

UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

Uncharted territory: Exploring the limits of splicing fidelity and the role of splicing in disease

Permalink

<https://escholarship.org/uc/item/3zf1b9jv>

Author

Reynolds, Derrick J

Publication Date

2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Uncharted territory: Exploring the limits of splicing fidelity
and the role of splicing in disease

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biomedical Sciences

by

Derrick James Reynolds

Dissertation Committee:
Professor Klemens J Hertel, Chair
Professor Marian L. Waterman
Professor Yongsheng Shi

2018

TABLE OF CONTENTS

LIST OF FIGURES	iii
LIST OF TABLES	v
ACKNOWLEDGEMENTS	vi
CURRICULUM VITAE	vii
ABSTRACT OF THE DISSERTATION	x
CHAPTER 1	1
INTRODUCTION	1
Pre-mRNA Splicing	1
Splicing Regulation	5
Alternative Splicing	9
The Fidelity of Splicing	14
Splicing and Disease	16
Summary	20
References	22
CHAPTER 2	26
Ultra-deep sequencing reveals splicing as a sequence driven high-fidelity process, constrained by splicing efficiency.	26
SUMMARY	26
INTRODUCTION	27
RESULTS	29
DISCUSSION	62
MATERIALS AND METHODS	72
REFERENCES	75
CHAPTER 3	79
Alternative Splicing in Breast Cancer: A result of aberrant expression of splicing regulators not splicing fidelity.	79
SUMMARY	79
INTRODUCTION	80
RESULTS	85
DISCUSSION	122
MATERIALS AND METHODS	126
REFERENCES	129
CHAPTER 4	138
PERSPECTIVES	138
Splicing Fidelity is High But Fragile	138
Impact of Splicing in Disease	141

LIST OF FIGURES

Figure 1.1	Model for stepwise assembly of the spliceosome	3
Figure 1.2	Model for stepwise assembly of factors influencing exon definition 6	6
Figure 1.3.	RNA secondary structures influence alternative splicing	10
Figure 1.4.	Types of alternative splicing	12
Figure 1.5.	Therapeutic strategy for treatment of Spinal Muscular Atrophy	18
Figure 2.1.	Mutation scheme for SMN1 exon 7 library	31
Figure 2.2.	Analysis of splicing error rates in DNA and RNA reads by error length	35
Figure 2.3.	Mutation rate by position in DNA input reads	38
Figure 2.4.	Cryptic splice site usage	42
Figure 2.5.	Mutant influence on cryptic splice site usage	52
Figure 2.6.	Cryptic SS usage with excluded exon 7	61
Figure 3.1.	Splicing Fidelity in Breast Cancer	87
Figure 3.2.	Log2 Fold Change gene expression of splicing regulators in breast cancer cell lines compared to MCF10A	92
Figure 3.3.	Venn diagram showing differentially spliced cassette exon events in ER- cell lines compared to MCF10A	99
Figure 3.4.	Venn diagram showing differentially spliced cassette exon events in ER+ cell lines compared to MCF10A	101
Figure 3.5.	Venn diagram showing common differentially spliced cassette exon events between ER+ and	107

Figure 3.6.	ER- cell lines when compared to MCF10A Exon features determined by the splicing code to be important in cassette exon inclusion in estrogen receptor positive cell lines	112
Figure 3.7.	Exon features determined by the splicing code to be important in cassette exon exclusion in estrogen receptor positive cell lines	114
Figure 3.8.	Exon features determined by the splicing code to be important in cassette exon inclusion in BT474	116
Figure 3.9.	Exon features determined by the splicing code to be important in cassette exon exclusion in BT474	119
Figure 3.10.	Composite model of exon features determined by the splicing code to be important in cassette exon inclusion and exclusion	121

LIST OF TABLES

Table 2.1.	Average Error Rate – input DNA	40
Table 2.1.	Analysis of splicing error rates in DNA and RNA reads by error length	44
Table 2.3.	5' SS Cryptic Splicing	47
Table 2.4.	Mutant Influence on Splicing Fidelity – Cryptic 3' SS AG/CCTCTGGN10...CAG GAA	51
Table 2.5.	Mutant Influence on Splicing Fidelity - 3'SS AG GGTTTCAG/ACA	55
Table 2.6.	Mutant Influence on Splicing Fidelity – 5'SS GA GTAA/GTCTGC	56
Table 2.7.	Mutation Derived Cryptic 5'SS At Position 27 Results In Truncated Exon 7	57
Table 2.8.	Skipped Exon 7 Cryptic Splicing	60
Table 3.1.	Gene expression is upregulated in subset of splicing related genes	91
Table 3.2.	Gene expression is upregulated in transcription and translation related genes	95
Table 3.3.	Difference between the average fold change of splicing related-genes in ER+ and ER- cancer cell lines compared to MCF10A	98
Table 3.4.	Differentially spliced cassette exons in ER+ cell lines compared to MCF10A	103
Table 3.5.	Differentially spliced cassette exons in ER- cell lines compared to MCF10A	105
Table 3.6.	Differentially spliced events in breast cancer cell lines compared to MCF10A	110

ACKNOWLEDGEMENTS

I would like to express my heartfelt appreciation and admiration for my advisor and committee chair Dr. Klemens (Dicki) Hertel. I feel lucky to have been supported and mentored by such a kind and insightful scientist.

I would like to thank my committee members, Drs. Marian Waterman and Yongsheng Shi, for their help in navigating my projects, and providing invaluable suggestions both in science and in life.

I would like to thank fellow Hertelians, Drs. Will Mueller and Anke Busch. Will, you literally pointed me in the right direction, both in joining the lab and in my projects. Anke, you were extremely generous with your time and knowledge, and I count myself lucky to have known you.

I would like to thank the many friends that I have come to know in my time here. I would especially like to thank Jake Biesinger, for his patience and time, in teaching me the mysteries of bioinformatics. To Michael Salmans, for convincing me to come to Irvine in the first place, and for being a great friend along the way. To James Yu, your willingness to help me in any capacity has been unmatched, and I am extremely grateful for your friendship.

Finally, I would like to thank my family. My wife Rachel, for joining me on this journey, both in life and two graduate programs; come what may, you have been there for me, forever and always. My Mom and Dad, Laurie and Dave, for standing as examples of loving parents, steadfast and honest; I can't imagine I would have made it this far without your guidance and support. To my younger siblings, Dane, Devin and Lisa: while you may have always had to look up to me in stature, know that surely I look to you for examples of how to be more patient, a better parent, and better person. Lastly my daughter, Isabella, you have been an inspiration to me everyday; I hope that this work provides you a love of lifelong learning.

CURRICULUM VITAE

Derrick Reynolds

EDUCATION

University of California – Irvine
PhD: Biomedical Sciences
2010-2018

Brigham Young University
MS: Genetics and Biotechnology
2007-2009

Brigham Young University
BS: Genetics and Biotechnology
2003-2007
Minor: Management

RESEARCH EXPERIENCE

Graduate Research Assistant: Department of Microbiology and Molecular Genetics
University of California-Irvine
(Sept. 2010 – 2018)

Graduate Research Assistant: Department of Plant and Wildlife Sciences
Brigham Young University
(Aug. 2007- Dec. 2009)

Undergraduate Student Project Lead: Department of Plant and Animal Science
Brigham Young University
(Aug. 2005-Jun. 2006; Aug. 2006-Apr 2007)

HSCI Summer Intern: Harvard Stem Cell Institute,
Harvard University
(June 2006-August 2006)

Plant Genomics Intern: Division of Plant Sciences,
University of Missouri-Columbia (June 2005 –July 2005)

Student Project Assistant: Department of Plant/Animal Science,
Brigham Young University
(Jan. 2005-Jun. 2005)

TEACHING EXPERIENCE

Teaching assistant, Developmental and Cell Biology Lab, Department of Developmental and Cell Biology, University of California-Irvine, Irvine, CA (2014)

Teaching assistant, Genetics, Department of Plant and Wildlife Sciences, Brigham Young University, Provo, UT (2006-2008)

PUBLICATIONS

Raney, J.A., Reynolds, D.J., Elzinga, D.B., Page, J., Udall, J.A., Jellen, EN, Bonifacio, A., Fairbanks, D.J., and Maughan, P.J. (2014). Transcriptome Analysis of Drought Induced Stress in *Chenopodium quinoa*. *American Journal of Plant Sciences* 05(03):338-357

Movassat, M., Crabb, T. L., Busch, A., Yao, C., Reynolds, D.J., Shi, Y., and Hertel, K.J. (2016). Coupling between alternative polyadenylation and alternative splicing is limited to terminal introns. *RNA Biology* Vol. 13 , Iss. 7

Zou, D., McSweeney, C., Sebastian, A., Reynolds, D.J., Dong, F., Zhou, Y., Deng, D., Wang, Y., Liu, L., Zhu, J., Zou, J., Shi, Y., Albert, I., and Mao, Y. (2015). A critical role of RBM8a in proliferation and differentiation of embryonic neural progenitors. *Neural Development* Vol. 10, Iss. 18

Reynolds, D.J., Mueller, W., and Hertel, K.J. Ultra-deep sequencing of SMN1 exon 7 reveals low intrinsic splicing error rate and patterns of splicing errors. (In Preparation)

PRESENTATIONS

Alternative Splicing in breast cancer: A result of aberrant regulation not loss of splicing fidelity. 20th International Conference on Intelligent Systems for Molecular Biology, Long Beach, CA (July 2012).

Development, Assembly and Characterization of a *Chenopodium quinoa* EST Collection. Plant and Animal Genomes Conference XVII, San Diego, CA (Jan. 2009).

Oral/Poster: Endothelial differentiation of human embryonic stem cells (hESC) Harvard Stem Cell Institute Internship Program – Cambridge, MA. (Aug 2006).

Genetic Mapping of QTL Conditioning Resistance to Soybean Cyst Nematode in Soybean PI464925B. 2005 Summer Undergraduate Research and Creative Achievements Forum. University of Missouri-Columbia (Aug. 2005).

GRANTS AND FELLOWSHIPS

Biomedical Informatics Training Program – Honorary Member (2011-2013)

ABSTRACT OF THE DISSERTATION

Uncharted territory: Exploring the limits of splicing fidelity
and the role of splicing in disease

By

Derrick James Reynolds

Doctor of Philosophy in Biomedical Sciences

University of California, Irvine, 2018

Professor Klemens J. Hertel, Chair

Pre-mRNA splicing is required for the generation of functional mRNA transcripts. Splicing requires a large dynamic complex called the spliceosome to excise introns and ligate exons. This process requires accurate splice site selection. Various splice sites can be used within a gene to generate many different isoforms. Alternative pre-mRNA splicing is one of the most efficient systems to diversify the proteome. The ability to use different splice sites requires strict and fluid regulation to ensure the correct splicing combinations to safeguard against nonfunctional or deleterious protein isoforms. However this flexibility also permits the possibility of incorrect splice site selection leading to a loss of splicing fidelity.

Using next-generation sequencing on a three-exon min-gene to generate millions of spliced transcripts, we determined the splicing fidelity of each splice site down to the nucleotide level. We demonstrated that the 3' splice site is more prone to splicing errors than the 5' splice site. The rate at which the incorrect splice site was selected ranged between 1 in 631 and 1 in 131,611. Splice site recognition is highly reliant on the

sequence of the mRNA. We demonstrated that single point mutations within the exon could have a drastic effect on splicing fidelity.

Using a bioinformatics approach, we examine the gene expression of splicing related genes in estrogen positive and estrogen negative breast cancer cell lines. We also surveyed the degree of alternative splicing in breast cancer, examining both between breast cancer and normal cell types and between different breast cancers. This analysis confirmed some known cancer related splicing programs and suggest the presence of new targets Using a splicing code we identified a generalized model suggesting that sequence conservation and some splicing regulators are the main components in cassette exon inclusion or exclusion.

CHAPTER 1

INTRODUCTION

In eukaryotic gene expression, pre-messenger RNA (pre-mRNA) is transcribed from DNA in the nucleus of the cell. These pre-mRNAs then undergo several processing steps to generate a mature mRNA transcript. One of these mRNA processing steps is splicing, where introns (the non-coding intragenic regions of the pre-mRNA) are removed and exons are ligated together. The fully spliced mRNA transcript is then shuttled to the cytoplasm to be translated into a protein. This chapter describes the splicing process, the importance of splicing in gene expression, the complex relationship between splicing regulation and splicing fidelity and how this dynamic process connects with disease.

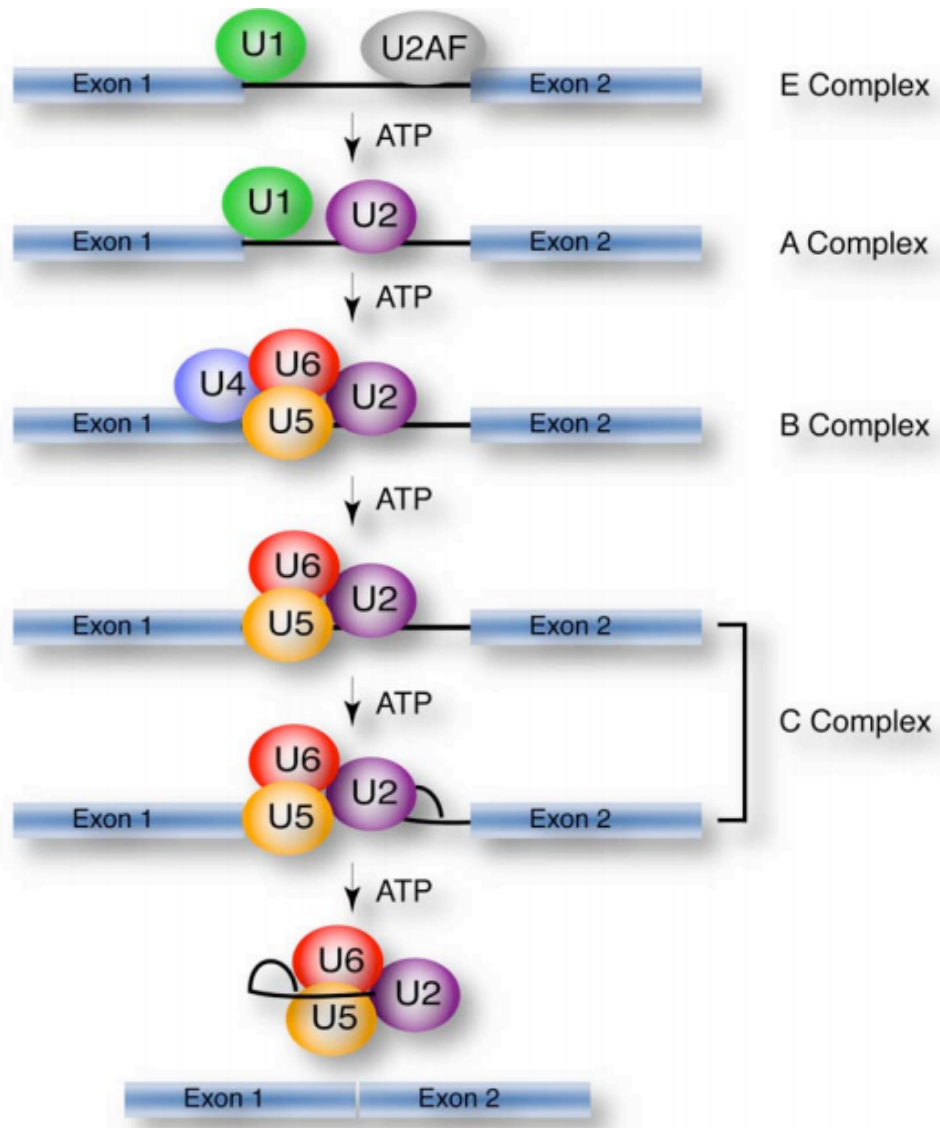
Pre-mRNA Splicing

The splicing of pre-mRNA is catalyzed by the major spliceosome (Black 2003; Wahl et al. 2009). The spliceosome is comprised of five core components, the U1, U2, U4, U5, and U6 small nuclear ribonucleoproteins (snRNPs), along with U2AF1 (U2 small nuclear RNA auxiliary factor 1 or U2AF35), U2AF2 (U2 small nuclear RNA auxiliary factor 2 or U2AF65) (Graveley et al. 2001) and splicing factor 1 (SF1) (Black 2003) which are required for the dynamic assembly and disassembly of more than 300 proteins in a number of different complexes during the splicing process (Wahl et al. 2009).

These spliceosomal complexes, which can form co-transcriptionally or after transcription is completed (Merkhofer et al. 2014), assemble in a sequential pattern: H-complex → E-complex → A-complex → B-complex → C-complex → final spliced mRNA products. H-complex occurs as the pre-mRNA is synthesized, where the RNA is bound non-specifically by hnRNPs (heterogeneous nuclear ribonucleoprotein) (Jurica and Moore 2003).

The recognition of splice sites constitutes the first step of the spliceosomal assembly process. E-complex formation commits the pre-mRNA to splicing (Figure 1.1), and it is defined by U1snRNP binding to the yag|GURAGU (where Y=pyrimidine, R=A/G, and the | denotes the actual splice site) sequence at the 5' splice site (5'SS) through base pairing with the snRNA (small nuclear RNA) component of U1snRNP (Reed 1996; Black 2003; Dou et al. 2006). The 3' splice site (3'SS), also referred to as the splice acceptor site, has three distinct sequence elements generally found within 40 nucleotides upstream of the defined intron|exon junction. These three sequence elements consist of the branch point sequence (BPS), the polypyrimidine tract (PPT), and the actual 3'SS. The canonical 3'SS is defined by the YAG|N sequence at the intron|exon junction. The BPS is a highly degenerate consensus sequence YNYURAY (where Y=A/C and R=A/G) surrounding the conserved branchpoint adenosine (Reed 1996). The PPT is of a variable length characterized by a high percentage

Figure 1.1 Model for stepwise assembly of the spliceosome. Mature spliced mRNA transcripts are generated through a series of spliceosomal complexes that assemble on the pre-mRNA. The core snRNPs of the spliceosome are shown (U1, U2, U4, U5, U6) at each step of the complex ending with the ligated exons and the removed intron (black line).



of pyrimidines (C or U nucleotides) (Reed 1996). Splicing factor 1 (SF1) binds to the BPS, and U2AF1 binds loosely at the 3'SS and U2AF2 binds to the PPT. A-complex occurs when U2snRNP becomes stably associated to the BPS in an ATP-dependent reaction. The U4/U5•U6 tri-snRNP is then recruited to generate B-complex, at which point U1snRNP dissociates. The catalytic spliceosome, or C-complex is formed by the rearrangement of the U4/U5•U6 tri-snRNP, which enables the excision of introns and ligation of exons through two sequential transesterification reactions. Briefly, in the first reaction, the adenosine in the downstream BPS performs a nucleophilic attack on the 5'SS creating an intron lariat intermediate by 2',5'-phosphodiester linkage. The second step is the attack on the 3'SS by the 5'SS causing the removal of the intron lariat and the formation of the fully spliced mRNA transcript.

Splicing Regulation

Given the ability of the spliceosome to generate many different transcript isoforms from a single gene, there must be some way for the splicing machinery to determine the selection of the correct splice sites. Indeed, there are many elements that contribute to exon recognition (Figure 1.2). The splicing of internal exons is influenced by spliceosomal recognition of consensus sequence elements at the 5'SS, 3'SS and BPS regions, by the presence of splicing regulatory elements (SRE), by the intron/exon architecture and by RNA secondary structure (Nilsen and Graveley 2010; Hertel 2008).

Figure 1.2. Model for stepwise assembly of factors influencing exon definition.



Splicing efficiency is dependent on many variable components. Some of these include 3' and 5' splice site strength, splicing regulatory elements (SREs), such as intronic and exonic splicing enhancers (ISEs, ESEs) and intronic and exonic splicing silencers (ISSs, ESSs), and local RNA secondary structures. Each component contributes to the overall affinity of the spliceosome to the exon and, thus, the level of exon inclusion. Figure adapted from (Hertel 2008).



Splice-site strength

YYYYYNYAG 3' ss
 YAG GURAGU 5' ss

Binding sites for exonic splicing regulators

-  Exonic Splicing Enhancer (ESE)
-  Exonic Splicing Silencer (ESS)

Binding sites for intronic splicing regulators

-  Intronic Splicing Enhancer (ISE)
-  Intronic Splicing Silencer (ISS)

Local secondary structures



Splice site strength is generally defined by how closely a 5'SS or 3'SS mirrors the respective consensus splice site. In a 5'SS the complementarity of the splice site to the U1 snRNA determines the strength of the splice site. A 5'SS with high complementarity to U1 snRNA is defined as having a "strong" 5'SS, while low complementarity equates to a "weak" 5'SS. 3'SS strength is determined in a similar fashion, where a longer PPT or higher percentage of pyrimidines in the PPT creates a higher affinity binding site for U2AF1 and U2AF2, and thus, a "strong" 3'SS. There have been several attempts to assign a numerical score to splice site strength (Zhang and Chasin 2004; Yeo and Burge 2004). Indeed, it has been demonstrated that splice site strength scores correlate well with splicing efficiency (Hicks et al. 2010). In our analyses we follow the MaxEntScan method defined by Yeo et al. to determine splice site strength scores.

SRE binding sites modulate spliceosomal recognition of splice sites based on adjacent sequence elements. This collection of cis-acting elements includes intronic and exonic splicing enhancers (ISEs, ESEs) and intronic and exonic splicing silencers (ISSs, ESSs). These binding sites recruit auxiliary splicing factors including hnRNPs (heterogeneous nuclear ribonucleoproteins) and SR proteins (serine/arginine-rich proteins) (Long and Caceres 2009; Han et al. 2010). When bound to the pre-mRNA their interactions with spliceosomal components influence the efficiency of spliceosomal assembly, thus modulating intron excision.

Single-stranded RNA molecules (such as pre-mRNA) fold into extraordinarily complicated secondary and tertiary structures as a result of intramolecular base pairing. These local RNA structures can have a profound effect on splicing regulation (Shepard and Hertel 2008). Figure 1.3 demonstrates the many ways RNA secondary structure

can affect splicing regulation: splice site suppression; the occlusion/exposure of cis-acting regulatory elements; ADAR-mediated splice site selection; approximation of cis-elements; “looping-out”; dsRNA-mediated steric hindrance; and competition between RNA secondary structures (Jin et al. 2011).

Alternative Splicing

The human genome encodes approximately 20,400 genes (Genome Reference Consortium 2017) while the estimated proteome is well over 100,000 before post-translational modifications. This numerical discrepancy is due the alternative splicing of pre-messenger RNA. Alternative splicing, also referred to as differential splicing, is the process that contributes to transcript variation in a highly coordinated and complex fashion through the inclusion or exclusion of whole or partial exons and intronic regions from a final processed mRNA (Kornblihtt et al. 2013). Next-generation sequencing studies estimate that ~86-88% of multi-exonic genes undergo some form of alternative splicing (Pan et al. 2008; Wang et al. 2008; Chen et al. 2014). There are several forms of alternative splicing, sometimes referred to as splicing events, including cassette or skipped exons (SE), alternative 3' splice sites (A3SS), alternative 5' splice sites (A5SS), mutually exclusive exons (MXE), and retained introns (RI) (Figure 1.4).

Figure 1.3. RNA secondary structures influence alternative splicing. Constitutive exons are colored grey; alternative exons are colored blue. Introns are black lines. Red lines indicate intramolecular base pairing. A hypothetical example of each type of RNA secondary structure influence is shown: A) splice site suppression; B) the occlusion/exposure of cis-acting regulatory elements; C) ADAR-mediated splice site selection; D) approximation of cis-elements; E) “looping-out”; F) dsRNA-mediated steric hindrance; and G) competition between RNA secondary structures. Figure adapted from (Jin et al. 2011).

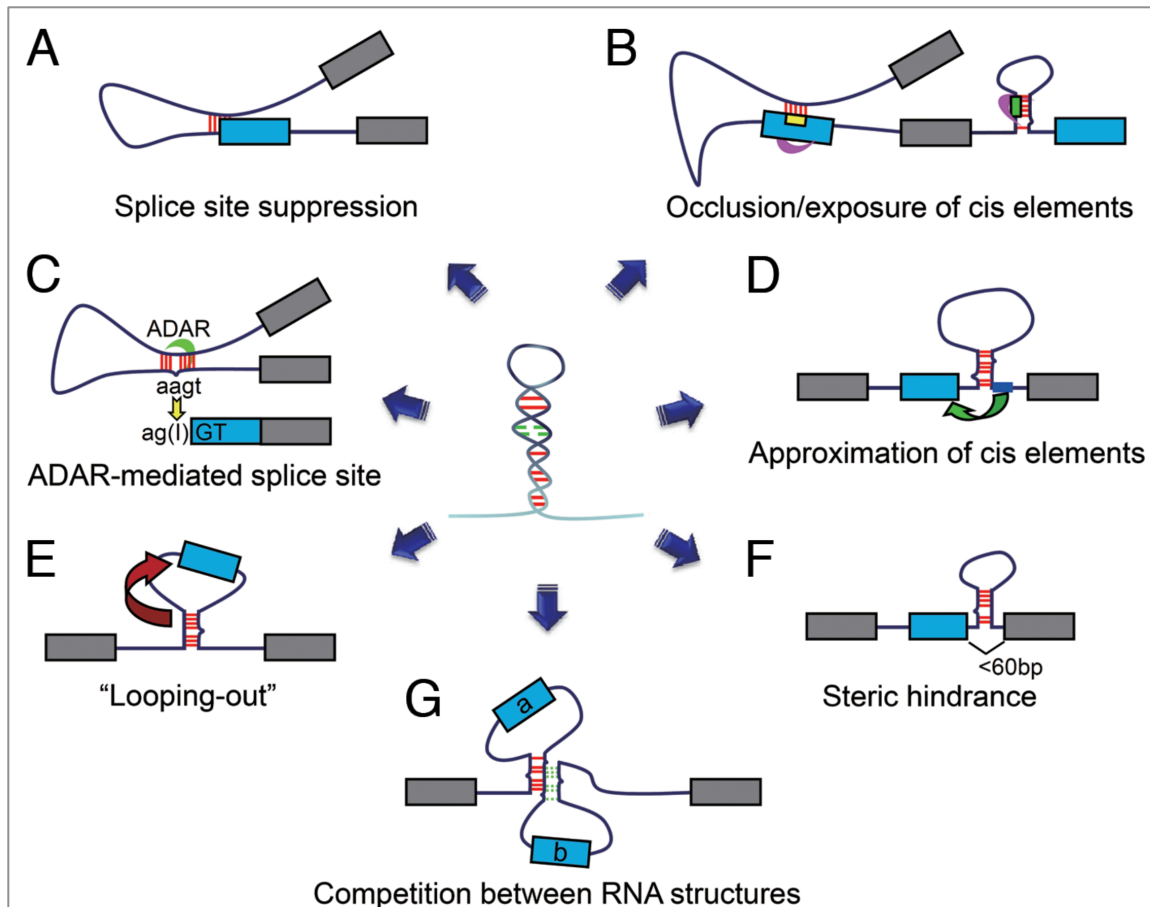
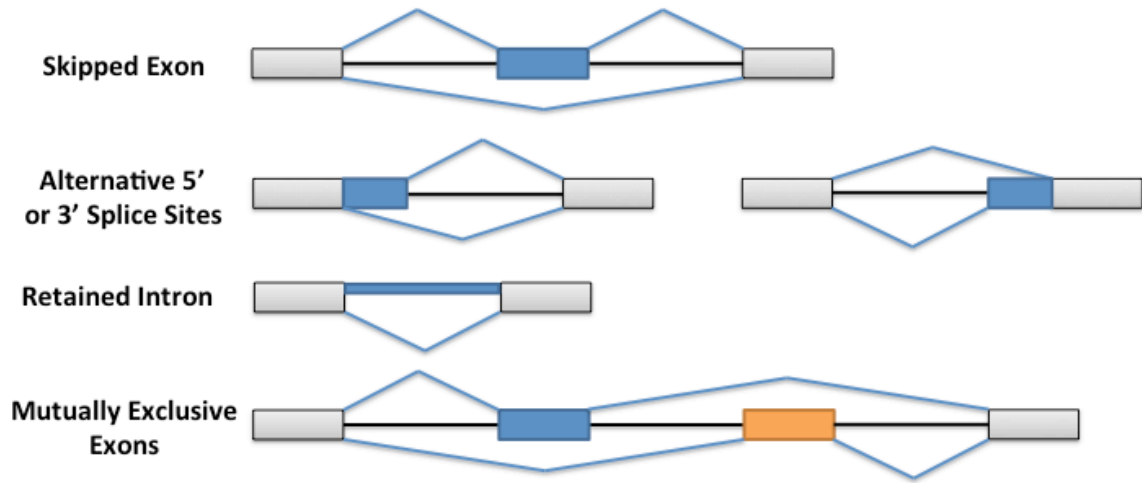


Figure 1.4. Types of alternative splicing. Constitutive exons are colored grey; alternative exons are colored blue or yellow. Introns are black lines. Skipped Exon is an alternative splicing event in which an entire exon is skipped in the final transcript. Alternative 3' Splice Site usage occurs when the canonical 5'SS is used, but an alternative 3'SS is chosen. The 3'SS can be located either upstream or downstream from the dominant splice site. Alternative 5' Splice Site usage happens when the same 3'SS is used, but an alternative 5'SS is used. The 3'SS can be located either upstream or downstream from the dominant splice site. Mutually Exclusive Exon usage occurs when one exon or another are used, but never together. Retained intron usage occurs when an entire intron is included in the mRNA.



Each of these alternative splicing events can also occur in tandem with each other, sometimes referred to as local splicing variations (LSV), creating even more complex mRNA transcript variants (Vaquero-Garcia et al. 2016).

A quick glance across the phylogenetic landscape reveals that there is a higher number of alternative splicing events in higher order organisms, suggesting a correlation between alternative splicing and species complexity (Chen et al. 2014). The transcriptome can also be altered and regulated in a tissue-dependent manner by alternative splicing (Smith 2008; Merkin et al. 2012; Barbosa-Morais et al. 2012). Furthermore, alternative splicing can be used as control switches in cell fate and gene expression by changing transcript isoforms. One such example are the MBNL proteins as repressors of embryonic cell-specific alternative splicing and reprogramming (Han et al. 2013).

The Fidelity of Splicing

Considering the complexity of the splicing machinery, the number of processing steps and the variability of splice sites and splicing regulators, it is amazing that splicing occurs in such a stable manner. Clearly, splicing fidelity must be maintained to avoid introducing errors and deleterious transcripts during gene expression. Estimates of the splicing error rate range from 1 in 100 to 1 in 100,000 (Fox-Walsh and Hertel 2009; Pickrell et al. 2010). Based on these studies, it has been suggested that splicing accuracy is merely limited by Pol II transcription error rates (Fox-Walsh and Hertel 2009; Mellert et al. 2011).

To preserve splicing fidelity, the spliceosome (like transcription and translation) uses kinetic proofreading mechanisms during splice site selection. The most well-studied of these include the DExD/H box ATPases Prp16 and Prp22 (Semlow and Staley 2012). These proofreading mechanisms occur during the spliceosomal assembly process and splice site selection. Prp16 preferentially represses suboptimal transcripts by destabilizing the first transesterification step of the spliceosome by directly competing with 5'SS cleavage (Semlow and Staley 2012). This proofreading also helps Prp16 discriminate against slow splicing substrates (Koodathingal and Staley 2013). In a similar kinetic proofreading mechanism, Prp22 antagonizes primarily suboptimal 3'SS by competing directly with exon ligation (Mayas et al. 2006; Semlow and Staley 2012).

Again, the actual sequence of the pre-mRNA is very important to splice site selection, splicing fidelity, proofreading, and the regulation of splicing by SREs. While it has been previously shown that the spliceosome pairs exons with a high degree of accuracy that may be limited by the quality of pre-mRNAs generated by RNA pol II (Fox-Walsh and Hertel 2009; Mellert et al. 2011; Pickrell et al. 2010), these studies were all constrained by the limits of RT-PCR (unable to determine sequence of individual transcripts) or genome-wide RNA-seq approaches (individual transcripts identified at a shallow depth compared to the entirety of the transcriptome). It is currently impossible to capture every splicing error, but the identification of rare and ultra-rare spliced transcript variants, both canonical (expected) and cryptic (unexpected), could lead to a greater understanding of every aspect of splicing.

Splicing and Disease

Mistakes in splicing or its regulation can be harmful to the cell and play a role in multiple human diseases. According to the Human Gene Mutation Database (HGMD release 2014.4), mutations that disrupt normal splicing have been estimated to account for up to a third of all disease-causing mutations (Padgett 2012; Singh and Cooper 2012; Dagueuet et al. 2015). Because the splicing process is so complex, there are a myriad of ways that splicing defects can occur. Perhaps the most common splice altering mutations are in the pre-mRNA, located in core consensus sequences such as the 5'SS, 3'SS, BPS, or within one or more SREs. One example of this type of splice altering mutation is a G→A substitution at position +1 in intron 7 of *PINK1*. The mutation destroys the consensus 5'SS and thereby activates a cryptic splice site ultimately resulting in *PINK1* exon 7 skipping. This defective *PINK1* transcript leads to early-onset Parkinson's disease.

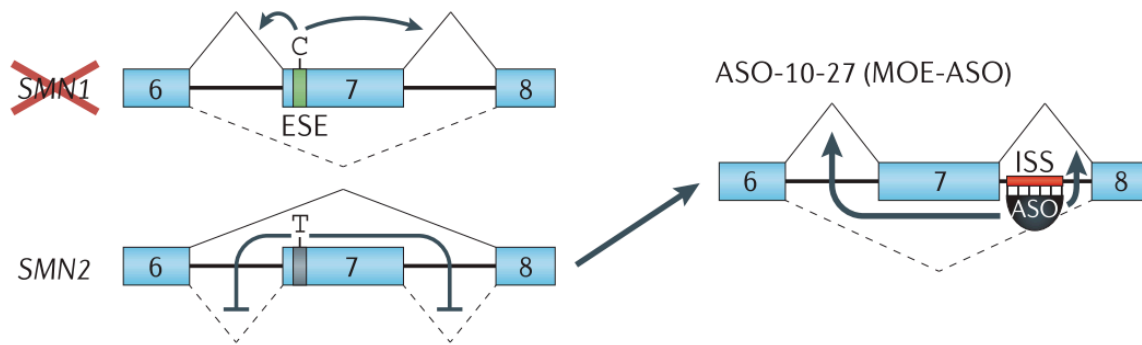
Disease causing mutations can also be found within the spliceosomal machinery itself. Retinitis pigmentosa is caused by mutations in *PRPF6* or *SNRNP200*, affecting spliceosome assembly or a decreased proofreading mechanism respectively (Tanackovic et al. 2011; Cvačková et al. 2014). Likewise, a mutation in the key splicing factor *U2AF1* alters 3'SS recognition and leads to myelodysplastic syndromes (Shirai et al. 2015).

Spinal muscular atrophy (SMA) is perhaps the most common and well-studied disease caused by splicing defects. SMA is the leading cause of hereditary infant mortality occurring in 1 in ~10,000 live births. SMA is an autosomal recessive neurodegenerative disease affecting the motor neurons of the individual, caused by

loss-of-function mutations and/or deletions in the survival of motor neuron 1 (*SMN1*) gene, which encodes the SMN protein required for the generation of snRNPs. A paralogous gene, *SMN2*, is nearly identical to *SMN1*, differing only by a single nucleotide, a C→T transition at position 6 in exon 7, which alters an SRSF1 ESE resulting in the exclusion of exon 7 (Cartegni et al. 2006; Lorson et al. 1999). This truncated SMN protein cannot compensate for a mutated or defective *SMN1*, resulting in SMA. Recent studies support the usage of an antisense oligonucleotide (ASO) for the treatment of SMA, by blocking an ISS in the *SMN2* intron 7 to increase SMN protein levels in affected children (Figure 1.5) (Lim and Hertel 2001; Hua et al. 2008; Chiriboga et al. 2016; Schoch and Miller 2017).

Alternative splicing is gaining recognition in the key role it plays in tumorigenesis and cancer progression. Recently, a The Cancer Genome Atlas (TCGA) Project analysis demonstrated that single nucleotide variants (SNVs) that cause intron retention are enriched in tumor suppressors, suggesting that intron retention is a common tumor suppressor inactivation mechanism (Jung et al. 2015). Further analyses of the TCGA have identified additional common alternative splicing events and affected splicing factors (Tsai et al. 2015; Sebestyén et al. 2015, 2016).

Figure 1.5. Therapeutic strategy for treatment of Spinal Muscular Atrophy. The inactivation or deletion of survival of motor neuron 1 (SMN1), which produces the majority of the SMN protein, leads to spinal muscular atrophy (SMA). SMN2 produces little SMN protein due to a C→T transition in exon 7 that leads to exon 7 skipping. ASO-10-27 (an antisense oligonucleotide) blocks an intronic splicing silencer (ISS, red bar) to enhance SMN protein production by SMN2. Figure adapted from (Scotti and Swanson 2015).



Previous studies have linked individual splicing regulators, such as SR proteins and hnRNPs, to cancer (Karni et al. 2007; Lefave et al. 2011; Anczuków et al. 2012, 2015). For example, *ESRP1* and *ESRP2* are epithelial cell-type-specific regulators of *FGFR2* splicing, a gene which is responsible for epithelial-mesenchymal transition (EMT) (Warzecha et al. 2009). These new discoveries suggest that dysregulation of alternative splicing can be regarded as one of the molecular hallmarks of cancer (Oltean and Bates 2014).

Summary

Previous research on pre-mRNA splicing describes a complex process that is the result of a coordinated effort of spliceosomal assembly, regulation and fidelity. The following work highlights novel insights into splicing fidelity, splicing regulation and differential splicing in the analysis of disease models. Chapter 2 analyzes the fidelity of pre-mRNA splicing across 3 sets of exon|intron junctions at an unprecedented level of depth using the SMN1 exon 7 mini-gene construct. This study discovered new unannotated alternative splice sites and cryptic splice sites, while reinforcing the impressive fidelity of pre-mRNA splicing. Additionally, we measured the effects that synonymous mutations in SMN1 exon 7 have on splicing fidelity when compared to wild-type. These combined results stress the importance of the pre-mRNA sequence in determining splicing efficiency and alternative splicing pattern. Chapter 3 describes the potential for splicing fidelity and splicing regulation as possible determinants of alternative splicing in breast cancer. Differential alternative splicing was examined between estrogen receptor positive and estrogen receptor negative breast cancer cell types, revealing the presence of known and unknown MBNL isoforms. Through the use

of a splicing code, a common model for exon inclusion and exclusion in breast cancer was revealed. Chapter 4 discusses the results of this dissertation in the context of the current views of alternative splicing. It further explores potential challenges that face further study of splicing fidelity and alternative splicing in disease.

References

- Anczuków O, Akerman M, Cléry A, Wu J, Shen C, Shirole NH, Raimer A, Sun S, Jensen MA, Hua Y, et al. 2015. SRSF1-Regulated Alternative Splicing in Breast Cancer. *Mol Cell* **60**: 105–117.
<http://linkinghub.elsevier.com/retrieve/pii/S1097276515007017>.
- Anczuków O, Rosenberg AZ, Akerman M, Das S, Zhan L, Karni R, Muthuswamy SK, Krainer AR. 2012. The splicing factor SRSF1 regulates apoptosis and proliferation to promote mammary epithelial cell transformation. *Nat Struct Mol Biol* **19**: 220–8.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3272117&tool=pmcentrez&rendertype=abstract> (Accessed March 9, 2012).
- Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, et al. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**: 1587–93.
<http://www.ncbi.nlm.nih.gov/pubmed/23258890> (Accessed May 22, 2013).
- Black DL. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72**: 291–336. <http://www.ncbi.nlm.nih.gov/pubmed/12626338> (Accessed June 14, 2011).
- Cartegni L, Hastings ML, Calarco JA, de Stanchina E, Krainer AR. 2006. Determinants of exon 7 splicing in the spinal muscular atrophy genes, SMN1 and SMN2. *Am J Hum Genet* **78**: 63–77.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1380224&tool=pmcentrez&rendertype=abstract>.
- Chen L, Bush SJ, Tovar-Corona JM, Castillo-Morales A, Urrutia AO. 2014. Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity. *Mol Biol Evol* **31**: 1402–1413.
<https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msu083>.
- Chiriboga CA, Swoboda KJ, Darras BT, Iannaccone ST, Montes J, De Vivo DC, Norris DA, Bennett CF, Bishop KM. 2016. Results from a phase 1 study of nusinersen (ISIS-SMN Rx) in children with spinal muscular atrophy. *Neurology* **86**: 890–897.
<http://www.neurology.org/lookup/doi/10.1212/WNL.0000000000002445>.
- Cvačková Z, Matějů D, Staněk D. 2014. Retinitis pigmentosa mutations of SNRNP200 enhance cryptic splice-site Recognition. *Hum Mutat*.
- Daguenet E, Dujardin G, Valcárcel J. 2015. The pathogenicity of splicing defects: mechanistic insights into pre-mRNA processing inform novel therapeutic approaches. *EMBO Rep* **16**: 1640 LP-1655.
<http://embor.embopress.org/content/16/12/1640.abstract>.
- Dou Y, Fox-Walsh KL, Baldi PF, Hertel KJ. 2006. Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. *Bioinformatics* **12**: 2047–2056.
- Fox-Walsh KL, Hertel KJ. 2009. Splice-site pairing is an intrinsically high fidelity process. *Proc Natl Acad Sci U S A* **106**: 1766–71.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2644112&tool=pmcentrez&rendertype=abstract>.
- Genome Reference Consortium. 2017. GRCh38.p12. *NCBI*.

- Graveley BR, Hertel KJ, Maniatis T. 2001. The role of U2AF35 and U2AF65 in enhancer-dependent splicing. *RNA* **7**: 806–18. <http://www.ncbi.nlm.nih.gov/pubmed/11421359>.
- Han H, Irimia M, Ross PJ, Sung H-K, Alipanahi B, David L, Golipour A, Gabut M, Michael IP, Nachman EN, et al. 2013. MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* 5–12. <http://www.nature.com/doi/10.1038/nature12270> (Accessed June 5, 2013).
- Han SP, Tang YH, Smith R. 2010. Functional diversity of the hnRNPs: past, present and perspectives. *Biochem J* **430**: 379–92. <http://www.ncbi.nlm.nih.gov/pubmed/20795951> (Accessed June 29, 2011).
- Hertel KJ. 2008. Combinatorial control of exon recognition. *J Biol Chem* **283**: 1211–5. <http://www.ncbi.nlm.nih.gov/pubmed/18024426> (Accessed July 15, 2011).
- Hicks MJ, Mueller WF, Shepard PJ, Hertel KJ. 2010. Competing Upstream 5' Splice Sites Enhance the Rate of Proximal Splicing. *Mol Cell Biol* **30**: 1878–1886. <http://mcb.asm.org/cgi/doi/10.1128/MCB.01071-09>.
- Hua Y, Vickers TA, Okunola HL, Bennett CF, Krainer AR. 2008. Antisense Masking of an hnRNP A1/A2 Intronic Splicing Silencer Corrects SMN2 Splicing in Transgenic Mice. *Am J Hum Genet* **82**: 834–848. <http://linkinghub.elsevier.com/retrieve/pii/S0002929708001638>.
- Jin Y, Yang Y, Zhang P. 2011. New insights into RNA secondary structure in the alternative splicing of pre-mRNAs. *RNA Biol* **8**: 450–457.
- Jung H, Lee D, Lee J, Park D, Kim YJ, Park W-Y, Hong D, Park PJ, Lee E. 2015. Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat Genet* **47**: 1242–1248. <http://dx.doi.org/10.1038/ng.3414>.
- Jurica MS, Moore MJ. 2003. Pre-mRNA splicing: awash in a sea of proteins. *Mol Cell* **12**: 5–14. <http://www.ncbi.nlm.nih.gov/pubmed/12887888>.
- Karni R, de Stanchina E, Lowe SW, Sinha R, Mu D, Krainer AR. 2007. The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nat Struct Mol Biol* **14**: 185–93. <http://www.ncbi.nlm.nih.gov/pubmed/17310252> (Accessed July 12, 2012).
- Koodathingal P, Staley JP. 2013. Splicing fidelity. *RNA Biol* **10**: 1073–1079. <http://www.tandfonline.com/doi/abs/10.4161/rna.25245>.
- Kornblihtt AR, Schor IE, Alló M, Dujardin G, Petrillo E, Muñoz MJ. 2013. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat Rev Mol Cell Biol* **14**: 153–165. <http://www.nature.com/articles/nrm3525>.
- Lefave C V, Squatrito M, Vorlova S, Rocco GL, Brennan CW, Holland EC, Pan Y-X, Cartegni L. 2011. Splicing factor hnRNPH drives an oncogenic splicing switch in gliomas. *EMBO J* **30**: 4084–97. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3209773&tool=pmcentrez&rendertype=abstract> (Accessed March 12, 2012).
- Lim SR, Hertel KJ. 2001. Modulation of Survival Motor Neuron Pre-mRNA Splicing by Inhibition of Alternative 3' Splice Site Pairing. *J Biol Chem* **276**: 45476–45483.
- Long JC, Caceres JF. 2009. The SR protein family of splicing factors: master regulators of gene expression. *Biochem J* **417**: 15–27. <http://www.ncbi.nlm.nih.gov/pubmed/19061484> (Accessed June 17, 2011).
- Lorson CL, Hahnen E, Androphy EJ, Wirth B. 1999. A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proc Natl*

- Acad Sci U S A* **96**: 6307–11.
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC26877/>.
- Mayas RM, Maita H, Staley JP. 2006. Exon ligation is proofread by the DExD/H-box ATPase Prp22p. *Nat Struct Mol Biol* **13**: 482–90.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3729281&tool=pmcentrez&rendertype=abstract>.
- Mellert K, Uhl M, Högel J, Lamla M, Kemkemer R, Kaufmann D. 2011. Aberrant single exon skipping is not altered by age in exons of NF1, RABAC1, AATF or PCGF2 in human blood cells and fibroblasts. *Genes (Basel)* **2**: 562–577.
- Merkhofer EC, Hu P, Johnson TL. 2014. Introduction to cotranscriptional RNA splicing. ed. K.J. Hertel. *Methods Mol Biol* **1126**: 83–96.
<http://link.springer.com/10.1007/978-1-62703-980-2>.
- Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science (80-)* **338**: 1593–1599.
<http://www.sciencemag.org/cgi/doi/10.1126/science.1228186>.
- Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**: 457–63. <http://www.nature.com/articles/nature08909>.
- Oltean S, Bates DO. 2014. Hallmarks of alternative splicing in cancer. *Oncogene* **33**: 5311–5318. <http://dx.doi.org/10.1038/onc.2013.533>.
- Padgett RA. 2012. New connections between splicing and human disease. *Trends Genet* **28**: 147–154. <http://dx.doi.org/10.1016/j.tig.2012.01.001>.
- Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**: 1413–5. <http://www.ncbi.nlm.nih.gov/pubmed/18978789> (Accessed July 17, 2012).
- Pickrell JK, Pai A a, Gilad Y, Pritchard JK. 2010. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet* **6**: e1001236.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3000347&tool=pmcentrez&rendertype=abstract> (Accessed August 14, 2013).
- Reed R. 1996. Initial splice-site recognition and pairing during pre-mRNA splicing. *Curr Opin Genet Dev* **6**: 215–220.
<http://linkinghub.elsevier.com/retrieve/pii/S0959437X96800530>.
- Schoch KM, Miller TM. 2017. Antisense Oligonucleotides: Translation from Mouse Models to Human Neurodegenerative Diseases. *Neuron* **94**: 1056–1070.
<http://dx.doi.org/10.1016/j.neuron.2017.04.010>.
- Scotti MM, Swanson MS. 2015. RNA mis-splicing in disease. *Nat Rev Genet* **17**: 19–32.
<http://www.nature.com/doi/10.1038/nrg.2015.3>.
- Sebestyén E, Singh B, Miñana B, Pagès A, Mateo F, Pujana MA, Valcárcel J, Eyraas E. 2016. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Res* **26**: 732–744.
<http://genome.cshlp.org/lookup/doi/10.1101/gr.199935.115>.
- Sebestyén E, Zawisza M, Eyraas E. 2015. Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Res* **43**: 1345–1356. <http://academic.oup.com/nar/article/43/3/1345/2411389/Detection-of-recurrent-alternative-splicing>.
- Semlow DR, Staley JP. 2012. Staying on message: ensuring fidelity in pre-mRNA

- splicing. *Trends Biochem Sci* **37**: 263–273.
<http://dx.doi.org/10.1016/j.tibs.2012.04.001>.
- Shepard PJ, Hertel KJ. 2008. Conserved RNA secondary structures promote alternative splicing. *RNA* **14**: 1463–9. <http://www.rnajournal.org/cgi/doi/10.1261/rna.1069408>.
- Shirai CL, Ley JN, White BS, Kim S, Tibbitts J, Shao J, Ndonwi M, Wadugu B, Duncavage EJ, Okeyo-Owuor T, et al. 2015. Mutant U2AF1 Expression Alters Hematopoiesis and Pre-mRNA Splicing In Vivo. *Cancer Cell*.
- Singh RK, Cooper TA. 2012. Pre-mRNA splicing in disease and therapeutics. *Trends Mol Med* **18**: 472–482. <http://dx.doi.org/10.1016/j.molmed.2012.06.006>.
- Smith CWJ. 2008. *Alternative Splicing in the Postgenomic Era*, edited by Benjamin J. Blencowe and Brenton R. Graveley. 2007, XXIV, Springer, New York. ISBN: 978-0-387-77373-5. <http://www.rnajournal.org/cgi/doi/10.1261/rna.1340908>.
- Tanackovic G, Ransijn A, Ayuso C, Harper S, Berson EL, Rivolta C. 2011. A missense mutation in PRPF6 causes impairment of pre-mRNA splicing and autosomal-dominant retinitis pigmentosa. *Am J Hum Genet*.
- Tsai YS, Dominguez D, Gomez SM, Wang Z. 2015. Transcriptome-wide identification and study of cancer-specific splicing events across multiple tumors. *Oncotarget* **6**: 6825–6839. <http://www.oncotarget.com/fulltext/3145>.
- Vaquero-Garcia J, Barrera A, Gazzara MR, González-Vallinas J, Lahens NF, Hogenesch JB, Lynch KW, Barash Y. 2016. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* **5**: e11752. <http://www.ncbi.nlm.nih.gov/pubmed/26829591> (Accessed February 2, 2016).
- Wahl MC, Will CL, Lührmann R. 2009. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**: 701–18. <http://www.ncbi.nlm.nih.gov/pubmed/19239890> (Accessed July 21, 2011).
- Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–6.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2593745&tool=pmcentrez&rendertype=abstract> (Accessed June 14, 2011).
- Warzecha CC, Sato TK, Nabet B, Hogenesch JB, Carstens RP. 2009. ESRP1 and ESRP2 Are Epithelial Cell-Type-Specific Regulators of FGFR2 Splicing. *Mol Cell* **33**: 591–601.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2702247&tool=pmcentrez&rendertype=abstract> (Accessed July 27, 2011).
- Yeo G, Burge CB. 2004. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *J Comput Biol* **11**: 377–394.
<http://www.liebertonline.com/doi/abs/10.1089/1066527041410418>.
- Zhang XHF, Chasin LA. 2004. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev* **18**: 1241–1250.

CHAPTER 2

Ultra-deep sequencing reveals splicing as a sequence driven high-fidelity process, constrained by splicing efficiency.

SUMMARY

Alternative splicing diversifies mRNA transcripts in human cells. It has previously been shown by quantitative real-time PCR that the spliceosome pairs exons with a high degree of accuracy. Yet, pre-mRNA splicing error rates have not been deeply analyzed at the nucleotide level to determine the quantity and identity of rare splicing errors across splice junctions. Using ultra-deep sequencing we determined the splicing error for three splice junctions flanking exon 7 of SMN1 at single nucleotide resolution. After corrections for background noise introduced to the dataset by PCR amplification and sequencing steps, pre-mRNA splicing maintains a low overall error rate. We identified several previously unannotated splicing events across 3 exon|intron junctions in SMN1. We demonstrate the effects that mutations in SMN exon 7 have on splicing fidelity; modulating splicing efficiency by changing RNA secondary structures, altering the binding of regulatory proteins, and changing the 5' splice site strength. Mutations also create a truncated SMN exon 7 through the introduction of a *de novo* cryptic 5' splice site. These results underscore the impressive fidelity of pre-mRNA splicing and further demonstrate that splicing efficiency controlled by sequence context is the driving force behind splice site pairing.

INTRODUCTION

Splicing is a complex process requiring hundreds of proteins to work in concert with proper regulation (Wahl et al. 2009). A pre-mRNA transcript from a single gene can be alternatively spliced to generate many mRNA variants. Differential pre-mRNA processing contributes significantly to genetic variability; it is estimated that transcripts from ~86-88% of multi-exon genes undergo alternative splicing (Pan et al. 2008; Wang et al. 2008; Chen et al. 2014). Many mRNA isoforms are generated from a single gene as a result of splicing regulation, which may be due to required isoform ratios, systemic feedback, or tissue specific splicing (Kornblihtt et al. 2013; Barbosa-Morais et al. 2012). Other alternative mRNA isoforms may be the result of erroneous splice site pairing, sometimes referred to as cryptic splice sites, which may result in aberrant alternative mRNA isoforms (Buratti et al. 2007). It has been shown that the most common form of cryptic splice site activation occurs near the canonical splice site, mainly due to the U1 snRNP binding consensus sequence for 5' splice sites or duplicate YAG trinucleotides near 3' splice sites (Dou et al. 2006; Tsai et al. 2010).

To avoid these errors, there are several safeguards to ensure splicing fidelity. Like transcription and translation, splicing has an active proofreading mechanism, while additionally relying on sequence information to guide the spliceosome through the process. Prp16 (Koodathingal et al. 2010) and Prp22 (Mayas et al. 2006; Semlow and Staley 2012) provide proofreading mechanisms for the first and second sequential transesterification reactions of splicing, and may even remodel pre-mRNA to activate alternative splice sites (Semlow et al. 2016). Splicing regulatory element binding sites and the base-pairing of snRNPs to the pre-mRNA substrate lead to the selection of the

correct splice sites based on optimal adjacent sequence contexts (Nilsen and Graveley 2010). Even with these safeguards, splicing fidelity can be compromised when the sequence context for splice sites is suboptimal. All mRNA isoforms are subject to a number of quality control mechanisms, such as nonsense-mediated decay (NMD), nonstop decay (NSD), or no-go decay (NGD), however, not all aberrant mRNA isoforms are removed through these processes and could be translated.

Owing to the importance of splicing regulation, a large number of mis-splicing or splicing errors can result in different diseases (Scotti and Swanson 2015). According to the Human Gene Mutation Database (HGMD release 2014.4), mutations that disrupt normal splicing have been estimated to account for up to a third of all disease-causing mutations (Daguenet et al. 2015). It has been demonstrated that the spliceosome can pair constitutive exons with high fidelity at an error rate as low as one in 20,000 splicing events (Fox-Walsh and Hertel 2009; Mellert et al. 2011). Based on these studies, it was suggested that splicing accuracy is limited by Pol II transcription error rates (Fox-Walsh and Hertel 2009; Mellert et al. 2011). These RT-qPCR based studies are inherently limited to resolution at the exonic level, investigating only single exon skipping events based on EST annotation. Using genome-wide RNA-sequencing, similar error rates were observed (Pickrell et al. 2010), but it is still unclear whether splicing errors are the result of transcription errors, poor exon recognition mediated by weak splice sites and splicing regulatory elements, or whether errors are merely stochastic in nature. Additionally, the extent of aberrant mRNA splicing at the nucleotide level remains unknown. Using ultra deep sequencing we determined the splicing error for three splice junctions flanking exon 7 of SMN1 at single nucleotide resolution. We identified

previously unannotated splice sites, a potential microexon, potential transcription error-mediated splicing errors and the rate at which 5' splice sites with their inherently susceptible U1snRNP binding site incorrectly splice at positions 4 nucleotides upstream or downstream of the canonical splice site. Furthermore, we evaluated the effects that mutations in SMN exon 7 have on splicing fidelity.

RESULTS

Dataset For Ultra-Deep Analysis Of Splicing Fidelity

To determine the splicing error rate we used a recent dataset (Mueller et al. 2015) of *SMN1* exon 7 inclusion rates based on a synonymous position mutation library in the well-studied *SMN1* mini-gene, which spans exons 6-8 (Lorson et al. 1999; Lim and Hertel 2001; Singh et al. 2007) where exon 7 is included or excluded depending on splicing signals in the pre-mRNA (Figure 2.1). In a *SMN1* exon 7, neighboring codons in exon 7 were mutated to every possible combination of silent mutations within the context of a sliding hexamer window, a minimal binding site for splicing regulatory proteins (Mueller et al. 2015; Fairbrother et al. 2002). The resulting library of plasmids was transfected into HeLa cells and plasmid-specific mRNAs were analyzed by deep sequencing. The

Figure 2.1. Mutation scheme for *SMN1* exon 7 library

The *SMN1* mini-gene construct consists of exon 6, exon 7, and exon 8 with shortened introns 6 and 7. All possible silent mutations in exon 7 were created within the context of a sliding hexamer window. For example, the first two codons depicted are GGT TTC. All three mutations were made in GGT resulting in GGN and combined with all silent

mutations in TTC (TTT), resulting in eight combinations including the wild-type sequence. A transition 6C→T results in SMN1 exon 7 skipping.

GGT TTC AGA CAA AAT CAA AAA GAA GGA AGG TGC TCA CAT TCC TTA AAT TAA GGA

GGN TTY

TTY MGN

MGN CAR

ETC.

SMN1

Exon 6

T

C

Exon 7

Exon 8

data obtained from these library transfections were previously used to determine if synonymous mutations in exon 7 influence splicing. This study also resulted in the several million-fold sequencing of three exon|intron splice site junctions, *SMN1* exon6|exon7, exon7|exon8, and exon6|exon8. This extensive sequencing data allowed for an ultra-deep detection of low abundance local isoforms, including rare stochastic and non-stochastic splicing outcomes described below. Our observations and calculations of splicing fidelity are based on the wild-type *SMN1* mini-gene.

Total Splicing Error Rate

There are several explanations why splicing can occur in a non-canonical way. In this study, deviations from the expected canonical exon 7 inclusion splicing pathway (referred to here as cryptic splicing) could be the consequence of imperfections in the generation of the *SMN1* mini-gene mutation library, pre-splicing transcription errors, sequencing errors, or the activation of rarely used splice sites, such as *de novo* splice sites or the selection of microexons. Using the ultra-deep dataset, we determined how many of the alternatively spliced transcripts identified are a result of cryptic splice-site selection and how many may be due to other factors such as transcription errors, sequencing errors, or the splicing machinery failing to perform as expected. The ultra-deep sequencing of the *SMN1* mini-gene depicts several clear-cut examples of cryptic splice site selection, albeit at a very low rate. Out of a total of 6,469,446 reads that contained wild-type *SMN1* exon 7, there were 20,505 reads that contained some sort of error at either the exon6|exon7 or the exon7|exon8 junction, for a raw error rate of 3.2E-03 or 1 error for every 315 splicing events. At first glance this is a higher rate than that

of other gene expression steps, transcription and translation, each of which are characterized by error rates as low as $1.0E-05$ (Nesser et al. 2006; Jeon and Agarwal 1996; Imashimizu et al. 2013). Further examination of the dataset revealed that not every observed raw error could be counted as a result of aberrant splicing fidelity.

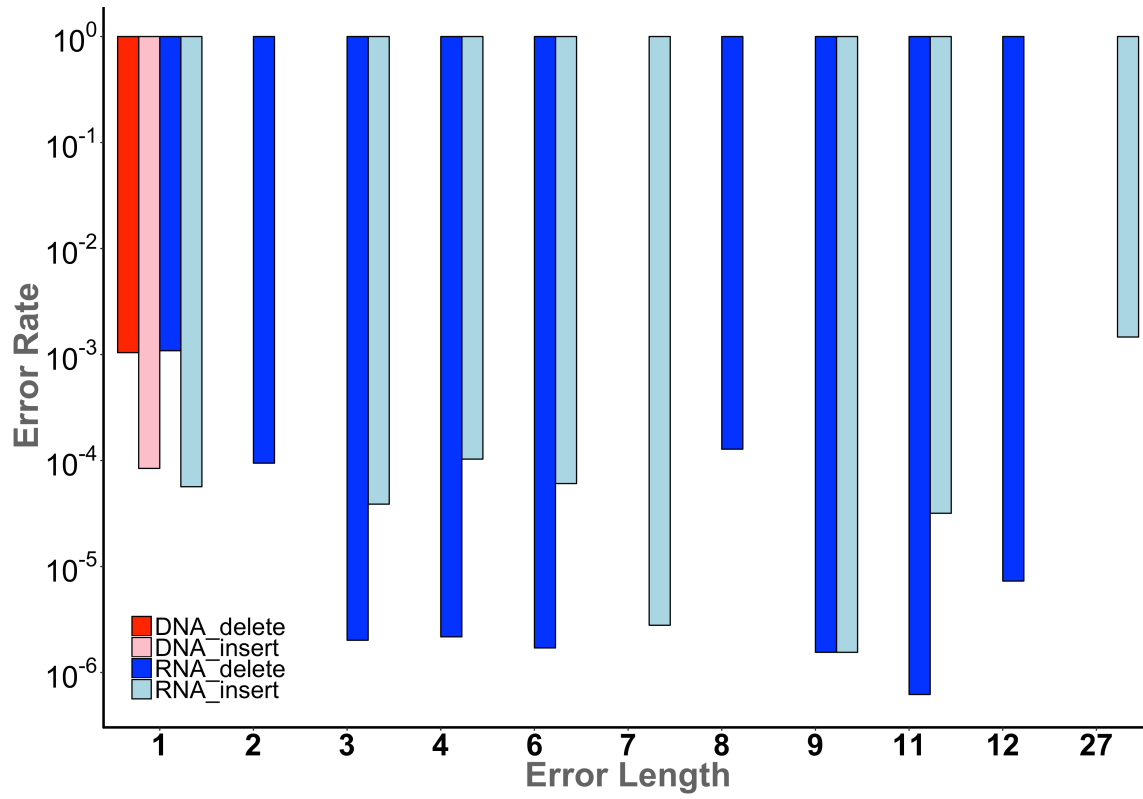
Control For Sequencing Errors

In addition to sequencing the RNA generated from our SMN mini-gene mutation library, we sequenced the transfected SMN mini-gene constructs themselves. This served as a control to demonstrate that sequence differences we detected in the mRNA reads are due RNA generation and processing. The most common sequence deviation from the RNA pool was the deletion of a single guanosine from a GGG triplet at the exon6|exon7 junction at a frequency of $1.1E-03$, accounting for nearly 1/3 of the total identified errors. However, the same deletion occurred within a GGG triplet at a nearly identical frequency at the intron6|exon 7 junction in their DNA counterparts ($1.0E-03$) (Figure 2.2). Similarly, a single guanosine insertion at this same site, producing a GGGG motif, occurs in 247 DNA reads and 320 RNA reads at rates $8.4E-05$ and $5.0E-05$, respectively (Figure 2.2). These observations strongly suggest that these single guanosine insertions and deletions derive from errors independent of splicing. Importantly, there were no errors in the DNA input reads that resulted in

Figure 2.2. Analysis of splicing error rates in DNA and RNA reads by error length

To detect possible reading frame preservation bias in splicing errors, an analysis of splicing error rates in DNA and RNA by insertion or deletion length in wild-type reads was performed. For example, the insertion of 4 nucleotides GTAA at the 5'SS of exon 7

and the insertion of 4 nucleotides ACAG at the 3'SS of exon 8 are combined as a total error rate for the insertion of 4 nucleotides. There is no consensus error length.



insertions or deletions of multiple consecutive nucleotides. We conclude that any RNA output reads with 2 or more nucleotides consecutively inserted or deleted are attributable to pre-mRNA processing errors or pre-existing sequence variations introduced in the library during its construction..

Control For Plasmid Generated Errors

The sequencing of the transfected SMN mini-gene constructs also serves as a control to demonstrate that errors we detect in the processed RNA reads are due to the generation and processing of the RNA, namely transcription and splicing, and were not already present in the DNA template. Due to the size constraints of the sequencing (100 nucleotide read length) and the location of the input DNA primers, we were only able to estimate the plasmid error rate for the region that was flanked and amplified by the DNA primers used. This region consists of exon7 and the adjacent 6 upstream and 10 downstream nucleotides (Figure 2.3). We found that while mutants that arose from errors in the SMN mini-gene construct do exist, they occur at a low rate, averaging $3.0E-04$ (Figure 2.3, Table 2.1). While infrequent, these library construction imperfections limit the sensitivity of splicing error detection.

Figure 2.3. Mutation rate by position in DNA input reads

A heatmap representing positional error rates in DNA input reads that lie within the 15 nucleotide primers on either side of the amplified section that was sequenced (100 nucleotides total). Positions within exon 7 that were purposely mutated in our library construction were omitted and set to zero.

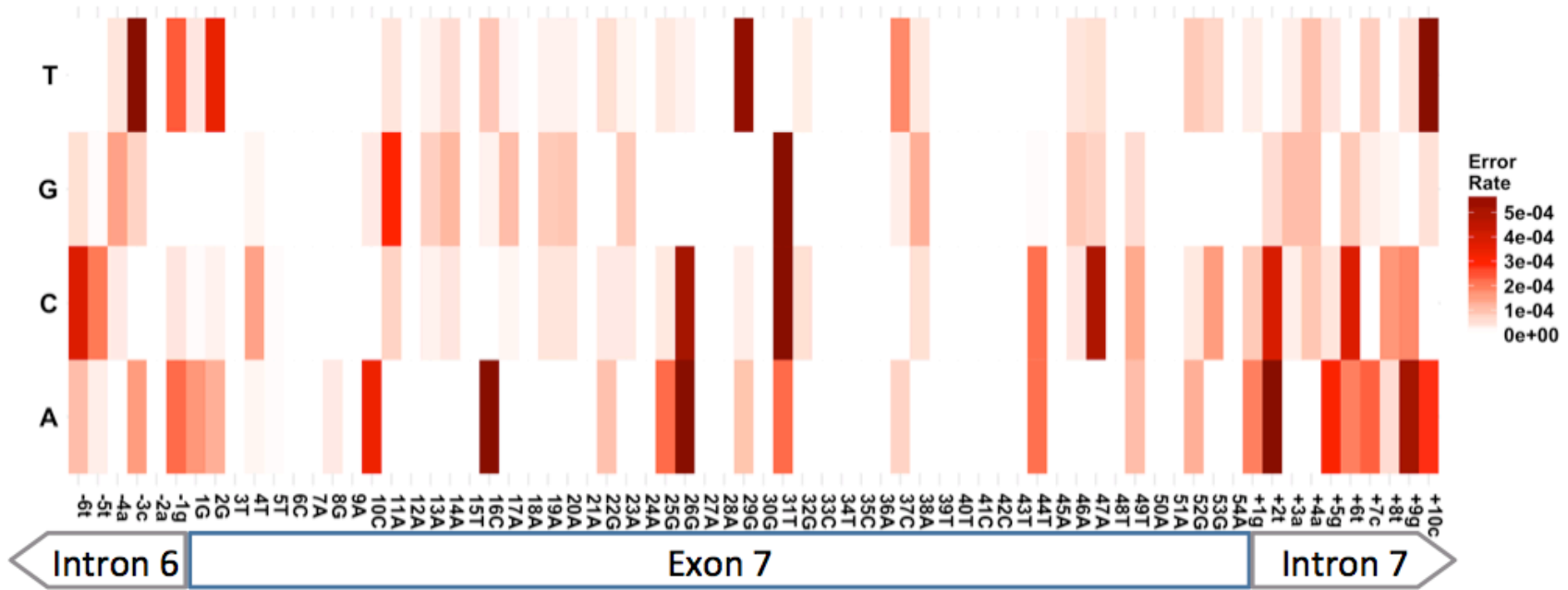


Table 2.1. Average Error Rate – input DNA

	Intron	Exon7	Total
A→C	3.6E-04	8.3E-05	1.6E-04
A→G	7.9E-04	1.1E-04	3.0E-04
A→T	3.9E-04	2.8E-05	1.3E-04
C→A	1.9E-04	4.2E-04	3.3E-04
C→G	4.5E-05	2.4E-05	3.2E-05
C→T	3.6E-04	4.2E-04	4.0E-04
G→A	3.1E-04	7.7E-04	6.3E-04
G→C	8.4E-05	9.7E-05	9.3E-05
G→T	8.5E-05	1.4E-04	1.2E-04
T→A	8.5E-04	1.1E-04	4.8E-04
T→C	3.0E-04	1.2E-03	7.5E-04
T→G	3.9E-05	1.9E-04	1.1E-04

Summary results for each nucleotide substitution across the intronic or exonic regions. The intronic region spans the regions 6 nucleotides upstream and the 10 nucleotides downstream of exon 7. The exonic region is based on those nucleotides within exon 7 that were not subjected to synonymous mutation. The calculated number is the error rate of the nucleotide listed first being substituted by the second nucleotide. The total is the summation of all substitution errors intronic and exonic.

Cryptic 3'-Splice Site Usage

An abundant example of cryptic splicing observed is an unannotated 3' splice site 27 nucleotides upstream of the canonical intron7|exon8 3' splice site (AG/CCTCTGGN₁₀...CAG|GA...; where the cryptic splice site is designated by a “|” and the canonical splice site is represented by a “|”) (Figure 2.4A). This novel splice site is used at a frequency of 1.5E-03 (Table 2.2) and it is characterized by a canonical AG dinucleotide that defines the 3' end of nearly every intron in metazoans (Horowitz 2012). However, a poorly defined upstream polypyrimidine tract prevents extensive usage of this cryptic 3' splice site (maximum entropy score (MES) = -1.62) (Yeo and Burge 2004). At a splice site usage rate of 1 in 680 transcripts, this cryptic splicing event is rare enough that it is only readily discovered using ultra-deep sequencing. The upstream location relative to the canonical 3' splice site polypyrimidine tract suggests that this splice site is acting independent of the canonical 3' splice site.

At the same intron7|exon8 3' splice site we also observed two additional lower frequency insertions. In 376 cases (error rate = 5.9E-05, Fig 2.4A) the ligation of exon 7 and exon 8 took place 6 nucleotides upstream of the canonical 3' splice site (AT/TTGCAG|GAA). The sequence upstream of this cryptic splice site is an AT dinucleotide, instead of the requisite AG. Thus, the low observed frequency of this cryptic 3' splice site selection could be the consequence of selecting a poorly defined AT/TT junction (MES = -1.72), it could have arisen by rare nucleotide mis-incorporation mediated by elongating pol II to change the junction to AG/TT (MES = 6.87), or it could be the consequence of low-level

Figure 2.4. Cryptic splice site usage

A) Cryptic 3' SS usage between exon 7 and exon 8. The green line represents the unannotated cryptic 3' SS event 27 nucleotides upstream of the canonical intron7|exon8 junction. The red dashed line represents the cryptic splicing observed that is due to transcription or library generation errors resulting in canonical AG dinucleotide sequences. **B)** Cryptic 3' SS usage between exon 6 and exon 7. The green line represents the unannotated cryptic 5' SS event at position 8 in exon 7. The red dashed line represents the cryptic splicing observed that is due to transcription or library generation errors. **C)** Cryptic 5' SS usage between exon 7 and exon 8. The green line represents the usage of the intrinsic cryptic 5' SS. **D)** Cryptic 5' SS usage between exon 6 and exon 7. The green line represents the usage of the intrinsic cryptic 5' SS.

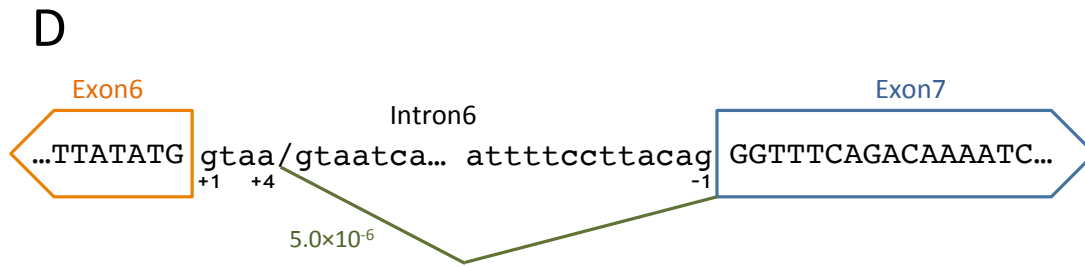
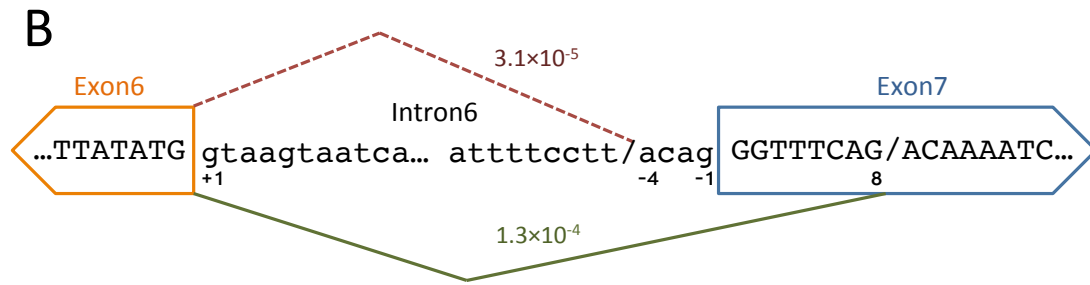
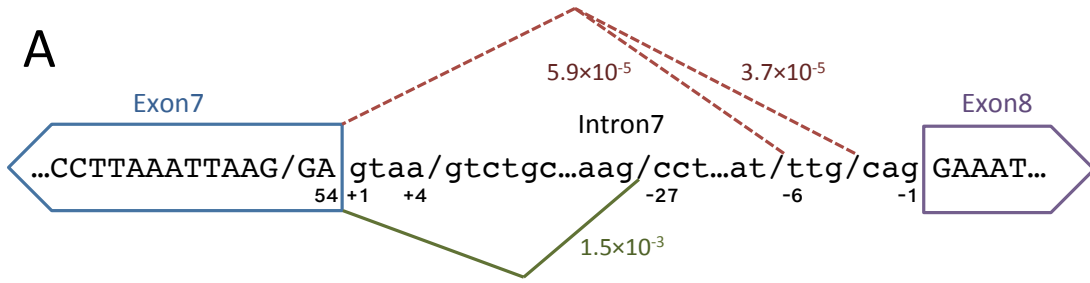


Table 2.2. 3' Cryptic Splicing

Cryptic Splice Site	Count	Error rate	Junction	Error Type	SS location vs canonical	MES
AG GGTTTCAG/ACA	825	1.3E-04	ex6 ex7	deletion	downstream	0.35
AT TTTCCTTACAG GGT	202	3.1E-05	ex6 ex7	insertion	upstream	2.24
AG/CCTCTGGN ₁₀ ...CAG GAA	9433	1.5E-03	ex7 ex8	insertion	upstream	-1.62
AT TTGCAG GAA	376	5.9E-05	ex7 ex8	insertion	upstream	6.87
TG/CAG GAA	238	3.7E-05	ex7 ex8	insertion	upstream	3.04

43

The cryptic splice site is designated by a “/” and the canonical splice site is represented by a “|”. Count refers to the number of wild type reads that contain the cryptic splice site, with their associated rate of occurrence. Junction refers to the location of the cryptic splice site. Error type refers to the result of the cryptic splicing, either an insertion or deletion of sequence from the canonical transcript. SS location vs. canonical refers to the position of the cryptic splice site relative to the canonical splice site. MES is the Maximum entropy score for the cryptic splice site.

nucleotide variations intrinsic to the mutation library that would also give rise to pre-mRNAs characterized by the improved AG/TT 3' splice site junction.

To distinguish between these possibilities, we compared the frequency at which alternative nucleotides were detected at invariable nucleotide positions across exon7 of our SMN1 mutant library (Figure 2.3). On average, the library generation imperfection resulted in T to G nucleotide changes at a frequency of 1.1E-04 (Figure 2.3, Table 2.1). Using this frequency as a measure for background noise, it is impossible to assign the (AT/TTGCAG|GAA) cryptic splicing event to any other cause but library defects.

At the same intron7|exon8 junction, we also observe 238 similar occurrences (error rate = 3.7E-05, Figure 2.4A, Table 2.2) where a possible nucleotide change 4 nucleotides upstream of the canonical 3' splice site (AT/CAG|GAA → AG/CAG|GAA) drastically changes the favorability of the splice site, (MES = -5.33 → 3.04). These nucleotide changes would create cryptic 3' splice sites that approximate those observed in annotated EST databases (Dou et al. 2006). Using the same arguments as above, it is most likely that the (AT/CAG|GAA) cryptic splicing event is observed because of library imperfections.

Analysis of 3' cryptic splicing at the exon6|exon7 junction highlights two events that are represented with reasonable frequency. We observed 825 reads (error rate = 1.3E-4, Figure 2.4B, Table 2.2) where the activation of a 3' splice site 8 nucleotides downstream of the canonical 3' splice site AG|GGTTTCAG/AC results in a truncated exon 7. This cryptic 3' splice site has a low MES (0.35), yet its activation does not rely on nucleotide changes at the new spliced junction. Based on these considerations it is

likely that cryptic AG selection intrinsic to the SMN1 wild-type sequence context mediates this cryptic splicing event.

A lower frequency event is represented by 202 reads (error rate = $3.1E-05$, Figure 2.4B, Table 2.2) where a cryptic 3' splice site (AT/TTTCCTTACAG|GGT) was selected for the intron6|exon7 junction 11 nucleotides upstream of the canonical 3' splice site. The wild-type sequence upstream of this cryptic splice site is an AT dinucleotide, instead of the requisite AG, again arguing that the selection of this sequence as a cryptic splice site is likely a consequence of library imperfections.

Intrinsic 5'-Splice Site Fidelity – The U1snRNP Binding Site Conundrum

Previous *in silico* sequence analyses have shown that 5' splice sites are often subject to cryptic splice site activation 4 nucleotides upstream or downstream from the canonical splice site due to the presence of the U1snRNP binding sequence (AG|**GURAGU**), which commonly includes a GU dinucleotide 4 nucleotides downstream from the canonical splice site (Dou et al. 2006). Our ultra-deep sequencing reveals the activation of an intrinsic cryptic exon7 5' splice site (GA|GTAA/GTCTGC) in 491 reads (error rate $7.7E-05$, Figure 2.4C, Table 2.3). While the canonical 5' splice site is reasonably strong splice site (MES = 8.57), it should be highly favored compared to this downstream intrinsic 5' splice site (MES = -7.82).

Table 2.3. 5' SS Cryptic Splicing

Cryptic Splice Site	Count	Error rate	Junction	Error Type	SS location vs canonical	MES
TG GTAA/GTAATC	32	5.0E-06	ex6 ex7	insertion	downstream	-1.24
GA GTAA/GTCTGC	491	7.7E-05	ex7 ex8	insertion	downstream	-7.82

The cryptic splice site is designated by a “|” and the canonical splice site is represented by a “|”. Count refers to the number of wild type reads that contain the cryptic splice site, with their associated rate of occurrence. Junction refers to the location of the cryptic splice site. Error type refers to the result of the cryptic splicing, either an insertion or deletion of sequence from the canonical transcript. SS location vs. canonical refers to the position of the cryptic splice site relative to the canonical splice site. MES is the Maximum entropy score for the cryptic splice site.

The selection of the cryptic 5' splice site at exon6|intron6 (TG|GTAA/GTAATC) was observed 32 times (error rate = 5.0E-06, Figure 2.4D, Table 2.3). This lower rate in cryptic splice site activation may be partially explained by a stronger canonical 5' splice site as determined by MES (11.01) even with the presence of a less unfavorable or weak intrinsic 5' splice site (MES = -1.24).

Microexon Discovery In SMN1

Another type of rare splicing variants observed are microexons. Many microexons, 3 to 30 nucleotides long, have not been annotated because of their rarity and size (Ustianenko et al. 2017). In our analysis we identified the presence of a microexon contained within intron 6. The microexon AG/ATCTGGG/GTAATGT is located 210 nucleotides upstream of the intron6|exon7 junction, and it was detected in 18 reads (2.8E-06). The microexon is flanked by weak splice sites (3' splice site, MES = 0.99; 5' splice site, MES = 4.85) and its usage does not rely on nucleotide changes at the splice junctions. Interestingly, this microexon uses the same 5' splice site as the recently discovered cryptic exon 7a (Yoshimoto et al. 2016). Thus, it is possible that the generation of the intron 6 microexon is an alternative splicing pathway in the generation of cryptic exon 7a.

Mutant Influence On Splicing Fidelity

Our library was created with synonymous mutations at all possible positions within a six-nucleotide window throughout SMN exon 7. Using the splice efficiency results from this mutant library we tested the hypothesis that positional mutants alter

canonical splice-site usage by increasing or decreasing cryptic splice-site usage. Here we refer to the ratio of cryptic splicing and canonical splicing in the mutant as compared to the wild-type as the Cryptic Splicing Value. There are several examples of mutations that show an increase in cryptic splicing. However, many of these same mutations also increase wild-type exon 7 inclusion levels, confounding the number of increased cryptic splice site usage reads with an increased total number of exon 7 inclusion reads. For instance, mutations at positions 42C→T+43T→C+45A→G has a Cryptic Splicing Value of 2.0 for splicing of AG/CCTCTGGN₁₀...CAG|GA at the intron7|exon8 junction. However, this same set of mutations is also responsible for a 2.1-fold increase in exon 7 inclusion according to its Inclusion Index Value (Mueller et al. 2015). Therefore, the increased number of reads containing cryptic splicing compared to wild-type is inherently tied to the increase of exon 7 inclusion by this same set of mutations.

To identify mutations that preferentially influence cryptic splice-site usage, we focused only on high incidence cryptic splice events and compared the Cryptic Splicing Value with the Inclusion Index Value. The result of this comparison is referred to as the Mutant Influence Value. Statistical significance of the Mutant Influence Value was tested using a difference of log odds ratio approach coupled with the Benjamini-Hochberg procedure to control the false discovery rate. In addition, a threshold was imposed that a cryptic splice event must have or be expected to have a minimum of 10 cryptically spliced reads (based on the wild-type cryptic splicing rate) to avoid outsized conclusions based on small sample size. Several mutant positions had statistically significant effects on cryptic and canonical splicing.

Cryptic 3' Splice Site Activation At The Exon7|Exon8 Junction

A frequently used cryptic splice site is located upstream of the canonical intron7|exon8 3' splice site (Figure 2.4A). Our analysis reveals that multiple mutations within exon 7 result in altered cryptic splice site usage, either increasing or decreasing its selection (Table 2.4). A significant decrease in cryptic 3' splice site usage occurs with the mutants 54A→C or 54A→G, which reside within the exon 7 5' splice site. Combinations of these mutations with 50A→G result in similar if slightly lesser effects. Other combinatorial mutations in the region 39 through 45 generally lead to further decreases in cryptic splice site usage (Table 2.4, Figure 2.5A). Conversely, the combinatorial mutation 3T→G+6C→T results in greater cryptic 3' splice site usage. Interestingly, the most influential mutants identified cluster to either the 5' or the 3' end of exon 7.

Cryptic Splicing (GGTTTCAG Deletion) At The Intron6|Exon7 Junction

When compared to wild-type cryptic 3' splicing at the intron6|exon7 junction AG|GGTTTCAG/AC is significantly reduced by the exon 7 mutations 28A→C,

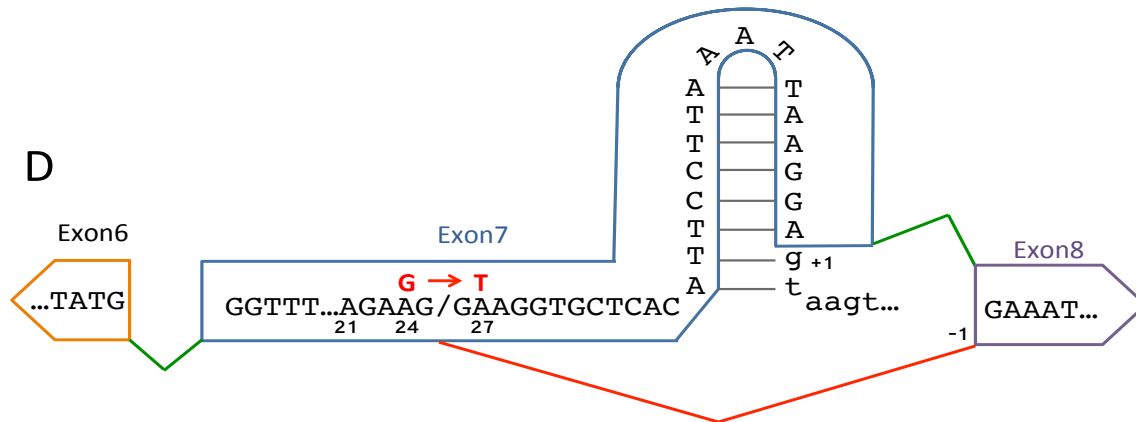
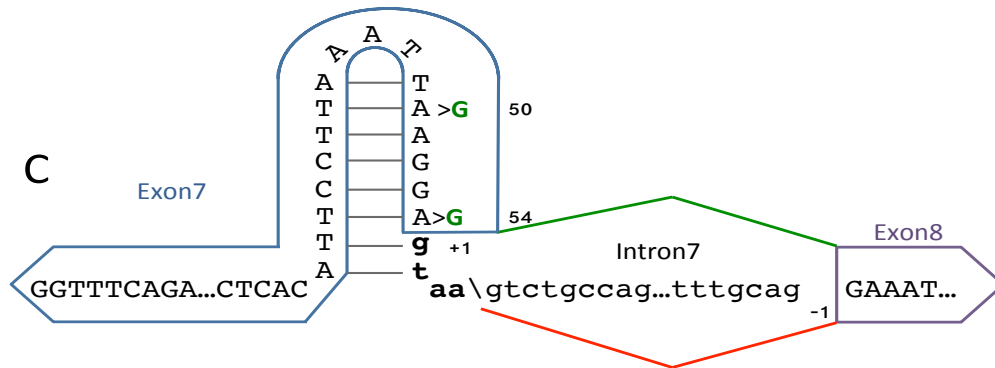
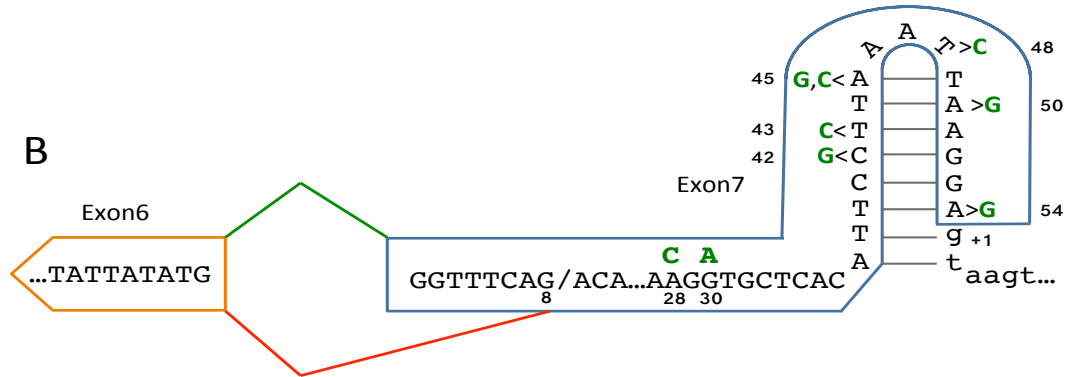
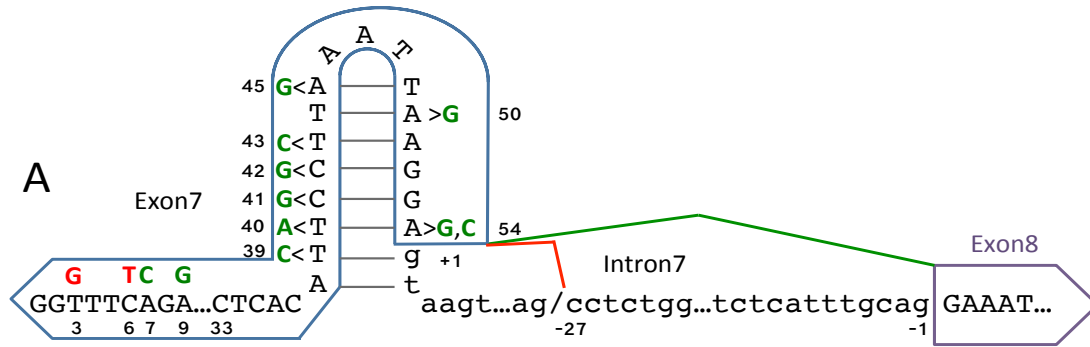
Table 2.4. Mutant Influence on Splicing Fidelity – Cryptic 3' SS AG/CCTCTGGN₁₀...CAG|GAA

Mutation	# Cryptic Spliced Reads	# Normal Spliced Reads	Inclusion Index Value	Cryptic Splicing Value	Mutant Influence Value	Mutant Influence Type
Wildtype	9433	6449627	1.0	1.0	0.0	N/A
54A→C	14	35679	1.7	0.2	1.5	SS Strength
54A→G	9	12474	3.3	0.5	2.8	SS Strength
3T→G+6C→T	34	19447	0.1	1.2	-1.1	SRE
7A→C+9A→G	1535	880320	1.6	1.2	0.4	SRE
39T→C+40T→A+41C→G	49	34670	2.2	1.0	1.2	- TSL2
42C→G+43T→C	23	16141	2.3	1.0	1.3	- TSL2
42C→G+43T→C+45A→G	22	10846	2.8	1.4	1.4	- TSL2
42C→G+45A→G	102	49001	2.3	1.4	0.9	- TSL2
50A→G+54A→C	27	41501	1.6	0.4	1.2	SS Strength
50A→G+54A→G	56	75269	3.0	0.5	2.5	SS Strength

Mutant Influence Values are shaded green for the canonical, white for no influence and the red for cryptic. Significance at Benjamini-Hochberg FDR = 0.2. SS Strength = a change in splice site strength, SRE refers to the alteration of a splicing regulatory element, and - TSL2 refers to the weakening of Terminal Stem Loop 2 within exon 7.

Figure 2.5. Mutant influence on cryptic splice site usage

A) Mutant influence on cryptic 3' SS usage between exon 7 and exon 8. The green line represents canonical SS usage, while the red line represents cryptic 3' SS usage 27 nucleotides upstream of the canonical intron7|exon8 junction. Mutations at positions in green influence more canonical SS usage, while mutations at position in red influence more cryptic SS usage. **B)** Mutant influence on cryptic 3' SS usage between exon 6 and exon 7. The green line represents canonical SS usage, while the red line represents cryptic 5' SS usage 4 nucleotides downstream of the canonical intron6|exon7 junction. Mutations at positions in green influence more canonical SS usage. **C)** Mutant influence on cryptic 5' SS usage between exon 7 and exon 8. The green line represents canonical SS usage, while the red line represents cryptic 5' SS usage 4 nucleotides downstream of the canonical exon7|intron7 junction. Mutations at positions in green influence more canonical SS usage. **D)** Mutant at position 27 creates 5' SS. The green line represents canonical SS usage, while the red line represents cryptic 5' SS usage. Mutations at positions in red influence more cryptic 5' SS usage.



30G→A, and 45A→G (Table 2.5, Figure 2.5B). As was observed for other cryptic splicing events, combinations of mutants generally preserve the overall effect single mutants have. By contrast, no mutant results in significantly increased cryptic splicing compared to wild-type.

Cryptic Splicing (GTAA Insertion) At The Exon7|Intron8 Junction

Another abundant cryptic splice site that is affected by mutation is the retention of GTAA by activation of an intrinsic 5' cryptic splice site (GA|GTAA/GTCTGC) for exon 7. A significant decrease in cryptic 5' splice site usage occurs with the combinatorial mutant 50A→G + 54A→G (Table 2.6, Fig 2.5C).

Mutations Create A Highly Efficient De Novo Cryptic 5' Splice Site

By far the most abundant example of cryptic splicing in our dataset is the truncation of exon 7 to the first 25 nucleotides. This 5' cryptic splicing event (AAG/GAAGGT) results in 309,793 reads containing a truncated exon 7 with no other mutations (Table 2.7, Figure 2.5D). However, at a splice site usage rate of 1 in ~21 wild type transcripts, this cryptic splicing event is common enough that it should have been readily discovered and annotated without using ultra-deep sequencing. This cryptic splicing event is most likely the result of one or more mutations downstream of position 25. An analysis of the effects of other mutations on this truncated exon 7 was performed. The difficulty in this analysis is that any mutation that occurs at the 26th through the 54th positions of exon 7

Table 2.5. Mutant Influence on Splicing Fidelity - 3'SS AG|GGTTTCAG/ACA

Mutation	# Cryptic Spliced Reads	# Normal Spliced Reads	Inclusion Index Value	Cryptic Splicing Value	Mutant Influence Value	Mutant Influence Type
Wildtype	825	6449627	1.0	1.0	0.0	N/A
28A→C	38	519479	1.5	0.5	1.0	SRE
30G→A	82	1378053	1.4	0.4	1.0	SRE
45A→G	6	138179	2.0	0.3	1.7	- TSL2
42C→G+45A→G	5	49001	2.3	0.8	1.5	- TSL2
43T→C+45A→C	3	41164	2.4	0.6	1.8	- TSL2
43T→C+45A→C+48T→C	5	62155	2.3	0.6	1.7	- TSL2
43T→C+45A→G	4	39199	2.5	0.8	1.7	- TSL2
43T→C+45A→G+48T→C	2	39842	2.5	0.4	2.1	- TSL2
50A→G+54A→G	5	75269	2.9	0.5	2.4	SS Strength

54

Mutant Influence Values are shaded green for the canonical, white for no influence and the red for cryptic. Significance at Benjamini-Hochberg FDR = 0.2. SS Strength = a change in splice site strength, SRE refers to the alteration of a splicing regulatory element, and +/- TSL2 refers to the weakening of Terminal Stem Loop 2 within exon 7.

Table 2.6. Mutant Influence on Splicing Fidelity – 5'SS GA|GTAA/GTCTGC

Mutation	# Cryptic Spliced Reads	# Normal Spliced Reads	Inclusion Index Value	Cryptic Splicing Value	Mutant Influence Value	Mutant Influence Type
Wildtype	491	6449627	1.0	1.0	0.0	N/A
50A→G+54A→G	2	75269	3.0	0.4	2.6	SS Strength
54A→G#	0	12474	3.3	N/A	N/A	SS Strength?
50A→G#	7	143757	1.5	0.6	0.9	N/A

Mutant Influence Values are shaded green for the canonical, white for no influence and the red for cryptic. Significance at Benjamini-Hochberg FDR = 0.2. SS Strength = a change in splice site strength. #=Not statistically significant.

Table 2.7. Mutation Derived Cryptic 5'SS At Position 27 Results In Truncated Exon 7

Mutation	# Cryptic Spliced Reads	# Normal Spliced Reads	Inclusion Index Value	Cryptic Splicing Value	Mutant Influence Value	Mutant Influence Type
Wildtype	309793	6449627	1.1	218.5	-217.4	5'SS
21A→G+24A→G#	9	145745	1.2	0.3	0.9	N/A
9A→G	230	1046328	1.0	1.0	0.00	N/A
24A→G	9624	289440	0.8	151.3	-150.5	5'SS

Mutations at position 27 create a highly efficient cryptic 5' splice site. Calculations are based off of splicing neutral 9A→G instead of wildtype. Mutant Influence Values are shaded green for the canonical, white for no influence and the red for cryptic. Significance at Benjamini-Hochberg FDR = 0.2. 5'SS = Creation of highly efficient 5'SS. #=Not statistically significant.

cannot be accurately assessed, as the mutation will be omitted from the read as a result of the truncation of exon 7. The second obstacle in the analysis is that the wild-type reads cannot be used as the comparative baseline. In this instance we normalized to the splicing neutral 9A→G (Pedrotti et al. 2010). Normalization to splicing neutral 15A→C provided similar results (not shown). The mutation 24A→G occurred in far more truncated exon 7 reads (9,624) than any other mutant (Table 2.7, Figure 2.5D). But the combinatorial mutation 21A→G+24A→G did not occur at an increased rate in truncated exon 7 reads, when compared to splicing neutral mutations 9A→G and 15T→C. We propose that the truncated exon 7 reads containing the mutation 24A→G are actually the result of the 24A→G+27A→T combinatorial mutant (and probably to a lesser extent 24A→G+27A→C), as these mutations exhibit the same exon 7 inclusion behavior as the single mutations 27A→T and 27A→C in a previous study (Mueller et al. 2015).

Cryptic Splicing During Skipping Of Exon 7

The mini-gene used in our analysis also generates mRNA transcripts with skipped exon 7. We recovered ~4 million reads that include exon 6 and exon 8, but skip exon 7 in our dataset. While we are unable to directly validate the junctions observed in these skipped exon 7 reads, they can be compared to the exon 7 inclusion reads. In general, exon 7 exclusion events are susceptible to the same cryptic splicing errors at similar rates. The cryptic 3' splice site upstream of exon 8 (AG/CCTCTGGN₁₀...CAG|GA...) is observed at a frequency almost identical to the usage rate seen for exon 7 inclusion events (1.5E-03 vs 1.5E-03) (Table 2.2, Table 2.8, Figure 2.6A). Other cryptic splice site selection events at the exon 6|intron 6 or the

intron7|exon8 junctions are also observed at similar frequencies (Tables 2.2, 2.3, 2.8, Figure 2.6A-B). A surprising result in the analysis of exon 7 exclusion transcripts was the discovery of a GA dinucleotide frequently inserted between exon 6 and exon 8 (1257 occurrences, error rate = 3.0E-04, Table 5). It is possible that this frequent inclusion event is triggered through recursive splicing (Figure 2.6C).

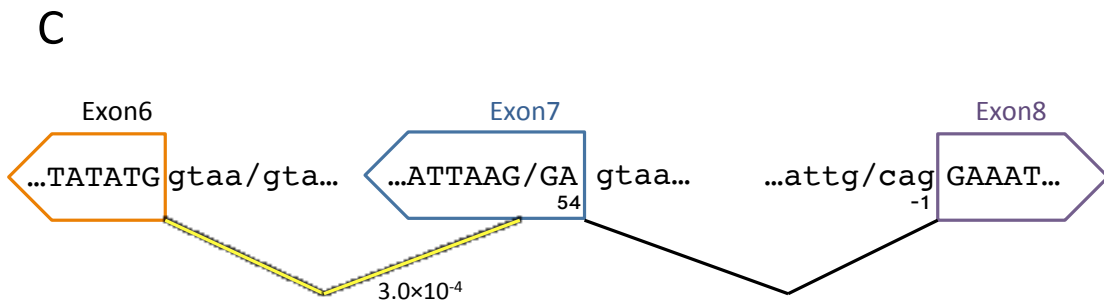
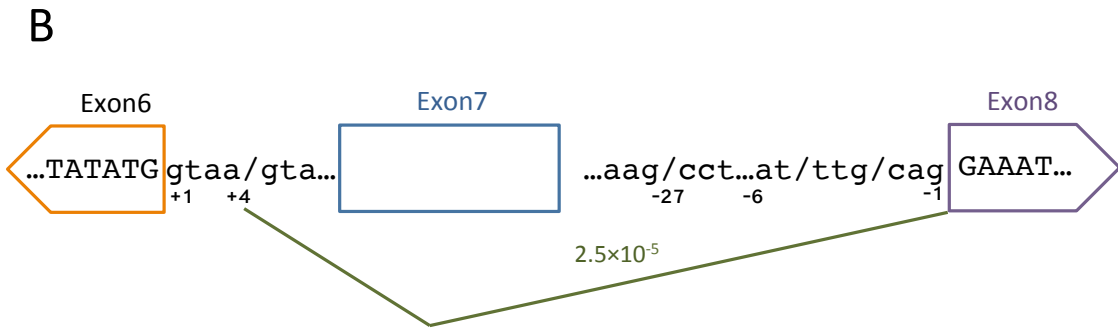
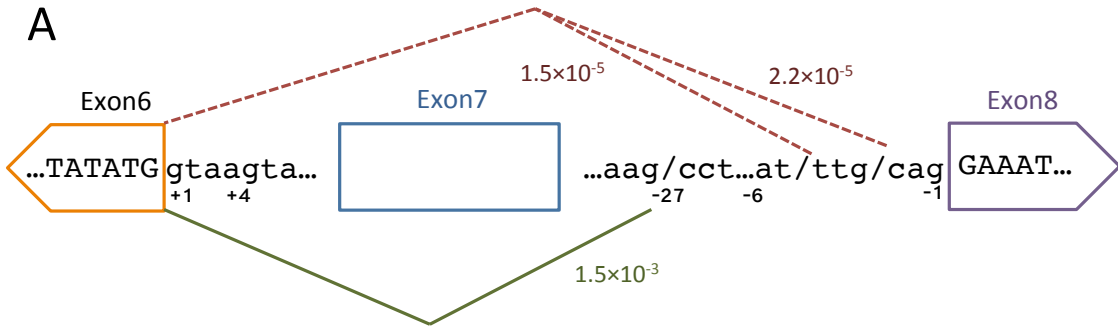
Table 2.8. Skipped Exon 7 Cryptic Splicing

Cryptic Splice Site	Count	Error rate	Junction	Error Type	SS location vs canonical
AG/CCTCTGGN ₁₀ ...CAG GAA	6372	1.5E-03	ex6 ex8	insertion	upstream
AG/GA GTAAGT	1257	3.0E-04	ex6 ex8	insertion	recursive splicing
TG GTAA/GTAATC	107	2.5E-05	ex6 ex8	insertion	downstream
TG/CAG GAA	92	2.2E-05	ex6 ex8	insertion	upstream
AT/TTGCAG GAA	63	1.5E-05	ex6 ex8	insertion	upstream
AG/ATCTGGG/GTAATGT	3	7.0E-07	ex6 ex8	insertion	microexon

The cryptic splice site is designated by a “/” and the canonical splice site is represented by a “|”. Count refers to the number of skipped exon 7 reads that contain the cryptic splice site, with their associated rate of occurrence. Junction refers to the location of the cryptic splice site. Error type refers to the result of the cryptic splicing, either an insertion or deletion of sequence from the canonical transcript. . SS location vs. canonical refers to the position of the cryptic splice site relative to the canonical splice site.

Figure 2.6. Cryptic SS usage with excluded exon 7.

A) Cryptic 3' SS usage between exon 6 and exon 8. The green line represents the unannotated cryptic 3' SS event 27 nucleotides upstream of the canonical intron7|exon8 junction. The red dashed line represents the cryptic splicing observed that is due to transcription or library generation errors resulting in canonical AG dinucleotide sequences. B) Cryptic 5' SS usage between exon 6 and exon 8. The green line represents the usage of the intrinsic cryptic 5' SS. C) Recursive splicing resulting in GA dinucleotide insertion between exon 6 and exon 8.



DISCUSSION

Cryptic Splicing Detection Is Limited By Sequence Context And Sequencing Accuracy

The most common spliced sequence deviation in our dataset was the deletion of a single guanosine from a GGG triplet at the exon6|exon7 junction. However, the DNA reads, which display frequencies of single G insertions or deletions at the intron6|exon7 junction similar to those detected at the exon6|exon7 junction have not been subjected to RNA processing, which suggests that the insertion or deletion of a single guanosine at GGG positions in the RNA reads is most likely not the result of a splicing error. While these observed insertion/deletion error rates are higher than previous Illumina HiSeq sequencing error studies (Minoche et al. 2011; Nakamura et al. 2011), the GGG sequence context may be more prone to sequencing errors when compared to genome-wide observations. GC-rich regions, in particular GG homopolymers, are subject to higher error rates (Schirmer et al. 2016), which may explain the elevated error rates of the intron6|exon7 and exon6|exon7 junctions that harbor GGG motifs. It is also possible that during the library construction plasmids were created with an inserted/deleted guanosine at one of these positions due to the same GGG motif error tendencies. The deletion of a guanosine from the intron6|exon7 junction has no predicted deleterious effect on 3' splicing; in fact this "mutant" sequence results in a more favorable 3' splice site strength (MES = 11.33 → 12.23), which may be why we see such a strong positive correlation in DNA and RNA reads with this particular deletion. If the guanosine deletion were deleterious, we would expect to only find the deletion in the DNA reads. The insertion of a guanosine at this site, while at an order of magnitude less than the

observed guanosine deletion, occurs at similar rates in both DNA and RNA reads and likewise does not introduce obvious changes in splicing efficiency (MES = 11.33 → 11.10). Additionally, we do not detect other frequent insertions or deletions in our DNA reads. Furthermore, insertions and deletions found only in the RNA reads cannot be attributed to library mistakes as the generated exon/exon junction is made after intron removal. These observations support the idea that the numerous guanosine insertion/deletion errors found at the intron6|exon7 junction are attributable to sequencing errors and library construction errors inherently associated with GGG motifs.

It has been reported that error rates of splicing are comparable to the error rates of transcription, and that it may in fact be limiting factor for splicing fidelity (Fox-Walsh and Hertel 2009; Mellert et al. 2011). Nearly every splicing event we captured that deviates from the canonical exon 7 can be reasonably explained, either by known splicing mechanisms, or known or potential errors introduced upstream of splicing, which include our library construction and transcription. In fact, while the errors in our library limit our ability to accurately pinpoint the exact rates for all cryptic splicing sites, they do provide evidence that the spliceosome is bound to act according to the context of the surrounding sequence.

In our study, the most abundant cryptic splice site usage, in both quantity of sites and usage frequency of those sites, is found in the selection of 3' splice sites. This may be due to the two-step process of 3' splice site pairing, where there are more variables for the spliceosome to consider, than the U1 snRNP binding site controlled by the 5' splice site. Given the low abundance of cryptic splicing detected, we did not find

evidence for multiple cryptic splice events within a single transcript. This observation suggests that upstream cryptic splicing does not increase the likelihood of downstream cryptic splicing, at least as far as our mini-gene approach can decipher. Our results demonstrate that cryptic splicing usage is constrained by the sequence surrounding putative splice sites and that cryptic splicing is more likely to occur at 3' splice sites.

Cryptic 5' Splice Sites Do Not Preserve Reading Frame

The SMN1 mini-gene is not subject to mRNA surveillance methods such as NMD, as it lacks the features of a full-length transcript. In our study we hoped to discern raw splicing error rates, independent of transcript correctional surveillance methods, such as NMD. This allows us to detect all potential splicing errors and cryptic splicing. Dou et al. postulated that downstream cryptic 5' splice sites at the +4 position are subjected to less selective pressures as they are not part of the coding region. In our study we found this to be the case, as only cryptic 5' splice sites located within the canonical intronic region were selected at a statistically significant level. Additionally, only a handful of cryptic 5' splice sites (not enough to meet statistical analysis thresholds) were detected that reside within canonical exon 6 or exon 7. Furthermore, within the context of *SMN1* exon 7, a downstream +4 cryptic 5'splice site will be positioned after the natural stop codon, which is located in exon 7. While this cryptic splicing event would contribute to lengthening the 3'UTR by 4 nucleotides, it would not disrupt the coding region of the gene, nor be subject to NMD. This would require less selective pressures to be in place to ensure biological controls at the 5' splice site of exon 7. Our findings support this hypothesis, as the number of cryptic 5' spliced events

at exon 6 is 10-fold less than those detected for exon 7. As expected from the independence of NMD, we do not detect a direct reading frame preference for the splicing errors observed. Thus, the splicing events observed, cryptic or canonical, are all a result of a contextual sequence dependent process not beholden to frame preservation.

Skipped Exon 7

While the preferential usage of cryptic splicing is the result of several different mutant influences, the fact that they occur at similar frequencies in exon 7 inclusion and exclusion types suggests that splice site selection occurs at each splice site independently. The most surprising result in the analysis of exon 7 exclusion transcripts was the discovery of a GA dinucleotide frequently inserted between exon 6 and exon 8 (Table 2.8, Figure 2.6C). As this error occurs at a higher rate than our established library construction error rate, it is highly likely that this event is the result of cryptic splice site usage. This unexpected isoform may be generated through a form of recursive splicing, where all but the last two nucleotides of exon 7 are lost to a cryptic 3' splice site (AAG/GA|GTAAGT) contained at position 52. It is possible that this cryptic 3' splice site is selected and once intron 6 and the first 52 nucleotides of exon 7 are excised, the last two nucleotides of exon 7 are ligated to exon 6. This rare new isoform again solidifies our expectations of finding previously unannotated cryptic splice events in our deep sequencing dataset.

Mutations Identify Splicing Fidelity As Sequence Driven Process Influenced By Splicing Efficiency

Our analysis on the effect of mutations on cryptic splice site usage found that three abundant cryptic splice sites were indeed subjected to altered usage rates by several different mutations that seem to converge into a few major types of splicing influencers.

3' Cryptic Splice Site At The Exon7|Exon8 Junction

Mutational effects on cryptic 3' splicing at the exon7|exon8 junction (Table 4, Figure 5A) can be explained by three separate factors that are expected to alter the splicing efficiency. The first factor is the manipulation of splicing regulatory elements (SREs) within exon 7. The exon 7 mutant 3T→G+6C→T was implicated as a mutation that affects cryptic splice site usage and greatly reduces exon 7 inclusion (Mueller et al. 2015). It is well known that the mutation 6C→T in *SMN1* results in decreased inclusion of exon 7 (Lorson et al. 1999; Cartegni et al. 2006). It has been shown that combinatorial mutations at position 3T→A or G synergize with 6T to further decrease exon 7 inclusion levels (Mueller et al. 2015; Singh et al. 2017). What is interesting is that this mutation near the 3' splice site of exon 7 has such a marked effect on downstream cryptic 3' splice site usage. We propose that this increase in cryptic splice site usage is due to the destruction of SRE binding sites that result in inefficient splicing of exon 7, therefore promoting cryptic 3' splice site selection at the exon7|exon8 junction.

The impact of mutational influence on SREs is again manifested in the mutation 7A→C+9A→G, which shows modest decreases in cryptic splice site usage. The

mutation 7A→C has been shown to inhibit the binding of Sam68, which promotes exon 7 inclusion, while a mutation at position 9 is splicing inclusion neutral (Pedrotti et al. 2010). It is possible that the inhibition of Sam68 or some other regulatory protein promotes exon 7 definition, decreasing recognition lag times, which can lead to cryptic splice site selection.

The second factor affecting cryptic splicing at the exon7|exon8 junction is RNA secondary structure. All significant combinatorial mutations containing positions from 39 to 45 lie within the reported exon 7 inhibitory terminal stem-loop 2 (TSL2) (Singh et al. 2007). These mutations are predicted to simply disrupt the stability of this inhibitory RNA hairpin (Singh et al. 2007; Singh and Singh 2011), thereby promoting the canonical splicing pathway (Figure 2.5A). An interesting case arises in that 39T→C+40T→A+41C→G decreases cryptic splicing. The mutation at position 39 strengthens TSL2 (Figure 2.5A), potentially making the canonical 5' splice site less accessible. By contrast, the consecutive mismatched base pairing at positions 40 and 41 weaken TSL2 more than the matched pairing at position 39 strengthens it, thus promoting canonical splicing. Combinatorial mutations at other TSL2 positions create mismatches within the stem and result in decreased cryptic splice site usage, again most likely due to the disruption of TSL2.

The third factor involved in the mutational influence of splicing efficiency is the direct altering of splice site strength. Reduction of cryptic splicing for mutations 54A→C, and 54A→G (Table 2.4 Figure 2.5A) can be explained by an expected increase in exon 7 5' splice site strength (wild-type MES = 8.57 increases to 54A→C (MES = 9.39) and 54A→G (MES = 9.65)). The strengthening of the 5' splice site on exon 7 provides

increased exon definition, thus increasing splicing efficiency that favors the canonical pathway.

Cryptic Splicing (GGTTTCAG Deletion) At The Intron6|Exon7 Junction

The same three factors (SREs, RNA secondary structure, and splice site strength) that affect the exon7|exon8 junction similarly play a role in cryptic 3' splicing at the exon6|exon7 junction (Table 2.5, Figure 2.5B) The mutations of 28A→C, 30G→A, and the combination of both mutations reside within a conserved tract of exon 7 (Singh et al. 2007) that is directly adjacent to a SRE, the Tra2-β1 binding site, (Singh et al. 2004), a splicing enhancer (Hofmann et al. 2000; Watermann et al. 2006). Disruption of the TSL2 RNA secondary structure by combinatorial mutations between positions 42 and 50 generally decreases cryptic splicing at the exon6|exon7 junction. Similarly, splice site mutations (positions 50-54) increase canonical splicing and decrease cryptic splicing. The mutation 50A→G+54A→G significantly reduces cryptic 3' splicing in exon7. This mutation results in a modest increase in exon 7 5' splice site strength from a MES of 8.57 to 9.65. Taken alone, the decrease in cryptic splice site usage caused by mutation 50A→G+54A→G is difficult to explain. However, examining the individual mutations may provide clarity. Interestingly, the mutation 54A→G results in the detection of a single read with the cryptic splice site usage, while the mutation 50A→G alone does not significantly change cryptic splice site usage. As mentioned previously, the mutation 54A→G directly increases the 5' splice site strength. Even though 54A→G is not statistically significant in this case due to not meeting read count thresholds, we

can infer that the mutation 54A→G is the driving force behind the increases in canonical splice site usage.

Cryptic Splicing (GTAA Insertion) At The Exon7|Intron8 Junction

The only mutation that significantly influences cryptic 5' splice site usage in exon 7 is 50A→G+54A→G (Table 2.6, Figure 2.5C). Again, 54A→G is not statistically significant, as not a single read was detected with this cryptic splice site usage. In this case, and potentially others, the relative rarity of this cryptic splice event may have contributed to the lack of significant influential mutations. However, the appearance of 50A→G+54A→G as a significant mutation in determining cryptic splice site usage indicates the importance of splice site strength on splicing efficiency and its effect on splicing fidelity. As has been demonstrated throughout this analysis increasing the strength of the canonical 5' splice site results in greater inclusion of exon, and decreased usage of the cryptic splice site.

In short, changes in the amount of cryptic splicing can be explained by its inverse relationship with canonical splicing. Mutations that increase canonical splice site selection reduce cryptic splice site activation. By contrast, mutations that reduce canonical splice site recognition increase cryptic splice site selection by modifying SRE binding sites, splice site strength or RNA secondary structures.

Mutation At Position 27 Creates A Highly Efficient Cryptic 5' Splice Site

While the influence of the previously mentioned mutations on cryptic splicing are fairly straightforward, the most abundant cryptic splicing event is much more difficult to decipher. Although we detected 309,793 reads as having a cryptic splice site at position 25 resulting in a truncated exon 7, we showed that this cryptic splice event was not a previously annotated splice site. We propose that this cryptic splicing event is the result of the creation of a *de novo* cryptic 5' splice site predominantly by the mutation 27A→T and to a lesser extent, 27A→C (Table 2.7, Figure 2.5D). The mutation 27A→T creates a strong (MES = 10.29) 5' splice site AAG/GTAGGT. While the mutation 27A→C does not result in a strong splice site (MES = 2.53), splice sites have been known to use a GC instead of the canonical GU dinucleotide (Burset et al. 2000).

It was previously shown that combining mutants 27A→T and 27A→C with 28A→C decreased the level of exon 7 exclusion. We propose that the mechanism of this partial rescue is the weakening of the *de novo* cryptic 5' splice site created by mutations at position 27. Indeed the splice site strength of 27A→T+28A→C is decreased (MES = 7.13) below that of the downstream canonical 5' splice site (MES = 8.57). These results show how splicing efficiencies can be radically changed under the influence of a single mutation.

In summary, the mutational analysis provides evidence for our proposed model that the balance between cryptic and canonical splicing is determined by splicing efficiency. This idea that splicing efficiency is important to splicing fidelity extends beyond our SMN1 model. Several factors, including splice site strength and RNA secondary structure play an active role in determining the balance of splicing kinetics. The appearance of mutations at many sites within exon 7 that affect several different

cryptic splice sites highlights the difficult tightrope the spliceosome must walk, where some positions within an exon are important factors in splicing regulation at both the 3' and 5' splice sites.

MATERIALS AND METHODS

Cell Culture, Sequencing Library Preparation

The creation and sequencing of the mutant library was executed by our lab (Mueller et al. 2015), in brief, HeLa cells were used for creation and transfection of the SMN1 Exon 7 mutant library. These were maintained in monolayer at 37°C in Dulbecco's high glucose modified Eagle's medium (Invitrogen) supplemented with 10% fetal bovine serum, 4mM L-Glutamate, and 1mM Na-Pyruvate. Cell confluence was maintained at ~80% or less before splitting cells. Cells were transfected according to manufacturer's specifications for Lipofectamine 2000 (Invitrogen) for plate sizes of 10cm, 15cm, and 6-well plates with 3cm wells.

Bioinformatic Analysis of Splicing Fidelity for SMN1 mini-gene

We obtained 54,780,073 single-end reads of 100 nucleotides from the sequencing run. These reads were aligned to a custom index consisting of genomic SMN1 exons 6 to 8, spliced mRNA sequence consisting of exons 6, 7, and 8, and all exon 7 mutants that were introduced into the library. The reads were classified as either input reads, which would be the unprocessed DNA based reads, and output reads which would comprise all reads that were sequenced from processed mRNAs, which would include undergoing transcription and splicing. We used custom Python scripts to identify reads with wild-type exon 7 and all mutant exon 7 types. Regular expression search functions were employed to search for multiple anchor sequences associated with exon 6, exon 7, exon 7 mutants, and exon 8. In order to determine the existence of splicing errors we checked each read that contained these anchor sequences for either unexpected

additional sequence inserted between exon anchor sequences or sequence deleted from the expected anchor sequences that result in partial anchor sequences. Error rates were calculated as percentages where the total number of normal reads divided by the total number of reads that contained each distinct error. Reads where anchor sequences for exon 6 and exon 8 were found, but no anchor sequences from exon 7 were considered to be cryptically spliced reads where exon 7 was excised. Reads that did not contain anchor sequences or with multiple quality score based errors resulting in ambiguous nucleotides were discarded.

Mutation Position Effects on Splicing Fidelity

In order to calculate the effect of mutations on splicing fidelity, taking into account the rarity of events, we utilized odds ratios (OR) to determine those mutations that significantly change the ratios of splicing errors compared to the error rates observed in the wild-type SMN exon 7. We calculated the OR for each SMN1 exon 7 mutant type, by taking the rate of each distinct cryptic splicing event and divided it by the rate that the corresponding cryptic splicing event occurs in wild-type SMN1 exon 7. This we refer to as the Cryptic Splicing Value. To normalize the influence of exon 7 inclusion rates on the Cryptic Splicing Value, we took the difference between the published Inclusion Index Value, creating the Mutant Influence Value. We then took the absolute value of the Mutational Influence Value and calculated the standard error. A z-score statistic was calculated and used to determine the p-value for the difference between the Cryptic Splicing Value and the Inclusion Index Value. To account for multiples testing problems, the Benjamini-Hochberg procedure was used at a level of 0.2, to control the false

discovery rate. Additionally, a minimum of 10 cryptically spliced reads (based on the wild-type cryptic splicing rate) threshold was imposed to avoid outsized conclusions based on small sample size.

REFERENCES

- Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, et al. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**: 1587–93.
<http://www.ncbi.nlm.nih.gov/pubmed/23258890> (Accessed May 22, 2013).
- Buratti E, Chivers M, Královičová J, Romano M, Baralle M, Krainer AR, Vořechovský I. 2007. Aberrant 5' splice sites in human disease genes: Mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Res* **35**: 4250–4263.
- Burset M, Seledtsov IA, Solovyev V V. 2000. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res* **28**: 4364–75.
<http://www.ncbi.nlm.nih.gov/pubmed/11058137> (Accessed April 24, 2018).
- Cartegni L, Hastings ML, Calarco JA, de Stanchina E, Krainer AR. 2006. Determinants of exon 7 splicing in the spinal muscular atrophy genes, SMN1 and SMN2. *Am J Hum Genet* **78**: 63–77.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1380224&tool=pmcentrez&rendertype=abstract>.
- Chen L, Bush SJ, Tovar-Corona JM, Castillo-Morales A, Urrutia AO. 2014. Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity. *Mol Biol Evol* **31**: 1402–1413.
<https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msu083>.
- Daguenet E, Dujardin G, Valcárcel J. 2015. The pathogenicity of splicing defects: mechanistic insights into pre-mRNA processing inform novel therapeutic approaches. *EMBO Rep* **16**: 1640 LP-1655.
<http://embor.embopress.org/content/16/12/1640.abstract>.
- Dou Y, Fox-Walsh KL, Baldi PF, Hertel KJ. 2006. Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. *Bioinformatics* **12**: 2047–2056.
- Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science (80-)* **297**: 1007–1013.
- Fox-Walsh KL, Hertel KJ. 2009. Splice-site pairing is an intrinsically high fidelity process. *Proc Natl Acad Sci U S A* **106**: 1766–71.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2644112&tool=pmcentrez&rendertype=abstract>.
- Hofmann Y, Lorson CL, Stamm S, Androphy EJ, Wirth B. 2000. Htra2-beta 1 stimulates an exonic splicing enhancer and can restore full-length SMN expression to survival motor neuron 2 (SMN2). *Proc Natl Acad Sci* **97**: 9618–9623.
<http://www.pnas.org/cgi/doi/10.1073/pnas.160181697>.
- Horowitz DS. 2012. The mechanism of the second step of pre-mRNA splicing. *Wiley Interdiscip Rev RNA* **3**: 331–350.
- Imashimizu M, Oshima T, Lubkowska L, Kashlev M. 2013. Direct assessment of transcription fidelity by high-resolution RNA sequencing. *Nucleic Acids Res* **41**: 9090–9104.
- Jeon C, Agarwal K. 1996. Fidelity of RNA polymerase II transcription controlled by

- elongation factor TFIIIS. *Proc Natl Acad Sci U S A* **93**: 13677–13682.
- Koodathingal P, Novak T, Piccirilli JA, Staley JP. 2010. The DEAH box ATPases Prp16 and Prp43 cooperate to proofread 5' splice site cleavage during Pre-mRNA splicing. *Mol Cell* **39**: 385–395. <http://dx.doi.org/10.1016/j.molcel.2010.07.014>.
- Kornblihtt AR, Schor IE, Alló M, Dujardin G, Petrillo E, Muñoz MJ. 2013. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat Rev Mol Cell Biol* **14**: 153–165. <http://www.nature.com/articles/nrm3525>.
- Lim SR, Hertel KJ. 2001. Modulation of Survival Motor Neuron Pre-mRNA Splicing by Inhibition of Alternative 3' Splice Site Pairing. *J Biol Chem* **276**: 45476–45483.
- Lorson CL, Hahnen E, Androphy EJ, Wirth B. 1999. A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proc Natl Acad Sci U S A* **96**: 6307–11. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC26877/>.
- Mayas RM, Maita H, Staley JP. 2006. Exon ligation is proofread by the DExD/H-box ATPase Prp22p. *Nat Struct Mol Biol* **13**: 482–90. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3729281&tool=pmcentrez&rendertype=abstract>.
- Mellert K, Uhl M, Högel J, Lamla M, Kemkemer R, Kaufmann D. 2011. Aberrant single exon skipping is not altered by age in exons of NF1, RABAC1, AATF or PCGF2 in human blood cells and fibroblasts. *Genes (Basel)* **2**: 562–577.
- Minoche AE, Dohm JC, Himmelbauer H. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol* **12**: R112. <http://dx.doi.org/10.1186/gb-2011-12-11-r112>.
- Mueller WF, Larsen LSZ, Garibaldi A, Hatfield GW, Hertel KJ. 2015. The silent sway of splicing by synonymous substitutions. *J Biol Chem* **290**: 27700–27711.
- Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, et al. 2011. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* **39**. <http://dx.doi.org/10.1093/nar/gkr344>.
- Nesser NK, Peterson DO, Hawley DK. 2006. RNA polymerase II subunit Rpb9 is important for transcriptional fidelity in vivo. *Proc Natl Acad Sci* **103**: 3268–3273. <http://www.pnas.org/cgi/doi/10.1073/pnas.0511330103>.
- Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**: 457–63. <http://www.nature.com/articles/nature08909>.
- Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**: 1413–5. <http://www.ncbi.nlm.nih.gov/pubmed/18978789> (Accessed July 17, 2012).
- Pedrotti S, Bielli P, Paronetto MP, Ciccocanti F, Fimia GM, Stamm S, Manley JL, Sette C. 2010. The splicing regulator Sam68 binds to a novel exonic splicing silencer and functions in SMN2 alternative splicing in spinal muscular atrophy. *EMBO J* **29**: 1235–1247. <http://dx.doi.org/10.1038/emboj.2010.19>.
- Pickrell JK, Pai A a, Gilad Y, Pritchard JK. 2010. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet* **6**: e1001236. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3000347&tool=pmcentrez&rendertype=abstract> (Accessed August 14, 2013).
- Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. 2016. Illumina error profiles:

- resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* **17**: 125.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4787001&tool=pmcentrez&rendertype=abstract>.
- Scotti MM, Swanson MS. 2015. RNA mis-splicing in disease. *Nat Rev Genet* **17**: 19–32.
<http://www.nature.com/doi/10.1038/nrg.2015.3>.
- Semlow DR, Blanco MR, Walter NG, Staley JP. 2016. Spliceosomal DEAH-Box ATPases Remodel Pre-mRNA to Activate Alternative Splice Sites. *Cell* **164**: 985–998. <http://linkinghub.elsevier.com/retrieve/pii/S0092867416300022>.
- Semlow DR, Staley JP. 2012. Staying on message: ensuring fidelity in pre-mRNA splicing. *Trends Biochem Sci* **37**: 263–273.
<http://dx.doi.org/10.1016/j.tibs.2012.04.001>.
- Singh NN, Androphy EJ, Singh RN. 2004. In vivo selection reveals combinatorial controls that define a critical exon in the spinal muscular atrophy genes. *RNA* **10**: 1291–1305.
- Singh NN, Howell MD, Singh RN. 2017. Transcriptional and Splicing Regulation of Spinal Muscular Atrophy Genes. In *Spinal Muscular Atrophy*, pp. 75–97, Elsevier <http://linkinghub.elsevier.com/retrieve/pii/B9780128036853000057> (Accessed April 19, 2018).
- Singh NN, Singh RN. 2011. Alternative splicing in spinal muscular atrophy underscores the role of an intron definition model. *RNA Biol* **8**: 600–606.
<http://www.tandfonline.com/doi/abs/10.4161/rna.8.4.16224>.
- Singh NN, Singh RN, Androphy EJ. 2007. Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes. *Nucleic Acids Res* **35**: 371–389. <https://www.ncbi.nlm.nih.gov/pubmed/17170000>.
- Tsai K-W, Chan W-C, Hsu C-N, Lin W-C. 2010. Sequence features involved in the mechanism of 3' splice junction wobbling. *BMC Mol Biol* **11**: 34.
<http://www.ncbi.nlm.nih.gov/pubmed/20459675> (Accessed November 23, 2018).
- Ustianenko D, Weyn-Vanhentenryck SM, Zhang C. 2017. Microexons: discovery, regulation, and function. *Wiley Interdiscip Rev RNA* e1418.
<http://doi.wiley.com/10.1002/wrna.1418>.
- Wahl MC, Will CL, Lührmann R. 2009. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**: 701–18. <http://www.ncbi.nlm.nih.gov/pubmed/19239890> (Accessed July 21, 2011).
- Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–476. <http://dx.doi.org/10.1038/nature07509>.
- Watermann DO, Tang Y, Zur Hausen A, Jäger M, Stamm S, Stickeler E. 2006. Splicing factor Tra2-beta1 is specifically induced in breast cancer and regulates alternative splicing of the CD44 gene. *Cancer Res* **66**: 4774–80.
<http://www.ncbi.nlm.nih.gov/pubmed/16651431> (Accessed November 23, 2018).
- Yeo G, Burge CB. 2004. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *J Comput Biol* **11**: 377–394.
<http://www.liebertonline.com/doi/abs/10.1089/1066527041410418>.
- Yoshimoto S, Harahap NIF, Hamamura Y, Ar Rochmah M, Shima A, Morisada N, Shinohara M, Saito T, Saito K, Lai PS, et al. 2016. Alternative splicing of a cryptic

exon embedded in intron 6 of SMN1 and SMN2. *Hum Genome Var* **3**: 16040.
<http://www.nature.com/articles/hgv201640>.

CHAPTER 3

Alternative Splicing in Breast Cancer: A result of aberrant expression of splicing regulators not splicing fidelity.

SUMMARY

There are many alternative spliced mRNA transcript isoforms associated with cancer progression and metastasis. The aberrant splicing observed in disease states can be the result of spliceosomal mistakes through the modification of the splicing machinery itself or the dysregulation of splicing. The sheer amount of alternative splicing observed in breast cancer is difficult to assign as a result of spliceosomal mistakes. Here we describe a bioinformatics analysis to detect the gene expression profiles of splicing regulators and determine the extent of differential alternative splicing in estrogen receptor positive and estrogen receptor negative breast cancer cell types. We identified common sources of splicing dysregulation, including known drivers of cancer progression that may be responsible for widespread aberrant splicing. Our results align with current breast cancer knowledge, and suggest potential alternatively spliced isoforms that could be used as biological markers of breast cancer or potential therapeutic targets.

INTRODUCTION

Over half a million women die of breast cancer each year and it is the most common cause of cancer death among women under the age 50 years (Jemal et al. 2010). 90% of cancer deaths are a direct result of metastasis (Jia et al. 2015; Gupta and Massagué 2006). Because breast cancer is recognized as a complex disease network, most research focused on the identification of gene expression markers and profiles to establish early detection screens and better prognosis predictions for breast cancer (Morrow and Hortobagyi 2009). There are several breast cancer classifications, but the four major molecular subtypes are luminal, HER2-enriched, basal-like and normal-like (Perou et al. 2000; Sørlie et al. 2001; Parker et al. 2009). The expression of the estrogen receptor has been highly correlated with luminal subtypes of breast cancer (Sørlie et al. 2003). Indeed, approximately 75% of all breast cancers are categorized as luminal and express the estrogen receptor (Abe et al. 2005). Estrogen receptor positive (ER+) breast cancer has been treated by the selective estrogen-receptor modulator drug tamoxifen for nearly 20 years with varying levels of success (Fisher et al. 1998; Detre et al. 2017). The Cancer Genome Atlas Network analyzed ~800 breast cancers using multiple platforms of next-generation sequencing and microarrays to determine several genetic and molecular makeup of breast cancer, including DNA copy number, DNA methylation, gene expression and transcriptomic variance (Koboldt et al. 2012). This study also reinforced the ER+ and ER- molecular subtypes of breast cancer (Koboldt et al. 2012).

Recent work has also demonstrated that significant changes in alternative splicing accompany breast cancer, suggesting that an understanding of alternative

splicing unique to breast cancer could greatly increase prognosis accuracy (Venables et al. 2008; Sveen et al. 2016). These observations open new avenues of research in basic and translational molecular oncology.

Alternative splicing provides an additional layer of genomic complexity by producing multiple mRNAs and protein variants from any given gene (Grosso et al. 2008). Differential pre-mRNA processing contributes tremendously to genetic variability. It is estimated that transcripts from ~95% of multi-exon genes undergo alternative splicing (Pan et al. 2008). Splicing is carried out by spliceosomes, which are comprised primarily of small nuclear RNAs, (snRNAs) and snRNA associated proteins (snRNPs). It has been shown previously that the spliceosome pairs exons with a high degree of accuracy that may be limited by the quality of pre-mRNAs generated by RNA pol II (Fox-Walsh and Hertel 2009). If genes involved in maintaining the fidelity of splice site pairing are perturbed, it can result in a loss of splicing fidelity as has been demonstrated for spinal muscular atrophy (SMA) (Fox-Walsh and Hertel 2009). Spliceosomal activity is regulated by a number of components including hnRNPs (heterogeneous nuclear ribonucleoproteins) and SR proteins (serine/arginine-rich proteins) (Long and Caceres 2009). SR proteins and hnRNPs are RNA-binding proteins involved in the binding of nascent transcripts, alternative splicing and translational regulation (Long and Caceres 2009),(Han et al. 2010). Generally, splicing regulation is balanced by the activities of splicing enhancers and repressors. With the availability of RNA-seq, genome-wide approaches have been used to determine the prevalence of splicing in cancer. Recent studies indicate that each of the six hallmarks of cancer (sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative

immortality, inducing angiogenesis, and activating invasion and metastasis) (Hanahan and Weinberg 2011) can be affected by alternative splicing (Oltean and Bates 2014).

Complementing studies have indicated that there is a link between the mis-expression of SR proteins and the development of cancerous tissues (Long and Caceres 2009). Some regulators of alternative splicing have been observed to be expressed at elevated levels in cancer cells, potentially serving as markers for patient prognosis and the severity of malignancy (Tockman et al. 1997). Indeed, it has been shown that hnRNP H drives an oncogenic splicing switch in gliomas (Lefave et al. 2011). Other reports also correlate elevated expression of certain SR proteins with cancer (Ghigna et al. 1998). *SRSF1* is a proto-oncogene (Karni et al. 2007), is a critical transcriptional target of MYC and regulates apoptosis and proliferation to promote mammary epithelial cell transformation (Anczuków et al. 2012).

In addition to these core splicing regulators, several studies describe an ESRP-regulated splicing program involved in epithelial-mesenchymal transition (EMT), which may allow for an initiation of metastasis in cancer progression (Warzecha et al. 2010, 2009; Shapiro et al. 2011). *ESRP1* controls a *CD44* mRNA isoform switch from multiple isoforms to a standard isoform required for EMT in breast cancer (Brown and Reinke 2011). These observations suggest that members of splicing regulatory networks have the ability to function either as proto-oncogenes or as tumor suppressors.

Recent genome-wide analyses (Huelga et al. 2012) have shown that the expression change of only one of these splicing regulators can have profound effects on the expression of other regulators and the alternative splicing that they regulate. Therefore, identifying alternative splicing regulatory networks in breast cancer and their

possible role in cancer progression is needed to evaluate their ability to serve as powerful markers for cancer classification and outcome measures. Furthermore, understanding the molecular network involved in mediating breast cancer-specific alternative splicing will likely point to new alternative targets for breast cancer therapy.

In biology many molecular pathways interact to fine-tune gene expression. Understanding this interplay between pathways permits the derivation of codes (genetic code, histone code, epigenetic code, etc.), which seek to provide reliable predictions of expected behavior based on a set of overarching rules specific to that biological context. For instance, a splicing code was developed to predict the regulation of splicing in human tissues (Xiong et al. 2014). The splicing code uses hundreds of sequence-based features to predict the splicing behavior of an exon in different biological contexts such as different tissues, normal/cancer, and treated/untreated. Given a set of exons with known splicing outcomes, the splicing code can be trained to predict the splicing outcomes for all exons within that same biological context by summing the information gain from each feature. The contribution of a feature's information gain on predicted exon splicing behaviors can be used to highlight features that may figure prominently in a model of a tissue specific splicing program (Barash et al. 2010; Xiong et al. 2011). This is a clever strategy to sift through the numerous genes that can be identified using high throughput experiments. For example, this approach helped to identify MBNL proteins as repressors of embryonic cell specific alternative splicing and reprogramming (Han et al. 2013).

Several reports have linked the dysregulation of alternative pre-mRNA splicing with breast cancer (Karni et al. 2007; Anczuków et al. 2012; Shapiro et al. 2011;

Warzecha et al. 2009). While these studies focused on individual genes, high-throughput approaches are beginning to highlight the extent of alternative splicing in breast cancer. Recently, the analysis of 105 patients from the TCGA database identified 9 splicing factors involved in aberrant splicing in breast cancer (Wen et al. 2015).

If alternative splicing in breast cancer is a regulated process, similar to the regulation of tissue-specific splicing, it should exhibit a normal error rate of splicing and only certain genes under an altered regulatory message would be differentially spliced. However, if cancer-specific alternative splicing is the result of altered splicing fidelity, the production of many minor non-specific splice variants would be observed.

Here we describe the computational analysis of high-throughput experimental data to identify breast cancer-specific alternative splicing and its potential regulators. Based on unpublished work from our lab, breast-cancer specific alternative splicing is not caused by a change in the fidelity of the splicing reaction. Therefore, we sought to determine if the differential expression of splicing regulators perturbs normal splicing patterns in cancerous cells. These alternate splicing patterns may generate mRNA transcripts that are more favorable to cancer formation, either in general, or specific to cancer subtypes. We demonstrate that the altered splicing profile in breast cancer cell lines can mainly be attributed to the up-regulation of splicing factors, in particular SR proteins and the cell specific splicing regulators ESRP1 and ESRP2.

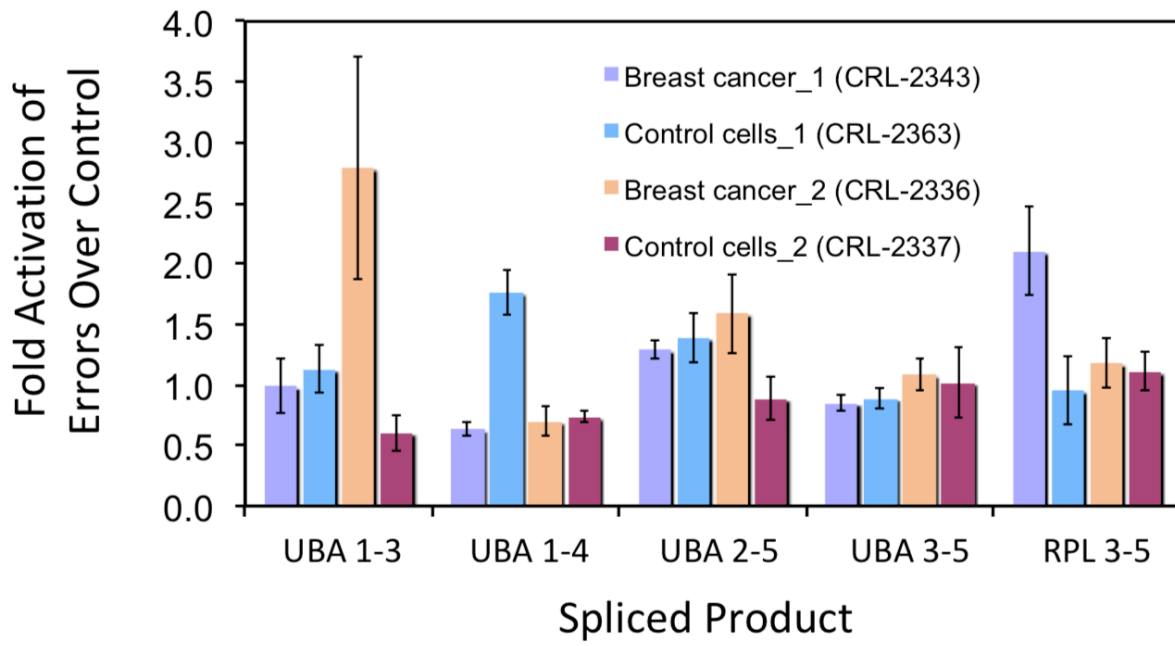
RESULTS

Splicing fidelity in breast cancer (unpublished data generated by Dr. Shu-Ning Hsu)

To examine whether the splicing fidelity is altered in breast cancer we performed splicing error analysis on paired normal and breast cancer cell lines. We used the A-52 ribosomal protein fusion product 1 (*UBA52*) and ribosomal protein-L 23 (*RPL23*) genes as fidelity readouts, two highly conserved and ubiquitously expressed genes, which are constitutively spliced (Fox-Walsh and Hertel 2009).

Unlike the case for SMA, the real-time PCR splicing error rate analysis showed no uniform pattern between breast cancer cell lines and their corresponding normal cell lines (Figure 3.1). For some of the analyzed splicing events the error rate increased while for others events it decreased. These results suggest that the splicing differences observed are not cause by an alteration of the intrinsic ability of the spliceosome to carry out intron removal at

Figure 3.1. Splicing Fidelity in Breast Cancer. Two breast cancer cell lines (CRL-2343, CRL-2336) and their respective normal matches (CRL-2363, CRL-2337) were evaluated for mis-spliced junctions all normalized to the splicing fidelity of HeLa cells. UBA 1-3 represents the splice junction of exon1 with exon3 in the constitutive gene *UBA52*.



its highest efficiency. Rather, the data suggest that the aberrant alternative splicing observed in breast cancer is a result of mis-regulation of splicing regulatory networks.

Analysis of gene expression in breast cancer cells

We predicted that abnormal splicing in luminal breast cancer is a result of differential expression of splicing regulatory factors. To evaluate the extent of breast cancer-specific alternative splicing and to measure differences in the expression of splicing factors, the gene expression profiles of breast cancer cells were compared to normal breast tissue using mRNA-Seq data generated from 3 luminal breast cancer cell lines (MCF7, T47D, BT474) and normal breast tissue (Wang et al. 2008). The mapped data sets were normalized with in-house bioinformatics tools using the Fragments Per Kilobase of exon model per Million mapped reads (FPKM) methodology (Mortazavi et al. 2008; Trapnell et al. 2010). FPKM values were compared between normal breast tissue and breast cancer cell lines to measure differential gene expression as a fold change of the average of the 3 breast cancer cell lines and the normal breast tissue sample. Evaluating genes with the largest differential expression between normal and the breast cancer cell lines is relatively straightforward. However, the genes with the largest fold change are not always functionally important and their identity may not allow us to assess if dysregulation of splicing networks is responsible for aberrant alternative splicing in breast cancer. To evaluate alternative splicing in breast cancer, a list of 280 known splicing regulatory and spliceosome component genes was derived using amiGO (Carbon et al. 2009), an online gene ontology tool.

Based on fold change, the expression of 90% of these splicing-related genes were observed to be higher in breast cancer cell lines when compared to normal breast tissue, including 52% which were expressed at a level greater than 2-fold higher (Table 3.1). For example, among splicing regulators, just a representative subset of splicing-related genes, the higher expression in breast cancer compared to normal is striking (Figure 3.2). These results could either be due to a higher genome-wide level of expression in breast cancer cells or part of a larger regulated network being perturbed in breast cancer. Evaluation of expression changes of all genes showed that while there was a general upshift in gene expression in breast cancer (58% of expressed genes are expressed at a higher level in the breast cancer cell lines than the normal tissue) it was not nearly as extreme as the shift shown in the list of splicing components and regulators.

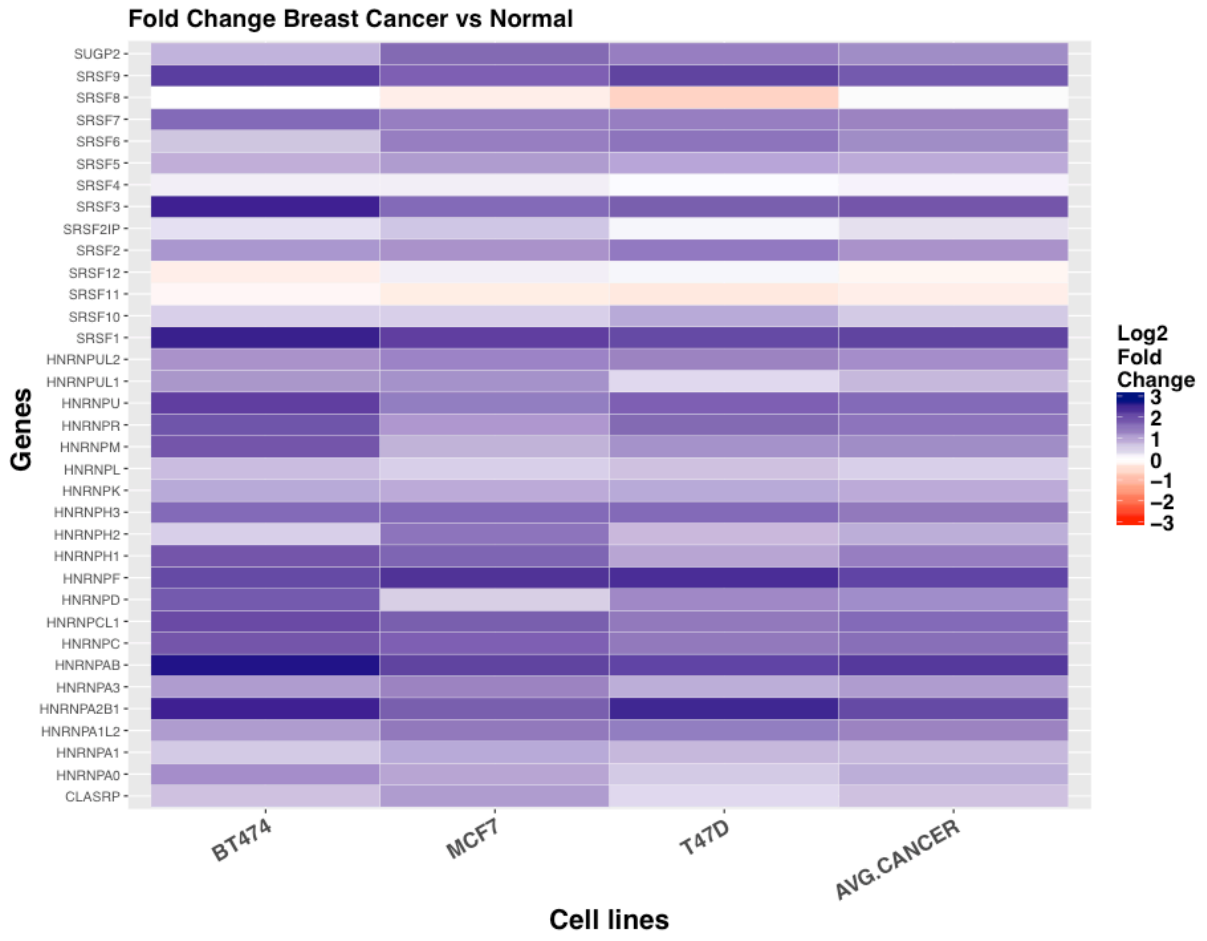
An identical analysis was carried out for another mRNA-Seq dataset (Sun et al. 2011) generated from 7 luminal breast cancer cell lines, 4 of which are Estrogen Receptor Positive (ER+) and 3 which are Estrogen Receptor Negative (ER-). This dataset includes the same cell lines (in bold) as the Wang et al. dataset (ER+: **MCF7**, **T47D**, **BT474**, ZR751; ER-: BT20, MDAMB231, MDAMB468) and the non-tumorigenic luminal breast cell line MCF10A as a control. The analysis of the Sun et al. dataset showed a similar trend of upregulated gene expression among spliceosomal components. While the

Table 3.1. Gene expression is upregulated in subset of splicing related genes.

Common is the average of the combination of breast cancer cell lines common to both the Wang et al. and Sun et al. datasets, which happen to be 3 of the 4 ER+ cell lines from the Sun et al. dataset. ER+ and ER- are the combined breast cancer cell lines from the Sun et al datasets with the presence or absence of the estrogen receptor.

Dataset	All Genes (29174 genes)			Splicing Related (280 genes)		
	Total Expressed	Upregulated	>2-fold change	Total Expressed	Upregulated	>2-fold change
Wang et al.	22886	58%	31%	277	90%	52%
Sun et al.	22317	67%	29%	278	92%	25%
Common	21905	63%	28%	278	90%	21%
ER+	22078	64%	28%	278	91%	21%
ER-	22086	64%	28%	278	92%	32%

Figure 3.2. Log2 Fold Change gene expression of splicing regulators in breast cancer cell lines compared to MCF10A. Upregulated expression is shown in purple, no change in expression is white, and downregulated expression is shown in red.



expression of 92% of splicing-related genes were upregulated, only 25% exhibited an upregulation in gene expression greater than two-fold when compared to MCF10A. Within the entire transcriptome, 67% of expressed genes are upregulated with respect to the normal breast tissue. While the change in the expression of splicing related genes is similar between both datasets, the magnitude of differences observed is lower. This could be due in part to different sources of control “normal type” used in the datasets; the immortalization of MCF10A cells compared to the more senescent normal breast tissue sample in the Wang et al dataset could exhibit gene expression signatures more in line with a fast-growing tumorigenic cell line. It is also possible that the larger set of breast cancer cell lines evaluated in the Sun et al. dataset reduced overall expression differences.

To compare the expression pattern observed in splicing-related genes to the expression pattern of other gene expression networks, we evaluated the gene expression of 81 genes associated with transcription. Similar to the pattern of expression observed in splicing related genes, 71 of the transcription-related genes (89%) were upregulated in breast cancer. The gene expression observed in a list of 452 translation-related genes exhibited a similar result, with 383 (85%) genes being upregulated (Table 3.2). The differential expression of spliceosomal components and splicing regulatory factors coupled with the splicing error analysis experiments suggests that altered splicing in breast cancer is a result of dysregulation, and not a result of changes in the splicing fidelity.

Table 3.2 Gene expression is upregulated in transcription and translation related genes.

Transcription (81 genes)			Translation (452 genes)		
Total Expressed	Upregulated	>2-fold change	Total Expressed	Upregulated	>2 -fold change
81	89%	21%	437	85%	34%

Differential gene expression of splicing related genes based on ER-delineated cell lines reveals few differences

The larger and more recent dataset (Sun et al. 2011) allows for the evaluation of the gene expression signatures and transcriptomic variation across cell lines that either express or lack the expression of the therapeutically significant estrogen receptor (ER). ER+ and ER- breast cancer cells exhibit different behaviors and gene signaling cascades. The comparison of the average gene expression of splicing related genes in ER+ and ER- cell lines showed very few differences. Many of the observed differences in the computed average were due to a single cell line exhibiting a remarkable difference, while the other cell lines within the ER-delineated groups were expressed at similar levels. To identify splicing related genes that may play a role in ER+/ER- cancer cell dynamics, we filtered the genes to those that exhibited an average expression difference of breast cancer cell lines compared to MCF10A greater than 2-fold change between ER+ and ER- cell lines. To control for highly differentially expressed outliers found in only one cell line, the standard deviation of the expression of the grouped cell lines could not exceed the observed difference. This process identified 2 splicing related-genes that are uniformly differentially expressed between ER+ and ER- cell types (Table 3.3). Both *ESRP2* (Epithelial Splicing Regulatory Protein 2) and *SF3B3* (Splicing Factor 3b Subunit 3) are expressed at a much higher level in ER+ cell lines.

Differential alternative splicing in MCF10A and breast cancer cell lines

It is known that breast cancer elicits alternative splicing alterations. We hypothesized that differential splicing in ER+ and ER- luminal breast cancer cell lines could improve the categorization of breast cancers and possibly serve as biological markers. To

determine breast cancer-specific differentially spliced events the RNA-seq datasets were analyzed using the computational tool MISO (Mixture of ISOforms) (Katz et al. 2010). We calculated the percentage of spliced-in (PSI or Ψ) values of every exon of genes in the MISO catalogue of splicing events and computed the Δ PSI ($\Delta\Psi$) values as Ψ breast cancer cell lines – Ψ MCF10A. The application of stringent filters to identify skipped exons, the most abundant alternative splicing event type, resulted in the identification of 78 events common and specific to the ER+ cell lines and 78 events common and specific to the ER- cell lines (Figures 3.3 and 3.4, Tables 3.4 and 3.5). A comparison of the ER+ and ER- cell lines revealed an overlap of 16 alternatively spliced skipped exon events common to all breast cancer cell types (Figure 3.5). Interestingly, among the genes harboring these common exon skipping events, 8 have been reported to associate with

Table 3.3. Difference between the average fold change of splicing related-genes in ER+ and ER- cancer cell lines compared to MCF10A.

Gene	AVG ER+	AVG ER-	Difference ER+/ER-
<i>SF3B3</i>	4.13	1.65	2.48
<i>ESRP2</i>	8.35	2.73	5.63

Figure 3.3. Venn diagram showing differentially spliced cassette exon events in ER- cell lines compared to MCF10A.

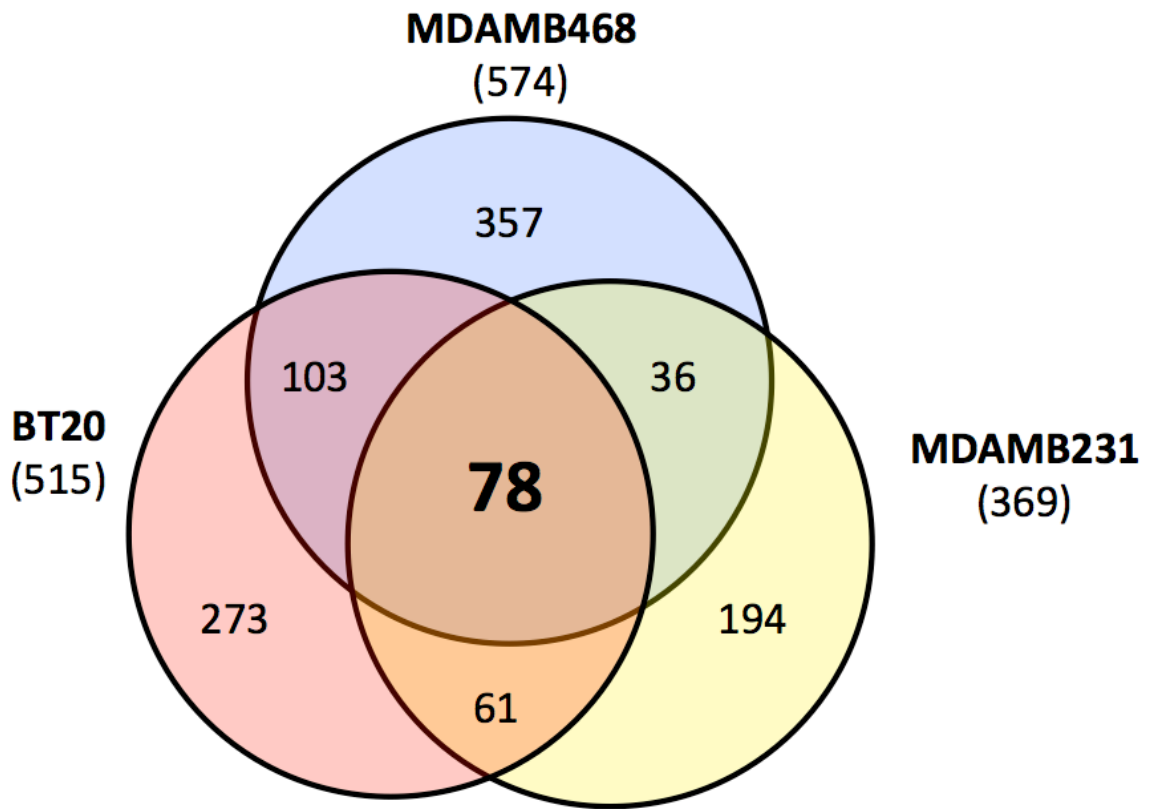


Figure 3.4. Venn diagram showing differentially spliced cassette exon events in ER+ cell lines compared to MCF10A.

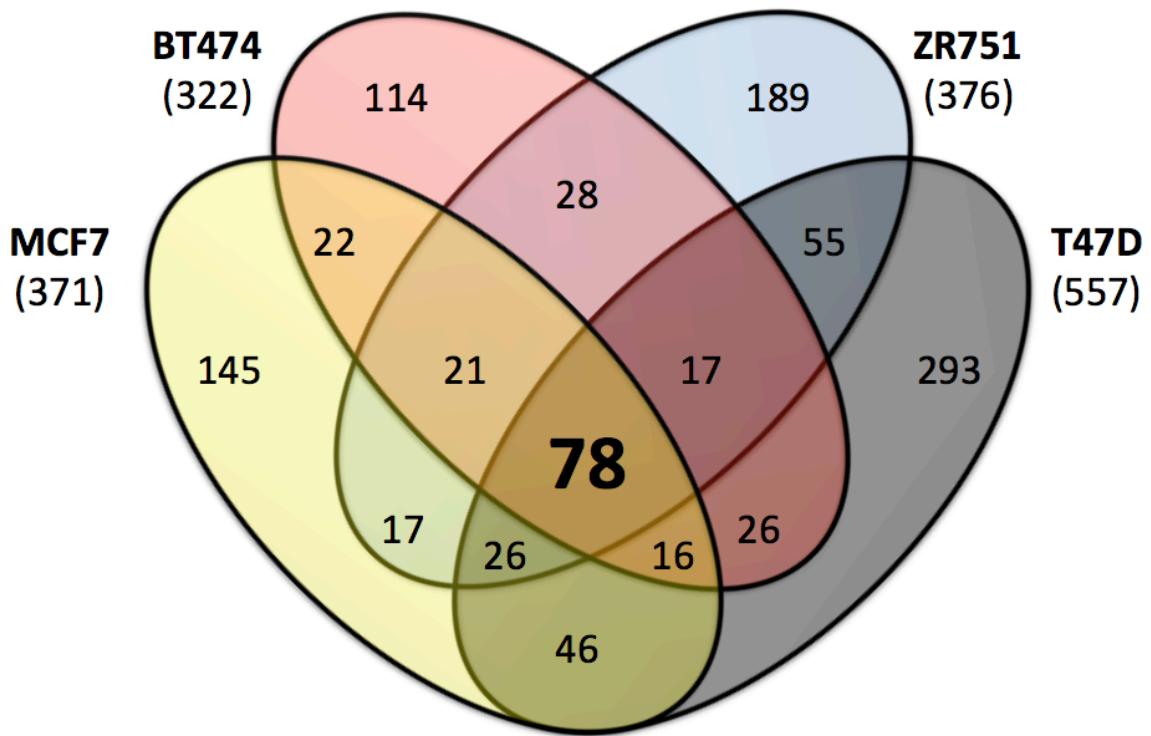


Table 3.4. Differentially spliced cassette exons in ER+ cell lines compared to MCF10A.

Skipped Exon Splicing Event	Ψ ER+	Ψ MCF10A	$\Delta\Psi$	Gene Name
chr10:27044584:27044670	0.87	0.47	0.4	ABI1
chr10:111892063:111892158	0.24	0.95	-0.71	ADD3
chr16:24950685:24950918	0.91	0.44	0.47	ARHGAP17
chrX:138813810:138813914	0.56	0.95	-0.39	ATP11C
chr11:61557257:61557462	0.22	0.80	-0.58	C11orf10
chr22:24939810:24940051	0.73	0.97	-0.24	C22orf13
chr5:2752736:2752868	0.58	0.17	0.41	C5orf38
chr5:2752794:2752868	0.42	0.06	0.36	C5orf38
chr3:191092851:191093378	0.13	0.74	-0.61	CCDC50
chr11:35232793:35232996	0.23	0.75	-0.52	CD44
chr1:22400587:22400712	0.81	0.33	0.48	CDC42
chr1:9797556:9797612	0.99	0.04	0.95	CLSTN1
chr1:9816539:9816568	0.45	0.03	0.42	CLSTN1
chr16:85813985:85814079	0.69	0.97	-0.28	COX4NB
chr22:38691393:38691453	0.08	0.43	-0.35	CSNK1E
chr11:57558966:57559145	0.63	0.04	0.59	CTNND1
chr11:57558857:57559145	0.77	0.02	0.75	CTNND1
chr11:57558857:57559145	0.57	0.04	0.53	CTNND1
chr11:57558966:57559145	0.54	0.06	0.48	CTNND1
chr8:11721885:11721972	0.40	0.07	0.33	CTSB
chr8:11718915:11718988	0.39	0.03	0.36	CTSB
chr1:68947729:68948580	0.16	0.74	-0.58	DEPDC1-V1
chr1:68948324:68948580	0.05	0.51	-0.46	DEPDC1-V1
chr5:140967791:140967817	0.97	0.61	0.36	DIAPH1
chr22:29725701:29725709	0.30	0.04	0.26	DKFZp686A01208
chr17:74086410:74086478	0.11	0.81	-0.70	EXOC7
chr17:78075610:78075689	0.53	0.95	-0.42	GAA
chr17:78075610:78075724	0.34	0.81	-0.47	GAA
chr12:54676863:54677018	0.43	0.08	0.35	HNRNPA1
chr14:105181621:105181677	0.64	0.90	-0.26	INF2
chr17:17083921:17083983	0.98	0.37	0.61	KIAA0864
chr1:115280092:115280184	0.85	0.47	0.38	KIAA0885
chr10:86259631:86259715	0.39	0.12	0.27	KIAA1128
chr14:51223210:51225348	0.95	0.20	0.75	KIAA1565
chr3:57911572:57911661	0.83	0.12	0.71	KIAA1601
chr6:17771345:17771449	0.33	0.95	-0.62	KIN13A
chr19:8315994:8316133	0.39	0.90	-0.51	LASS4
chr3:37132958:37133029	0.83	0.13	0.70	LRRFIP2

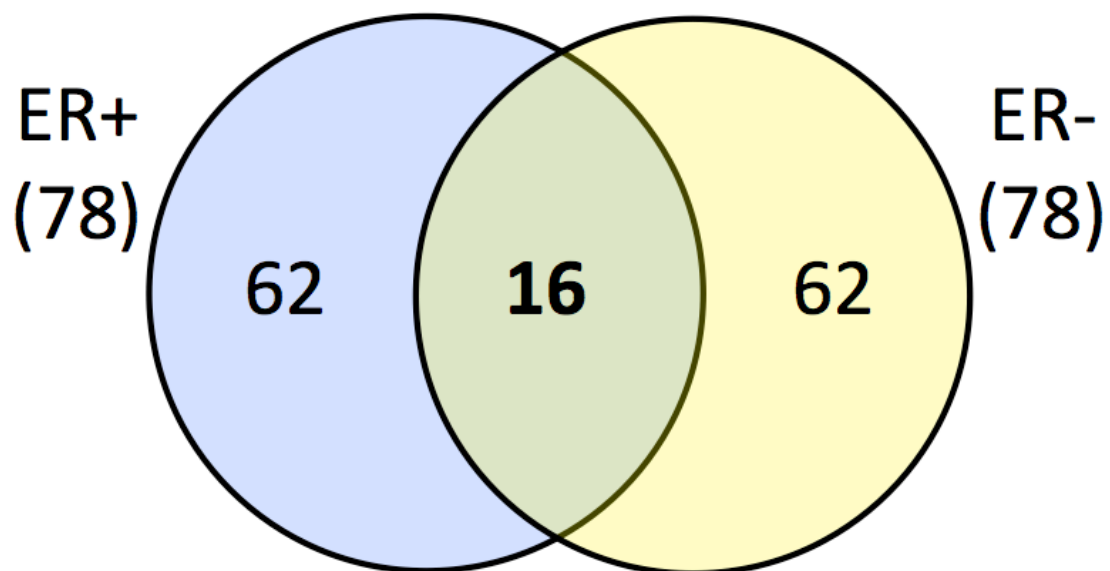
chr20:35927166:35927282	0.95	0.39	0.56	MANBAL
chr14:103964839:103964865	0.72	0.03	0.69	MARK3
chr14:103966493:103966537	0.50	0.03	0.47	MARK3
chr3:152164493:152164546	0.28	0.95	-0.67	MBNL1
chr12:56554410:56554454	0.81	0.28	0.53	MLC-3
chr2:238443207:238443290	0.43	0.11	0.32	MLPH
chr12:124811955:124812179	0.54	0.95	-0.41	NCOR2
chr17:8366638:8366672	0.70	0.20	0.50	NDEL1
chr21:44323292:44324386	0.78	0.28	0.50	NDUFV3
chr14:73745989:73746132	0.26	0.88	-0.62	NUMB
chr7:540068:540136	0.09	0.83	-0.74	PDGFA
chr8:144996672:145000052	0.46	0.81	-0.35	PLEC1
chr1:16047824:16047883	0.88	0.35	0.53	PLEKHM2
chr22:42997976:42998113	0.89	0.44	0.45	POLDIP3
chr2:128610499:128610679	0.36	0.82	-0.46	POLR2D
chr15:91512309:91512350	0.56	0.17	0.39	PRC1
chr1:201965275:201965537	0.45	0.82	-0.37	RNPEP
chr4:152021637:152021740	0.15	0.72	-0.57	RPS3A
chr14:94854897:94855000	0.71	0.33	0.38	SERPINA1
chr14:94854897:94855000	0.69	0.28	0.41	SERPINA1
chr14:94854897:94854997	0.69	0.29	0.40	SERPINA1
chr4:48396593:48396670	0.61	0.31	0.30	SLAIN2
chr10:105770574:105770666	0.15	0.95	-0.80	SLK
chr2:27594136:27594335	0.31	0.77	-0.46	SNX17
chr18:12459754:12459927	0.05	0.88	-0.83	SPIRE1
chr9:131355262:131355321	0.64	0.95	-0.31	SPTAN1
chr17:17726832:17726921	0.08	0.84	-0.76	SREBF1
chr12:131280540:131280665	0.19	0.67	-0.48	STX2
chr13:114285938:114286220	0.06	0.35	-0.29	TFDP1
chr22:50964430:50964585	0.83	0.31	0.52	TYMP
chr17:16285216:16285444	0.62	1.00	-0.38	UBB
chr4:76716489:76716509	0.07	0.85	-0.78	USO1
chr10:75280666:75280785	0.48	0.86	-0.38	USP54
chr14:100841620:100841687	0.79	0.04	0.75	WARS
chr14:100840473:100840581	0.73	0.03	0.70	WARS
chr14:100841620:100841740	0.84	0.04	0.80	WARS
chr14:100841620:100841743	0.82	0.02	0.80	WARS
chr12:988739:989197	0.39	0.05	0.34	WNK1
chr16:3335059:3335239	0.15	0.53	-0.38	ZNF263
chr1:71531361:71531435	0.82	0.46	0.36	ZRANB2

Table 3.5. Differentially spliced cassette exons in ER- cell lines compared to MCF10A.

Splicing Event	Ψ ER-	Ψ MCF10A	$\Delta\Psi$	Gene Name
chr7:73151259:73151440	0.52	0.85	-0.33	ABHD11
chr10:27044584:27044670	0.87	0.31	0.56	ABI1
chr1:155033239:155033308	0.39	0.93	-0.54	ADAM15
chr6:3264443:3264559	0.61	0.01	0.60	AK096219
chr16:30078206:30078359	0.51	0.01	0.50	ALDOA
chr15:60688350:60688626	0.09	0.60	-0.51	ANXA2
chr16:24939005:24939053	0.47	0.10	0.37	ARHGAP17
chr17:79826932:79826951	0.97	0.48	0.49	ARHGDI1A
chr7:12727260:12727353	0.40	0.02	0.38	ARL4A
chr21:42622679:42622828	0.28	0.52	-0.24	BACE2
chr6:31607277:31607423	0.74	0.38	0.36	BAT3
chr10:73979812:73980137	0.57	0.23	0.34	C10orf104
chr11:61557257:61557462	0.22	0.96	-0.74	C11orf10
chr17:16342842:16343017	0.67	0.38	0.29	C17orf45
chr20:35236293:35236403	0.61	0.12	0.49	C20orf24
chr6:160208775:160208903	0.97	0.62	0.35	CCT1
chr1:207940952:207940996	0.55	0.02	0.53	CD46
chr1:22400587:22400712	0.81	0.09	0.72	CDC42
chr5:148897357:148897440	0.67	0.08	0.59	CSNK1A1
chr8:11721885:11721972	0.40	0.13	0.27	CTSB
chr8:11718915:11718988	0.39	0.05	0.34	CTSB
chr6:31506717:31506836	0.55	0.01	0.54	DKFzp547B159
chr10:14595321:14595386	0.09	0.48	-0.39	FAM107B
chr4:187511522:187511557	0.42	0.03	0.39	FAT1
chr6:31804076:31804294	0.79	0.32	0.47	G8
chr6:31804072:31804294	0.77	0.31	0.46	G8
chr11:62401782:62401847	0.66	0.13	0.53	GANAB
chr6:138733220:138733393	0.36	0.01	0.35	HEBP2
chr20:36647407:36647546	0.52	0.11	0.41	KIAA0406
chr17:17083921:17083983	0.98	0.29	0.69	KIAA0864
chr1:115280092:115280184	0.85	0.15	0.70	KIAA0885
chr14:51223210:51225348	0.95	0.46	0.49	KIAA1565
chr20:35927166:35927282	0.95	0.30	0.65	MANBAL
chr3:152165514:152165562	0.97	0.13	0.84	MBNL1
chr13:98018713:98018807	0.49	0.83	-0.34	MBNL2
chr2:228193394:228193505	0.74	0.96	-0.22	MFF
chr12:56554410:56554454	0.81	0.20	0.61	MLC-3
chr9:6495585:6497103	0.65	0.15	0.50	NIRF

chr10:105153956:105154151	0.81	0.00	0.81	PD04912
chr21:45175358:45175645	0.21	0.47	-0.26	PDXK
chr12:53689623:53690059	0.11	0.49	-0.38	PFDN5
chr12:53689623:53690251	0.07	0.33	-0.26	PFDN5
chr2:128610499:128610679	0.36	0.79	-0.43	POLR2D
chr17:6916638:6916835	0.54	0.05	0.49	RNASEK
chr1:201965275:201965537	0.45	0.83	-0.38	RNPEP
chrX:100650323:100650445	0.14	1.00	-0.86	RPL36A
chr5:40834063:40834139	0.36	0.00	0.36	RPL37
chr10:79799962:79799983	0.40	0.05	0.35	RPS24
chr4:152022127:152022237	0.99	0.62	0.37	RPS3A
chr1:53416427:53416558	0.14	0.64	-0.50	SCP2
chr8:144889722:144889784	0.55	0.10	0.45	SCRIB
chr6:36567598:36568053	0.88	0.11	0.77	SFRS3
chr11:62651461:62651584	0.57	0.02	0.55	SLC3A2
chr15:66787668:66787757	0.09	0.58	-0.49	SNAPC5
chr19:49605371:49606844	0.91	0.14	0.77	SNRP70
chr9:91077411:91077664	0.97	0.64	0.33	SPIN1
chr9:91077407:91077664	0.96	0.58	0.38	SPIN1
chr9:130672230:130672807	0.20	0.65	-0.45	ST6GALNAC4
chr2:74056532:74056637	0.66	0.29	0.37	STAMBP
chr20:47782534:47782822	0.47	0.15	0.32	STAU1
chr8:54891083:54891257	0.60	0.00	0.60	TCEA1
chr13:114285938:114286220	0.06	0.64	-0.58	TFDP1
chr15:30011981:30012220	0.20	0.50	-0.30	TJP1
chr14:103596077:103596191	0.10	0.32	-0.22	TNFAIP2
chr2:73959711:73959827	0.46	0.01	0.45	TPRKB
chr12:104682709:104682818	0.59	0.80	-0.21	TR
chr11:2423069:2423377	0.83	0.44	0.39	TSSC4
chr22:50964430:50964585	0.83	0.07	0.76	TYMP
chr22:50964430:50964570	0.88	0.11	0.77	TYMP
chr17:16285216:16285444	0.62	0.99	-0.37	UBB
chr14:100841620:100841883	0.78	0.36	0.42	WARS
chr14:100840473:100840581	0.73	0.01	0.72	WARS
chr14:100841620:100841740	0.84	0.34	0.50	WARS
chr14:100841620:100841743	0.82	0.35	0.47	WARS
chr3:49049068:49049174	0.17	0.71	-0.54	WDR6
chr9:74978386:74978522	0.31	0.54	-0.23	ZFAND5

Figure 3.5. Venn diagram showing common differentially spliced cassette exon events between ER+ and ER- cell lines when compared to MCF10A.



cancer (Table 3.6). Similarly, all other alternative splicing events that we analyzed including alternative 3' splice sites (A3SS), alternative 5' splice sites (A5SS), alternative first exons (AFE), alternative last exons (ALE), mutually exclusive exons (MXE), and retained introns (RI) were observed to occur in both ER+ and ER- cell lines separately and with varying degree of overlap. Thus, these common alternative splicing events could be targets for isoform specific-marker detection and for functional analysis.

The genes associated with unique differential splicing in the ER+ or ER- only breast cancer subtypes could provide useful insights that can be exploited as diagnostic and prognostic markers and, therefore, should also be functionally analyzed. In summary, the comparative gene and mRNA isoform approach shows great promise in identifying biologically significant differentially spliced events specific to breast cancer and in different breast cancer subtypes.

A splicing code reveals features important to cassette exon inclusion and exclusion

To gain insights into the features that support inclusion or exclusion of cassette exons between normal and breast cancer cell lines, a splicing code (Busch and Hertel 2015) based on a support vector machine (SVM) algorithm was applied to our differentially spliced events results. The absolute number of events available for testing is important in this and all machine learning predictions. For this reason, only the cassette exons had a sufficient

Table 3.6. Differentially spliced events in breast cancer cell lines compared to MCF10A.

Splicing Event	Genes With Previously Identified Role in Cancer	ER+ only	ER- only	ER+/ER-
A3SS	C22orf13(Iorns et al. 2012), HNRNPD(Pont et al. 2012), DAXX(Li et al. 2013)	12	15	6
A5SS	TYMP(Goto et al. 2012), SMARCC2(Shain and Pollack 2013)	4	20	3
AFE	WARS(Kim et al. 2011; Wakasugi et al. 2002), BAT4(Ramirez et al. 2010), DDX39B(Kubota et al. 2012), RGS3(Shi et al. 2012), IMP3(Samanta et al. 2012)	22	36	11
ALE	SERF1A(Mustacchi et al. 2013), BAT4 (Ramirez et al. 2010), KIF1B (Henrich et al. 2012)	12	48	8
MXE	DRG1(Baig et al. 2012), USP10(Yuan et al. 2010)	26	61	2
RI	GNL3(Liu et al. 2010), MINK1(Venables et al. 2013), EMD(Capo-chichi et al. 2009)	7	34	4
SE	WARS(Kim et al. 2011; Wakasugi et al. 2002), CDC42(Johnson et al. 2010), MANBAL(Muraoka et al. 2012), ABI1(Chen et al. 2010), CTSB(Nouh et al. 2011), NIN(Olson et al. 2011), TYMP(Goto et al. 2012), TFDP1(Melchor et al. 2009)	78	78	16

number of events for the splicing code analysis to provide useful predictors. Using the list of all differentially spliced cassette exons between ER+ and ER- cell lines and MCF10A as determined by our filtered MISO analysis, the splicing code was trained and used to predict the expected splicing outcome, either inclusion or exclusion. The top features that provided the most information for the prediction of either exon inclusion or exon exclusion were then extracted. A threshold for information gain was determined to be $> \sim 0.08$ based on the entropy separation feature selection model (Chandrashekar and Sahin 2014; Alhaj et al. 2016).

The features most predictive of differential alternative splicing in ER+ cell lines are the average sequence conservation 50 nt upstream of the 3'SS and 50 nt downstream of the 5'SS (Figures 3.6, 3.7). Essentially, the phylogenetic conservation of the intronic sequence flanking the cassette exon of interest is the most important factor for either inclusion or exclusion of those differentially spliced exons. Unfortunately, the levels of information gain assigned to the sequence features for ER- cell lines are so low (< 0.08) that they cannot be trusted as reliable predictions.

The sequence feature analysis of differentially spliced events determined by MISO in each individual breast cancer cell line reveals new possible predictors of exon inclusion and exclusion. For example, the use of the splicing code on BT474 yields the same intronic sequence conservation that flanks the 3' and 5'SS at a higher level than that seen in the ER+/- cell lines for both the inclusion and exclusion of cassette exons (Figure 3.8). However, new features

Figure 3.6. Exon features determined by the splicing code to be important in cassette exon inclusion in estrogen receptor positive cell lines.

ER+ cassette exon inclusion

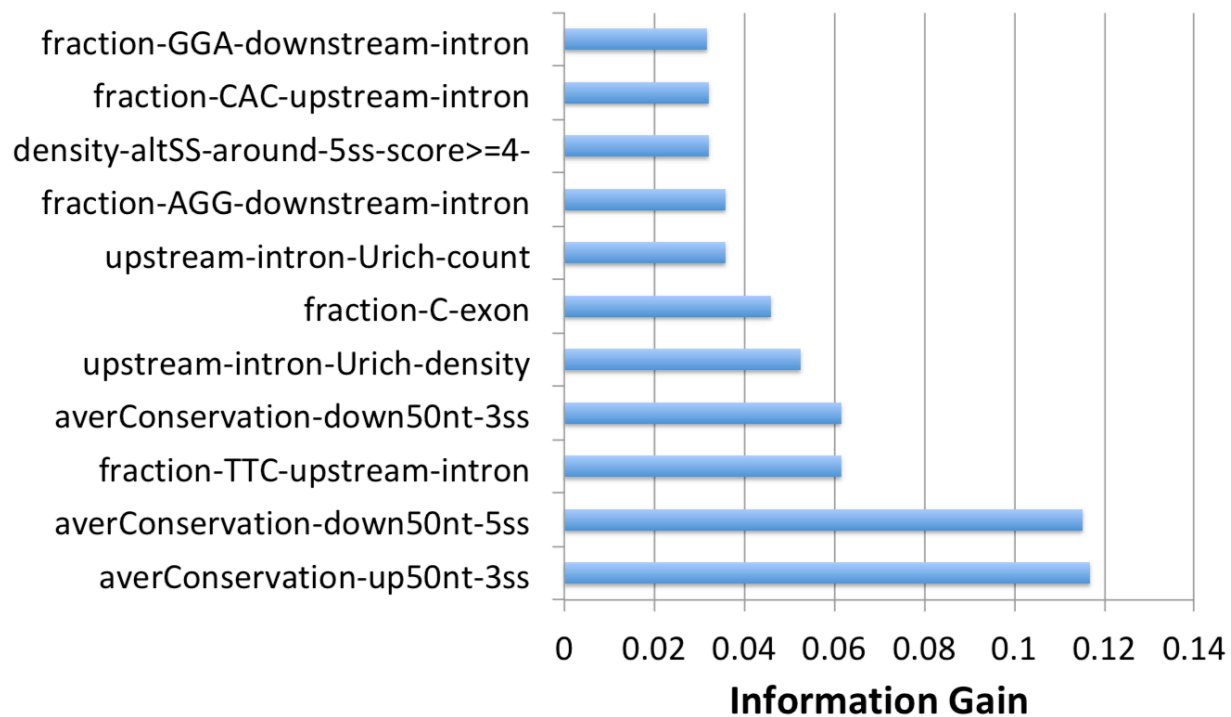


Figure 3.7. Exon features determined by the splicing code to be important in cassette exon exclusion in estrogen receptor positive cell lines.

ER+ cassette exon exclusion

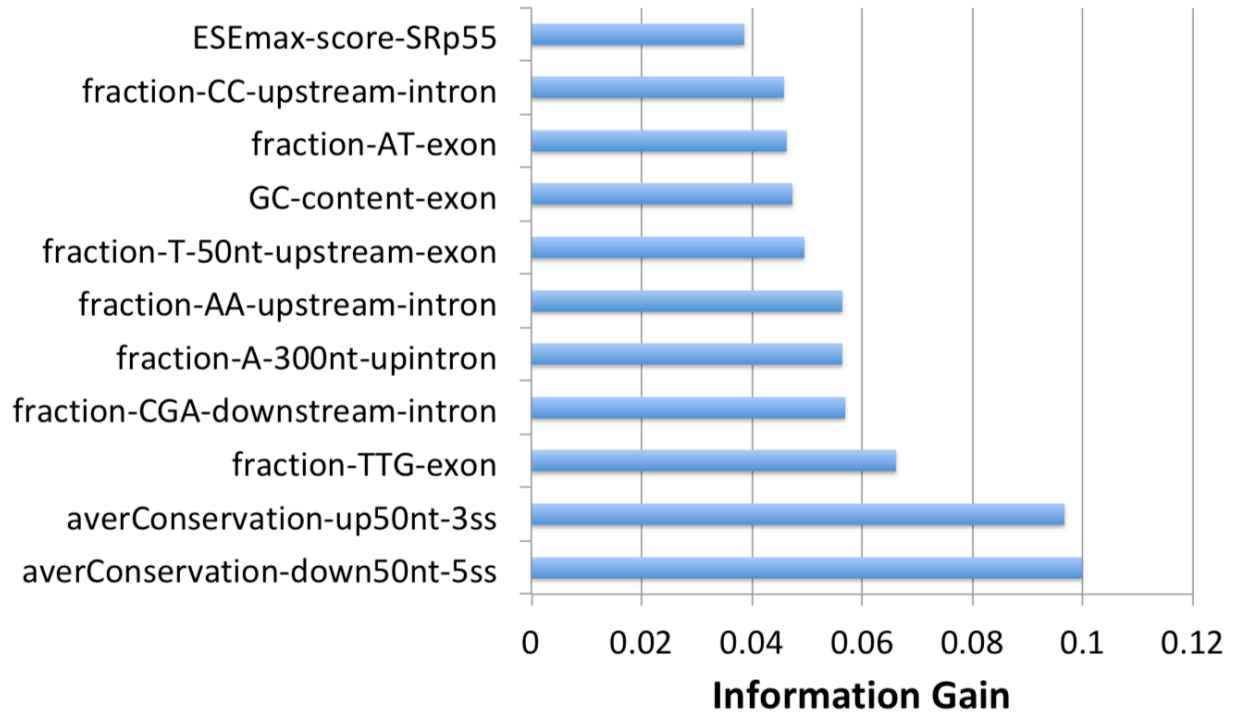
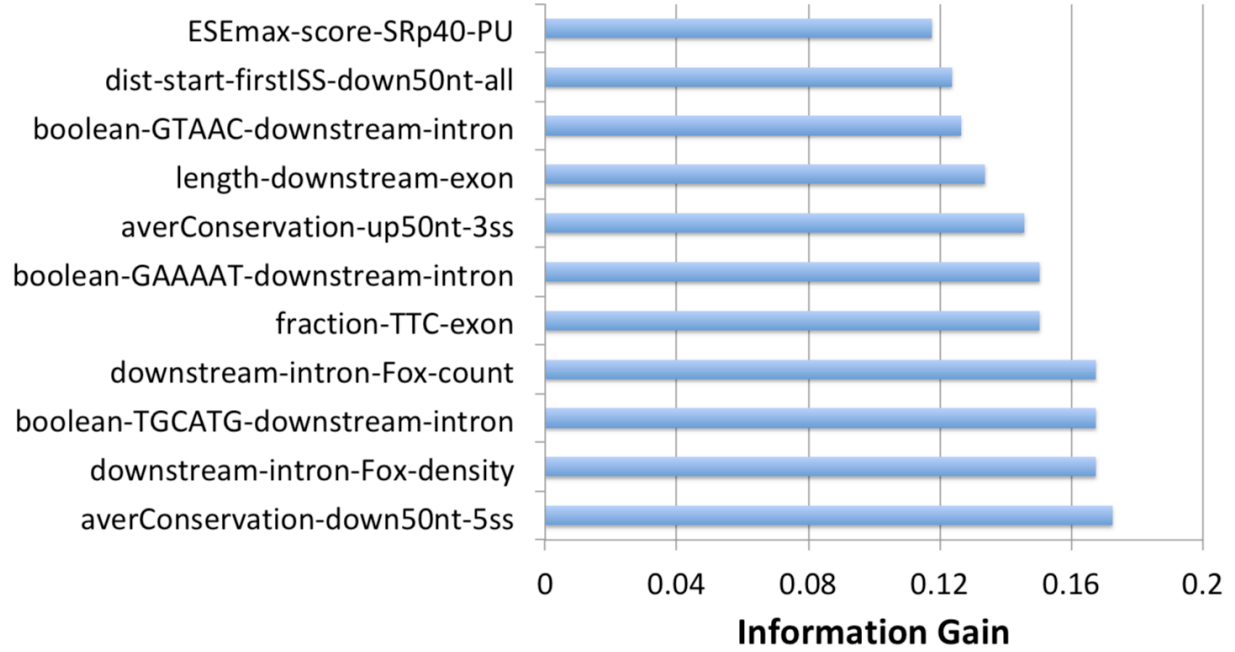


Figure 3.8. Exon features determined by the splicing code to be important in cassette exon inclusion in BT474.

BT474 cassette exon inclusion



are identified at levels of information gain that provide a higher confidence of their validity. For instance, the count and density binding motifs in the downstream intron are predicted to be important for inclusion of differentially spliced exon in the cell line BT474 (Figure 3.9). Interestingly, similar patterns are observed in each other cell line evaluated. The average sequence conservation of the 50 nt downstream of the 5'SS is identified as a top feature for the inclusion of cassette exons in all cell lines, while only BT20 cells do not classify average sequence conservation of the 50 nt upstream of the 3'SS as important for exon inclusion. These same features are only identified in BT474, ZR751 and MDAMB468 as important to exon exclusion. The drawback to these individual cell line analyses is that the increased information gains observed may be due to the smaller number of differentially expressed exons for each individual cell line.

By combining the results from the splicing code feature evaluation for each individual cell line, we were able to propose possible broad models for the regulation of exon inclusion and exclusion (Figure 3.10). The presence of exonic binding sites of splicing regulatory SR proteins *SRSF1* and *SRSF2* and some intronic and exonic trimeric sequences were found to be most important to promote cassette exon inclusion in breast cancer. The exonic binding site of *SRSF5*, exon length and phylogenetic conservation 50bp up- and downstream of the 3' and 5' splice sites were found to potentially participate in mediating the exclusion of cassette exons in breast cancer. These results demonstrate that it may be possible to determine splicing-specific features that figure prominently in breast cancer.

Figure 3.9. Exon features determined by the splicing code to be important in cassette exon exclusion in BT474.

BT474 cassette exon exclusion

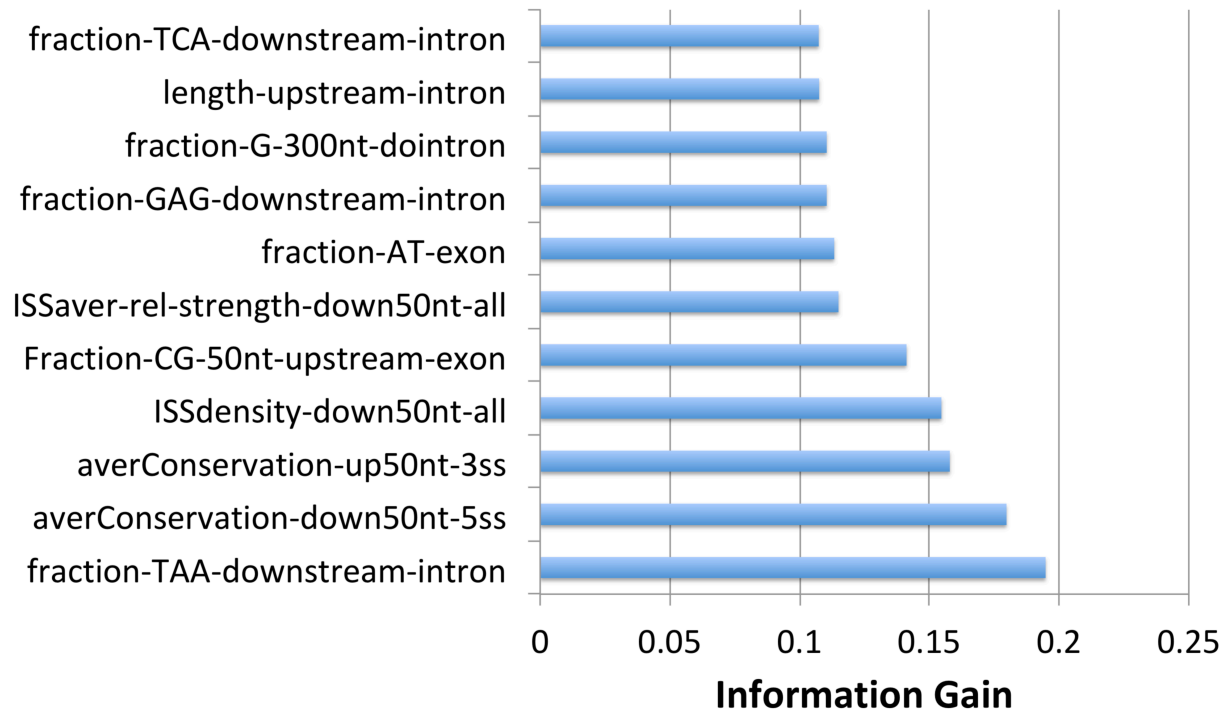
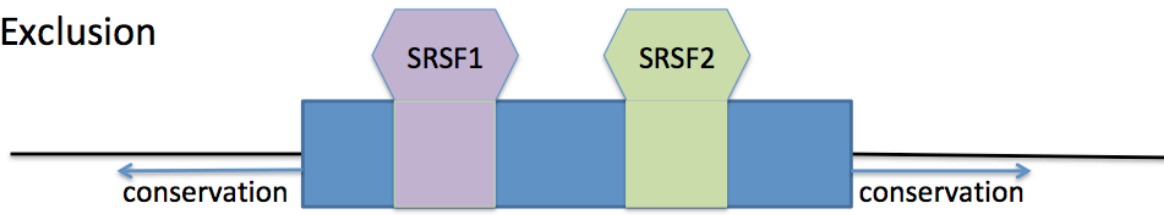
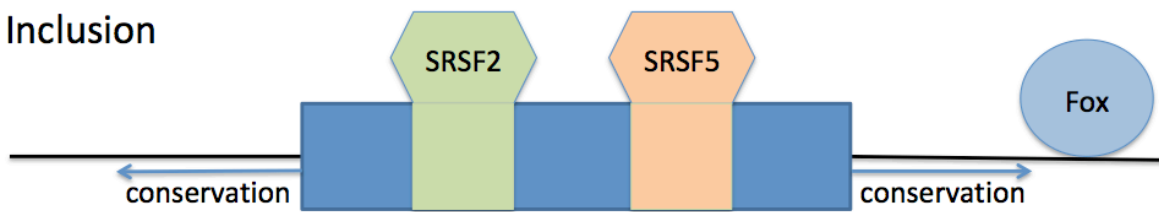


Figure 3.10. Composite model of exon features determined by the splicing code to be important in cassette exon inclusion and exclusion.

Exclusion



Inclusion



DISCUSSION

Here we present an mRNA sequencing-based study that reports the variations in gene expression and alternative splicing in ER+ and ER- breast cancer transcriptomes. Transcriptome analysis in ER+ and ER- cell lines both reveal a marked upward shift in gene expression in nearly all splicing-related genes, including spliceosome components and splicing regulators. Previous studies have linked individual splicing regulators, such as SR proteins and hnRNPs, to cancer (Karni et al. 2007; Lefave et al. 2011; Anczuków et al. 2012, 2015). Recently, an integrative genome-wide analysis revealed cooperative regulation of alternative splicing by hnRNP proteins (Huelga et al. 2012) and our results indicate a global network of splicing dysregulation exists in breast cancer that is not limited to a select few genes.

ESRP2 and *SF3B3* were the only splicing-related genes implicated on a gene expression level as significantly differentially expressed between the observed ER+ and ER- cell lines. *ESRP2* is an epithelial cell-type-specific regulator of *FGFR2* splicing which is responsible for EMT (Warzecha et al. 2009). *ESRP2* has also been shown to regulate the splicing of *CD44* and *CTNND1* (p120-Catenin) (Warzecha et al. 2009). Markers of EMT are strongly correlated with high-grade disease and low expression of ER/PR receptors (Willipinski-Stapelfeldt et al. 2005). Our data supports these findings in which the exon 13 of *CD44* and the exon 3 of *CTNND1* are differentially spliced in our ER+ cell lines (Table 3.4).

The significantly upregulated expression levels of *SF3B3* that we observed have previously been shown to correlate with prognosis and endocrine resistance in estrogen receptor-positive breast cancer treated with tamoxifen (Gökmen-Polar et al. 2015).

Additionally, alternative splicing of *EZH2* pre-mRNA by *SF3B3* has also been shown to contribute to the tumorigenic potential of renal cancer (Chen et al. 2017). This work outlines how the unique intertwined prospects of future breast cancer research in concert with mRNA splicing in the context of ER expression are observed in our analysis.

Many of the gene expression signatures and the differential splicing events identified by our study align with existing knowledge on breast cancer genetics. Among the genes with differential splicing events common to all breast cancer cell lines, many have previously been linked to cancer, either directly or indirectly (Table 3.6). However, many of these studies have not fully explored the relationship of alternative splicing in these genes to cancers. Additionally, there are several genes identified by our study that have no known function.

The observation of upregulated RNA processing-associated genes begs the question, is the upregulation of mRNA processing genes a result of some growth requirement of cancer, or is the overexpression of these transcripts a direct cause of the cancer? We hypothesize that the changes in alternative splicing, and the expression of splicing regulators and mRNA processing genes are a mixture of early splicing changes leading to a cancerous transformation, and later changes to maintain cancer phenotype. We currently have only evaluated broadly related breast cancer cell types divided solely by the presence of the ER. Many of these cell lines fall into additional breast cancer classifications. For instance MCF7 and T47D cell lines are Luminal A, with an ER+, progesterone receptor+/- (PR) and HER2- immunoprofile while BT474 and ZR751 are Luminal B with an ER+, progesterone receptor+/- (PR), and HER2+ (Holliday and

Speirs 2011). Subsequent transcriptome-wide splicing analysis could identify gene expression signatures and differentially spliced events that strongly correlate with HER2 or other breast cancer classifiers.

While RNA-binding MBNL proteins have been shown to regulate embryonic stem cell-specific alternative splicing (Han et al. 2013), they have been previously identified as being involved in mRNA export and stability and MBNL is a key player in the disease mechanism of myotonic dystrophy (Goers et al. 2010; Tran et al. 2011; Masuda et al. 2012; Konieczny et al. 2014; Sznajder et al. 2016; Timchenko 2013). Recently, the role of MBNL in cancer has been examined (Fish et al. 2016; Singh et al. 2018). Alternative splicing plays an important role in *MBNL1* as studies have highlighted exons 3 and 5 for its function and localization (Tran et al. 2011; Edge et al. 2013).

In our differential splicing analysis, exon 5 of *MBNL1* has a Ψ value of 0.97 for ER- breast cancer cell lines while MCF10A exhibits a Ψ of 0.13 for a prominent $\Delta\Psi$ of 0.84 (Table 3.5). A very recent study implicates exon 7 of *MBNL1* as being differentially included in prostate cancer cell lines and patient samples (Tabaglio et al. 2018). As in the study by Tabaglio et al. , the overall expression of *MBNL1* transcripts in our analyses (both ER+ and ER-) was downregulated (data not shown), consistent with its described role as a tumor suppressor (Sebestyén et al. 2016).

In light of the importance of different *MBNL1* isoforms, our observed exclusion of exon 4 in *MBNL1* in ER+ breast cancer cell lines may also be particularly worthy of further study. Our data show a Ψ value of 0.28 for ER+ breast cancer cell lines while MCF10A exhibits a Ψ of 0.95 for a striking $\Delta\Psi$ of -0.67 (Table 3.4). In summary, our

analyses agree with previous cancer-specific alternative splicing findings, but they also serve as a jumping point for new alternative splicing investigations in breast cancer.

Our use of a splicing code (Busch and Hertel 2015) for common splicing feature extraction provided us with a simple model of how observed differentially spliced cassette exon events in ER+ and ER- breast cancer cell lines can be broadly linked through a few key splicing determiners and regulators. We identified splicing regulators *SRSF1*, *SRSF2* and intronic splice site conservation as important features for cassette exon exclusion, while *SRSF2*, *SRSF5* and intronic splice site conservation as important features for cassette exon inclusion. Each of these splicing regulators were highly expressed in the breast cancer cell lines compared to MCF10A. But, the splicing code we used did not include all splicing regulators as potential features. It stands to reason that the inclusion of additional splicing regulators would yield more specific splicing regulator combinations responsible for differential splicing events in breast cancer. However, the differential expression of spliceosomal components and regulatory factors coupled with the splicing error analysis experiments (Figure 3.1) suggests that differential alternative splicing in breast cancer is a result of a dysregulation of splicing, and not altered splicing fidelity.

MATERIALS AND METHODS

Splicing Fidelity Assay

All splicing fidelity assays were performed as previously described (Fox-Walsh and Hertel 2009).

mRNA-seq data acquisition

mRNA-seq reads were obtained from the Short Read Archive section of Gene Expression Omnibus (GEO) at the NCBI, and were previously used to analyze gene expression, CpG island methylation and gene copy number in breast cancer (Wang et al. 2008; Sun et al. 2011). Sequence read data for the were obtained from the Short Read Archive section of GEO at NCBI under accession numbers GSE12946 and SRA002355.1. Sequencing analysis was performed by the UCI Genomics High Throughput Facility using ELAND.

Alignment of short reads to genome and transcriptome.

We mapped the datasets to the human genome (GRCh37/hg19) using Tophat2 and Bowtie2 using the default parameters except the Sun et al. dataset where the paired-end reads option was utilized. The data sets were normalized with in-house bioinformatics tools using the Fragments Per Kilobase of exon model per Million mapped reads (FPKM) methodology (Mortazavi et al. 2008; Trapnell et al. 2010). FPKM values were compared between normal breast and breast cancer cell lines to measure differential expression.

Detection of alternatively spliced exons.

We used the mixture of isoforms (MISO) framework (Katz et al. 2010). We identified alternative splicing events using Version 2 of the human hg19 annotations (compiled June 2013). Splicing events evaluated include the inclusion/exclusion of skipped exons, alternative 3' and alternative 5' splice sites, retained introns, alternative first exons, alternative last exons, and tandem UTRs. We filtered the results file to find only events with: (a) the sum of inclusion and exclusion reads is at least 10, (b) the $\Delta \Psi$ is at least 0.20, (c) the Bayes factor is at least 10 and (a)-(c) are true in both of the samples.

Feature analysis of differentially spliced events

We applied a machine learning pipeline based on an implementation of a SVM in WEKA (version 3.6.2) as outlined by Busch and Hertel (2015) to our differentially spliced events data from MISO. We determined that an information gain value of ~ 0.08 or less constituted background noise, where no individual feature had more appreciable information gain than any random feature (Chandrashekar and Sahin 2014; Alhaj et al. 2016)

Creation of training sets

We created 2 training sets of internal exons with known splicing behavior (constitutive and cassette) to train the SVM as previously described (Busch and Hertel 2015). The training sets of exons were obtained through strict filtering of the UCSC Genome Browser (GRCh37/hg19) (Meyer et al. 2013) by HEXEvent (Busch and Hertel 2013). The set of constitutive exons includes internal human exons that are not involved in any

type of alternative splicing and that are supported by at least 20 ESTs. Cassette exons in our sets are internal exons that show only one type of alternative splicing. All exons have a minimal length of 23 nt and a minimal length of neighboring introns of 78 nt.

Feature Extraction

A total of 262 (1072) sequence features were analyzed on all training set exons and the differentially expressed exon sets for individual cell lines and ER+/- cell line groups. These features included splice site strength, exon/intron architecture, local secondary structures and splicing regulator binding sites as previously outlined (Busch and Hertel 2015). An additional 810 sequence features including: motif clusters (Yeo et al. 2007), additional sequence features (Barash et al. 2010) and all possible 2-mers and 3-mers were also evaluated.

REFERENCES

- Abe O, Abe R, Enomoto K, Kikuchi K, Koyama H, Masuda H, Nomura Y, Sakai K, Sugimachi K, Tominaga T, et al. 2005. Effects of radiotherapy and of differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: An overview of the randomised trials. *Lancet*.
- Alhaj TA, Siraj MM, Zainal A, Elshoush HT, Elhaj F. 2016. Feature Selection Using Information Gain for Improved Structural-Based Alert Correlation. *PLoS One* **11**: e0166017. <http://www.ncbi.nlm.nih.gov/pubmed/27893821> (Accessed November 22, 2018).
- Anczuków O, Akerman M, Cléry A, Wu J, Shen C, Shirole NH, Raimer A, Sun S, Jensen MA, Hua Y, et al. 2015. SRSF1-Regulated Alternative Splicing in Breast Cancer. *Mol Cell* **60**: 105–117. <http://linkinghub.elsevier.com/retrieve/pii/S1097276515007017>.
- Anczuków O, Rosenberg AZ, Akerman M, Das S, Zhan L, Karni R, Muthuswamy SK, Krainer AR. 2012. The splicing factor SRSF1 regulates apoptosis and proliferation to promote mammary epithelial cell transformation. *Nat Struct Mol Biol* **19**: 220–8. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3272117&tool=pmcentrez&rendertype=abstract> (Accessed March 9, 2012).
- Baig RM, Sanders AJ, Kayani MA, Jiang WG. 2012. Association of Differentiation-Related Gene-1 (DRG1) with Breast Cancer Survival and in Vitro Impact of DRG1 Suppression. *Cancers (Basel)* **4**: 658–672. <http://www.mdpi.com/2072-6694/4/3/658/> (Accessed September 7, 2013).
- Barash Y, Calarco J a, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ. 2010. Deciphering the splicing code. *Nature* **465**: 53–9. <http://www.ncbi.nlm.nih.gov/pubmed/20445623> (Accessed May 21, 2013).
- Brown R, Reinke L. 2011. CD44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression. *J Clin ...* **121**. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049398/> (Accessed September 9, 2013).
- Busch A, Hertel KJ. 2013. HEXEvent: A database of human EXon splicing Events. *Nucleic Acids Res* **41**.
- Busch A, Hertel KJ. 2015. Splicing predictions reliably classify different types of alternative splicing. *RNA* **21**: 813–823. <http://rnajournal.cshlp.org/content/early/2015/03/24/rna.048769.114>.
- Capo-chichi CD, Cai KQ, Testa JR, Godwin AK, Xu X-X. 2009. Loss of GATA6 leads to nuclear deformation and aneuploidy in ovarian cancer. *Mol Cell Biol* **29**: 4766–77. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2725711&tool=pmcentrez&rendertype=abstract> (Accessed September 7, 2013).
- Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S. 2009. AmiGO: online access to ontology and annotation data. *Bioinformatics* **25**: 288–9. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2639003&tool=pmcentrez&rendertype=abstract> (Accessed March 6, 2013).

- Chandrashekar G, Sahin F. 2014. A survey on feature selection methods. *Comput Electr Eng* **40**: 16–28.
<https://www.sciencedirect.com/science/article/pii/S0045790613003066?via%3Dihub> (Accessed November 22, 2018).
- Chen H, Wu X, Pan ZK, Huang S. 2010. Integrity of SOS1/EPS8/ABI1 tri-complex determines ovarian cancer metastasis. *Cancer Res* **70**: 9979–90.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3059077&tool=pmcentrez&rendertype=abstract> (Accessed August 13, 2013).
- Chen K, Xiao H, Zeng J, Yu G, Zhou H, Huang C, Yao W, Xiao W, Hu J, Guan W, et al. 2017. Alternative Splicing of EZH2 pre-mRNA by SF3B3 Contributes to the Tumorigenic Potential of Renal Cancer. *Clin Cancer Res* **23**: 3428–3441.
<http://clincancerres.aacrjournals.org/lookup/doi/10.1158/1078-0432.CCR-16-2020>.
- Detre SI, Ashley S, Mohammed K, Smith IE, Powles TJ, Dowsett M. 2017. Immunohistochemical Phenotype of Breast Cancer during 25-Year Follow-up of the Royal Marsden Tamoxifen Prevention Trial. *Cancer Prev Res* **10**: 171–176.
<http://cancerpreventionresearch.aacrjournals.org/lookup/doi/10.1158/1940-6207.CAPR-16-0247-T>.
- Edge C, Gooding C, Smith CWJ. 2013. Dissecting domains necessary for activation and repression of splicing by muscleblind-like protein 1. *BMC Mol Biol* **14**: 29.
<http://bmcmolbiol.biomedcentral.com/articles/10.1186/1471-2199-14-29>.
- Fish L, Pencheva N, Goodarzi H, Tran H, Yoshida M, Tavazoie SF. 2016. Muscleblind-like 1 suppresses breast cancer metastatic colonization and stabilizes metastasis suppressor transcripts. *Genes Dev* **30**: 386–398.
<http://genesdev.cshlp.org/lookup/doi/10.1101/gad.270645.115>.
- Fisher B, Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM, Vogel V, Robidoux A, Dimitrov N, Atkins J, et al. 1998. Tamoxifen for Prevention of Breast Cancer: Report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *JNCI J Natl Cancer Inst* **90**: 1371–1388. <http://breast-cancer-research.biomedcentral.com/articles/10.1186/bcr2889>.
- Fox-Walsh KL, Hertel KJ. 2009. Splice-site pairing is an intrinsically high fidelity process. *Proc Natl Acad Sci* **106**: 1766–1771.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2644112&tool=pmcentrez&rendertype=abstract>.
- Ghigna C, Moroni M, Porta C, Adenocarcinomas HC, Porta G, Riva S, Biamonti G. 1998. Altered Expression of Heterogeneous Nuclear Ribonucleoproteins and SR Factors in Human Colon Adenocarcinomas Altered Expression of Heterogeneous Nuclear Ribonucleoproteins and SR Factors in. 5818–5824.
- Goers ES, Purcell J, Voelker RB, Gates DP, Berglund JA. 2010. MBNL1 binds GC motifs embedded in pyrimidines to regulate alternative splicing. *Nucleic Acids Res* **38**: 2467–2484. <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkp1209>.
- Gökmen-Polar Y, Neelamraju Y, Goswami CP, Gu X, Nallamothe G, Janga SC, Badve S. 2015. Expression levels of SF3B3 correlate with prognosis and endocrine

- resistance in estrogen receptor-positive breast cancer. *Mod Pathol* **28**: 677–685. <http://www.nature.com/articles/modpathol2014146>.
- Goto T, Shinmura K, Yokomizo K, Sakuraba K, Kitamura Y, Shirahata A, Saito M, Kigawa G, Nemoto H, Sanada Y, et al. 2012. Expression levels of thymidylate synthase, dihydropyrimidine dehydrogenase, and thymidine phosphorylase in patients with colorectal cancer. *Anticancer Res* **32**: 1757–62. <http://www.ncbi.nlm.nih.gov/pubmed/22593457>.
- Grosso AR, Martins S, Carmo-Fonseca M. 2008. The emerging role of splicing factors in cancer. *EMBO Rep* **9**: 1087–93. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2581861&tool=pmcentrez&rendertype=abstract> (Accessed August 15, 2011).
- Gupta GP, Massagué J. 2006. Cancer Metastasis: Building a Framework. *Cell*.
- Han H, Irimia M, Ross PJ, Sung H-K, Alipanahi B, David L, Golipour A, Gabut M, Michael IP, Nachman EN, et al. 2013. MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* 5–12. <http://www.nature.com/doi/10.1038/nature12270> (Accessed June 5, 2013).
- Han SP, Tang YH, Smith R. 2010. Functional diversity of the hnRNPs: past, present and perspectives. *Biochem J* **430**: 379–92. <http://www.ncbi.nlm.nih.gov/pubmed/20795951> (Accessed June 29, 2011).
- Hanahan D, Weinberg RA. 2011. Hallmarks of cancer: the next generation. *Cell* **144**: 646–74. <https://doi.org/10.1016/j.cell.2011.02.013>.
- Henrich K-O, Schwab M, Westermann F. 2012. 1P36 Tumor Suppression--a Matter of Dosage? *Cancer Res* **72**: 6079–88. <http://www.ncbi.nlm.nih.gov/pubmed/23172308> (Accessed September 7, 2013).
- Holliday DL, Speirs V. 2011. Choosing the right cell line for breast cancer research. *Breast Cancer Res* **13**: 215. <http://breast-cancer-research.biomedcentral.com/articles/10.1186/bcr2889>.
- Huelga SC, Vu AQ, Arnold JD, Liang TY, Liu PP, Yan BY, Donohue JP, Shiue L, Hoon S, Brenner S, et al. 2012. Integrative Genome-wide Analysis Reveals Cooperative Regulation of Alternative Splicing by hnRNP Proteins. *Cell Rep* **1**: 167–178. <http://linkinghub.elsevier.com/retrieve/pii/S2211124712000435> (Accessed February 17, 2012).
- Iorns E, Ward TM, Dean S, Jegg A, Thomas D, Murugaesu N, Sims D, Mitsopoulos C, Fenwick K, Kozarewa I, et al. 2012. Whole genome in vivo RNAi screening identifies the leukemia inhibitory factor receptor as a novel breast tumor suppressor. *Breast Cancer Res Treat* **135**: 79–91. <http://www.ncbi.nlm.nih.gov/pubmed/22535017> (Accessed September 7, 2013).
- Jemal A, Siegel R, Xu J, Ward E. 2010. Cancer Statistics, 2010. *CA Cancer J Clin* **60**: 277–300. <http://dx.doi.org/10.3322/caac.20073>.
- Jia D, Jolly MK, Boareto M, Parsana P, Mooney SM, Pienta KJ, Levine H, Ben-Jacob E. 2015. OVOL guides the epithelial-hybrid-mesenchymal transition. *Oncotarget* **6**: 15436–15448. <http://www.oncotarget.com/fulltext/3623>.

- Johnson E, Seachrist DD, DeLeon-Rodriguez CM, Lozada KL, Miedler J, Abdul-Karim FW, Keri R a. 2010. HER2/ErbB2-induced breast cancer cell migration and invasion require p120 catenin activation of Rac1 and Cdc42. *J Biol Chem* **285**: 29491–501.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2937981&tool=pmcentrez&rendertype=abstract> (Accessed August 11, 2013).
- Karni R, de Stanchina E, Lowe SW, Sinha R, Mu D, Krainer AR. 2007. The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nat Struct Mol Biol* **14**: 185–93. <http://www.ncbi.nlm.nih.gov/pubmed/17310252> (Accessed July 12, 2012).
- Katz Y, Wang ET, Airoidi EM, Burge CB. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**: 1009–15.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3037023&tool=pmcentrez&rendertype=abstract> (Accessed May 23, 2013).
- Kim S, You S, Hwang D. 2011. Aminoacyl-tRNA synthetases and tumorigenesis: more than housekeeping. *Nat Rev Cancer* **11**: 708–18.
<http://www.ncbi.nlm.nih.gov/pubmed/21941282> (Accessed August 19, 2013).
- Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, Fulton LL, Dooling DJ, Ding L, Mardis ER, et al. 2012. Comprehensive molecular portraits of human breast tumours. *Nature* **490**: 61–70.
<http://www.nature.com/doi/10.1038/nature11412>.
- Konieczny P, Stepniak-Konieczna E, Sobczak K. 2014. MBNL proteins and their target RNAs, interaction and splicing regulation. *Nucleic Acids Res*.
- Kubota D, Okubo T, Saito T, Suehara Y, Yoshida A, Kikuta K, Tsuda H, Katai H, Shimada Y, Kaneko K, et al. 2012. Validation study on pftin and ATP-dependent RNA helicase DDX39 as prognostic biomarkers in gastrointestinal stromal tumour. *Jpn J Clin Oncol* **42**: 730–41. <http://www.ncbi.nlm.nih.gov/pubmed/22723667> (Accessed September 7, 2013).
- Lefave C V, Squatrito M, Vorlova S, Rocco GL, Brennan CW, Holland EC, Pan Y-X, Cartegni L. 2011. Splicing factor hnRNPH drives an oncogenic splicing switch in gliomas. *EMBO J* **30**: 4084–97.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3209773&tool=pmcentrez&rendertype=abstract> (Accessed March 12, 2012).
- Li J, Gu L, Zhang H, Liu T, Tian D, Zhou M, Zhou S. 2013. Berberine represses DAXX gene transcription and induces cancer cell apoptosis. *Lab Invest* **93**: 354–64.
<http://www.ncbi.nlm.nih.gov/pubmed/23295648> (Accessed September 7, 2013).
- Liu R, Zhang Z, Xu Y. 2010. Downregulation of nucleostemin causes G1 cell cycle arrest via a p53-independent pathway in prostate cancer PC-3 cells. *Urol Int* **85**: 221–7. <http://www.ncbi.nlm.nih.gov/pubmed/20664182> (Accessed September 7, 2013).
- Long JC, Caceres JF. 2009. The SR protein family of splicing factors: master regulators of gene expression. *Biochem J* **417**: 15–27.
<http://www.ncbi.nlm.nih.gov/pubmed/19061484> (Accessed June 17, 2011).

- Masuda A, Andersen HS, Doktor TK, Okamoto T, Ito M, Andresen BS, Ohno K. 2012. CUGBP1 and MBNL1 preferentially bind to 3' UTRs and facilitate mRNA decay. *Sci Rep*.
- Melchor L, Saucedo-Cuevas LP, Muñoz-Repeto I, Rodríguez-Pinilla SM, Honrado E, Campoverde A, Palacios J, Nathanson KL, García MJ, Benítez J. 2009. Comprehensive characterization of the DNA amplification at 13q34 in human breast cancer reveals TFDP1 and CUL4A as likely candidate target genes. *Breast Cancer Res* **11**: R86. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2815550&tool=pmcentrez&rendertype=abstract> (Accessed September 7, 2013).
- Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, et al. 2013. The UCSC Genome Browser database: Extensions and updates 2013. *Nucleic Acids Res*.
- Morrow PKH, Hortobagyi GN. 2009. Management of breast cancer in the genome era. *Annu Rev Med* **60**: 153–65. <http://www.ncbi.nlm.nih.gov/pubmed/19630569> (Accessed March 5, 2013).
- Mortazavi A, Williams BA, Mccue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 1–8.
- Muraoka S, Kume H, Watanabe S, Adachi J, Kuwano M, Sato M, Kawasaki N, Kodera Y, Ishitobi M, Inaji H, et al. 2012. Strategy for SRM-based Verification of Biomarker Candidates Discovered by iTRAQ Method in Limited Breast Cancer Tissue Samples. *J Proteome Res* **11**: 4201–4210. <http://dx.doi.org/10.1021/pr300322q>.
- Mustacchi G, Sormani MP, Bruzzi P, Gennari A, Zanconati F, Bonifacio D, Monzoni A, Morandi L. 2013. Identification and validation of a new set of five genes for prediction of risk in early breast cancer. *Int J Mol Sci* **14**: 9686–702. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3676806&tool=pmcentrez&rendertype=abstract> (Accessed August 23, 2013).
- Nouh M a, Mohamed MM, El-Shinawi M, Shaalan M a, Cavallo-Medved D, Khaled HM, Sloane BF. 2011. Cathepsin B: a potential prognostic marker for inflammatory breast cancer. *J Transl Med* **9**: 1. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3022726&tool=pmcentrez&rendertype=abstract> (Accessed September 5, 2013).
- Olson JE, Wang X, Pankratz VS, Fredericksen ZS, Vachon CM, Vierkant R a, Cerhan JR, Couch FJ. 2011. Centrosome-related genes, genetic variation, and risk of breast cancer. *Breast Cancer Res Treat* **125**: 221–8. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2997159&tool=pmcentrez&rendertype=abstract> (Accessed September 5, 2013).
- Oltean S, Bates DO. 2014. Hallmarks of alternative splicing in cancer. *Oncogene* **33**: 5311–5318. <http://dx.doi.org/10.1038/onc.2013.533>.
- Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**: 1413–5. <http://www.ncbi.nlm.nih.gov/pubmed/18978789> (Accessed July 17, 2012).

- Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, et al. 2009. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *J Clin Oncol* **27**: 1160–1167. <http://ascopubs.org/doi/10.1200/JCO.2008.18.1370>.
- Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees C a, Pollack JR, Ross DT, Johnsen H, Akslen L a, et al. 2000. Molecular portraits of human breast tumours. *Nature* **406**: 747–752. <http://www.ncbi.nlm.nih.gov/pubmed/23000897>.
- Pont AR, Sadri N, Hsiao SJ, Smith S, Schneider RJ. 2012. mRNA decay factor AUF1 maintains normal aging, telomere maintenance, and suppression of senescence by activation of telomerase transcription. *Mol Cell* **47**: 5–15. <http://www.ncbi.nlm.nih.gov/pubmed/22633954> (Accessed August 9, 2013).
- Ramirez AB, Loch CM, Zhang Y, Liu Y, Wang X, Wayner E a, Sargent JE, Sibani S, Hainsworth E, Mendoza E a, et al. 2010. Use of a single-chain antibody library for ovarian cancer biomarker discovery. *Mol Cell Proteomics* **9**: 1449–60. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2938096&tool=pmcentrez&rendertype=abstract> (Accessed September 7, 2013).
- Samanta S, Sharma VM, Khan a, Mercurio a M. 2012. Regulation of IMP3 by EGFR signaling and repression by ER β : implications for triple-negative breast cancer. *Oncogene* **31**: 4689–97. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3337950&tool=pmcentrez&rendertype=abstract> (Accessed September 7, 2013).
- Sebestyén E, Singh B, Miñana B, Pagès A, Mateo F, Pujana MA, Valcárcel J, Eyras E. 2016. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Res* **26**: 732–744. <http://genome.cshlp.org/lookup/doi/10.1101/gr.199935.115>.
- Shain AH, Pollack JR. 2013. The spectrum of SWI/SNF mutations, ubiquitous in human cancers. *PLoS One* **8**: e55119. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3552954&tool=pmcentrez&rendertype=abstract> (Accessed August 12, 2013).
- Shapiro IM, Cheng AW, Flytzanis NC, Balsamo M, Condeelis JS, Oktay MH, Burge CB, Gertler FB. 2011. An EMT-Driven Alternative Splicing Program Occurs in Human Breast Cancer and Modulates Cellular Phenotype. *PLoS Genet* **7**: e1002218. <http://www.ncbi.nlm.nih.gov/pubmed/21876675> (Accessed August 31, 2011).
- Shi C-S, Huang N-N, Kehrl JH. 2012. Regulator of G-protein signaling 3 isoform 1 (PDZ-RGS3) enhances canonical Wnt signaling and promotes epithelial mesenchymal transition. *J Biol Chem* **287**: 33480–7. <http://www.ncbi.nlm.nih.gov/pubmed/22859293> (Accessed August 16, 2013).
- Singh B, Trincado JL, Tatlow P, Piccolo SR, Eyras E. 2018. Genome Sequencing and RNA-Motif Analysis Reveal Novel Damaging Noncoding Mutations in Human Tumors. *Mol Cancer Res* **16**: 1112–1124. <http://mcr.aacrjournals.org/lookup/doi/10.1158/1541-7786.MCR-17-0601>.
- Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. 2001. Gene expression patterns of breast carcinomas

- distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98**: 10869–74.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=58566&tool=pmcentrez&rendertype=abstract>.
- Sørli T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, et al. 2003. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci* **100**: 8418–8423.
<http://www.pnas.org/cgi/doi/10.1073/pnas.0932692100>.
- Sun Z, Asmann YW, Kalari KR, Bot B, Eckel-Passow JE, Baker TR, Carr JM, Khrebtukova I, Luo S, Zhang L, et al. 2011. Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing. *PLoS One* **6**: e17490.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3045451&tool=pmcentrez&rendertype=abstract> (Accessed March 13, 2013).
- Sveen A, Kilpinen S, Ruusulehto A, Lothe RA, Skotheim RI. 2016. Aberrant RNA splicing in cancer; Expression changes and driver mutations of splicing factor genes. *Oncogene* **35**: 2413–2427. <http://dx.doi.org/10.1038/onc.2015.318>.
- Sznajder ŁJ, Michalak M, Taylor K, Cywoniuk P, Kabza M, Wojtkowiak-Szlachcic A, Matłoka M, Konieczny P, Sobczak K. 2016. Mechanistic determinants of MBNL activity. *Nucleic Acids Res* **44**: 10326–10342. <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw915>.
- Tabaglio T, Low DH, Teo WKL, Goy PA, Cywoniuk P, Wollmann H, Ho J, Tan D, Aw J, Pavesi A, et al. 2018. MBNL1 alternative splicing isoforms play opposing roles in cancer. *Life Sci Alliance* **1**: e201800157. <http://www.life-science-alliance.org/lookup/doi/10.26508/lsa.201800157>.
- Timchenko L. 2013. Molecular mechanisms of muscle atrophy in myotonic dystrophies. *Int J Biochem Cell Biol*.
- Tockman MS, Mulshine JL, Piantadosi S, Erozan YS, Gupta PK, Ruckdeschel JC, Taylor PR, Zhukov T, Zhou WH, Qiao YL, et al. 1997. Prospective detection of preclinical lung cancer: results from two studies of heterogeneous nuclear ribonucleoprotein A2/B1 overexpression. *Clin Cancer Res* **3**: 2237–2246.
<http://clincancerres.aacrjournals.org/content/3/12/2237.abstract>.
- Tran H, Gourrier N, Lemercier-Neuillet C, Dhaenens CM, Vautrin A, Fernandez-Gomez FJ, Arandel L, Carpentier C, Obriot H, Eddarkaoui S, et al. 2011. Analysis of exonic regions involved in nuclear localization, splicing activity, and dimerization of muscleblind-like-1 isoforms. *J Biol Chem*.
- Trapnell C, Williams B a, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511–5. <http://www.ncbi.nlm.nih.gov/pubmed/20436464> (Accessed July 5, 2011).
- Venables JP, Brosseau J-P, Gadea G, Klinck R, Prinos P, Beaulieu J-F, Lapointe E, Durand M, Thibault P, Tremblay K, et al. 2013. RBFOX2 is an important regulator

- of mesenchymal tissue-specific splicing in both normal and cancer tissues. *Mol Cell Biol* **33**: 396–405.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3554129&tool=pmcentrez&rendertype=abstract> (Accessed August 9, 2013).
- Venables JP, Klinck R, Bramard A, Inkel L, Dufresne-Martin G, Koh C, Gervais-Bird J, Lapointe E, Froehlich U, Durand M, et al. 2008. Identification of alternative splicing markers for breast cancer. *Cancer Res* **68**: 9525–31.
<http://www.ncbi.nlm.nih.gov/pubmed/19010929> (Accessed August 7, 2011).
- Wakasugi K, Slike BM, Hood J, Otani A, Ewalt KL, Friedlander M, Cheresh D a, Schimmel P. 2002. A human aminoacyl-tRNA synthetase as a regulator of angiogenesis. *Proc Natl Acad Sci U S A* **99**: 173–7.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=117534&tool=pmcentrez&rendertype=abstract>.
- Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–6.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2593745&tool=pmcentrez&rendertype=abstract> (Accessed June 14, 2011).
- Warzecha CC, Jiang P, Amirikian K, Dittmar K a, Lu H, Shen S, Guo W, Xing Y, Carstens RP. 2010. An ESRP-regulated splicing programme is abrogated during the epithelial-mesenchymal transition. *EMBO J* **29**: 3286–300.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2957203&tool=pmcentrez&rendertype=abstract> (Accessed June 10, 2011).
- Warzecha CC, Sato TK, Nabet B, Hogenesch JB, Carstens RP. 2009. ESRP1 and ESRP2 Are Epithelial Cell-Type-Specific Regulators of FGFR2 Splicing. *Mol Cell* **33**: 591–601.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2702247&tool=pmcentrez&rendertype=abstract> (Accessed July 27, 2011).
- Wen J, Toomer KH, Chen Z, Cai X. 2015. Genome-wide analysis of alternative transcripts in human breast cancer. *Breast Cancer Res Treat*.
<http://www.ncbi.nlm.nih.gov/pubmed/25913416> (Accessed April 29, 2015).
- Willipinski-Stapelfeldt B, Riethdorf S, Assmann V, Woelfle U, Rau T, Sauter G, Heukeshoven J, Pantel K. 2005. Changes in cytoskeletal protein composition indicative of an epithelial-mesenchymal transition in human micrometastatic and primary breast carcinoma cells. *Clin Cancer Res* **11**: 8006–8014.
- Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, et al. 2014. The human splicing code reveals new insights into the genetic determinants of disease. *Science (80-)* **1254806**. <http://www.sciencemag.org/content/347/6218/1254806.full>.
- Xiong HY, Barash Y, Frey BJ. 2011. Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics* **27**: 2554–62.
<http://www.ncbi.nlm.nih.gov/pubmed/21803804>.
- Yeo GW, Van Nostrand EL, Liang TY. 2007. Discovery and analysis of evolutionarily

conserved intronic splicing regulatory elements. *PLoS Genet.*

Yuan J, Luo K, Zhang L, Cheville JC, Lou Z. 2010. USP10 regulates p53 localization and stability by deubiquitinating p53. *Cell* **140**: 384–96.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2820153&tool=pmcentrez&rendertype=abstract> (Accessed August 26, 2013).

CHAPTER 4

PERSPECTIVES

Pre-mRNA splicing is perhaps the single greatest factor in the development and complexity of the transcriptome and by extension the proteome. What was once considered noise or an artifact is now firmly entrenched in all aspects of gene expression and human health. There is a correlation between the amount of alternative splicing and species complexity, as the majority of multi-exon genes in humans exhibit this capability (Chen et al. 2014). The sheer number of potential alternative spliced transcripts raises important questions, such as how much of observed alternative splicing is intended or biologically relevant? Can alternative splicing be used to decipher the health of a cell or organism on a molecular level? In order to answer these questions, the first problems that must be solved are the types and rates of erroneous splicing.

Splicing Fidelity is High But Fragile

Previous research investigating the fidelity of splicing estimate the error rate to be somewhere between 1 in 100 and 1 in 100,000 (Fox-Walsh and Hertel 2009; Pickrell et al. 2010). Our results place splicing errors rates from 1 error in 5479 wild-type spliced transcripts at the 3'SS of intron 6 to 1 error in 131,611 wild-type spliced transcripts at the 5'SS of intron 6. Likewise, the 5'SS of intron 7 (1 error in 5750 wild-type spliced transcripts) is much less error prone than the 3'SS (1 error in 631 wild-type spliced transcripts). Within the context of this limited investigation, this endorses the current thought that in general the selection of the 3'SS is more vulnerable to error due to the

extra complexity and degeneracy of 3'SSs (Vořechovský 2006; Buratti et al. 2007; Stepankiw et al. 2015). The measured splicing error rates span a wide range, whereas the fidelity of other biological processes like transcription and translation are much more static. However, the abundance of competing signals that regulate splicing and the incredible diversity and complexity of the sequence variables justifies this observed range in splicing error rates. As mentioned earlier and throughout this dissertation, devoid of a defective piece of the splicing machinery, splicing is highly sequence dependent.

The unexpected contribution to our dataset was the errors introduced during the construction of the library and the synonymous mutations within exon 7. These mutations, both purposeful and accidental, show how single nucleotide changes can completely alter the spliceosomal splice site selection. The most obvious example was the A→T mutation that we determined to occur at position 27 in the middle of exon 7 that created a 5' splice site with a high splice site strength score. Based on the number of 27A→T mutant DNA input reads (150,992 reads) and the Inclusion Index Values normalized to wild-type exon 7 as described by Mueller et al., we would have expected approximately 328,742 RNA output reads with the 27A→T mutation. Instead we only observed 1,333 27A→T exon 7 mutant reads in the RNA output. Yet, we observed 309,793 RNA output reads that were spliced at position 25 in exon 7, appearing to have used this cryptic 5' splice site created by the 27A→T mutation. If we compare the expected RNA output reads (1,333) with the observed RNA output reads (~309,793), it becomes clear that the splicing efficiency for this cryptic 5' splice site is ~99.5, or that the canonical 5' splice site is used only 1 in ~232 times in 27A→T mutants.

It is remarkable to observe how a single nucleotide variant (SNV) can have such an impact on splice site selection of an important gene such as *SMN1*.

When the 5' splice site is completely complementary to the U1 snRNA with no cis-acting splicing silencers in the adjacent sequence, but perhaps in the presence of some cis-acting splicing enhancers, the fidelity of the 5' splice site selection should be nearly perfect. The same logic can be applied to the 3' splice site, even if it is inherently more susceptible to alternative splicing, if all of the sequence matches the consensus for the splice site sequence, branch point sequence with a strong polypyrimidine tract. In our analysis we observed canonical, alternative and cryptic splicing. Most of the observed non-canonically spliced reads could be explained by the selection of weak splice sites, or potentially canonical splice sites being generated by transcriptional error. However, even beyond these circumstances, we detected some aberrant spliced reads. However, these extremely rare events were not reported because their frequency was below next-generation sequencing error thresholds. Could these unexplainable reads represent the absolute error rate of the splicing machinery?

Taken together, the error rate analysis raises interesting theoretical questions, such as what constitutes alternative splicing as opposed to cryptic splicing? Are these cryptic spliced reads really just background noise, which serve no purpose and, therefore, should be used to approximate the rate of splicing fidelity? What becomes apparent from the analysis is that each splice site set or each exon is its own independent evolutionary unit. Part of evolution, in the transcriptomic context, requires the ability to test the fitness or viability of the new transcripts. If there were no flexibility

within the process of splicing, the transcriptome and by extension the proteome, would lose an important pathway to adaptability.

The balance and interplay of cis-acting elements and trans-acting factors that ultimately configure the potential of an exonic sequence to be included into the final mRNA isoform is extremely complex. Attempts to combine our knowledge of these variables has been referred to as the splicing code, which represents a computational attempt to predict splicing outcomes(Wang and Cooper 2007; Barash et al. 2010; Xiong et al. 2015; Busch and Hertel 2015). While each permutation of the splicing predictor gets better, each attempt has ended in very specific categorical splicing codes. Machine learning can and has helped in this regard, however, there are so many specific variables and contexts that a generalizable model may not be possible without running the danger of severely overfitting the models. Perhaps categorization is the best answer to a workable splicing code. Rather than a generalizable splicing code, there may be a plethora of more specific splicing codes for every permutation of tissue type, developmental stage and disease type.

Impact of Splicing in Disease

Based on the impact that SNVs have on disease-causing splicing mutations and the large number of genomic variations in each individual, a personalized version of a splicing code might be required. According to the 1000 Genomes Project, there are over 84.7 million SNPs, while the average genome differs from the reference genomes by 4-5 million SNPs (Gibbs et al. 2015). How many of those SNPs are actually creating diverse splicing environments is still being determined. These levels of splicing

complexity could then be taken to the next level. Cancer is undoubtedly a very complex disease, and aberrant splicing is a one of the many hallmarks of cancer (Ladomery 2013; Oltean and Bates 2014).

Using an unbiased approach, we examined the role that splicing in might play in breast cancer. We found through computational analysis that there are a number of common alternative splicing events that occur. Many of the genes we identified have been previously implicated in cancer, yet outside of splicing differences. Thus, alternative splicing of the identified genes may provide additional insights into cancer progression/diagnosis. We observed a uniform and occasionally dramatic shift in the gene expression of spliceosomal and splicing regulatory genes. We explored the differences in gene expression of these splicing-related genes between estrogen receptor positive (ER+) and estrogen receptor negative (ER-) breast cancers. The two genes we discovered that were always differentially expressed (*ESRP2*, *SF3B3*), have previously been implicated in epithelial-mesenchymal transition in breast cancer and as a marker in renal cancer, respectively. We suggested that transcriptomic analyses that include monitoring alternative splicing should be employed to evaluate other cancer categorization types, as these could yield new candidate genes to study in the battle against cancer.

We also discovered several differential-splicing events in our ER+ and ER- breast cancer comparisons. Many of these have been shown to important to tumorigenesis and cancer progression. Of particular interest are the MBNL proteins. Our results validate previous studies implicating MBNL1 exons 3, 5 and 7 as having important roles in different cancers (Tran et al. 2011; Sebestyén et al. 2016; Fish et al.

2016; Singh et al. 2018; Tabaglio et al. 2018). Based on our analysis we suggest further studies to understand the alternative splicing of exon 4 in MBNL1.

Much of our work focused on skipped exon splicing event types, as they were the most abundant in the databases evaluated. The subsequent rise in knowledge of different event types suggests that a new analysis using the most advanced programs assessing differential alternative splicing could yield new and unstudied splicing events in alternatively spliced genes. These could be used to further categorize and subtype cancers on a transcript isoform level. Such improved alternative splicing information could be an important prognostic tools given the rise in splicing targeting therapeutics, such as the antisense oligonucleotides used to treat spinal muscular atrophy and the small molecule strategies against myotonic dystrophy type 1 (Schoch and Miller 2017; Childs-Disney et al. 2013).

While there are several cutting-edge computational programs available to accurately decipher local alternative splicing isoforms (Katz et al. 2010; Shen et al. 2012; Vaquero-Garcia et al. 2016) However, until sequencing technology has the ability to capture full-length transcripts at an adequate level of depth, we will have to rely upon stitching together shorter reads and relying on junction reads to define alternative splicing variance. This will remain a challenge in the study of alternative splicing and splicing networks.

REFERENCES

- Barash Y, Calarco J a, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ. 2010. Deciphering the splicing code. *Nature* **465**: 53–9. <http://www.ncbi.nlm.nih.gov/pubmed/20445623> (Accessed May 21, 2013).
- Buratti E, Chivers M, Královičová J, Romano M, Baralle M, Krainer AR, Vořechovský I. 2007. Aberrant 5' splice sites in human disease genes: Mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Res* **35**: 4250–4263.
- Busch A, Hertel KJ. 2015. Splicing predictions reliably classify different types of alternative splicing. *RNA* **21**: 813–823. <http://rnajournal.cshlp.org/content/early/2015/03/24/rna.048769.114>.
- Chen L, Bush SJ, Tovar-Corona JM, Castillo-Morales A, Urrutia AO. 2014. Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity. *Mol Biol Evol* **31**: 1402–1413. <https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msu083>.
- Childs-Disney JL, Stepniak-Konieczna E, Tran T, Yildirim I, Park H, Chen CZ, Hoskins J, Southall N, Marugan JJ, Patnaik S, et al. 2013. Induction and reversal of myotonic dystrophy type 1 pre-mRNA splicing defects by small molecules. *Nat Commun* **4**: 2044. <http://www.ncbi.nlm.nih.gov/pubmed/23806903> (Accessed November 26, 2018).
- Fish L, Pencheva N, Goodarzi H, Tran H, Yoshida M, Tavazoie SF. 2016. Muscleblind-like 1 suppresses breast cancer metastatic colonization and stabilizes metastasis suppressor transcripts. *Genes Dev* **30**: 386–398. <http://genesdev.cshlp.org/lookup/doi/10.1101/gad.270645.115>.
- Fox-Walsh KL, Hertel KJ. 2009. Splice-site pairing is an intrinsically high fidelity process. *Proc Natl Acad Sci* **106**: 1766–1771. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2644112&tool=pmcentrez&rendertype=abstract>.
- Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, Lee S, Muzny D, Reid JG, Zhu Y, et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74. <http://www.nature.com/articles/nature15393> (Accessed November 26, 2018).
- Ladomery M. 2013. Aberrant alternative splicing is another hallmark of cancer. *Int J Cell Biol*.
- Oltean S, Bates DO. 2014. Hallmarks of alternative splicing in cancer. *Oncogene* **33**: 5311–5318. <http://dx.doi.org/10.1038/onc.2013.533>.
- Pickrell JK, Pai A a, Gilad Y, Pritchard JK. 2010. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet* **6**: e1001236. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3000347&tool=pmcentrez&rendertype=abstract> (Accessed August 14, 2013).
- Schoch KM, Miller TM. 2017. Antisense Oligonucleotides: Translation from Mouse Models to Human Neurodegenerative Diseases. *Neuron* **94**: 1056–1070. <http://dx.doi.org/10.1016/j.neuron.2017.04.010>.
- Sebestyén E, Singh B, Miñana B, Pagès A, Mateo F, Pujana MA, Valcárcel J, Eyraas E. 2016. Large-scale analysis of genome and transcriptome alterations in multiple

- tumors unveils novel cancer-relevant splicing networks. *Genome Res* **26**: 732–744. <http://genome.cshlp.org/lookup/doi/10.1101/gr.199935.115>.
- Singh B, Trincado JL, Tatlow P, Piccolo SR, Eyras E. 2018. Genome Sequencing and RNA-Motif Analysis Reveal Novel Damaging Noncoding Mutations in Human Tumors. *Mol Cancer Res* **16**: 1112–1124. <http://mcr.aacrjournals.org/lookup/doi/10.1158/1541-7786.MCR-17-0601>.
- Stepankiw N, Raghavan M, Fogarty EA, Grimson A, Pleiss JA. 2015. Widespread alternative and aberrant splicing revealed by lariat sequencing. *Nucleic Acids Res* **43**: 8488–8501. <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv763>.
- Tabaglio T, Low DH, Teo WKL, Goy PA, Cywoniuk P, Wollmann H, Ho J, Tan D, Aw J, Pavesi A, et al. 2018. MBNL1 alternative splicing isoforms play opposing roles in cancer. *Life Sci Alliance* **1**: e201800157. <http://www.life-science-alliance.org/lookup/doi/10.26508/lsa.201800157>.
- Tran H, Gourrier N, Lemercier-Neuillet C, Dhaenens CM, Vautrin A, Fernandez-Gomez FJ, Arandel L, Carpentier C, Obriot H, Eddarkaoui S, et al. 2011. Analysis of exonic regions involved in nuclear localization, splicing activity, and dimerization of muscleblind-like-1 isoforms. *J Biol Chem*.
- Vořechovský I. 2006. Aberrant 3' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Res* **34**: 4630–4641. <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkl535> (Accessed November 26, 2018).
- Wang G-S, Cooper TA. 2007. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8**: 749–761. <http://www.nature.com/doi/10.1038/nrg2164>.
- Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, et al. 2015. The human splicing code reveals new insights into the genetic determinants of disease. *Science (80-)* **347**: 1254806–1254806. <http://www.sciencemag.org/cgi/doi/10.1126/science.1254806>.

