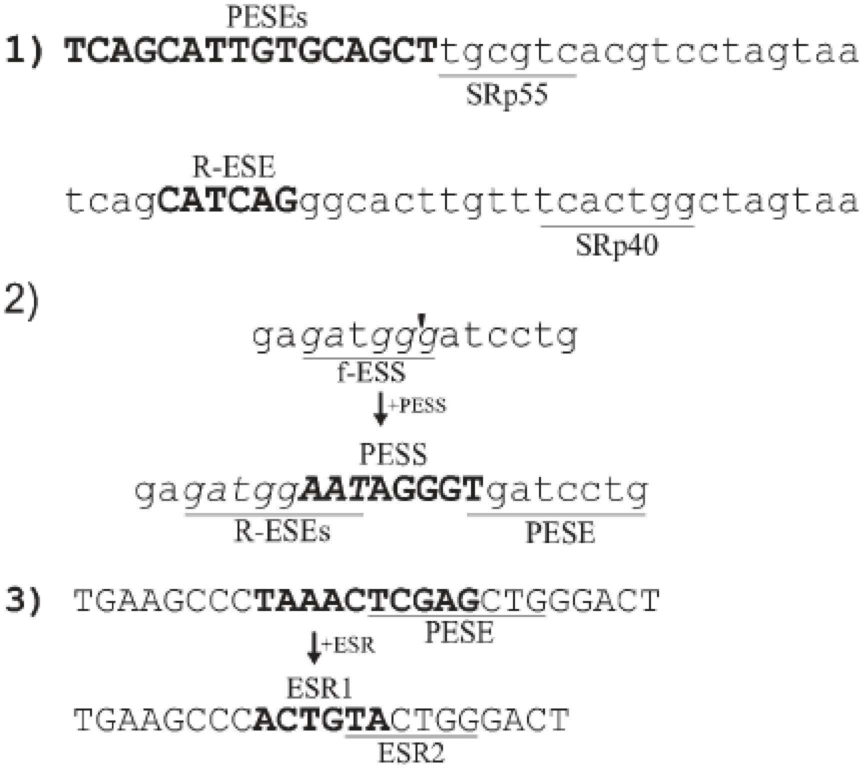


Figure 2.1. Examples of ambiguity in the identification or testing of splicing regulatory motifs. 1) Top: A functional SELEX selected sequence (Liu et al. 1998) that conferred responsiveness to SRp55. The match to the derived ESEfinder (Cartegni et al. 2003) SRp55 consensus sequence is underlined. The sequence also contains overlapping predicted PESE motifs (bold). Bottom: A functional SELEX selected sequence that conferred responsiveness to SRp40, with the SRp40 ESEfinder motif underlined. The sequence also contains a RESCUE-ESE (Fairbrother et al. 2002) motif (bold). These sequences are taken from those underlying ESEfinder (v.2; http://rulai.cshl.edu/tools/ESE2/) and were provided by Adrian Krainer. A similar analysis of all the sequences used by ESEfinder is presented in Supplementary Fig. S2.1. 2) Testing of a predicted PESS by its insertion into a test exon (*thbs4* exon 13) by Zhang and Chasin (Zhang and Chasin 2004). The bold sequence at the bottom was inserted into a BamHI site (arrowhead) in a test exon. Beyond the addition of the PESS, a fashex3 ESS (f-ESS) was disrupted (underlined in top sequence), a PESE was created (underlined in the bottom sequence), and 2 overlapping RESCUE-ESEs were created (R-ESEs, underlined in the bottom sequence). 3) Testing a predicted exonic splicing regulator (ESR, bold 6-mer at bottom) by substituting it for a 10-mer (bold at top) in a test exon (Goren et al. 2006). Besides the addition of the ESR, a PESE (underlined at top) was disrupted and an additional ESR (underlined at bottom) was created.

1997) can exert a strong influence on splicing efficiency. This context can be viewed in molecular terms by the quality and proximity of splicing regulatory motifs relative to each other and relative to the splice sites. A straightforward molecular genetic approach to test such a model would involve varying these parameters, but due to the density of regulatory motifs such variations would almost always change several parameters at once, and so confound the interpretation. One way to get around this problem might be to search for rare exons containing just a few well defined regulatory motifs that are each separated by sequences predicted to have no effect on splicing, i.e., neutral sequences. Another way would be to construct such exons in silico and then in vitro, using as building blocks known motifs that have enhancing, silencing or neutral effects. Here we have used the latter approach, using a prototype ESE, ESS and a putatively neutral 8-mer motif. We have assembled exons with random combinations of these elements placed between a constant pair of natural 3' and 5' splice sites. These "designer exons" have a general requirement for the ESE modules to achieve efficient splicing and are inhibited by the inclusion of the ESS modules. Despite their apparently simplified modular organization, splicing of these designer exons exhibits a complex dependence on the exact pattern of the ESEs and ESSs present.

Results

Design of the exons

We used 3 modules to build designer exons: an exonic splicing enhancer (ESE), an exonic splicing silencer (ESS), and a neutral sequence. Each module consisted of one particular

8-nt sequence chosen from among putative ESEs and ESSs and neutral sequences we previously identified on the basis of their overrepresentation in exons vs. human transcript regions that do not undergo splicing (Zhang and Chasin 2004). Libraries of exons consisting of multiple instances of one ESE and one ESS motif were created by using linkers with complementary overhangs (Fig. 2.2A) for the random ligation of the synthetic sequences. The linkers were designed to create a neutral motif upon ligation (Fig. 2.2B); thus the same neutral spacer is in place between each and every enhancer or silencer motif. The ligation products were inserted between 3’ and 5’ splice sites taken from intron 1 and intron 3, respectively, of the Chinese hamster dihydrofolate reductase (*dhfr*) gene. The exons so formed constituted the middle exon of a 3-exon minigene (Fig. 2.2C) with the 5’ exon being *dhfr* exon 1 (with its promoter) and the 3’ exon consisting of the fused *dhfr* exons 4 through 6 (with the first *dhfr* polyA site). Each designer exon also contains a neutral sequence at each end of the stretch of modules, generated as part of the insertion process (Fig. 2.2C).

The first and key step in the experimental design was to select an appropriate combination of ESE, ESS and neutral sequence modules. We required the combination of ESE/ESS/neutral sequences to meet several criteria. The first, and only essential, criterion was that the concatenation of these modules in any order should not yield any sequence that falls outside a neutral range (see below). Second, their concatenation should not produce any in-frame stop codons, to rule out nonsense codon mediated decay (NMD) as a factor. Third, the ESE and ESS should contain distinct restriction sites, to facilitate the determination of their order by partial digestion.

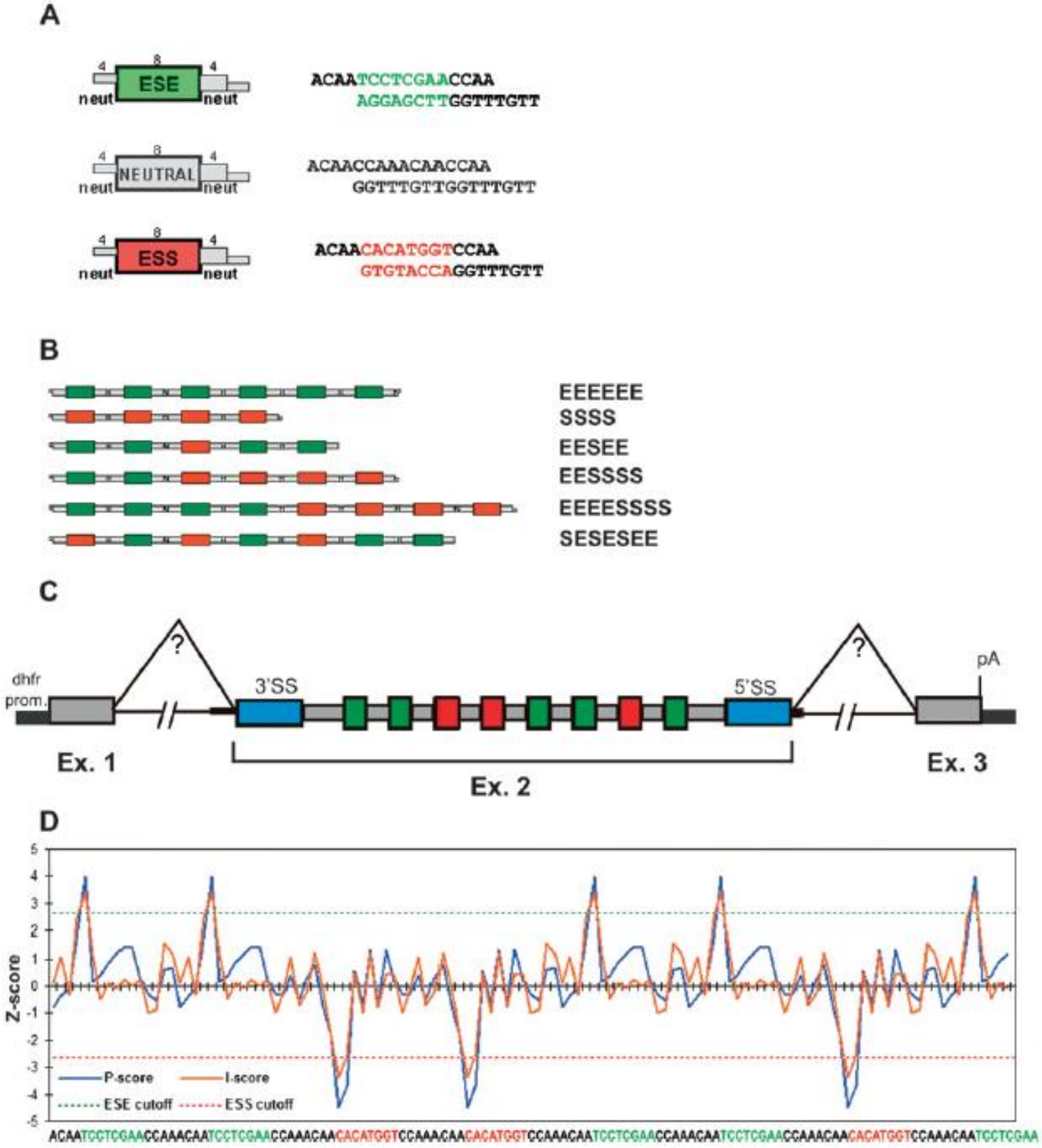


Figure 2.2. Construction of designer exons. A. Cartoons of E, N, and S modules showing single-stranded ends used for ligation, along with the actual sequences. B. Color-coded examples of possible designer exons (green, ESE; red, ESS, gray, neutral) along with the abbreviated notation used (E,S,N). Note that the abbreviated notation does not indicate the neutral 8-mer that lies between each E and S module and at each end. C. Diagram of a designer exon within the test minigene used. Exon 1 is exon 1 of the Chinese hamster *dhfr* gene, exon 3 comprises the fused exons 4 to 6 of the *dhfr* gene. The 3' and 5' splice sites (SS) are from *dhfr* introns 1 and 3, respectively. D. Z-score profile of all 8-mers in a possible designer exon (EESSEESE). The E and S modules generate salient signals over an otherwise unremarkable landscape.

We previously devised a scoring scheme and identified lists of octamers as putative splicing enhancers and silencers (PESE and PESSs, (Zhang and Chasin 2004)). In that work, each octamer was assigned two z-scores based on its over/under-representation in internal non-coding exons versus: 1) pseudo exons and 2) 5’-UTRs of intronless genes. The two z-scores were called the P-score and I-score, respectively. Underrepresented octamers were assigned negative z-scores. An octamer was called a PESE if both scores were greater than 2.62 or a PESS if both scores were lower than -2.62. Based on these criteria, we collected a list of ~2000 PESEs and ~1000 PESSs. We searched this list for PESE, neutral and PESS sequence combinations that fulfilled the criteria discussed in the above paragraph, requiring the neutral spacer sequence to have a z-score with an absolute value of less than 1.8. From among millions of 8-mer combinations, only about 3 dozen met all the criteria. We chose the following 3 sequences to build designer exons: TCCTCGAA (an ESE, P-score +3.99, I-score +3.44), CCAAACAA (a neutral sequence: P-score -0.28, I-score -0.98) and CACATGGT (an ESS, P-score -4.50, I-score -3.38), which we term “E”, “N”, and “S” for brevity. An example of the distribution of these scores across all of the 8-mers of a typical designer exon is shown in Fig. 2.2D.

Enhancers are required for the efficient splicing of designer exons

We first tested these sequences for their effect on splicing by inserting each singly into a BamHI site in the central exon (*chuk* exon 8) of a 3-exon minigene, and measuring the proportion exon inclusion (included/(included + skipped)) after transfection into human 293 cells and semi-quantitative PCR, as described in our previous study (Zhang and Chasin 2004). As expected, the E sequence promoted splicing of a poorly spliced version of the *chuk*8 exon, the S sequence inhibited splicing of a well spliced *chuk*8 exon, and the N sequence had little effect on either type of exon (data not shown). We next constructed homogeneous designer exons made up of multiple copies of just a single type of motif (E, S, or N), constructed as described above and in Fig. 2.2. Designer exons made up of E modules spliced very well, whereas those made up of S modules showed little or no splicing. Designer exons made up exclusively of concatenated N modules were also very poorly spliced: an exon with 3 modules (seven N 8-mers counting the spacers) was included 13% of the time and an exon with 5 modules (eleven N 8-mers counting the spacers) showed almost no inclusion. Thus these designer exons in this context require an enhancer for efficient splicing. We then went on to assemble a large number of additional designer exons carrying both E and S modules and test their splicing efficiency after transfection into 293 cells.

Splicing of designer exons carrying randomly combined E and S motifs

We ligated E and S motifs at various ratios, inserted the ligation products into the 3-exon minigene vector and isolated 139 clones. The number and order of the modules was quickly determined by PCR amplification of the plasmid region spanning the designer exon using a fluorescently labeled primer followed by partial digestion with TaqI (TCGA) and with CviAII (CATG), as these sites are present in the E and S modules, respectively. From the ladder of fluorescent bands seen after electrophoresis, the arrangement of modules could be deduced (Fig. 2.3A). Splicing was then measured after transfection of 293 cells with plasmid DNA and using semi-quantitative radioactive RT-PCR (Chen and Chasin 1993) to quantify molecules that

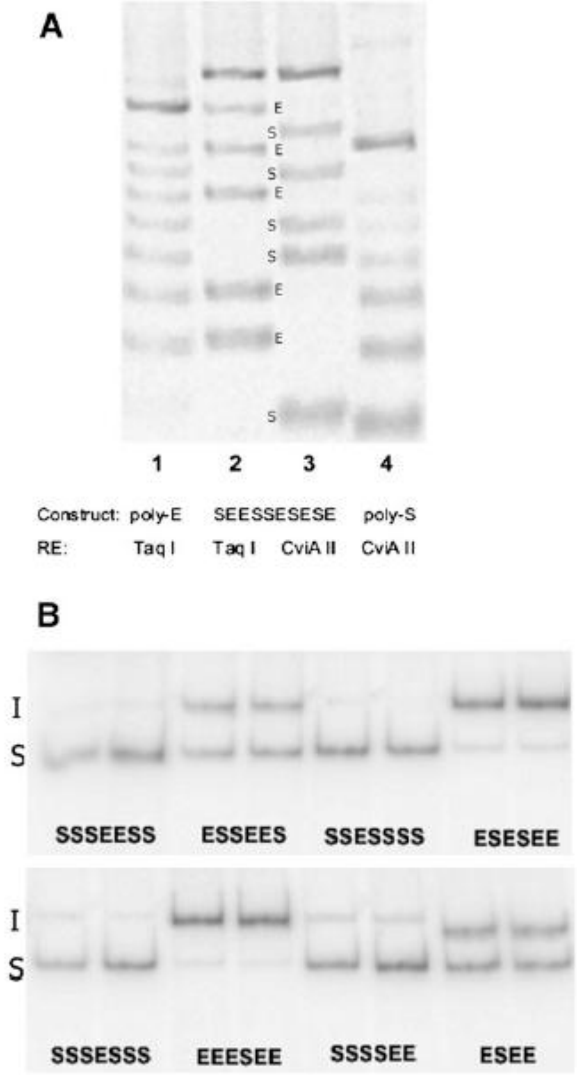


Figure 2.3. Determination of designer exon genotypes and phenotypes. A. Module order screening. Plasmid DNA from designer exon clones was PCR amplified using one fluorescently tagged primer and then cleaved with the diagnostic restriction enzymes (RE) TaqI, which cuts in the E module or CViAII, which cuts in the S module. Lane 1 represents a clone with 7 E modules and no S modules cut with TaqI and lane 4 represents a clone with 6 S modules and no E modules, cut with CviAII; these lanes serve here as standards. Lanes 2 and 3 represent the same 10-module clone (SEESSESESE) cut with either TaqI (lane 2) or CviAII (lane 3). The relative positions of the bands (labeled E or S) allow the order of E and S modules to be read directly from the gel. All constructs used for analysis were subsequently DNA sequenced. B. Splicing phenotype measurement. Plasmids harboring designer exons were transfected into HEK 293 cells and the mRNA products were amplified by radioactive RT-PCR. The relative amounts of molecules that included (I) or skipped (S) the designer exon were determined by Phosphorimaging. The analysis of 8 representative clones is shown, with two independent transfections for each clone. The sequence of modules present in each exon is shown below each pair of lanes.

included or skipped the designer exon. An example of these results is shown in Fig. 2.3B. The splicing efficiencies of 139 exons are presented in Fig. 2.4, where they are ranked according to the proportion of designer exon inclusion along and with a graphical depiction of their structure. Reading from left to right it can be seen by eye that in general splicing efficiency decreases as the number of Es (green boxes) per exon decreases and as the number of Ss (red boxes) increases. It should be kept in mind that in between each E and/or S there exists an N module that is not depicted. Less easily seen but also discernible is a tendency for splicing efficiency to decrease with exon length.

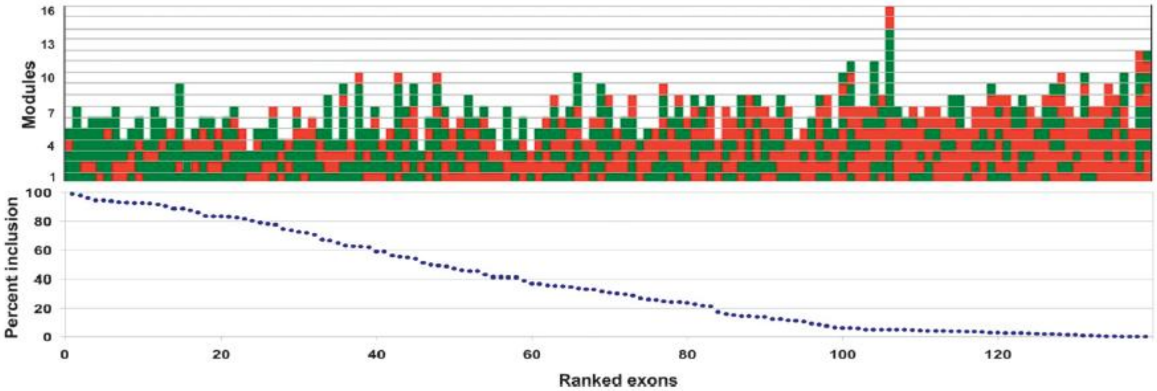


Figure 2. 4. Splicing of designer exons. Bottom: Splicing of 139 designer exons ranked by percent exon inclusion. Exon inclusion is defined as: 100 x included/(included+ skipped), and the value is the average of at least 2 independent transfections (average SE= 16%). Top: Structure of the corresponding designer exons. Each colored rectangle represents an E (green) or S (red) module. The 5' end of the exon is at the bottom. Exon inclusion levels for these designer exon are presented in tabular form in Supplementary Table 2.1.

A more quantitative assessment of correlations between splicing efficiency and designer exon structure was made by calculating Pearson's correlation of determination, R^2, from scatter plots of the data (Fig. 2.5). Each of the 6 charts in Fig. 2.5 tests a hypothesis about the dependence of splicing on these regulatory sequences. The first is that splicing is proportional to the absolute number of enhancer modules in an exon. We found a significant, if weak,

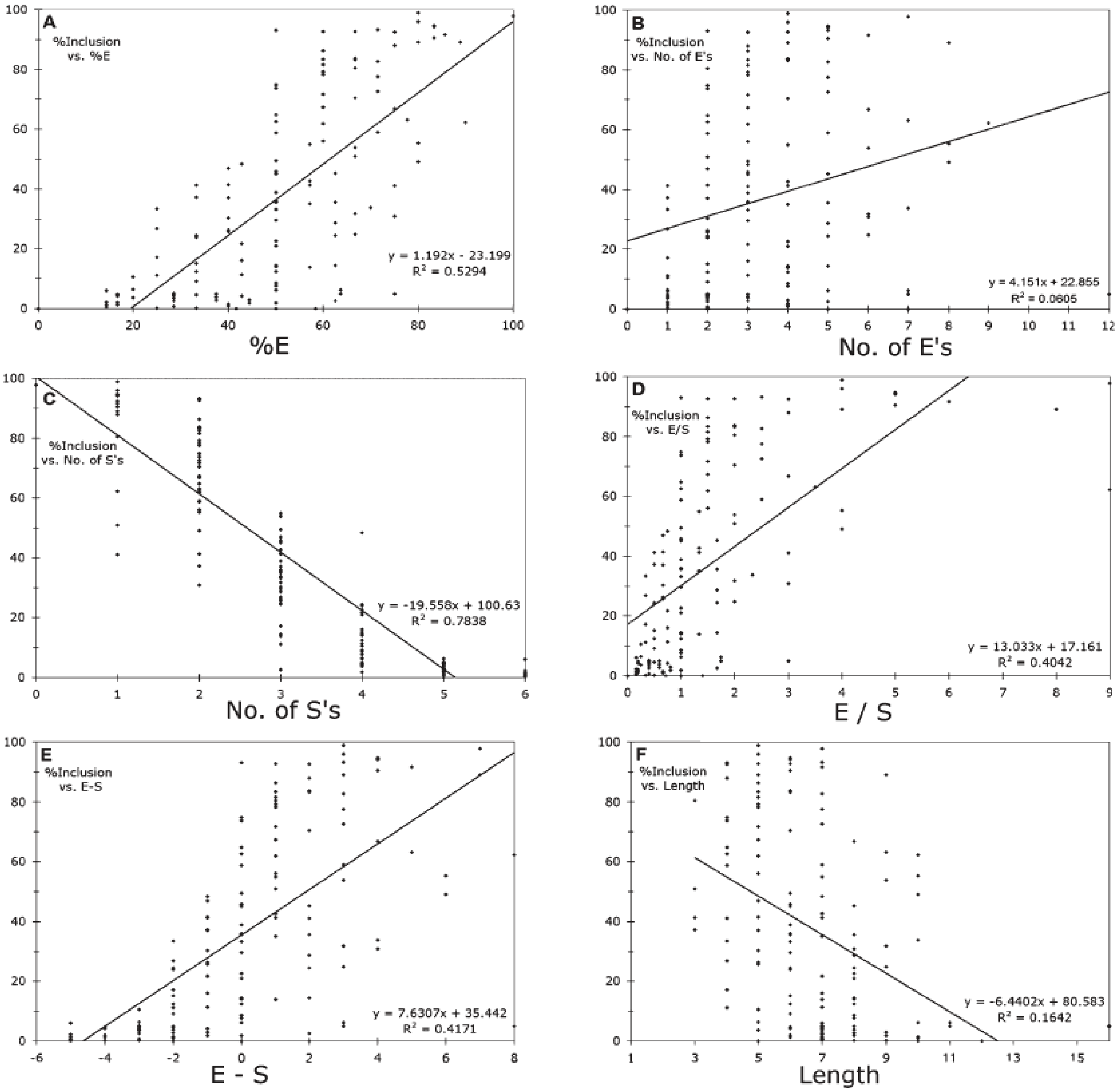


Figure 2.5. Correlations between exon inclusion and splicing regulatory elements. Straight lines were fitted by a linear regression (Excel). A) Percent enhancers (100 x E/(E+S)); ; B) Number of enhancers; C) Number of silencers; D) Ratio of enhancers to silencers; E) Number of enhancers minus number of silencers; F) Exon length (number of E's plus S's).

correlation between and splicing and the number of Es (R^2 = 0.06, p=0.004, t-test, Fig. 2.5B). The correlation was much stronger when we considered the proportion of Es in an exon (R^2 = 0.53, p < 3e-24, Fig. 2.5A). A converse hypothesis is that it is the silencers that play the most important role in determining splicing efficiency. Indeed, the number of Ss per exon produced a much stronger (negative) correlation with inclusion rate (R^2 = 0.78, p < 5e-47, Fig. 2.5C) than the number of Es. The correlation for the proportion of Ss is the same as for the proportion of Es by definition here, although of opposite sign.

The factors governing splicing decisions, or the splicing code, have often been ascribed to a balance between positive and negative factors, so we tested the effect of combining the E and S content of each exon by calculating the E/S ratio and the E-S difference. To our surprise, these variables were less correlated with splicing (R^2 = 0.40 and 0.42 respectively, Fig. 2.5D and 2.5E) than the proportion of Es or Ss considered independently, suggesting that combining Es and Ss in these ways added more noise than information. We also examined exon length, a variable related to E or S content, and found a weaker but highly significant negative correlation with splicing (R^2 = 0.16, p < 1e-6, Fig. 2.5F). The strongest correlation seen was with the number of Ss (Fig. 2.5C), indicating that silencing was the most important factor at play in these exons. However, the negative effect of the number of Ss is actually a measurement of two variables: percentage of Ss and length, since longer exons will tend to have more Ss. The effect of Ss normalized for exon length can be seen in the plot of %E (Fig. 2.5A), which is equivalent to 1-%S, and which shows a lesser but still strong correlation (0.53 for %E vs. 0.78 for number of Ss).

Splicing of designer exons carrying no silencers

Although the correlation coefficients for %E and number of Ss indicate that most of the variance can be explained by a linear relationship between inclusion and these variables, the fact remains that there is considerable scatter among these points. For example, in the plot with the best correlation (Fig. 2.5C), exons all having 3 Ss yielded inclusion levels ranging from 3% to 55%. We considered the possibility that complexities inherent in antagonism between E's and S's tend to produce a metastable state, and that a better correlation between splicing and Es would be seen in the absence of added silencers. Designer exons were therefore constructed by randomly ligating E and N modules. Here again we point out that there were always additional N modules as spacers between each pair of named modules, plus one at each end: i.e., the designation ENE represents the sequence nEnNnEn, where the lower case n is formed in the course of construction and is the same sequence as N. Twenty-two EN exons were analyzed for splicing. The results are shown in Fig. 2.6A, which displays the inclusion levels that correspond to the specific exon structures. The correlation between inclusion and %E was somewhat better for these E+N exons ($R^2 = 0.75$, $p < 6e-13$, Fig. 2.6B) compared to E+S exons ($R^2 = 0.53$, $p < 3e-6$, Fig. 2.5A), but there was still considerable scatter: at E = 50% the inclusion levels ranged from 38% to 94%.

Consideration of predicted secondary structure

It is possible that many of the E and S sequences included in these exons are not available for enhancing or silencing splicing because they are sequestered in the double-stranded stems of secondary structures. Anecdotal examples of secondary structure affecting splicing are many (reviewed in (Buratti and Baralle 2004)) and a survey of functional ESEs showed that these tend to remain single stranded (Hiller et al. 2007). If only the single stranded Es in our designer exons were functional, then we might see a better correlation between inclusion and this subset of Es.

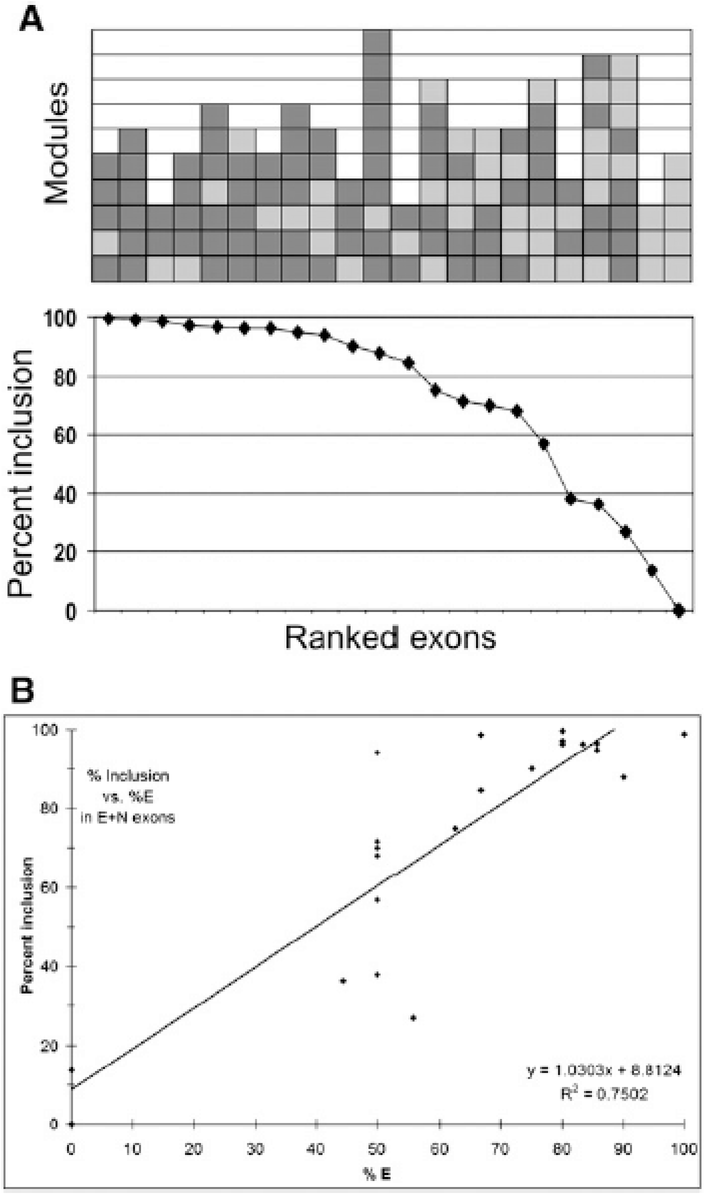


Figure 2.6. Splicing of exons designed without silencers. A. Bottom: Splicing of 22 designer exons ranked by percent exon inclusion. Top: Structure of the corresponding designer exons. Each colored rectangle represents an E (dark gray) or N (light gray) module. An additional N module is present between each E and N module but these are not depicted. The 5’ end of the exon is at the bottom. B. Correlation between exon inclusion and the proportion of E elements in an exon. Exon inclusion levels for these designer exons are presented in tabular form in Supplementary Table 2.1.

We used RNAstructure (Mathews, 2006) to fold each of the E+S designer exons and then assigned each base in the E, N, and S modules a probability of being in a double stranded stem (s), a single stranded loop (l), or a single stranded interstem region (i). The designer exons as a whole did not form exceptionally stable secondary structures; for the most stable structures the average free energy value per nucleotide was -0.22 kcal/mole compared to -0.21 for scrambled versions. Correlation coefficients were calculated between exon inclusion level and each of these 9 variables (E_s, E_l, E_i; N_s, N_l; S_s, S_l, S_i). As can be seen in Table 2.1, none of the R^2 values for the proportion of E or S nucleotides that are confined to stems, loops or interstem regions was appreciably greater than the value for the E or S nucleotides as a whole. These results do not support the idea that variable secondary structures underlie the wide ranges of inclusion levels for designer exons that have the same proportion of E or S modules.

Table 2.1. Correlation coefficients between inclusion level and the proportion of module bases found in different types of predicted secondary structures

Region	% in Es	% in Ns	% in Ss	% in entire exon (E+N+S)
stems	0.428[a]	-0.339	-0.482	-0.132
loops	0.390	0.034	-0.562	-0.202
interstems	0.347	0.381	-0.152	0.238
All 3	0.529	0.148	-0.529	

[a] Numbers throughout represent Pearson's R^2 values, with a negative sign denoting a negative correlation.

Discussion

Responses to enhancers and silencers

The concatenation of single specific enhancer and silencer modules has been used here to construct exons that are much less complex than their natural counterparts. The plainness of these exons has allowed us to test some simple hypotheses regarding internal exon recognition in pre-mRNA splicing.

The first hypothesis is that enhancers are necessary for efficient splicing, and it is supported by our results. Designer exons consisting solely of neutral sequences spliced poorly if at all; incorporation of enhancers was required to achieve inclusion levels near 100%. As yet, this conclusion is limited to the context we provided, the natural splice sites and intronic flanks found in the *dhfr* gene. These 3' and 5' splice sites are of average or above average strength (i.e., agreement with the consensuses), with consensus values (Senapathy et al. 1990; Zhang et al. 2005b) of 81 and 88 respectively. It is likely that provision of stronger splice site sequences could obviate the need for an enhancer (see for example (Ram et al. 2008)). Nonetheless, most natural exons do not have splice sites stronger than those used here.

The second hypothesis is that splicing efficiency increases in proportion to the number of enhancer elements, and it is less directly supported by our data. Hertel and Maniatis showed that in vitro splicing of a 2-exon transcript responded linearly to the addition of multiple enhancer elements downstream of the 3' splice site (Hertel and Maniatis 1998). Our results for an internal exon agree with this finding in that a highly significant correlation ($R^2 = 0.53$) to a linear model was found for exon inclusion vs. the number of enhancers per exon if the data were normalized for exon length differences (%E, Fig. 2.5A). However, the great splicing variability seen among exons having the same proportion of enhancers belies the simple model in which the mere presence of an enhancer sequence adds linearly to the probability of binding an activator protein which in turn leads to a proportional increase in splicing.

We tried to take into account the possibility that some of the included motifs were being sequestered in secondary structures, but our test did not provide support for this idea. In particular, the simple notion that E's are much more effective when present as single-stranded targets was not substantiated. This negative result cannot be considered conclusive, as our ability to predict the in vivo secondary and tertiary structures of RNA molecules is limited. RNA folding in vivo may be influenced by RNA binding proteins that unwind, compete with or enhance RNA-RNA interactions. Some of this secondary structure analysis was rather surprising, suggesting that S modules were more effective in inhibiting splicing when present in stems and loops compared to interstem regions (Table 2.1). It is possible that the particular S motif we used is a better target when presented in double-strand form. The N modules also showed an inhibitory effect when present in stems but a stimulatory effect when present in interstem regions (Table 2.1). These effects could be indirect, for instance by N sequences forming a stem that places an S module in a loop. These ideas are amenable to experimental testing.

The third hypothesis is that splicing is the result of a balance between positive and negative elements. This oft-quoted inference has gained wide acceptance based mostly on its reasonableness, but has rarely been tested using multiple elements. Our data has supported this idea in a general sense: whether the balance is considered the difference between enhancer and silencer content or their ratio, R^2 values of 0.4 were obtained. These ways of defining balance proved no better than that given by the proportion of enhancer motifs (%E), which intrinsically also compares enhancer to silencer content (%E =100x E/(E+S)) in this system. However, the scatter in the data suggests that splicing is highly dependent on the relative positions of the E and S motifs, beyond their relative proportions. As yet we cannot posit a straightforward mathematical model capable of predicting splicing patterns based on the positions of the E and S modules. Table 2.2 illustrates this difficulty by showing 3 examples of pairs of compositional isoforms exhibiting 2- to 10-fold differences in splicing efficiency despite having identical E/S ratios and E-S differences.

Table 2.2. Designer exon pairs with the same composition but different splicing behavior.

	Modules	Exon length (nts)	Percent inclusion
1	ESESESE	126	55
	SEEEESS	126	14
2	EESE	78	93
	ESEE	78	41
3	EESESSSE	142	21
	ESSSEEES	142	2

Designer exons as a model system

The synthetic biology used to create designer exons results in molecules that are unlike any found in living cells. One might argue that we run the danger of being misled by the analysis of such artificial molecules; their behavior will tell us little about the rules governing the splicing of their more complex natural counterparts. We contend the opposite, that a prerequisite to understanding how individual elements combine to yield an emergent property requires an understanding of how the individual elements act, first alone, and then in simple combinations. The way in which transcriptional regulatory signals combine to produce what is often a binary decision is in many ways analogous to the splicing decision problem, and can also be explored using synthetic promoters (Alper et al. 2005; Ligr et al. 2006). Additional examples of this approach lie in the de novo design of proteins (Kuhlman et al. 2003) and cell membranes (Tanaka and Sackmann 2005).

We have not implemented an orchestrated combinatorial approach in these first experiments, but rather relied on seeing simple patterns emerge from randomly assembled molecules. Such was not the case, pointing to our ignorance of important parameters yet to be defined. We can speculate on several ways in which parameters may have been hidden in these designer exons. First, the assumption that the N modules are truly neutral may be wrong. The N module was chosen from among 10000 sequences predicted to be neutral and was evaluated in an arbitrarily chosen test exon; in a different context it may not act neutrally. Second, we may have been overzealous in isolating the E motifs from each other and from the S motifs. The inclusion of 8-nt neutral spacers between each motif may be precluding important interactions that require close apposition. In this regard, we see no reason to assume that fundamentally neutral spacers are not occupied by RNA-binding proteins and so act passively to prevent positive interactions. Third, the multiplicity of exon lengths that were produced by random ligation added a confounding factor in the interpretation of the results.

Many of these problems will be solved by synthesizing designer exons in a more deliberate fashion. To this end we are developing methods to readily synthesize exons of defined size, and adding specific modules one or two at a time to produce a stable of exactly designed molecules. For instance, the placement of an E at 10 evenly spaced positions within a 160 nt exon will answer the question of whether proximity to a splice site is required for ESE action and whether there is specificity of a given E for a 3' or 5' splice site. This information can then be extended to real exons to weigh the potential of embedded ESEs and so improve exon prediction and the prediction of alternative splicing efficiency. The interplay between splice site strength and ESE and ESS effectiveness can also be weighed first in designer exons and then applied to natural exons. Finally, an all N vector a with a unique restriction site will provide a convenient and perhaps more reliable vector for testing candidate ESEs gathered from real genes. Thus despite the complexities revealed by this first generation analysis, we think that the properties of these synthetic molecules are likely to be useful for discovering properties of their natural counterparts.

Materials and Methods

Construction of designer exons

A starting plasmid was pDCH1P12D, which contains a Chinese hamster *dhfr* minigene consisting of exon 1, a hybrid intron1+3, exons 4 through 6 of the cDNA and the first natural *dhfr* polyA site in exon 6; a NotI site was added near the center of the intron, which also contains a unique NheI site. The chuk8 minigene used for the initial test of candidate motifs was constructed by inserting exon 8 of the human *chuk* gene into the NotI site, either with about 50 nt of intronic flanks or with just its splice sites. The construction of these *chuk*8 minigenes has been described previously (Zhang et al. 2005a; Zhang et al. 2005c). To construct designer exons, the two strands of E, S or N modules were first mixed in an annealing buffer (30mM pH7.4 HEPES, 2mM MgOAc, 100mM K_2OAc) at equal concentration (~5ug/ul). The mixtures were heated to 95^oC and then gradually cooled at room temperature. The annealed E, S or N modules were mixed at different ratios (1:3, 1:1 or 3:1) with T4 ligase (10 U in a 20 ul reaction mixture, 0.5 ug of each module). The ligation mix was subsequently electrophoresed in a 2% agarose gel and the products corresponding in length to 3 to 16 modules were extracted from the gel. These fragments were then ligated as a pool to the constant 5' and 3' modules containing an upstream EagI or downstream NheI restriction site, respectively. The product of this second ligation was then subjected to PCR with primers targeting the constant 5' and 3' modules. The PCR products were cut with EagI and NheI and inserted into pDCH1P12D that had been cut with NotI and NheI, the latter located 300 nt downstream the Not I site of pDCH1P12D. The resulting minigenes have the designer exon located between a 300 bp upstream intron and a 600 bp downs tream intron. The E and S module sequence of individual transformant colonies was determined by PCR amplification of the region spanning the designer exon using one primer 5’ end labeled with Cy5 followed by partial separate digestions with TaqI and CviAII, for which there are restriction sites in the E and S sequences respectively. After electrophoresis, the fluorescent bands were visualized using a Phosphorimager Storm (Fig. 2.3A). The sequence of the E and S modules could be read from the partial digest patterns. All unique designer exons chosen for analysis were subsequently sequenced (GeneWiz).

Measurement of splicing

Human HEK293 cells were transfected with plasmid DNA using Lipofectamine 2000 (Invitrogen) according to the manufacturer’s protocol. After 24 hours, total RNA was isolated using RNAwiz (Ambion), reverse transcribed with Omniscript and random hexamers from QIAGEN and subjected to radioactive semiquantitative PCR as described previously (Chen and Chasin 1993). Percent inclusion was calculated as 100 x included/(included + skipped), using Phosphoimager counts of the indicated electrophoretic bands taking into account the number of labeled bases in each molecule. Each transfection was performed at least twice; the average standard error for biological replicates was 16%.

Secondary structure analysis

The 139 E+S designer exon sequences were folded from -100 to +100 relative to the exon ends using RNAstructure 4.5 (Mathews 2006) for DOS with default settings. The output of this program included the 20 most stable structures in .ct table format. The .ct values from all 20 structures were converted to Vienna dot-bracket depiction. The average designation of each base in each of the 3 structural categories (stem, loop, and interstem) was recorded, along with its identity as part of an E, N or S motif. Pearson's correlation coefficient (r) and coefficient of determination (R^2) were calculated for % inclusion vs. % of motif bases in each of the various predicted structural classes using a custom written computer program.

Authors' Contributions

LC and MA conceived the study. MA performed preliminary experiments and designed the protocol for designer exon construction and measurement. XZ refined the protocol, made all designer exons with ESE/ESS combinations and measured their inclusion level. SK made all designer exons with ESE/neutral combinations and measured their inclusion level. XZ and LC performed the analysis of the results. LC wrote the text and MA revised it.

Acknowledgements

We thank Ashira Lubkin for a critical reading of the manuscript. This work was supported by a grant from the NIH (GM072740) to L.A.C.

References

Alper H, Fischer C, Nevoigt E, Stephanopoulos G. 2005. Tuning genetic control through promoter engineering. *Proc Natl Acad Sci U S A* **102**(36): 12678-12683.

Black DL. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72**: 291-336.

Bourgeois CF, Lejeune F, Stevenin J. 2004. Broad specificity of SR (serine/arginine) proteins in the regulation of alternative splicing of pre-messenger RNA. *Prog Nucleic Acid Res Mol Biol* **78**: 37-88.

Buratti E, Baralle FE. 2004. Influence of RNA secondary structure on the pre-mRNA splicing process. *Mol Cell Biol* **24**(24): 10505-10514.

Carothers AM, Urlaub G, Grunberger D, Chasin LA. 1993. Splicing mutants and their second-site suppressors at the dihydrofolate reductase locus in Chinese hamster ovary cells. *Mol Cell Biol* **13**(8): 5085-5098.

Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR. 2003. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res* **31**(13): 3568-3571.

Chasin LA. 2007. Searching for Splicing Motifs, in "Alternative Splicing in the Postgenomic Era, " B. Blencowe and B. Graveley, eds., pp. 85-106. Landes Bioscience, Austin, TX 78701

Chen IT, Chasin LA. 1993. Direct selection for mutations affecting specific splice sites in a hamster dihydrofolate reductase minigene. *Mol Cell Biol* **13**(1): 289-300.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* **297**(5583): 1007-1013.

Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G. 2006. Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers. *Mol Cell* **22**(6): 769-781.

Heisenberg W. 1927. Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik*, **43**: 172-198.

Hertel KJ, Maniatis T. 1998. The function of multisite splicing enhancers. *Mol Cell* **1**(3): 449-455.

Hiller M, Zhang Z, Backofen R, Stamm S. 2007. Pre-mRNA secondary structures influence exon recognition. *PLoS Genet* **3**(11): e204.

Hwang DY, Cohen JB. 1997. U1 small nuclear RNA-promoted exon selection requires a minimal distance between the position of U1 binding and the 3' splice site across the exon. *Mol Cell Biol* **17**(12): 7099-7107.

Kanopka A, Muhlemann O, Akusjarvi G. 1996. Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. *Nature* **381**(6582): 535-538.

Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. 2003. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**(5649): 1364-1368.

Ladd AN, Cooper TA. 2002. Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol* **3**(11): reviews0008.

Lear AL, Eperon LP, Wheatley IM, Eperon IC. 1990. Hierarchy for 5' splice site preference determined in vivo. *J Mol Biol* **211**(1): 103-115.

Ligr M, Siddharthan R, Cross FR, Siggia ED. 2006. Gene expression from random libraries of yeast promoters. *Genetics* **172**(4): 2113-2122.

Liu HX, Zhang M, Krainer AR. 1998. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev* **12**(13): 1998-2012.

Mathews DH. 2006. RNA secondary structure analysis using RNAstructure. *Curr Protoc Bioinformatics* **Chapter 12**: Unit 12 16.

Mayeda A, Screaton GR, Chandler SD, Fu XD, Krainer AR. 1999. Substrate specificities of SR proteins in constitutive splicing are determined by their RNA recognition motifs and composite pre-mRNA exonic elements. *Mol Cell Biol* **19**(3): 1853-1863.

Pozzoli U, Sironi M. 2005. Silencers regulate both constitutive and alternative splicing events in mammals. *Cell Mol Life Sci* **62**(14): 1579-1604.

Ram O, Schwartz S, Ast G. 2008. Multifactorial interplay controls the splicing profile of Alu-derived exons. *Mol Cell Biol* **28**(10): 3513-3525.

Senapathy P, Shapiro MB, Harris NL. 1990. Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. *Methods Enzymol* **183**: 252-278.

Sun H, Chasin LA. 2000. Multiple splicing defects in an intronic false exon. *Mol Cell Biol* **20**(17): 6414-6425.

Tanaka M, Sackmann E. 2005. Polymer-supported membranes as models of the cell surface. *Nature* **437**(7059): 656-663.

Zhang XH, Arias MA, Ke S, Chasin LA. 2009. Splicing of designer exons reveals unexpected complexity in pre-mRNA splicing. *RNA* **15**(3): 367-376.

Zhang XH, Chasin LA. 2004. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev* **18**(11): 1241-1250.

Zhang XH, Kangsamaksin T, Chao MS, Banerjee JK, Chasin LA. 2005a. Exon inclusion is dependent on predictable exonic splicing enhancers. *Mol Cell Biol* **25**(16): 7323-7332.

Zhang XH, Leslie CS, Chasin LA. 2005b. Computational searches for splicing signals. *Methods* **37**(4): 292-305.

-. 2005c. Dichotomous splicing signals in exon flanks. *Genome Res* **15**(6): 768-779.

Zheng ZM. 2004. Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression. *J Biomed Sci* **11**(3): 278-294.

Supplemental Material

Table S2.1. Inclusion level for the set of designer exons used in this chapter.

Clone #	Sequence	Modules	Mean	SEM
	S-E combinations			
7-14-16	EEEEEEE	7	0.978	0.006
5-2-82	EEESE	5	0.989	0.007
7-13-34	EEEEEEEEES	10	0.623	0.183
7-11-22	ESEEEEEEE	9	0.890	0.033
7-53	EESEEEE	7	0.916	0.004
7-85	ESEEEE	6	0.947	0.013
5-2-90	SEEEEE	6	0.946	0.023
6-9-19	EEEESE	6	0.941	0.011
5-46	EEESEE	6	0.906	0.011
31-30	ESEEE	5	0.960	0.006
6-16-89	EEEES	5	0.890	0.036
30-21	EESE	4	0.926	0.015
31-40	EEES	4	0.879	0.024
27-24	ESEE	4	0.411	0.003
3-27-8	SEE	3	0.804	0.082
3-23-10	ESE	3	0.509	0.020
7-11-3	SEEEEEEEES	10	0.553	0.105
7-14-35	EEEEEESEES	10	0.491	0.109
7-14-26	EESEEEESE	9	0.631	0.099
6-13-31	ESSEEEEE	8	0.667	0.016
7-14-8	ESEEEEES	8	0.309	0.059
7-11-19	SSEEEEE	7	0.933	0.022
6-16-87	EEESESE	7	0.827	0.081

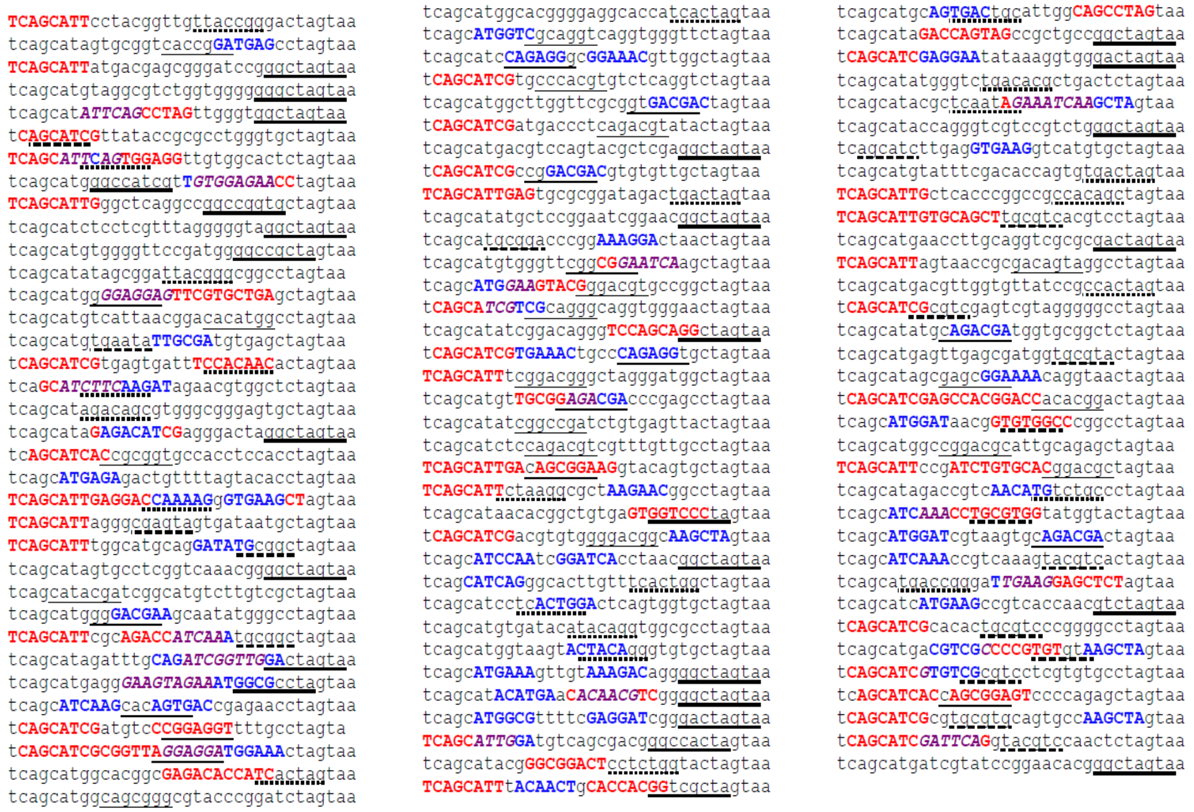
7-13-64	ESEEEES	7	0.776	0.025
31-14	EEEESES	7	0.727	0.008
5-2-6	SEEEESE	7	0.589	0.004
29-60	SEESEE	6	0.927	0.006
7-11-26	EEESSE	6	0.838	
5-39	ESESEE	6	0.835	0.032
30-31	EESSEE	6	0.833	0.023
7-11-72	EEEESS	6	0.705	0.044
25-3	ESSEE	5	0.927	0.001
6-9-87	ESESE	5	0.862	0.041
27-35	ESEES	5	0.835	0.039
7-72	EEESS	5	0.815	0.040
7-69	SEESE	5	0.792	0.033
6-16-65	SSEEE	5	0.783	0.077
6-9-42	SEEES	5	0.718	0.029
22-4	EESSE	5	0.673	0.039
30-27	SESEE	5	0.619	0.019
7-74	EESES	5	0.560	0.038
25-10	SSEE	4	0.932	0.017
30-8	SESE	4	0.748	0.007
31-16	ESES	4	0.737	0.013
31-34	ESSE	4	0.648	0.017
31-24	EESS	4	0.626	0.016
3-23-7	SEES	4	0.588	0.090
31-36	SES	3	0.413	0.013
31-42	ESS	3	0.373	0.005
6-14-72	SEEESESEEE	10	0.338	0.084
7-11-8	EEEESSSEE	9	0.538	0.071

29-86	SESEESEEE	9	0.317	0.003
6-13-71	EEESESEES	9	0.248	0.039
6-16-6	SSEESEEE	8	0.452	0.061
6-14-39	ESESEEES	8	0.357	0.065
22-19	EEESEESS	8	0.286	0.030
6-14-25	EESSEEES	8	0.244	0.055
5-2-35	EEEEESSS	8	0.143	0.031
5-2-7	SEESSEEE	8	0.025	0.002
6-14-38	ESESESE	7	0.549	0.036
29-30	SEESESE	7	0.427	0.046
5-2-80	SSSEEEE	7	0.412	0.005
6-13-42	EESSSEE	7	0.350	0.033
8-7	SEEEESS	7	0.139	0.004
6-9-55	SESESE	6	0.495	0.029
5-2-33	SESEES	6	0.458	0.016
6-13-46	SSEESE	6	0.451	0.030
5-2-16	SSESEE	6	0.390	0.011
5-31	ESSEES	6	0.359	0.036
6-13-9	ESESES	6	0.355	0.035
6-16-70	EESSES	6	0.331	0.066
3-23-2	ESSSEE	6	0.296	0.073
30-25	SEESS	5	0.469	0.009
6-9-60	SESSE	5	0.413	0.023
7-60	SSESE	5	0.370	0.014
8-16	ESSES	5	0.304	0.039
27-58	ESSSE	5	0.262	0.056
6-16-76	EESSS	5	0.258	0.053
31-31	SSES	4	0.334	0.005

7-27	SSSE	4	0.269	0.034
7-8	SESS	4	0.172	0.008
30-16	ESSS	4	0.112	0.014
7-14-20	EEEESSEEEEEEEESS	16	0.049	0.031
30-15	ESEEESSEESE	11	0.062	0.025
29-17	EESSSESEEEE	11	0.050	0.004
6-9-89	SSEESESE	8	0.226	0.018
5-2-59	EESESSSE	8	0.211	0.019
31-10	SESSSEEE	8	0.144	0.008
6-9-6	EESSSESE	8	0.140	0.015
6-9-17	ESSESSEE	8	0.123	0.008
7-12	ESEESSSE	8	0.085	0.007
6-14-64	ESESEESS	8	0.076	0.013
6-9-64	ESSSEEES	8	0.019	0.011
6-13-88	SSESESE	7	0.484	0.019
29-96	SSSESEE	7	0.217	0.005
6-14-76	ESEESSS	7	0.161	0.033
6-9-12	ESSSEES	7	0.113	0.001
5-2-62	EESSESS	7	0.042	0.006
7-76	SESSSE	6	0.244	0.030
6-14-30	ESSSES	6	0.239	0.057
25-13	ESSESS	6	0.151	0.013
8-23	SESESS	6	0.124	0.010
27-31	SSSSEE	6	0.091	0.022
7-57	ESESSS	6	0.045	0.004
22-12	SSESS	5	0.107	0.020
3-23-4	SSSSE	5	0.065	0.011
5-2-48	ESSSS	5	0.037	0.013

5-2-12	SEEESEESEESS	12	0.000	0.000
25-5	SEESSESESE	10	0.063	0.005
6-9-9	SSESESSEEE	10	0.000	0.000
6-13-32	EESESSSSE	9	0.030	0.003
5-2-11	SSSESEESE	9	0.017	0.000
6-9-28	SESSSEES	8	0.049	0.011
6-16-68	SSSESSEE	8	0.039	0.018
6-20-30	SSSSSEEE	8	0.038	0.025
6-16-22	ESSSESES	8	0.028	0.007
6-9-77	ESEESSSS	8	0.027	0.018
6-14-73	SSSESES	7	0.050	0.012
7-84	SSSESSE	7	0.048	0.007
6-9-32	ESSSSES	7	0.046	0.010
27-39	SSESESS	7	0.042	0.002
6-14-23	SESESSS	7	0.036	0.008
5-21	SSSEESS	7	0.035	0.005
7-75	SSSSSEE	7	0.026	0.007
27-23	ESSSESS	7	0.007	0.003
6-14-47	SSSSSE	6	0.048	0.004
7-11-46	SSSESS	6	0.042	0.001
7-6	SSSSES	6	0.020	0.002
6-13-65	SSESSS	6	0.012	0.012
3-23-3	SSSSS	5	0.000	0.000
6-16-92	SESESESSES	10	0.015	0.007
6-9-83	ESSSSSEESE	10	0.009	0.009
29-47	SSESEESSS	9	0.002	0.002
6-14-88	SSSSSEES	8	0.002	0.002
5-40	SSSESSS	7	0.060	0.030

5-37	SSESSSS	7	0.022	0.009
6-20-38	SESSSSS	7	0.012	0.008
6-16-96	SSSSESS	7	0.006	0.006
6-13-72	SSSESEESSESE	12	0.000	0.000
	N-E Combinations			
7-20-10	ENEEE	5	0.996	0.004
7-20-46	EEEEEE	6	0.989	0.011
7-20-33	NEE	3	0.986	0.014
7-19-28	NEEEE	5	0.971	0.001
7-19-23	EEENEEE	7	0.965	0.002
7-19-47	EEEEEN	6	0.964	0.003
7-20-32	EENEE	5	0.962	0.009
7-19-21	EENEEEE	7	0.947	0.004
7-19-8	ENNNEE	6	0.941	0.014
7-20-47	NEEE	4	0.901	0.016
7-20-20	EENEEEEEEE	10	0.879	0.015
7-20-11	ENE	3	0.846	0.094
7-20-31	NEENEEEN	8	0.749	0.040
7-19-17	EENNEN	6	0.715	0.015
7-20-26	EEENNN	6	0.700	0.053
7-19-36	ENNENE	6	0.679	0.072
7-20-37	NNNEEEEN	8	0.569	0.034
7-19-6	NENE	4	0.379	0.010
7-19-43	NEENNNENE	9	0.362	0.104
7-19-40	EEEENENNN	9	0.269	0.079
7-20-9	NNN	3	0.137	0.029
MA1	NNNNN	5	0.0	



```
TCAGCATTcctacggttgttaccgggactagtaa
tcagcatagtgcggtcaccgGATGAGcctagtaa
TCAGCATTatgacgagcgggatccgggctagtaa
tcagcatgtaggcgtctggtgggggggctagtaa
tcagcatATTCAGCCTAGttgggtggctagtaa
tCAGCATCGttataccgcgcctgggtgctagtaa
TCAGCATTCAGTGGAGGttgtggcactctagtaa
tcagcatgggccatcgtTGTGGAGAACCtagtaa
TCAGCATTGggctcaggccggccggtgctagtaa
tcagcatctcctcgtttaggggtaggctagtaa
tcagcatgtggggttccgatggggccgctagtaa
tcagcatatagcggattacgggcggcctagtaa
tcagcatggGGAGGAGTTCGTGCTGAgctagtaa
tcagcatgtcattaacggacacatggcctagtaa
tcagcatgtgaataTTGCGAtgtgagctagtaa
tCAGCATCGtgagtgattTCCACAACactagtaa
tcaGCATCTTCAAGATagaacgtggctctagtaa
tcagcatagacagcgtgggcgggagtgctagtaa
tcagcataGAGACATCGagggactaggctagtaa
tcAGCATCACcgcggtgccacctccacctagtaa
tcagcATGAGAgactgttttagtacacctagtaa
TCAGCATTGAGGACCAAAAGgGTGAAGCTagtaa
TCAGCATTagggcgagtagtgataatgctagtaa
TCAGCATTtggcatgcagGATATGcggctagtaa
tcagcatagtgcctcggtcaaacggggctagtaa
tcagcatacgatcggcatgtcttgtcgctagtaa
tcagcatgggGACGAAgcaatatgggcctagtaa
TCAGCATTcgcAGACCATCAAAtgcggctagtaa
tcagcatagatttgCAGATCGGTTGGActagtaa
tcagcatgaggGAAGTAGAAATGGCGcctagtaa
tcagcATCAAGcacAGTGACcgagaacctagtaa
tCAGCATCGatgtcCCGGAGGTtttgcctagta
tCAGCATCGCGGTTAGGAGGATGGAAActagtaa
tcagcatggcacggcGAGACACCATCactagtaa
tcagcatggcagcgggcgtacccggatctagtaa

tcagcatggcacggggaggcaccatcactagtaa
tcagcATGGTCgcaggtcaggtgggttctagtaa
tcagcatcCAGAGGgcGGAAACgttggctagtaa
tCAGCATCGtgcccacgtgtctcaggtctagtaa
tcagcatggcttggttcgcggtGACGACtagtaa
tCAGCATCGatgaccctcagacgtatactagtaa
tcagcatgacgtccagtacgctcgaggctagtaa
tCAGCATCGccgGACGACgtgtgttgctagtaa
TCAGCATTGAGtgcgcggatagactgactagtaa
tcagcatatgctccggaatcggaacggctagtaa
tcagcatgcggacccggAAAGGActaactagtaa
tcagcatgtgggttcggCGGAATCAagctagtaa
tcagcATGGAAGTACGggacgtgccggctagtaa
tCAGCATCGTCGcagggcaggtgggaactagtaa
tcagcatatcggacagggTCCAGCAGGctagtaa
tCAGCATCGTGAAACtgccCAGAGGtgctagtaa
TCAGCATTtcggacgggctagggatggctagtaa
tcagcatgtTGCGGAGACGAcccgagcctagtaa
tcagcatatcggccgatctgtgagttactagtaa
tcagcatctccagacgtcgtttgttgcctagtaa
TCAGCATTGAcAGCGGAAGgtacagtgctagtaa
TCAGCATTctaaggcgctAAGAACggcctagtaa
tcagcataacacggctgtgaGTGGTCCCtagtaa
tCAGCATCGacgtgtggggacggcAAGCTAgtaa
tcagcATCCAAtcGGATCAcctaacggctagtaa
tcagCATCAGggcacttgtttcactggctagtaa
tcagcatcctcACTGGActcagtggtgctagtaa
tcagcatgtgatacatacaggtggcgcctagtaa
tcagcatggtaagtACTACAgggtgtgctagtaa
tcagcATGAAAgttgtAAAGACaggggctagtaa
tcagcatACATGAaCACAACGTCggggctagtaa
tcagcATGGCGttttcGAGGATcgggactagtaa
TCAGCATTGGAtgtcagcgacgggccactagtaa
tcagcatacgGGCGGACTcctctggtactagtaa
TCAGCATTtACAACTgCACCACGGtcgctagtaa

tcagcatgcAGTGACtgcattggCAGCCTAGtaa
tcagcataGACCAGTAGccgctgccggctagtaa
tCAGCATCGAGGAAtataaaggtgggactagtaa
tcagcatatgggtctgacacgctgactctagtaa
tcagcatacgctcaatAGAAATCAAGCTAgtaa
tcagcataccagggtcgtccgtctgggctagtaa
tcagcatcttgagGTGAAGgtcatgtgctagtaa
tcagcatgtatttcgacaccagtgtgactagtaa
TCAGCATTGctcacccggccgccacagctagtaa
TCAGCATTGTGCAGCTtgcgtcacgtcctagtaa
tcagcatgaaccttgcaggtcgcgcgactagtaa
TCAGCATTagtaaccgcgacagtaggcctagtaa
tcagcatgacgttggtgttatccgccactagtaa
tCAGCATCGcgtcgagtcgtagggggcctagtaa
tcagcatatgcAGACGAtggtgcggctctagtaa
tcagcatgagttgagcgatggtgcgtactagtaa
tcagcatagcgagcGGAAAAcaggtaactagtaa
tCAGCATCGAGCCACGGACCacacggactagtaa
tcagcATGGATaacgGTGTGGCCcggcctagtaa
tcagcatggccggacgcattgcagagctagtaa
TCAGCATTccgATCTGTGCACggacgctagtaa
tcagcatagaccgtcAACATGtctgccctagtaa
tcagcATCAAACCTGCGTGGtatggtactagtaa
tcagcATGGATcgtaagtgcAGACGActagtaa
tcagcATCAAAccgtcaaagtacgtcactagtaa
tcagcatgaccgggaTTGAAGGAGCTCTagtaa
tcagcatcATGAAGccgtcaccaacgtctagtaa
tCAGCATCGcacactgcgtcccggggcctagtaa
tcagcatgaCGTCGCCCCGTGTgtAAGCTAgtaa
tCAGCATCGTGTCGcgtcctcgtgtgcctagtaa
tcAGCATCACcAGCGGAGTccccagagctagtaa
tCAGCATCGcgtgcgtgcagtgccAAGCTAgtaa
tCAGCATCGATTCAGgtacgtccaactctagtaa
tcagcatgatcgtatccggaacacgggctagtaa
```

Figure. S2.1. RESCUE-ESEs and PESEs in sequences found by SR protein responsive functional SELEX (Liu et al. 2000; Liu et al. 1998). The ~34-mers shown include a central 20 nt derived from a random insert, plus constant flanks of 7 nt each; the 20-mers represent the sequences underlying ESEfinder (Cartegni et al. 2003), kindly provided by Adrian Krainer. RESCUE-ESEs are in bold blue, PESEs are in bold red, overlaps between the two are in violet italics and the top ESEfinder sequences (largest increment over the default threshold) are underlined (thin = ASF/SF2, thick = SC35, dotted = SRp40, and dashed = SRp55).

Chapter 3

This chapter is part of a manuscript submitted for publication (Arias et al. 2013)

A Reductionist Approach to Splicing of Designer Exons: Single-Parameter Perturbations

Introduction

For many genes, transcription produces pre-mRNA molecules that include exons and introns; the introns are removed and the exons are spliced together. Machinery in the cell is able to identify the boundaries between exons and introns with extreme accuracy. Early studies showed that the sequences at these boundaries are fundamental contributors and consensus sequences were published (Mount 1982). However, it was later realized that these sequences by themselves are not enough since many sequences that resemble the consensus are ignored in the process of splicing while others that show less similarity are used (Sun and Chasin 2000).

Two alternative models have been implicit in thinking about the early recognition of sites (De Conti et al. 2013). In the first model, intron definition, each intron is recognized as a unit and removed; the exons are joined as a byproduct. In the second model, exon definition, each exon is recognized as a unit and joined to another similarly recognized exon. Hence, the ends of the intervening intron must be paired, requiring intron definition. The first model leads to studies of the recognition of the boundaries across the intron while the second leads to studies of the recognition of the boundaries across the exon. These considerations are particularly informative for internal exons for it has been suggested that exons flanked on both sides by introns longer than about 250 nt rely on exon definition for their inclusion while shorter introns can be removed solely by intron definition. More than 75% of human exons belong to the former category, stressing the importance of studying exon definition (Fox-Walsh et al. 2005).

A protocol to obtain nuclear extracts and the subsequent development of *in vitro* splicing have facilitated research into many aspects of splicing (Dignam et al. 1983; Krainer et al. 1984). However, this tool works well only with short introns and systems with only a single such intron are routinely used. Accordingly, systems with internal exons that are surrounded by long introns have usually been abbreviated by removing large chunks of the introns and frequently further abridged to comprise only two exons. Even with these restrictions, the rate of intron removal is lower *in vitro* than *in vivo* (Hicks et al. 2005) and is lower yet for long introns *in vitro* (Lazarev and Manley 2007). These limitations in splicing long introns are present even in current transcription-splicing coupled systems (Lazarev and Manley 2007). Therefore, modifications to the *in vitro* assay or the development of new tools to complement it are needed for the study of exon definition.

Several factors affect inclusion of an exon in the final mRNA molecule, including the strengths of the 5' and 3' splice sites and the presence of regulatory sequences both in the exon (exonic splicing enhancers, ESEs; and exonic splicing silencers, ESSs) and in the intron (intronic splicing enhancers, ISEs; and intronic splicing silencers, ISSs). More recently, the importance of transcription kinetics (Dujardin et al. 2013) and the influence of chromatin structure (Luco et al. 2011) have been recognized. Many of these factors have been targeted by systematic studies (Graveley et al. 1998; Luco et al. 2010; Shepard et al. 2011). Genomic studies have given us an exquisitely detailed picture of chromatin and the RNA make-up of cells under a variety of conditions; understanding how that picture is realized represents a fundamental goal and remains a challenge for which new approaches may be required (Roca et al. 2013).

We have chosen to explore these issues using a reductionist approach, attempting to segregate individual parameters governing splicing so as to identify fundamental biophysical principles and parameters involved. Toward this end we have created simplified exon sequences of our own design (“designer exons” or DEs). A key feature in the design of these exons was the ability to vary the parameters of exon length, ESE/ESS number and ESE/ESS position without otherwise changing the sequence characteristics of the exon. We found that the relationships linking these parameters to splicing display both simple and complex behaviors.

Results

DEs: effect of size

The exon definition model for splice site recognition (Berget 1995) maintains that internal exons will be chosen for inclusion only if they have acceptable splice sites at both ends, suggesting a physical interaction between the two ends of the exon. Thus the distance between the two ends of the exon could be an important parameter for the realization of this interaction. Consistent with this idea internal exon size in humans is limited, with less than 4% being greater than 300 nt (Berget 1995). In most previous experiments on the effect of exon size in splicing the experimental expansion of exons changed the quality as well as the length of the test exons, opening up the possibility that splicing was being affected by parameters other than length alone

(Chen and Chasin 1994; Sterner et al. 1996). Since DEs can be expanded by adding identical sequence modules, chosen to avoid exonic regulatory elements, the contribution of parameters other than length should be diminished.

To assess the effect of size on exon inclusion we constructed a series of 3-exon minigenes containing a DE as the central exon (Fig. 3.1 and Supplemental Fig. S3.1). The DE is composed of 6 nucleotides that are parts of the splice sites or necessary “linker” sequences with the remainder of the DE being comprised exclusively of repeats of the reference sequence CCAAACAA. This sequence, previously termed “neutral,” is predicted to be neither an exonic splicing enhancer (ESE) nor an exonic splicing silencer (ESS) and had relatively little effect on the splicing of a test exon (Zhang et al. 2009). More importantly this sequence has the property

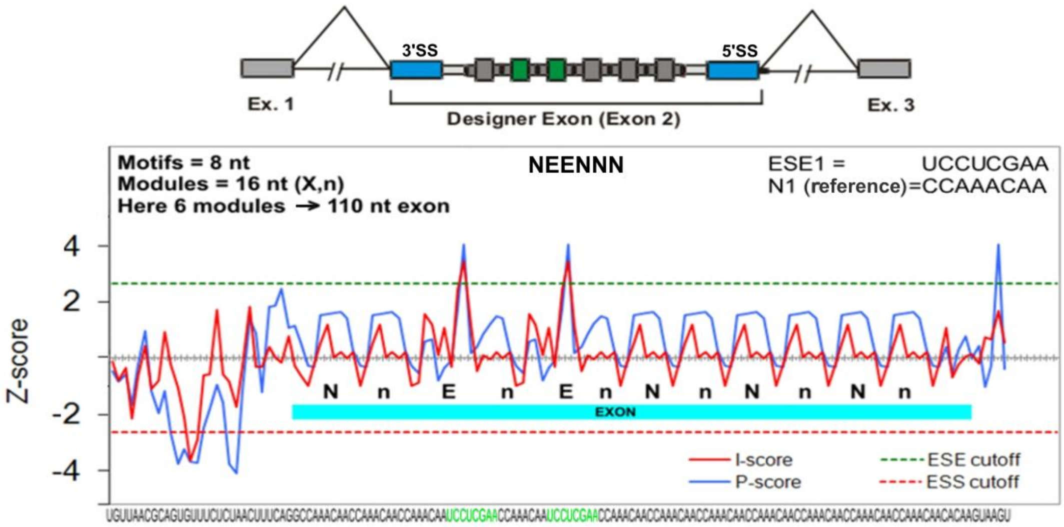


Figure 3.1. Construction of designer exons (example). From the bottom: The RNA sequence with two 8 nt ESE motifs in green. Above that is a plot of the computationally predicted enhancer/silencer strengths of each overlapping 8-mer using two different criteria: red or blue (Zhang and Chasin 2004). The dashed lines indicate cutoffs used for classifying a sequence as an ESE (green) or ESS (red). The exon is indicated by a blue bar, where E refers to the ESE motif and N or n refers to the reference motif, with the lower case indicating its use as a spacer. At the top of the panel is an abbreviated version of the motif composition in which the spacer n motifs have been omitted. Finally, at the top is a cartoon showing the overall structure of a minigene containing a DE, with the splice sites in blue and the ESE motifs in green.

of creating neither a predicted ESE nor a predicted ESS when all overlapping 8-mers created by self-concatenation are considered (Zhang et al. 2009). The exon sizes used here ranged from 14 nt to 302 nt. In preliminary experiments, we found that the level of exon inclusion was too low to be informative using our original DE framework. Changing the sequence at the 3'SS from the original wild type UCUCUAACUUUCAG/G to UCUCUUUUUUUCAG/G or the sequence at the 5'SS from the original wild type CAA/GUAAGU to CAG/GUAAGU resulted in useful levels of splicing. Splicing was assessed in HEK293 cells after transient transfection or site-specific permanent transfection followed by RT-QPCR; the same location was used in all permanent transfection experiments (see Methods).

As shown in Fig.3.2, the curves describing the inclusion of DEs display an optimum size range for exon inclusion (percent spliced in, or psi), with inclusion levels dropping off dramatically both above and below this range. The optimum range depends on the nature of the splice site sequences, being about 45 to 80 nt for exons with a strong 3'SS and 80 to 110 nt for those with a strong 5'SS. Interestingly, not just the optimum but the entire curve was shifted according to the splice site sequences used. Although exon inclusion efficiencies differ at many points depending on splice site sequences, the shapes of the curves are remarkably similar. A stronger 3'SS favors the inclusion of shorter exons whereas a stronger 5'SS favors the inclusion of longer exons. For example, compare the effects of the different splice site sequences on DEs of 46 nt and 142 nt in Fig. 3.2. To assess the extensibility of these results to a chromosomal context, we engineered a cell line where DEs could be placed by stable transfections exclusively at a defined location in the genome (see Supplemental Material); we call these exons

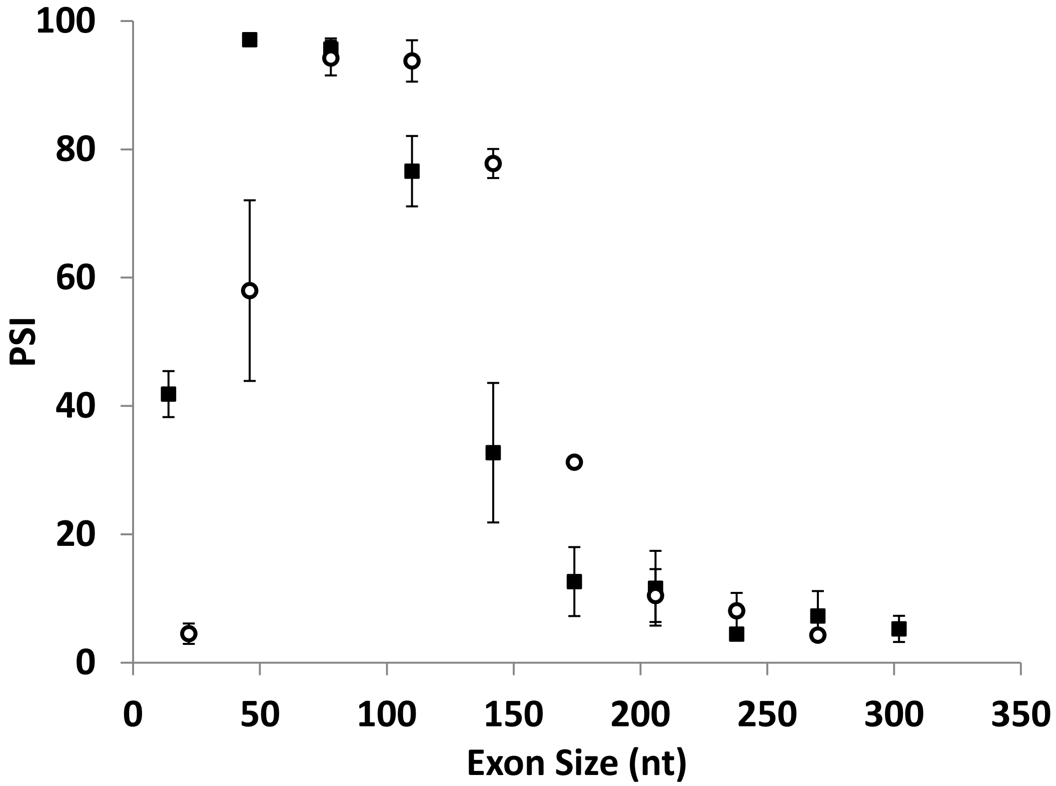


Figure 3.2. Exon inclusion has an optimum size range. Inclusion levels (psi) of DEs in transient transfections. DEs consist of reference sequences and have either a strong 3'SS (filled symbols) or a strong 5'SS (open symbols). See Supplemental Fig. S3.2 for inclusion levels of DEs in a chromosomal context. Error bars: SEM, n≥3.

chromosomal DEs. A series of exons bearing the strong 3'SS yielded a curve closely resembling that for transient transfections (Supplemental Fig. S3.2). The interdependence between the quality and position of a splice site sequence and size-dependent efficiency of exon inclusion is surprising and will be revisited in the Discussion.

DEs: effect of ESE position

Interactions that take place in exon definition have been shown to be facilitated by exonic splicing enhancers (Blencowe 2000; Chasin 2007). To assess the effect of enhancers on DE inclusion, we chose as the baseline a DE of 110 nt made up exclusively of reference sequence repeats and carrying wild type splice sites, SS Set 7 (see Table 3.1). This exon yielded a psi of about 7%, a suitably low value for observing the effect of enhancers. As a prototype ESE we used the sequence UCCUCGAA. Like the reference sequence this sequence has the property of creating neither a predicted ESS nor a predicted ESE within the overlaps created by its insertion into the baseline DE. Importantly, this ESE is the same prototype we used in our initial study of designer exons (Zhang et al. 2009). This ESE was always inserted into the baseline DE such that

Table 3.1. Effect of splice site sequences on exon inclusion.

SS Set no.			Consensus value[1]		Consensus value difference[2]	psi with no ESEs	psi range	Max p-value	Min p-value among all 15 single ESE pairwise comparisons
	3’ SS	5’SS	3’ SS	5’SS			1 ESE, all positions	single vs. no ESE	
1	UCUCUUUUUUUCAG/G	CAG/GUAAGU	93.1	99.9	-6.8	95	ND[3]	–	–
2	UCUCUAACUUUCAG/G	CAG/GUAAGU	81	99.9	-18.9	94	ND	–	–
3	UCUCUUUUUUUCAG/G	CAA/GUAAGU	93.1	88.4	4.7	77	ND	–	–
4	UCUCUUUUUUUCAG/G	CAA/GUGAGU	93.1	83.4	9.7	26	59 – 72	0.004*[4]	0.10
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	87.4	88.4	-1	49	86 – 90	0.0001*	0.27
6	UCUCUAAUUUUCAG/G	CAA/GUAAGU	82.6	88.4	-5.8	11	25 – 37	0.05	0.41
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	81	88.4	-7.4	7	18 – 34	0.001*	0.16

[1] Based on a modification of the method presented by Shapiro and Senapathy (Shapiro and Senapathy 1987; Zhang et al. 2005)

[2] Defined as the difference between the 3’SS and 5’SS consensus values.

[3] Not done.

[4] Asterisks mark statistically significant differences ($p < 0.05$)

it was flanked by two reference sequences, so as to satisfy the conditions for which it was designed and to keep constant the local context defined by the flanking 8 nucleotides. Thus we can consider the resulting DEs as being comprised of 16 nt modules each consisting of the reference 8-mer followed by either the ESE or another reference 8-mer. The baseline DE provides 6 evenly spaced non-overlapping positions at which a 16 nt ESE module can be substituted. Later we will describe similar constructs bearing ESS sequences. To define the composition of a DE we will use the notation E for ESE, S for ESS and N for the reference sequence. So, for instance, we refer to the placement of an ESE at the second of the 6 available positions as NENNNN.

We substituted a sole ESE at each of the 6 evenly spaced positions in the baseline DE and measured splicing after transient transfection. At each position the presence of the ESE caused a 3- to 4-fold increase ($p<0.01$) with respect to the baseline DE (Fig. 3.3A). Similar increases were seen for these exons in the chromosomal context (Supplemental Fig. S3.3A). The ESE was effective at each of the 6 positions and there was no statistically significant difference between positions, except for 3 modest differences in permanent transfections (maximum psi difference of 5%).

Enhancers are often thought of as acting by enhancing the recruitment of components of the splicing machinery to a nearby splice site, and as such would be expected to show a position effect, being more important close to a weak splice site. The lack of a position effect here could be due to the incorrectness of this argument or to the possibility that the ESE is equally effective at enhancing the use of the 3’SS and the 5’SS, and so only appears to be position independent.

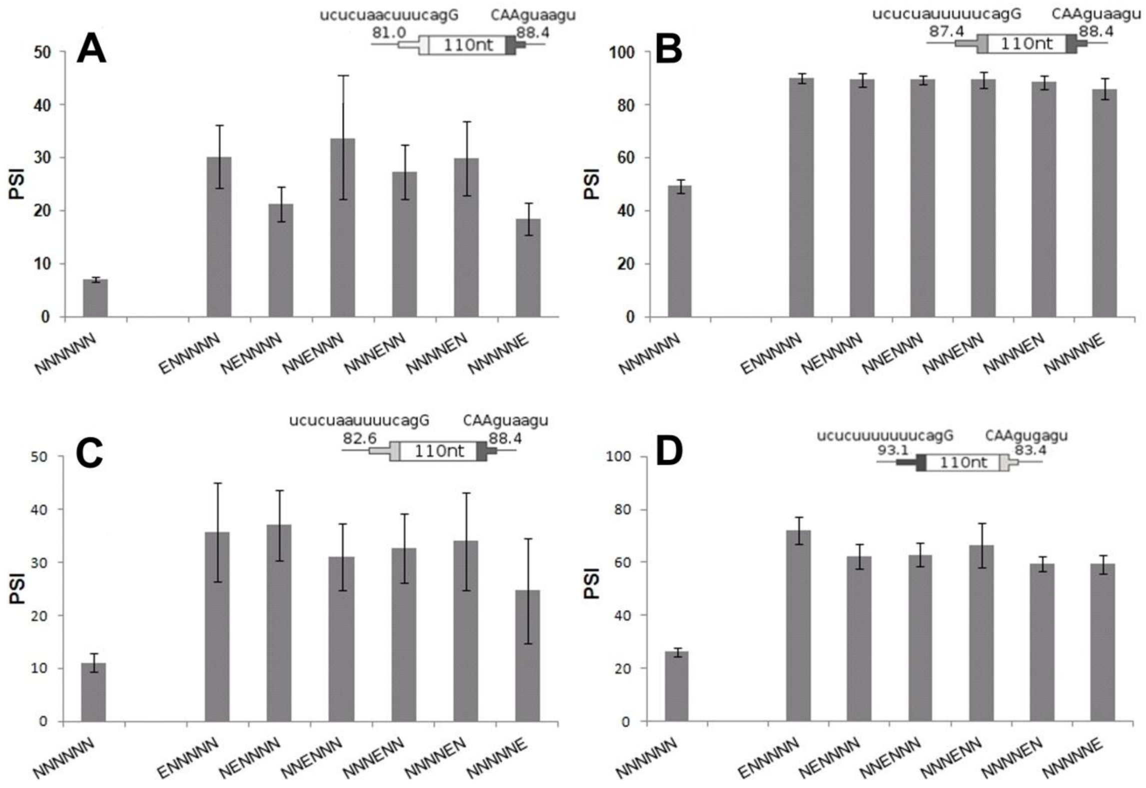


Figure 3.3. Addition of a single ESE enhances inclusion level and is position independent. The cartoons show the consensus values for splice site strengths used. A. Position variation in DEs with SS Set 7. B. Position variation in DEs with SS Set 5. C. Position variation in DEs with SS Set 6. D. Position variation in DEs with SS Set 4. Error bars: SEM, n≥3 except panel C, where n=2. In all cases the psi of DEs with an ESE are significantly different from that without ESEs (t-test, p<0.01), except for the last position in panel C (p=0.05). None of the 90 pairwise comparisons between ESEs at different positions showed significant differences (t-test, p>0.05). See also Table 3.1. See Supplemental Fig. S3.3 for inclusion levels of DEs in a chromosomal context.

To distinguish between these two ideas, we manipulated one or the other of the DE splice sites so as to create a range of differences between the two in terms of strength. This last is taken to be the agreement to the consensus, expressed as the consensus value (Zhang et al. 2005).

We tested 7 combinations of three 5’ and four 3’ SS sequences using transient transfection. We started with a DE with two relatively strong splice sites, having consensus values of 93.1 and 88.4 for the 3’SS and 5’SS respectively. This exon was efficiently included even without an ESE (psi of ~80%) and so was not useful for evaluating enhancement (Table 3.1, SS Set 3). We then weakened one or the other of the splice sites so as to produce a range of disparities between the 3’SS and the 5’SS strengths; the differences in strength (3’SS minus 5’SS) for the 4 tested pairs were +10, -1, -6 and -7. Weakening either the 3’SS or 5’SS reduced the psi to values from 7% to 49% so that the effect of adding an ESE could be evaluated (Table 3.1, SS Sets 4 through 7). Once again no statistically significant difference was found for the effect of the ESE at the various positions: p-values were greater than 0.05 for all pairwise comparisons (Fig. 3.3B, 3.3C and 3.3D; and Table 3.1, last column). SS Set 5 was also tested in a chromosomal context; addition of a single ESE produced statistically significant increases of similar magnitude to those found using transient transfections and was once again position independent (Supplemental Fig. S3.3B). These observations provide no support for the existence of a position effect for the enhancement by the ESE in these DEs and so likewise provide no support for the recruitment model for ESE action.

DEs with single ESEs have inclusion levels that are much higher than those of their corresponding baseline DEs. Moreover, the absolute increments in psi produced by adding a single ESE were much higher in the series shown in Fig. 3.3B and 3.3D (~40%) than that observed when the original splice sites were used (~20%, Fig. 3.3A). That is, the magnitude of the effect produced by an ESE depended on the splice site combination present.

DEs: effect of multiple ESEs

The sequence of our DEs allowed us to add an ESE while diminishing the chance of creating other regulatory sequences within overlapping sequences. It also allowed us to add multiple copies of an ESE while not adding any sequences that were not already present in a single ESE DE. It has been shown that the ESE strength or number inversely correlates with splice site strength in mammalian exons, i.e., ESEs can compensate for weak splice sites (Xiao et al. 2007; Ke et al. 2008). In addition, Hertel and Maniatis (1998) showed that the use of multiple downstream enhancer elements increased the use of a 3'SS in an additive manner when tested *in vitro* (Hertel and Maniatis 1998). We asked whether such additivity also holds true for the definition of an internal exon *in vivo*.

To assess the effect of multiple enhancers in a single exon, splicing of DEs with 0, 1, 2, 3 or 6 ESEs was measured using transient transfections. For these experiments we used SS Set 7 in Table 3.1, which was the same set used in our previous study of randomly constructed DEs (Zhang et al. 2009). The data for no ESEs and 1 ESE at all possible positions was shown in Fig. 3.3A. The analogous data for all 36 combinations of positions for 2, 3 and 6 ESEs are shown in Fig. 3.4 and Supplemental Table S3.1. As was the case for 1 ESE, there was no strong or consistent position effect when 2 or 3 ESEs were present. Psi values increased with the number of ESEs in a linear manner up to 3 ESEs ($R^2 = 0.82$) and leveled off when 6 ESEs were included. Ascribing the last point to saturation, these results are consistent with the additive model. The slope in the linear range was a moderate 20% per ESE added; this kind of limited enhancement enabled testing the effect of multiple ESEs.

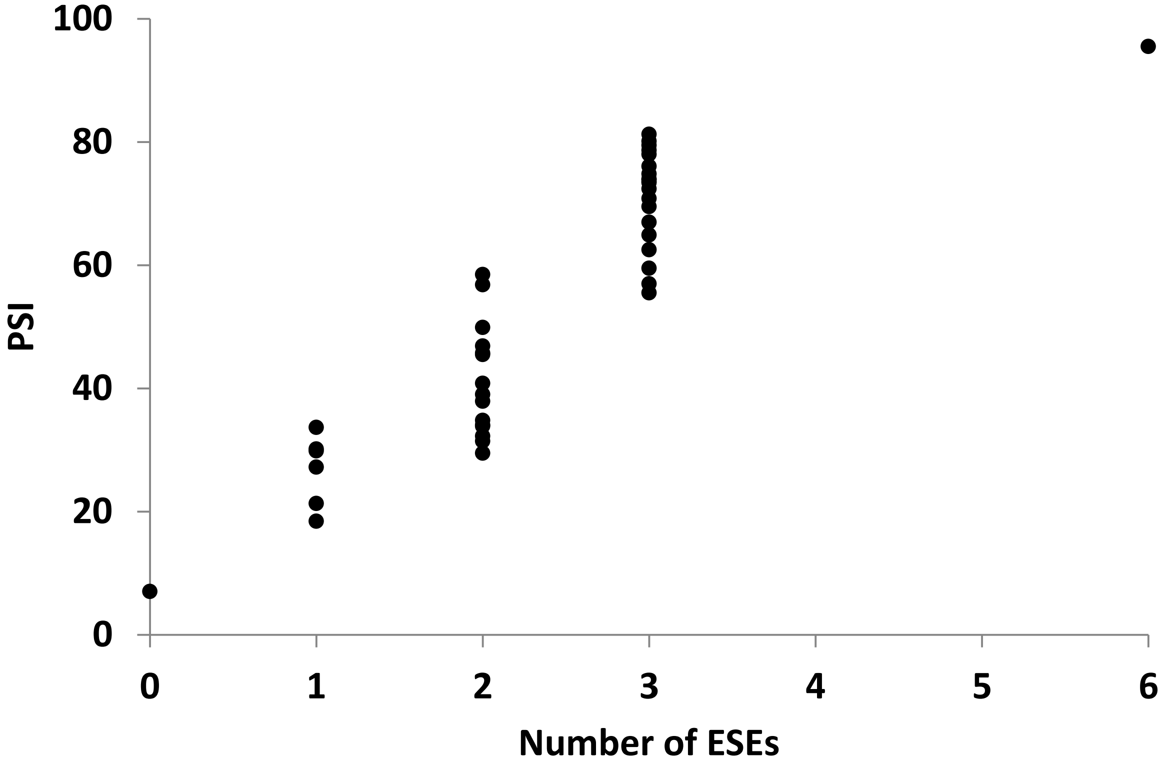


Figure 3.4. Inclusion levels of DEs increase with the number of ESEs present. The psi for all possible DE permutations with 0, 1, 2, 3, or 6 ESEs was measured (n≥3).

DEs: effect of ESS position

A similar analysis for the effect of an exon silencer sequence (ESS) was performed. In this case SS Set 5 (Table 3.1) was used in order to provide a psi of about 50% as a baseline value from which decreases could be measured. The ESS sequence, CACAUGGU, was carefully chosen so as to not create any other predicted splicing regulatory sequence when placed in the DE; this same ESS was used in our previous study (Zhang et al. 2009). Placement of a single ESS at positions 2, 3, 5, or 6 reduced the psi significantly (p values from 0.003 to 0.031, Fig. 3.5). There was no effect at position 1; variability at position 4 did not allow a conclusion.

Repeating this experiment in a chromosomal context yielded similar results (Supplemental Fig. S3.4). These results indicate some differences between positions, a conclusion that is supported by considering the effects of multiple ESSs (see below).

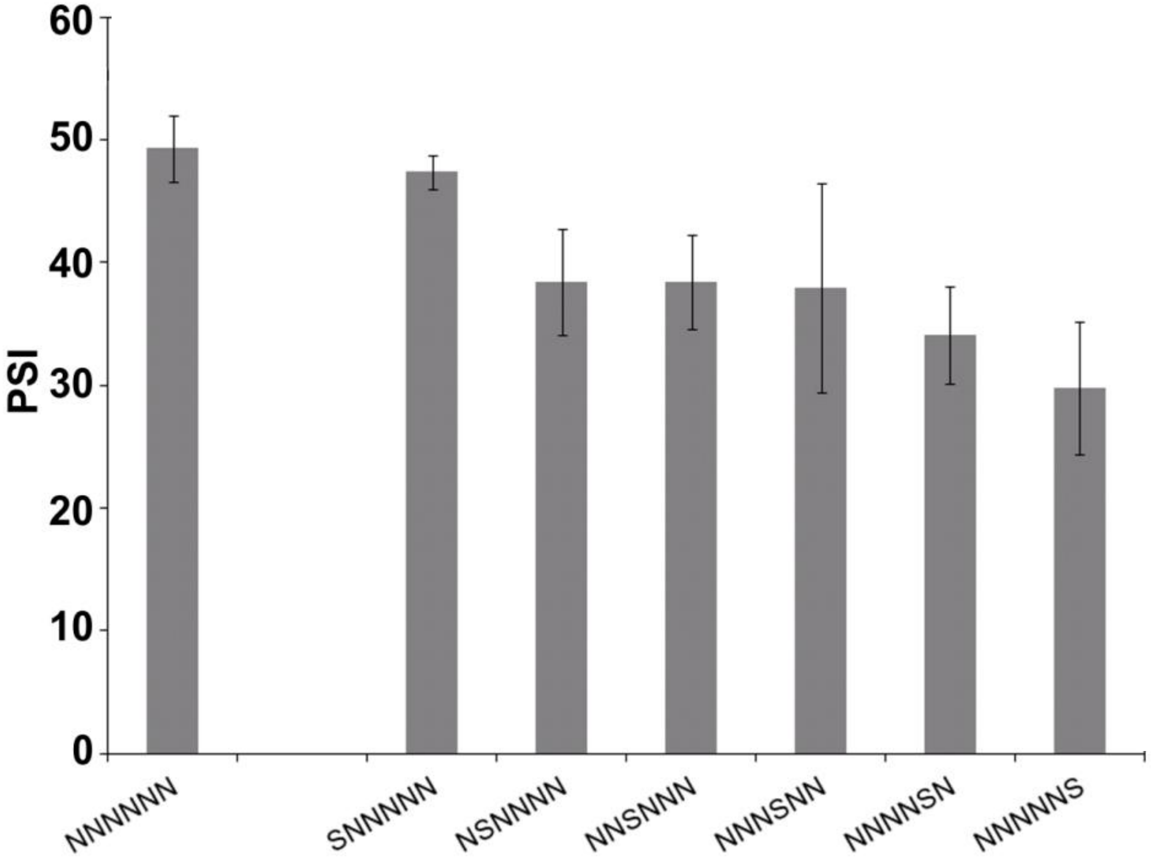


Figure 3.5. Addition of a single ESS decreases inclusion level and shows some position dependence. The psi for DEs with a single ESS are shown for transient transfections. Error bars: SEM, n≥3. See Supplemental Fig. S3.4 for inclusion levels of DEs in a chromosomal context.

DEs: effect of multiple ESSs

We next measured the effect of multiple ESSs, once again with the question of additivity in mind. The results of including 2, 3 or 6 ESSs in all 36 possible combinations of positions are summarized in Fig. 3.6A and shown in detail in Supplemental Table S3.1, which includes psi values for each positional combination. Psi decreased with ESS number in a reasonably linear manner up to 3 ESSs (R^2=0.68); 6 ESSs resulted in 10-fold silencing, but this single point shows signs of saturation (Fig. 3.6A). These results are consistent with an additive model in which each ESS contributes about a 12% drop in psi.

The simple relationship between ESS number and psi described above does not take into account the lack of an effect for an ESS at position 1 (seen in Fig. 3.5). To address this issue we performed a different test of additivity, one that allows each ESS to exert a characteristic position effect. The psi of a multi-ESS DE was predicted by summing the effects of the individually positioned ESSs as measured in the single-ESS DE experiment:

3.1. Predicted psi=baseline + $\sum_{i=1}^{6} P_i \cdot (\mathrm{psi}(i) - \mathrm{baseline})$

where baseline is the psi of the baseline DE with no ESSs, *i* is an index number for positions 1 through 6, P*i* is 1 if an ESS is present at position *i* and 0 otherwise, and psi(*i*) is the measured psi for a DE bearing a single ESS at position *i*. The observed psi measurements for all thirty-five 2- and 3-ESS DEs show a good agreement to these linear combination predictions (R^2= 0.80, Fig. 3.6B). In contrast, when we assumed that all positions were equivalent and used the average value for all the single-ESS DEs to predict psi then the R^2 value dropped to 0.56, supporting the position dependence observed in Fig. 3.5. To explore this further, we examined the contributions of individual positions to this position effect by averaging all positions but one while retaining the position-specific contribution of the latter. Retaining the position-specific contribution of the first or last positions increased the R^2 value from 0.56 to 0.68 or 0.72, respectively, while such retention at the internal positions 2 to 5 produced no increase in R^2. Thus it appears that

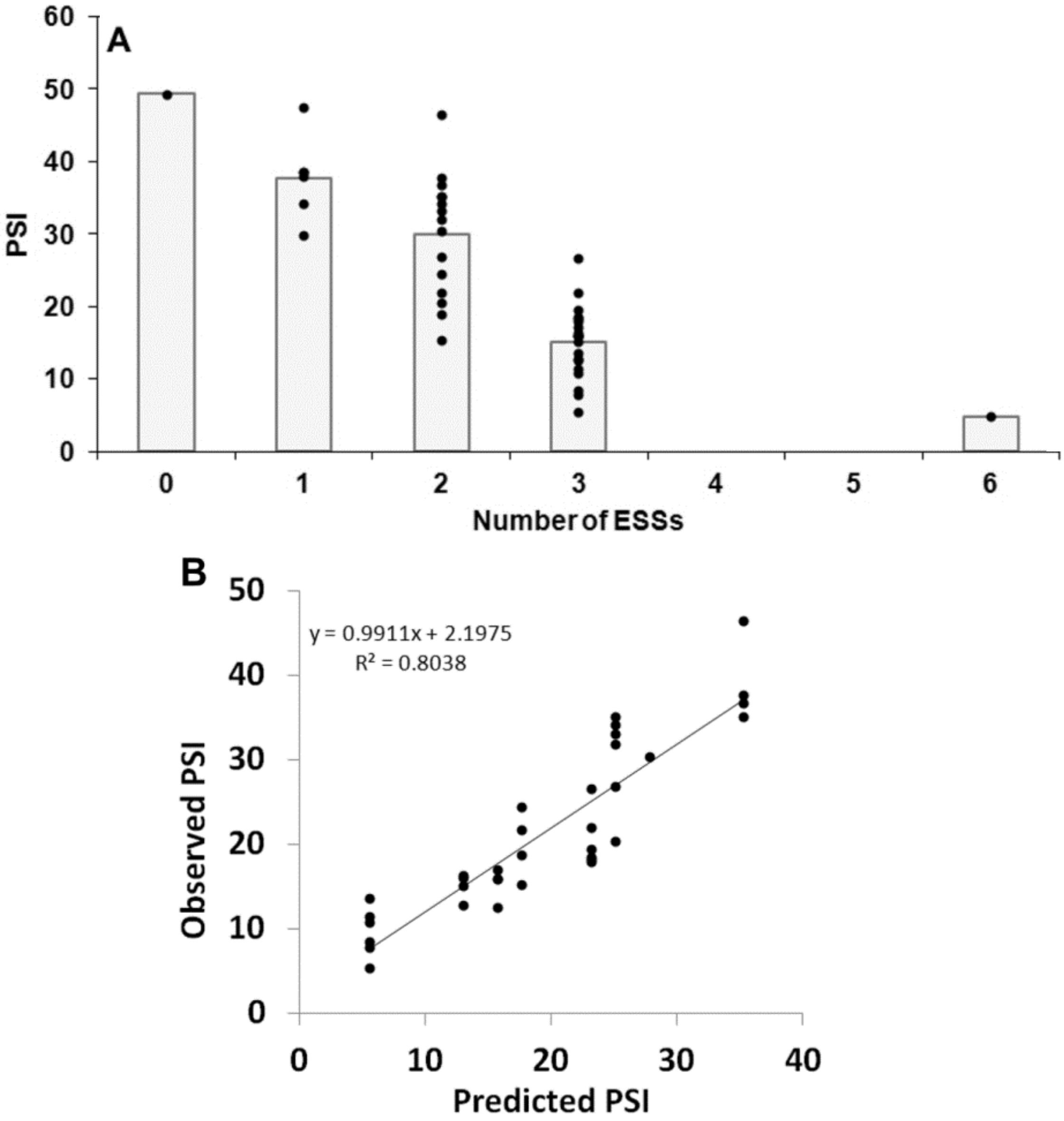


Figure 3.6. Inclusion levels of DEs decrease with the number of ESSs present. A. The psi for all possible permutations with 0, 1, 2, 3, or 6 ESSs were measured (n≥3). The columns depict the average. B. The psi for all possible permutations with 2 and 3 ESSs were plotted against predictions based on the addition of the individual position effects of each ESS as measured in the single ESS experiments (Fig. 3.5).

positional information is important only for the 2 terminal positions. Indeed, retention of the position effect of 1 and 6 alone returned the R^2 value to 0.80, the same as the value reached using all positional information. Taking all this data into account, it appears that an ESS at the first position has no effect, an ESS at the last position is the most effective and ESSs in the middle positions have equivalent intermediate effects that are independent of their positions.

Discussion

We have described the splicing phenotypes of exons of our own design, each principally comprised of just 1 or 2 prototype 8-base sequence modules that represent an ESE, an ESS or a reference sequence that resembles neither. The splicing effects of exon size, ESE content and ESS content were independently and systematically evaluated in the context of a three-exon minigene. We found that there is a major effect of size on splicing. Both small and large exons are spliced less efficiently than exons of intermediate size. Surprisingly, when we used different splice site sequences, we found a striking difference in exon size dependence. One set showed a better efficiency for long exons while the other was better for short exons; that is, one dependency was shifted relative to the other. Using a DE of a fixed size, the ESE sequence used increased psi in all positions tested. Interestingly, the magnitude of this effect was not position dependent, even when the 3’SS or 5’SS was purposely weakened. Moreover, when multiple ESEs were used the effect increased proportionately before showing saturation as psi approached 100%. The ESS sequence, on the other hand, displayed some position dependence. Its effect was maximal when placed close to the 5’SS but showed almost no effect near the 3’SS. Intermediate positions showed a uniform intermediate effect. When multiple ESSs were present their combined effects increased proportionately with signs of saturation as the psi approached 0%. Thus neither the ESEs nor the ESSs used here showed signs of cooperative behavior.

Irrespective of the model used, the shift between these two curves implies that comparing the strengths of 5’SS sequences might be more complex than previously thought. Finally, compared to these simplified exons, natural exons may be influenced by other factors. For instance, the collision rate between the exon ends (see Chapter 5) could be increased (or decreased) by additional protein-protein interactions, by the formation of secondary structures or by a scaffolding platform. In this respect, DEs can provide a framework for investigating such possible influences.

Effect of ESEs

The idea that ESEs act by recruitment of the splicing machinery (Kohtz et al. 1994; Staknis and Reed 1994) is supported by evidence of interactions between activator proteins that bind ESEs and some of the proteins involved in the early steps of splicing (Hoffman and Grabowski 1992; Wu and Maniatis 1993; Kohtz et al. 1994; Staknis and Reed 1994). It has often been assumed therefore that this interaction should display a position effect: the closer the binding site for the activator to the splice site, the more efficient it should be in recruiting the splicing machinery to that site. Since we saw no evidence for such a position effect we propose that the action of the ESE is stabilization of an otherwise volatile interaction between U2AF and U1 snRNP (see Chapter 5). Position independence is also suggested by the finding that SRSF1 can contact a branchpoint sequence across a distance of 50 nt (Shen et al. 2004).

ESS number and position effect

We showed that the effect of multiple ESSs could be predicted by their linear combination as long as the particular characteristics of positions 1 and 6 were taken into account (Fig. 3.6B). The position effect seen for ESSs suggests that ESSs may act by destabilizing bound U1 snRNP or even blocking its binding rather than by affecting an exon definition complex. However, other possibilities remain. One of such possibilities includes a destabilizing effect of the proteins binding ESS on proteins binding the reference sequence (see Chapter 6). If the displaced proteins have a positive effect on splicing, the effect of ESS sequences would appear to be negative. Further studies using different ESS/SS combinations and exploring mechanisms such as competition for binding RNA could be tried using the present system as a starting point.

Materials and Methods

Plasmids

A “drafting” plasmid, pAL-SB, containing type IIS restriction sites flanking CCAAACA allowed BsmBI and BsaI (NEB) to generate the appropriate overhangs for seamless building-block additions on either the upstream or downstream side, respectively, of the DE under construction. Building blocks composed of combinations of two synthetic modules: NN, NE, EN, EE, NS, SN and SS were added by sequential ligations. The finished DEs were amplified by PCR and digested with BbvI (NEB) to generate overhangs compatible with appropriate receiving plasmids. Each receiving plasmid contained a modified dhfr minigene controlled by a tet-responsive promoter and a SV40 polyA site and with the first start codon being a Kozak sequence placed in exon 3. Each receiving plasmid had a specific SS set and, in place of a DE, a specifically designed removable sequence/adapter: RA. Using BveI (Fermentas), this RA was removed generating appropriate overhangs for seamless incorporation of the DEs constructed in the pAL-SB plasmid. This scheme and some variations were used to generate the DE-containing minigenes (see Supplemental Material).

The plasmid employed in the generation of the cell line used for chromosomal incorporations, pMA-FW, contains a kanamycin resistance gene for initial selection, a promoterless puromycin resistance gene for subsequent selection of site-specific recombinations with DE-containing plasmids, a phiC31 attP site and only the downstream half of the modified dhfr minigene. pMA-IC contains a CMV promoter to drive the puromycin resistance gene after site-specific recombination, the upstream half of the modified dhfr minigene with a replaceable exon 2 for reconstitution of the full minigene, and an attB site (Supplemental Fig. S3.5; see Supplemental Material). DEs were transferred to this plasmid by replacing the initial exon 2.

An actin-skipped coupled-standard was generated by incorporating cDNA for both gamma actin and DE-skipped mRNA in the same plasmid. Purified plasmid was digested with EcoO109I (NEB) to generate a solution with equimolar amounts of each type of molecule. This solution provided a standard for relative quantification through QPCR. Included-skipped equimolar coupled-standards were analogously constructed and used to calibrate the psi measurements (Supplemental Fig. S3.6; see Supplemental Material for details).

Psi measurement

RNA was extracted from transfected cells and reverse transcribed. Serial dilutions of the equimolar coupled-standard were used for QPCR quantification and the ratio of DE-skipped to DE-included was obtained (S/I, or SOI). This ratio was used to obtain psi by the formula psi=100/(1+SOI). A similar protocol was followed for stable transfections including gamma actin quantification (see Supplemental Material).

Transfections

Transient transfections were performed in modified HEK293 cells carrying a tTA gene (cMA-HEK293-tTA). RNA was extracted after 25 hours. Stable transfections were performed in cMA-FW cells using a DE-containing pMA-IC plasmid and the plasmid coding for the site-specific recombinase pPGKPhiC31obpA (Addgene). After puromycin selection, the resulting site-specific recombinants were pooled and grown for RNA extraction (see Supplemental Material).

Cell lines

HEK293 cells were stably transfected with a plasmid coding for the tet-Off trans-activator (Gossen and Bujard 1992). A clone, cMA-HEK293-tTA, was chosen and used for all transient transfections. cMA-HEK293-tTA cells were electroporated using linearized pMA-FW plasmid. Clone cMA-FW was selected as one that had incorporated a single genomic copy of pMA-FW, had a high level of expression and had an adequate level of recombination. This clone was used for all site-specific recombinations (see Supplemental Material).

Authors’ Contributions

MA and LC conceived the study and designed the experiments. AL designed and made plasmid pAL-SB. MA designed the protocols, carried out the experiments and performed the analysis. LC and MA wrote the text.

Acknowledgments

We thank Vincent Anquetil, Jonathan Cacciatore and Shengdong Ke for valuable discussions and Dennis Weiss for the same plus a thoughtful critique of the manuscript. This work was supported by a grant from the NIH (GM072740) to L.A.C.

References

Arias MA, Lubkin AL, Chasin LA. 2013. Splicing of designer exons informs a biophysical model for exon definition. *Manuscript submitted*.

Berget SM. 1995. Exon recognition in vertebrate splicing. *J Biol Chem* **270**(6): 2411-2414.

Blencowe BJ. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci* **25**(3): 106-110.

Chasin LA. 2007. Searching for splicing motifs. *Adv Exp Med Biol* **623**: 85-106.

Chen IT, Chasin LA. 1994. Large exon size does not limit splicing in vivo. *Mol Cell Biol* **14**(3): 2140-2146.

De Conti L, Baralle M, Buratti E. 2013. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA* **4**(1): 49-60.

Dignam JD, Lebovitz RM, Roeder RG. 1983. Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res* **11**(5): 1475-1489.

Dujardin G, Lafaille C, Petrillo E, Buggiano V, Gomez Acuna LI, Fiszbein A, Godoy Herz MA, Nieto Moreno N, Munoz MJ, Allo M et al. 2013. Transcriptional elongation and alternative splicing. *Biochim Biophys Acta* **1829**(1): 134-140.

Fox-Walsh KL, Dou Y, Lam BJ, Hung SP, Baldi PF, Hertel KJ. 2005. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proceedings of the National Academy of Sciences of the United States of America* **102**(45): 16176-16181.

Gossen M, Bujard H. 1992. Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proceedings of the National Academy of Sciences of the United States of America* **89**(12): 5547-5551.

Graveley BR, Hertel KJ, Maniatis T. 1998. A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. *EMBO J* **17**(22): 6747-6756.

Groth AC, Olivares EC, Thyagarajan B, Calos MP. 2000. A phage integrase directs efficient site-specific integration in human cells. *Proceedings of the National Academy of Sciences of the United States of America* **97**(11): 5995-6000.

Hertel KJ, Maniatis T. 1998. The function of multisite splicing enhancers. *Mol Cell* **1**(3): 449-455.

Hicks MJ, Lam BJ, Hertel KJ. 2005. Analyzing mechanisms of alternative pre-mRNA splicing using in vitro splicing assays. *Methods* **37**(4): 306-313.

Hoffman BE, Grabowski PJ. 1992. U1 snRNP targets an essential splicing factor, U2AF65, to the 3' splice site by a network of interactions spanning the exon. *Genes Dev* **6**(12B): 2554-2568.

Ke S, Zhang XH, Chasin LA. 2008. Positive selection acting on splicing motifs reflects compensatory evolution. *Genome Res* **18**(4): 533-543.

Kohtz JD, Jamison SF, Will CL, Zuo P, Luhrmann R, Garcia-Blanco MA, Manley JL. 1994. Protein-protein interactions and 5'-splice-site recognition in mammalian mRNA precursors. *Nature* **368**(6467): 119-124.

Krainer AR, Maniatis T, Ruskin B, Green MR. 1984. Normal and mutant human beta-globin pre-mRNAs are faithfully and efficiently spliced in vitro. *Cell* **36**(4): 993-1005.

Lazarev D, Manley JL. 2007. Concurrent splicing and transcription are not sufficient to enhance splicing efficiency. *RNA* **13**(9): 1546-1557.

Luco RF, Allo M, Schor IE, Kornblihtt AR, Misteli T. 2011. Epigenetics in alternative pre-mRNA splicing. *Cell* **144**(1): 16-26.

Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. 2010. Regulation of alternative splicing by histone modifications. *Science* **327**(5968): 996-1000.

Mount SM. 1982. A catalogue of splice junction sequences. *Nucleic Acids Res* **10**(2): 459-472.

Roca X, Krainer AR, Eperon IC. 2013. Pick one, but be quick: 5' splice sites and the problems of too many choices. *Genes Dev* **27**(2): 129-144.

Shapiro MB, Senapathy P. 1987. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res* **15**(17): 7155-7174.

Shen H, Kan JL, Green MR. 2004. Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. *Mol Cell* **13**(3): 367-376.

Shepard PJ, Choi EA, Busch A, Hertel KJ. 2011. Efficient internal exon recognition depends on near equal contributions from the 3' and 5' splice sites. *Nucleic Acids Res* **39**(20): 8928-8937.

Staknis D, Reed R. 1994. SR proteins promote the first specific recognition of Pre-mRNA and are present together with the U1 small nuclear ribonucleoprotein particle in a general splicing enhancer complex. *Mol Cell Biol* **14**(11): 7670-7682.

Sterner DA, Carlo T, Berget SM. 1996. Architectural limits on split genes. *Proceedings of the National Academy of Sciences of the United States of America* **93**(26): 15081-15085.

Sun H, Chasin LA. 2000. Multiple splicing defects in an intronic false exon. *Mol Cell Biol* **20**(17): 6414-6425.

Wu JY, Maniatis T. 1993. Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell* **75**(6): 1061-1070.

Xiao X, Wang Z, Jang M, Burge CB. 2007. Coevolutionary networks of splicing cis-regulatory elements. *Proceedings of the National Academy of Sciences of the United States of America* **104**(47): 18583-18588.

Zhang XH, Arias MA, Ke S, Chasin LA. 2009. Splicing of designer exons reveals unexpected complexity in pre-mRNA splicing. *RNA* **15**(3): 367-376.

Zhang XH, Chasin LA. 2004. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev* **18**(11): 1241-1250.

Zhang XH, Leslie CS, Chasin LA. 2005. Computational searches for splicing signals. *Methods* **37**(4): 292-305.

Supplemental Material

Materials and methods

Double stranded oligomers

Sense and antisense oligomers were purchased from either Invitrogen or Fisher Scientific and annealed by mixing them together at a concentration of 40 uM in 300 mM sodium acetate. These mixtures were placed in a bath of boiling water for 5 min and allowed to slowly cool down to room temperature. The annealed oligomers were phosphorylated at a final concentration of 100 nM with T4 polynucleotide kinase from New England Biolabs (NEB) by following the manufacturer's protocol. We call these molecules phosphorylated double stranded oligomers or P-ds-oligos.

Removable Adapters

Removable adapters or RAs are sequences that contain recognition sites for type IIS restriction enzymes (REs) that cut at both ends of the adapter. Due to the nature of type IIS enzymes, the sequence of the overhangs generated can be chosen essentially without restrictions. Two kinds of removable adapters were designed. RAs of the first kind (RA-I) are removed by a single type IIS RE that cuts on both sides of the adapter. RAs of the second kind (RA-II) allow independently controlled cuts on either end: one type IIS RE cuts on one side while a different type IIS RE cuts on the other side.

Plasmids

Fig. S3.1 shows the features of the modified dhfr minigene used to harbor the DEs.

All modifications performed on plasmids were verified by sequencing the appropriate regions (Genewiz).

A “drafting” plasmid, pAL-SB, was derived from pEGFP-C3 (Addgene) to facilitate the construction of DEs. This plasmid contains an adapter that allows the use of type IIS enzymes BsmBI and BsaI to add building blocks at either flank of the DE in progress, but it does not contain a dhfr minigene. The finished DEs can be copied and pasted into any of the receiving plasmids (see below). In order to provide flexibility for future extensions, BfuAI sites were removed from pEGFP-C3. For this purpose, nested PCR was performed using two primer pairs: oligo36 and oligo37, and oligo38 and oligo39; the oligo36 and oligo39 primers were used for the final amplification, which appended temporary BsaI sites at both ends to generate the appropriate overhangs. The products were cut with BsaI and ligated into pEGFP-C3 which was previously digested with BfuAI. This was followed by transformation of DH5-alpha competent cells and selection in kanamycin (Sigma-Aldrich). Successful clones were selected by evaluating digestion patterns with BfuAI. The intended use of BsaI for DE construction required removal of the BsaI site from pEGFP-C3. To remove it, a PCR fragment was obtained using primers oligo40 and oligo41; this PCR fragment and the previously modified pEGFP-C3 plasmid were digested with BsaI and EcoO109I, mixed and ligated together. After these preparations and in order to add the appropriate adapter, the oligo42 primer was designed. Along with oligo43, it was used to amplify a fragment from pEGFP-C3. Both the adapter-containing PCR fragment and the plasmid were digested with PstI (NEB) and HindIII (NEB), mixed and ligated together to obtain pAL-SB.

As a starting point for all the dhfr minigene containing plasmids, pMA-Universal was made from plasmid pUHD10-3 (Gossen and Bujard 1992). The whole dhfr minigene was copied from the pD12 plasmid (Zhang et al. 2009) and integrated into pUHD10-3 by placing it under the control of the tet-responsive promoter with a SV40 polyA signal for cleavage and polyadenylation. During the transfer, all the ATGs in exon 1 were eliminated, the first out-of-frame ATG in exon 3 was eliminated, and the following in-frame ATG was modified to conform to the Kozak sequence. These modifications were performed to reduce possible translation effects of modifying the middle exon. Additionally, the DE was substituted with an RA-II. The RA-II employed relies on BfuAI and BtgZI for its function. Therefore, the BtgZI site present in pUHD10-3 was removed. We call the dhfr minigene in pMA-Universal the modified dhfr minigene; its sequence is included below. The details for its generation follow. For the removal of the BtgZI site, the plasmid was cut with BtgZI (NEB) and NgoMIV (NEB) and dephosphorylated; P-ds-oligo oligo1/oligo2 was ligated to this plasmid using T4 ligase (NEB) by following the manufacturer’s protocol. The dhfr minigene was transferred from the pD12 plasmid (Zhang et al. 2009) and simultaneously modified in five stages using PCR and P-ds-oligo ligations as described below. An intermediate plasmid, piMA-F5, was obtained by PCR amplification of fragment F5 (oligo3 and oligo4), digestion with XbaI and MluI and ligation into the modified pUHD10-3 after its digestion with XbaI and BtgI and dephosphorylation. This was followed by transformation of DH5-alpha competent cells and selection in ampicillin (Sigma-

Aldrich). A clone with piMA-F5 was chosen by verification of the expected sizes of appropriate PCR products. Similarly, fragment 4 (oligo5 and oligo6) and fragment 3 (oligo7 and oligo8) were sequentially added using BfuAI (NEB) and SphI (NEB) for the digestions of the PCR products and BtgZI (NEB) followed by SphI for the plasmids. Fragment 2 was added as P-ds-oligo oligo9/oligo10 to the previous plasmid digested with BtgZI followed by SphI. Fragment 1 was added as P-ds-oligo oligo11/oligo12 to the resulting plasmid after digestion with BtgZI and BsiWI. This new plasmid was digested with NotI (NEB) and NheI (NEB) and ligations with P-ds-oligo oligo13/oligo14 generated pMA-Universal.

A series of plasmids containing an RA-II were generated: the construction plasmids. These plasmids contain the modified dhfr minigene and an RA-II surrounded by an appropriate SS set to allow "on-site" construction of DEs (see below). These plasmids are derived from pMA-Universal. For SS Set 7, the 5'-SS, the polypyrimidine tract, and the 3'-SS were added by three sequential rounds of ligation, transformation and selection using the restriction sites for NheI, for NotI and SphI, and for BtgZI and SphI, respectively, and three pairs of P-ds-oligos: oligo15/oligo16 for the 5'-SS, oligo17/oligo18 for the polypyrimidine tract, and oligo19/oligo20 for the 3'-SS. A similar procedure was used for SS Set 3, but oligo21/oligo22 was used for the second ligation. For SS Set 5, the 5'-SS, and the polypyrimidine tract together with the 3'-SS were added sequentially using the restriction sites for NheI, and for NotI and SphI respectively and two pairs of P-ds-oligos: oligo23/oligo24 for the 5'-SS, and oligo25/oligo26 for the rest. For SS Set 6, a similar approach was used but oligo27/oligo28 was used for the second ligation.

The receiving pMA plasmids contain an RA-I and were used for incorporating the DEs made in the pAL-SB plasmid into an modified dhfr minigene. Each receiving plasmid contains a SS set. For SS Set 5, P-ds-oligo oligo32/oligo33 was ligated into the pMA-Universal plasmid after digesting the latter with NheI and NotI. The intermediate plasmid containing the 5'SS of SS Set 7 described in the previous paragraph was digested with SphI and NotI and ligated to P-ds-oligo oligo34/oligo35 to generate the receiving pMA plasmid for SS Set 3. The RA-I used in the receiving pMA plasmids is different from the RA-II in the plasmids that allow stepwise construction of the DE and it leaves different overhangs upon its removal.

The pMA-FW plasmid provided the basis for incorporation of the modified dhfr minigenes into the genome. It contains a kanamycin resistance gene for initial selection of the cell line, a promoterless puromycin gene for subsequent selection of site-specific recombinations with DE-containing plasmids, an attP site for site-specific recombination and only the downstream half of the modified dhfr minigene. This plasmid was derived from pEGFP-C3. The CMV promoter and the EGFP gene were cut out with AseI and BamHI and in its stead a promoterless puromycin resistance gene was ligated by amplification from ptTA (a kind gift from Jim Manley) using primers oligo44 and oligo45 and digestion with AseI and BamHI. This new plasmid was digested with XhoI (NEB), dephosphorylated and ligated to P-ds-oligo oligo46/oligo47, which provides an attP site for PhiC31 recombinase (Groth et al. 2000). Several clones were sequenced and, of the two orientations possible for the attP site, the one in which oligo46 was on the sense strand of the puromycin gene was chosen. This intermediate plasmid was digested with AseI and XhoI. The downstream half of the minigene starting in the middle of intron 2 (1 bp downstream from the EcoRI site) and including 100 bp downstream from the polyA site was amplified from pMA-Universal using the primers oligo48 and oligo49, digested with AseI and XhoI, mixed with the digested intermediate plasmid and ligated to obtain pMA-FW.

Plasmid pMA-IC allows reconstitution of a fully functional puromycin resistance gene and a DE-containing modified dhfr minigene upon site-specific recombination with the sequence from pMA-FW (Fig. S3.5). The DE-containing plasmids for site-specific recombinations contained a CMV promoter to drive the puromycin resistance gene after site-specific recombination, the upstream half of the modified dhfr minigene including the DE for reconstitution of the modified dhfr minigene, and an attB site for site-specific recombination. An “empty” pMA-IC plasmid was constructed from a pMA-Universal derived plasmid which contained an irrelevant sequence between the NotI and the NheI sites. The CMV promoter was amplified from pEGFP-C3 using oligo50 and oligo51; both the pMA-Universal derived plasmid and the PCR product were cut with XbaI (NEB) and EcoRI (NEB) and ligated together. This step removed the downstream half of the minigene. An attB site for PhiC31 recombinase (Groth et al. 2000) was ligated into the XhoI site of the modified plasmid as P-ds-oligo oligo52/oligo53. Of the two orientations possible, the one in which oligo52 was on the sense strand of the partial dhfr minigene was chosen. The BtgZI site was removed to enable future extensions by digesting the previous plasmid with NcoI (NEB) and BsaAI (NEB) and ligating P-ds-oligo oligo54/oligo55.

To serve as the basis for the coupled-standards, the plasmid piS-Std was generated, which contained the skipped cDNA for the modified dhfr minigene. The cDNA of a transient transfection with a DE of 110nt (SS Set 7) composed exclusively of reference sequences was used for PCR amplification using primers oligo56 and oligo6. The PCR fragments obtained were digested with BfuAI and BsiWI and ligated into the plasmid piMA-F5 previously digested sequentially with BsiWI and BtgZI. Plasmid piS-Std was selected by the size of the products in appropriately chosen PCR amplifications. An adapter to facilitate subsequent ligations was added to generate piS-StdwAd by digestion with NcoI and XbaI, dephosphorylation and ligation of P-ds-oligo oligo57/oligo58. For generating the Gamma Actin coupled-standard, piSActin-Std, cDNA generated from MA-tTA cells by reverse transcription with primer oligo61 was amplified using primers oligo62 and oligo63. The PCR product and plasmid piS-StdwAd were digested with EcoRI and NotI and ligated together. For generating the coupled-standard for SS Sets 1, 2 and 4, piSI-CAG-Std, cDNA from a transient transfection using a DE of 110nt composed exclusively of reference sequences and SS Set 1 was amplified using primers oligo59 and oligo60. This PCR product and plasmid piS-StdwAd were digested with EcoRI and NotI and ligated together. The coupled-standard for SS Sets 3, 5, 6 and 7, piSI-CAA-Std, was made analogously from a transient transfection using a DE with SS Set 3: a mutation of A to G at position 64 of the DE was deemed innocuous and accepted.

DE construction

Most DEs were constructed in a stepwise fashion by ligating P-ds-oligos oligo64/oligo65 (NN), oligo66/oligo67 (EE), oligo68/oligo69 (EN), oligo70/oligo71 (NE), oligo72/oligo73 (SS), oligo74/oligo75 (SN), and oligo76/oligo77 (NS) into pAL-SB or the RA-II-containing construction plasmids (previous section). For the pAL-SB plasmids, the appropriate plasmids were digested with either BsmBI (to add a building block upstream of the DE in progress) or BsaI (to add a building block downstream). The final DEs were amplified with primers oligo78 and oligo42, digested with BbvI and ligated to the appropriate receiving pMA plasmid after removing its RA-I by digestion with BfuAI or its isoschizomer BveI (Fermentas). For the RA-II-containing construction plasmids, appropriate plasmids were digested with BfuAI (to add a building block downstream of the RA), BtgZI (to add a building block upstream) or both (to remove the RA or replace it with a building block). RA-II-containing construction plasmids with SS Sets 3 and 7 were digested with NheI and BtgZI to incorporate 22 bp DEs by ligating a P-ds-oligo oligo79/oligo80 or oligo81/oligo82 as appropriate. Constructs using SS Set 1 and SS Set 2 were made by amplifying the corresponding DEs from plasmids with SS Sets 3 and 7, respectively, using primers oligo29 and oligo30, digesting both the PCR products and pMA-Universal with NheI and NotI and ligating them together. By following this protocol, DEs using SS Set 4 were made by amplifying the corresponding DEs from plasmids with SS Set 3 using primers oligo29 and oligo31.

For generating the DE-containing pMA-IC plasmids, DEs were amplified by PCR from the appropriate modified dhfr minigenes using oligo29 and oligo83, digested with NotI and EcoRI and ligated into the pMA-IC plasmid, which was previously digested with NotI and EcoRI and dephosphorylated.

Psi measurement

RNA was extracted from transiently transfected cells using the RNA Spin Mini kit (GE Healthcare) and quantified using a Nanodrop 1000 (Thermo Scientific). Lack of degradation was assessed by gel electrophoresis. Reverse transcription (RT) was performed using an Omniscript kit (QIAGEN) in 10 µl reactions with 400 ng of RNA for each sample using the primer oligo84 at 100 nM. To measure the ratio between the mRNA molecules that skip the DE and those that contain it for SS Sets 3, 5, 6 and 7, the appropriate coupled-standard was prepared from plasmid piSI-CAA-Std by digestion with EcoO109I (NEB) followed by inactivation. The concentration of this digested plasmid was approximately 10^{10} plasmid molecules per µl based on absorbance measurements. This solution was diluted to approximately 10^{8} molecules per µl and a dilution series was prepared: 10-fold dilution per step. The starting solution was labeled as having exactly 10^{8} arbitrary units. Given that the coupled-standard plasmid concatenates a molecule that skips the DE and a molecule that includes it (see Supplemental Fig. S3.6), each diluted solution contains equimolar amounts of each, which enables accurate calibration by QPCR of one type of molecule relative to the other. (Furthermore, all coupled-standards were calibrated to each other by means of the common "skipped mRNA" region to further allow comparisons among standards.) QPCR was performed in 20 µl reactions that included 400 nM of forward and reverse primers, 2 µl of a 1:5 dilution of the RT product for each sample and 10 µl of 2X Power Green QPCR Master Mix (Applied Biosystems) using a 7300 PCR System (Applied Biosystems) according to the manufacturer's protocol. The data was analyzed using the software provided by the manufacturer. The primer sets used in QPCR reactions share the reverse primer oligo83 and include either oligo85 as the forward primer to detect the molecules that contain the DE or oligo86 to detect molecules that skip it. Since there is at least a 100-fold reduction in cross-detection of the molecules that skip the DE with the primers that detect its inclusion and vice versa (data not shown), this scheme provides the ratio (termed the SOI) of molecules that skip to those that include the DE for each sample; SOI should not be affected by differences in efficiency between the sets of primers used. The psi was obtained by the formula psi=100/(1+SOI). For SS Sets 1, 2 and 4, the coupled-standard derived from the plasmid piSI-CAG-Std was used along with oligo87 as the forward primer for the detection of inclusion.

Importantly, because of the placement of the QPCR primers all amplified products consist of identical sequences for each SS set and in particular are independent of the E, S and N combinations used, thus ensuring equal PCR efficiencies.

To assess the expression levels of the minigenes in stable transfections, the Gamma Actin primer oligo61 was added to the RT reaction. To quantify the mRNA levels for Gamma Actin, the coupled-standard derived from piSActin-Std was used in QPCR reactions. Comparisons between Gamma Actin mRNA and mRNA for the minigene are affected by the relative efficiency of the two reverse transcription primers, disallowing a direct comparison. However, normalization to Gamma Actin mRNA enables direct comparisons for the transcription levels of the minigene between samples.

Transfections

For transient transfections, cMA-HEK293-tTA cells were grown in 10 cm dishes to ~80% confluence. Cells from each dish were plated in 6 wells of a 6-well plate and incubated at 37°C for 24 hours. Transient transfections were performed using 600 ng of plasmid and 4 µl of Lipofectamine 2000 (Invitrogen) using Opti-MEM I (Invitrogen) according to the manufacturer’s protocol. Cells were incubated at 37°C for 25 hours before RNA extraction.

For stable transfections, cMA-FW cells were grown in 10 cm dishes to ~80% confluence. Transfections were performed using 2.4 µg of the DE-containing pMA-IC plasmid, 15 µg of pPGKPhiC31obpA plasmid (Addgene) and 30 µl of Lipofectamine 2000 using Opti-MEM I according to the manufacturer’s protocol. Successful PhiC31 site-specific recombinations (see

Supplemental Fig. S3.5) were selected after 72 hours of incubation at 37°C by adding puromycin (Sigma-Aldrich) to a final concentration of 4.2 µg/ml. In effect, only site-specific recombination allows reconstitution of the minigene (Supplemental Fig. S3.5). The puromycin containing medium was changed every 5 days. After ~3 weeks of puromycin exposure, the surviving clones were pooled and allowed to grow in 6-well dishes before RNA extraction.

Cell lines

HEK 293 cells were modified to express the tet-Off trans-activator (Gossen and Bujard 1992) by co-transfecting 1 µg of pUHD15-1 plasmid and 0.1 µg of pLi082 plasmid, which provides hygromycin resistance. Clones were grown in 100 µg/ml hygromycin and recloned. Individual clones were chosen and expression of a tetracycline-response-element controlled minigene was evaluated. A clone, cMA-HEK293-tTA, that displayed adequate expression levels and a good response to doxycycline was chosen (data not shown). This clone was used for all transient transfections.

This cell line was used to generate the cMA-FW by incorporation of pMA-FW digested with MluI and transfected by electroporation using Nucleofector II (Lonza) according to the manufacturer's protocol. Cells were incubated for 48 hours at 37°C and successful genomic incorporations of the transfected DNA were selected by adding G418 (Invitrogen) at a final concentration of 500 µg/ml. Site-specific recombination into these cells was evaluated with a pMA-IC plasmid containing the DE NNNENE. One clone cMA-FW was selected that provided adequate levels of expression for the reconstituted minigene and generated an acceptable number of colonies after puromycin selection (data not shown). The presence of a single genomic copy of pMA-FW was evaluated by verifying the full disruption of attP sites in puromycin surviving colonies by PCR using primers oligo83 and oligo88 (data not shown): full disruption in multiple independent site-specific recombinations, evidenced by the absence of PCR products, is expected only if a single attP site is present since reconstitution of a single puromycin resistance gene suffices for survival. This result was confirmed by using a Southern blot (data not shown). Also genomic DNA was digested with NspI, diluted and ligated to obtain DNA circles; inverse PCR was then performed using nested primer pairs: oligo83 and oligo44 in the first PCR reaction and oligo89 and oligo90 in the second. These products were cloned into the Not I site of pMA-Universal and sequenced. This information allowed mapping of the genomic integration point to PLEKHG1 in chromosome 6, specifically 141 bp before its 23 nt exon (i.e., intron 14 in NM_001029884.1). The location was verified by detection of PCR products that crossed the 2 ends of the integration site in the genomic DNA using primer pairs oligo91 with oligo89 and oligo92 with oligo93. Additionally, the size profile observed in the Southern Blot coincided with that predicted from integration at this genomic location.

Since the minigene was integrated into the sense strand of the PLEKHG1gene, we were concerned about the possibility that fusion transcripts would be synthesized in which a PLEKHG1 exon was spliced to a DE, leading to a counterfeit measurement of inclusion. However, no such fused mRNAs were detected by PCR using oligo94 (in the PLEKHG1 sequence) and oligo83 (in dhfr exon 3) probably due to the presence of a SV40 polyA site in pMA-FW upstream of the minigene.

Sequence of pMA-Universal

Shown below is the sequence inserted into pUHD10-3. Regions of exons 1 and 3 are shown in blue. The regions of introns 1 and 2 that were used are shown in gray. The restriction sites used for incorporation of DEs are indicated: the NotI site is shown in magenta and the NheI site is highlighted in yellow. The removable adapter is highlighted in green. The first and last four nucleotides of the entire sequence correspond to the overhangs added for cloning. The 3 nucleotides, TAC, that follow the first four were added to facilitate the transfer.

```
CGCGTACGGTTCGACCGCTGAACTGCATCGTCGCCGTGTCCCAGAATAAGGGCATCGGCAAGAACGGAGACCTTCCC
TGGCCAAAGCTCAGgtactggctggattgggttagggaaaccgaggcggttcgctgaatcgggtcgagcacttggcg
gagacgcgcgggccaactacttagggacagtcatgaggggtaggcccgccggctgcagcccttgcccatgcccgcgg
tgatccccatgctgtgccagcctttgcccagaggcgctctagctgggagcaaagtccggtcactgggcagcaccacc
ccccggacttgcatgggtagccgctgagatggagcctgagcacacgcggccgccgcatgcaacatcgcacctgctag
ctggccagtgagatccaagaatcttcctgtctctgctgatccactgataggattacaagtacatgccaccaagccca
gcttcctcttaccaggtgctggggaccaaacttaggccctcattcctacacagtgaatacttgactttgttatcacc
caaccctaataaataactcactatccaaacaagttgaaacccttagaattctgtgttgctccagcatgatgttgtgg
taaacgttaatacaataagatgcacaggtcataagtgcacattagctaagtgttgacaaagacttagacctacataa
cttaaccctattagccctccagaaagttcctcattctccattccaggcaactttcatcacaccacatcatgtacaac
tactattgaagttgttttccactatagatacaatgagatgtcacatacggctttgtgttttgatttgcaagtaccaa
tcgagtatgaaatatggagtggatattggacattggccaccatctaaatactttgtgttaaaagaattggttttcat
aatttgttttgtactgactgctggctagtcagattacctgactagtatggacaggattttgcaataatcataattct
tttttcagGGAACCACCACAAGGAGCTCATTTTCTTGCCAAAAGTCTGGACGAAGCCTTAAAACTTATTGAACAACC
AGAGTTAGCAGATAAAGTGGAGCTGTCATGGTTTGGATAGTTGGAGGCAGTTCCGTTTACAAGGAAGCCATGAATCA
GCCAGGCCATCTCAGACTCTTTGTGACAAGGATCATGCAGGAATTTGAAAGTGACACGTTCTTCCCAGAAATTGATT
TGGAGAAATATAAACTTCTCCCAGAGTACCCAGGGGTCCTTTCTGAAGTCCAGGAGGAAAAAGGCATCAAGTATAAA
TTTGAAGTCTATGAGAAGAAAGGCTAACAGAAAGATACTTGCTGATTGACTTCAAGTTCTACTGCTTTCCTCCTAAA
ATTATGCATTTTTACAAGACCATGGGACTTGTGTTGGCTTTAGATCTATGAGTTATTCTTTCTTTAGAGAGGGATAG
TTAGGAAGATGTATTTGTTTTGTGGTACCAGAGATGGAACCTGGGATCCTGTGCATCCTGGGCAACTGTTGTACTCT
AAGCCACTCCCCAAAGTCATGCCCCAGCCCCTGTATAATTCTAAACAATTAGAATTATTTTCATTTTCATTAGTCTA
ACCAGGTTATCTAG
```

Supplemental tables

Supplemental Table S3.1. Exon inclusion of DEs for size, ESE and ESS perturbation.

SS Set No.	3’ SS	5’SS	Exon size (nt)	Internal Name	Code	psi	Std error
			Size perturbation				
3	UCUCUUUUUUUCAG/G	CAA/GUAAGU	14	i6un	0N	42	4
3	UCUCUUUUUUUCAG/G	CAA/GUAAGU	46	i6uNN	2N	97	1
3	UCUCUUUUUUUCAG/G	CAA/GUAAGU	78	i6u(N)x4	4N	96	2
3	UCUCUUUUUUUCAG/G	CAA/GUAAGU	110	i6u(N)x6	6N	76	5
3	UCUCUUUUUUUCAG/G	CAA/GUAAGU	142	i6u(N)x8	8N	33	11
3	UCUCUUUUUUUCAG/G	CAA/GUAAGU	174	i6u(N)x10	10N	13	5
3	UCUCUUUUUUUCAG/G	CAA/GUAAGU	206	i6u(N)x12	12N	12	6
3	UCUCUUUUUUUCAG/G	CAA/GUAAGU	238	i6u(N)x14	14N	4	1
3	UCUCUUUUUUUCAG/G	CAA/GUAAGU	270	i6u(N)x16	16N	7	4
3	UCUCUUUUUUUCAG/G	CAA/GUAAGU	302	i6u(N)x18	18N	5	2
2	UCUCUAACUUUCAG/G	CAG/GUAAGU	22	innm1I3	½N	4	2
2	UCUCUAACUUUCAG/G	CAG/GUAAGU	46	iNNm1I3	2N	58	14
2	UCUCUAACUUUCAG/G	CAG/GUAAGU	78	i(N)x4m1I3	4N	94	3
2	UCUCUAACUUUCAG/G	CAG/GUAAGU	110	i(N)x6m1I3	6N	94	3
2	UCUCUAACUUUCAG/G	CAG/GUAAGU	142	i(N)x8m1I3	8N	78	2
2	UCUCUAACUUUCAG/G	CAG/GUAAGU	174	i(N)x10m1I3	10N	31	1
2	UCUCUAACUUUCAG/G	CAG/GUAAGU	206	i(N)x12m1I3	12N	10	4
2	UCUCUAACUUUCAG/G	CAG/GUAAGU	238	i(N)x14m1I3	14N	8	3
2	UCUCUAACUUUCAG/G	CAG/GUAAGU	270	i(N)x16m1I3	16N	4	0
			ESE Perturbation				
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i555	NNNNNN	7	0
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i255	ENNNNN	30	6
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i455	NENNNN	21	3

7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i525	NNENNN	34	11
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i545	NNNENN	27	5
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i552	NNNNEN	30	7
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i554	NNNNNE	18	3
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i155	EENNNN	34	6
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i225	ENENNN	38	8
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i245	ENNENN	57	6
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i252	ENNNEN	59	6
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i254	ENNNNE	35	6
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i425	NEENNN	47	6
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i445	NENENN	50	4
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i452	NENNEN	41	6
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i454	NENNNE	46	6
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i515	NNEENN	46	5
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i522	NNENEN	39	3
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i524	NNENNE	34	4
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i542	NNNEEN	32	4
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i544	NNNENE	31	4
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i551	NNNNEE	29	5
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i125	EEENNN	74	6
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i145	EENENN	80	7
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i152	EENNEN	75	6
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i154	EENNNE	56	8
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i215	ENEENN	81	8
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i222	ENENEN	79	6
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i224	ENENNE	57	3
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i242	ENNEEN	79	5
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i244	ENNENE	65	3
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i251	ENNNEE	60	4

7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i415	NEEENN	73	4
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i422	NEENEN	74	5
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i424	NEENNE	71	5
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i442	NENEEN	76	6
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i444	NENENE	67	5
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i451	NENNEE	78	4
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i512	NNEEEN	74	7
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i514	NNEENE	72	11
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i521	NNENEE	70	11
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i541	NNNEEE	63	13
7	UCUCUAACUUUCAG/G	CAA/GUAAGU	110	i4u555	EEEEEE	96	0
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u255	ENNNNN	90	2
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u455	NENNNN	90	2
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u525	NNENNN	90	2
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u545	NNNENN	89	3
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u552	NNNNEN	89	2
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u554	NNNNNE	86	4
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u111	EEEEEE	98	0
ESS Perturbation							
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u555	NNNNNN	49	3
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u855	SNNNNN	47	1
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u655	NSNNNN	38	4
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u585	NNSNNN	38	4
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u565	NNNSNN	38	9
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u558	NNNNSN	34	4
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u556	NNNNNS	30	5
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u955	SSNNNN	37	2
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u885	SNSNNN	35	2
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u865	SNNSNN	46	1

5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u858	SNNNSN	38	4
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u856	SNNNNS	30	3
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u685	NSSNNN	35	2
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u665	NSNSNN	33	2
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u658	NSNNSN	34	2
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u656	NSNNNS	24	2
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u595	NNSSNN	32	3
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u588	NNSNSN	20	2
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u586	NNSNNS	22	4
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u568	NNNSSN	27	4
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u566	NNNSNS	19	2
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u559	NNNNSS	15	3
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u985	SSSNNN	27	5
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u965	SSNSNN	19	1
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u958	SSNNSN	18	1
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u956	SSNNNS	16	2
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u895	SNSSNN	18	2
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u888	SNSNSN	22	1
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u886	SNSNNS	13	1
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u868	SNNSSN	18	1
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u866	SNNSNS	17	6
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u859	SNNNSS	16	6
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u695	NSSSNN	16	5
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u688	NSSNSN	15	5
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u686	NSSNNS	8	1
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u668	NSNSSN	16	5
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u666	NSNSNS	5	2
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u659	NSNNSS	8	1
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u598	NNSSSN	13	4

5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u596	NNSSNS	11	3
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u589	NNSNSS	11	2
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u569	NNNSSS	14	3
5	UCUCUAUUUUUCAG/G	CAA/GUAAGU	110	i4u999	SSSSSS	5	0

Supplemental Table S3.2. Oligomers used.

Oligomer	Sequence	Primary purposes
oligo1	CGCGCGACCTTCAGCATTG	Removal of BtgZI site on pUHD10-3
oligo2	CCGGCAATGCTGAAGGTCG	Removal of BtgZI site on pUHD10-3
oligo3	AAACGCGTACGGCGATGCCGCATGCAAGCTGTCATGGTTTGGATAGTTGG	Primer for fragment 5 of the modified dhfr minigene
oligo4	CCTCTAGATAACCTGGTTAGACTAATG	Primer for fragment 5 of the modified dhfr minigene
oligo5	CCGCATGCAAAGCCTTAAAACTTATTGAACAACC	Primer for fragment 4 of the modified dhfr minigene
oligo6	CCCCCCACCTGCAAAACAGCTCCACTTTATCTGCTAACTCTGG	Primer for fragment 4 of the modified dhfr minigene
oligo7	CCGCATGCAAAGCTCAGGTACTGGCTGGATTGGG	Primer for fragment 3 of the modified dhfr minigene
oligo8	CCCCCCACCTGCAAAAGGCTTCGTCCAGACTTTTGGCAAG	Primer for fragment 3 of the modified dhfr minigene
oligo9	CAAAGGGCATCGGCAAGAACGGAGACCTTCCCTGGCCAA	Primer for fragment 2 of the modified dhfr minigene
oligo10	AGCTTTGGCCAGGGAAGGTCTCCGTTCTTGCCGATGCCCTTTGCATG	Primer for fragment 2 of the modified dhfr minigene
oligo11	GTACGGTTCGACCGCTGAACTGCATCGTCGCCGTGTCCCAGAATA	Primer for fragment 1 of the modified dhfr minigene
oligo12	CCCTTATTCTGGGACACGGCGACGATGCAGTTCAGCGGTCGAACC	Primer for fragment 1 of the modified dhfr minigene
oligo13	GGCCGCCGCATGCAACATCGCACCTG	Addition of the universal RA
oligo14	CTAGCAGGTGCGATGTTGCATGCGGC	Addition of the universal RA
oligo15	CTAGGAAACAACACAAGTAAGTG	Addition of the 5’SS for SS Set 7 to Universal RA
oligo16	CTAGCACTTACTTGTGTTGTTTC	Addition of the 5’SS for SS Set 7 to Universal RA
oligo17	GGCCGCTGTTAACGCAGTGTTTCTCTAACTTTAAGCATG	Addition of the polypyrimidine tract for SS Set 7 to Universal RA
oligo18	CTTAAAGTTAGAGAAACACTGCGTTAACAGC	Addition of the polypyrimidine tract for SS Set 7 to Universal RA

oligo19	CTTTCAGGCCAAACGGGCATG	Addition of the 3’SS for SS Set 7 to Universal RA
oligo20	CCCGTTTGGCCTG	Addition of the 3’SS for SS Set 7 to Universal RA
oligo21	GGCCGCTGTTAACGCAGTGTTTCTCTTTTTTTAAGCATG	Addition of the polypyrimidine tract for SS Set 3 to Universal RA
oligo22	CTTAAAAAAAGAGAAACACTGCGTTAACAGC	Addition of the polypyrimidine tract for SS Set 3 to Universal RA
oligo23	CTAGGAAACAACACAAGTAAGTG	Addition of the 5’SS for SS Set 5 to Universal RA
oligo24	CTAGCACTTACTTGTGTTGTTTC	Addition of the 5’SS for SS Set 5 to Universal RA
oligo25	GGCCGCTGTTAACGCAGTGTTTCTCTATTTTTCAGGCCAAACGGGCATG	Addition of the polypyrimidine tract and 3’SS for SS Set 5 to Universal RA
oligo26	CCCGTTTGGCCTGAAAAATAGAGAAACACTGCGTTAACAGC	Addition of the polypyrimidine tract and 3’SS for SS Set 5 to Universal RA
oligo27	GGCCGCTGTTAACGCAGTGTTTCTCTAATTTTCAGGCCAAACGGGCATG	Addition of the polypyrimidine tract and 3’SS for SS Set 6 to Universal RA
oligo28	CCCGTTTGGCCTGAAAATTAGAGAAACACTGCGTTAACAGC	Addition of the polypyrimidine tract and 3’SS for SS Set 6 to Universal RA
oligo29	GCCAACTACTTAGGGACAGT	Common primer to transfer DEs
oligo30	CACTGGCCAGCTAGCACTTACCTGTGTTGTTTG	Primer to transfer DEs while adding a consensus 5’SS
oligo31	CACTGGCCAGCTAGCACTCACTTGTGTTGTTTG	Primer to transfer DEs while adding a wild-type 5’SS
oligo32	GGCCGCTGTTAACGCAGTGTTTCTCTATTTTTCAGGCCAAACAGGGCGCAG GTGCATGCACCTGCTAGGAAACAACACAAGTAAGTG	Generation of the receiving plasmid with SS Set 5
oligo33	CTAGCACTTACTTGTGTTGTTTCCTAGCAGGTGCATGCACCTGCGCCCTGT TTGGCCTGAAAAATAGAGAAACACTGCGTTAACAGC	Generation of the receiving plasmid with SS Set 5

oligo34	GGCCGCTGTTAACGCAGTGTTTCTCTTTTTTTCAGGCCAAACAGGGCGCAG GTGCATG	Generation of the receiving plasmid with SS Set 3
oligo35	CACCTGCGCCCTGTTTGGCCTGAAAAAAAGAGAAACACTGCGTTAACAGC	Generation of the receiving plasmid with SS Set 3
oligo36	AAAAAAGGTCTCGGATTGCACGCTGGTTCTCCGGCCGCTTGGGT	Primer from set 1 in nested PCR to remove BfuAI sites from EGFP-C3 plasmid
oligo37	TCGAATGGGCACGTAGCCGGATCAAGCGTATGCA	Primer from set 1 in nested PCR to remove BfuAI sites from EGFP-C3 plasmid
oligo38	TGATCCGGCTACGTGCCCATTCGACCACCAAGCGAAACA	Primer from set 2 in nested PCR to remove BfuAI sites from EGFP-C3 plasmid
oligo39	AAAAAAGGTCTCGTCGTGATGCCAGGTTGGGCGTCGCTTGGT	Primer from set 2 in nested PCR to remove BfuAI sites from EGFP-C3 plasmid
oligo40	AAAAAAGGTCTCCCCACCCAGACCCCATTGGGGCCAATA	Primer for PCR to remove the BsaI site from EGFP-C3 plasmid
oligo41	TATGGCAGGGCCTGCCGCCCCGA	Primer for PCR to remove the BsaI site from EGFP-C3 plasmid
oligo42	GCAAAGACCCCAACGAGAAGCGCGA	Addition of the linker to generate the pAL-SB plasmid
oligo43	CCCCCTGCAGCAGCCGTCTCCAAACAGAGACCAGCTGCAAGCTTGAGCTCG AGATCTGAGTA	Addition of the linker to generate the pAL-SB plasmid
oligo44	CCCCCCATTAATCCCCCCTCGAGCCACCATGACCGAGTACAAGCCCA	Amplification of the promoterless puromycin gene
oligo45	GGGGATCCTCAGGCACCGGGCTTGCGGGT	Amplification of the promoterless puromycin gene
oligo46	TCGAGCCCCAACTGGGGTAACCTTTGAGTTCTCTCAGTTGGGGG	Incorporation of the attP site to generate pMA-FW
oligo47	TCGACCCCCAACTGAGAGAACTCAAAGGTTACCCCAGTTGGGGC	Incorporation of the attP site to generate pMA-FW
oligo48	CCCTCGAGTGTTGCTCCAGCATGATGTTGT	Amplification and transfer of the second half of the modified dhfr minigene to generate pMA-FW

oligo49	GGGGGGATTAATAGACGACGAGGCTTGCAGGATCAT	Amplification and transfer of the second half of the modified dhfr minigene to generate pMA-FW
oligo50	CCCCCCTCTAGATCATAGCCCATATATGGAGTTCCGCGT	Amplification and transfer of the CMV promoter
oligo51	TGAATTCCGGATCTGACGGTTCACTAAACCA	Amplification and transfer of the CMV promoter
oligo52	AATTCGCGCCCGGGGAGCCCAAGGGCACGCCCTGGCACC	Incorporation of the attB site to generate pMA-IC
oligo53	AATTGGTGCCAGGGCGTGCCCTTGGGCTCCCCGGGCGCG	Incorporation of the attB site to generate pMA-IC
oligo54	CATGGTAATAGCCATGACTAATAC	Removal of BtgZI site to generate pMA-IC plasmid
oligo55	GTATTAGTCATGGCTATTAC	Removal of BtgZI site to generate pMA-IC plasmid
oligo56	CCCGTACGGTTCGACCGCTGAACTGCATCG	Amplification of cDNA to generate the standard plasmids
oligo57	CATGGACGAATTCCCCAAAGCGGCCGCAA	Addition of an adapter to generate the coupled-standard plasmids
oligo58	CTAGTTGCGGCCGCTTTGGGGAATTCGTC	Addition of an adapter to generate the coupled-standard plasmids
oligo59	CCCCCGCGGCCGCAAAGATCCAGCCTCCGCGTA	Amplification of cDNA to generate the coupled-standard plasmids
oligo60	CCCCCGAATTCAAAACACAAGTCCCATGGTCTTGTA	Amplification of cDNA to generate the coupled-standard plasmids
oligo61	GCATTTGCGGTGGACG	Reverse transcription primer for Gamma Actin
oligo62	AAACCGCGGCCGCTCGTGCGTGACATTAAGGAGA	Amplification of Gamma Actin cDNA for coupled-standard generation
oligo63	AAACCGAATTCGCATTTGCGGTGGACG	Amplification of Gamma Actin cDNA for coupled-standard generation
oligo64	AAACAACCAAACAACCAAACAACCAAACAACC	Generation of a building block containing two reference sequences (NN)

oligo65	GTTTGGTTGTTTGGTTGTTTGGTTGTTTGGTT	Generation of a building block containing two reference sequences (NN)
oligo66	AAACAATCCTCGAACCAAACAATCCTCGAACC	Generation of a building block containing two enhancer sequences (EE)
oligo67	GTTTGGTTCGAGGATTGTTTGGTTCGAGGATT	Generation of a building block containing two enhancer sequences (EE)
oligo68	AAACAATCCTCGAACCAAACAACCAAACAACC	Generation of a building block containing an enhancer and a reference sequence (EN)
oligo69	GTTTGGTTGTTTGGTTGTTTGGTTCGAGGATT	Generation of a building block containing an enhancer and a reference sequence (EN)
oligo70	AAACAACCAAACAACCAAACAATCCTCGAACC	Generation of a building block containing a reference and an enhancer sequence (NE)
oligo71	GTTTGGTTCGAGGATTGTTTGGTTGTTTGGTT	Generation of a building block containing a reference and an enhancer sequence (NE)
oligo72	AAACAACACATGGTCCAAACAACACATGGTCC	Generation of a building block containing two silencer sequences (SS)
oligo73	GTTTGGACCATGTGTTGTTTGGACCATGTGTT	Generation of a building block containing two silencer sequences (SS)
oligo74	AAACAACACATGGTCCAAACAACCAAACAACC	Generation of a building block containing a silencer and a reference sequence (SN)
oligo75	GTTTGGTTGTTTGGTTGTTTGGACCATGTGTT	Generation of a building block containing a silencer and a reference sequence (SN)
oligo76	AAACAACCAAACAACCAAACAACACATGGTCC	Generation of a building block containing a reference and a silencer sequence (NS)
oligo77	GTTTGGACCATGTGTTGTTTGGTTGTTTGGTT	Generation of a building block containing a reference and a silencer sequence (NS)
oligo78	GCGGTACCGTCGACTTCAGCAGCCGT	Transfer of the finished DEs to a receiving plasmid

oligo79	AAACAACCAAACAACACAGGTAAGTG	Generation of 22nt DEs: SS Set 3
oligo80	CTAGCACTTACCTGTGTTGTTTGGTT	Generation of 22nt DEs: SS Set 3
oligo81	AAACAACCAAACAACACAAGTAAGTG	Generation of 22nt DEs: SS Set 7
oligo82	CTAGCACTTACTTGTGTTGTTTGGTT	Generation of 22nt DEs: SS Set 7
oligo83	GGAACTGCCTCCAACTATCCAA	Transfer of DEs to pMA-IC for stable transfections and shared QPCR reverse primer for the modified dhfr minigene
oligo84	AGAGTCTGAGATGGCCTGGCT	Reverse transcription primer for the modified dhfr minigene
oligo85	CAAACAACACAAGGAACCACCA	QPCR forward primer for molecules including DEs with wild type 5’SS
oligo86	GCCAAAGCTCAGGGAACCA	QPCR forward primer for molecules skipping the DEs
oligo87	CAAACAACACAGGGAACCACC	QPCR forward primer for molecules including DEs with consensus 5’SS
oligo88	TCATGGTGGTTCGACCCCCAA	Primer for verification of disruption of attP sites
oligo89	AAAAGCGGCCGCGTGCCTGAGGATCGGATCTA	Primer for the second PCR amplification in the nested PCR used to map the genomic incorporation of pMA-FW
oligo90	AAAAGCGGCCGCGGCGGTAATACGGTTATCCA	Primer for the second PCR amplification in the nested PCR used to map the genomic incorporation of pMA-FW
oligo91	TCATCTTTACATAATTGTCATGGCAT	Primer for verification of the genomic location of pMA-FW
oligo92	CAACACTCAACCCTATCTCGGTCTA	Primer for verification of the genomic location of pMA-FW
oligo93	CTGAAGTGAACATTTCCAAGTAAGAA	Primer for verification of the genomic location of pMA-FW

oligo94	GCTCTAAAGAAGGTTCTGCTCCAT	Primer for detection of hybrid mRNA molecules

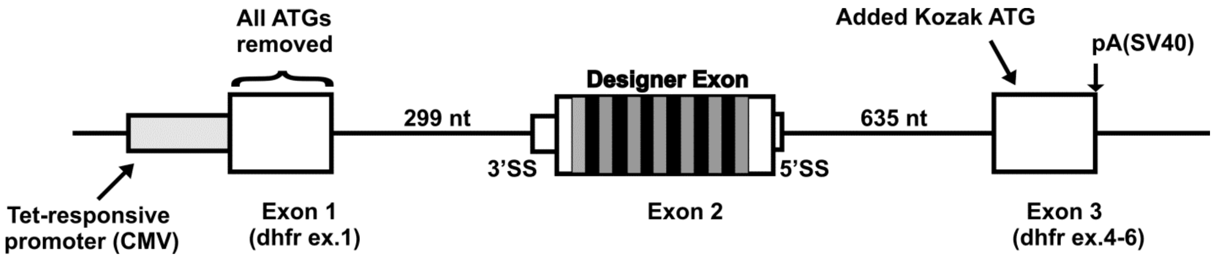


Figure S3.1. Cartoon of a typical DE-containing minigene.

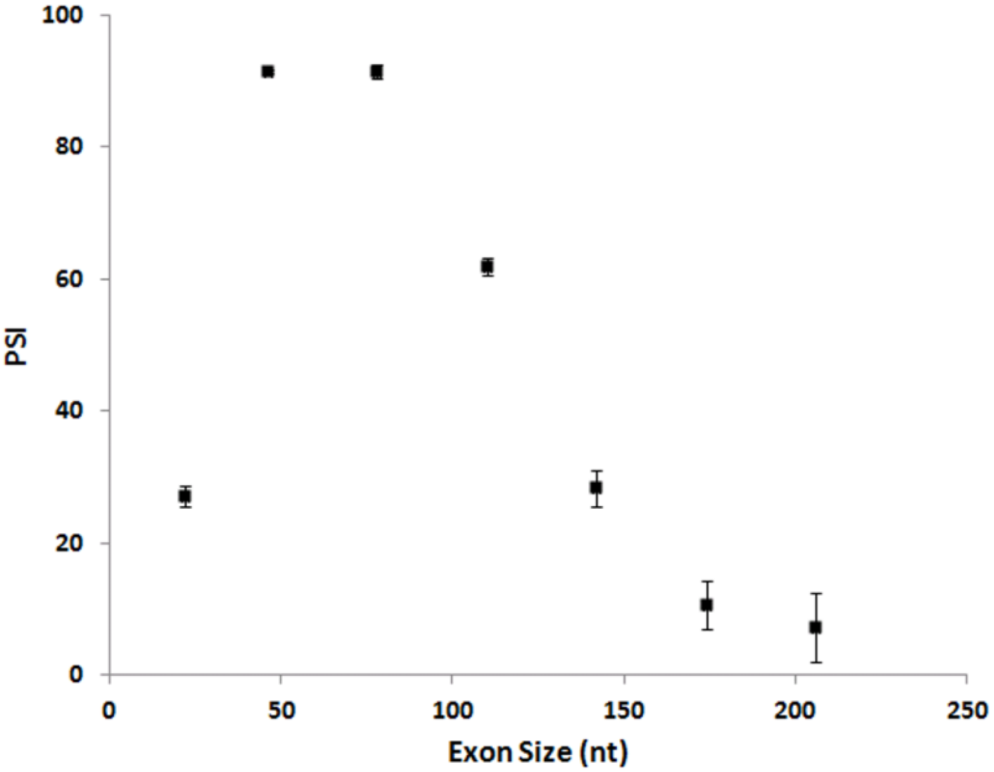


Figure S3.2. Exon inclusion has an optimum size range in a chromosomal context. Inclusion levels (psi) of DEs of various sizes in a chromosomal context. DEs consist of reference sequences and have a strong 3'SS. Error bars: SEM, n≥3.

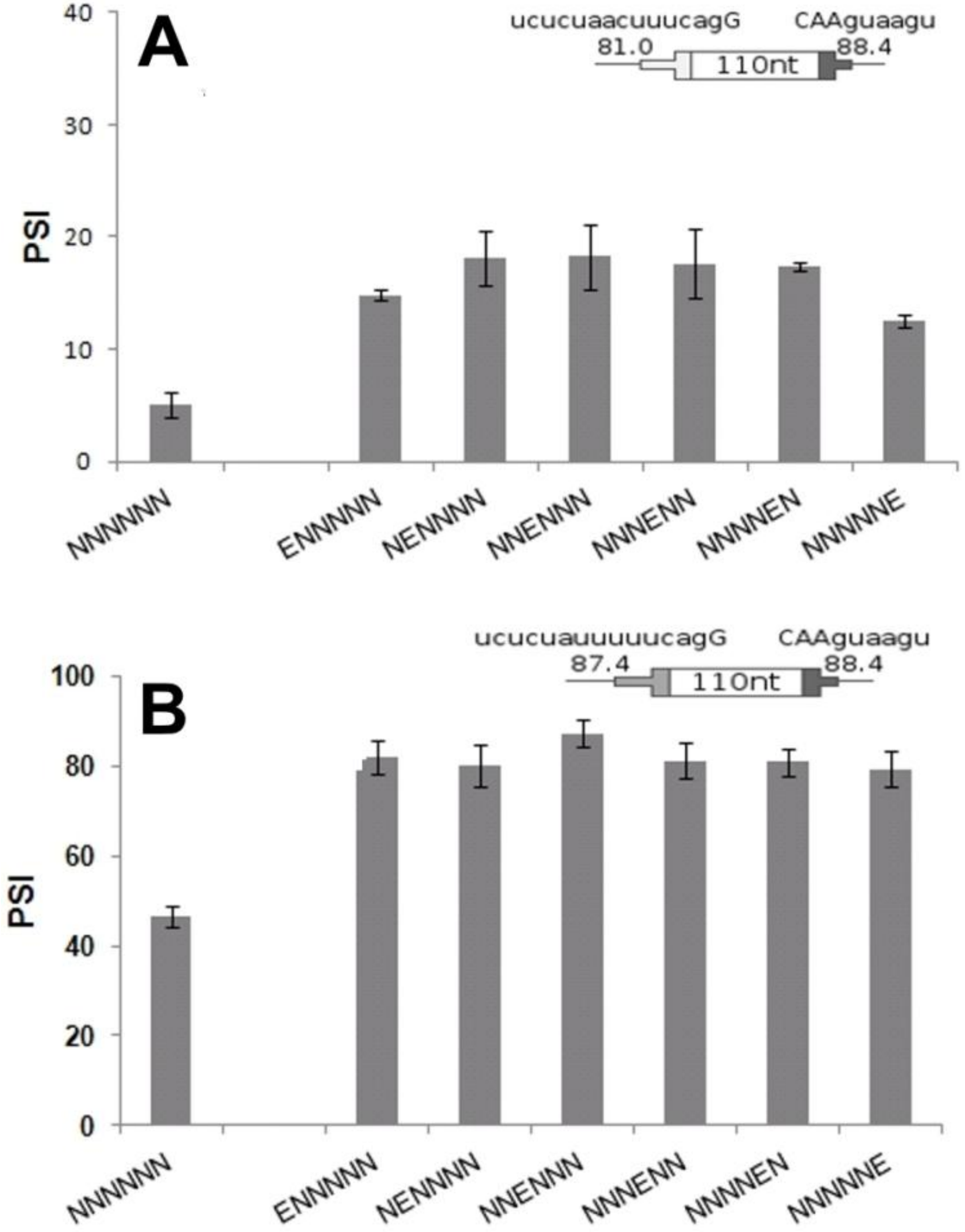


Figure S3.3. Addition of a single ESE enhances inclusion level and is position independent in a chromosomal context. The cartoons show the consensus values for splice site strengths used. A. Position variation in DEs with SS Set 7. B. Position variation in DEs with SS Set 5. Error bars: SEM, n≥3. In all panels the psi of DEs with an ESE are significantly different from that without an ESE (t-test, p<0.01).

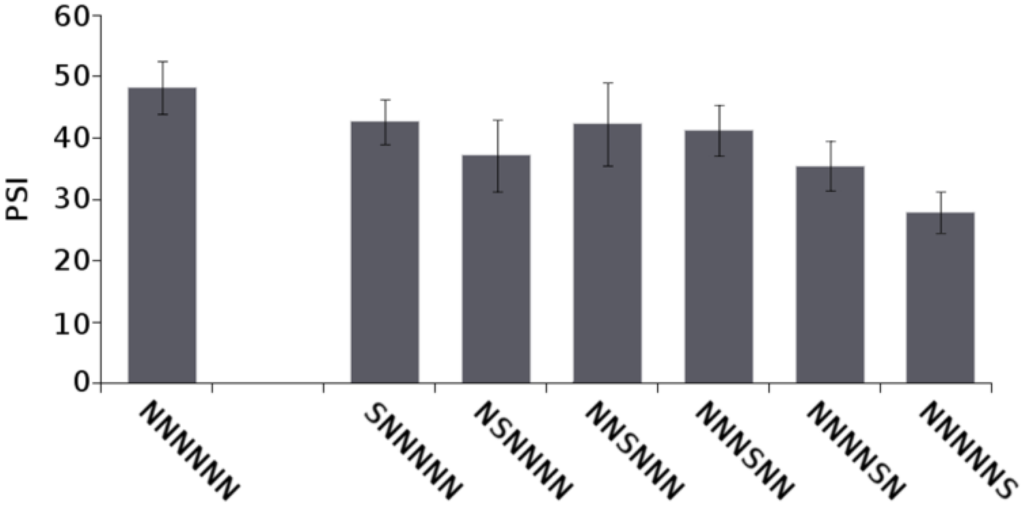


Figure S3.4. Addition of a single ESS decreases inclusion level and shows some position dependence in a chromosomal context. The psi for DEs with a single ESS are shown for stable transfections. Error bars: SEM, n≥3.

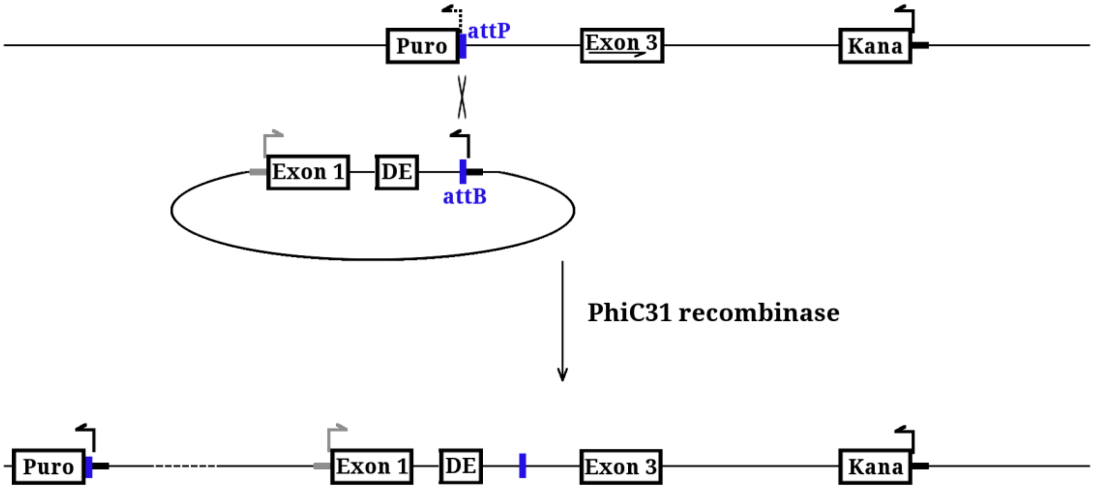


Figure S3.5. Site-specific recombination reconstitutes the minigene in a specific location of the genome. Using the kanamycin resistance gene (Kana) through selection with G418, an attP site has been incorporated in the genome of HEK 293 cells along with the downstream half of the modified dhfr minigene and a promoterless copy of a gene conferring puromycin resistance (Puro). After transient transfection with a plasmid incorporating the upstream half of the minigene as well as a promoter for the puromycin resistance gene, along with a gene for PhiC31 recombinase, puromycin-resistant site-specific recombinants can be isolated that have reconstituted the minigene as well as the puromycin resistance gene. Exons are indicated with boxes while introns and intergenic regions are indicated by thin lines. The promoters are indicated with thick horizontal lines: gray for the minigene and black for the puromycin and G418 resistance genes. The direction of transcription is indicated by bent arrows; the dashed arrow indicates the nominal direction of transcription for the promoterless puromycin resistance gene. For exon 3 of the minigene, the nominal direction of transcription is indicated with a horizontal arrow. The PhiC31 recombination sites are indicated by blue vertical lines.

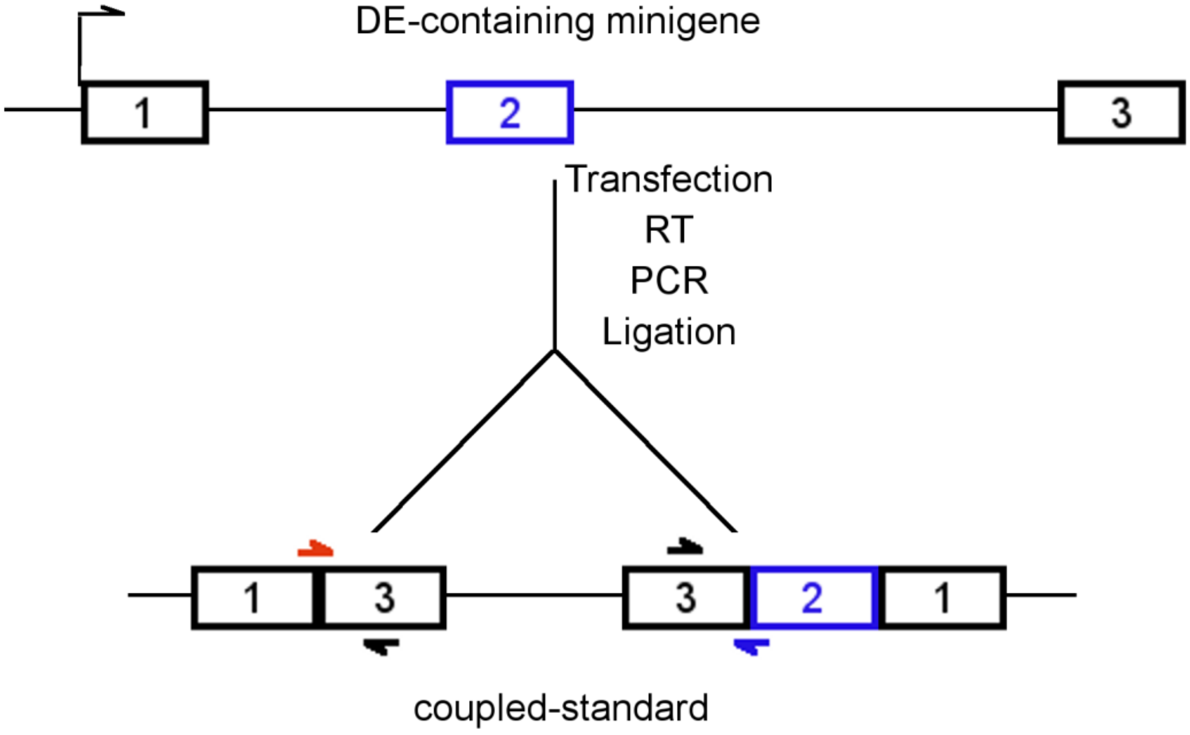


Figure S3.6. Coupled-standards incorporate two cDNAs into a single molecule. A reverse transcribed copy of the mRNA with the DE spliced in (included) as well as a copy without it (skipped) have been incorporated into the same plasmid molecule by sequential ligations. Digestions of this plasmid are therefore guaranteed to have equimolar amounts of both species. A dilution series of these molecules was used as a standard in QPCR reactions. The primers used for QPCR of the standards and the experimental samples are indicated with arrows: black, shared primer; blue, joint primer for detection of included molecules; and red, joint primer for detection of skipped molecules. See Detailed Materials and Methods above for details.

Chapter 4

Parts of this chapter were submitted as part of a manuscript (Arias et al. 2013)

Derivations of the Equations to Model Splicing

Introduction

Transcripts of many genes in higher eukaryotes are interrupted by sequences that are removed to generate mRNA molecules. These sequences are called introns and those sequences that are spliced together to form the mRNA molecule are called exons. This splicing process occurs with very high accuracy regarding the identification of the ends of the exons (Fox-Walsh and Hertel 2009). Many decades of research have seen continuous progress in understanding this phenomenon. Early on the exon/intron junctions were identified as functional sequences (Mount 1982). This was followed by the identification of functional sequences inside of the exons that could have either a facilitator effect on the proper inclusion of the exon in which they reside, Exonic Splicing Enhancers or ESEs, or a silencing effect that negatively affected exon inclusion, Exonic Splicing Silencers or ESSs.

There are two models that have been proposed regarding how the exons are paired up (Berget 1995; De Conti et al. 2013). The first model is called intron definition and it postulates that the ends of the intron are recognized and paired making the intron the unit of recognition. In this case, the exons are defined by the introns that flank them. The second model is called exon definition and it postulates that the exon is recognized as a unit. Exons are subsequently paired up in a collinear manner and in this way the introns are defined. Both models could be affecting splicing in a single organism and it has been proposed that small introns are recognized by intron definition while short exons are recognized by exon definition (Fox-Walsh et al. 2005). In higher eukaryotes it has been noted that most exons are relatively short while the flanking introns are relatively long, making exon definition the predominant mechanism (Fox-Walsh et al. 2005).

One of the main tools available to study splicing is *in vitro* splicing (Krainer et al. 1984). It entails mainly the placement of *in vitro* pre-synthesized RNA in an environment similar to the cell nucleus, in the form of nuclear extract obtained from cultured cells. Under appropriate conditions RNA splices and, while some of the characteristics observed *in vivo* are reproduced, others are not. For example, this type of assay forces a sequential relationship between transcription and splicing. However, removal of introns has been shown to occur co-transcriptionally (Kessler et al. 1993; Wada et al. 2009). Indeed, in many cases it has been shown to occur shortly after the two spliced exons have been synthesized and to be independent on the length of the downstream intron (Wada et al. 2009). This simultaneity makes it possible for one process to affect the other. It has been shown for example that changes in the rate of transcription elongation can affect how often an exon is included (de la Mata et al. 2003). Another aspect that deserves mention is the time required for splicing to occur. In *in vitro* assays, the time required for splicing is at least an order of magnitude greater than *in vivo* (Hicks et al. 2005). Transcription/splicing assays have been developed but these shortcomings have not been addressed properly (Lazarev and Manley 2007).

The decision to include a specific exon in the mRNA molecule occurs early as an irreversible step (Lim and Hertel 2004). This step precedes or coincides with formation of complex A, occurs after complex E formation and pairs the two ends of what constitutes the intron to be removed. The formation of an exon definition complex should precede these events and is usually associated with complex E formation. This complex depends on the presence of both splice sites defining the exon. It is these early steps the ones affected by specific sequences in many of the exons studied (Black 2003; House and Lynch 2006; Chen and Manley 2009). However, other steps of the splicing process might be affected.

Early observations in Miller chromatin spreads of *D. melanogaster* embryos provided evidence for co-transcriptional removal of introns (Osheim et al. 1985; Osheim et al. 1988). This has been confirmed by studies of long genes (Singh and Padgett 2009; Wada et al. 2009). These and other studies indicate that splicing takes place mere minutes after transcription (Kessler et al. 1993; Singh and Padgett 2009; Wada et al. 2009). For long genes this allows the removal of the introns to occur co-transcriptionally. For short genes on the other hand, the nominal time required for transcription might be too short for the removal to occur. Importantly though, several of the early steps in the splicing process should have had enough time to occur making the splicing process co-transcriptional.

Here the development of a novel mathematical tool that allows testing of mechanisms for splicing is presented. For its development, a focus was put on exon definition since it is the predominant mechanism in humans (Fox-Walsh et al. 2005). The problem was simplified by studying designer exons (DEs), exons of our own making made of combinations of a small number of modules. This allowed the exploration of general mechanisms that would affect equally these simplified exons and the natural ones, while avoiding the complexities present in the latter. Within this framework a general equation is obtained and used as a testing ground for different ideas about exon recognition. The effect of size is modeled as mediated by tethered collisions of exon ends that when successful allow the formation of a complex enabling exon definition. ESEs are modeled as having an effect on the stability of this complex. The resulting model will be tested in Chapter 5.

A Biophysical Model to Explain Splicing Decisions

We embarked on the design of our own exons so as to be able to examine individual parameters that govern splicing decisions. While this reductionist approach dispenses with the complexity of natural exons, it has the advantage of making fundamental principles discernible. Having varied parameters in over 150 DEs (see Chapter 3), we sought to develop a biophysical model that could explain these data. The goal of this model is to relate the observed psi to the parameters that have been varied in these DEs.

The biophysical model is centered on exon definition as a decisive step in the recognition of most splice sites and assumes that this step requires the formation of an RNA-protein complex on the exon of interest. The number of pre-mRNA molecules with a complex is determined by the balance between assembly and disassembly, which can be described by overall association and dissociation rate constants. Once assembled, complexed molecules can then proceed to a state of commitment to exon inclusion delineating the commitment progress (Fig. 4.1A).

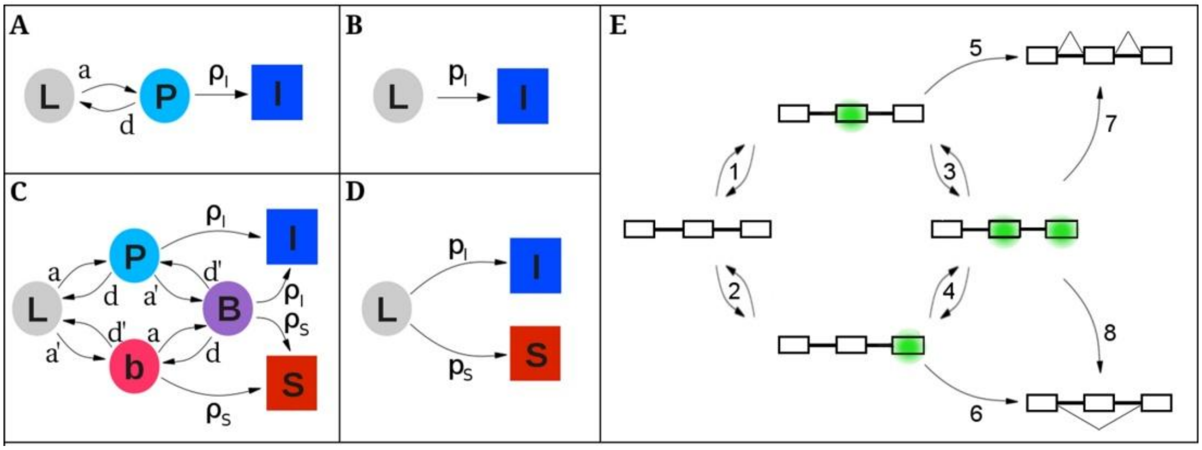


Figure 4.1. Complex kinetics can be described in simpler terms. The squares and circles represent different states of a pre-mRNA molecule: L, "naked" transcript; P, exon of interest in an exon definition complex (EDC) with the downstream exon either not present or present but not in an EDC; b, downstream exon in an EDC with the exon of interest not in an EDC; B, both exons in EDCs; I (inclusion) and S (skipping) represent molecules that have either committed to or achieved their respective splicing outcomes. The arrows represent transitions between states, and are labeled with rate constants: a and d, association and dissociation, respectively, of the complex on the exon of interest; a' and d', the same for the downstream exon; ρ_I and ρ_S, commitment to inclusion and skipping, respectively, of the exon of interest. A. Model for the splicing reactions before time τ. Importantly, the transition from P to I is independent of the presence of exon 3. B. Simplified model before time τ; p_I amalgamates a, d and ρ_I. C. Model for the splicing reactions after time τ. D. Model after time τ simplified analogously to panel B. See Supplemental Material for details. E. Cartoon showing the states implied in panel C for a pre-mRNA molecule depicting EDCs (green). Steps 1 to 4 represent the formation or loss of EDCs; steps 5 to 8 represent commitments to the splicing outcome shown.

First time period

We start with a set of assumptions that are described in Box 4.1 and focus on a group of pre-mRNA molecules (conceptually "tagged") that are all in the same state of synthesis. To consider the choice between inclusion and skipping, it seems necessary that the exon

Box 4.1: Model assumptions and definitions

General equations were obtained based on the following conditions and assumptions:

1. The cell or system studied is at steady state.
2. We consider all the pre-mRNA molecules of interest that have started transcription within the same negligibly small time interval as “tagged”; it is this “pulse-tagged” cohort of molecules whose fate will be analyzed.
3. For simplicity, we assume that the transcription rate is the same for all of these molecules.
4. We define time zero as the time at which the exon of interest has been synthesized and made available for splicing.
5. Each tagged molecule contains at least one internal exon, one of which is the exon of interest. We will assess inclusion and skipping of this exon in all the tagged molecules.
6. There is a complex that forms on the exon of interest that is an obligatory intermediate for exon inclusion. We consider this complex to be an exon definition complex. At any given point in time we define P as the number of tagged molecules with this complex and I as the number that are committed to inclusion, taken to be splicing to the committed exon that lies immediately upstream.
7. The exons flanking the exon of interest are constitutive. In particular, we assume that by the time the exon of interest is made available for splicing, the upstream exon is already committed.
8. Exon definition can commit an internal exon to inclusion whether or not the downstream exon has been synthesized.
9. We assume first order kinetics for all transitions between states. In particular, $dI/dt = \rho_I * P$, where ρ_I is the rate at which these molecules commit to the included pathway.
10. All tagged molecules will follow one of two pathways: inclusion or skipping; we will consider only the decision between these 2 possibilities.

downstream be synthesized; we define time τ as the time required to make this downstream exon available for splicing. For times prior to τ, there are 3 types of pre-mRNA molecules with respect to the exon of interest: naked, L; complexed, P; committed to inclusion, I (Fig. 4.1A). A set of differential equations relates the number of tagged P, I and L molecules starting at $t = 0$:

4.1 $dL/dt = d\,P - a\,L$

4.2 $dP/dt = a\,L - (d+\rho_I)\,P$

4.3 $dI/dt = \rho_I\,P$

It is a cohort of previously tagged molecules that is being followed, so rates of synthesis need not be considered.

In equations 4.1 through 4.3, defining F as the number of uncommitted molecules, $F = L + P$, and taking the Laplace transform, indicated as $X = X$ (s) for any function X(t), for the equations 4.1 and 4.2, we get

4.4 $(s+a)\,F = (s+d+a)\,P + F_0 - P_0$

4.5 $(s+a+d+\rho_I)\,P = a\,F + P_0$

where F_0 and P_0 represent initial values for F(t) and P(t) respectively.

Solving for *F*,

4.6 $[s^2 + (a+d+\rho_I)\,s + a\rho_I]\,F = (s+a+d+\rho_I)\,F_0 - \rho_I\,P_0$

Using partial fractions, we obtain

4.7 $F = [(r_2\,F_0 - \rho_I\,P_0) / (s - r_1) - (r_1\,F_0 - \rho_I\,P_0) / (s - r_2)] / (\,r_2 - r_1\,)$

with r_1 & r_2 ($r_1 \geq r_2$) the roots of the quadratic equation

4.8 $s^2 + (d+a+\rho_I)\,s + \rho_I\,a = 0$

Notice that $a+d+\rho_I$ is greater than $a+\rho_I > 0$ and that $\rho_I\, a > 0$, which implies that both roots are real and negative.

This implies that

4.9 $F(t) = [(r_2 F_0 – \rho_I P_0) e^{r1\, t} – (r_1 F_0 – \rho_I P_0) e^{r2\, t}] / (r_2 – r_1)$

Now, solving for P

4.10 $[s^2 + (a+d+\rho_I) s+ \rho_I a] P = a F_0 + s P_0$

Therefore,

4.11 $P = [-(r_1 P_0 + r F_0) / (s – r_1) + (r_2 P_0 + r F_0) / (s – r_2)] / (r_2 – r_1)$

This yields

4.12 $P(t) = [-(r_1 P_0 + r F_0) e^{r1\, t} + (r_2 P_0 + r F_0) e^{r2\, t}] / (r_2 – r_1)$

Evaluating these equations at time τ and noting that $I(t) = L_0 – F(t)$ where L_0 is the initial value for $L(t)$, we obtain

4.13 $F_\tau = [(r_2 F_0 – \rho_I P_0) e^{r1\, \tau} – (r_1 F_0 – \rho_I P_0) e^{r2\, \tau}] / (r_2 – r_1)$

4.14 $P_\tau = [(a F_0 + r_2 P_0) e^{r2\, \tau} – (a F_0 + r_1 P_0) e^{r1\, \tau}] / (r_2 – r_1)$

4.15 $I_\tau = L_0 - F_\tau$

where the notation X_τ represents $X(t)$ at time τ.

If we assume that at the beginning of the observation period no complexes have formed, then $P_0 = 0$ and $F_0 = L_0$. If we further assume that the assembly and/or the dissociation of the complex occurs much faster than commitment, so that $d+a >> \rho_I$, we obtain

4.16 $r_2 \approx -(d+a)$,

4.17 $r_1 \approx -\rho_I\, a / (d+a)$ and

4.18 $r_2 - r_1 \approx -(d+a)$.

Defining p_I as

4.19 $p_I = \rho_I \;/ (1+d/a)$

we get

4.20 $F_\tau \approx L_0\, e^{-pI\,\tau}$

Therefore the system can now be approximated by the state diagram shown in Fig 4.1B.

Second time period

For times starting at time τ the molecules can consider splicing the upstream exon to the downstream exon; i.e., skipping the exon of interest (Fig. 4.1C and E). To minimize the complexity of notation below, we define a new reference time t' that sets time τ to zero: $t' = t - \tau$. From the state diagram shown in Fig. 4.1C, the following equations are obtained for $t' > 0$:

4.21 $dL/dt' = d\,P + d'\,b - (a+a')\,L$

4.22 $dP/dt' = a\,L + d'\,B - (d+a'+\rho_I)\,P$

4.23 $db/dt' = a'\,L + d\,B - (d'+a+\rho_S)\,b$

4.24 $dB/dt' = a'\,P + a\,b - (d+d'+\rho_I+\rho_S)\,P$

4.25 $dI/dt' = \rho_I\,(P+B)$

4.26 $dS/dt' = \rho_S\,(b+B)$

The Laplace transform of the first four equations, indicated as $X = X(s)$ for any function $X(t')$, was taken yielding

4.27 $sL - L_\tau = d\,P + d'\,b - (a+a')\,L$

4.28 $sP - P_\tau = a\,L + d'\,B - (d+a'+\rho_I)\,P$

4.29 $sb - b_\tau = a'\,L + d\,B - (d'+a+\rho_S)\,b$

4.30 $sB - B_\tau = a'\,P + a\,b - (d+d'+\rho_I+\rho_S)\,B$

where the notation X_τ represents X(t') at t' = 0 (i.e., at t = τ).

Although we are most interested in the probability of exon inclusion, it is easier to calculate S, and its final expression actually provides more insight into the roles of the different parameters. I becomes simply all the tagged molecules not included in S. Therefore we will focus on an expression for S_∞. Let's define Б = b+B. Notice that according to the previous assumptions, the value of $S(t') = 0$ for $t' \leq 0$ and, according to the final value theorem and equation 4.26, as $t' \rightarrow \infty$, $S(t')$ approaches $S_\infty = \rho_S \lim_{s\rightarrow 0} Б(s) = \rho_S\, Б_0$, where the notation X_0 represents $X(s)$ evaluated at $s = 0$. Substituting L = F – P and b = Б – B, and assuming no tagged molecules contain the second complex at t' = 0, $b_\tau = B_\tau = 0$ which implies $Б_\tau = 0$, we obtain

4.31 $s\,(F - P) - F_\tau + P_\tau = d\,P + d'\,(Б - B) - (a+a')\,(F - P)$

4.32 $sP - P_\tau = a\,(F - P) + d'\,B - (d+a'+\rho_I)\,P$

4.33 $s\,(Б - B) = a'\,(F - P) + d\,B - (d'+a+\rho_S)\,(Б - B)$

4.34 $sB = a'P + a\,(Б - B) - (d+d'+\rho_I+\rho_S)\,B$

Taking s = 0, the equations become

4.35 $(a+a')\,F_0 = (d+a+a')\,P_0 + d'\,\textit{Б}_0 - d'\,B_0 + F_\tau - P_\tau$

4.36 $(d+a+a'+\rho_I)\,P_0 = a\,F_0 + d'\,B_0 + P_\tau$

4.37 $(d'+a+\rho_S)\,\textit{Б}_0 = a'\,F_0 + (d+d'+a+\rho_S)\,B_0 - a'\,P_0$

4.38 $(d+d'+a+\rho_I+\rho_S)\,B_0 = a'\,P_0 + a\,\textit{Б}_0$

Substituting F_0 from equation 4.36 into equations 4.35 and 4.37, we get

4.39 $[a'\,(d+a+a'+\rho_I) + a\,\rho_I]\,P_0 = d'\,a\,\textit{Б}_0 + d'\,a'\,B_0 + a\,F_\tau + a'\,P_\tau$

4.40 $(d'+a+\rho_S)\,a\,\textit{Б}_0 = (d+a'+\rho_I)\,a'\,P_0 + [(d+d'+a+\rho_S)\,a - d'\,a']\,B_0 - a'\,P_\tau$

Substituting P_0 from equation 4.38 into these equations, they become

4.41 $\{[a'\,(d+a+a'+\rho_I) + a\,\rho_I]\,(d+d'+a+\rho_I+\rho_S) - d'\,a'^2\}fc\,B_0 = [a'\,(d+d'+a+a'+\rho_I) + a\,\rho_I]\,a\,\textit{Б}_0 + a\,a'\,F_\tau + a'^2\,P_\tau$

4.42 $(d+d'+a+a'+\rho_I+\rho_S)\,a\,\textit{Б}_0 = [(d+d'+a+\rho_I+\rho_S)\,(d+a'+\rho_I) + (d+d'+a+\rho_S)\,a - d'\,a']\,B_0 - a'\,P_\tau$

Taking $\alpha = d+d'+a+a'$ and $\beta = \alpha\,(d+a) + (\alpha+d)\,\rho_I + (d+a+a')\,\rho_S + (\rho_I+\rho_S)\,\rho_I$, these equations simplify to

4.43 $\{\beta\,a' + a\,[\alpha\,\rho_I+(\rho_I + \rho_S)\,\rho_I]\}\,B_0 = [\alpha\,a' + (a+a')\,\rho_I]\,a\,\textit{Б}_0 + a\,a'\,F_\tau + a'^2\,P_\tau$

4.44 $(\alpha+\rho_I+\rho_S)\,a\,\textit{Б}_0 + a'\,P_\tau = \beta\,B_0$

Substituting B_0 from equation 4.44 into equation 4.43, taking $\gamma = \alpha+\rho_I+\rho_S$ and substituting $S_\infty = \rho_S\,\textit{Б}_0$ in the final equation, we get

4.45 $\{\alpha\ [(d'+a')\ a\rho_I + (d+a)\ a'\rho_S] + (a+a')\ (a\rho_I^2+\gamma\rho_I\rho_S+a'\rho_S^2) + (d'a\rho_I+da'\rho_S)\ (\rho_I+\rho_S)\}\ S_\infty = a'\rho_S\ [\beta F_\tau - \gamma\rho_I P_\tau]$

This, along with equations 4.13 and 4.14, provide the general solution for S_∞. However a more useful expression can be obtained if we assume that assembly and/or dissociation of the complexes on both exons occur much faster than commitment for either pathway: i.e., $d+a >> \rho_I$, $d+a >> \rho_S$, $d'+a' >> \rho_I$ and $d'+a' >> \rho_{S.}$ This approximation yields:

4.46 $S_\infty \approx L_0\ e^{-pI\,\tau}\ (d+a)\ a'\rho_S\ /\ [(d'+a')\ a\rho_I + (d+a)\ a'\rho_S]$

Using p_I as defined previously and defining p_S analogously as

4.47 $p_S = \rho_S\ /\ (1 + d'\ /\ a')$

yields

4.48 $S_\infty \approx L_0\ e^{-pI\,\tau}\ p_S/(p_S+p_I)$

This situation can be summarized with the state diagram shown in Fig. 4.1D for $t > \tau$, with the initial condition $L_\tau = L_0\ e^{-pI\,\tau}$. To model the system at all times requires only three constants, namely τ, p_I and p_S (see Fig. 4.1B and 4.1D).

If the rates of degradation of the included and skipped molecules are similar, equation 4.48 provides approximations for the fraction of skipped and, by subtraction, of included untagged molecules at steady state. The form of equation 4.48 lends itself to intuitive interpretation, and the focus on S provided insight into the roles of the different parameters (see below). The exponential decay term describes the commitment to inclusion during the pre-τ interval: molecules no longer available for skipping. The remaining fraction reflects the competition after time τ between inclusion and skipping among those molecules capable of either. At this point the model predicts splicing outcomes in terms of an unspecified exon definition complex and of the ratios of rate constants p_I and p_S. We now turn to relating these terms to biophysical processes and to use the resulting model to predict psi values.

Modeling the DEs

Equation 4.48 should be applicable to the definition of any internal exon. In the case of natural exons there are many factors that could be in play and that are poorly understood. For instance, protein-protein interactions and pre-mRNA secondary or tertiary structure could well determine ρ, a, d and/or τ. We did not consider such factors in applying this model to DEs, which represent a simplified framework for testing the validity of the model and for building more refined versions.

In order to apply equation 4.48 to the DE data, we needed to model τ, p_I and p_S. We consider τ and p_S to be constant for all DEs used, τ depending on the transcription time and p_S depending on the downstream exon. Thus we are left with p_I, which is ρ_I / (1+d/a). A physical model for ρ_I is challenging, as this term describes the conversion of an initial complex to a commitment complex. It is not yet understood what commitment entails or how it is achieved. We therefore decided to focus on the formation of the initial complex itself, asking whether the effect of exon size, ESEs and ESSs on its formation (a/d) can explain our data. That is, we assume that ρ_I, the rate constant for the conversion of an exon with an assembled complex to a committed exon, remains constant with respect to these 3 parameters. Equation 4.48 can be rewritten as equation 4.49, which combines those terms that are not resolvable by the experiments we carried out and which serves as the proving ground for fitting the data to the model:

4.49 $\text{pso} \approx 100\, e^{-T/(1+D)}/(1+C/(1+D))$

where pso denotes percent spliced out (i.e., skipped), $T=\rho_I\, \tau$, $C=\rho_I/p_S$ and $D=d/a$. We then focused on how all the different DE configurations affect D, the ratio of the disassociation and assembly rate constants of the initial complex, while T and C were taken to be constant.

We first sought an expression relating size and D, modeling the formation of an exon-spanning complex. There is evidence for indirect interactions between the macromolecules at the two ends of an exon (Wu and Maniatis 1993), and the motifs present in the intermediate protein involved are present in subunits at the two ends, opening the possibility for an RNA-binding activated direct interaction. We reasoned that in the simplest case, the formation of this complex is proportional to the probability of the two tethered ends of the exon having undergone a productive collision, which occurs when both ends of the exon are suitably occupied and they approach each other in the correct orientation through thermal movements. The ends will then be at a fitting distance, y_i from each other as shown in Fig. 4.2A for the direct interaction and analogously for the indirect one. Assuming the RNA behaves as a worm-like chain with contour length much greater than persistence length, the probability for a given end-to-end distance as a function of exon size can be obtained using a Gaussian approximation (Becker et al. 2010). Using this approximation, the ends of the molecule while inside the range of distances within which attractive and repulsive forces become important can be modeled. Taking this range to be small with respect to the fitting distance, y_i, and applying the mean-value theorem for integrals, the collision probability can be estimated with the formula

4.50 $P(Y_i,x) \approx k_i\, Y_i^2\, x^{-3/2}\, e^{-3/2\, Y_i^2/x}$

Here x is contour length of the exon in nm and is determined by the exon size in nucleotides (see

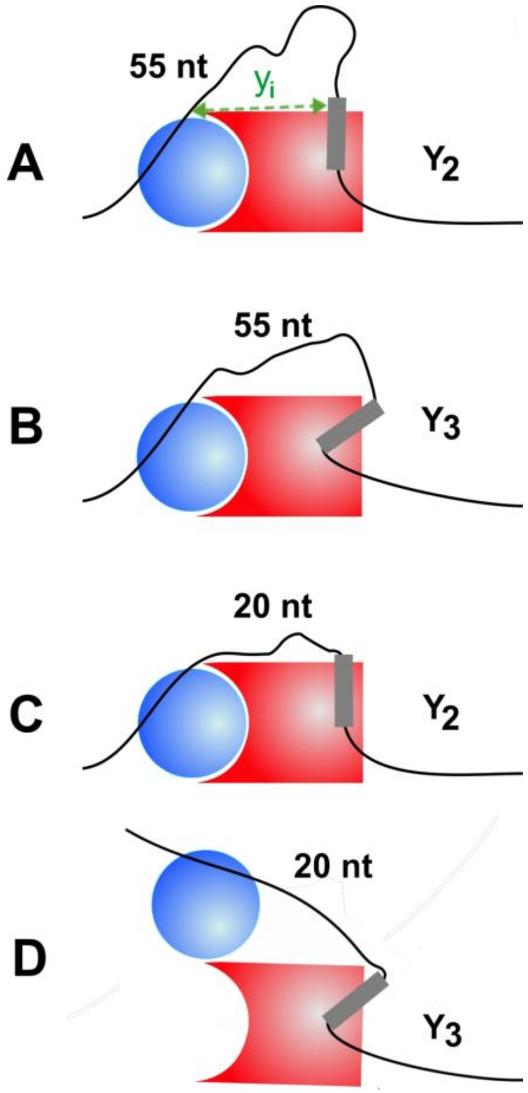


Figure 4.2. Modeling exon end-to-end contact in exon definition. RNA molecules are bound by U2AF proteins (blue circles) at the 3’ SS and U1 snRNPs (red rectangles with semicircular sockets) at the 5’ SS (gray rectangle). A. After a collision between the two ends of the exon, the exon definition complex forms. Communication between these two ends is mediated by protein-protein interaction. The arrow line indicates the fitting distance y_i, which is the distance between the outermost points in the exon that are unconstrained by protein binding. Here, SS Set 2 is modeled (i=2). B. A change in the point or angle at which the pre-mRNA extends from a binding protein can increase y_i and consequently the minimum exon size that allows proper protein-protein contact. Here, SS Set 3 is modeled. C. Same as A but with a shorter exon. D. Same as B but showing the impaired recognition of a short exon.

equation 4.55), the index i refers to the splice sites used (4 sets, Table 3.1 in Chapter 3: sets 2, 3, 5 and 7), Y_i is the distance y_i divided by the square root of the Kuhn length for an RNA molecule, which is a measure of stiffness and assuming a cationic concentration equivalent to ~300 mM should be approximately ~3.0 nm (Chen et al. 2012). The catch-all constant k_i depends on the Kuhn length, the range of distances within which attractive and repulsive forces become important and the chance that a collision will result in an association and is independent of the length of the exon in question. Although the values of these parameters are unknown, we consider them as constant for any set of splice sites. D is inversely proportional to a, and so will be inversely proportional to $P(Y_i,x)$.

As shown in Fig. 4.2B, different SS sets are allowed to have different geometries in their interaction. Therefore, a different fitting distance can be used for each SS set. For example, in Fig. 4.2B the fitting distance for SS Set 3 is modeled as being greater than that for SS Set 2 due to differences in the geometry of the interaction between U1 snRNP and the 5’ SS of the exon of interest. Fig. 4.2C and D show the effects of these fitting distances on short exons. The fitting distance for SS Set 2 allows the interaction to occur in exons as short as 30 nt. However, the fitting distance and the geometry required for SS Set 3 make the interaction between the two ends impossible for these short exons. As will be seen in chapter 5, this difference in the fitting distance is enough to explain the shift observed in Fig. 3.2 between the curves relating psi and size for different SS sets.

Due to the ability of ESE-binding activators to interact with proteins at either end of the exon, the RNA itself and other SR proteins, we modeled the effect of enhancers by assuming that they act by increasing the stability of the exon definition complex. In this case, the rate of dissociation should be proportional to the rate at which random collisions transfer kinetic energy greater than a threshold, $E_{threshold}$, to the complex. The addition of a single ESE was taken to increase this energy threshold by a fixed constant amount ΔE. Any additional ESE will increase this energy threshold by an additional ΔE.

For a simplified analysis, we considered the collision between the complex on the exon of interest, C, and a molecule, M. This collision transfers enough kinetic energy to cause dissociation of C if the collision is head-on and the relative kinetic energy of M is higher than a threshold. However, if the collision is not head-on, then the geometry of the collision should be taken into account. As an approximation, C and M were modeled as spheres; the angles between the collision trajectory and the tangent plane at the site of contact determine the energy that is transferred. An analogous situation is found when modeling reactive encounters (Atkins and de Paula 2002): following a traditional analysis of such situations, an equation for the rate of dissociation d was obtained

4.51 $d \approx W\, e^{-Ethreshold/(kT)}$

where W is a proportionality constant that takes into account all speeds and collision angles, k is the Boltzmann constant, T is the absolute temperature and $E_{threshold}$ is the energy necessary to cause dissociation of C. If the complex is modified so that the energy required to cause its dissociation becomes $E'_{threshold} = E_{threshold} + E_{enh}$, then the new dissociation constant, d_E, becomes

4.52 $d_E \approx W\, e^{-(Ethreshold+Eenh)/(kT)} = W\, e^{-Ethreshold/(kT)}\, e^{-Eenh/(kT)} = d\, e^{-Eenh/(kT)}$

The addition of a single enhancer, then, modified the dissociation rate by a factor of $c_E = e^{-Eenh/(kT)} < 1$, and

4.53 $d_E \approx d\, c_E$

Repeating the analysis to account for the addition of n identical enhancers yielded

4.54 $d_n \approx d\, c_E^n$

(These results could be generalized by making $c_E = \gamma\, e^{-Eenh/(kT)}$, which allows γ to account for other parameters such as occupancy.)

Consequently, each ESE affects D by the factor c_E and n of those sequences affect it by c_E^n. To be consistent with the results observed for the ESE under consideration, this effect was taken to be independent of position. In this simple scenario, multiple enhancers were modeled as independent, leading to an exponential dependence of D on the number of enhancers present.

A similar approach was taken for modeling the ESSs, which are considered to be disruptive to the complex and therefore decrease its stability. Since the ESS used showed a position-dependent effect, we divided the ESSs into 3 categories based on their position: first, last and remaining intermediate positions. As in the case of the ESEs, multiple ESSs were modeled as independent of each other.

The effect of the reference sequences on stability also had to be considered, for it is unknown if they should be modeled as enhancers, silencers or something else. However, since the effect of replacing reference sequences with ESEs was shown to be position-independent, the effect of individual reference sequences should also be position-independent. Extending the analogy with ESEs and ESSs, multiple reference sequences in a single exon were modeled as independent.

Taking all of this into account and modeling these effects as independent of each other gave the following approximation for D in equation 4.49:

$$4.55 \quad D \approx K_i\ Y_i^{-2}\ c_E^{nE}\ c_R^{nR}\ c_F^{nF}\ c_L^{nL}\ c_I^{nI}\ Z^{3/2}\ e^{3Yi^2/Z}$$

where Z is the size of the DE in nucleotides figuring 2 nt/nm (Chen et al. 2012), Y_i is $y_i/\sqrt{\text{KuhnLength}}$, n_I is the number of non-terminal ESSs, n_F and n_L are 1 if the first or last position, respectively, is occupied by an ESS and 0 otherwise, n_R is the total number of reference sequences present, and n_E is the number of ESEs. The c constants represent destabilization coefficients for the ESSs (c_F, c_L, c_I), reference sequences (c_R) and ESEs (c_E). K_i is a constant that combines all the constants generated by each of the individual terms; the index i refers to the set of splice sites present.

Discussion

The derivation of the commitment progression equation, which assumed no mechanisms for formation or dissociation of the exon definition complex, was based on simple assumptions. The physical model is centered on exon definition as a decisive step in the recognition of most splice sites and assumes that this step requires the formation of an RNA-protein complex on the exon of interest. A general solution was found but its complexity precluded intuitive analyses and interpretations. For this reason, simplifying assumptions were made and a simpler and more intuitive equation was obtained. This equation had two components: one derived from the period in which skipping is not an option yet and the other when the two fates, inclusion and skipping, are available. Interestingly, using this equation in Chapter 3 suggested that the second time period could be safely ignored making F_τ in equation 4.23 or its approximation in equation 4.20 an appropriate estimate for S_∞.

The formation of exon definition complex was modeled as dependent on an indirect physical interaction across the exon of its ends. The mathematical treatment of this scenario was fairly intuitive and included treating the RNA molecule as a polyelectrolyte. The dissociation of this complex was modeled based on collisions with random molecules. The effects of ESEs were incorporated as providing stability to the complex against these collisions. All aspects that went into these equations were combined by assuming independence of the different actions which might be an oversimplification but worked well here (Chapter 3).

Importantly, an emphasis was placed on mechanisms that would affect both DEs and natural exons alike. The mechanisms proposed align well with published observations as described in Chapter 3. Both the general commitment progression equation as well as the mechanistic equations contributed to this.

Some published mechanisms were not considered and might be attempted later. One of them, the recruitment model for ESE action, was briefly addressed in Chapter 3 with only limited success in predicting splicing outcomes. Another one is the scanning hypothesis for exon/intron recognition (Kuhne et al. 1983; Lang and Spritz 1983; Borensztajn et al. 2006). In the exon definition variant of this hypothesis, putative machinery, upon finding one end of an exon, scans the RNA molecule searching for the other end. When it finds it according to some criteria, an exon is defined. This mechanism was not explored because the tethered exon ends collision hypothesis presented here is more intuitive, does not require cellular machinery for which no evidence exists, does not require as many exceptions when analyzing exons with multiple splice site options, and provides good predictions for the observations available.

Authors’ Contributions

MA conceived and performed all the calculations and analysis described. LC and MA wrote the text.

Acknowledgements

Special thanks to Kyle Jurado for reviewing the detailed derivation of equation 4.48. Thanks to the Chasin Lab for useful discussions.

References

Arias MA, Lubkin AL, Chasin LA. 2013. Splicing of designer exons informs a biophysical model for exon definition. *Manuscript submitted*.

Atkins P, de Paula J. 2002. *Physical Chemistry*. W. H. Freeman and Company, New York.

Becker NB, Rosa A, Everaers R. 2010. The radial distribution function of worm-like chains. *Eur Phys J E Soft Matter* **32**(1): 53-69.

Berget SM. 1995. Exon recognition in vertebrate splicing. *J Biol Chem* **270**(6): 2411-2414.

Black DL. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72**: 291-336.

Borensztajn K, Sobrier ML, Duquesnoy P, Fischer AM, Tapon-Bretaudiere J, Amselem S. 2006. Oriented scanning is the leading mechanism underlying 5' splice site selection in mammals. *PLoS Genet* **2**(9): e138.

Chen H, Meisburger SP, Pabit SA, Sutton JL, Webb WW, Pollack L. 2012. Ionic strength-dependent persistence lengths of single-stranded RNA and DNA. *Proceedings of the National Academy of Sciences of the United States of America* **109**(3): 799-804.

Chen M, Manley JL. 2009. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10**(11): 741-754.

De Conti L, Baralle M, Buratti E. 2013. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA* **4**(1): 49-60.

de la Mata M, Alonso CR, Kadener S, Fededa JP, Blaustein M, Pelisch F, Cramer P, Bentley D, Kornblihtt AR. 2003. A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell* **12**(2): 525-532.

Fox-Walsh KL, Dou Y, Lam BJ, Hung SP, Baldi PF, Hertel KJ. 2005. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proceedings of the National Academy of Sciences of the United States of America* **102**(45): 16176-16181.

Fox-Walsh KL, Hertel KJ. 2009. Splice-site pairing is an intrinsically high fidelity process. *Proceedings of the National Academy of Sciences of the United States of America* **106**(6): 1766-1771.

Hicks MJ, Lam BJ, Hertel KJ. 2005. Analyzing mechanisms of alternative pre-mRNA splicing using in vitro splicing assays. *Methods* **37**(4): 306-313.

House AE, Lynch KW. 2006. An exonic splicing silencer represses spliceosome assembly after ATP-dependent exon recognition. *Nat Struct Mol Biol* **13**(10): 937-944.

Kessler O, Jiang Y, Chasin LA. 1993. Order of intron removal during splicing of endogenous adenine phosphoribosyltransferase and dihydrofolate reductase pre-mRNA. *Mol Cell Biol* **13**(10): 6211-6222.

Krainer AR, Maniatis T, Ruskin B, Green MR. 1984. Normal and mutant human beta-globin pre-mRNAs are faithfully and efficiently spliced in vitro. *Cell* **36**(4): 993-1005.

Kuhne T, Wieringa B, Reiser J, Weissmann C. 1983. Evidence against a scanning model of RNA splicing. *EMBO J* **2**(5): 727-733.

Lang KM, Spritz RA. 1983. RNA splice site selection: evidence for a 5' leads to 3' scanning model. *Science* **220**(4604): 1351-1355.

Lazarev D, Manley JL. 2007. Concurrent splicing and transcription are not sufficient to enhance splicing efficiency. *RNA* **13**(9): 1546-1557.

Lim SR, Hertel KJ. 2004. Commitment to splice site pairing coincides with A complex formation. *Mol Cell* **15**(3): 477-483.

Mount SM. 1982. A catalogue of splice junction sequences. *Nucleic Acids Res* **10**(2): 459-472.

Osheim YN, Miller OL, Jr., Beyer AL. 1985. RNP particles at splice junction sequences on Drosophila chorion transcripts. *Cell* **43**(1): 143-151.

-. 1988. Visualization of Drosophila melanogaster chorion genes undergoing amplification. *Mol Cell Biol* **8**(7): 2811-2821.

Singh J, Padgett RA. 2009. Rates of in situ transcription and splicing in large human genes. *Nat Struct Mol Biol* **16**(11): 1128-1133.

Wada Y, Ohta Y, Xu M, Tsutsumi S, Minami T, Inoue K, Komura D, Kitakami J, Oshida N, Papantonis A et al. 2009. A wave of nascent transcription on activated human genes. *Proceedings of the National Academy of Sciences of the United States of America* **106**(43): 18357-18361.

Wu JY, Maniatis T. 1993. Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell* **75**(6): 1061-1070.

Chapter 5

This chapter is part of a manuscript submitted for publication (Arias et al. 2013)

Splicing of Designer Exons Informs a Biophysical Model for Exon Definition

Introduction

Transcription generates pre-mRNA molecules that are then processed to produce mRNA. Modifications of the pre-mRNA molecule include capping, cleavage, polyadenylation, and splicing. The latter refers to the removal of usually long stretches of RNA designated introns yielding a concatenation of the flanking sequences designated exons in mRNA molecules. This process occurs with great fidelity and therefore requires precise definitions of the sequences to be removed and/or the sequences to be kept. Early on, the sequences at the boundaries between exons and introns were identified as having a fundamental role in this process (Mount 1982). However, these sequences proved insufficient to define the exons/introns of transcrpts (Sun and Chasin 2000).

For studying the early recognition of splice sites, two alternative models have been put forth (De Conti et al. 2013). In the first model, intron definition, the introns are recognized and removed; the exons are joined as a byproduct. In the second model, exon definition, the exons are recognized and joined to one another. This requires a subsequent intron definition for the ends of each intron must be paired. In this study we focus on exon definition and in particular on exon recognition. It has been suggested that this paradigm applies to short exons, smaller than about 250 nt flanked by long introns, longer than about 250 nt, a category that includes most exons present in humans (Fox-Walsh et al. 2005).

As the pre-mRNA is being synthesized, some of the events involved in splicing can be taking place. Indeed, there is evidence that in some cases the introns are removed before transcription finishes (Osheim et al. 1985; Osheim et al. 1988; Singh and Padgett 2009; Wada et al. 2009). More importantly, critical events such as exon recognition should then be taking place as the pre-mRNA is being synthesized. This possibility is substantiated by the finding that slowing down the RNA polymerase II affects splicing decisions (Roberts et al. 1998; de la Mata et al. 2003; Munoz et al. 2009). These observations have led to the kinetic model of splicing (Dujardin et al. 2013). For these reasons co-transcriptional splicing was incorporated in a model for splicing presented in Chapter 4.

Exon recognition involves binding of U1 to the 5' SS, which is the earliest event characterized in *in vitro* reactions (Hoskins et al. 2011). However, there are reports that on the upstream end of exons U2AF65 binds shortly after the 3' SS is synthesized (Ujvari and Luse 2004), making it likely that this event precedes U1 snRNP binding *in vivo*. These two ends of the exon are therefore recognized early. They also provide near equal contributions toward efficient exon recognition(Shepard et al. 2011). One possibility for this symmetry is a direct or indirect interaction. Evidence for interactions of proteins at both ends of the exon and SC25, an SR protein, was presented by Wu and Maniatis (Wu and Maniatis 1993). Therefore it was proposed in Chapter 4 that this interaction determines the effect of size on inclusion level and an appropriate model was presented.

Sequences in addition to the ones found at the ends of exons were identified early on as having an important role in exon recognition (Reed and Maniatis 1986; Mardon et al. 1987; Cooper and Ordahl 1989; Tsai et al. 1989). The sequences in the exons themselves were named exonic splicing enhancers (ESEs) when they had a positive effect on inclusion or exonic splicing silencers (ESSs) when the effect was negative. Shortly afterwards, proteins binding some of these sequences were found (Ge and Manley 1990; Krainer et al. 1990). It has been hypothesized that these and similar proteins act by increasing the likelihood that U1 snRNP or U2AF would bind to their respective sequences (Fu 1995). Further experiments have shown inadequacies in the model by requiring that U1 snRNP be recruited with U2AF (Lam and Hertel 2002). These results are easily reconciled with the ESEs having a stabilization effect on the exon definition complex. Networks of interactions including these proteins and components of the early splicing machinery have been shown (Hoffman and Grabowski 1992; Wu and Maniatis 1993). Importantly, ESEs can have further effects in downstream reactions in the splicing process. However, this chapter focuses exclusively on exon recognition. An appropriate model incorporating these ideas was developed in Chapter 4.

In this chapter the performance of the model developed in Chapter 4 is evaluated using the data obtained for single-perturbation experiments in Chapter 3 to gather the information necessary to use it. After finding appropriate constants for the different parameters in the model, the performance for the single-perturbation data is evaluated to assess the correctness of the procedure to obtain the paramenters. Finally, the model is used to predict the inclusion levels of the designer exons composed of ESE/ESS combinations.

Results

A biophysical model to explain splicing decisions

We embarked on the design of our own exons so as to be able to examine individual parameters that govern splicing decisions. While this reductionist approach dispenses with the complexity of natural exons, it has the advantage of making fundamental principles discernible. Having varied parameters in over 150 DEs, we next sought to develop a biophysical model that could explain these data. The goal of this model is to relate the observed psi to the parameters that have been varied in these DEs.

The physical model is centered on exon definition as a decisive step in the recognition of most splice sites and assumes that this step requires the formation of an RNA-protein complex on the exon of interest. The number of pre-mRNA molecules with a complex is determined by the balance between assembly and disassembly, which can be described by overall association and dissociation rate constants. Once assembled, complexed molecules can then proceed to a state of commitment to exon inclusion (Fig. 4.1A).

We start with a set of assumptions that are described in Chapter 4, specifically in Box 4.1, and focus on a group of pre-mRNA molecules (conceptually “tagged”) that are all in the same state of synthesis. To consider the choice between inclusion and skipping, it seems necessary that the exon downstream be synthesized; we define time τ as the time required to make this downstream exon available for splicing. For times prior to τ, there are 3 types of pre-mRNA molecules with respect to the exon of interest: naked, L; complexed, P; committed to inclusion, I (Fig. 4.1A). A set of differential equations relates the number of tagged P, I and L molecules starting at $t = 0$:

3.2. $dL/dt = d\,P - a\,L$

3.3. $dP/dt = a\,L - (d+\rho_I)\,P$

3.4. $dI/dt = \rho_I\,P$

It is a cohort of previously tagged molecules that is being followed, so rates of synthesis need not be considered.

For times starting at time τ the molecules can consider splicing the upstream exon to the downstream exon; i.e., skipping the exon of interest (Fig. 4.1C and E). A set of differential equations, analogous to the set for the pre-τ period, describes this situation (see Chapter 4). Although we are most interested in the probability of exon inclusion, it is easier to calculate the probability of exon skipping. The general solutions as well as some approximations and intermediate results are presented in Chapter 4. Equation 5.4 describes the fraction of tagged molecules that skip the exon:

3.5. $S_\infty / L_0 \approx e^{-pI\,\tau}\, p_S/(p_S+p_I)$

where L_0 represents the total number of molecules that were initially tagged, S_∞ is the final number of skipped molecules, $p_I = 1/(1+d/a)$ and $p_S = 1/(1+d'/a')$, and a, d, a’, and d’ can be seen in Fig. 4.1.

If the rates of degradation of the included and skipped molecules are similar, equation 5.4 provides approximations for the fraction of skipped and, by subtraction, of included untagged molecules at steady state. The form of equation 5.4 lends itself to intuitive interpretation, and the focus on S provided insight into the roles of the different parameters (see below). The exponential decay term describes the commitment to inclusion during the pre-τ interval: molecules no longer available for skipping. The remaining fraction reflects the competition after time τ between inclusion and skipping among those molecules capable of either. At this point the model predicts splicing outcomes in terms of an unspecified exon definition complex and of the ratios of rate constants p_I and p_S. We now turn to relating these terms to biophysical processes and to use the resulting model to predict psi values.

Modeling the DEs

Equation 5.4 should be applicable to the definition of any internal exon. In the case of natural exons there are many factors that could be in play and that are poorly understood. For instance, protein-protein interactions and pre-mRNA secondary or tertiary structure could well determine ρ, a, d and/or τ. We did not consider such factors in applying this model to DEs, which represent a simplified framework for testing the validity of the model and for building more refined versions.

In order to apply equation 5.4 to the DE data, we needed to model τ, p_I and p_S. We consider τ and p_S to be constant for all DEs used, τ depending on the transcription time and p_S depending on the downstream exon. Thus we are left with p_I, which is ρ_I / (1+d/a). A physical model for ρ_I is challenging, as this term describes the conversion of an initial complex to a commitment complex. It is not yet understood what commitment entails or how it is achieved. We therefore decided to focus on the formation of the initial complex itself, asking whether the effect of exon size, ESEs and ESSs on its formation (a/d) can explain our data. That is, we assume that ρ_I, the rate constant for the conversion of an exon with an assembled complex to a committed exon, remains constant with respect to these 3 parameters. Equation 5.4 can be rewritten as equation 5.5, which combines those terms that are not resolvable by the experiments we carried out and which serves as the proving ground for fitting the data to the model:

3.6. $pso \approx 100\, e^{-T/(1+D)}/(1+C/(1+D))$

where pso denotes percent spliced out (i.e., skipped), $T=\rho_I\, \tau$, $C=\rho_I/p_S$ and $D=d/a$. We then focused on how all the different DE configurations affect D, the ratio of the disassociation and assembly rate constants of the initial complex, while T and C were taken to be constant.

We first sought an expression relating size and D, modeling the formation of an exon-spanning complex. We reasoned that in the simplest case, the formation of this complex is proportional to the probability of the two tethered ends of the exon having undergone a productive collision, which occurs when both ends of the exon are suitably occupied and they approach each other in the correct orientation through thermal movements. The ends will then be at a fitting distance, y_i from each other as shown in Fig. 4.2A. See Chapter 4 for the translation of these ideas and those below into equation 5.6.

We modeled the effect of enhancers by assuming that they act by increasing the stability of the exon definition complex. Note that this choice is in contradistinction to other possibilities such as recruitment or catalysis. In this simple scenario, multiple enhancers are modeled as independent, leading to an exponential dependence of D on the number of enhancers present. A similar approach was taken for modeling the ESSs, which are considered to be disruptive to the complex and therefore decrease its stability. Since the ESS used showed a position-dependent effect, we divided the ESSs into 3 categories based on their position: first, last and remaining intermediate positions. As in the case of the ESEs, multiple ESSs were modeled as independent of each other.

The effect of the reference sequences on stability also had to be considered, for it is unknown if they should be modeled as enhancers, silencers or something else. However, since the effect of replacing reference sequences with ESEs was shown to be position-independent, the effect of individual reference sequences should also be position-independent. Extending the analogy with ESEs and ESSs, multiple reference sequences in a single exon were modeled as independent.

Taking all of this into account and modeling these effects as independent of each other gave the following approximation for D in equation 5.5:

3.7. $D \approx K_i\, Y_i^{-2}\, c_E^{nE}\, c_R^{nR}\, c_F^{nF}\, c_L^{nL}\, c_I^{nI}\, Z^{3/2}\, e^{3Yi^2/Z}$

where Z is the size of the DE in nucleotides figuring 2 nt/nm (Chen et al. 2012), Y_i is $y_i/\sqrt{}$KuhnLength, n_I is the number of non-terminal ESSs, n_F and n_L are 1 if the first or last position, respectively, is occupied by an ESS and 0 otherwise, n_R is the total number of reference sequences present, and n_E is the number of ESEs. The c constants represent destabilization coefficients for the ESSs (c_F, c_L, c_I), reference sequences (c_R) and ESEs (c_E). K_i is a constant that combines all the constants generated by each of the individual terms; the index i refers to the set of splice sites present.

To optimize the values for K_i, y_i, and the c constants in equation 5.6 we used the BFGS algorithm for minimizing the sum of the squared differences between predicted and observed pso values. Due to the lack of size perturbation data for some of the splice site sets, it was necessary to assume the parameters y_5 and y_7, which are related to the distance between the two ends of the exon in the complex, to be equal to a discoverable y; y_3 was chosen and turned out to be appropriate (see below). The data used for optimization are described fully in Materials and Methods and shown in Supplemental Table S3.1. The parameter set that emerged is shown in Table 5.1.

Table 5.1. Best fit for parameters in equations 5.5 and 5.6

Parameter	Value	Parameter	Value
T	5.24	c_E	0.611
C	22.5×10^{-6}	c_R	1.48
K_2	1.36×10^{-5}	c_F	1.57
K_3	1.70×10^{-4}	c_L	3.04
K_5	4.76×10^{-4}	c_I	2.26
K_7	3.36×10^{-3}	y_2	21.6
		y_3	12.0

Testing the model

One test for this model is how well the equation fits the data that was used for its optimization. That is, can this model, built on biophysical principles, fit the data; and if so how good is the fit? Using the parameter set in Table 5.1 and equations 5.5 and 5.6, the 112 data points of single-parameter perturbation data (size, ESE number and position, ESS number and position) were predicted very accurately (R^2 = 0.94, Table 5.2 and Fig. 5.1). Importantly, the slope of the fit was 1.04, very close to the expected 1.00; and the intercept for observed psi was - 0.96%, again very close to the expected zero. A visual way to assess the accuracy achieved is to

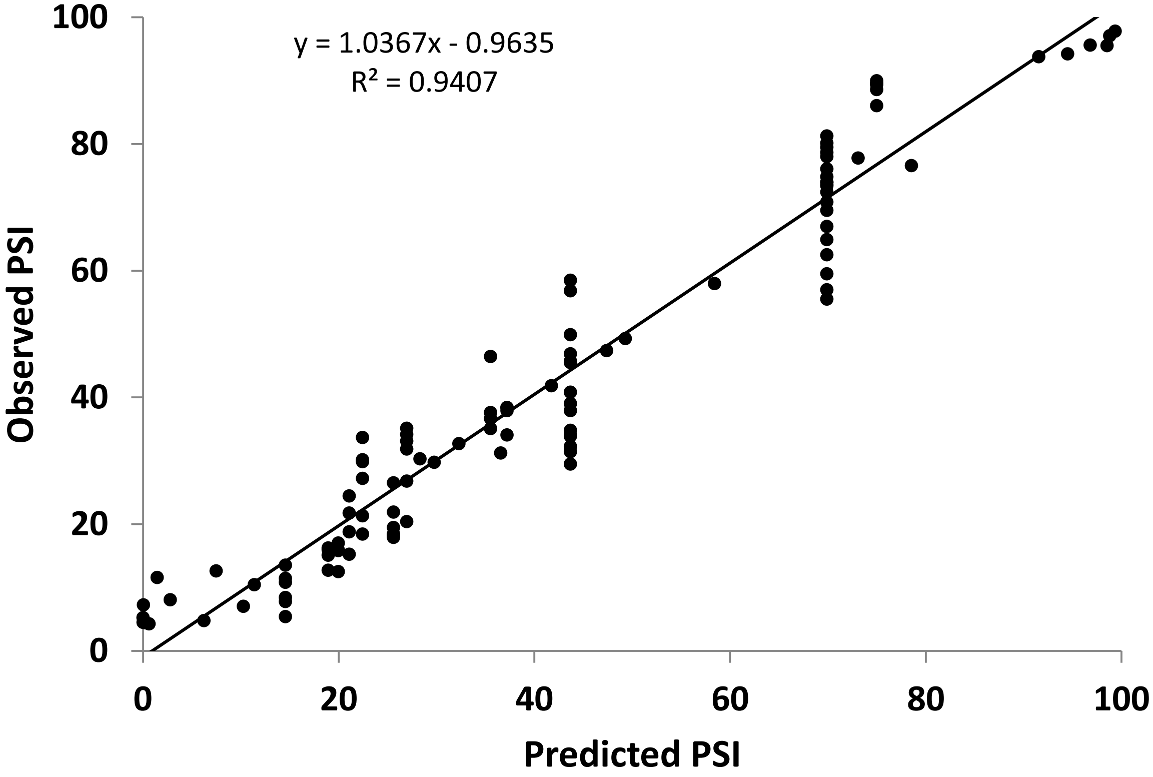


Figure 5.1. The model accurately predicts the inclusion levels of DEs with single parameter perturbations. Psi were predicted using the sequences of the exons and equations 5.5 and 5.6. Values used were those derived for the single parameter experiments described here (Table 5.1). A) Psi for DEs described in Chapter 3.

examine the correspondence between the points in Fig. 3.2 and 3.4 and the curves in Fig. 5.2A

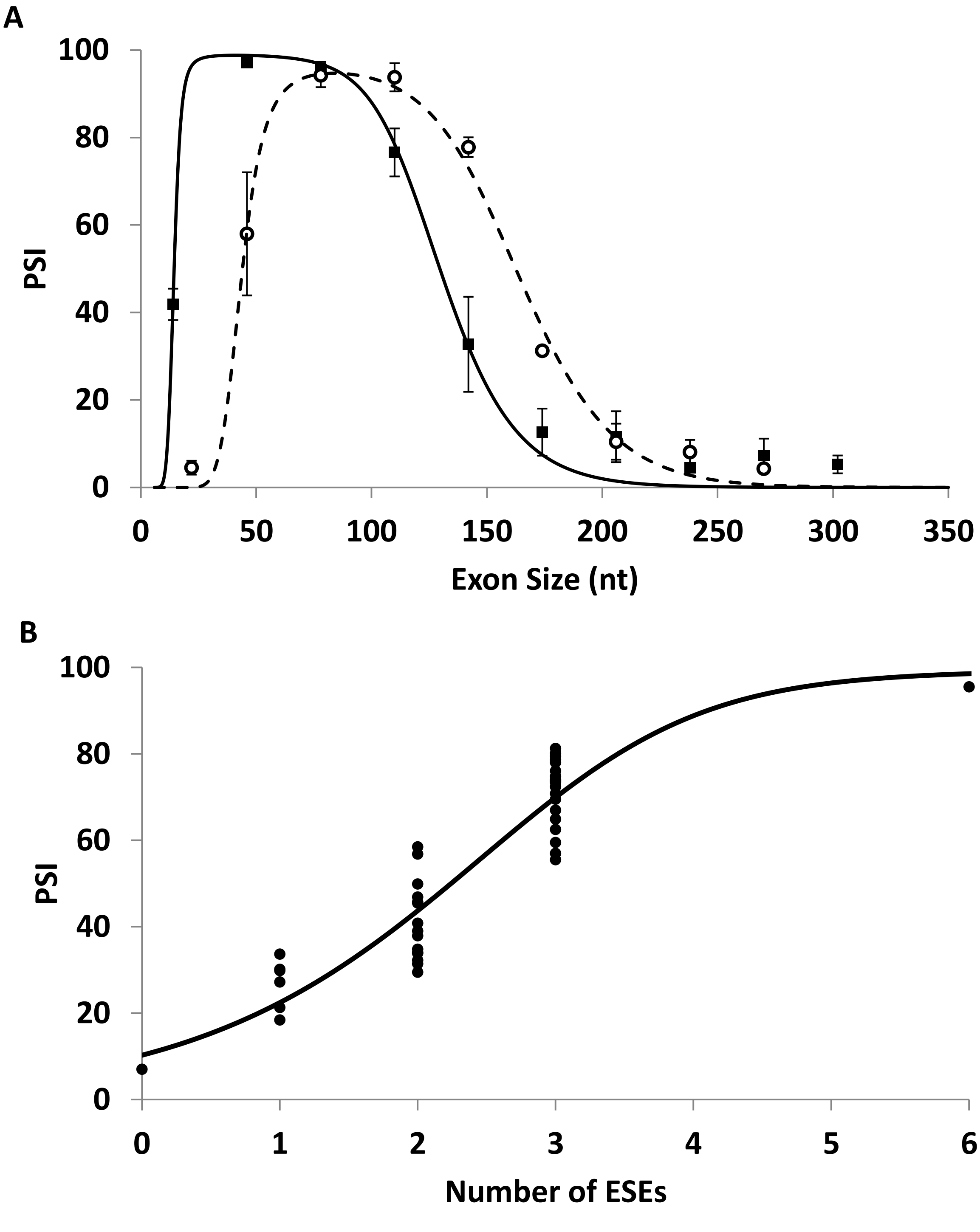


Figure 5.2. The model predicts the size and ESE dependences accurately. Curves drawn using the model described in the text were superimposed on the data points shown in Fig. 3.2 (A) and Fig 3.4 (B). A. Size distributions for SS Set 2 (filled symbols and solid line) and SS Set 3 (open symbols and dashed line). B. Psi variations with ESE number and the predictions made by the model.

and B, for size and ESEs respectively; these curves were drawn according to the predictions of the model and not by a heuristic fit to the points. The agreement between the points and the curves is excellent in both cases. We could not draw a comparable curve for ESSs due to position dependence. However, the accuracy of the observed vs. predicted values can be visualized in Fig.

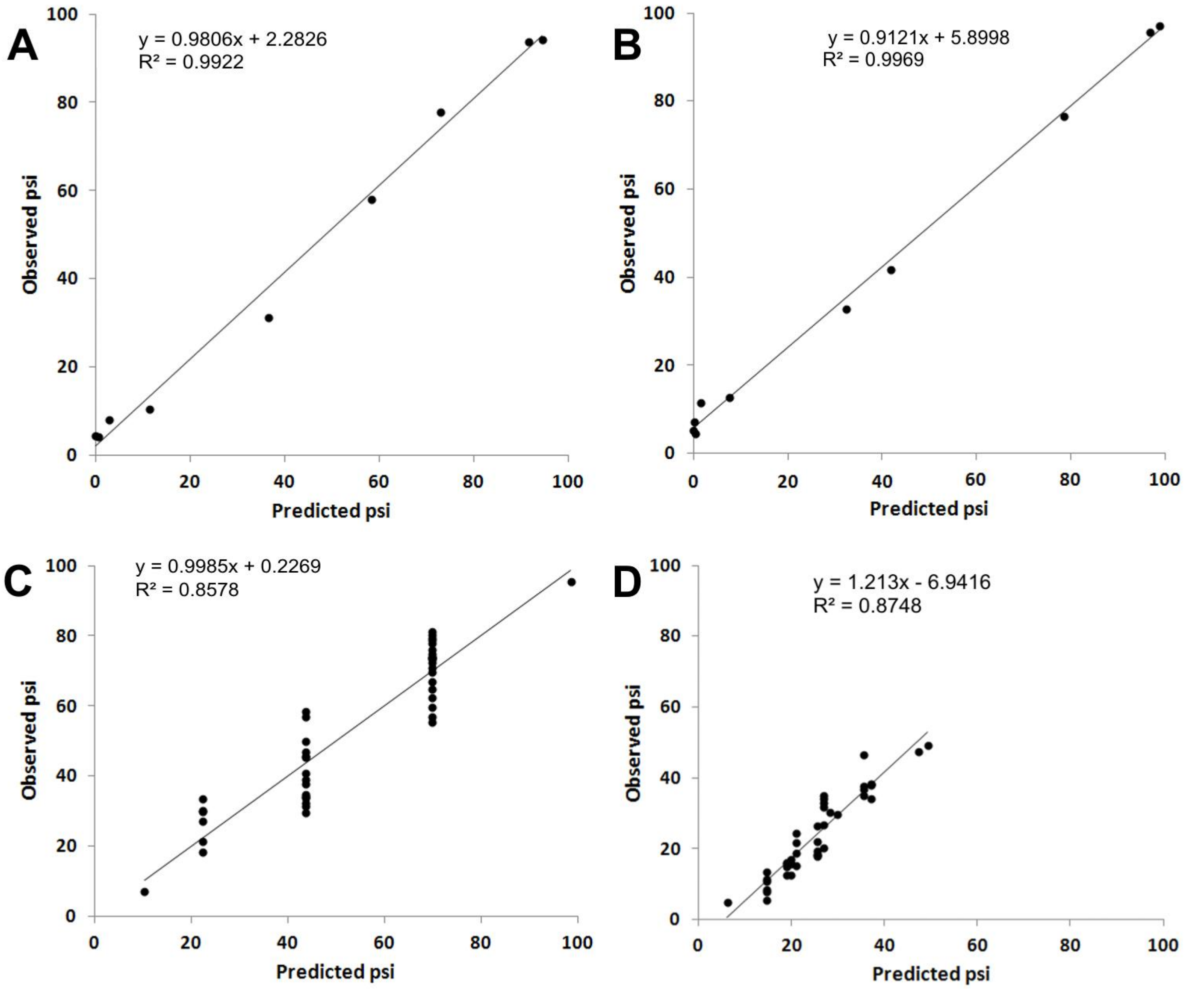


Figure 5.3. The model accurately predicts the inclusion levels of DEs for each parameter examined. The values in Table 5.1 were used to predict the psi; these values were optimized using a condensed and abridged version of the data presented in Chapter 3. A. Exons of different sizes using SS Set 2. B. Exons of different sizes using SS Set 3. C. Exons with 0, 1, 2, 3 or 6 ESEs in all positional combinations. SS Set 7 was used. D. Exons with 0, 1, 2, 3 or 6 ESSs in all positional combinations. SS Set 5 was used.

5.3 for all 3 parameters separately. We conclude that the values shown in Table 5.1 along with equations 5.5 and 5.6 provide a satisfactory model for the single-parameter perturbation data.

A more demanding validation is to test the power of this model on a set of data that was not used for its generation. In the present experiments we purposely examined size, ESEs, and ESSs separately so as to be able to focus on the role of each individual parameter in these DEs. In a previous experiment we examined a much more complex set of 142 DEs that were comprised of exons of varying size and randomly mixed ESE and ESS composition (Zhang et al. 2009). These DEs ranged from 62 to 270 nt in length and included sequences such as SES, SSSE, EEESEE, etc. We asked whether our model could explain the behavior of these more complex DEs, despite the fact that it was optimized without using any exon in which an ESE and an ESS were present together. We refer to these previously studied DEs as “complex DEs.”

Table 5.2. Evaluation of the model

	Single-parameter perturbations	Complex designer exons	y_7 and K_7 fitted to complex DEs
R^2	0.94	0.86	0.86
Slope	1.04	0.95	0.95
Intercept	-0.96%	0.69%	1.29%

Complex DEs differ in two additional ways from the present set of DEs: 1) In the present DEs, a different promoter and polyadenylation site were incorporated, as well as some additional mutations in the first and last exons (see Methods); and 2) semiquantitative endpoint RT-PCR was used in the older experiments as opposed to RT-QPCR used here. These caveats notwithstanding, the model worked quite well in predicting this untouched data, generating an R^2 of 0.86, a slope of 0.95 and an intercept of 0.69% (Table 5.2 and Fig. 5.4). The high correlation indicated by the R^2 is complemented by the high accuracy implied by the match to expected slope and intercept, providing substantial additional support to the model. Although the R^2 value achieved was gratifying, some points were evidently not accurately predicted. There are two types of explanations for such discrepancies. The first is technical, due to the different contexts and methods used and to simple experimental error. The second may be due to limitations in the

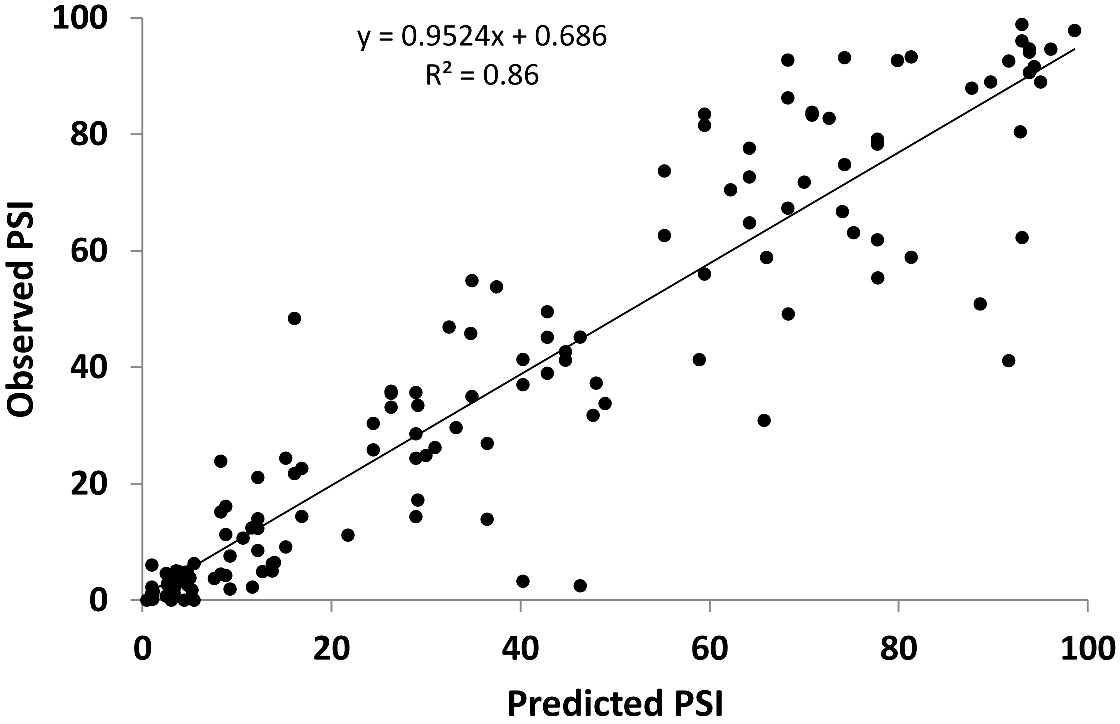


Figure 5.4. The model accurately predicts the inclusion levels of DEs. Psi were predicted using the sequences of the exons and equations 5.5 and 5.6. Values used were those derived for the single parameter experiments described here (Table 5.1). The graph shows psi for more complex DEs harboring combinations of ESEs and ESSs reported in Chapter 2.

current model, which does not take into account possible ESE/ESS interactions or a role for secondary structures.

We addressed three anticipated sources of discrepancy between the old and new data. First, because we examined the size dependence using SS Sets 2 and 3 (Table 3.1) we were able to discover the fitting distance y_2 and y_3 (Table 5.1). Since we did not have a fitting distance (y_7) for the splice site set used to generate the complex DEs, we tried setting it equal to y_2 or to y_3. Either value accurately predicted the results for ESE and ESS variation, as expected for fixed-size DEs. However, in predicting the observations of the complex DEs that differ in size, y_3 was clearly superior (R^2 of 0.86 vs. 0.64). To further explore this issue, we asked what the optimal value for y_7 was for predicting the complex DEs. The BFGS routine was used to optimize y_7 and K_7, while keeping all other parameters constant. The optimized value for y_7 was in fact close to that of y_3 (11.2 vs. 12.0) and quite different from that of y_2 (21.6). Moreover, as shown in Table 5.2, the BFGS-optimized y_7 performed no better than y_3. This suggests a similarity between SS Sets 3 and 7, while it also implies a difference between these and Set 2. As should be evident from Table 3.1, these relationships point to the 5'SS as being the determinant factor in the shift observed in Fig. 5.2A for this limited collection of three SS sets (see Discussion). Second, since the effects of ESEs and ESSs were studied separately here, we evaluated the assumption of independence used to predict the splicing seen in the complex DEs. We studied ESE and ESS action here exclusively in 110 nt exons; therefore, we only examined complex DEs of this length to eliminate any confounding effect of size. As shown in Fig. 5.5, prediction for the splicing of these complex DEs is excellent, with an R^2 of 0.96, a slope of 1.02 and an intercept of 3.29%, supporting the independent action of these two regulatory elements.

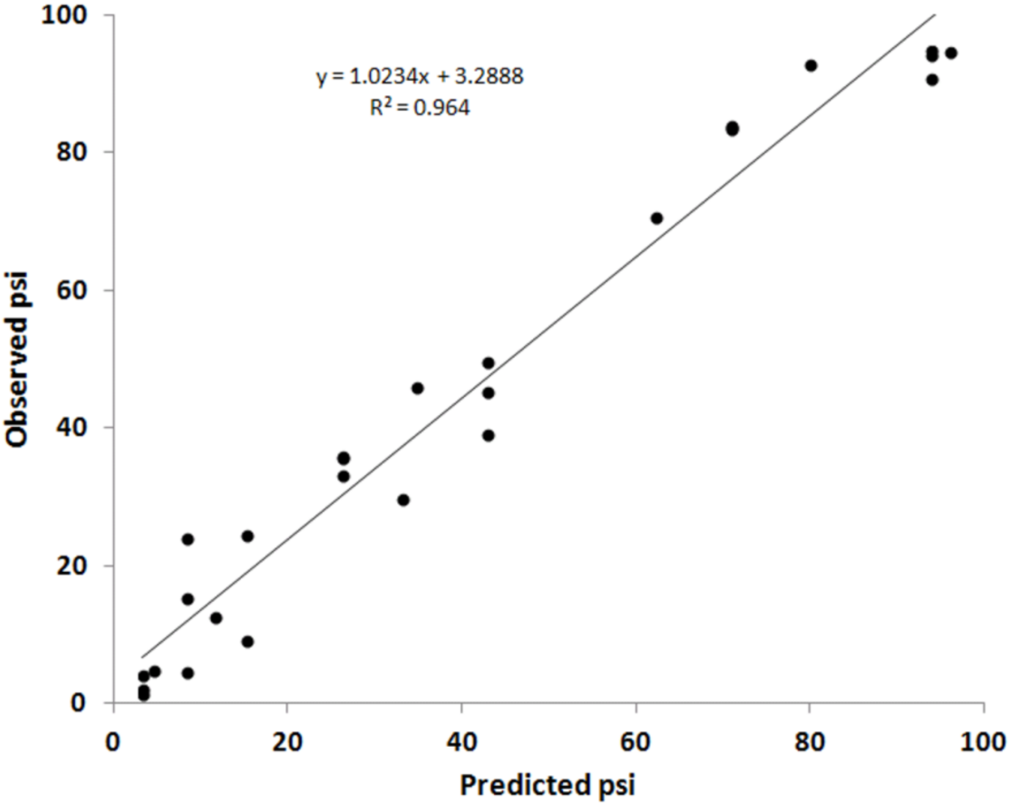


Figure 5.5. The combinations of ESEs and ESSs are accurately modeled in 110 nt exons. The predictions of complex DEs of 110 nt were assessed separately to evaluate the performance of the model for combining ESEs and ESSs while removing the effect of size.

Third, extending this analysis to other sizes we found that even though predictions for the great majority of complex DEs in all size classes showed good correlations (data not shown), a systematic trend was revealed in their accuracy. Beyond ~140 nt the observed values progressively fell short of the predictions at a rate of about 1% per additional nt as shown in Fig. 5.6. We interpret this distortion as being due to a drop-off in PCR efficiency for longer templates, an artifact that is expected in the older data but was avoided by the use of QPCR in the present study. Taking all these results together, the good fit seen suggests that the possible omission of some biological factors in the model is not having a substantial effect on any of these DEs.

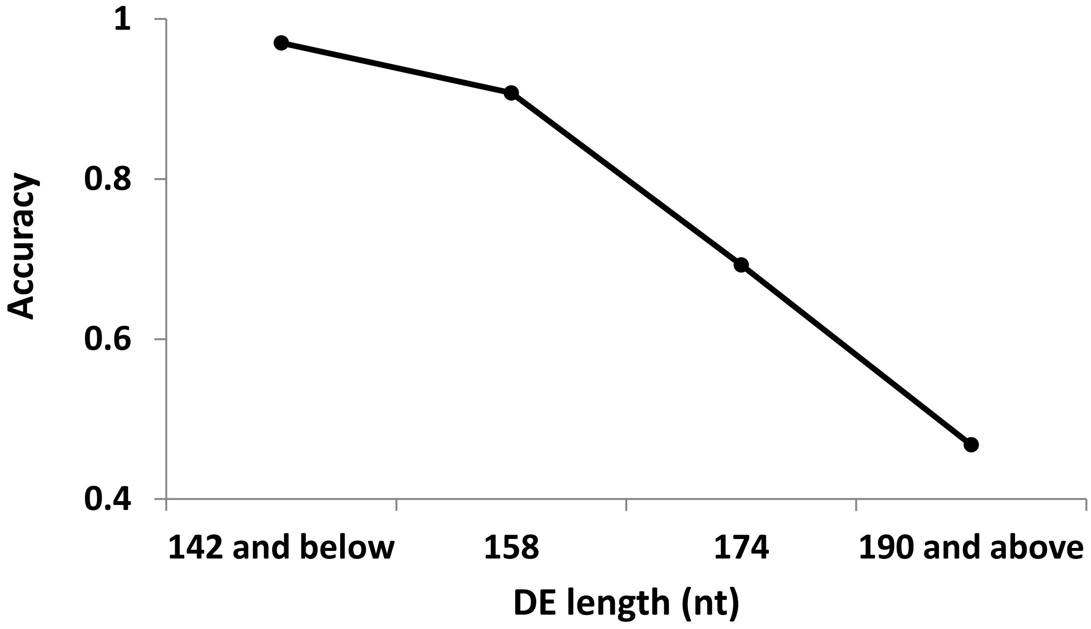


Figure 5.6. The observed psi for complex DEs progressively falls short of prediction as sizes increase above 140 nt. A best fit was performed comparing observed vs. predicted psi for complex DEs of the indicated sizes. Although R^2 values were high in all, accuracy, as reflected in the slope of the fit, decreased with size. This artifact is expected when we consider the decrease in the efficiency of PCR with size for the included (but not the skipped) mRNAs in the measurements for the complex DEs.

Discussion

In chapter 3, we described the splicing phenotypes of exons of our own design, each principally comprised of just 1 or 2 prototype 8-base sequence modules that represent an ESE, an ESS or a reference sequence that resembles neither. This approach enabled us to focus on each of these parameters individually. We found an optimal size range for inclusion that depended on the splice site sequences used. A stronger 5’ splice site favored inclusion of longer exons while disfavoring the inclusion of shorter exons. In contrast, a stronger 3’ splice site favored inclusion of shorter exons while disfavoring the inclusion of longer exons, the opposite effect. That is, surprisingly, changes that worked for improving the splicing of longer exons did not work for improving the splicing of shorter exons, and vice versa. An ESE enhanced splicing uniformly from any position along the exon; the expected position dependence was not seen. The ESS showed some position dependence: the most downstream position tested was the most effective while the most upstream position had little effect; intermediate positions showed uniform intermediate effects. Multiple enhancers or silencers showed additive enhancing or silencing effects respectively.

Many systems used to study splicing, especially in vitro splicing, use 2-exon substrates or substrates with short (<200 nt) introns, favoring intron definition rather than exon definition (Talerico and Berget 1994; Fox-Walsh et al. 2005). In contrast, we sought here to focus on exon definition, and so studied splicing of an internal exon and used longer introns (~300 and ~600 nt). Importantly, weakening either splice site of the internal exon resulted only in increased skipping, as expected with exon-definition, with no signs of intron retention (data not shown). Prompted by these results indicating that DEs were recognized by exon definition, we devised a general equation for exon definition that incorporated several intermediate states along a splicing pathway (equation 5.4). It is noteworthy that this equation predicts that lengthening τ, the time available for commitment exclusively to the included fate (e.g., by slowing synthesis), should increase psi; this kinetic effect has been observed previously in exon-definition systems (Dujardin et al. 2013). Using the equations developed in Chapter 4, we also explored the potential of intuitive but novel mechanisms to explain our observations. While these observations have been obtained using a simplified exon we expect the underlying mechanisms to be equally applicable to the definition of natural exons since they are based on straightforward biophysical assumptions and are indeed supported by previous studies (see below).

Modeling the effect of size

It has been suggested that there is an interaction between U2AF and U1 snRNP not only across the intron (Michaud and Reed 1993) but also across the exon (Hoffman and Grabowski 1992; Reed 1996). We modeled this interaction across the exon as an exon definition complex. Tethered collisions were used to model the formation of this complex (Fig. 4.2). Not all collisions will be productive; both ends of the exon must approach each other in the correct orientation in order to interact. The probability of a productive collision was modeled assuming the RNA behaves as a flexible worm-like chain. After the bound RNA sequences at the ends of this chain become associated the physical distance between these two ends becomes fixed (the fitting distance y_i in Fig. 4.2). The emerging equations predict that splicing efficiency should decrease for short exons and for long exons: if an exon is very short no collisions may be possible while for long exons the chance of a collision between the ends is low. Effects of length have been observed previously (Black 1991; Dominski and Kole 1991; Peterson et al. 1994; Sterner et al. 1996; Hwang and Cohen 1997; Borensztajn et al. 2006) as well as in these systematic DE results. By optimizing the fitting distance independently for each set of SSs used, we found that a difference in this parameter could predict the surprising shift seen in Fig. 5.2A. The optimal size range shift is explained solely by the difference in these two distances. However, a low psi would be expected at this peak for SS Set 3 but a high K_3 relative to K_2 compensates for this. The values for y_2 and y_3 in Table 5.1 are in the size range of the RNP complexes posited (Kastner and Luhrmann 1989; Pomeranz Krummel et al. 2009; Weber et al. 2010). Importantly, in this limited collection of three SS sets, the 5’ SS is the determinant factor, since y_i changes substantially only when the 5’ SS is changed. Three possibilities come to mind to explain this SS sequence dependence: 1) a large conformational change in one or more of the proteins bound to these sequences. Although a difference of 9.6 nm (Table 5.1) seems large, protuberances of this size have been seen in U1 snRNP (Kastner and Luhrmann 1989; Pomeranz Krummel et al. 2009; Weber et al. 2010). 2) A small conformational change that enables one or more proteins to recruit an additional “bulging” factor. 3) A sequence dependent change in the point or angle at which the pre-mRNA extends from U1 snRNP (Fig. 4.2B-D). A possible example of this last option can be seen by comparing the crystal structures of different nucleic acid sequences bound by U1 snRNP described by Pomeranz Krummel et al., (PDB ID 3CW1, 2009) and by Weber et al., (PDB ID 3PGW , 2010): the two different nucleic acid molecules extend from the U1 snRNP in a different manner. Irrespective of the model used, the shift between these two curves implies that comparing the strengths of 5’SS sequences might be more complex than previously thought.

Several similarities have been noted between the size restrictions for exons in exon definition and those for introns in intron definition. For example, introns longer than ~300 nt are disfavored in organisms relying mostly on intron definition (*D. melanogaster*) and exons longer than ~300 nt are disfavored in organisms relying mostly on exon definition (humans) (Sterner et al. 1996; Fox-Walsh et al. 2005; Xiao et al. 2007). Moreover, Garcia-Blanco and colleagues presented evidence supporting pairing of the ends of introns via three dimensional diffusion (Pasman and Garcia-Blanco 1996), a mechanism similar to that proposed here for exon end pairing. Interestingly, the size distributions of short introns in human and Drosophila are greatly disjoint (Fig. 3.6 in (Fox-Walsh et al. 2005)). The optimum size for splicing is greater in human (90 nt) than in Drosphila (75 nt) nuclear extracts (Guo et al. 1993). These observations suggest that a size dependence similar to that in Fig 5.2A could explain this species difference by assuming tethered end collisions across the intron with a different y_i for each organism. This difference could be dictated by differences in the size and/or number of the proteins involved.

Modeling the effect of ESEs

Recruitment of the splicing machinery as a mode of action for ESEs (Kohtz et al. 1994; Staknis and Reed 1994) is supported by evidence of interactions between activator proteins that bind ESEs and some of the proteins involved in the early steps of splicing (Hoffman and Grabowski 1992; Kohtz et al. 1994; Staknis and Reed 1994). There is an expected position effect that this interaction should display: the closer the binding site for the activator to the splice site, the more efficient it should be in recruiting the splicing machinery to that site. The absence of this effect prompted us to model the action of the ESE as simply stabilizing an otherwise volatile interaction between U2AF and U1 snRNP.

Previously observed increases in theyield of splicing complexes (Hoffman and Grabowski 1992) can be explained by stabilization as well as by recruitment. Changes in stability, expressed as the rate of dissociation (d within equation 5.4), should respond exponentially to the number of ESEs. This stability model predicts a sigmoidal curve but with a near linear relationship between psi and the number of ESEs over much of the range examined and accounts for the saturation effect when more than 4 ESEs are used (Fig. 5.2B). In contrast, recruitment depends on a change in binding probability of the splicing machinery, which is expected to be linear with respect to ESE number. This linearity could be incorporated in a (the association rate constant) within equation 5.4 (see Supplemental Material). The resulting recruitment model led to a fit for predicting the results on the single-parameter-perturbation data (an R^2 of 0.92, a slope of 0.90, and an intercept of 5.62%) that was nearly as good as the stability model (Fig. 5.1). However, it produced a negative exponential shaped curve that did not fit the ESE data as well (See Supplemental Fig. S5.1) and unlike the stability model it performed poorly for the complex DEs (R^2 of 0.37 compared to 0.86). In particular, for the constant size class of 110 nt which isolates the effect of ESE/ESS combinations, even though an acceptable R^2 of 0.84 was obtained, a slope of 1.78 and an intercept of -63% revealed a flawed performance compared to the stability model (compare Supplemental Fig. S5.2 and Fig. 5.5).

Thus although recruitment may play a role, we conclude that stabilization is the dominant feature. It is interesting to note that the model used here could account for the dependence of *in vitro* splicing efficiency on the number of doublesex enhancers (compare Supplemental Fig. S5.3 to Fig. 2D in (Hertel and Maniatis 1998)). This agreement with long-established data suggests that these results using a prototype ESE of our own design reflect general mechanisms involved in splicing and may not be limited to internal exons. Recruitment and stabilization are not at all mutually exclusive; one can imagine recruitment of a factor followed by stabilization of the binding of that factor and/or the subsequent stabilization of a full exon definition complex.

ESS number and position effect

In Chapter 3 we showed that the effect of multiple ESSs could be predicted by their linear combination as long as the particular characteristics of positions 1 and 6 were taken into account (Fig. 3.6B). For modeling, we contented ourselves with considering the action of ESSs to be opposite that of ESEs; that is, as destabilizing elements. Although only the data for single-ESS DEs were used to optimize the model, the effect of multiple ESSs (which included the saturating case of 6 ESSs) was accurately predicted (Supplemental Fig. S5.4) these predictions were in fact more highly correlated ($R^2 = 0.82$) than simply summing the effects of the individual ESSs (equation 3.1; $R^2 = 0.73$ when the 6 ESS data point was included). The position effect seen for ESSs suggests that ESSs may act by destabilizing bound U1 snRNP or even blocking its binding rather than by affecting an exon definition complex. Further studies using different ESS/splice sites combinations and incorporating mechanisms such as competition for binding RNA into the equations could be tried using the present model as a template.

Mechanistic interpretations

These results indicate that modeling a proposed irreversible step in exon recognition (exon commitment) coupling statistical mechanics and a reductionist approach is enough to make accurate predictions for the psi of combinations of the sequences presented in Chapters 2 and 3. This result agrees with those of previous studies showing that the sequences studied, which affected splicing, exert a decisive change on early assembly of the spliceosome, preceeding even complex A formation. It could be envisioned that once the initial hurdle of exon recognition is surpassed, effects on subsequent steps in the splicing reaction might delay the final outcome but not affect it significantly. Thus, “commitment” becomes an appropriate description for this early milestone and is reminiscent of cell differentiation and promoter clearance in transcription. Indeed this commitment step might be related to the probability of the exon being captured by an exon hub, a function that could be fulfilled by the CTD (see below).

The values of the optimized equation coefficients used in the model (Table 5.1) show expected characteristics as well as some surprises. The coefficients for dissociation for ESEs (c_E) and ESSs (c_F, c_L and c_I) were less and greater than unity, respectively, as expected. We expected the coefficient for the reference sequence (c_R) to be close to unity if it were neutral, but obtained a value of 1.5, which leads to a substantial decrease in the predicted psi (data not shown). Therefore this reference sequence has a negative effect on the formation of the exon definition complex.

K_i is a catch-all constant in equation 5.6 that notably includes the effect of SS “strength.” SS Set 3 differs from Set 5 by only a single base in the 3’SS (see Table 3.1) and results in a 2.8-fold increase in K_i. Set 5 differs from Set 7 by two bases in the 3’ SS and results in a 7.1-fold increase in K_i. Differences in the 5’ SS were found to be substantially greater. Set 2 differs from Set 7 by only a single base in the 5’ SS yet results in a ~250-fold increase in K_i. The greater effect of the 5’ SS suggests a more critical role of its sequence, as has been suggested before (Xiao et al. 2007).

Finally, T and C in equation 5.5 provide an indication of the contributions of the pre-τ and the post-τ phases. The value for T represents the commitment to inclusion that takes place even before the third exon is synthesized while the value for C models the period after the third exon becomes available. As shown in Table 5.1, C is several orders of magnitude smaller than T, implying that by the time competition becomes possible, essentially no additional molecules commit to inclusion (i.e., all remaining molecules will be committed to skipping). Indeed, setting C = 0 does not change the performance of the model (data not shown). This surprising result could be due to an unexplained relative weakness of the DEs used compared to the downstream exon; or, more intriguingly, to a mechanism that was not considered in the model: that there is a restricted window of commitment time that is shorter than τ. Consequently, molecules that have not committed to inclusion within this window of time can no longer do so; paradoxically, they are, by default, “committed” to skipping even before the downstream exon is synthesized. This shifts the critical time period from minutes (Kessler et al. 1993; Singh and Padgett 2009; Wada et al. 2009) to seconds, arguing that most of the time spent before the spliced product is formed is spent after the commitment step has been taken. As a matter of fact this window of commitment might be dictated by the time the exon containing the exon definition complex can be captured by a putative hub. In the case of the hub being the CTD of the RNA polymerase II, this might be related to the time at which transcription on the subsequent intron tethers the exon too far to ensure a collision of the exon definition complex and the CTD. Four regimens are then established: the time involved in exon definition complex formation, which should be in the order of milliseconds at most (Chen et al. 2012; Hyeon and Thirumalai 2012); the time required for commitment, seconds (as suggested in this paper); the time required to generate the spliced product, a few minutes (Kessler et al. 1993; Singh and Padgett 2009; Wada et al. 2009), and the time required to generate the final mRNA molecule, up to several hours.

Materials and Methods

Parameter optimization

A BFGS algorithm adapted from Press et al (Press et al. 2007) implementing walls to force all parameters to be non-negative and using explicit gradient was written in Perl for minimizing the sum of the squared differences between observed and predicted pso (equations 3.6 and 3.7; see Supplemental Material).

Authors' Contributions

MA conceived the study and approach. MA carried out the calculations and performed the analysis. LC and MA wrote the text.

Acknowledgments

We thank Vincent Anquetil and Jonathan Cacciatore for valuable discussions and Dennis Weiss for the same plus a thoughtful critique of the manuscript. This work was supported by a grant from the NIH (GM072740) to L.A.C.

References

Arias MA, Lubkin AL, Chasin LA. 2013. Splicing of designer exons informs a biophysical model for exon definition. *Manuscript submitted*.

Black DL. 1991. Does steric interference between splice sites block the splicing of a short c-src neuron-specific exon in non-neuronal cells? *Genes Dev* **5**(3): 389-402.

Borensztajn K, Sobrier ML, Duquesnoy P, Fischer AM, Tapon-Bretaudiere J, Amselem S. 2006. Oriented scanning is the leading mechanism underlying 5' splice site selection in mammals. *PLoS Genet* **2**(9): e138.

Chen H, Meisburger SP, Pabit SA, Sutton JL, Webb WW, Pollack L. 2012. Ionic strength-dependent persistence lengths of single-stranded RNA and DNA. *Proceedings of the National Academy of Sciences of the United States of America* **109**(3): 799-804.

Cooper TA, Ordahl CP. 1989. Nucleotide substitutions within the cardiac troponin T alternative exon disrupt pre-mRNA alternative splicing. *Nucleic Acids Res* **17**(19): 7905-7921.

De Conti L, Baralle M, Buratti E. 2013. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA* **4**(1): 49-60.

de la Mata M, Alonso CR, Kadener S, Fededa JP, Blaustein M, Pelisch F, Cramer P, Bentley D, Kornblihtt AR. 2003. A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell* **12**(2): 525-532.

Dominski Z, Kole R. 1991. Selection of splice sites in pre-mRNAs with short internal exons. *Mol Cell Biol* **11**(12): 6075-6083.

Dujardin G, Lafaille C, Petrillo E, Buggiano V, Gomez Acuna LI, Fiszbein A, Godoy Herz MA, Nieto Moreno N, Munoz MJ, Allo M et al. 2013. Transcriptional elongation and alternative splicing. *Biochim Biophys Acta* **1829**(1): 134-140.

Fox-Walsh KL, Dou Y, Lam BJ, Hung SP, Baldi PF, Hertel KJ. 2005. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proceedings of the National Academy of Sciences of the United States of America* **102**(45): 16176-16181.

Fu XD. 1995. The superfamily of arginine/serine-rich splicing factors. *RNA* **1**(7): 663-680.

Ge H, Manley JL. 1990. A protein factor, ASF, controls cell-specific alternative splicing of SV40 early pre-mRNA in vitro. *Cell* **62**(1): 25-34.

Guo M, Lo PC, Mount SM. 1993. Species-specific signals for the splicing of a short Drosophila intron in vitro. *Mol Cell Biol* **13**(2): 1104-1118.

Hertel KJ, Maniatis T. 1998. The function of multisite splicing enhancers. *Mol Cell* **1**(3): 449-455.

Hoffman BE, Grabowski PJ. 1992. U1 snRNP targets an essential splicing factor, U2AF65, to the 3' splice site by a network of interactions spanning the exon. *Genes Dev* **6**(12B): 2554-2568.

Hoskins AA, Friedman LJ, Gallagher SS, Crawford DJ, Anderson EG, Wombacher R, Ramirez N, Cornish VW, Gelles J, Moore MJ. 2011. Ordered and dynamic assembly of single spliceosomes. *Science* **331**(6022): 1289-1295.

Hwang DY, Cohen JB. 1997. U1 small nuclear RNA-promoted exon selection requires a minimal distance between the position of U1 binding and the 3' splice site across the exon. *Mol Cell Biol* **17**(12): 7099-7107.

Hyeon C, Thirumalai D. 2012. Chain length determines the folding rates of RNA. *Biophys J* **102**(3): L11-13.

Kastner B, Luhrmann R. 1989. Electron microscopy of U1 small nuclear ribonucleoprotein particles: shape of the particle and position of the 5' RNA terminus. *EMBO J* **8**(1): 277-286.

Kessler O, Jiang Y, Chasin LA. 1993. Order of intron removal during splicing of endogenous adenine phosphoribosyltransferase and dihydrofolate reductase pre-mRNA. *Mol Cell Biol* **13**(10): 6211-6222.

Kohtz JD, Jamison SF, Will CL, Zuo P, Luhrmann R, Garcia-Blanco MA, Manley JL. 1994. Protein-protein interactions and 5'-splice-site recognition in mammalian mRNA precursors. *Nature* **368**(6467): 119-124.

Krainer AR, Conway GC, Kozak D. 1990. Purification and characterization of pre-mRNA splicing factor SF2 from HeLa cells. *Genes Dev* **4**(7): 1158-1171.

Lam BJ, Hertel KJ. 2002. A general role for splicing enhancers in exon definition. *RNA* **8**(10): 1233-1241.

Mardon HJ, Sebastio G, Baralle FE. 1987. A role for exon sequences in alternative splicing of the human fibronectin gene. *Nucleic Acids Res* **15**(19): 7725-7733.

Michaud S, Reed R. 1993. A functional association between the 5' and 3' splice site is established in the earliest prespliceosome complex (E) in mammals. *Genes Dev* **7**(6): 1008-1020.

Mount SM. 1982. A catalogue of splice junction sequences. *Nucleic Acids Res* **10**(2): 459-472.

Munoz MJ, Perez Santangelo MS, Paronetto MP, de la Mata M, Pelisch F, Boireau S, Glover-Cutter K, Ben-Dov C, Blaustein M, Lozano JJ et al. 2009. DNA damage regulates alternative splicing through inhibition of RNA polymerase II elongation. *Cell* **137**(4): 708-720.

Osheim YN, Miller OL, Jr., Beyer AL. 1985. RNP particles at splice junction sequences on Drosophila chorion transcripts. *Cell* **43**(1): 143-151.

-. 1988. Visualization of Drosophila melanogaster chorion genes undergoing amplification. *Mol Cell Biol* **8**(7): 2811-2821.

Pasman Z, Garcia-Blanco MA. 1996. The 5' and 3' splice sites come together via a three dimensional diffusion mechanism. *Nucleic Acids Res* **24**(9): 1638-1645.

Peterson ML, Bryman MB, Peiter M, Cowan C. 1994. Exon size affects competition between splicing and cleavage-polyadenylation in the immunoglobulin mu gene. *Mol Cell Biol* **14**(1): 77-86.

Pomeranz Krummel DA, Oubridge C, Leung AK, Li J, Nagai K. 2009. Crystal structure of human spliceosomal U1 snRNP at 5.5 A resolution. *Nature* **458**(7237): 475-480.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 2007. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, New York.

Reed R. 1996. Initial splice-site recognition and pairing during pre-mRNA splicing. *Curr Opin Genet Dev* **6**(2): 215-220.

Reed R, Maniatis T. 1986. A role for exon sequences and splice-site proximity in splice-site selection. *Cell* **46**(5): 681-690.

Roberts GC, Gooding C, Mak HY, Proudfoot NJ, Smith CW. 1998. Co-transcriptional commitment to alternative splice site selection. *Nucleic Acids Res* **26**(24): 5568-5572.

Shepard PJ, Choi EA, Busch A, Hertel KJ. 2011. Efficient internal exon recognition depends on near equal contributions from the 3' and 5' splice sites. *Nucleic Acids Res* **39**(20): 8928-8937.

Singh J, Padgett RA. 2009. Rates of in situ transcription and splicing in large human genes. *Nat Struct Mol Biol* **16**(11): 1128-1133.

Staknis D, Reed R. 1994. SR proteins promote the first specific recognition of Pre-mRNA and are present together with the U1 small nuclear ribonucleoprotein particle in a general splicing enhancer complex. *Mol Cell Biol* **14**(11): 7670-7682.

Sterner DA, Carlo T, Berget SM. 1996. Architectural limits on split genes. *Proceedings of the National Academy of Sciences of the United States of America* **93**(26): 15081-15085.

Sun H, Chasin LA. 2000. Multiple splicing defects in an intronic false exon. *Mol Cell Biol* **20**(17): 6414-6425.

Talerico M, Berget SM. 1994. Intron definition in splicing of small Drosophila introns. *Mol Cell Biol* **14**(5): 3434-3445.

Tsai AY, Streuli M, Saito H. 1989. Integrity of the exon 6 sequence is essential for tissue-specific alternative splicing of human leukocyte common antigen pre-mRNA. *Mol Cell Biol* **9**(10): 4550-4555.

Ujvari A, Luse DS. 2004. Newly Initiated RNA encounters a factor involved in splicing immediately upon emerging from within RNA polymerase II. *J Biol Chem* **279**(48): 49773-49779.

Wada Y, Ohta Y, Xu M, Tsutsumi S, Minami T, Inoue K, Komura D, Kitakami J, Oshida N, Papantonis A et al. 2009. A wave of nascent transcription on activated human genes. *Proceedings of the National Academy of Sciences of the United States of America* **106**(43): 18357-18361.

Weber G, Trowitzsch S, Kastner B, Luhrmann R, Wahl MC. 2010. Functional organization of the Sm core in the crystal structure of human U1 snRNP. *EMBO J* **29**(24): 4172-4184.

Wu JY, Maniatis T. 1993. Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell* **75**(6): 1061-1070.

Xiao X, Wang Z, Jang M, Burge CB. 2007. Coevolutionary networks of splicing cis-regulatory elements. *Proceedings of the National Academy of Sciences of the United States of America* **104**(47): 18583-18588.

Zhang XH, Arias MA, Ke S, Chasin LA. 2009. Splicing of designer exons reveals unexpected complexity in pre-mRNA splicing. *RNA* **15**(3): 367-376.

Supplemental Material

Modeling recruitment

To model recruitment, we made use of the linearity predicted previously (Hertel and Maniatis 1998) and assumed that the rate of association would be affected by the number of enhancers in a linear manner. The same was assumed for the silencers and reference sequences, generating the approximation

S1. $D \approx K_i\ Y_i^{-2}\ c_R^{nR}\ Z^{3/2}\ e^{3Yi^2/Z}\ /\ (1+c_{E*}n_E+\ c_{F*}n_F\ +c_{L*}n_L\ +c_{I*}n_I)$

Preliminary attempts using this equation gave values for T, C, K_2, K_3 and K_5 of the order of thousands and for K_7 of the order of millions, suggesting that the rate of dissociation was much greater than the rate of association and convergence was difficult to achieve. To solve this issue, a was assumed to be neglible compared to d in equation 5.5 to generate

S2. $pso \approx 100\ e^{-T/D}/(1+C/D)$

The data available cannot be used to separate the contributions of T, C and K_i in equation S5.2. However, assuming a value for T allows C and K_i to be optimized. We decided to retain the value for T obtained with the stability model: 5.24 (see Table 5.1). After a first round of optimization, a second round was performed using as input only DEs for which a positive prediction was obtained in the first round. Additionally, in order to mimic the effects of saturation, any predicted value above 100% was taken to be 100% and any negative value was taken to be 0%. The optimized values for the model are shown in Supplemental Table S5.1. Using this model a good fit was obtained for the single-parameter-perturbation data (R^2 = 0.92, slope = 0.90 and intercept = 5.62%). However, the predictions of the complex DEs were poor (R^2 = 0.37, slope = 1.18, intercept = -35.7%). In particular for 110 nt DEs, even though the R^2 was acceptable (0.84), the poor predictions as indicated by a slope of 1.78 and an intercept of -63%, suggest that the combination of ESEs and ESSs in a single exon is not modeled appropriately. We considered the possibility that these regulatory sequences affect both splice sites by squaring each contribution; this modification did not improve R^2 for either the input data or the more complex DEs (data not shown).

Materials and methods

Model optimization

A Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS) adapted from Press et al. (Press et al. 2007) implementing walls to force all optimized values to be non-negative and using explicit gradient was written in Perl for minimizing the sum of the squared differences between observed and predicted pso (equation 5.5). All values for the model were seeded as 1 except for T, C and Y_i. T was allowed to vary between 0 and 10 in steps of 1, C between 10^{-8} and 100 with a factor of 10 between steps and Y_3 between 1 and 25 in steps of 1. Y_2 and Y_3 were started with the same seed but subsequently allowed to vary independently; any other Y_i was assumed to be equal to either Y_3 or Y_2 as indicated in the Results. To improve convergence, the routine was modified to reset the direction for line minimization to that of steepest descent if the vector of values for the model did not change when a full step in the updated BFGS direction was taken. This modification reduces the number of seed sets for which the program crawls to a stop without reaching a convergent solution (a stalled run) and practically increases the number of convergent solutions by allowing otherwise stalled runs to converge. So as not to include the results of stalled runs, a set of values was taken as a solution if and only if, for a set of input data, considering the full set of seeds, it provided the minimum sum of the squared differences. This same minimum sum had to be obtained several times with all optimized values identical to at least 5 significant figures (using different seeds). The magnitude of the gradient had to be smaller than 10^{-7} in at least 2 cases. If no such solution was found the program was said to have failed to find a convergent solution. These criteria were met for all optimized values except for C, which was so low as to be negligible. In this case the value yielding the minimum sum with the minimum gradient was used. In fact, setting C equal to zero did not affect the results.

Using all the data points available (Supplemental Table S3.1), the program failed to converge on a set of values for the model. We reasoned that the multidimensional surface was too complex and relatively flat causing the program to crawl to a stop when exploring it. To address this issue, we simplified the data by using our observations that ESEs are position independent and that, using single-ESS DEs, the effects of multiple-ESS DEs can be predicted. We thus condensed all ESE results corresponding to a given number of ESEs by their average and removed the 36 data points corresponding to multiple-ESS DEs. (The data points for ESEs exclusively with SS Set 7 were used.) We also found it necessary to remove a single outlying point (SS Set 3, length = 206 in Fig. 3.2) from the 19 size perturbation points in order to achieve reproducible convergence. This point also did not agree with the data from permanent transfections (Supplemental Fig. S3.2). While this reduced and condensed set was used for optimizing the values for the model, the single-parameter-perturbation evaluation of the model was performed using all the 112 points available.

Supplemental Table S5.1. Values for the recruitment model.

T	C	K_2	K_3	K_5	K_7	c_E	c_R	c_F	c_L	c_I	y_2	y_3
5.24	4.60×10^{-13}	6.44×10^{-3}	1.90×10^{-2}	3.49×10^{-2}	0.946	5.56	-4.53×10^{-2}	-6.76×10^{-2}	-0.242	-0.175	15.5	8.73

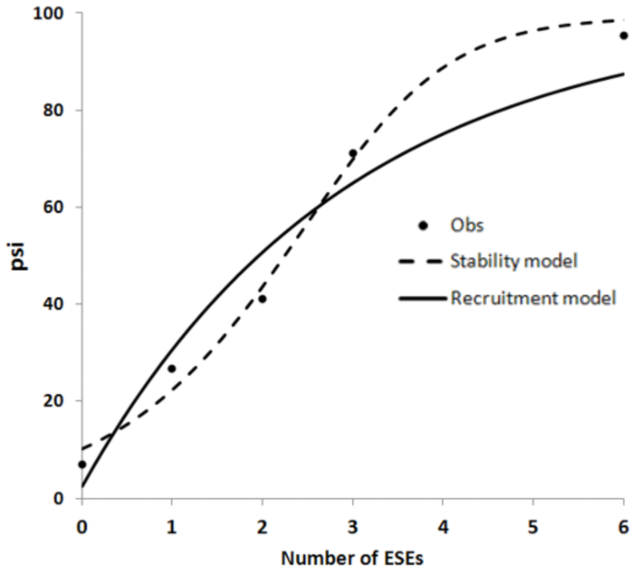


Figure S5.1. Increasing the number of ESEs yields distinctive curves for the stability and the recruitment models for ESE action. Adding ESEs generates a sigmoidal curve according to the stability model and a negative exponential curve according to the recruitment model. The stability model follows the experimental observations more closely.

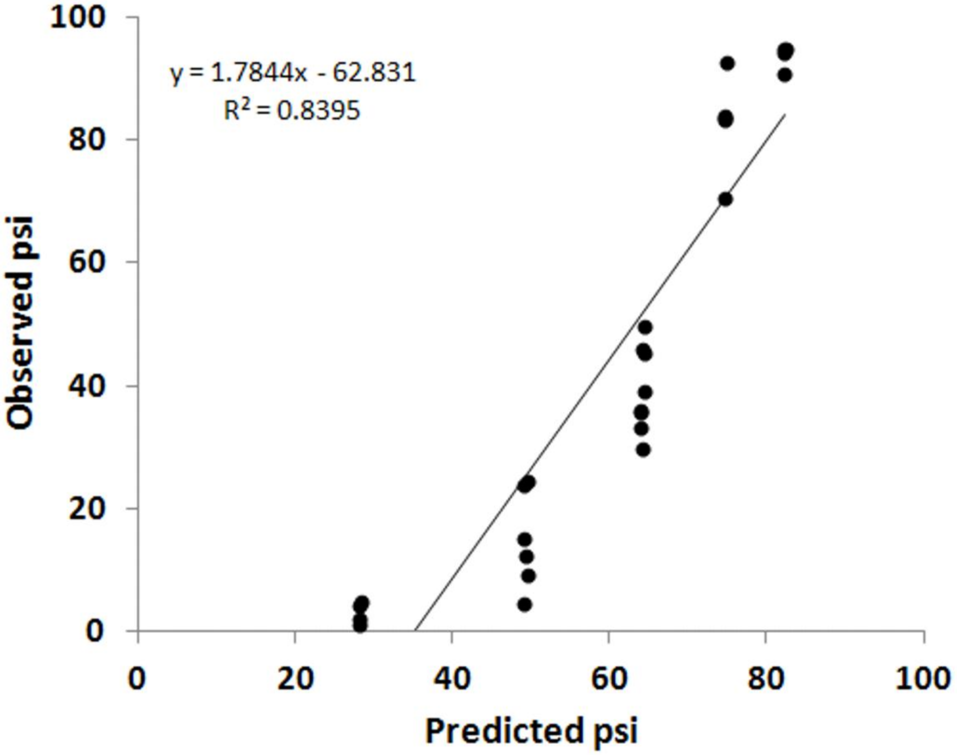


Figure S5.2. The recruitment model fails to accurately predict the effect of combining ESEs and ESSs in 110 nt exons. The predictions of complex DEs of 110 nt were assessed separately to evaluate the performance of the model for combining ESEs and ESSs while removing the effect of size. Although a high R^2 was achieved the slope and intercept deviate markedly from the expected values of 1 and 0%, respectively.

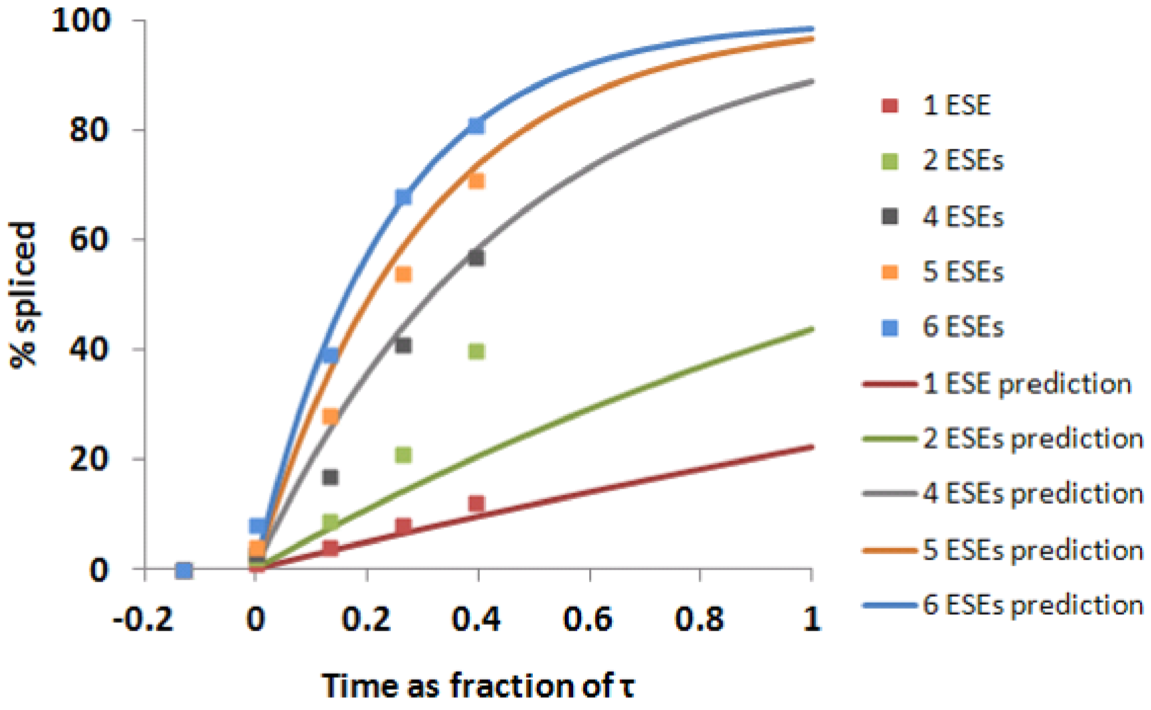


Figure S5.3. Comparison of time course experiments using constructs with multiple ESEs in (Hertel and Maniatis 1998) with time course predictions by the model described in the text. The observed data (points) were extracted from Fig. 2D of (Hertel and Maniatis 1998) and plotted assuming a splicing delay of 1 hr and a total time (τ) of 7.6 hrs. The values in Table 3.2 were used for the model (curves). The points and the curve corresponding to 1 ESE are shown in red, 2 ESEs in green, 4 ESEs in gray, 5 ESEs in orange, and 6 ESEs in blue.

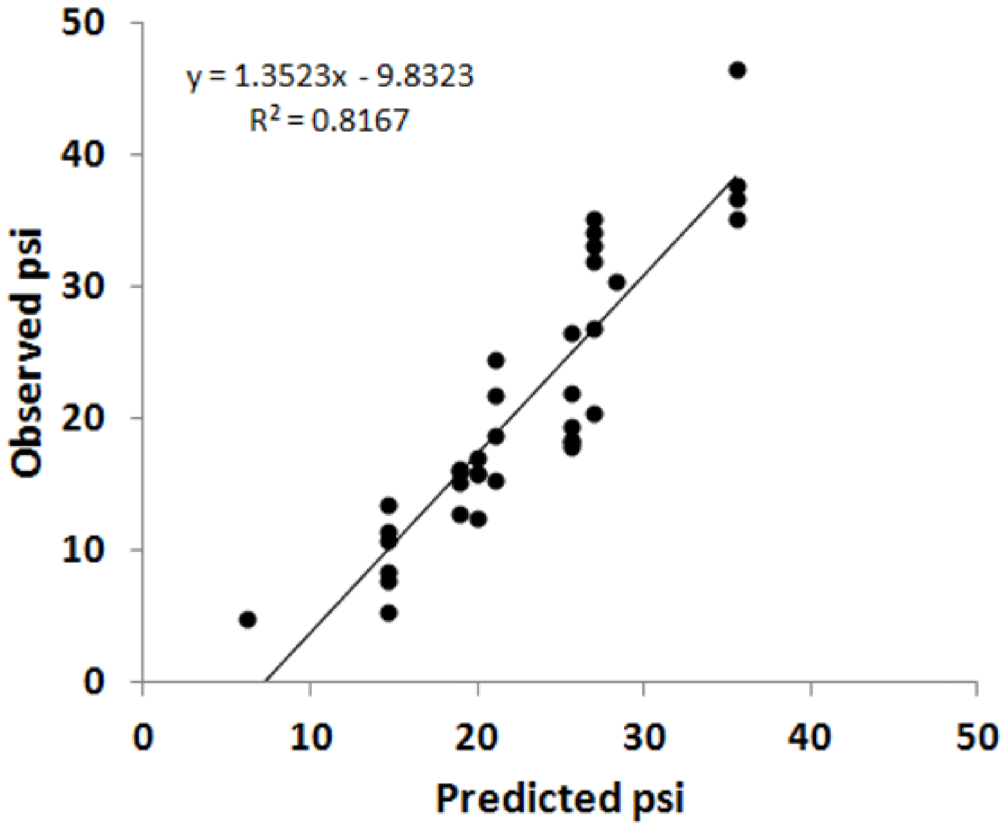


Figure S5.4. The psi of DEs with multiple ESSs as predicted by the model. The values in Table 3.2 were used to predict the psi for all constructs with multiple ESSs. These values were obtained using only single-ESS DEs. The predictions show a good level of correlation: R^2 = 0.82. Although the R^2 is high, the slope and intercept deviate somewhat from what is expected (1 and 0%, respectively) suggesting that the model could be further refined.

Chapter 6

Identifying Candidate Effectors

Introduction

In the late 1970s, it was discovered that genes of higher organisms are interrupted by non-coding intervening sequences (Berget et al. 1977; Chow et al. 1977). After the genetic information is copied into a pre-mRNA molecule, these sequences are removed to generate the mature mRNA. This surprising result has led to studies aiming at elucidating the mechanism that allows the recognition of the regions that are removed, introns, and the ones that are kept and spliced together, exons. The first functional sequences to be defined were at the exon/intron boundaries (Lerner et al. 1980; Mount and Steitz 1981; Mount 1982). However, by the late 1980s there was evidence that sequences inside of the exons themselves played a role (Reed and Maniatis 1986; Mardon et al. 1987; Cooper and Ordahl 1989; Tsai et al. 1989).

At the time of these studies, a new tool was being developed: *in vitro* splicing. In 1983 Dignam and colleagues developed a protein extraction procedure from the nucleus that allowed *in vitro* transcription to be performed efficiently (Dignam et al. 1983). Shortly thereafter, these nuclear extracts were used for splicing pre-mRNA molecules *in vitro* (Krainer et al. 1984). These developments allowed the discovery of the first splicing factor in higher eukaryotes (Ge and Manley 1990; Krainer et al. 1990). This factor, known as SRSF1, has two RNA binding domains known as RRM domains and a region rich in arginine and serine dipeptides known as RS domain (Ge et al. 1991; Krainer et al. 1991; Long and Caceres 2009). The RRM and RS domains became hallmarks of a set of proteins that was shown to affect splicing decisions by preventing exon skipping (Ibrahim et al. 2005). Further work characterized the RNA sequences recognized by many of them (Tacke and Manley 1995; Lynch and Maniatis 1996; Tacke et al. 1997; Liu et al. 1998a; Lou et al. 1998; Long and Caceres 2009) and currently high-throughput approaches are providing tools to refine our understanding of their sequence recognition specificity (Ray et al. 2009; Anko et al. 2012).

Another family of proteins was discovered for its binding to pre-mRNA; these proteins formed heterogeneous nuclear ribonucleoprotein complexes (Pinol-Roma et al. 1988). These so called hnRNPs have a diverse set of functions with respect to RNA metabolism (Han et al. 2010). However, many of these proteins have been characterized as sharing the function of splicing repressors (Martinez-Contreras et al. 2007). The binding sequences for many of these proteins have been elucidated (Han et al. 2010). Even though these proteins have been grouped as a family, their heterogeneity and the diversity of their functions warrant prudence when making generalizations.

In our study of splicing, we have taken a reductionist approach to understand the mechanism by which the proper splice sites are selected. These studies used designer exons made up of a few prototype modules of our own design (including an exonic splicing enhancer, ESE, and silencer, ESS) in a three exon minigene. In this chapter I focus on the identification of the proteins that bind these modules and that provide the functionality observed *in vivo*. Recent advances in mass spectrometry allowed us to efficiently make a fairly comprehensive list of candidates for each of the modules studied.

Results

Sequences used

Four sequences were chosen for proteomics studies. Three of them were introduced in Chapters 2 and 3 and comprised the building blocks for the construction of Designer Exons (DEs): a reference sequence, CCAAACAA; an ESE, UCCUCGAA; and an ESS, CACAUGGU. Another ESS was added for future extensions and was included in this study for it also served as an additional reference for some of the experiments that follow: ESS2, CACAUACA. The first three sequences were presented previously and their effects were described at length (Chapters 2 and 3).

The additional silencing sequence ESS2 was chosen in a similar manner to ESS, so as to not create any other predicted splicing regulatory sequence when placed in the DE. Preliminary tests substituting ESS2 for ESS, indicated that ESS2 had stronger silencer effects than ESS (data not shown) and that a strong set of splice sites was required to fully observe the effects of adding it to the DE. For this purpose, we strengthened the 3'SS of the DEs used (SS Set 5 in Table 3.1) and tested the effect of the ESS2 at all six positions defined previously along the exon. For reference purposes, a DE with a sole ESS at position 4 was included as well as a DE with only reference sequences. For ESS2, the psi decreased substantially from 92% to 42% or less for all positions (Fig. 6.1) and these differences were statistically significant (t-test, p<0.01). On the other hand, the effect of the original ESS was statistically significant (t-test, p<0.01) but modest: a reduction of 3%. This confirmed that ESS2 was stronger than ESS.

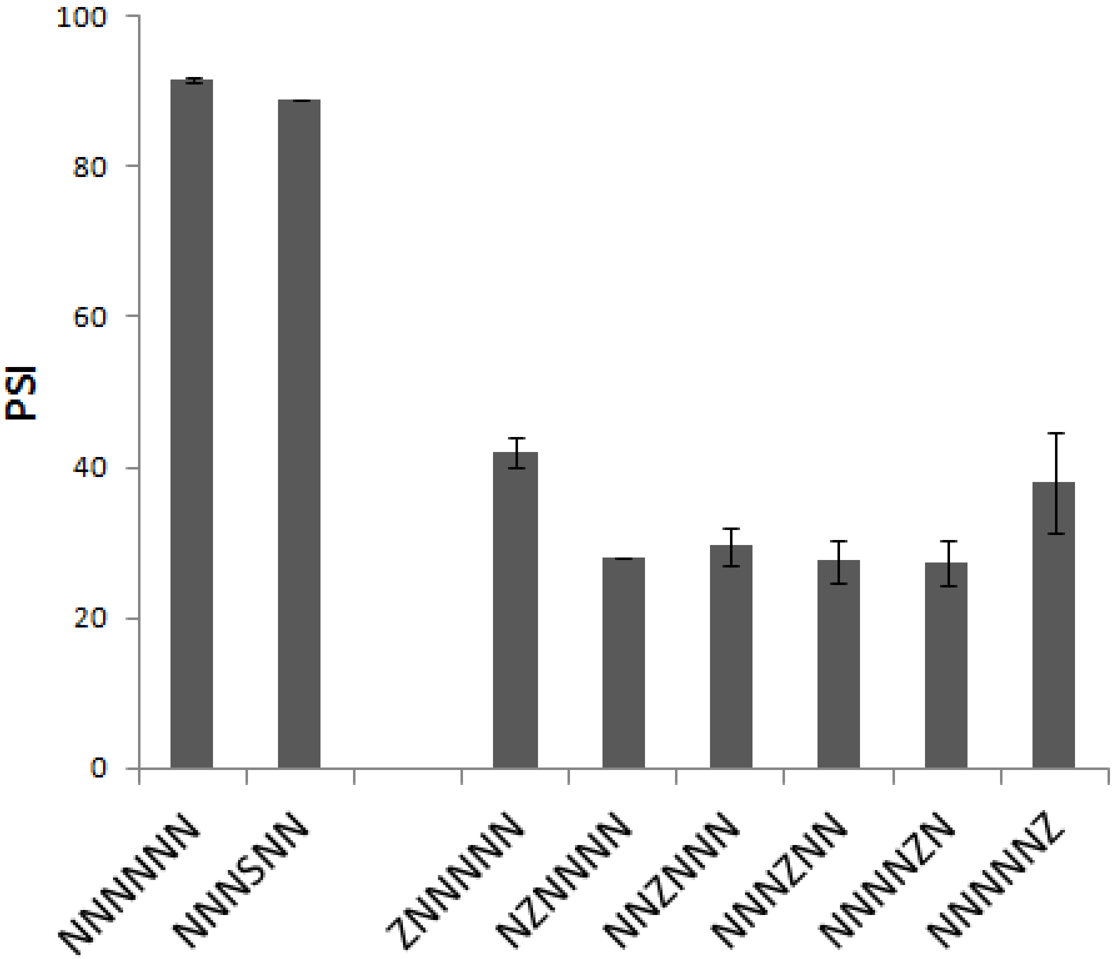


Figure 6.1. ESS2 has stronger silencer activity than ESS and works at all positions tested. ESS2 was placed in all six positions in a 110 nt DE. For all positions, the effect was substantial, reducing the psi from ~92% to ~42% or less. For reference purposes, an ESS was tested at position 4. The decrease was statistically significant (p<0.01) but modest: from ~92% to ~89%. n=3. Exons are indicated using a six letter code indicating the sequence present at each position: N, reference sequence; S, ESS; Z, ESS2.

The variation in the effect of the ESS2 with position was small: from a decrease of 50% at either end to a fairly uniform decrease of ~64% in the middle of the DE. However, a statistically significant difference was found between the first position and each of the four intermediate positions (t-test, $p<0.05$). No other statistically significant difference was found between the positions tested.

Interestingly, in the presence of a weaker 3'SS and a reference psi of ~50%, a reduction of ~10% was observed using the analogous construct for ESS (see Chapter 3). This indicates that the observed effect of ESS depended on the quality of the splice sites. A similar observation was made regarding the effect of ESEs in Chapter 3 and might be a general phenomenon. Indeed, this observation is consistent with saturation of the psi near 0% and 100%. Hence, the presence of strong splice sites generated a high psi value for the DE composed of reference sequences which in turn led to a saturation effect that resulted in a smaller reduction in psi upon the addition of an ESS. Alternatively, considering the ESS-containing DE, the high psi signals saturation. Hence, removal of the ESS causes an increase that is constricted due to its proximity to 100%.

Proteins bound to the ESE, ESS, ESS2 and reference sequences

We performed protein binding experiments using 32 nt biotinylated RNA molecules. Four different molecules were evaluated. The first three RNA molecules evaluated the ESE, ESS and ESS2 sequences separately using two copies of the corresponding sequence. These copies were placed in a sequence environment that duplicates that found in DEs by including a reference sequence in between copies, the 3' half of the reference sequence at the 5' end of the

RNA molecule and the 5’ half at the 3’ end (see Materials and Methods). An analogous fourth RNA molecule consisting exclusively of reference sequences was evaluated to identify proteins bound to the reference sequences themselves. These RNA molecules were bound to streptavidin beads and exposed to nuclear extract. After several washes, the proteins bound were released by digesting the RNA molecules with RNase A. (See Materials and Methods for further details.)

An initial assessment of the differential binding of proteins to the RNA molecules was performed using a sensitive zinc-imidazole reverse staining assay (Fernandez-Patron et al. 1992). However, it should be taken into account that due to the use of relatively high amounts of RNase A and BSA (see Materials and Methods), the presence of proteins at around 15, 35 and 70 kDa was obscured by their bands: the first two bands were observed when purified RNase A was run by itself with the second band probably due to dimerization (Crestfield and Fruchter 1967; Liu et al. 1998b); the third band was observed when purified BSA was run by itself (data not shown). Additionally, proteins of around 30kDa or less are not detected well with this reverse staining assay. Only low stringency washes were performed after exposing the RNA molecules to nuclear extract to ensure good coverage yielding a fairly “crowded” gel. In spite of this, some differences were observed (Fig. 6.2). Among the differences observed, a ~57 kDa band was present only in the ESE lane, ~80 and ~50 kDa bands were only present in the ESS lane and the ratio between the band at ~42 kDa and the band at ~41 kDa was higher for ESS2 indicating at least an enrichment or a depletion for proteins in one of these two bands. Further experiments without BSA, using pre-passed nuclear extract and running higher stringency washes yielded a bright band at ~68 kDa for ESS2 only (data not shown). This band was identified as hnRNP L through tandem mass spectrometry. In spite of these promising results, the presence of a high number of bands could potentially hide important differences.

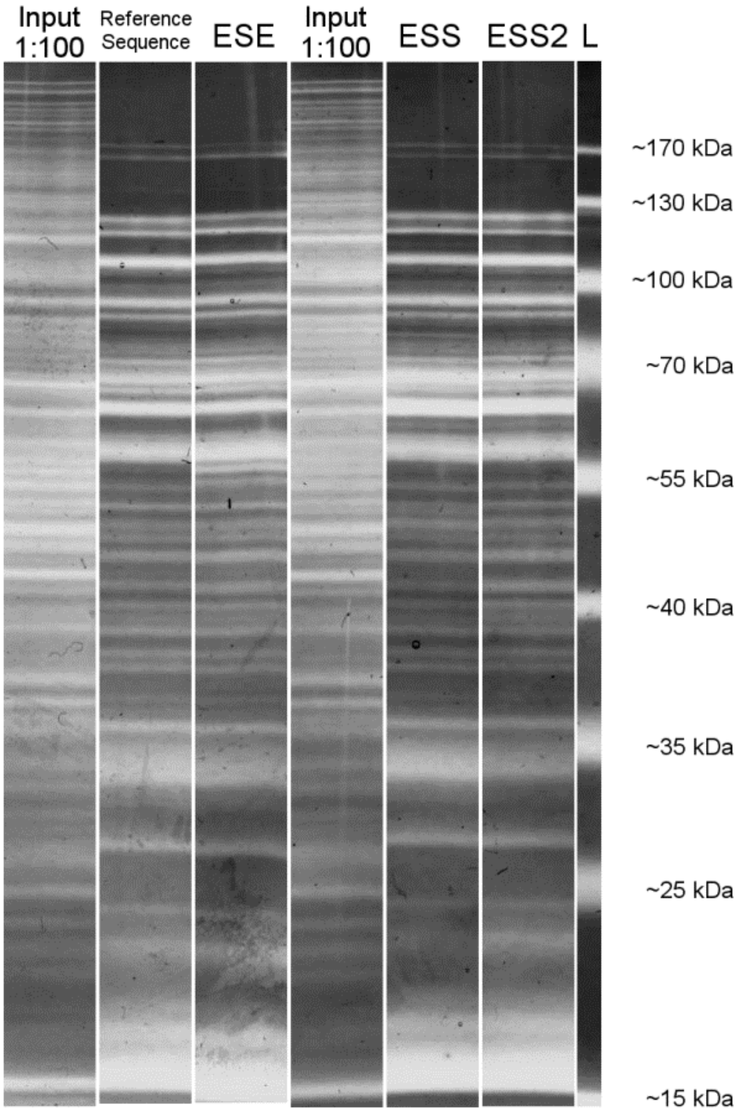


Figure 6.2. Proteins pulled down by RNA baits for the reference sequence, ESE, ESS and ESS2. RNA binding experiments were performed and the proteins released upon treatment with RNase A were electrophoresed in a 10% polyacrylamide gel and stained with zinc-imidazole; the bands are white on a dark background. The lanes shown correspond to two gels and were combined using the ladder and the input lanes for alignment.

To identify the proteins with differential binding as well as to make a more exhaustive comparison, the proteins bound to the RNA sequences were trypsinized on the beads themselves without the use of RNase A and the mixture of the resulting fragments was analyzed by liquid chromatography linked to mass spectrometry for the intact fragments alternating with mass spectrometry for their dissociation products (LC-MS/MS). Comparison of the resulting fingerprints with those expected for known proteins in human allowed identification of the proteins as well as label-free quantification of their abundance (Zhu et al. 2010). We then performed comparisons of the levels of the different proteins and identified those enriched for each one of the different RNA bait molecules (Materials and Methods).

A total of 26 proteins with differential enrichments were found (Table 6.1). A subset of the proteins identified showing their quantification value is shown in Fig. 6.3. Some of these proteins represent different isoforms from the same gene. The case of hnRNP D0 is particularly interesting: isoform 1/a/Dx9 was enriched for ESS2 while isoform 2/b/Dx4 was enriched for ESE (Fig. 6.3G and 6.3I). Furthermore, isoform 3/c/Dx7 was present in the results and barely missed the cut for statistical significance (data not shown): its enrichment profile was similar to that of isoform 1/a/Dx9. Isoforms 1 and 3 share an alternative exon 2, which is missing from isoform 2. Not all cases were as complex. For hnRNP A1, all three isoforms were identified as being enriched for the ESS2 RNA molecule. For some of the proteins in Table 6.1, only the isoform reported showed noticeable differences between the RNA baits: e.g., SRSF7 isoform 2 showed no enrichment between RNA molecules (data not shown). However, for some of them, other isoforms showed a similar profile but had, for at least one of the comparisons, a p-value higher than the threshold used, 7.3x10^{-5}: PTB/hnRNP I isoform 2 showed enrichment for ESE but the p-values were above-the-threshold: around 0.0002 for the three comparisons involved.

This set represents candidate proteins for providing the functional characteristics associated with each sequence. For the reference sequence, only two proteins showed enrichment: HqkI and hnRNP Q. For ESS, only one protein showed enrichment: CSTF1/CstF-50. Interestingly, eight proteins showed depletion: SRSF3, hnRNP K, hnRNP A0, hnRNP A2/B1, hnRNP A3, RNP L, MAP4 and p100 co-activator. For ESS2, many proteins showed enrichment: hnRNP proteins A0, A1, A2/B1, A3, D0 (isoform 1/a/Dx9), L and U, DEAH box protein 36, DAZ-associated protein 1, and p100 co-activator. Only one protein showed depletion: ZC3H4. For ESE, many proteins showed enrichment: hnRNP proteins E2, I/PTB, K and D0 (isoform 2/b/Dx4), Matrin-3, SRSF7/9G8, and DAZ-associated protein 1. Two proteins showed depletion: hnRNP L and PPIase PIN4.

These results explain some of the previous observations but not all. For ESS, the band at around 50 kDa could be due to CSTF1/CstF-50, but the band at ~80 kDa remains a mystery. For ESS2, the previously identified hnRNP L was detected anew. The higher intensity of the band at ~42 kDa could be due to DAZ-associated protein 1, which has a predicted molecular mass of 43 kDa. For ESE, the band at around 57 kDa can be explained by PTB/hnRNP I: isoform 1 has predicted molecular mass of ~57 kDa; a doublet would be expected if isoform 2 was considered as mentioned before for it has a predicted mass of 59 kDa. However, due to variation in protein migration due to post-translational modifications, more analyses would be required to assign the differences observed to the corresponding proteins.

Table 6.1. Proteins showing differential enrichment with the four RNA molecules.

Protein	Isoform reported[1]	Accession Number[2]	Reference Sequence[3]	ESS	ESS2	ESE
CSTF1/CstF-50	-	Q05048		Enrichment		
DAZ-associated protein 1	1	Q96EP5			Enrichment	Enrichment
DEAH box protein 36	2	Q9H2U1			Enrichment	
hnRNP A0	-	Q13151		Depletion	Enrichment	
hnRNP A1	A1-B	P09651			Enrichment	
hnRNP A1	A1-A	P09651			Enrichment	
hnRNP A1	2	P09651			Enrichment	
hnRNP A2/B1	B1	P22626		Depletion	Enrichment	
hnRNP A3	1	P51991		Depletion	Enrichment	
hnRNP D0	1/a/Dx9	Q14103			Enrichment	
HnRNP D0	2/b/Dx4	Q14103				Enrichment
hnRNP E2/Alpha-CP2	1	Q15366				Enrichment
hnRNP K	2	P61978		Depletion		Enrichment
hnRNP L	1	P14866			Enrichment	Depletion
hnRNP Q	1	O60506	Enrichment			
hnRNP U/SAF-A	Long	Q00839			Enrichment	
HqkI	6	Q96PU8	Enrichment			
MAP-4	6	P27816		Depletion		
Matrin-3	1	P43243				Enrichment
p100 co-activator/SND1	-	Q7KZF4		Depletion	Enrichment	
PPIase Pin4	2	Q9Y237				Depletion
PTB/hnRNP I	1	P26599				Enrichment
RNPL	-	P98179		Depletion		
SRSF3/SRp20	-	P84103		Depletion		
SRSF7/9G8	1	Q16629				Enrichment
ZC3H4	-	Q9UPT8			Depletion	

[1] The information for the isoform was obtained from the mass spectrometry report. When no isoform was indicated it was assumed that the canonical isoform was used: isoform 1 in most cases. For most proteins, the other isoforms were present in the output file confirming this fact and showed differences that were not statistically significant. However, for hnRNP E2, hnRNP L and Matrin-3 no other isoforms were found in the report; isoform 1 was assumed. For proteins for which there are no variants in the UniProt database a dash was used to indicate this fact.

[2] UniProt Database

[3] For hnRNP Q and HqkI, the other three RNA molecules showed depletion. This was interpreted as enrichment for the reference sequence.

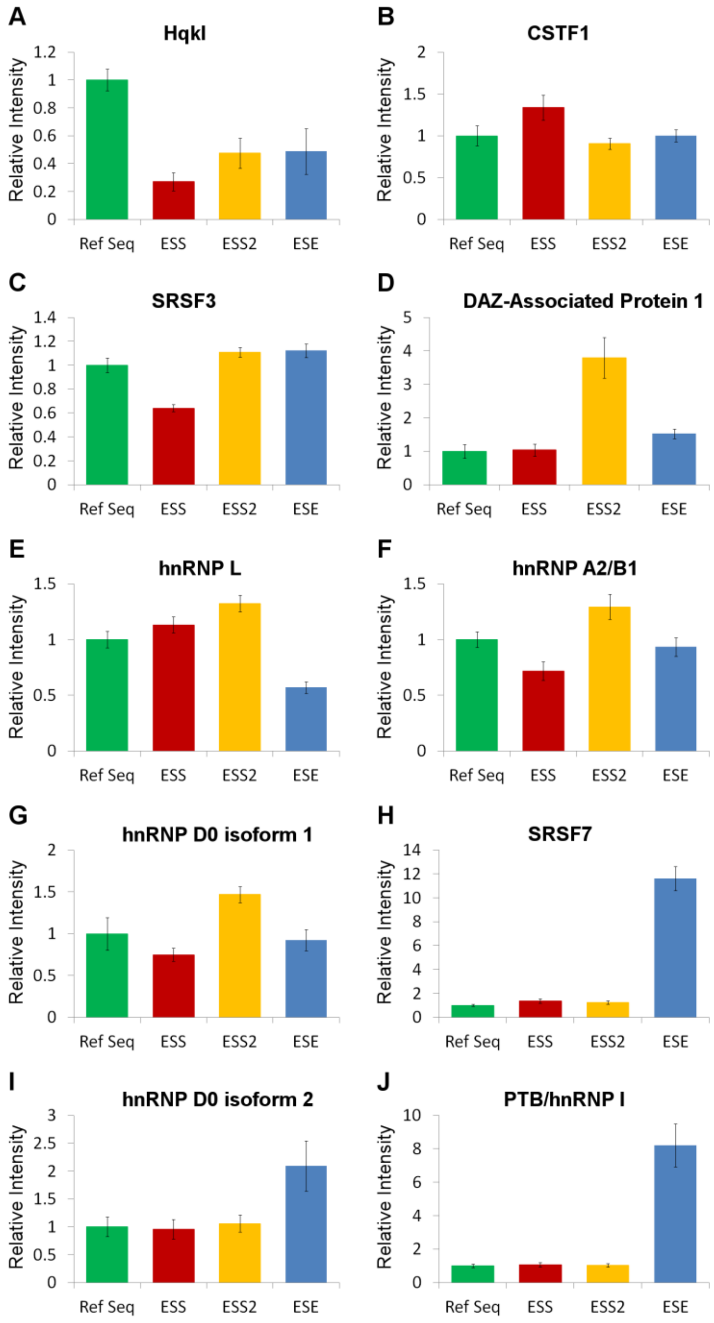


Figure 6.3. Proteins are enriched differentially for the four bait sequences. A. HqkI shows enrichment for the reference sequence and a non-statistically-significant depletion for the ESS bait. B. CSTF1/Cstf-50 shows enrichment for ESS. C. SRSF3 shows depletion for ESS. D. DAZ-associated protein 1 shows enrichment for ESS2 and ESE. E. hnRNP L shows significant enrichment for ESS2 and depletion for ESE. F. hnRNP A2/B1 shows enrichment for ESS2 and depletion for ESS. G. hnRNP D0 isoform 1 shows enrichment for ESS2. H. SRSF7 shows dramatic enrichment for ESE. I. hnRNP D0 isoform 2 shows enrichment for ESE. J. PTB/hnRNP I shows dramatic enrichment for ESE. (See Materials and Methods.)

In order to validate the results of these experiments, two proteins were selected for spot checking through western blots. The presence of PTB/hnRNP I and hnRNP L was assessed in the supernatants after treating the beads with RNase A. As shown in Fig. 6.4, the results of the western blot closely follow the results obtained through mass spectrometry. For PTB/hnRNP I the resolution of the scanner used to document the western blots was unable to capture the presence of an observed doublet for PTB/hnRNP I. This doublet was expected based on the mass spectrometry results. For ESS, the band for hnRNP L came out weaker than expected.

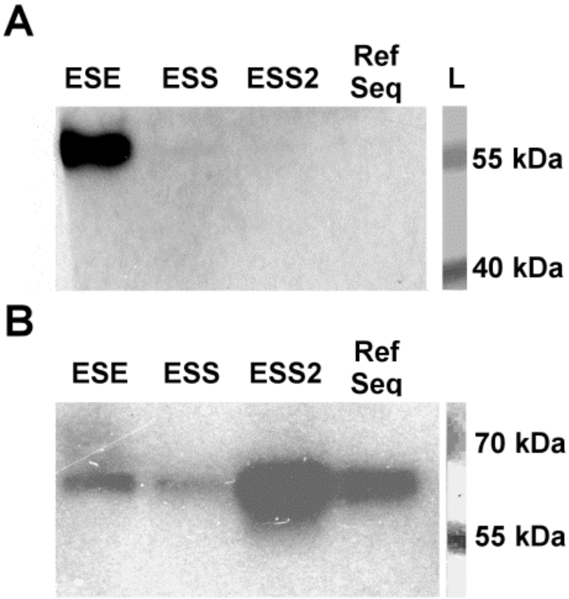


Figure 6.4. Western blots confirm the mass spectrometry results. A. A western blot for PTB/hnRNP I detects PTB only for the ESE bait as expected. B. A western blot for hnRNP L detects this protein for all baits but the intensity is much greater for ESS2.

From the results obtained, it is interesting that many of the proteins identified were missed when analyzing the gel. For instance, many of the proteins had molecular masses below 30 kDa: PPIase Pin4 and SRSF3 for example. Other proteins would require modifications to the zinc-imidazole assay for identification for they were hidden "behind" highly abundant proteins with similar migration rates: hnRNPL and many of the proteins at around 35 kDa are good examples.

To assess a link between the functional characteristics of the sequences used and the proteins found, *in vivo* experiments were planned. Preliminary siRNA experiments were performed for hnRNP K. As shown in Fig. 6.5A, a significant reduction in the level of hnRNP K mRNA was maintained after 5 and 6 days. Since this protein was enriched for the ESE bait, it was expected that knocking it down would reduce the psi of DEs that include an ESE. A decrease of ~10% was observed (Fig. 6.5B) and it was statistically significant after 6 days in comparisons with the three controls (t-test, p<0.05). This is seemingly consistent with hnRNP K having a positive effect on the ESE-containing DEs. However, a more dramatic reduction of ~20% was observed for DEs that did not contain the ESE. This difference was statistically significant after 5 and 6 days (t-test, p<0.05). These results indicate that hnRNP K has a positive effect on the inclusion of the DE both in the presence and in the absence of ESEs in the DE. The observations for the NNNNNN DE might be effected through hnRNP K actions not involving the DE per se. For the NNENNN DE, on the other hand, a comparison between knockdown cells indicates an ESE-linked increase of ~40%. This is higher than the increase of ~30% observed in the negative controls indicating a negative effect of hnRNP K through the ESE.

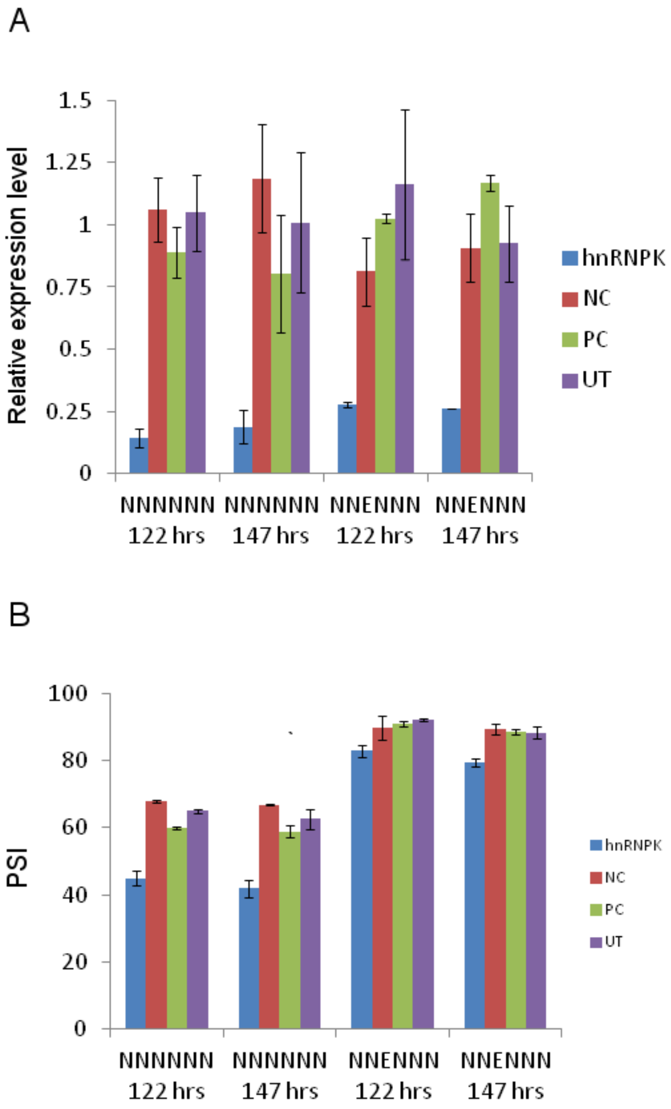


Figure 6.5. Knock down results make hnRNP K an unlikely candidate to explain the effect of ESE. A. Two cell lines were simultaneously tested with hnRNP K knock downs. These cell lines have a modified DHFR minigene in the same chromosomal location but contain different DEs. NNNNNN represents a DE containing no ESEs whereas NNENNN contains a single ESE on the third position (see Chapter 3). All DEs use SS Set 5 (see Chapter 3). Quantification of hnRNP K mRNA was performed at two different time points: 5 and 6 days after the initial transfection. hnRNP K was effectively knocked down. Three controls were used: NC, negative control using an innocuous siRNA molecule; PC, positive control for siRNA using cyclophilin B (no effect expected on hnRNP K); UT, untransfected cells. The expression data was normalized to the average of the three controls. B. PSI for the DHFR minigenes was measured for the cells with hnRNP K knock downs and for the controls. After knock down the effect of the ESE is more pronounced (see text). Error bars show range. n=2.

Discussion

We have introduced a new sequence, ESS2, and added it to the set of sequences we used in DEs. ESS2 showed a stronger silencer effect than ESS and this effect was present in all positions tested. Only minor decreases in its intensity were found in the terminal positions.

Identification of the proteins involved in splicing that are recruited by the different sequences studied here allows the whole set of results previously obtained to be connected to other studies in the literature. These connections might inform new approaches for the study of DEs or for the study of the protein themselves, which is not limited to our research group. Due to the ease and comprehensive nature of shotgun proteomics, such approach was used and several candidates were obtained for every sequence. All these candidates have to be evaluated carefully in order to assign functionality but priorities can be assigned to direct those efforts based on what is known about each candidate. A short list of the candidates is presented in Table 6.2.

Table 6.2. Short list of candidate proteins for providing the observed functionality of each sequence.

Reference Sequence	ESS	ESS2	ESE
HqkI	CSTF1/CstF-50 CSPF5 CSPF7 SRSF3/SRp20 p100 co-activator/SND1	hnRNP L	SRSF7/9G8 hnRNP L

For the ESE, many of the candidates obtained have shown effects that are opposite the ones required for the function studied here. Of these, four enriched candidates belong to the family of hnRNP proteins, which has many members with silencer effects (Han et al. 2010): hnRNP D0, hnRNP E2, PTB/hnRNP I and hnRNP K. In the case of hnRNP K, results consistent with such a silencer effect mediated by the ESE were obtained through siRNA experiments. A further candidate belonging to this family, hnRNP L, shows depletion, which is consistent with a positive effect on psi due to the presence of an ESE. The isoform for Pin4 reported as a depleted candidate includes a mitochondrial localization signal. Even though similarities exist between PIN4 and PIN1 (Sekerina et al. 2000), which is known to affect the conformation of the CTD in RNA polymerase II (Xu and Manley 2007) and potentially affect splicing, PIN4 is not predicted to serve a similar role but to function instead in ribosome biogenesis (Fujiyama-Nakamura et al. 2009) or in the mitochondria (Kessler et al. 2007). There is little information in the literature about the candidate DAZ-associated protein 1 at this time but it might be associated with intronic splicing enhancers (Pastor and Pagani 2011) and probably has silencing effects when placed in the exon (Goina et al. 2008) making it an unlikely candidate for ESE. Matrin-3 and PTB/hnRNP I have been isolated by their association with an U/C rich RNA sequence (Sharma 2008) and in the case of PTB/hnRNP I a sequence containing UCCU, which is found in ESE, was found to interact strongly with PTB/hnRNP I (Ray et al. 2009). The proteins hnRNP K, hnRNP E2 and SRSF7 were found to interact with Matrin-3 by yeast two-hybrid assays (Zeitz et al. 2009), suggesting an indirect interaction with the RNA bait. However, hnRNP K has also been shown to interact with Matrin-3 by co-immunoprecipitation in a RNA dependent fashion (Salton et al. 2011), making RNA mediated interactions a possible explanation for the yeast two-hybrid observations. Additionally, hnRNP E2 has been shown to interact with UC rich sequences (Yeap et al. 2002). All of these results are consistent with Matrin-3, hnRNP K, hnRNP E2 and PTB/hnRNP I interacting with U/C rich sequences similar to the ones in ESE. Binding of SRSF7 requires more consideration for the characterized binding sites are fairly different from the sequences present in ESE (Lynch and Maniatis 1996; Cavaloc et al. 1999; Schaal and Maniatis 1999). This might indicate that one of the other proteins is recruiting it or that the ESE bait contains a new type of binding sequence for SRSF7. In any case, SR proteins are associated with enhanced splicing and its recruitment/binding might provide the functionality observed for ESE. From this analysis, SRSF7 and hnRNP L emerge as the leading candidates to explain the enhancing effect of ESE.

For ESS the situation is more complicated. Only CSTF1 showed enrichment with the threshold used. Binding of this cleavage stimulation factor to the DE might interfere with the proper recruitment of the splicing machinery. Interestingly, relaxing the threshold to 0.001 uncovers only two other enriched proteins, CPSF5 and CPSF7, which show p-values of 0.0003 or lower and are functionally related to CSTF1. A possible participation of factors involved in pre-mRNA 3' processing is reminiscent of the role of CPSF1 in the inclusion exon 6 in IL7R, where CPSF1 seems to interfere with spliceosome binding without causing cleavage (Evsyukova et al. 2013) and the active role of “silenced” polyadenylation signals in transcripts (Almada et al. 2013). Additional candidates include the proteins that were depleted. However, hnRNP A0, hnRNP B1, hnRNP A3 and hnRNP K belong to the hnRNP family and due to their predicted silencing effect (Revil et al. 2009; Han et al. 2010), depletion would not explain the observed effects of ESS. Similarly, the depleted candidate RBM3 has been linked to splicing changes for CD44 (Zeng et al. 2013) but the proposed activity is of silencing, opposite that needed for a depleted candidate. SND1, on the other hand, has been shown to facilitate the assembly of the spliceosome (Yang et al. 2007) making it a likely candidate. For MAP4 more information on its roles in splicing would be needed. The other depleted candidate, SRSF3, has been shown to facilitate splicing when placed in exons (Long and Caceres 2009). Therefore, substituting a sequence that has low affinity for SRSF3 instead of one that has high affinity would decrease psi and would look as if a “silencer” had been added and would coincide with the treatment of ESS in the model in Chapters 4 and 5. A final candidate that might satisfy the requirements is hnRNP R which is enriched for the ESS sequence (maximum p-value of 0.0025) but did not reach the conservative threshold used. This protein has the fourth lowest p-value for a protein showing enrichment for ESS over the other sequences and migrates at around 80 kDa (Pinol-Roma et al. 1988), which might explain the band observed in the zinc-imidazole gel (Fig. 6.2). However, there is little information about its function in the literature. These considerations make CSTF1, CPSF5, CPSF7, SRSF3 and SND1 favored candidates to explain the effect of ESS.

The manner in which these proteins contribute to the silencing effect of ESS might be more complex than suggested in the previous paragraph. CTSF1 is part of a complex that participates in pre-mRNA 3’ end processing (Mandel et al. 2008). Binding of this complex to ESS might interfere with proteins binding to nearby regions explaining the depletion of many proteins observed in the mass spectrometry results for ESS (Table 6.1). In particular, it might interfere with binding of proteins in the flanking reference sequences. Since a binding sequence for SRSF3, UCAAC (Anko et al. 2012), is similar to ACAAC in the reference sequence, the substitution of ESS for a reference sequence in a DE would contribute to a decrease of bound SRSF3 in two ways: the substitution of a SRSF3 binding sequence with a non-binding sequence per se and the interference with SRSF3 binding to other flanking reference sequences. Furthermore, this interference might be related to transcription along the RNA molecule and could potentially explain the position effect observed for ESS in Chapter 3: proteins binding to parts of the molecule that are synthesized first might interfere with binding of proteins to parts of the molecule that are synthesized later. Therefore, U2AF65 having early access to the synthesized RNA (Ujvari and Luse 2004) might prevent a big complex from forming near the 5' end of the exon allowing SRSF3 binding to the abutting reference sequences. At the other end, the presence of ESS near the 3' end of the exon might interfere with U1 snRNP binding, the delay in synthesis effectively providing enough time for the complex to form and negatively affecting inclusion of the exon. Taking all this into consideration, the expected effect would then resemble Fig. 3.5: no significant effect for ESS when placed near the 5' end of the exon, a uniform intermediate effect in the middle of the exon and a stronger effect near the 3' end of the exon.

For ESS2, hnRNP L binds preferentially to CA and TA repetitions such as those present in ESS2. Indeed ESS2 is listed as a high scoring binding site for hnRNP L (Hung et al. 2008). The other hnRNP proteins reported would probably be recruited by bound hnRNP L molecules (Chiou et al. 2013). Additionally, DAZ-associated protein 1 has been shown to interact through its RRM domains with hnRNP U and hnRNP A1 (Yang et al. 2009), which is consistent with hnRNP L being the only protein directly bound to the ESS2 while the rest are recruited directly or indirectly by it. Information about a role for ZC3H4 in splicing was lacking. Therefore, its possible role in splicing DEs remains tentative. The presence of SND1 seems counterintuitive for it facilitates the assembly of spliceosomes. Taking all this into account, a very likely candidate for explaining the effects of ESS2 is hnRNP L.

Regarding the reference sequence, it was found that two proteins were enriched when it was used as bait: HqkI and hnRNP Q. The function of HqkI depends on where it binds on the RNA molecule: stabilization, localization and translation of mRNA when bound to the 3'-UTR (Saccomanno et al. 1999; Li et al. 2000; Lakiza et al. 2005; Zhao et al. 2010; Zearfoss et al. 2011) and facilitator of exon inclusion when bound to intronic splicing enhancers (Hall et al. 2013). Taking into account that several factors that have a silencing role when their binding sites are placed in exons have an enhancing role when placed in introns (Martinez-Contreras et al. 2006), we can hypothesize that HqkI might display a silencing effect when its binding sequence is placed in exons. This is consistent with the likely silencer effect predicted by the model for the reference sequences (Chapter 3). The other candidate, hnRNP Q, was characterized as a component of the spliceosome (Mourelatos et al. 2001) which also bound exon 7 in SMN2 and promoted its inclusion (Chen et al. 2008). However, the sequence involved included many uracils which are absent in the reference sequence, making its role in DEs uncertain. These considerations make HqkI the foremost candidate for binding to the reference sequence.

Materials and Methods

RNA molecules

The biotinylated RNA molecules were obtained from Dharmacon. The sequences were as follows:

reference sequence	5'–Bi–ACAACCAAACAACCAAACAACCAAACAACCAA–3',
ESS	5'–Bi–ACAACACAUGGUCCAAACAACACAUGGUCCAA–3',
ESS2	5'–Bi–ACAACACAUACACCAAACAACACAUACACCAA–3' and
ESE	5’–Bi–ACAAUCCUCGAACCAAACAAUCCUCGAACCAA–3’.

Protein binding experiment

The following buffers were prepared: Buffer B (20 mM HEPES-Na pH 7.9, 20% Glycerol, 42 mM (NH4)$_2$SO$_4$, 0.5 mM DTT and 0.2 mM EDTA), Buffer FW (2 mM MgCl$_2$, 20 mM KCl, 100 µM EDTA, 1mM DTT, 0.25 units/µl RNAseOUT—Invitrogen—, 0.4X Buffer B, 0.1% NP-40 and 0.1% w/v BSA—NEB), Buffer PW (2 mM MgCl$_2$, 20 mM KCl, 100 µM EDTA, 1mM DTT, 0.25 units/µl RNAseOUT, 0.4X Buffer B and 0.1% NP-40) and Buffer SMNNE (2 mM MgCl$_2$, 20 mM KCl, 100 µM EDTA, 1 mM DTT and 0.4X Buffer B).

For experiments other than shotgun mass spectrometry experiments, MyOne T1 beads (Invitrogen), 25 ul per sample, were prepared and washed according to the manufacturer’s protocol. Biotinylated RNA, 5 nmol per sample, was mixed with the beads and incubated for 15 min at room temperature using gentle rotation. Nuclear extract from HEK 293 cells (ProteinOne), 15 ul per sample, was diluted 3:4 in Buffer B, incubated for 5 to 10 min at 30°C and used to prepare splicing mix (20 mM creatine phosphate, 500 µM ATP, 2 mM $MgCl_2$, 20 mM KCl, 0.25 units/ul RNAseOUT, 100 uM EDTA, 1 mM DTT and diluted nuclear extract), 50ul per sample. After two washes in Buffer FW, the beads were placed in 50 ul of splicing mix at 30°C for 30 min using gentle rotation. After two washes with Buffer PW and three washes with Buffer SMNNE, the RNA was digested with 500 ng/µl of RNase A (Invitrogen) in Buffer SMNNE at 30°C for 20 min. The supernatant was retrieved and run in 10% polyacrylamide gels.

For shotgun mass spectrometry experiments, MyOne C1 beads (Invitrogen), 20 ul per sample, were prepared and washed according to the manufacturer's protocol. Biotinylated RNA, 4 nmol per sample, was mixed with the beads and incubated for 15 min at room temperature using gentle rotation. Nuclear extract from HEK 293 cells (ProteinOne), 12 ul per sample, was diluted 3:4 in Buffer B, incubated for 5 to 10 min at 30°C and used to prepare splicing mix (20 mM creatine phosphate, 500 µM ATP, 2 mM $MgCl_2$, 20 mM KCl, 0.25 units/ul RNAseOUT, 100 uM EDTA, 1 mM DTT and diluted nuclear extract), 40ul per sample. After two washes in Buffer PW, the beads were placed in 40 ul of splicing mix at 30°C for 30 min using gentle rotation. After two washes with Buffer PW and two washes with Buffer SMNNE, the beads were washed three times in 500 mM ammonium bicarbonate, resuspended in 30 µl of 500 mM ammonium bicarbonate and sent for mass spectrometry at the local facility.

Shotgun mass spectrometry and data processing

For shotgun mass spectrometry, three samples were prepared for each of the four RNA molecules for a total of 12 samples. LC-MS/MS was performed three times for each sample by the local facility and the proteins were identified yielding a total of 36 individual lists that included a quantification index. The samples were segregated into two groups based on principal component analysis of the quantification results (data not shown). These two groups showed different quantification characteristics which might interfere with direct comparisons (data not shown; see below). All three measurements for two samples for the reference sequence and the three measurements for one sample for the ESE constituted the first group. The second group contained the 27 remaining measurements corresponding to the remaining 9 samples, which included a sample for the reference sequence (see below). To reduce false discovery, proteins for which a sole fragment was identified were removed.

Each of the 36 lists was normalized to the total sum of its quantification index to allow comparison between samples in the same group. Subsequently, for each one of the two groups, the measurements for each protein were further normalized to the average of the measurements for the reference sequence for that protein. This normalization was performed to allow comparisons between different groups by at least partially compensating for their differing quantification characteristics. After these normalization steps all the resulting normalized measurements were combined. Proteins for which any measurement was missing were discarded. A list of 686 proteins was obtained. In order to identify proteins with sample specific enrichment, a two tailed t-test was performed using all available measurements for each possible pairing of RNA molecules. Proteins were chosen that had, for all comparisons, a p-value smaller than $7.3x10^{-5}$ (Bonferroni-corrected p-value corresponding to 0.05).

Zinc-imidazole reverse staining and gel documentation

The protocol from Fernandez-Patron et al. (Fernandez-Patron et al. 1992) was used. Briefly, after electrophoresis, the gel was equilibrated in 0.2 M imidazole with 0.1% w/v SDS for 15 min and then exposed to 0.3 M Zn sulfate for ~30 s until the bands became easily visible but were not too sharp. The gel was quickly transferred to a container with double-distilled water and rinsed 5 times for approximately 1 min each time. The gel was then placed in 0.5% w/v sodium carbonate and moved to a tight plastic bag; excess buffer was removed. This gel-containing bag was scanned in a Bio-5000 Microtek gel scanner using the "Transparent" option at high resolution.

Western blots

Western blots were performed using monoclonal mouse antibodies from Santa Cruz Biotechnology against PTB/hnRNP I (SH54) and hnRNP L (4D11). Detection was performed using secondary antibodies linked to alkaline phosphatase and CDP-Star (Roche).

siRNA

Preliminary experiments showed that even though the hnRNP K mRNA levels decreased by 48 hrs, the protein levels had only slightly decreased by 3 days (data not shown). The half life for many of the proteins in Table 6.1 is suspected to be long and the absurd results observed in some of the articles indicate difficulties in their measurements (Boisvert et al. 2012). Therefore, 5 and 6 day siRNA experiments were carried out with two sequential siRNA transfections (Bartlett and Davis 2006). Cells containing a chromosomally integrated minigene with either a NNNNNN or a

NNENNN DE using SS Set 5 (see Chapter 3) were seeded on 12-well dishes using 1 ml antibiotic-free MEM alpha modification (Hyclone) with 10% FBS serum (Atlanta Biologicals): 1 large dish (100 mm) 80% confluent was enough to seed 70 wells. After 24 hours, 60 pmol of siRNA (siGenome Smart Pool, Dharmacon) was transfected using DharmaFECT 1 according to the manufacuter's protocol. After 24 hours, the medium was changed and the cells were split: 60% for 5-day 12-well plate and 40% to 6-day 12-well plate. After 40 hours, the cells were transfected again. After 48 to 58 hours, RNA was extracted from the 5-day plate and the medium exchanged for the 6-day plate. After another 24 hours the RNA was extracted from the 6-day plate.

QPCR

The methodology presented in Chapter 3 was used for quantification of psi. Gamma actin was also quantified as explained in Chapter 3. Changes in hnRNP K were quantified using the delta-delta-ct methodology (Livak and Schmittgen 2001). The primer for reverse transcription was 5'-GCATTCTGTCAAAACCACCTCTT-3'. The primers for QPCR of hnRNP K mRNA were 5'-CACTGGGCGTCCGCGA-3' and 5'-TCATCCTTGATCTTATATCTGAGTCTCC-3'.

Authors' Contributions

Mauricio Arias and Larry Chasin planned the experiments. MA carried out the experiments, performeds the analyses and wrote the text.

Acknowledgements

We thank Lewis Brown for excellent technical support with the experiments involving mass spectrometry. Special thanks to Vincent Anquetil for technical advice. Thanks to the Chasin Lab. This work was supported by a grant from the NIH (GM072740) to L.A.C.

References

Almada AE, Wu X, Kriz AJ, Burge CB, Sharp PA. 2013. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**(7458): 360-363.

Anko ML, Muller-McNicoll M, Brandl H, Curk T, Gorup C, Henry I, Ule J, Neugebauer KM. 2012. The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. *Genome biology* **13**(3): R17.

Bartlett DW, Davis ME. 2006. Insights into the kinetics of siRNA-mediated gene silencing from live-cell and live-animal bioluminescent imaging. *Nucleic Acids Res* **34**(1): 322-333.

Berget SM, Moore C, Sharp PA. 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America* **74**(8): 3171-3175.

Boisvert FM, Ahmad Y, Gierlinski M, Charriere F, Lamont D, Scott M, Barton G, Lamond AI. 2012. A quantitative spatial proteomics analysis of proteome turnover in human cells. *Mol Cell Proteomics* **11**(3): M111 011429.

Cavaloc Y, Bourgeois CF, Kister L, Stevenin J. 1999. The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. *RNA* **5**(3): 468-483.

Chen HH, Chang JG, Lu RM, Peng TY, Tarn WY. 2008. The RNA binding protein hnRNP Q modulates the utilization of exon 7 in the survival motor neuron 2 (SMN2) gene. *Mol Cell Biol* **28**(22): 6929-6938.

Chiou NT, Shankarling G, Lynch KW. 2013. hnRNP L and hnRNP A1 induce extended U1 snRNA interactions with an exon to repress spliceosome assembly. *Mol Cell* **49**(5): 972-982.

Chow LT, Gelinas RE, Broker TR, Roberts RJ. 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**(1): 1-8.

Cooper TA, Ordahl CP. 1989. Nucleotide substitutions within the cardiac troponin T alternative exon disrupt pre-mRNA alternative splicing. *Nucleic Acids Res* **17**(19): 7905-7921.

Crestfield AM, Fruchter RG. 1967. The homologous and hybrid dimers of ribonuclease A and its carboxymethylhistidine derivatives. *J Biol Chem* **242**(14): 3279-3284.

Dignam JD, Lebovitz RM, Roeder RG. 1983. Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res* **11**(5): 1475-1489.

Evsyukova I, Bradrick SS, Gregory SG, Garcia-Blanco MA. 2013. Cleavage and polyadenylation specificity factor 1 (CPSF1) regulates alternative splicing of interleukin 7 receptor (IL7R) exon 6. *RNA* **19**(1): 103-115.

Fernandez-Patron C, Castellanos-Serra L, Rodriguez P. 1992. Reverse staining of sodium dodecyl sulfate polyacrylamide gels by imidazole-zinc salts: sensitive detection of unmodified proteins. *BioTechniques* **12**(4): 564-573.

Fujiyama-Nakamura S, Yoshikawa H, Homma K, Hayano T, Tsujimura-Takahashi T, Izumikawa K, Ishikawa H, Miyazawa N, Yanagida M, Miura Y et al. 2009. Parvulin (Par14), a peptidyl-prolyl cis-trans isomerase, is a novel rRNA processing factor that evolved in the metazoan lineage. *Mol Cell Proteomics* **8**(7): 1552-1565.

Ge H, Manley JL. 1990. A protein factor, ASF, controls cell-specific alternative splicing of SV40 early pre-mRNA in vitro. *Cell* **62**(1): 25-34.

Ge H, Zuo P, Manley JL. 1991. Primary structure of the human splicing factor ASF reveals similarities with Drosophila regulators. *Cell* **66**(2): 373-382.

Goina E, Skoko N, Pagani F. 2008. Binding of DAZAP1 and hnRNPA1/A2 to an exonic splicing silencer in a natural BRCA1 exon 18 mutant. *Mol Cell Biol* **28**(11): 3850-3860.

Hall MP, Nagel RJ, Fagg WS, Shiue L, Cline MS, Perriman RJ, Donohue JP, Ares M, Jr. 2013. Quaking and PTB control overlapping splicing regulatory networks during muscle cell differentiation. *RNA* **19**(5): 627-638.

Han SP, Tang YH, Smith R. 2010. Functional diversity of the hnRNPs: past, present and perspectives. *Biochem J* **430**(3): 379-392.

Hung LH, Heiner M, Hui J, Schreiner S, Benes V, Bindereif A. 2008. Diverse roles of hnRNP L in mammalian mRNA processing: a combined microarray and RNAi analysis. *RNA* **14**(2): 284-296.

Ibrahim EC, Schaal TD, Hertel KJ, Reed R, Maniatis T. 2005. Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. *Proceedings of the National Academy of Sciences of the United States of America* **102**(14): 5002-5007.

Kessler D, Papatheodorou P, Stratmann T, Dian EA, Hartmann-Fatu C, Rassow J, Bayer P, Mueller JW. 2007. The DNA binding parvulin Par17 is targeted to the mitochondrial matrix by a recently evolved prepeptide uniquely present in Hominidae. *BMC Biol* **5**: 37.

Krainer AR, Conway GC, Kozak D. 1990. Purification and characterization of pre-mRNA splicing factor SF2 from HeLa cells. *Genes Dev* **4**(7): 1158-1171.

Krainer AR, Maniatis T, Ruskin B, Green MR. 1984. Normal and mutant human beta-globin pre-mRNAs are faithfully and efficiently spliced in vitro. *Cell* **36**(4): 993-1005.

Krainer AR, Mayeda A, Kozak D, Binns G. 1991. Functional expression of cloned human splicing factor SF2: homology to RNA-binding proteins, U1 70K, and Drosophila splicing regulators. *Cell* **66**(2): 383-394.

Lakiza O, Frater L, Yoo Y, Villavicencio E, Walterhouse D, Goodwin EB, Iannaccone P. 2005. STAR proteins quaking-6 and GLD-1 regulate translation of the homologues GLI1 and tra-1 through a conserved RNA 3'UTR-based mechanism. *Dev Biol* **287**(1): 98-110.

Lerner MR, Boyle JA, Mount SM, Wolin SL, Steitz JA. 1980. Are snRNPs involved in splicing? *Nature* **283**(5743): 220-224.

Li Z, Zhang Y, Li D, Feng Y. 2000. Destabilization and mislocalization of myelin basic protein mRNAs in quaking dysmyelination lacking the QKI RNA-binding proteins. *J Neurosci* **20**(13): 4944-4953.

Liu HX, Zhang M, Krainer AR. 1998a. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev* **12**(13): 1998-2012.

Liu Y, Hart PJ, Schlunegger MP, Eisenberg D. 1998b. The crystal structure of a 3D domain-swapped dimer of RNase A at a 2.1-A resolution. *Proceedings of the National Academy of Sciences of the United States of America* **95**(7): 3437-3442.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**(4): 402-408.

Long JC, Caceres JF. 2009. The SR protein family of splicing factors: master regulators of gene expression. *Biochem J* **417**(1): 15-27.

Lou H, Neugebauer KM, Gagel RF, Berget SM. 1998. Regulation of alternative polyadenylation by U1 snRNPs and SRp20. *Mol Cell Biol* **18**(9): 4977-4985.

Lynch KW, Maniatis T. 1996. Assembly of specific SR protein complexes on distinct regulatory elements of the Drosophila doublesex splicing enhancer. *Genes Dev* **10**(16): 2089-2101.

Mandel CR, Bai Y, Tong L. 2008. Protein factors in pre-mRNA 3'-end processing. *Cell Mol Life Sci* **65**(7-8): 1099-1122.

Mardon HJ, Sebastio G, Baralle FE. 1987. A role for exon sequences in alternative splicing of the human fibronectin gene. *Nucleic Acids Res* **15**(19): 7725-7733.

Martinez-Contreras R, Cloutier P, Shkreta L, Fisette JF, Revil T, Chabot B. 2007. hnRNP proteins and splicing control. *Adv Exp Med Biol* **623**: 123-147.

Martinez-Contreras R, Fisette JF, Nasim FU, Madden R, Cordeau M, Chabot B. 2006. Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. *PLoS Biol* **4**(2): e21.

Mount SM. 1982. A catalogue of splice junction sequences. *Nucleic Acids Res* **10**(2): 459-472.

Mount SM, Steitz JA. 1981. Sequence of U1 RNA from Drosophila melanogaster: implications for U1 secondary structure and possible involvement in splicing. *Nucleic Acids Res* **9**(23): 6351-6368.

Mourelatos Z, Abel L, Yong J, Kataoka N, Dreyfuss G. 2001. SMN interacts with a novel family of hnRNP and spliceosomal proteins. *EMBO J* **20**(19): 5443-5452.

Pastor T, Pagani F. 2011. Interaction of hnRNPA1/A2 and DAZAP1 with an Alu-derived intronic splicing enhancer regulates ATM aberrant splicing. *PloS one* **6**(8): e23349.

Pinol-Roma S, Choi YD, Matunis MJ, Dreyfuss G. 1988. Immunopurification of heterogeneous nuclear ribonucleoprotein particles reveals an assortment of RNA-binding proteins. *Genes Dev* **2**(2): 215-227.

Ray D, Kazan H, Chan ET, Pena Castillo L, Chaudhry S, Talukder S, Blencowe BJ, Morris Q, Hughes TR. 2009. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol* **27**(7): 667-670.

Reed R, Maniatis T. 1986. A role for exon sequences and splice-site proximity in splice-site selection. *Cell* **46**(5): 681-690.

Revil T, Pelletier J, Toutant J, Cloutier A, Chabot B. 2009. Heterogeneous nuclear ribonucleoprotein K represses the production of pro-apoptotic Bcl-xS splice isoform. *J Biol Chem* **284**(32): 21458-21467.

Saccomanno L, Loushin C, Jan E, Punkay E, Artzt K, Goodwin EB. 1999. The STAR protein QKI-6 is a translational repressor. *Proceedings of the National Academy of Sciences of the United States of America* **96**(22): 12605-12610.

Salton M, Elkon R, Borodina T, Davydov A, Yaspo ML, Halperin E, Shiloh Y. 2011. Matrin 3 binds and stabilizes mRNA. *PloS one* **6**(8): e23882.

Schaal TD, Maniatis T. 1999. Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. *Mol Cell Biol* **19**(3): 1705-1719.

Sekerina E, Rahfeld JU, Muller J, Fanghanel J, Rascher C, Fischer G, Bayer P. 2000. NMR solution structure of hPar14 reveals similarity to the peptidyl prolyl cis/trans isomerase domain of the mitotic regulator hPin1 but indicates a different functionality of the protein. *J Mol Biol* **301**(4): 1003-1017.

Sharma S. 2008. Isolation of a sequence-specific RNA binding protein, polypyrimidine tract binding protein, using RNA affinity chromatography. *Methods in molecular biology* **488**: 1-8.

Tacke R, Chen Y, Manley JL. 1997. Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation: creation of an SRp40-specific splicing enhancer. *Proceedings of the National Academy of Sciences of the United States of America* **94**(4): 1148-1153.

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

Tsai AY, Streuli M, Saito H. 1989. Integrity of the exon 6 sequence is essential for tissue-specific alternative splicing of human leukocyte common antigen pre-mRNA. *Mol Cell Biol* **9**(10): 4550-4555.

Ujvari A, Luse DS. 2004. Newly Initiated RNA encounters a factor involved in splicing immediately upon emerging from within RNA polymerase II. *J Biol Chem* **279**(48): 49773-49779.

Xu YX, Manley JL. 2007. Pin1 modulates RNA polymerase II activity during the transcription cycle. *Genes Dev* **21**(22): 2950-2962.

Yang HT, Peggie M, Cohen P, Rousseau S. 2009. DAZAP1 interacts via its RNA-recognition motifs with the C-termini of other RNA-binding proteins. *Biochem Biophys Res Commun* **380**(3): 705-709.

Yang J, Valineva T, Hong J, Bu T, Yao Z, Jensen ON, Frilander MJ, Silvennoinen O. 2007. Transcriptional co-activator protein p100 interacts with snRNP proteins and facilitates the assembly of the spliceosome. *Nucleic Acids Res* **35(**13): 4485-4494.

Yeap BB, Voon DC, Vivian JP, McCulloch RK, Thomson AM, Giles KM, Czyzyk-Krzeska MF, Furneaux H, Wilce MC, Wilce JA et al. 2002. Novel binding of HuR and poly(C)-binding protein to a conserved UC-rich motif within the 3'-untranslated region of the androgen receptor messenger RNA. *J Biol Chem* **277**(30): 27183-27192.

Zearfoss NR, Clingman CC, Farley BM, McCoig LM, Ryder SP. 2011. Quaking regulates Hnrnpa1 expression through its 3' UTR in oligodendrocyte precursor cells. *PLoS Genet* **7**(1): e1001269.

Zeitz MJ, Malyavantham KS, Seifert B, Berezney R. 2009. Matrin 3: chromosomal distribution and protein interactions. *J Cell Biochem* **108**(1): 125-133.

Zeng Y, Wodzenski D, Gao D, Shiraishi T, Terada N, Li Y, Vander Griend DJ, Luo J, Kong C, Getzenberg RH et al. 2013. Stress-Response Protein RBM3 Attenuates the Stem-like Properties of Prostate Cancer Cells by Interfering with CD44 Variant Splicing. *Cancer Res* **73**(13): 4123-4133.

Zhao L, Mandler MD, Yi H, Feng Y. 2010. Quaking I controls a unique cytoplasmic pathway that regulates alternative splicing of myelin-associated glycoprotein. *Proceedings of the National Academy of Sciences of the United States of America* **107**(44): 19061-19066.

Zhu W, Smith JW, Huang CM. 2010. Mass spectrometry-based label-free quantitative proteomics. *J Biomed Biotechnol* **2010**: 840518.

Chapter 7

Conclusions

Designer exons were developed to unravel the complex aspects of natural phenomena. These exons diminish the risks inherent in modifying sequences by controlling and reducing the number of changes made while studying the effect of different parameters on splicing. DEs constructed by random combinations of ESE and ESS modules explored a broad range of sizes and ESE and ESS content, and were shown to generate a gamut of inclusion levels. These exons did not provide a systematic approach to understand splicing but explore their potential in this regard.

A systematic effort was then undertaken to understand the effect of three parameters separately: size, ESE composition and ESS composition. The relationships obtained between these parameters and splicing ratified previous knowledge but contained several surprising results. The effect of varying size on inclusion level showed the existence of an optimum size confirming previous observations. However, these studies also uncovered that the optimum size range shifted according to the splice site sequences used. For a certain size a specific combination of splice sites yielded a higher inclusion level than another. However, for a different size the opposite was true. This variation suggests that previous studies on the inherent "strength" of specific splice sites in a fixed context need to be reconsidered since repeating the experiment in a different fixed context might yield different results.

Regarding the prototypical sequences used, it was found that inclusion level is positively affected by the presence of ESE and that multiple ESEs showed increased effects. This relationship can be linearly approximated. However, contrary to common assumptions, it was found that positional dependence is not essential and that a linear approximation is not the best option for modeling this phenomenon. Importantly, it was found that functional sequences need not be close to the splice sites in order to affect splicing. Indeed it was found that all the sequences used are effective in the middle of the exons and that their effectiveness is sometimes greater when away from the splice sites: ESS2 for example. No support was found for the function of specific sequences varying drastically depending on its position along the exon. One possible explanation for position effects reported in the past is the introduction of different unintended changes depending on the sequences surrounding each target position in previous studies. Some of these changes might have a positive effect and some a negative one that masked the actual effect of the studied sequence.

The use of DEs was complemented by a mathematical treatment that allows the exploration of the predictive power of different mechanisms on splicing outcomes. A framework equation was obtained based on the existence of an exon definition complex by following a cohort of molecules that start synthesis in a negligibly small window of time. No details were assumed about the formation or dissociation of the exon definition complex. The obtained general solution for the model proved fairly complex for intuitive analysis. However, a much more manageable equation was found after certain assumptions were made. This equation had two components. The first describes the depletion of the uncommitted pool of pre-mRNA molecules by commitment to inclusion during the period in which skipping is not yet an option:

the time required to transcribe enough of the pre-mRNA molecule to make skipping an option defines this component. The second describes the competition between inclusion and skipping when they are available at the same time. Both the equations obtained as well as the approach used represent useful tools for the analysis of splicing outcomes, even if incipient efforts to characterize it.

Two mechanisms were tested using the framework equation. The first attempted to explain the splicing outcomes observed when size was varied. For this purpose, interactions across the exon involving its ends were used as the basis of the exon definition complex. For these interactions to occur, the ends were assumed to establish an indirect physical contact. This was modeled treating the RNA molecule as a polyelectrolyte and using statistical mechanics. For ESEs, their facility to form interactions with other proteins involved in the splicing process was used as a basis for proposing a role in the stability of the exon definition complex. This was modeled as changes in the probability of random collisions disrupting the exon definition complex. An analogous model was used for ESSs and reference sequences. All of the above considerations were combined by assuming the contributions were independent of each other. Good performance was obtained for the model both in terms of its ability to reproduce the single-parameter results as well as in terms of its ability to predict the inclusion level of more complex designer exons. A surprising result was that the competition component of the framework equation became insignificant while the time component was entirely responsible for the accurate predictions.

These results are interesting in several ways. First, the biophysical models used are intentionally of sufficient generality to apply to both DEs and natural exons alike. Moreover, these models aligned well with previously reported observations as explained in Chapters 4 and 5 and compare favorably with some other proposed ideas in the literature. Second, the combination of the different contributions was potentially treacherous because interactions between different functional elements require time and effort to sort out. However, good results were obtained by simply assuming independent contributions. Third, some of the results are suggestive of a new understanding for how splicing occurs. As mentioned in Chapter 5, the sufficiency of the time component in explaining the behavior of DEs suggests that no competition between fates is taking place. This is consistent with an intriguing scenario in which exons can commit to inclusion for only a brief window of time. Fourth, even though some exons with similar contents of ESE and ESS modules showed different outcomes, it was found that sufficiently accurate predictions were obtained using general mechanisms. This is reassuring for the focus of these studies is the understanding of natural exons; DEs are only a tool to accomplish that goal. Therefore, the mechanisms used to predict DE splicing give insight into splicing of natural exons avoiding the risks of overanalyzing a simplified system. In a similar vein, while the conclusions reached are constrained by the limited number of sequences used, they still provide a platform to explore the predictive power of new mechanisms in a more controlled system. These mechanisms can then be evaluated in the context of the results available in the literature.

The studies presented here are not aimed at understanding splicing in its entirety. However, they focus on what has been considered a crucial aspect of this process: exon recognition. It has been proposed that this is the discriminatory step in splicing outcomes. It was shown here that modeling this step is enough to predict accurately the behavior of DEs. It was further shown that other observations in the literature are consistent with this idea. This is surprising given the differences between what is modeled in the equations and what is present in many of the insightful studies performed *in vitro*. How can the *in vitro* assay, which ignores the length of natural introns, the order of availability of the different parts of the pre-mRNA molecule and the time dimension of splicing perform so well? A partial explanation could be provided by the expectation that some of the functional elements in pre-mRNA sequences have analogous roles in both intron definition as well as exon definition, explaining the convergence of the results between *in vitro* and *in vivo* results. However, this convergence does not necessarily mean that the observations *in vitro* are accurate representations of the situation *in vivo*.

The proteins identified as candidates are for the most part proteins that have been shown to affect splicing. Further analysis would identify the contribution of each one of these proteins to the effect of the different sequences studied on splicing. This would allow a connection to be established between these results and those of other research groups. However, the effort required and the promising leads in other aspects of this project precluded more research in this area at this time.

Currently new directions are being explored for future research. Many of the ideas highlighted here would require more experiments to become established. Priorities need to be set based to a great degree on currently available technology. For example, the time frame for the putative window of commitment time is expected to be in the seconds range complicating its study *in vivo*. The distances involved in the proposed exon-end interactions are expected to be around ~15 nm, which happens to be too long for FRET and too short for microscopy, even using super resolution techniques: a de facto blind spot. These problems are not unsolvable but their solution might require considerable effort, guesswork and creative solutions. Some of these results though suggest the existence of unexpected mechanisms that open a plethora of more accessible opportunities, even if they are riskier. For example, it can be surmised that a hub might exist and be anchored to the polymerase, maybe its CTD. Properly defined exons attach to this hub shortly after they are synthesized. In this way the small window of commitment time would be caused by the decrease in the probability of collision between the properly defined exons and the hub as the RNA tether between them is elongated by transcription. In this way exons that form an exon definition complex quickly have an advantage over those that take longer. Similarly, exons for which the exon definition complex is more stable have better odds of being captured when bumping into the hub, while those with unstable complexes might miss the few opportunities available to them. Safer options that are feasible with current technology include the study of how optimal range of exon size for splicing is affected by different splice site sequences. Another possibility involves studies to understand the effects of combining SR protein binding sites with hnRNP binding sites in the same DE.

To assess the length of the window of time for commitment, a modified version of the minigene used in Chapter 3 can be used. The downstream intron would be modified to have a long stretch of reference sequences ~500 nt. The middle exon would be a designer exon of ~20nt with SS Set 3 yielding an inclusion level of ~50%. In parallel, a stronger 5’ SS, AAGgtaagt for example, would be evaluated at different positions in the intron to establish the size-psi relationship for exon sizes from ~20 nt to ~400 nt. From the results in Chapter 3, this SS set should provide high psi for most of the range. This will be followed by competition experiments in which the stronger 5' SS will be added downstream of the weaker 5' SS. Since the weaker 5' SS is not able to commit the middle exon to inclusion in the full window of opportunity we can assume that there are exons being committed all along this time period. Depending on the distance between the competing sites, there will be a bigger or smaller overlap in the windows of opportunity of the two 5' SS. Due to the high efficiency of the downstream 5' SS, uncommitted molecules would be "stolen" from the upstream 5' SS reducing its ~50% share. When there is no overlap due to the time delay introduced by the elongation time required to synthesize the sequence separating the two 5' SSs, the ~50% share would remain invariant as the strong 5' SS is placed further downstream. This is the critical parameter to be found and can be expressed in nucleotides or, by simple conversion using the nominal elongation rate of the RNA polymerase II, in seconds.