

UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

The Impact of Pre-mRNA Splice Site Selection on mRNA Stability and Splicing Fidelity in Methionine-Dependent Cancer Cells

Permalink

<https://escholarship.org/uc/item/1bn4q8sf>

Author

Carranza, Francisco Gutierrez

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

The Impact of Pre-mRNA Splice Site Selection on mRNA Stability and Splicing Fidelity in
Methionine-Dependent Cancer Cells

DISSERTATION

Submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biomedical Sciences

By

Francisco Gutierrez Carranza

Dissertation Committee:
Professor Klemens J. Hertel, Chair
Professor Peter Kaiser
Professor Bert L. Semler

2023

Chapter 1 and Appendix B © 2022 RNA Biology
Appendix A © 2017 Methods in Molecular Biology
All other materials © 2023 Francisco Gutierrez Carranza

DEDICATION

To my parents: Leobardo and Aurelia Carranza.

My brothers: Jesus and Daniel.

Mis abuelos: Bulmaro y Estela Carranza, Joel y Maria Gutierrez

TABLE OF CONTENTS

	Page
LIST OF FIGURES	v
LIST OF TABLES	vii
ACKNOWLEDGEMENTS	viii
VITA	ix
ABSTRACT OF THE DISSERTATION	xii
CHAPTER 1: Introduction	1
Pre-mRNA splicing	1
Alternative splicing	5
Splicing regulation	7
Intron/Exon definition	10
mRNA degradation	12
Cancer and splicing	14
Methionine metabolism and cancer	14
Methylation and splicing	19
CHAPTER 2: Splice site proximity influences alternative exon definition	20
Summary	20
Introduction	21
Results	26
Discussion	36
Methods	48
CHAPTER 3: Genome-wide determination of mRNA and exon half-lives	53
Summary	53
Introduction	54
Results	56
Discussion	75
Materials and Methods	78
CHAPTER 4: Nutritional Control of Splicing Fidelity Contributes to Methionine Dependent Proliferation Defects in Cancer Cells	81
Introduction	82
Results	84
Discussion	100
Methods	103
CHAPTER 5: Perspectives	104
Introduction	104
Alternative splicing decisions are impacted by different modes of exon recognition	104
The influence of alternative splicing on mRNA degradation	106

Splicing fidelity is linked to nutrient availability in cancer cells and contributes to methionine dependence in cancer	108
REFERENCES	110
APPENDIX A: Isolation of Newly Transcribed RNA Using the Metabolic Label 4-thiouridine	121
APPENDIX B: Splice site proximity influences alternative exon definition	130

LIST OF FIGURES

	Page
Figure 1.1 Sequence Elements for Splice Site Recognition	2
Figure 1.2 Pre-mRNA splicing by the Major Spliceosome	4
Figure 1.3 The Types of Alternative Splicing Events	6
Figure 1.4 Splicing Regulatory Components	9
Figure 1.5 Intron Definition and Exon definition	11
Figure 1.6 The Hoffman effect	16
Figure 1.7 Methionine metabolism	18
Figure 2.1 Gene architecture and database	24
Figure 2.2 5' ss selection preference for different internal exon categories	27
Figure 2.3 Cross-exon selection of alternative 5' alternative splice sites	33
Figure 2.4 3'ss selection preference for different internal exon categories	35
Figure 2.5 Unifying model for the influence of splice site proximity in alternative exon Definition	39
Figure 2.6 The downstream 5' splice site is a functional splice site	41
Figure 2.7 GC content distribution for intron definition splice sites (SS) and exon definition splice sites (LL) based on the intron length-dependent classification used herein (SS<250 nts, LL>250 nts)	44
Figure 3.1 Experimental Design	56
Figure 3.2 Relationship between sequence length and mRNA half-life	59
Figure 3.3 The influence of exon length for terminal exons half-lives	62
Figure 3.4 Comparison between standard and outlier exons	64
Figure 3.5 Half-life distribution of standard, more stable outlier exons, and less stable outlier exons	65
Figure 3.6 Correlation between mRNA isoforms and RNA length	71
Figure 3.7 Correlation between exon degradation kinetics and sequence conservation	73
Figure 4.1 Methionine-dependent cells (MB468) experience the highest impact of methionine stress at 720 min in MET- HCY+ medium	85
Figure 4.2 Methionine stress in MB468 impacts splicing fidelity	88

Figure 4.3 Gene ontology analysis of alternatively spliced genes upon methionine stress in MB468 cells	90
Figure 4.4 Inverse relationship between overlapping skipped exon and intron retention events in MB468 and R8 cells	93
Figure 4.5 Methionine stress in MB468 cells leads to loss in SmD1 methylation and loss of splicing fidelity	95
Figure 4.6 Decreased methionine availability coupled with PRMT5 inhibition promotes cell death	97

LIST OF TABLES

	Page
Table 2.1. Alternative 5'ss selection and resulting exon length correlation	30
Table 3.1 Outlier exons are larger in size when compared to standard exons regardless of outlier stability	66
Table 3.2 Outlier exons are larger in size when compared to standard exons	68

ACKNOWLEDGMENTS

My deepest gratitude to UC Irvine's Minority Science Programs for introducing me to biological research. Thank you to Dr. Marlene de la Cruz and Dr. Luis Mota-Bravo for your kindness, trust, and continued mentorship.

I would like to thank my undergraduate mentors at UC Riverside. Thank you to the MARC U* program and the program director Dr. Ernest Martinez. Thank you, Dr. Morris Maduro and Gina Broitman-Maduro, for your mentorship. My years in the Maduro lab were crucial in my early development as a scientist. In addition, I want to thank my graduate student mentor in the Maduro lab, Dr. Hailey Choi. Aside from being a great scientific mentor you have been a great friend.

Many thanks to all the friends I have made along my academic journey, who have been with me through the high and lows. To my high school friends Matt, Andrew, and Jean, thank you for always hyping me up and most importantly being my friends. I Love and appreciate the whole Gonzales, Kim, and Crews family. Thank you to my 361 crew, my brothers and sister. Joey, Donna, Andre, and Andres. Thank you for always keeping me grounded and hyping me up at the same time. I am forever blessed to have such great friends cheering me on.

I am eternally grateful for the guidance and opportunities I have received from my PI and mentor Dr. Klemens (Dicki) Hertel. You have provided me with a lab/home where I could learn and thrive. Your support has meant the world to me, especially, when imposter syndrome has reared its ugly head. Thank you.

I would also like to thank the many members of the Hertel lab. "The girls" Dr. Angela Garibaldi and Dr. Maliheh Movassat thank you for being my bonus PIs and the big sisters I never knew I needed. Thank you to Dr. Wendy Ullmer for your guidance and sense of humor. Thank you to my "carnal" Dr. Hossein Shenasa. I look back fondly at our scientific and nonscientific discussions. Many thanks to Jessie Alteri for your friendship. I have an appreciation for our scientific discussions.

I would like to thank the members of my committee, Dr. Bert L. Semler and Dr. Peter Kaiser. I am grateful for your guidance, scientific discussion, and collaboration opportunities.

Lastly and most importantly, thank you and love you to my family. To my brother Jesus and Daniel, thank you for supporting me and keeping me grounded. Welcome to the family Danielle, Carmina, and the new baby Carranza. To my parents Aurelia and Leobardo, I will forever cherish all your unwavering support and love. Your sacrifices were worth it. Gracias a mis Abuelitos, Pienso en todos ustedes todos los días. Los extraño y los quiero mucho a todos. Les mando a todos un fuerte abrazo y beso.

This dissertation and all the work therein would not be possible without funding from the National Institutes of Health and the National Science Foundation. The text of this dissertation is a reprint of the material as it appears in RNA biology and Methods in Molecular Biology, used with permission from "Taylor & Francis" and Springer Nature publishing.

VITA

Francisco Gutierrez Carranza

EDUCATION

University of California, Irvine, CA 2023
Doctor of Philosophy, Biomedical Sciences

University of California, Riverside, CA 2014
Bachelor of Science, Biology

Fullerton College 2012
Associates of Science, Biology

RESEARCH EXPERIENCE

Ph.D. Graduate Student Researcher, Dr. Klemens J. Hertel, PhD 2016-2023
Department of Microbiology and Molecular Genetics
University of California, Irvine

Minority Health and Health Disparities International Research 2015
Training (MHIRT) Research Fellow, Dr. Jose Luis Martinez, PhD
Departamento de Biotecnología Microbiana
Centro Nacional de Biotecnología. Madrid, España

Undergraduate Student Researcher, Dr. Morris Maduro, PhD 2013-2015
Department of Biology
University of California, Riverside

Undergraduate Summer Researcher, Dr. Anthony A. James, PhD 2012
Department of Microbiology and Molecular Genetics
University of California, Irvine

Undergraduate Summer Researcher, Dr. Jose M Ranz, PhD 2011
Department of Ecology & Evolutionary Biology
University of California, Irvine

TEACHING EXPERIENCE

Bridges to the Baccalaureate Journal Club instructor Minority Science Programs University of California, Irvine	2018
Teaching Assistant University of California, Irvine	2021

CONFERENCES AND PRESENTATIONS

- RNA Society: Poster. May 2020
- Cold Spring Harbor: Pre-mRNA Processing: Talk, August 2019
- UC Irvine School of Medicine Grad Day: Poster Presentation, October 2019 & 2020
- Microbiology and Molecular Genetics Department Seminar: Talk, 2018, 2019, 2020
- UC Irvine RNA Club: Talk, 2018
- UC Irvine Brews & Brains: Talk, Fall 2017
- UC Riverside MARC U* Summer Symposium: Talk, Summer 2013 & 2014
- 19th International *C.elegans* Meeting at UCLA: Poster, Summer 2013
- Annual Biomedical Research Conference for Minority Students (ABRCMS), Poster, November 2011 & 2012

-

LEADERSHIP

- UC Irvine Department of Microbiology and Molecular Genetics: 2019-2020
Seminar selection committee
- UC Irvine Department of Microbiology and Molecular Genetics: 2018-2019
Graduate student representative
- UC Irvine DECADE SOM student representative 2019-2020
- UC Irvine DECADE Campus Coordinator 2018-2019

Fellowships/Awards

UCI LEAD, Excellence in Research and Health for the Latino Community Award, School of Medicine UCI School of Medicine – Travel Stipend Award	2021
NIH CREEDS Computational Genomics course	2018
NSF Graduate Research Fellowship Program	2017- 2020
NSF Bridges to the Doctorate Graduate Fellowship	2015- 2017
NIH sponsored Maximizing Access to Research Careers (MARC U*)	2013- 2014

Publications

1. **Carranza, F***, Shenasa, H*, & Hertel, K.J. Splice site proximity influences alternative exon definition. *RNA Biology* 19, No. 1, 829-840 (2022)
2. Clifton, B. D., **Carranza, F** et al. Understanding the early evolutionary stages of a tandem drosophila melanogaster-specific gene family: A structural and functional population study. *Mol. Biol. Evol.* 37, 2584–2600 (2020).
3. **Carranza, F***, Garibaldi, A*. & Hertel, K. J. Isolation of newly transcribed rna using the metabolic label 4-thiouridine. in *Methods in Molecular Biology* 1648, 169–176 (Humana Press Inc., 2017).
4. Maduro, M. F., **Carranza, F** et al. MED GATA factors promote robust development of the *C. elegans* endoderm. *Dev. Biol.* 404, 66–79 (2015).
5. Yeh, S.-D., **Carranza, F** et al. Functional evidence that a recently evolved *Drosophila* sperm-specific gene boosts sperm competition. *Proc. Natl. Acad. Sci. U. S. A.* 109, 2043–8 (2012).

* denotes equal contribution

Memberships

- RNA Society student member
- UC Irvine GPS-STEM

ABSTRACT OF THE DISSERTATION

The Impact of Pre-mRNA Splice Site Selection on mRNA Stability and Splicing Fidelity in Methionine-Dependent Cancer Cells

By

Francisco Gutierrez Carranza

Doctor of Philosophy in Biomedical Sciences

University of California, Irvine, 2023

Professor Klemens J. Hertel, Chair

Eukaryotic gene expression is an essential process for proper cell differentiation, development, and the cell's response to environmental signals. Pre-mRNA splicing is an important part of the eukaryotic gene expression program. Splicing contributes to protein diversity, gene expression regulation, evolution, and genetic diseases. Understanding the intricacies of pre-mRNA splicing is important for understanding gene expression and for the rational design and development of new therapies for genetic diseases.

The regulation of pre-mRNA splicing is a highly combinatorial process that relies on many cis- and trans-acting elements. Some of these elements include splice site strength and the intron-exon architecture. It is proposed that spliceosome assembly can either occur across the intron (referred to as intron definition) and across the exon (referred to as exon definition). Selecting between these modes of spliceosome assembly is thought to be dictated by intron architecture. Other studies have demonstrated that the proximity between the 5' splice site and its intronic 3' splice site plays a critical role in splice site selection. In chapter 2 we conducted a genome-wide computational analyses to evaluate the proximity rules in the context of intron and

exon definition. Our computational studies were complemented using designer mini-genes in cell transfection assays to evaluate the impact of splice site proximity in alternative splicing.

The ability to regulate gene expression allows the cell to adjust to its ever-changing needs and external cues. Aberrant regulation of gene expression is linked to diseases such as cancer. One major contributor to modulate gene expression is through the regulation of mRNA stability. A change in mRNA stability can lead to differing protein expression levels while alternative splicing primarily promotes protein diversity. The work described in Chapter 3 outlines what gene features impact mRNA stability and how alternative splicing can influence mRNA stability. Data was generated by conducting a 24-hour 4sU pulse-chase RNA-seq experiment. Our computational analyses allowed us to explore the relationship between mRNA stability and gene and/or exon length. In addition, we established a pipeline to derive exon and mRNA isoform half-lives to investigate the influence of alternative splicing on mRNA stability.

Cancer cells have been known to have unique metabolic needs for proliferation. One such need is the cancer cell's metabolic addiction to methionine, referred to as the "Hoffman effect." While the Hoffman effect has been observed in a wide array of cancer cells, the mechanisms by which it arises, and controls tumorigenesis are not fully understood. In chapter 4, gene expression analyses of methionine-dependent and independent cell lines reveal that splicing dysregulation is linked to methionine dependence. In particular, proper methylation of a general spliceosomal component is implicated as link between changes in splicing fidelity and the accessibility of exogenous methionine in cancer cells.

CHAPTER 1

Introduction

One of the most fascinating aspects of the eukaryotic gene expression is the existence of pre-mRNA splicing and, subsequently, alternative splicing. Pre-mRNA splicing is a co-transcriptional process that leads to the simultaneous excision of intron sequences and ligation of exon sequences to result in a mRNA transcript used for translation [1], [2]. Pre-mRNA splicing paves the way for mRNA and protein diversity, which is achieved through a process known as alternative splicing [3]. Alternative splicing allows human cells to transcribe ~215,000 mRNA isoforms from ~22,000 protein coding genes [4]. Alternative splicing is an important contributor to phenotypic complexity and organism complexity seen in higher multicellular eukaryotes [5], [6]. Additionally 15% of human hereditary diseases and cancers have been associated with aberrant splicing [7]. These facts alone make researching the mechanisms of RNA splicing an important study. In this chapter, I will focus on introducing the mechanism and determinants of spliceosome formation, the process of splice site selection, and the connection between splice site selection and spliceosome assembly. In addition, I will introduce the concept of methionine dependence in cancer cells and the impact of cancer on pre-mRNA splicing. Lastly in this chapter, I will introduce mRNA degradation and its relationship with RNA processing.

Pre-mRNA splicing

Pre-mRNA splicing is carried out by the spliceosome, an RNA-protein complex composed of 5 small nuclear ribonucleoproteins (snRNPs), U1, U2, U4, U5, and U6. These snRNPs consist of a snRNA that acts as scaffold, seven Sm proteins, and other associated proteins.

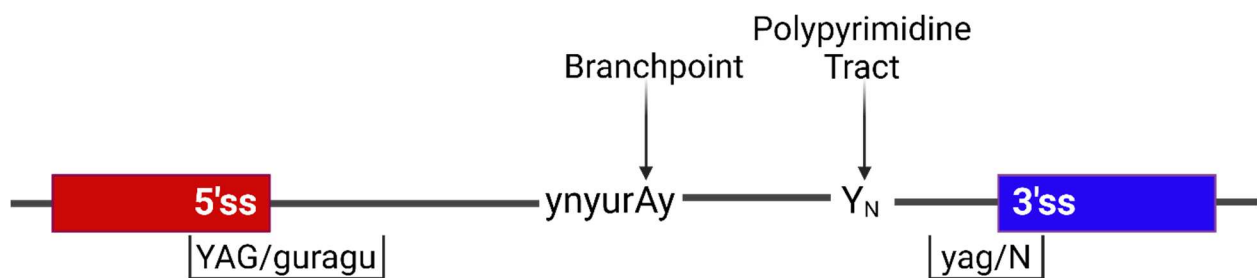


Figure 1.1 Sequence Elements for Splice Site Recognition.

A schematic of location and consensus sequence of the sequence elements required for spliceosomal assembly. The elements displayed are: 5' splice site demarking the exon/intron junction, branchpoint sequence, polypyrimidine tract, and 3' splice site located at the 3' of the intron boundary. The thin black line refers to the intron boundary, the red and blue box refer to the exon boundary. The "/" refers to the exon/intron junction with the capitalized sequence referring to the exon sequence at both the 5' and 3' splice site. Y refers to a pyrimidine (C or U nucleotide), R refers to a purine (A or G nucleotide), and N refers to any nucleotide.

Before stepwise spliceosome assembly can begin, splice site recognition of the 5' splice site (5'ss) and the 3' splice site (3'ss) must occur. The 5'ss is composed of a loosely conserved nine-nucleotide consensus sequence located at the 5' intron/exon junction, YAG/guragua (Y refers to a pyrimidine, R refers to a purine, "/" refers demarcates the boundary between the exonic sequence and intronic sequence, lower case letter refers to the intron, and uppercase refers to the exon sequence) (Figure 1.1). The 3'ss is defined by three sequence elements, a YAG trinucleotide located at the 3' intron/exon junction, an upstream polypyrimidine tract (PPT) (~20 nts), and the branch point sequence (BPS), which is located approximately 50 nts upstream of the 3' intron/exon junction [8]–[10] (Figure 1.1). Spliceosome assembly is initiated by U1 snRNP binding to the 5'ss via U1 snRNA complementarity to the 5'ss sequence and SF1/BBP protein and subunit U2AF binding to the BPS and the polypyrimidine tract, thus forming the E-complex [11], [12] (Figure 1.2). U2 snRNP replaces SF1/BBP to interact with the BPS in an ATP dependent manner to form A complex. Pre-catalytic B complex is formed with the incorporation of U4/U6, U5 tri-snRNP. RNP rearrangements result in an activated B complex triggering the release of U1 and U4 snRNPs. The activated spliceosome then carries out the first catalytic step of splicing with the help of RNA-dependent ATPase/helicases Prp28 and Brr2 to form the catalytic step 1 spliceosome (C complex). During the formation of C complex the phosphodiester bond at the 5'ss is attacked by the 2'-hydroxyl of the BPS adenosine. This creates a free 3'-hydroxyl at the upstream exon and an intron lariat at the 5' end of the downstream exon. During the subsequent step II of the splicing reaction (C* complex), the free 3'-hydroxyl of the upstream exon attacks

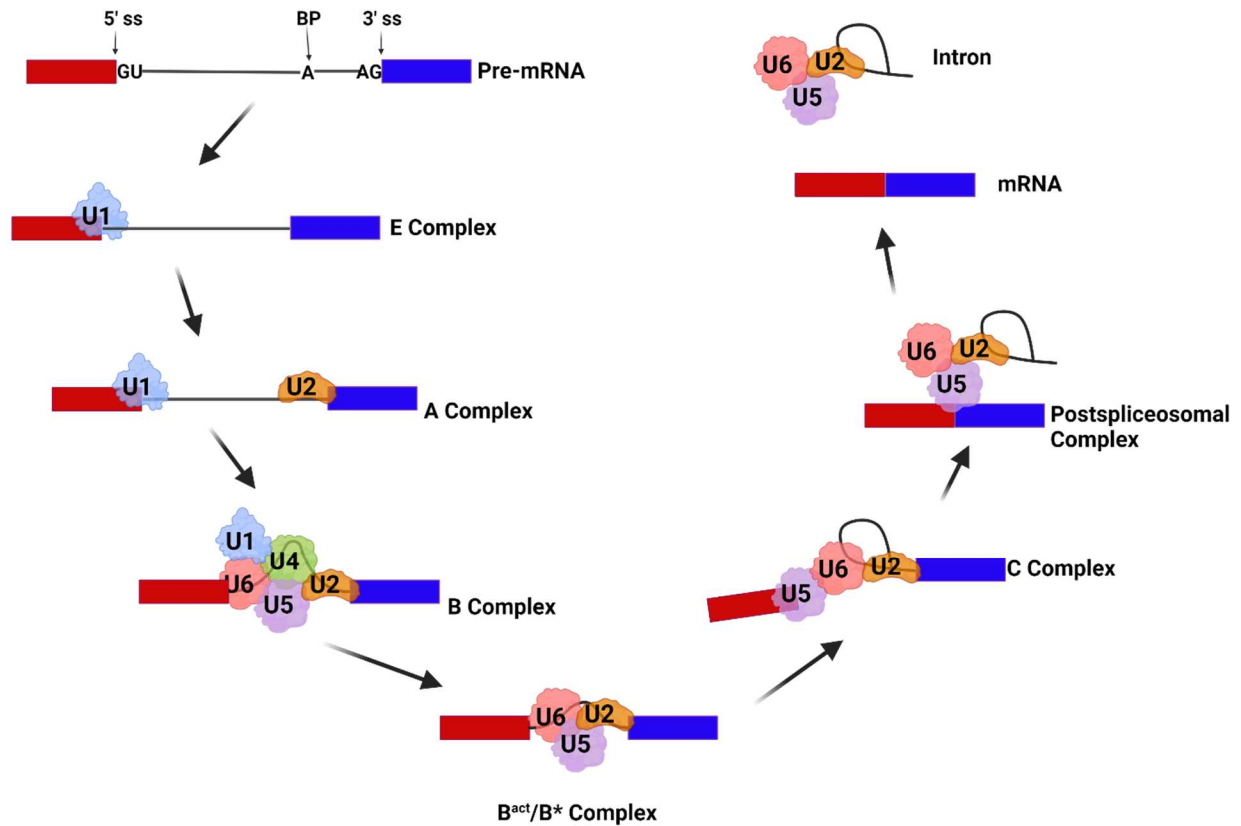


Figure 1.2 Pre-mRNA splicing by the Major Spliceosome.

This figure displays a step wise schematic of pre-mRNA splicing. The red and blue boxes indicated the exons, and the thick black line refers to the intron. BP refers to the branchpoint sequence.

the phosphodiester bond at the 3'ss, resulting in exon transesterification and excision of the intron lariat (Figure 1.2) [9], [12]–[15].

Alternative splicing

Constitutive splicing generates one mature mRNA from pre-mRNAs. A deviation from that is alternative splicing, a mechanism that allows for the generation of different mRNA isoforms from a single gene. It has been shown that ~95% of human genes undergo alternative splicing based on cell cycle, development, tissue origin, or signaling events, each of which are known to associate with changes in gene expression [3], [16]. It has also been postulated that alternative splicing plays an important role in the evolution of organism complexity [17]. Given its prevalence, alternative splicing significantly contributes to proteomic diversity in humans [18], [19]. Five different categories of alternative splicing (AS) exist: exon skipping or cassette exons (ES), mutually exclusive exons (MXE), intron retention (IR), alternative 5' splice site (A5SS), and alternative 3' splice site (A3SS) (Figure 1.3). The most frequent AS category is ES with an ~40% frequency in humans. MXE occurs at 10%, A3SS at 18%, A5SS at 7%, and IR at 5% [20]. AS may lead to mRNA degradation via the activation of nonsense-mediated decay or to the generation of unique protein isoforms [7], [21], [22]. AS may create abnormal or nonfunctional proteins, which in turn could contribute to genetic disease. To understand the impact alternative splicing has on gene expression, it is imperative to understand the underlying mechanism in this RNA processing step.

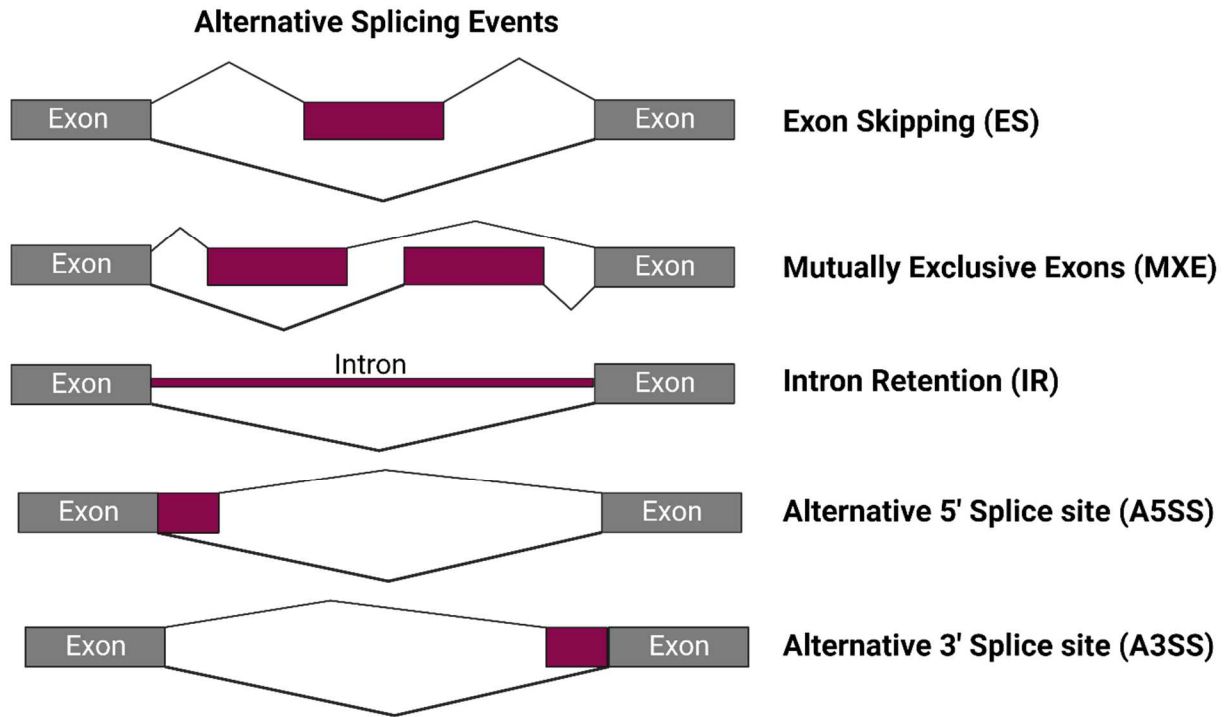


Figure 1.3 The Types of Alternative Splicing Events.

The lines depict the possible resulting mRNA due to alternative splicing. The alternative exon or retained intron is displayed in maroon. ES, the most common form of alternative splicing, results in the skipping of one or more exons in the final mRNA. Mutually exclusive exons refer to splicing of exon in a manner such that two or more splicing events are not independent. Intron retention refers to inclusion of one or more intron sequences in the final mRNA transcript. Alternative 5'ss refers to the use of an alternative 5'ss leading to different 3' boundary on the exon of interest. The opposite is seen in alternative 3'ss events where an alternative 3'ss is used leading to different 5' boundary on the exon of interest.

Splicing regulation

There are many determinants that impact pre-mRNA splicing outcomes. Exon recognition is a critical first step in pre-mRNA splicing. The strength of 5'ss and 3'ss play a major role in defining the boundary between exons and introns [9], [11], [23]. Cis-regulatory elements, such as splicing enhancer and silencing sequences (Figure 1.4), are known to recruit trans-acting splicing regulatory proteins to the pre-mRNA that promote or repress splicing. In addition to that, RNA secondary structure and the process of RNA polymerase II transcription has been shown to influence splice site selection [24].

The 5'ss sequence demarcates the 3' end of the exon and the 5' end of the downstream intron [25], [26]. The strength of the 5'ss is based on the complementarity between the 5'ss sequence and U1 snRNA [11]. The 3'ss strength is largely characterized by U2AF binding to the PPT [27]–[29]. These splice site recognition events are crucial as they signal the beginning of spliceosome assembly and formation of A complex [15]. Splice site strength can be assessed numerically by applying a maximum entropy-based method (MaxEntScan, [30]). MaxEntScan allows for the use of a scoring system where a positive score associates with stronger splice site while a negative score associates with a poorer splice site. Using MaxEntScan, it has been shown that 3' and 5' splice sites have a near equal impact on exon recognition. As expected stronger splice site strength results in better exon recognition [31]. This and other splice site strength scoring systems have highlighted the importance of splice site strength to predict splicing outcomes [11], [17], [30]–[32].

Splicing regulatory elements (SRE) are cis-acting regulatory sequence elements that are found near splice sites. They are binding sites for trans-acting splicing regulatory proteins that aid or hinder spliceosome assembly. There are four types of SREs that differ based on splicing impact

and location relative to the regulated exon: exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs), and lastly intronic splicing silencers (ISSs) (Figure 1.4) [4], [9], [23]. Most ESEs carry out their function by recruiting SR proteins to aid in splicing [33], [34]. SR proteins are members of the essential serine/arginine rich protein family. They interact with RNA via their RNA recognition motif (RRM) and with other proteins via their RS domain [34]. SR proteins not only regulate alternative splicing events but they have been shown to promote constitutive splicing [33]–[35]. One example of SR proteins promoting splicing is seen with SRSF2 and SRSF1 aiding in the recruitment of U1 to the 5'ss through its RS domain by associating with an ESE [34], [36]. Most frequently ESSs are binding sites for heterogeneous nuclear ribonucleoproteins (hnRNPs), proteins that contain one or more RNA-binding domain and a splicing inhibitory domain [9], [23], [37], [38]. HnRNPs can be described as splicing repressors or antagonists to SR proteins [37]–[41]. They can inhibit the recruitment of snRNPs by blocking U2 snRNP entry or U4/U6, U5 tri-snRNP recruitment [39], [42], [43]. The classical view of SR proteins is that of an activator while hnRNPs is that of a repressor of splicing. A study has challenged that belief; this study demonstrated that SR proteins and hnRNPs have positional-dependent functions. SR proteins enhance splicing when recruited to an exon and inhibit splicing when recruited to the adjacent intron. The opposite effect is observed for hnRNPs. HnRNPs were shown to be splicing activators when bound to the intron and splicing repressors when bound to exon [9], [44].

RNA secondary structures quickly form as the pre-mRNA is transcribed, thus potentially impacting the recognition of splice sites by changing their proximity to other regulatory sequences or by influencing the accessibility of SREs [45]–[47] [24], [45]–[48]. As most splicing

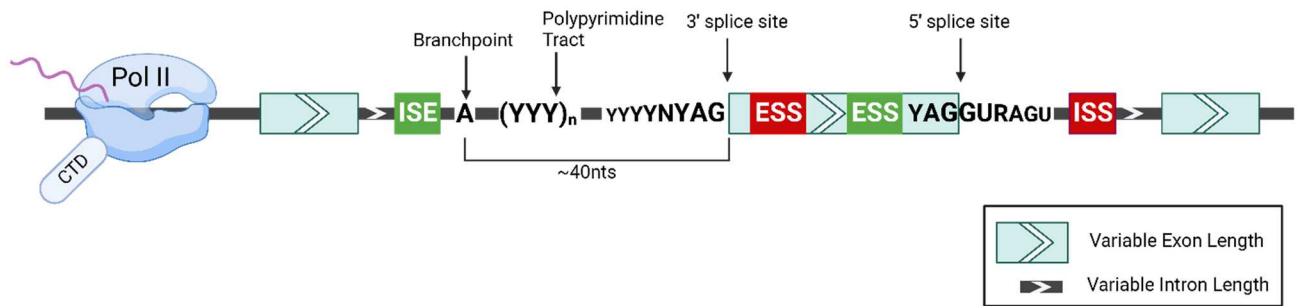


Figure 1.4 Splicing Regulatory Components.

Depiction of different regulatory sequence elements or regulatory factors. They include intron sequence enhancers (ISEs), intron sequence silencers (ISS's), exon sequence silencer (ESS's), and exon sequence enhancers (ESE's). This also includes splice site strength, intron/exon length, and transcription rate.

is co-transcriptional [49], [50], variations in the speed of RNA polymerase II transcription has been shown to influence splice site selection [51], [52]. A reduced elongation rate can result in an increased inclusion usage of cassette exons while the opposite is seen when elongation rates are increased [9], [52]. A recent study has also demonstrated a connection between RNA polymerase II transcription and RNA secondary structure formation [9], [47]. This study used a structural chemical probing method to demonstrate that the structural plasticity of nascent pre-mRNA transcripts differs depending on the elongation rate, thus modifying RNA splicing.

Intron/Exon definition

Another vital splicing determinant is the intron-exon architecture. It is defined by the length of the exons and introns, features of the pre-mRNA that play an important role in pre-mRNA splicing [9], [14], [49], [53]–[56]. The genomes of lower eukaryotes are characterized by large exons flanked by small introns [53]. In higher eukaryotes, genomes are characterized by long introns and short exons. The average intron length in most higher eukaryotes exceeds 1 or 2 kilobases while in lower eukaryotes the intron length varies from 50-500 nucleotides [57]. The average exon of yeast is ~1800 nts in length while the average vertebrate exon length is ~170 nts [14], [58]. Early studies investigating the impact of the intron-exon architecture in human cell lines demonstrated that large exons (> 500 nts) were skipped when flanked by large introns (> 500 nts). However, the same large exons were included when they were flanked by small introns [53]. These observations gave rise to the exon definition and the intron definition model (Figure 1.5). In intron definition, splice site recognition occurs across the intron. Intron definition occurs when an intron length is small (< 250nts) [59]. When introns are large, exon definition permits splice site recognition across the exon (> 250 nts) [59]. An initial cross exon-

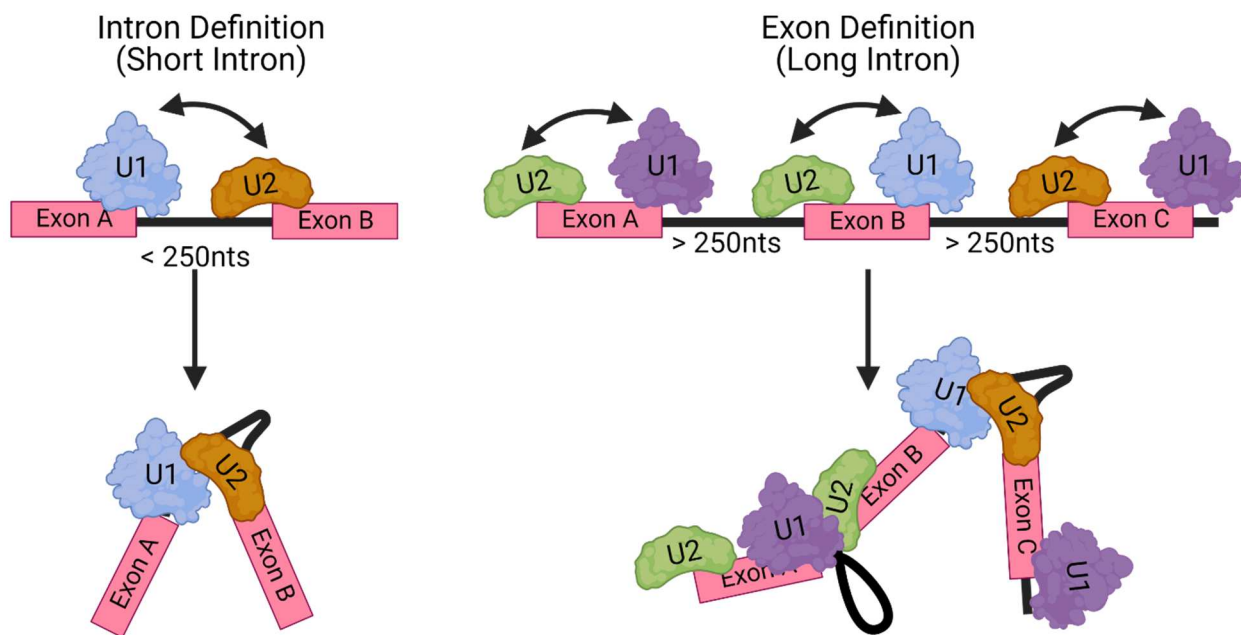


Figure 1.5 Intron Definition and Exon definition.

The two proposed modes of splice site recognition. During intron definition splice sites are recognized across the intron (left). Under exon definition (right) splice sites are initially recognized across the exon, followed by splice site juxtaposition.

complex of U1 and U2 is formed, followed by a cross-intron spliceosome complex formation to define the intron to be excised [14], [60]. *In-vitro* splicing assays in human cell extract demonstrated that splice site recognition across the intron is employed when introns are less than 250 nts in length. *Drosophila* exons flanked by long introns displayed a significantly higher frequency of alternative splicing than exons flanked by short introns. A similar trend is seen in the human genome but to a lesser extent [59]. Intron definition is described as the prominent form of splice site recognition in lower eukaryotes and is considered to be the default mode of splice site recognition [14], [59]. These results suggest that exon definition is the predominant form of splice site recognition in humans, and intron definition is the predominant form of splice site recognition in *Drosophila*. The results also highlighted the importance of the intron/exon architecture in AS, as has been reaffirmed by others [14], [56], [60], [61]. While originally seen as a concept, additional studies have presented structural evidence for exon definition as well [43], [62]–[64]. Most recently, a study showed that exon definition plays an important role in maintaining proper regulation of AS of a proto-oncogene [60]. The vastly larger intron lengths observed in the human genome, compared to the intron length seen in lower eukaryotes, has also been linked to increased rates of AS and enhanced organismal complexity [59], [65]. Interestingly, exon size is an evolutionary-conserved feature [66]. Combined these findings demonstrate the importance of the intron-exon architecture and its effect on splice site recognition. Studying the impact and interconnection in this complex splicing regulatory network remains a highly important field of study.

mRNA degradation

An important aspect of life is the cell's ability to regulate its gene expression program to adjust to changing needs and external cues. Dysregulation of gene expression programs is usually

studied through steady state expression comparisons between disease and healthy cells. There are two main contributors to establishing steady state RNA expression levels, transcriptional output and mRNA stability. Increased transcription rates lead to increased accumulation of mRNAs in the cell. Reduced activity of RNA polymerases decreases mRNA levels. This can be seen in leukemia where the transcription factor TAL1 limits the transcriptional output of important tumor suppressors to maintain proliferation [67], [68]. Changes in mRNA stability is another contributor to altered gene expression. Increased mRNA degradation rates result in fewer protein products made from that mRNA transcript. Accordingly, increased mRNA stability can elicit increased protein output. Altered mRNA stability has been described to be important in childhood osteosarcoma where XRN1, an important exoribonuclease involved in mRNA decay, is downregulated [69].

mRNA degradation is linked to translation as the poly-(A) tail shortens because of non-sense mediated decay (NMD), non-stop decay (NSD) or no-go decay NGD. NMD is a form of mRNA surveillance triggered by pre-mature termination codons [22]. Non-stop decay is activated for mRNAs lacking a stop codon [70]. No-go decay degrades mRNAs containing stalled ribosomes [71]. After several rounds of translation, the general pathway of mRNA degradation occurs via deadenylation by CCR4/NOT deadenylase containing enzyme complexes leading to 3' to 5' degradation. Alternatively, deadenylation can lead to 5' decapping via the decapping complex and finally a 5' to 3' mRNA digestion. Recent studies have also demonstrated the importance of codon optimality on translation speed, ultimately leading to different stability in mRNAs [72]. All degradation pathways can be impacted by alternative splicing. The relationship between mRNA stability and alternative splicing is one that still needs to be further explored.

Cancer and splicing

Aberrant pre-mRNA splicing has been proposed to be a cancer hallmark [73], [74]. A direct link between aberrant AS and cancer was demonstrated by the discovery of mutations in genes encoding pre-mRNA-splicing factors [75]–[77]. Since then, there has been a myriad of direct and indirect links of splicing alterations in cancer [73], [77]–[80]. This is especially highlighted in an analysis of ~8,000 tumors across 32 cancer types that demonstrated thousands of splicing variants not present in non-cancerous tissues [81]. Much aberrant splicing identified in cancer is associated with mutation in and/or altered expression of the splicing machinery [78]. For example, in chronic lymphocytic leukemia (CLL) a mutation in the U1 snRNA results in antagonistic changes in splicing impacting known cancer driver genes [82]. Other splicing factors with known mutations that are associated with cancer prognosis include SRSF2, U1 snRNA and U2AF1 [78]. Mining of the Cancer Genome Atlas revealed that putative cancer driving mutations occur in 119 genes encoding core splicing factors and regulators across 33 cancer types. Together they account for ~60% of the components associated with the splicing machinery [78], [83].

Methionine metabolism and cancer

Most cancer cells display a unique metabolic requirement known as the Warburg effect [84]. The Warburg effect describes the phenomenon by which cancer cells predominantly harness energy using aerobic glycolysis. This requires high uptake of glucose and fermentation of the resulting lactic acid [84], [85]. Aerobic glycolysis has been shown to be a requirement for tumor growth [84]. Another, less studied, metabolic requirement is the Hoffman effect, which describes a methionine dependence [86], [87]. Most cancer cells cannot proliferate in media where methionine is replaced with its precursor homocysteine. Normal cells do not display proliferation

defects in homocysteine media. Thus, proliferation of cancer cells requires exogenous methionine (Figure 1.6).

In mammalian cells methionine is an essential amino acid typically obtained through dietary intake [86]. In addition to functioning as a building block of proteins, methionine plays key roles in epigenetic control (*S*-adenosylmethionine or SAM), nuclear functions (polyamines), detoxification (glutathione), cellular membranes (phospholipids), and the modulation of nucleotide biosynthesis [86]. SAM is considered the cell's principal cellular methyl donor and acts as a cofactor in most methylation reactions, including epigenetic control. Once a methylation reaction occurs, SAM is irreversibly converted to *S*-adenosylhomocysteine (SAH). SAH is hydrolyzed to the metabolic precursor homocysteine (HCY) by removing adenosine from SAH. HCY is then remethylated to form methionine, which is subsequently converted to SAM to complete the methionine cycle [86], [88].

Another pathway for methionine generation is via the methionine salvage cycle [86]. SAM also functions as the sole donor of aminopropyl groups in polyamine synthesis. Polyamines are vital at high concentrations for cell proliferation and are often overexpressed in cancers [86]. Because of elevated polyamines levels, high concentrations of SAM are needed to maintain cell proliferation. Polyamine synthesis takes place within the context of the methionine salvage pathway. Decarboxylated SAM is converted to five'-deoxy-5'-methylthioadenosine (MTA). MTA is processed through the methionine salvage pathway to recycle adenine and methionine. One important aspect of MTA is its role as a competitive inhibitor of protein arginine N-methyltransferase 5 (PRMT5), which is often found to be overexpressed in cancer.

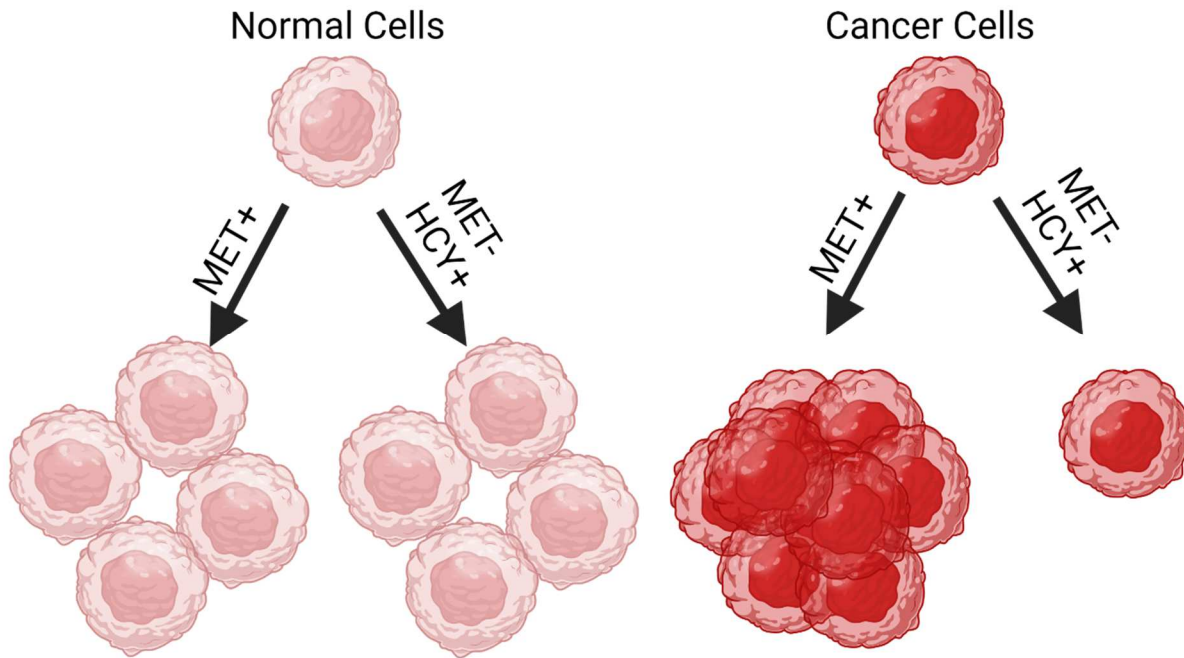


Figure 1.6 The Hoffman effect.

Non-transformed cells have the same proliferation rate in methionine (MET) supplemented media or methionine precursor homocysteine (HCY) supplemented media. Cancer cells, depicted on the right, experience proliferation defects in media absent of exogenous methionine, even when supplemented with homocysteine.

In addition, methylthioadenosine phosphorylase (MTAP), an enzyme involved in converting MTA to methionine, is often found to be deleted in tumors, thus, severely comprising methionine regeneration from the salvage pathway, and causing elevated levels of MTA [81].

These observations underscore the importance of methionine metabolism when trying to understand the addiction of cancer cells to exogenous methionine. Most cancer cells cannot proliferate and arrest in G1 phase when cultured in methionine-depleted medium supplemented by precursor homocysteine. Significant insights into understanding the Hoffman effect were made when studying the methylation potential in the cell which is best described as SAM/SAH ratio [89]. Tracer experiments using a breast cancer line (MDA-MB468) grown in methionine depleted (-MET) medium containing deuterium-labeled homocysteine showed that homocysteine was redirected towards the trans-sulfuration pathway and glutathione synthesis, away from the methionine cycle (Figure 1.7). This redirection resulted in a lower SAM/SAH ratio [89]. In addition, SAM supplementation to methionine depleted and homocysteine supplemented (MET-HCY+) medium has been demonstrated to alleviate the proliferation defects observed for methionine dependent cells [90]. Combined these studies highlight the importance of SAM and the methylation potential in mediating the Hoffman effect. It has been postulated that the SAM/SAH ratio acts as a measure for a metabolic checkpoint during the cell cycle. If the SAM/SAH ratio is too low, the cell cycle remains arrested at this checkpoint and apoptosis is initiated [86].

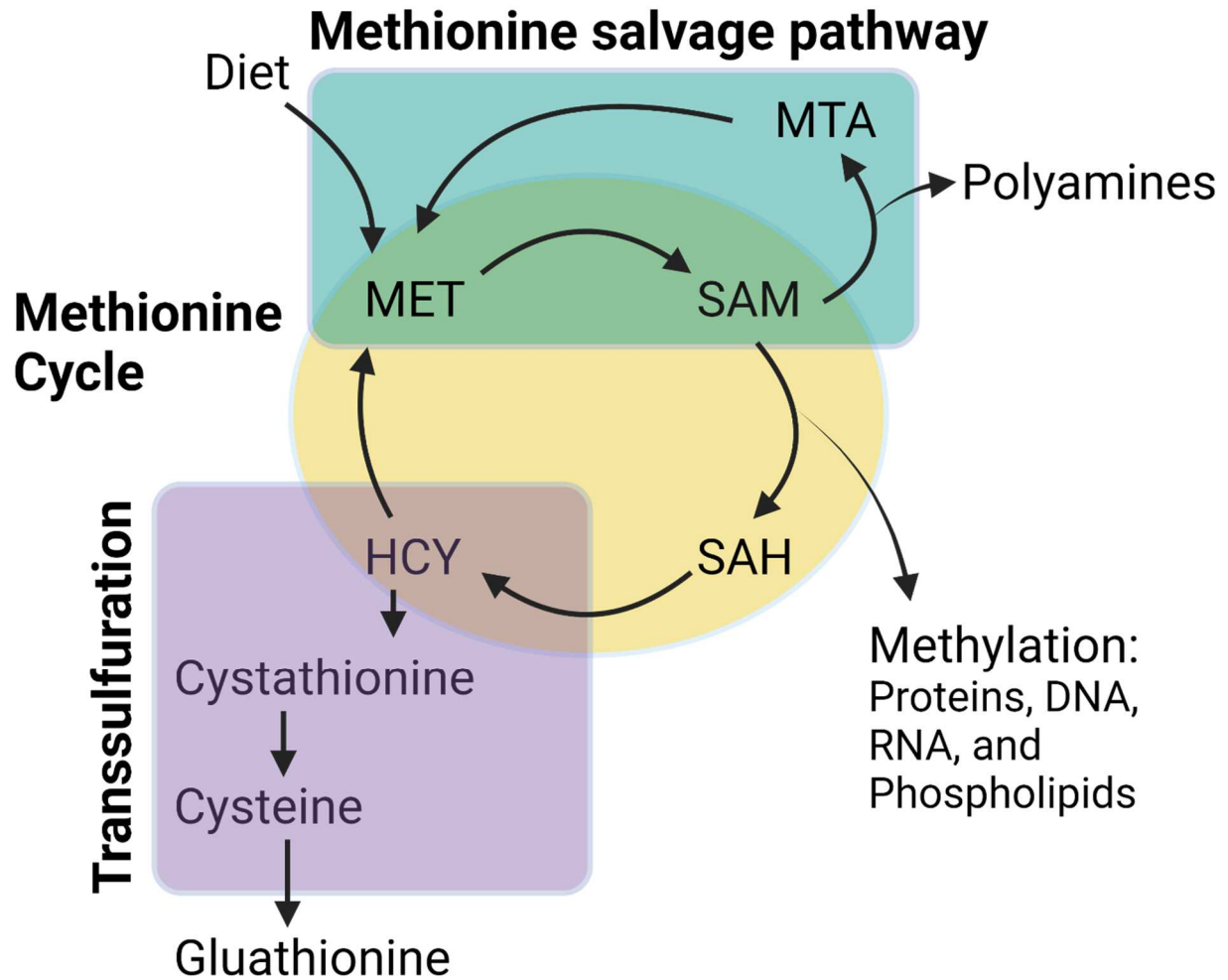


Figure 1.7 Methionine metabolism.

The metabolic connections between the methionine cycle primarily responsible in maintain methylation potential in the cell. The methionine salvage pathway, responsible for polyamine synthesis and methionine regeneration. Lastly, the transsulfuration pathway responsible for generation of glutathione needed to combat oxidation.

Methylation and splicing

Considering that many splicing factors are methylated, methyltransferases have been known to play vital roles in splicing. For example, PRMT5 symmetrically dimethylates spliceosomal proteins Sm D1, D3 and B/B' on their C-terminus [86]. This methylation drastically increases the binding affinity of these Sm proteins to the SMN complex, which is responsible for the assembly of snRNPs. The conditional knockout of PRMT5 in mouse stem cells led to wide-spread aberrant splicing [91]. A recent profiling of the PRMT4/5/7 methylome demonstrated that these methyltransferases regulate the binding of splicing factors (hnRNPA1) to the pre-mRNA and subsequently AS [80]. However, the connection between changes in AS and the altered methylation potential triggered by the Hoffman effect has not yet been explored.

CHAPTER 2

Splice site proximity influences alternative exon definition

Summary

Alternative splicing enables higher eukaryotes to expand mRNA diversity from a finite number of genes through highly combinatorial splice site selection mechanisms that are influenced by the sequence of competing splice sites, cis-regulatory elements binding trans-acting factors, the length of exons and introns harboring alternative splice sites and RNA secondary structures at putative splice junctions. To test the hypothesis that the intron definition or exon definition modes of splice site recognition direct the selection of alternative splice patterns, we created a database of alternative splice site usage (ALTssDB). When alternative splice sites are embedded within short introns (intron definition), the 5' and 3' splice sites closest to each other across the intron preferentially pair, consistent with previous observations. However, when alternative splice sites are embedded within large flanking introns (exon definition), the 5' and 3' splice sites closest to each other across the exon are preferentially selected. Thus, alternative splicing decisions are influenced by the intron and exon definition modes of splice site recognition. The results demonstrate that the spliceosome pairs splice sites that are closest in proximity within the unit of initial splice site selection.

Introduction

Pre-mRNA splicing is an essential step in eukaryotic gene expression that involves the excision of intronic sequences and the transesterification of exonic sequences by the spliceosome to generate protein coding mRNAs. Alternative exon inclusion is possible through a process known as alternative splicing. At least 95% of human genes undergo alternative splicing in response to cell cycle, developmental, tissue-specific or signaling cues. Alternative splicing increases proteomic diversity from a limited genome in a regulated fashion [3]. Thus, pre-mRNA splicing impacts gene expression [92].

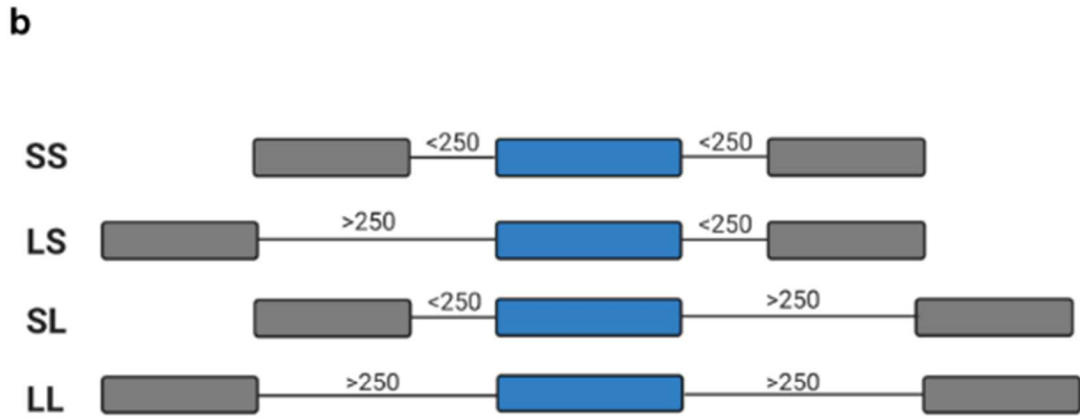
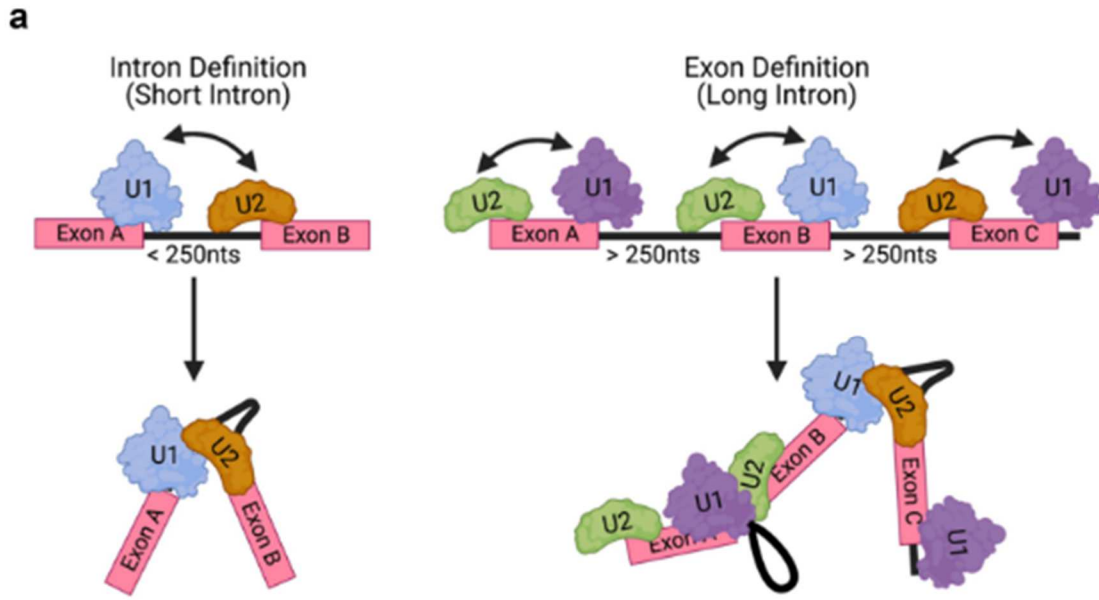
The recognition of splice junctions by the spliceosome initiates the splicing reaction. The 5' splice site (5'ss) is defined by a nine-nucleotide consensus sequence that spans the exon/intron junction at the 5' end of each intron. The 3' splice site (3'ss) includes three sequence elements found within an approximately 40 nucleotides (nts) stretch, upstream of the 3' intron/exon junction. These include the intron/exon junction sequence, which contains the essential AG dinucleotide at the 3' end of the intronic sequence, the polypyrimidine tract (PPT), a region containing 15–20 pyrimidines located upstream of the intron/exon junction and the branch point sequence, a highly degenerate sequence that contains a conserved adenosine located upstream of the PPT.

Exon recognition is a highly combinatorial process that is known to be influenced by many cis- and trans-acting features. These include splicing enhancers, silencers, RNA secondary structure, the intron-exon architecture, and the sequence context of splice junctions [23], [38], [44]. The strength of splice sites is determined by how well they conform to consensus splice junction motifs that function in recruiting U1 snRNP to the 5'ss and U2AF to the 3'ss. Consensus similarity scores, derived from the modelling of short sequence motifs using the maximum-

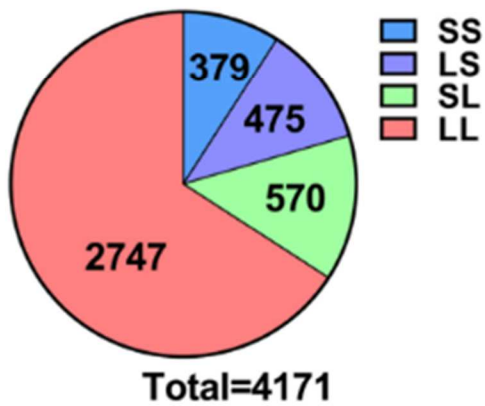
entropy principle (MaxEnt), define splice site strength numerically [30]. Splice sites are known to act synergistically and combined 5' and 3'ss scores are a much better predictor for exon inclusion than either splice site score alone [31]. Importantly, the ability of an exon to undergo various forms of alternative splicing is heavily influenced by the strength of its splice sites [93].

Another crucial factor in splice site selection is the genomic architecture [53], [59], [66], [94]. The genomes in lower eukaryotes are characterized almost exclusively by the presence of short introns (<250 nts). By contrast, human genes harbor long introns, with >87% of introns longer than 250 nts [59]. This different genomic architecture has been shown to contribute significantly to the manner in which spliceosomal assembly occurs. The two proposed mechanisms through which splice sites are recognized are referred to as the exon or intron definition mode of splice site recognition (Figure 2.1a). During intron definition, the spliceosome assembles across the intron that will be excised. Under conditions that promote exon definition, initial splice site recognition is postulated to occur across the exon. This initial recognition is predicted to be followed by an additional splice site juxta-positioning step to induce intron excision. *In vitro* splicing and transfection experiments of designer minigenes demonstrated that the transition between intron and exon definition occurs at an intron length of approximately 250 nts [59]. Thus, splice sites that are flanked by large introns (>250 nts) are recognized through exon definition, while intron-defined splice sites are associated with small flanking introns (<250 nts). It is currently unknown how exon and intron definition influence alternative splice site selection.

Understanding the relationship between the splice site strength and intron-exon architecture splicing determinants has been a longstanding goal in deciphering the splicing code. The mechanisms utilized by the spliceosome to select the correct splice site in the presence of



c Alternative 5' ss



d Alternative 3' ss

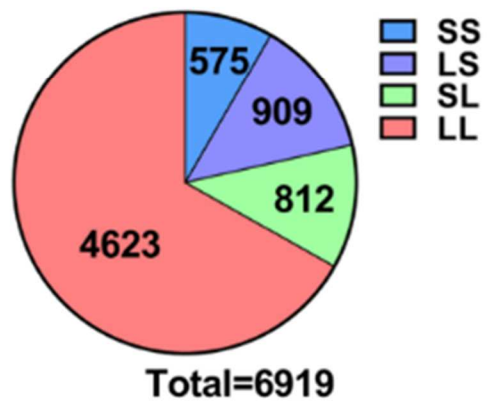


Figure 2.1 Gene architecture and database.

(a) The two proposed modes of splice site recognition. During intron definition splice sites are recognized across the intron (left). Under exon definition (right) splice sites are initially recognized across the exon, followed by splice site juxtaposition. (b) ALTssDB categories of internal exons as defined by flanking intron size. S stands for short (less than 250 nts), L stands for long (greater than 250 nts). (c) and (d) Distribution of ALTssDB internal exon categories for alternative 5' (c) and alternative 3' (d) splice site events.

multiple nearby cryptic or alternative splice sites are still not completely understood. Differences in intron-exon architecture and splice site strength are known to be important in mediating alternative splice site selection [93]. A series of classical experiments demonstrated that the proximity between the 5' and 3' splice sites, across the intron, plays a crucial role in splice site preference [94]. Reed and Maniatis showed that the splice site closest to its intronic splicing partner was favoured over a distal competing splice site [94]. Thus, in the case of competing alternative 5' splice sites, the downstream 5'ss was preferred because it was more proximal to the pairing 3'ss. Similarly, between competing 3' splice sites, the upstream 3'ss was chosen. These observations suggest that in the absence of confounding factors, shorter distances between splice sites are favoured during intron-defined splicing. This may be because splice site pairing is more efficient across shorter distances. These experiments established a splice site selection proximity rule (for clarity referred to as the intron-centric proximity rule); however, it is unclear how dominant it is within the hierarchical nature of known splicing determinants.

In this study, we carried out computational analyses to assess the impact of the intron-centric proximity rule. We demonstrate that the intron-centric proximity rule is generally applicable for the intron definition mode of splice site definition. For the exon definition mode of splice site definition, we observe an exon-centric proximity rule that deviates from the classical intron-centric proximity rule. The 5' and 3' splice sites closest to each other across the exon are preferentially selected. Thus, when the unit of splice site definition is across the intron (intron definition), the 5' and 3' splice sites closest to each other across the intron preferentially pair. When the unit of splice site recognition is the exon (exon definition), the 5' and 3' splice sites closest to each other across the exon are preferentially selected. Our results provide evidence that

alternative splicing decisions are influenced by the intron and exon definition modes of splice site recognition.

Results

The influence of intron-exon architecture on 5' splice site selection

To determine the impact of the intron-exon architecture and splice site strength on splice site selection, we created a database of alternative splice sites (ALTssDB) using the Human Exon Splicing Event Database HEXEvent [95], the Intron DB [96] and GeneBase [97]. MaxEntScan, a computational tool, was used to assign splice site scores [6]. To minimize variability, we focused on competing alternative 5' or 3' splice site pairs of internal exons with only one alternative splice pattern. Thus, ALTssDB catalogs pairs of alternative 5' splice sites competing for a common 3'ss or pairs of alternative 3' splice sites competing for a common 5'ss. ALTssDB reports the location of the major splice site and its competing alternative 3' or 5' splice site, corresponding exon sizes, usage levels, splice site scores and flanking intron lengths. Using these filters, ALTssDB captures 4,171 human 5' ss competition events and 6,919 human 3'ss competition events (Figure 2.1b-d).

We first tested whether the intron-centric proximity rule holds true when evaluating all alternative 5'ss events transcriptome-wide (Figure 2.2a). In agreement with the intron-centric proximity rule expectation that the downstream 5'ss should be selected over a competing upstream 5'ss, we observed a preference for downstream 5'ss selection in ~60% (2,497) of the alternative 5'ss splicing events (Figure 2.2b, left bar).

To evaluate whether the 'intron definition' or 'exon definition' mode of splice site selection influence adherence to the intron-centric proximity rule, we parsed the 5'ss dataset into

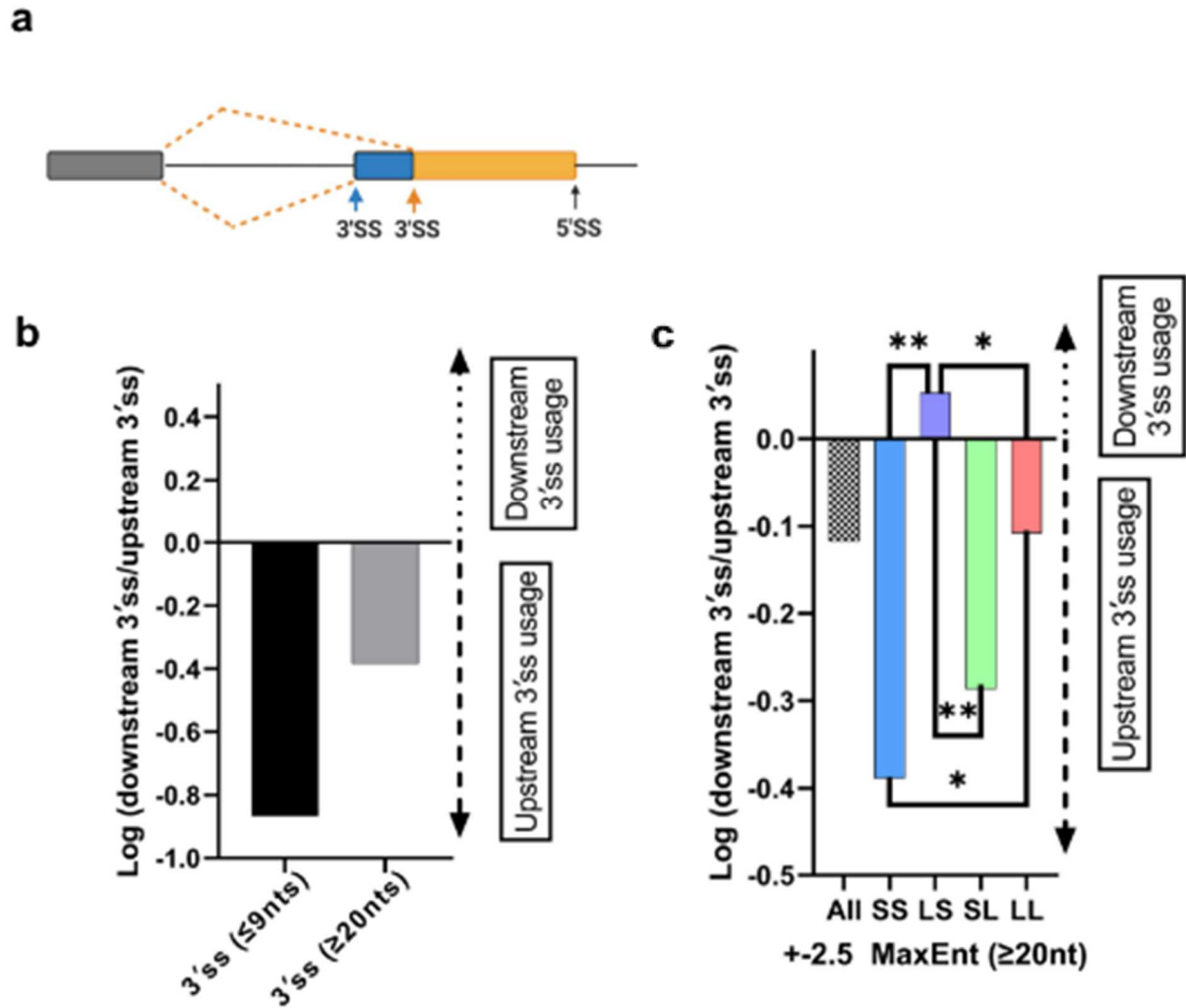


Figure 2.2 5' ss selection preference for different internal exon categories.

(a) Model depicting alternative 5'ss patterns. (b) Bar graph depicting the preference for downstream or upstream 5' ss selection for different internal exon categories (sample size = 379 SS, 2,747 LL, 570 LS, 475 SL). A positive log ratio represents downstream 5' ss preference, a negative log ratio represents upstream 5' ss preference. (c) Splice site selection preference for alternative 5' splicing events with near equal splice-site strength scores ($\Delta \pm 2.5$ MaxEnt, sample size = 88 SS, 725 LL, 104 LS, 143 SL), Fisher's exact test was performed. (b, c), ** $p < 0.01$, * $p < 0.05$.

intron definition events (379 SS), exon definition events (2,747 LL), and hybrid events (570 LS, 475 SL) (Figure 2.1b and c). For the purpose of alternative 5'ss selection analysis, the hybrid architectural class LS was categorized as intron defined because the 5'ss is adjacent to a short intron and U1 snRNP binding to the 5'ss at the exon/intron junction initiates early spliceosome formation [9]. By analogy, the architectural class SL was considered exon defined because the 5'ss is contained within a long intron. Surprisingly, in all four intron architecture classes, the majority of events still displayed a preference for the downstream 5'ss, consistent with the intron-centric proximity rule, albeit to varying degrees (Figure 2.2b). For example, the downstream 5'ss is selected more frequently for intron definition events (represented by SS, LS) when compared to exon definition events (represented by LL, SL). These varying degrees of preference suggest that the intron definition mode of splice site selection adheres more stringently to the intron-centric proximity rule.

The influence of intron-exon architecture on 5'ss selection in the absence of splice site strength differences

One important determinant that may mask the influence of splice site proximity is the difference in the splice site strength of competing splice sites. To determine the impact of splice site strength on alternative 5'ss selection, we compared the splice strength of the major 5'ss versus the alternative 5'ss. In 86% of the events evaluated the 5'ss with a higher predicted splice strength was the dominant 5'ss, irrespective of whether the exon was predicted to be recognized through exon definition (LL, SL) (85%) or intron definition (SS, LS) (90%) events. These results support the notion that splice site strength is a strong determinant in alternative 5'ss selection.

To determine how the exon and intron definition modes of splice site selection influence alternative splicing the impact of splice site strength differences was minimized computationally.

This was achieved by isolating 5'ss competition events with near equal splice site scores ($\Delta\text{MaxEnt} = \pm 2.5$), resulting in 88 SS, 725 LL, 104 LS, and 143 SL events. Interestingly, when this splice site strength filter was applied, we observed that the upstream 5'ss is preferentially selected in 60% of competition events, inconsistent with the expectations of the intron-centric proximity rule (Figure 2.2c, left bar). Strict intron definition events (SS category) display a downstream 5'ss selection preference, consistent with the intron-centric proximity rule, while strict exon definition events (LL) display a preference for the upstream 5' splice site (Figure 2.2c). The upstream preference under exon definition is inconsistent with the intron-centric proximity rule but consistent with an exon-centric proximity rule. These biases are heightened in the hybrid categories SL (upstream preference) and LS (downstream preference) (Figure 2.2c). These results suggest that for exon definition events the upstream 5'ss, which is proximal across the exon to the upstream 3'ss, is favored. By contrast, for intron definition events, the 5'ss proximal across the intron to the downstream 3'ss is favored.

The influence of exon size on 5'ss selection

It is known that exon size can influence splice site selection [53], [59]. To determine the influence of exon size on splice site selection, we compared splice patterns between three different exon size groups, exons smaller than 50 nts, exons between 50–250 nts in length, and exon longer than 250 nts. These cutoffs were chosen based on natural exon size distributions. We then calculated how frequently the major isoform contains the stronger 5'ss for the three different exon size classes (Table 2.1). When the major and the alternative exons are smaller than 50 nts, splice preference is driven almost exclusively by the stronger splice site score (Table 2.1). This preference weakens when the usage of the alternative 5'ss generates an exon greater than 50 nts. Thus, differences in exon size contribute to splice site selection, with a preference for generating

		Exon size generated with major splice site usage		
		Ex ≤50 nts	50 < Ex ≤250 nts	Ex >250 nts
Exon size generated with minor splice site usage	Ex ≤50 nts	98% ^{a1}	86% ^a	100% ^a
	50 < Ex ≤250 nts	73% ^b	87% ^{a2}	88% ^a
	Ex >250 nts	75% ^b	58% ^b	77% ^{b3}

Table 2.1. Alternative 5'ss selection and resulting exon length correlation.

The table reports how frequently the major isoform contains the stronger splice site when the alternative splice site lies within one of three different exon size classes. ^{a,b}Within a column, means without a common superscript differ ($p < 0.05$) between size categories in each column. ¹29% preference for the upstream 5'ss. ²42% preference for the upstream 5'ss. ³44% preference for the upstream 5'ss.

shorter exons. A similar trend is observed for alternative patterns of major exons within the 50–250 nts range. The selection of alternative exons larger than 250 nts is much less likely to be driven by splice site differences. These data provide evidence that exon size contributes to splice site selection with a preference for defining smaller exons.

Experimental verification of genome-wide computational analysis

To test whether the proposed exon-centric proximity rule can be confirmed experimentally, we tested five minigenes that contain an internal exon with two competing 5' splice sites of identical strength (MaxEnt 10.9, CAG/guaagu) and one 3' splice site with a MaxEnt of 12.56 (uguccuuuuuuuccacag/ CUG) (Figure 2.3a). All minigenes were designed to be recognized through exon definition (flanking intron size of 365 nts) and differ only in the resulting internal exon size. Cell transfection experiments demonstrated that for all constructs tested the upstream 5'ss was chosen exclusively (Figure 2.3b), consistent with the computational analysis demonstrating that upstream 5' splices sites are favoured under exon definition. To test if this splice site preference is altered when both competing splice sites are weakened, we mutated both 5' splice sites to have a MaxEnt score of -0.5 (GAG/ guguca). In the larger exon constructs (L and XL), this resulted in preferential internal exon skipping. In the M and S constructs, the upstream 5'ss maintained its preference (Figure 2.3c). These results demonstrate that in an isogenic exon definition context the 5'ss most proximal to the upstream 3'ss is favoured, supporting the computational analysis of an exon definition 'cross-exon proximity' preference.

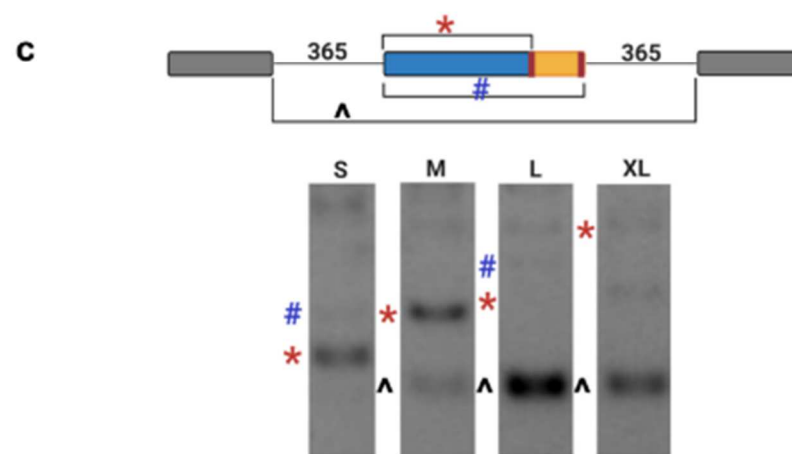
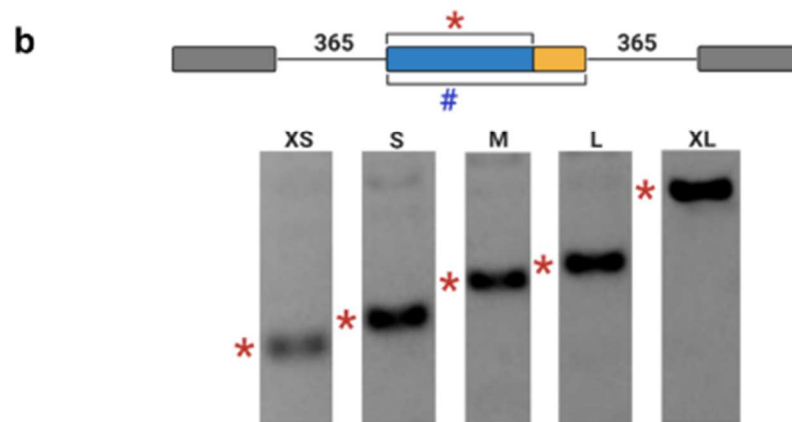
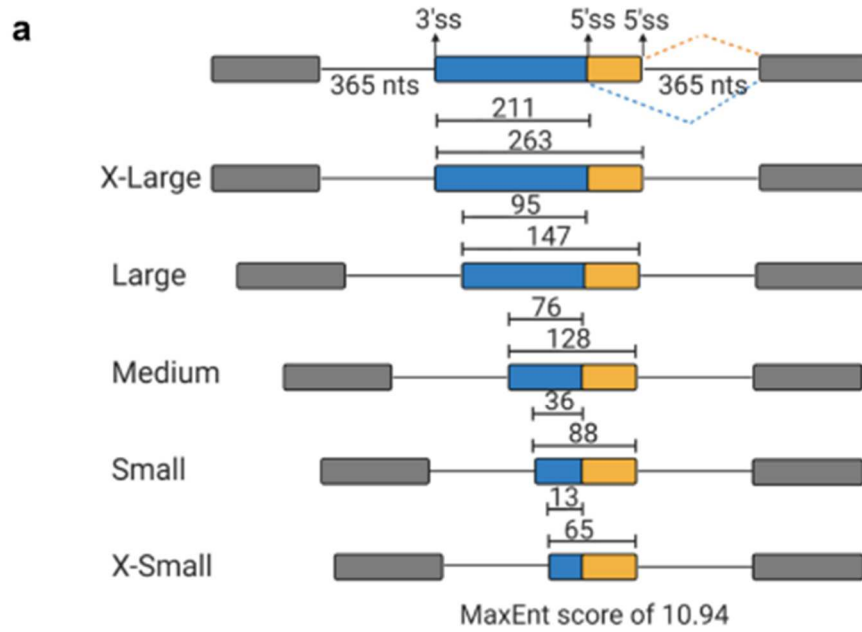


Figure 2.3 Cross-exon selection of alternative 5' alternative splice sites.

(a) Schematic of exon-defined mini-gene constructs with identical splice site strength (CAG/guaagu, MaxEnt = 10.9) used in transfection experiments. The size of the resulting internal exon is indicated for upstream (blue) and downstream 5' ss selection. (b) Representative image of ethidium bromide stained agarose gel splicing analysis. Bands denoting upstream (red symbol) or downstream (blue symbol) 5'ss usage are marked to the left of the image. (c) Splicing outcome of minigene constructs with identical but weakened competing 5' ss (GAG/guguca, MaxEnt = -0.5). Bands denoting upstream 5'ss usage (red symbol), downstream (blue symbol) 5' ss usage, or exon skipping (black symbol) are marked to the left of the image.

The influence of intron architecture on 3'ss selection

To investigate the impact of intron size and splice site strength on 3'ss selection we built a 3'ss dataset analogous to the 5'ss dataset described above (Figure 2.4a). For our analysis, we took into consideration that the 3'ss is recognized during the first and the second steps of splicing. Prior to the first step of splicing, the polypyrimidine tract is bound U2AF, which subsequently recruits U2 snRNP to the branch point. After the first step of splicing, the 3' splice junction YAG/N is selected before the exons are ligated via a transesterification reaction. It has been demonstrated that competing 3' splice sites in close proximity (up to 9 nts) are selected during the second step of splicing after identical first step definition [28]. Alternative 3' splice sites further apart (greater than 1220 nts) are typically defined during initial splice site recognition using different polypyrimidine tract and branch points. Thus, we split the 3'ss dataset into 'first step recognition' (≥ 20 nts apart from one another, 3839 events) and 'second step recognition' events (≤ 9 nts apart from one another, 2317 events). Both 3'ss event groups show a preference for upstream 3'ss usage, consistent with the intron-centric proximity rule (Figure 2.4b). This preference is particularly strong for the second step alternative 3'ss events. Filtering to obtain competing 3'ss pairs with comparable strengths and categorizing these events into intron (S/S, 69 events) or exon definition (L/L, 664) events again demonstrated the influence of the intron architecture on 3'ss selection (Figure 2.4c). The strong upstream 3'ss preference observed for intron defined events (S/S) is significantly reduced when splice sites are selected in the exon definition mode (L/L). Consistent with our 5'ss analysis, the hybrid classes (SL, 88 events and LS, 147 events) display more extreme splice site preferences relative to the SS and LL classes, with SL mimicking intron definition and LS mimicking exon definition behaviour.

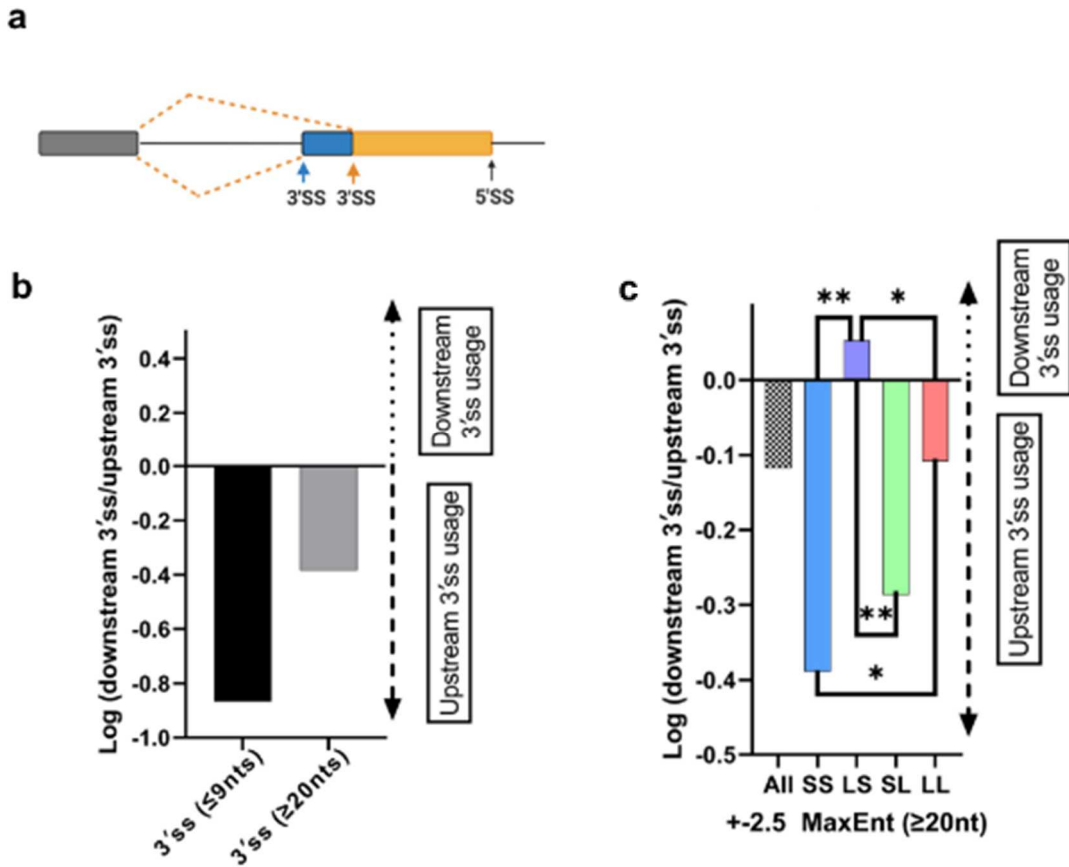


Figure 2.4 3'ss selection preference for different internal exon categories.

(a) Model depicting alternative 3'ss patterns. (b) Bar graph displaying the 3'ss preference for first (>20 nts distance between competing 3'ssplice sites, 2317 events) or second (<9 nts distance between competing splice sites, 3839 events) step selection. A positive log ratio represents downstream 3'ss preference, a negative log ratio represents upstream 3'ss preference. (c) Bar graph depicting the preference for downstream or upstream 3'ss selection with near equal splice-site strength scores for different internal exon categories ($\Delta \pm 2.5$ MaxEnt sample size = 69 SS, 664 LL, 88 SL, 147 LS). Fisher's exact test was performed. (b, c), ** $p < 0.01$, * $p < 0.05$.

Together, our transcriptome-wide analyses demonstrate that the mode of splice site selection critically influences splice site choice. For intron definition, splice sites closest across the intron are preferentially selected. Under exon definition, the selection of splice sites closest across the internal exon are favoured. These results suggest that the gene architecture influences alternative splicing by promoting splice site recognition via the intron or exon definition pathway.

Discussion

The regulation of pre-mRNA splicing is a combinatorial process that is controlled by splice site sequences, cis-regulatory elements binding trans-acting factors, the intron-exon architecture, and RNA secondary structure among other features [9]. Two mechanisms of splice site recognition have been proposed within the broader concept of intron-exon architecture. It has been postulated that under the intron definition splice sites are recognized across the intron, making the intron the initial unit recognized by the spliceosome. In an alternative mode of splice site recognition, splice sites are postulated to be initially recognized across the exon in a process called exon definition. Once the exon is defined as the initial unit of splice site recognition, subsequent structural rearrangements are predicted to recognize and pair the upstream and downstream splice sites across flanking introns [56]. The mechanisms of intron and exon definition have been studied in the field for almost 30 years [14], [53]–[56], [62], [94], [98].

Early evidence that the length of introns and exons is important came from size constraints on exon inclusion from minigenes that were transfected in cell culture. Large exons were efficiently spliced when flanked by short introns, consistent with an intron definition mechanism. However, when intron lengths were increased exons were only included efficiently if they were relatively short, less than ~500 nts long. The latter observation suggests that the

early spliceosome has a limited ‘wing-span’ when the exon is the unit of initial splice site recognition. Subsequently, biochemical studies demonstrated that intron definition is more efficient and that the rate of splicing for exon defined substrates is considerably slower. This study identified intron length as the primary determinant in the mode of splice site recognition employed by the early spliceosome and placed the transition from intron definition to exon-definition at the point when flanking introns become longer than 200–250 nts [59].

Another classical study used *in vitro* splicing assays to demonstrate that alternative splice site choice is influenced by the proximity between the pairing splice sites. When two splice sites are in competition, the splice site proximal to the intron is preferred. As a result, this proximity bias induces the preferential excision of the smaller intron [94]. This study and the pioneering study from Sterner and Berget when analysed together suggest that in the context of splice site competition, selection of proximal splice sites across an intron may allow the intron to be recognized through intron definition, while the selection of the distal splice site may lead to a larger unit of initial splice site recognition that may change the mode of splice site recognition all together [53], [94]. In broader terms, the findings by Reed and Maniatis [94] indicated that perhaps proximity across the initial unit of splice site recognition would drive splice site selection and influence alternative splicing. We set out to determine whether the proximity of splice sites across the proposed initial unit of splice site recognition may provide genome-wide evidence for the two modes of splice site recognition and elucidate their roles in alternative splicing.

Our analysis of the alternative splicing events captured in ALTssDB permitted the derivation of several important conclusions. First, the intron-centric proximity rule observed by Reed and Maniatis is maintained within the context of the intron definition mode of splice site

recognition [94]. In the context of exon definition, we observe an exon-centric proximity rule, where the proximity between 5' and 3' splice sites across the exon dictates splice site preference. Alternative exons subject to the intron-centric proximity rule undergo removal of the smaller intron and selection of the larger exon. Conversely, alternative exons subject to the exon-centric proximity rule undergo removal of the larger intron and selection of the smaller exon. Initially, these observations may appear inconsistent with each other, yet they highlight a commonality of spliceosomal assembly across the smallest unit of initial splice site recognition. For the intron definition mode of splice site recognition this unit is the intron, meaning the spliceosome assembles around the 5' and 3' splice sites that define the intron to be excised (Figure 2.5, top cartoon). For the exon definition mode of splice site recognition, the unit of recognition is the exon, meaning that initial splice site recognition by the spliceosome occurs across the exon (Figure 2.5, bottom cartoon). In both modes of splice site recognition, a preference for splice site selection that promotes the definition of the smaller initial recognition unit (as defined by the number of nucleotides) is observed. Thus, the proximity of 5' and 3' splice sites within the unit of initial recognition determines preferential splice site selection (Figure 2.5). We therefore conclude that an additional mechanism of alternative splicing can be the proximity of splice sites across the initial unit of definition.

Since the initial concepts of intron and exon definition were introduced, generating supporting evidence for the existence of these two proposed modes of splice site recognition has been challenging. Initial studies were limited to select cases where insights were gained from transfected designer minigenes or *in vitro* transcribed RNAs spliced using the nuclear extract system [14], [53]–[55], [59], [98]. These studies, while mechanistically enlightening, could not be extrapolated to the entire transcriptome.

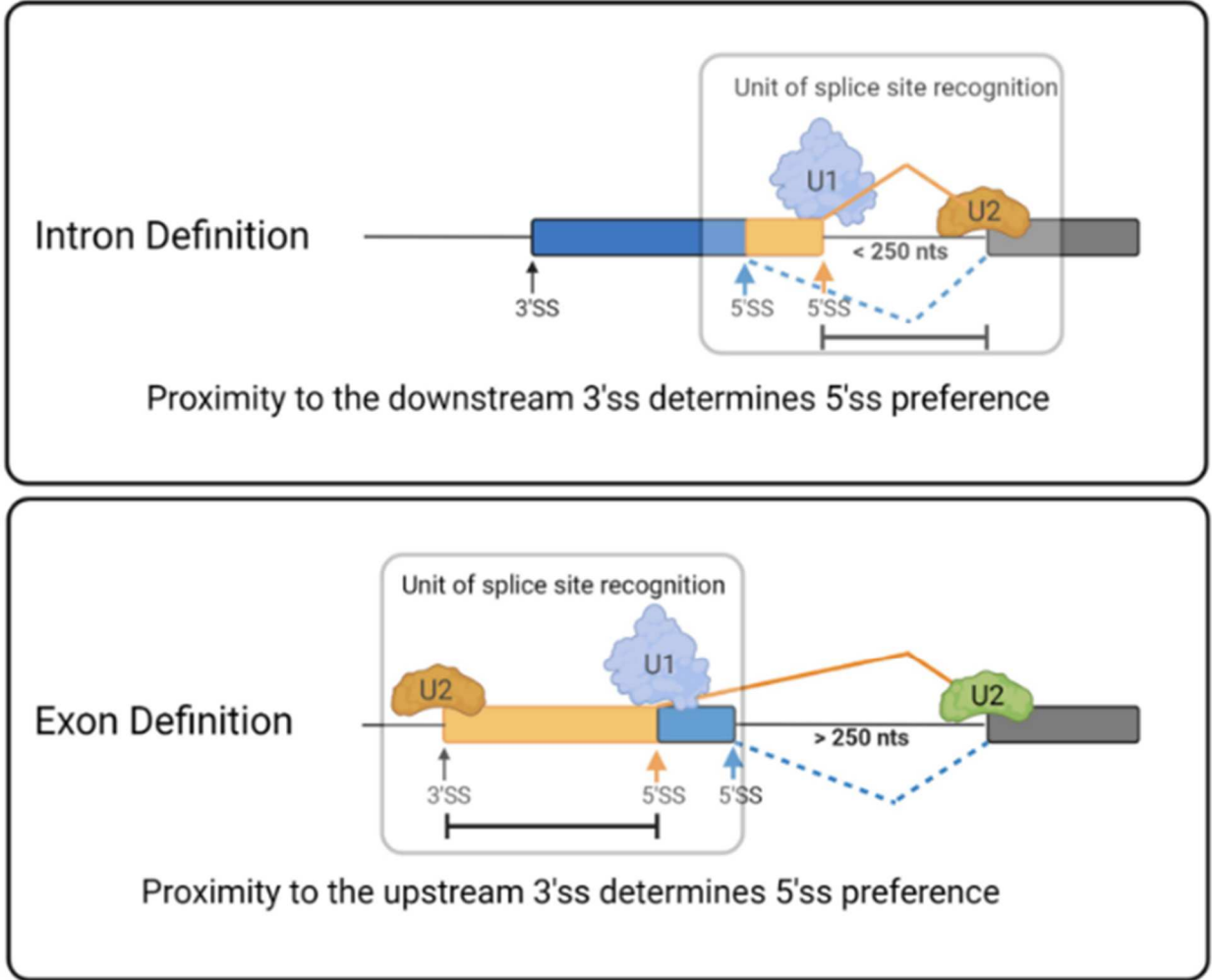


Figure 2.5 Unifying model for the influence of splice site proximity in alternative exon definition.

Depending on the size of flanking introns the splice sites of internal exons are initially recognized across the intron (top – intron definition) or across the exon (bottom – exon definition). In both scenarios, the 5' and 3' splice sites closest to each other across the unit of initial splice site recognition are preferentially selected. Thus, in intron definition 5' and 3' splice sites across the intron are preferentially selected. In exon definition, 5' and 3' splice sites across the exon are preferentially selected.

Recent analyses of *in vivo* splicing kinetics offer more comprehensive insights into the mechanisms of exon recognition. These studies lend support to the notion that exon and intron definition events display different global splicing kinetics. They also raise questions about the generality of exon definition and intron definition [49], [99], [100]. A single molecule intron tracking technique was used to determine the amount of splicing as a function of RNA polymerase II position along the gene. This technique and an orthogonal nanopore-based variation found splicing rates to be strikingly fast in *Saccharomyces cerevisiae* [49] demonstrating that 50% of splicing can be completed 1.4 seconds after 3'ss synthesis for the genes studied. The onset of splicing for a subset of the analysed genes was detected only 26 nucleotides after transcription of the 3'ss. The observation that splicing can be completed before the entire exon is transcribed is consistent with an intron definition mechanism in *Saccharomyces cerevisiae*, but begs the question is exon definition possible in lower eukaryote? The average *Saccharomyces cerevisiae* exon is ~1400 bases suggesting that exon definition would be highly unlikely for those genes where splicing rates were calculated to occur on the order of several seconds [58]. However, a recently proposed unifying model provides evidence for exon definition in *Saccharomyces cerevisiae* [64]. Electron microscopy analyses suggest that the splicing factor Prp40 can bridge the 5'ss bound U1 snRNP and branch point sequence bound BBP/Mud2 (SF1/U2AF65 homologs) either across the intron or across the exon to define E-complex. Structural evidence for exon definition in *Saccharomyces cerevisiae* was supported by genetic and biochemical analysis, which included the circularization of single exon constructs in yeast splicing extracts. The latter study provides strong structural, biochemical, and *in vivo* evidence for exon definition, even in *Saccharomyces cerevisiae*, where most splice sites would be expected to be recognized through intron definition [64].

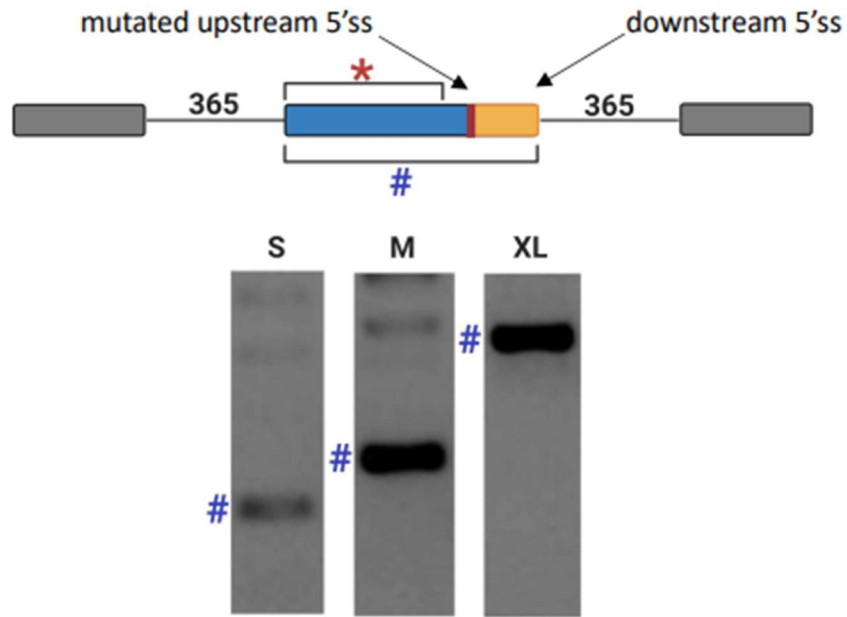


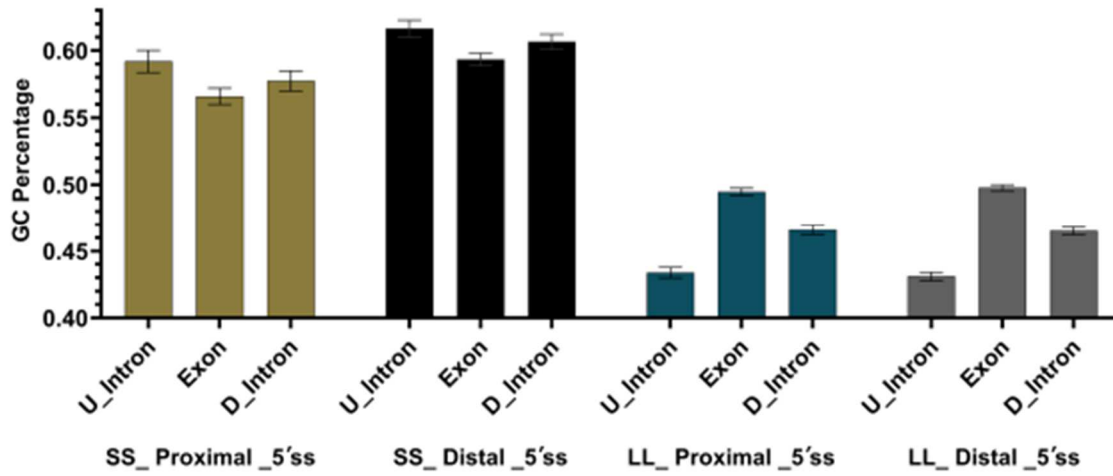
Figure 2.6 The downstream 5' splice site is a functional splice site.

Cartoon and representative image of the splicing outcome of minigenes containing a mutated and non-functional upstream 5' splice site (MaxEnt = -5.2, UCG/gucgau). Bands denoting downstream (purple symbol) or upstream (red symbol) 5' splice site usage are marked to the left. Spliced products were separated using ethidium bromide-stained agarose gels.

Regarding the intron-exon architecture of higher eukaryotes, ligation of 3' adapters and long read nanopore sequencing of nascent RNA were used to determine the splicing rates in *Drosophila* and human cells [100]. The nano-COP method determined that the majority of splicing in *Drosophila* occurs within 2 kilobases, once the 3'ss has been transcribed. This is in contrast to human cells where the majority of splicing is completed ~4 kilobases past the 3'ss [100]. The rate of splicing calculated from nano-COP is consistent with previous $t_{1/2}$ measurements that are ~2 minutes for *Drosophila* and 714 minutes for mammalian cells [99], [101]–[103]. Interestingly, nano-COP found that *Drosophila* introns less than 100 nts in length were spliced more quickly than introns greater than 300 nts, suggesting that intron definition is more efficient than exon definition [100]. These results are supported by an earlier study that used progressive metabolic labelling and also found a local maximum of splicing rates for introns that were 60–70 nts long [99]. However, the latter study also found that a subset of very long introns (>2944 nts) was spliced even more quickly with a $t_{1/2}$ of ~1.5 minutes suggesting gene level and pathway-specific splicing programmes may have evolved to utilize the rapid splicing that very long exon-defined introns undergo. Taken together these kinetic measurements suggest that while exon definition is broadly less efficient and intron definition is broadly more efficient as was first shown by Fox-Walsh and Hertel [59], exceptions do exist.

Recent investigations provide further support that both intron definition and exon definition occur *in vivo* [104], [105]. However, these studies present evidence that the mechanism by which splice sites are initially recognized is dictated by the difference in GC content, referred to as GC differential, between the exon and the flanking introns. Specifically, two architectures are described, referred to as the 'differential architecture' and the 'leveled architecture' [104], [105]. 'Differential architecture' exons have a low GC content, and their

A



B

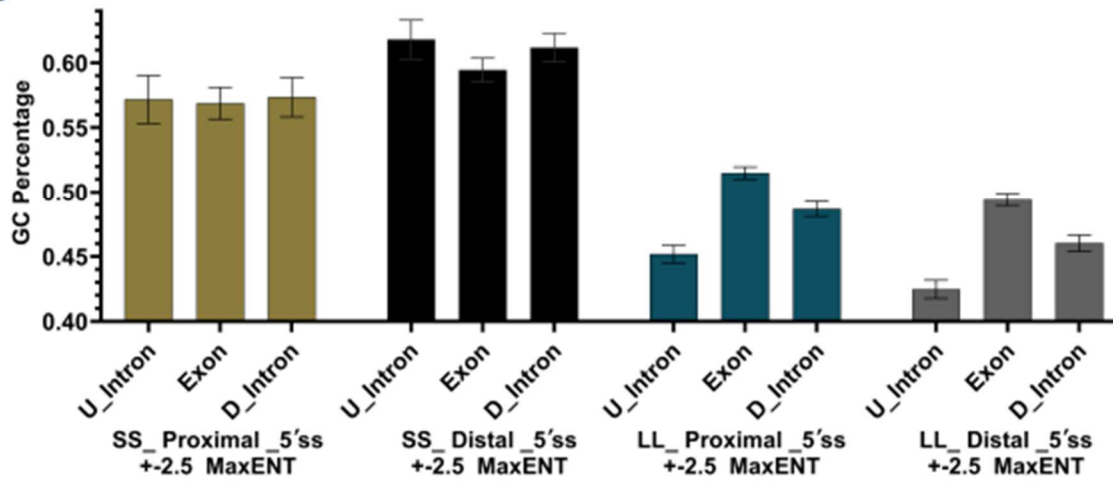


Figure 2.7 GC content distribution for intron definition splice sites (SS) and exon definition splice sites (LL) based on the intron length-dependent classification used herein (SS<250 nts, LL>250 nts).

(A) The data summarized in the graphs represent all alternatively spliced SS and LL 5' splice sites captured by ALTssDB. The GC content as defined by Amit et al [104] is displayed for intron definition (SS, forest green and black) and exon definition (LL, crimson blue and grey) events. GC content for each architectural class was computed for the exon and intron when the exon/intron junction is defined by the major 5' splice site (SS, forest green and LL, crimson blue) or the minor 5' splice site (SS, black and LL, grey) of the alternatively spliced exon. The comparison between SS and LL architectural definitions shows a striking GC content difference as was observed by Amit et al [104]. The profiles do not change significantly when the major or the minor 5' splice site is used. (B) The data summarized in the graphs represent alternatively spliced SS and LL 5' splice sites with near equal splice site scores (± 2.5 MaxEnt). Bar graph definitions are as described in (A). Filtering the alternative 5' splice site analysis by near equal splice site scores does not significantly change the outcome. The comparison between SS and LL architectural definitions highlights GC content difference as was observed by Amit et al [104] and the profiles do not change when the major or the minor 5' splice site is used.

flanking introns have an even lower GC content. The 'leveled architecture' exons are characterized by a high GC content, less difference in the GC content of flanking introns and short introns. The former class of exons was demonstrated to be localized to the nuclear periphery and recognized through exon definition while the latter was shown to be localized to the nuclear centre and recognized through intron definition. Altering the GC content between exon and the downstream intron can be used to alter the mode splice site recognition, without changing the length of the intron. These observations suggest that intron length may not be the determining factor in the mode of splice site recognition for exon definition [104], [105]. It will be important for future exon definition studies to consider the GC content across the exon and flanking introns. For example, a recent analysis of high-throughput mutagenesis data for an alternatively spliced exon in the proto-oncogene *RON* demonstrated that the alternatively spliced exon is recognized through exon definition, even though it is flanked by short introns on either side (87 and 80 nts) [60]. Thus, splice sites of short introns can be recognized through exon definition, perhaps because the unique GC content that typifies exon definition splice sites.

Finally, a recent transcriptome-wide study demonstrated that introns that undergo efficient co-transcriptional splicing have sharp structural transitions across the intron-exon boundary [47]. These introns display a peak of RNA structure downstream of the 5'ss and upstream of the 3'ss. Furthermore, some introns displayed enhanced co-transcriptional splicing under conditions where the elongation rate of RNA polymerase II was slowed down genome-wide, a process that promotes increased RNA folding. The latter group of introns had significantly steeper structural transitions when transcription was slow [47]. GC content is an indicator of the potential to form RNA secondary structures [46]. Thus, it may be the case that

the differential architecture associated with exon definition is driven partially by the propensity for RNA secondary structure formation that can help delineate the intron-exon boundary.

We set out to determine the degree of agreement between the intron length-dependent definitions of ‘intron-defined’ and ‘exon-defined’ splice sites with the ‘leveled’ and ‘differential’ architecture. We calculated GC content differentials between the LL and SS architectural classes of alternatively spliced 5’ and 3’s exons. Remarkably, the intron length defined LL and SS categories closely resemble the ‘differential’ and ‘leveled’ architectures respectively [104], [105] (Figure 2.7). Thus, the GC content, as defined the Amit et al. [104], of long introns (>250 nts) differs significantly from the GC content of short introns (<250 nts), suggesting that GC content or intron size definitions are comparable approaches to define exon and intron definition modes of splice site recognition. This notion is supported by evolutionary analyses that show the emergence of a distinct differential GC architecture as intron lengths increased through vertebrate evolution [65]. Thus, the emergence of the ‘differential architecture’ may be a co-evolutionary adaptation to define exons in the context of expanding introns. To evaluate whether the use of proximal or distal splice sites changes ‘leveled’ and ‘differential’ architecture designations, we calculated GC content for alternatively spliced exons captured by ALTssDB. Interestingly, the resulting GC differential does not change significantly (Figure 2.7), suggesting that alternative splice site selection is not dependent on differential GC content but contingent on defining the smallest unit of initial recognition.

Collectively, the results of our transcriptome-wide analysis of alternative splice site usage provide evidence that exon definition and intron definition do occur transcriptome-wide. When exons are flanked by long introns, the spliceosome tends to favour splice sites located internally within the exon being defined. By contrast, the spliceosome tends to move into the intron for

splice site definition for exons flanked by short introns. These observations suggest that the spliceosome can define the exon and the intron independently.

Our computational analysis of alternative 3'ss events permitted an evaluation of alternative 3'ss selection in the context of first or second step recognition. Initial 3'ss selection is mainly driven by the strength of the polypyrimidine tract and the presence of a consensus branch point. Upon recruitment of U2 snRNP to the branch point and tri-snRNP incorporation, the first step of the splicing reaction is initiated without engaging the 3'ss junction. After spliceosome rearrangements, the 3'ss junction is selected during the second step of splicing as the spliceosome aligns the AG/N intron/ exon junction into the active site. It is well established that competing 3'AGs in close proximity (less than 9 nts) use the same upstream polypyrimidine tract and branch point and that their selection is directed during the second step of splicing [28]. Interestingly, our analysis of alternative 3'ss selection in close proximity demonstrated that the upstream AG/N junction is almost exclusively chosen over the downstream AG/N. Thus, it appears that aligning the closest AG/N 3'ss junction is the default pathway of second step splice junction selection (Figure 2.4b).

The intron-exon architecture of genes is a major driver of splice site selection. Since the initial postulation of these two modes of splice site recognition, various forms of evidence have been presented, often in form of kinetic principles supporting intron or exon definition. However, measurements of splicing rates as a function of intron length do not constitute direct evidence of alternative spliceosomal assembly pathways. The ability of yeast E-complex to assemble across the intron or the exon is perhaps the strongest evidence yet for exon definition. Our study provides support for exon definition by demonstrating the spliceosome favours internal splice sites within exons when the splice site strengths of competing sites are comparable. This

suggests that the exon is being defined and not the intron. This study provides a unifying model for splice site selection, whereby the spliceosome assembles across the smallest unit of initial splice site recognition. In the case of intron definition, this entails removal of smaller introns and inclusion of larger exons. Indeed, studying the evolutionary trends in intron-exon architecture, lower eukaryotes tend to have larger exons and smaller introns. Upon intron expansion and a gradual shift towards exon defined gene architecture, the initial unit of splice site recognition often tends to be the exons. This may be due to the increased number of decoy splice signals associated with larger genome sizes. It would therefore be expected that the smaller exons would be favoured in higher eukaryotes. This trend is also broadly observed from yeast to humans. It is possible that the exon-centric proximity rule is an evolutionary adaptation to accurately recognize exons surrounded by long stretches of intronic sequence. Our results not only provide *in vivo* and transcriptome-wide evidence for exon definition, they also demonstrate that exon and intron definition influence alternative splicing in the context of alternative 5' or 3' splice site competition.

Methods

Construction of ALTssDB

ALTssDB was created using EST data from the Human Exon splicing Events (HEXEvent) database [95]. HEXEvent contains information regarding the location of competing splice sites, the resulting exon sizes, alternative splice site usage levels and the gene associated with each mRNA. The HEXEvent data was filtered to obtain a dataset comprising of only pairs of competing 5' and 3' splice sites separately. This database was subsequently modified to include splice site junction information and splice site strength scores using MaxEntScan [30]. Although other approaches exist to evaluate the strength of 5 splice sites [25], [106],

MaxEntScan is the preferred tool as it also permits comparable splice site score derivation for 3 splice sites. Using an R script and IntronDB dataset, (a database detailing eukaryotic intron features) flanking intron lengths were added to the database [96]. Alternative splicing events were further filtered to include only events that have 10 or more EST counts. The data was filtered into four categories according to intron length and included: both flanking introns around the exon of interest being short (<250 nts, SS), both flanking introns being long (>250 nts, LL), the upstream intron being short and downstream intron being long (SL) or the upstream intron being long and the downstream intron being short (LS). ALTssDB does not differentiate between canonical U2 introns and U12type introns. Given their rarity and limited involvement in alternative splicing beyond intron retention, it is anticipated that U12-type introns are not well represented in ALTssDB [107].

ALTssDB does not distinguish between isoforms that originate from a tissue specific splice switch or disease comparison. It lists all known splice patterns for a particular exon, independent of origin. EST data was used to build AltssDB to obtain high enough numbers of alternative splice site choices within the human genome to carry out all analyses. While datasets for tissue-specific splicing are available, the quantity of significant alternative splice site events is limiting.

Plasmid design

Five minigene constructs were designed containing three exons and two introns. The plasmid design was based primarily on previously validated constructs used to study splice site strength [31]. The internal exon was designed to contain two functional competing 5' splice sites (CAG/guaagu), with equal MaxEnt scores MES of 10.9, separated by 52 nucleotides. The sequence preceding the upstream 5' splice site was progressively shortened (Figure 2.1b).

Additional constructs were created where the MES of both competing 5' splice sites were changed from 10.9 to -0.5 (GAG/guguca) for S, M, L, and XL plasmid. Lastly, the upstream 5' splice sites were changed from MaxEnt = 10.9 to MaxEnt = -5.2 (UCG/gucgau) for the S, M, and XL to show that the downstream 5'ss is viable (Figure 2.6).

Cloning protocols to change splice site strength sequences

To linearize the plasmids, 10 nanograms (ng) of plasmid DNA obtained by midiprep was amplified using divergent primers. PCR reactions were carried out with NEB® Phusion® polymerase in 50 μ L according to NEB protocols. DH5 α E.coli midiprep derived plasmids in the PCR reaction were digested with 40 units of DpnI according to NEB protocols. Plasmids were purified with Zymo DNA clean and concentrator™ kit and DNA concentrations were obtained using a nanodrop 2000 instrument. For each construct, 0.03 picomoles of linearized plasmid DNA was mixed with a 10X molar ratio of phosphorylated double stranded DNA inserts, purchased from IDT, in 20 μ L ligation reactions. Synthetic inserts were cloned into linearized vectors using T4 ligase according to the standard NEB protocol and 10 μ L of the ligation reaction was transformed using in house DH5 α E.coli cells. Colonies were screened using PCR to detect the correct size insert. Colonies with the correct size insert were grown from 3 mL cultures to 20 mL cultures and underwent midiprep DNA extraction. Plasmid DNA from each colony was sequenced to ensure the correct orientation of inserts.

Cell transfections and RT-PCR Analysis

Transfection experiments were performed in triplicate using HeLa cells. 1 mL of 0.1×10^6 cells/mL was plated into each well of 12 well plates. Cell confluency was checked 24 hours later and 1 μ g of plasmid DNA was transfected according to Bioland Scientific's BioT protocol.

Cells were harvested 48 hours post-transfection. Each well was washed two times with phosphate buffered saline (PBS) and subsequently RNA was extracted with the standard Trizol™ protocol. RNA pellets were resuspended in 50 µL water and put through ZYMO RNA Clean and Concentrator™ columns. Sample volumes were adjusted to 80 µL, yielding RNA concentrations of ≤ 200 ng/µL. DNase digestion was performed with Turbo™ DNase (Ambion®) according to Ambion's protocol in 100 µL reactions. RNA was subsequently extracted with 100 µL phenol: chloroform and the aqueous phase was put through ZYMO RNA Clean and Concentrator™ columns. DNase digested and purified RNA samples were resuspended in 25 µL. A nanodrop 2000 instrument was used to obtain RNA concentrations. Reverse transcription reactions were carried out in 20 µL using 250 ng of total RNA and 200 ng of OligodT18 primer according to SuperScript™ III protocol. PCR primers are as followed: first exon forward primer (5'cggtcgtcctcactctcttc3') and third exon reverse primer (5'agatccccaaggactcaaaga3'). PCR primers were designed that bound the flanking exons and thus would detect upstream, proximal or downstream, distal 5' splice site usage.

PCR reactions contained 5 µL cDNA (10% vol:vol), 0.2 mM dNTPs, 0.2 µM of each primer, 1.5 mM MgCl₂ and 0.25 units taq polymerase (Apex Bioresearch). Semi-quantitative PCR using long extension times to limit PCR product size bias was carried out to demonstrate that the ratio of upstream and downstream splice site usage, or the alternative exon skipping pattern, remained constant throughout the dynamic linear range of the amplification reaction (data not shown). Based on these results 25 cycles of PCR were performed for each sample and 5 µL was subsequently loaded onto a 2% agarose gel and stained with ethidium bromide. Agarose gels were run at 150 V for 1 hour in 1X Tris-Borate EDTA (TBE).

Calculating splice site selection preference

5'ss selection preference was determined by calculating the log ratio of the number of splice site events preferring the downstream 5'ss over the upstream 5'ss. 3' splice site selection preference was determined by calculating the log ratio of the number of splice site events preferring the downstream 3'ss over the upstream 3'ss.

CHAPTER 3

Genome-wide determination of mRNA and exon half-lives

Summary

Dysregulation of gene expression often associates with disease, a state that is generally evaluated by comparing the steady state levels of mRNA using transcriptomics. A more insightful approach would consider the dynamic nature of mRNA expression. The level of mRNA expression can be influenced by transcriptional activity, mRNA stability, or a combination thereof. This chapter focuses on determining the kinetics of mRNA degradation. In addition, this chapter considers the impact of alternative splicing on mRNA stability. We carried out a 24-hour 4sU pulse/chase experiment in HepG2 cells. RNA-seq was performed on the metabolically labeled mRNAs to derive canonical mRNA, individual exon, and mRNA isoform half-lives. Our analysis allowed us to identify a positive relationship between gene/exon/mRNA isoform length and mRNA half-life. Additionally, our 4sU seq pipeline allowed us to identify exons that are kinetic outliers to their neighboring exons. These outlier exons are part of unique mRNA isoforms most likely generated through alternative splicing; they tend to be larger in size, and they have less sequence and size conservation when compared to their standard exon counterpart. Our studies have introduced a new method pipeline to study the interconnections between alternative splicing and mRNA stability.

Introduction

Establishing appropriate gene expression programs is crucial for proper cellular function. The most common approach to evaluate aberrant expression programs is by carrying out mRNA-sequencing, comparing the steady-state expression patterns of a normal cell vs an aberrant cell. Such steady-state studies can be thought of as observing gene expression through a static lens or a snapshot. Traditional steady-state analysis can only conclude if a gene is being upregulated or downregulated relative to another cell line's expression. While these studies can be insightful, they offer a limited amount of information on the dynamic nature of gene expression. There are two main contributors to mediating changes in gene expression. The first is RNA synthesis by RNA polymerase. Increased transcription rates lead to increased accumulation of mRNAs in the cell. Reduced activity of RNA polymerases decreases mRNA levels. Integrated in mRNA synthesis is the co-transcriptional process of pre-mRNA processing. Steady-state studies are unable to capture pre-mRNA processing rates. The second contributor to gene expression, often overlooked, is mRNA stability. A shorter mRNA half-life leads to less translation of that mRNA transcript, and increased mRNA stability can elicit increased protein output from a single mRNA. One aspect of mRNA degradation that is not well understood is the relationship between mRNA degradation and alternative splicing. Alternative splicing is a regulatory mechanism that leads to the creation of multiple mRNA isoforms from a single gene through the inclusion or exclusion of different exons. This can generate different mRNA isoforms with unique 5' and 3' untranslated regions (UTRs) varying in coding sequence or length. These differences can affect mRNA stability in many ways. For example, alternative splicing leading to a frameshift in the resulting mRNA often activates nonsense mediated decay (NMD). Thus, stabilizing and

destabilizing sequence elements can be included in the final mRNA based on which isoform is being expressed [108].

Studying gene expression through a dynamic lens of an RNA from “birth to death” will provide a better understanding of gene expression control, thus providing insights into underlying expression mechanisms. The lessons learned from this dynamic view of gene expression can in principle be applied to other cells or disease states to understand regulatory circuits and to potentially devise new therapeutic approaches to combat disease. Our lab has recently conducted a set of RNA 4-Thio-Uridine (4sU) metabolic labeling next generation sequencing (NGS) experiments to capture and study transcription rates and splicing rates (Garibaldi et al. unpublished). Previous mRNA degradation studies used transcriptional inhibitors such as actinomycin-D and DRB [109], [110] to track transcribed RNAs. However, such global transcription arrest approaches are invasive to the cell. In addition, these inhibitors have been found to alter the stability of certain mRNAs [109]. While metabolic labeling methods such as 5-Ethynyluridine or 5-Bromo-Uridine can provide an alternative to transcription arrests, they have been found to be toxic over prolonged usage [109]. Thus, previous RNA degradation studies relied on short metabolic labeling pulsing and indirectly deriving degradation rates from the transcription rates. By contrast, the uridine derivative 4sU can effectively enter the cell and is tolerable for prolonged periods, thus allowing us to directly capture mRNA degradation rates [110], [111].

This chapter establishes a method using the 4sU metabolic labeling technology to study mRNA degradation and the impact alternative splicing may have on mRNA stability. We explored the relationship between mRNA stability and sequence length and how the splicing of

certain exons can impact that relationship. Additionally, we investigated the relationship between exon sequence and length conservation with RNA stability.

Results

Experimental design to determine mRNA degradation rates

To obtain transcriptome-wide mRNA half-lives we designed a pulse-chase experiment using the uridine analog 4-thiouridine (4sU) to metabolically label nascent RNA (Figure 3.1a). HepG2 cells were pulsed for 24 hours in media spiked with 40 μ M 4sU. This pulse period took place over 24 hours to reach steady-state levels of 4sU labeled mRNAs transcriptome wide. After the pulsing period the 4sU media was replaced with fresh cell media lacking 4sU to initiate chase conditions. Total RNA was extracted at time points 0hr, 1hr, 3hr, 6hr, 9hr, 12hr, 18hr, and 24hr after the addition of the fresh media. The labeled RNA was then isolated via thiol-specific biotinylation and a streptavidin column pulldown. The 4sU labeled RNA was used to create a poly-(A) selected cDNA library for sequencing. Each timepoint was spiked with an equal amount of an *in vitro* transcribed 4sU labeled Yeast Act1 mRNA prior to performing the pulldown to control for pulldown efficiencies and to permit normalization of the sequencing data. The sequencing data for each time point was aligned to genes or exons and the resulting counts were normalized to the spike-ins (Figure 3.1a). The time dependent series of read counts was used to create a degradation profile based on first-order decay kinetics according to $[mRNA] = [mRNA]_0 \cdot e^{-kt}$, where $[mRNA]_0$ is the initial abundance of the mRNA and k is the observed rate of mRNA degradation. Corresponding half-lives were calculated using $t_{1/2} = \ln(2)/k$ (Figure 3.1b).

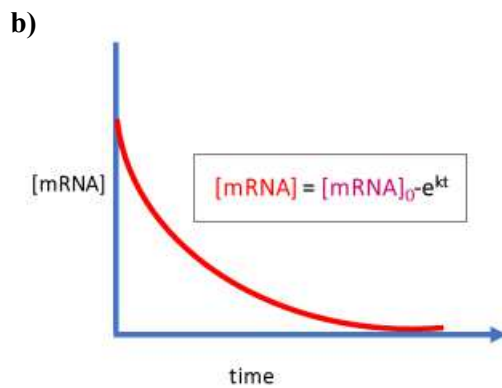
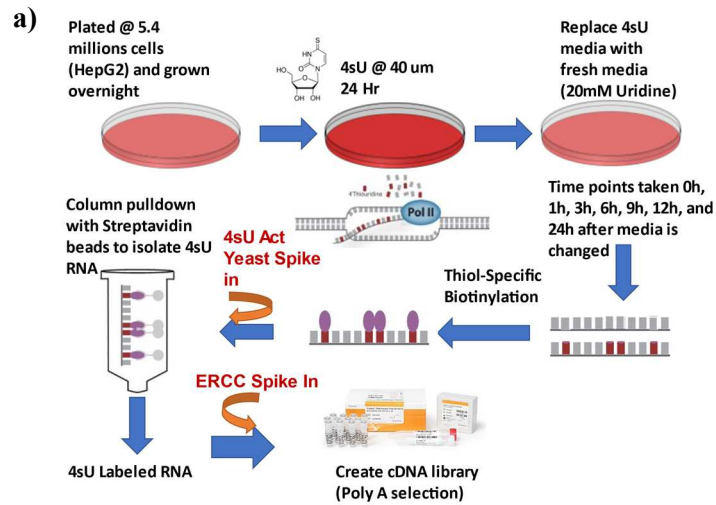


Figure 3.1 Experimental Design.

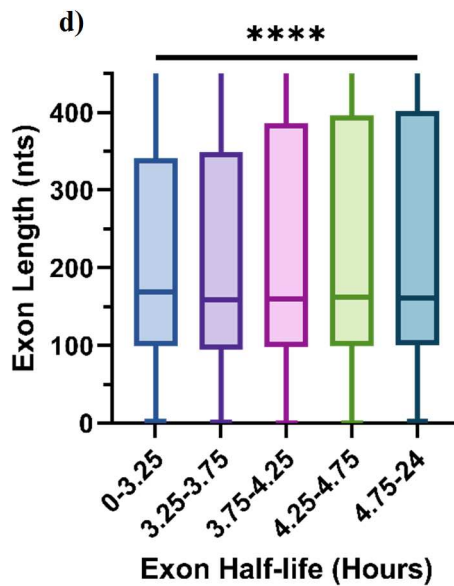
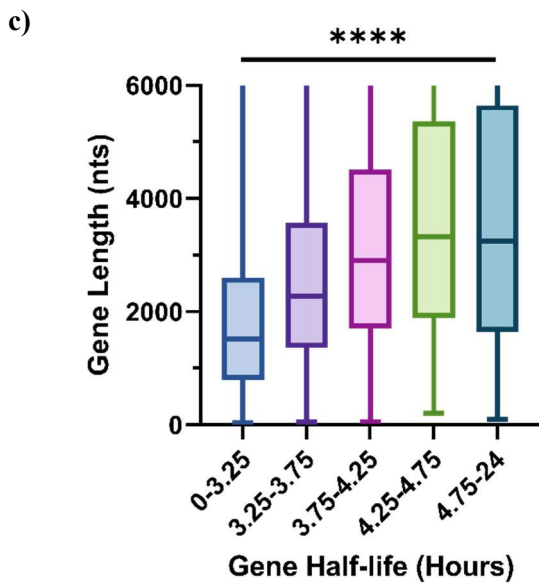
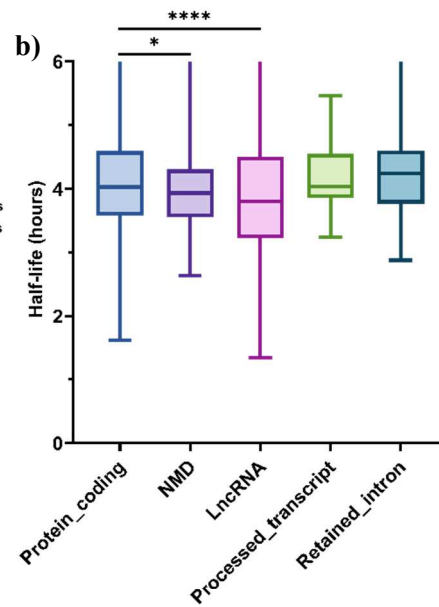
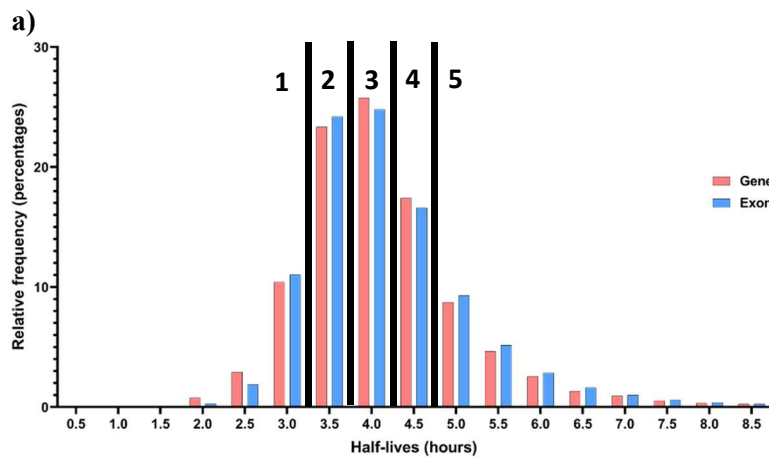
(a) 4sU pulse chase experimental design. (b) First-order decay profile and equation to derive mRNA degradation rates.

Gene and Exon Half-lives in HEPG2 Cells

The sequencing data was separately aligned to the canonical gene sequence and to all known exon sequences. This allowed us to obtain the canonical gene degradation rate and the degradation rate of every expressed exon. The derived gene and exon degradation rates were passed through a filter only keeping rates whose 1-hour timepoint has at least ≥ 50 normalized counts and a time series regression fit of R squared value ≥ 0.60 . Using these filters, 20,443 gene degradation profiles were captured. 65% of the genes captured are protein coding, 20% are long non-coding RNAs, and the remaining 15% are made up of other non-coding RNA. Our focus was on protein coding and long noncoding RNAs, which make up 17,404 gene degradation profiles. The average half-life of the mRNAs captured is 4.16 hours. Using the same filters, the exon analysis resulted in the capture of 264,963 exon half-lives with an average exon half-life of 4.21 hours. These observations correspond well with the estimated half-life range from recent studies using a similar technique [112] (Figure 3.2a).

We first asked if there is a difference in mRNA stability based on GENCODE transcript type classification [113]. The transcript types captured in the gene degradation profile include protein coding, non-sense mediated decay (NMD), long non-coding RNA (lncRNA), processed transcripts (unclassified transcripts that do not contain an ORF), and retained introns (Figure 3.2b). As expected, the NMD transcript degradation profile displayed a slightly lower, yet significant, mRNA mean half-life. The same is also true for lncRNAs when compared to protein coding RNAs.

To evaluate the difference between fast- and slow-degrading mRNAs, the data of all expressed genes and all expressed exons independent of transcript type was parsed into 5 equally weighted bins based on gene or exon half-lives. The bin boundaries are shown as black lines in



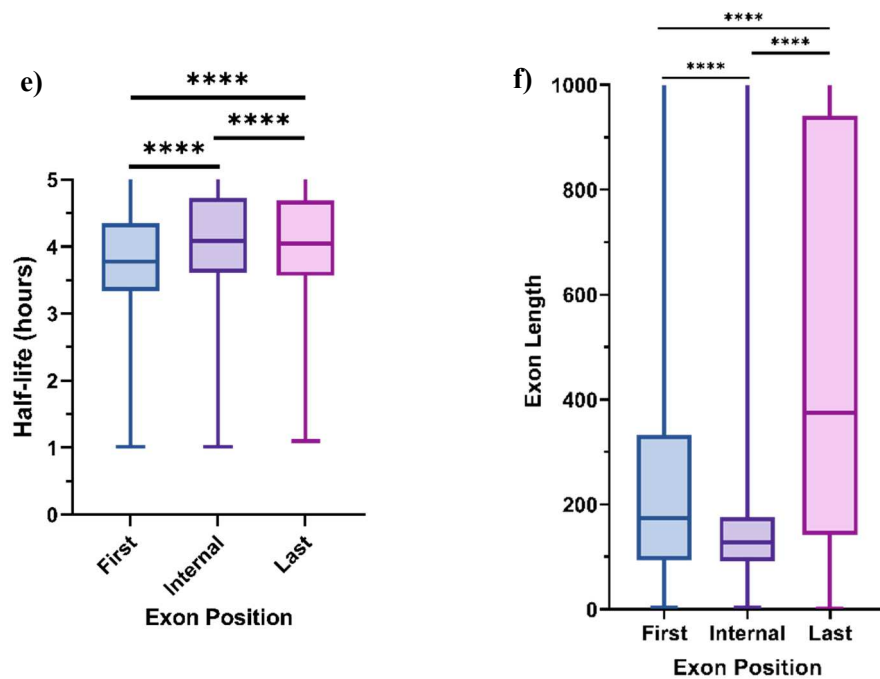


Figure 3.2 Relationship between sequence length and mRNA half-life.

(a) Gene and exon half-life histogram. The black bars represent population-weighted bins for downstream analyses. (b) Mean gene half-life compared across different RNA transcript types. (c) Bar plot analysis correlating gene length across 5 different gene half-life bins. (d) Bar plot analysis correlating exon length across 5 different exon half-life bins. (e) Bar plot analysis correlating exon half-life with exon positions. (d) Bar plot analysis correlating exon length with exon positions. (**** $P \leq 0.0001$ & * $P \leq 0.01$)

Figure 3.2a. Bin 1 consists of the fastest decaying genes or exons and subsequent bins are progressively more stable. Using this bin data, we assessed the relationship between mRNA length and mRNA half-life. We found a positive correlation between gene length and mRNA half-life. Fast decaying RNAs are characterized by a smaller gene length (Figure 3.2c). This is surprising as a previous study has implied a negative correlation between mRNA length and mRNA half-life [114]. Next, we evaluated the relationship between exon half-life and exon length. As was seen for the gene analysis, a positive correlation between the average exon length and exon half-life is observed (Figure 3.2d).

Published literature suggests that mRNA degradation occurs primarily 5' to 3' via XRN exoribonucleases [69]. Interestingly, our analysis shows that the first exon position is on average less stable when compared to the internal and last exon positions (Figure 3.2e). The last exon is also significantly longer than the first exon (Figure 3.2f). Together, these results support the notion that longer exons and transcripts are more stable.

The first and last exons define the 5' and 3' UTR. These untranslated regions have been shown to provide binding sites for trans-acting factors involved in regulating mRNA stability [115]. To determine the relationship between exon length and exon half-life for each exon type, we performed a binning analysis of the first, internal, and terminal exons (Figure 3.3). The half-lives of first and last exons are more impacted by exon length than internal exons, displaying a strong positive relationship between exon length and exon half-life (Figure 3.3a, b, e & f). Internal exons do not display a meaningful correlation between exon length and exon half-life (Figure 3.3c-d). These results demonstrate that the length of the 5' and 3' UTR is important in mediating mRNA half-life.

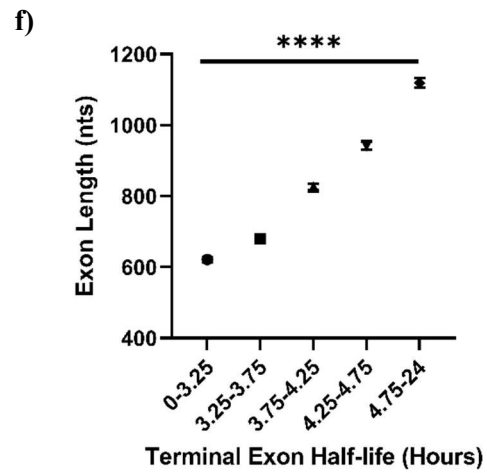
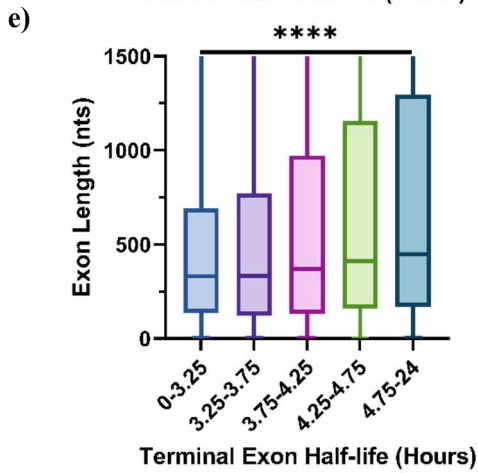
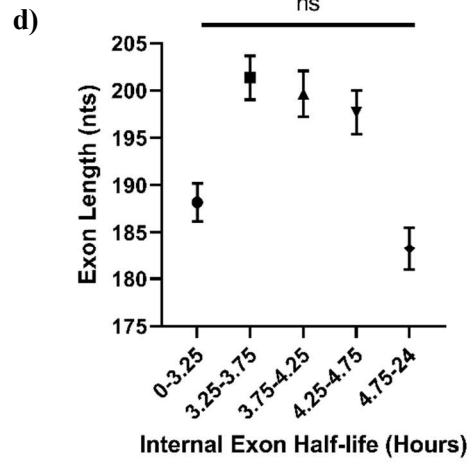
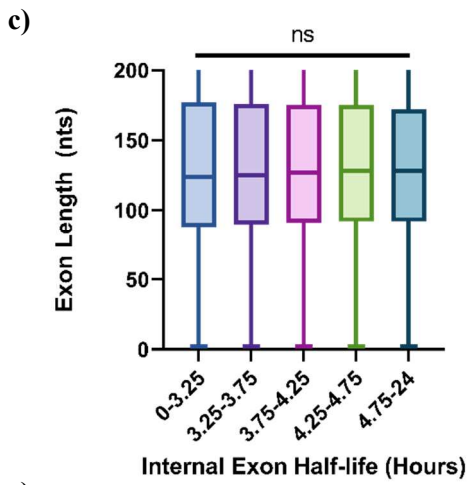
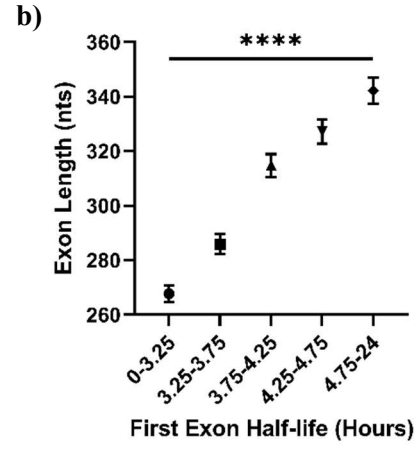
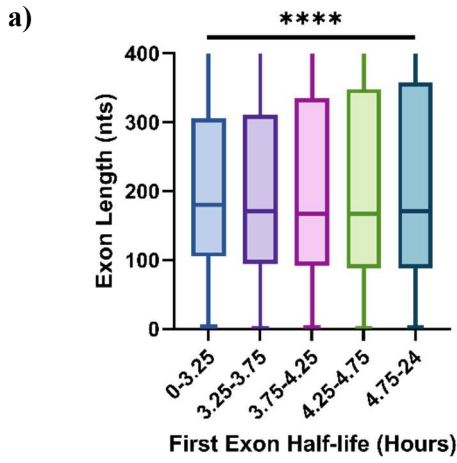


Figure 3.3 The influence of exon length for terminal exons half-lives.

(a) Bar plot analysis correlating exon length with binned first exon half-lives. (b) Means scatterplot of binned first exon half-lives as a function of exon length. (c) Bar plot analysis correlating exon length with binned internal exon half-lives. (d) Means scatterplot of binned internal exon half-lives as a function of exon length. (e) Bar plot analysis correlating exon length with binned terminal exon half-lives. (f) Means scatterplot of binned terminal exon half-lives as a function of exon length. (**** $P \leq 0.0001$ & not significant)

Outlier Exons Tend to be Larger in Size

Alternative splicing is critical for proper gene expression. Previous transcriptome-wide studies in lower eukaryotes have highlighted the influence of alternative splicing on mRNA stability, which has yet to be explored in mammalian cells [108]. One approach to evaluate the impact of alternative splicing on mRNA stability is to identify exons that display significantly different stability kinetics when compared to the rest of the exons within a gene. Combining the data from the gene alignment and exon alignment allowed us to identify outlier exons by comparing the canonical gene degradation half-life to each individual exon half-life. An exon was defined as an outlier exon if the exon half-life deviates two standard deviations from its overall gene half-life. Exons within the two standard deviation range were considered standard exons. Of the 264,790 exons captured we identified 22,084 outlier exons. While there is no significant difference in average half-life between the outlier exons and the standard exons (Figure 3.4b), their half-life is more widely distributed when compared to the standard exon's half-lives (Figure 3.4a, b). A comparison of exon length revealed that the outlier exons are significantly larger in size than standard exons (Figure 3.4c). To further investigate the size difference in outlier exons, we parsed both exon datasets into more or less stable outlier exons when compared to its corresponding gene half-life (Figure 3.5a). Interestingly, outlier exons are typically larger in size than standard exons regardless of whether they have a faster or slower half-life (Figure 3.5b and Table 3.1), although more strikingly so for the more stable outlier exons. 57% of the outlier exons are of the more stable type, whereas 43% are of the less stable outlier type. This data suggests that the longer an exon is, the more likely it is to be an outlier exon at both ends of exon stability.

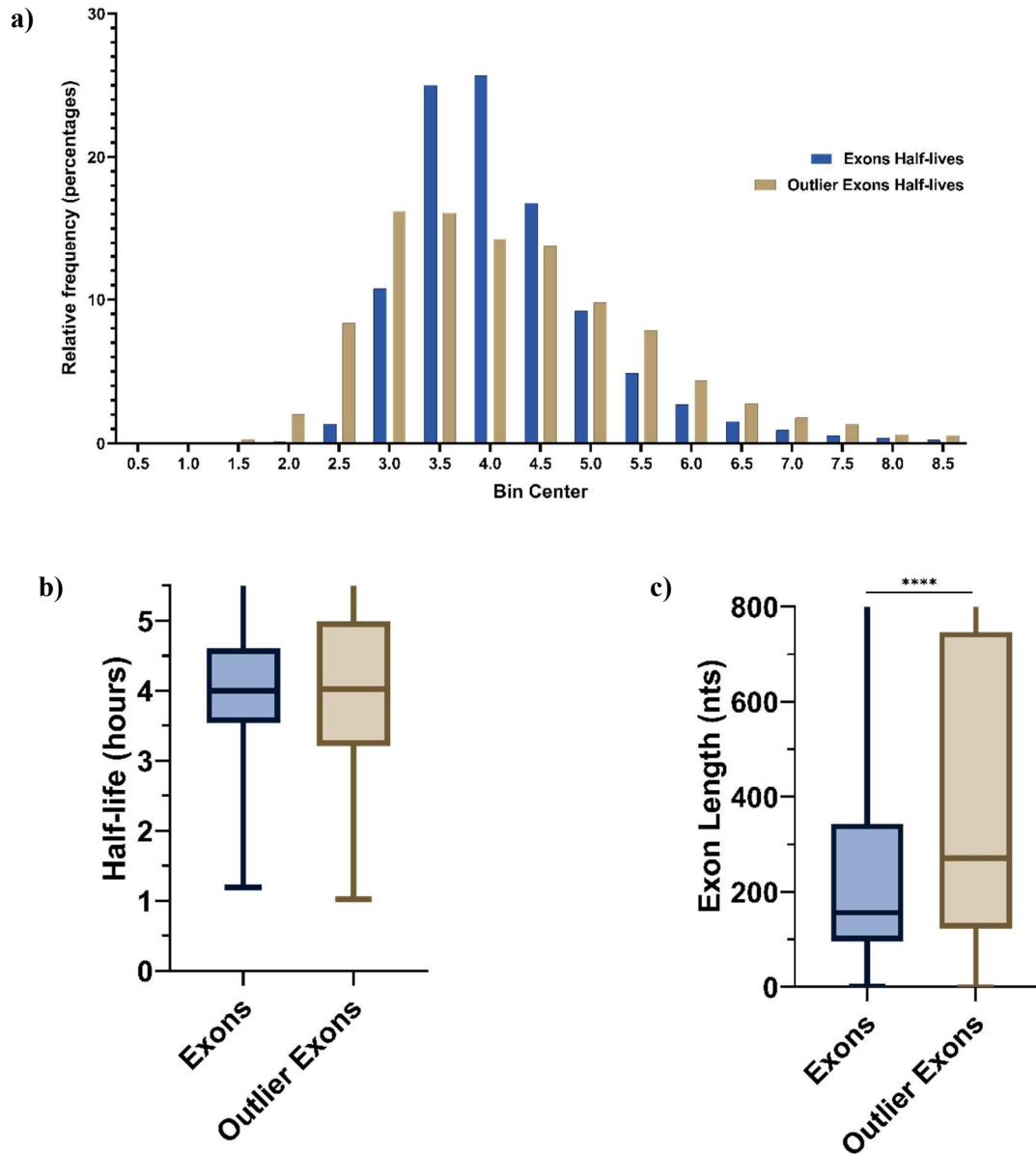


Figure 3.4 Comparison between standard and outlier exons.

(a) Standard and outlier exon half-life histogram. (b) Bar plot analysis correlating standard and outlier exons with half-lives. (c) Bar plot analysis correlating standard and outlier exons with exon length. (**** $P \leq 0.0001$).

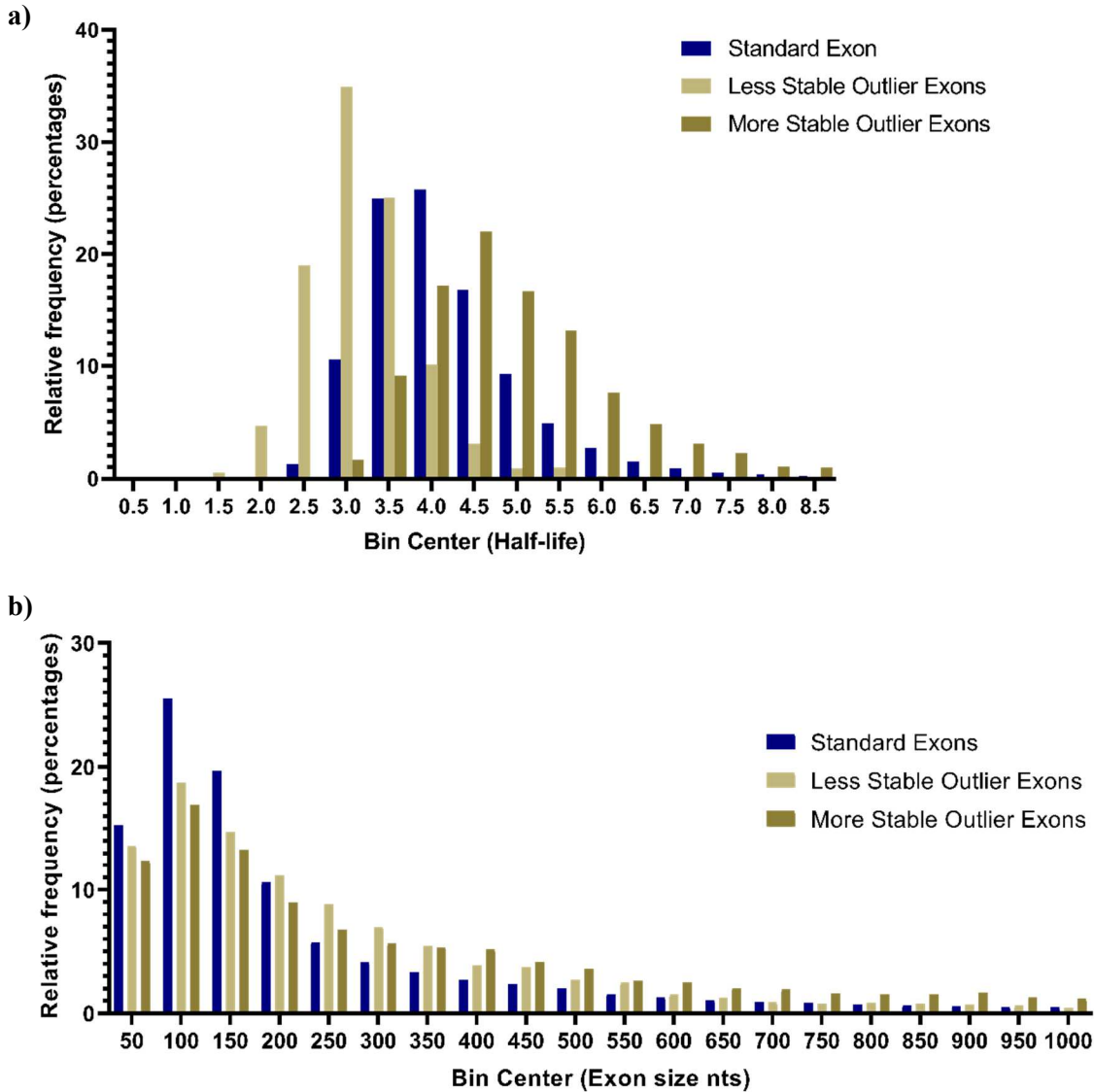


Figure 3.5 Half-life distribution of standard, more stable outlier exons, and less stable outlier exons.

(a) Exon half-life histogram. (b) Exon length histogram

Exon Type	Average Half-life (Hours)	Exon Length (nts)
Standard Exon	4.21	403.76
Less stable Outlier Exons	3.18****	421.33****
More stable Outlier Exons	5.07****	1101.86****

Table 3.1 Outlier exons are larger in size when compared to standard exons regardless of outlier stability.

Mean exon half-life and mean exon length analysis. T-test was performed against standard exon half-life and length (**** $P \leq 0.0001$)

Most efforts aimed at identifying regulatory elements affecting mRNA stability have focused on 3' UTR sequence features and the length of an mRNA [50], [83], [116]. We therefore compared the exon half-life difference between standard and outlier exons based on exon type (first, internal, last). This analysis demonstrated that the average length of outlier exons is significantly higher than that of standard exon for every exon type (Table 3.2). We conclude that outlier exons are generally longer in length than standard exons.

Isoform Analysis

The outlier exon analysis provided some insight into the role alternative splicing plays in mRNA stability. A complementary analysis is an mRNA isoform quantification of the 4sU HepG2 dataset. This allows us to determine which mRNA isoforms are most prevalent and what sequence features may play a role in their stability. We used SALMON as a computational tool to investigate mRNA isoforms. SALMON is a fast aligner designed for full-length isoform quantification from bulk mRNA sequencing. Keeping the same cutoff parameters used for the gene and exon count, we computed the degradation profiles for 39,834 putative mRNA isoforms. Using the canonical gene degradation profiles as reference we identified 18,835 mRNA isoforms that display half-lives that fall 2 standard deviations from the gene degradation half-life (Figure 3.6a). The outlier isoforms are characterized by a significantly lower half-life than that of a standard isoform (Figure 3.6a, c). In addition, the outlier isoforms have a significantly lower transcript length than that of the standard isoforms (Figure 3.6b). To further investigate the relationship between the standard isoforms and outlier isoforms, the isoform dataset was binned into less or more stable isoforms relative to the overall gene degradation profile. A comparison

Exon Type	Average Half-life (Hours)	Exon Length (nts)
Standard First Exons	4.00	295.26
Less stable Outlier First Exons	3.24****	326.98****
More stable Outlier First Exons	4.80****	575.79****
Internal Exons		
Standard Internal Exons	4.31	190.81
Less stable Outlier Internal Exons	3.11****	243.66****
More stable Outlier Internal Exons	5.25****	267.29****
Terminal Exons		
Standard Terminal Exons	3.98	307.61
Less stable Outlier Terminal Exons	3.16****	738.98****
More stable Outlier Terminal Exons	5.09****	1598.76****

Table 3.2 Outlier exons are larger in size when compared to standard exons.

Mean positional exon half-life and mean positional exon length analysis. T-test was performed against standard positional exon half-life and length (**** P ≤ 0.0001)

of the standard isoforms vs, the less stable outlier isoforms with isoform length displayed a positive relationship (Figure 3.6d). Additionally, the more stable isoform dataset displayed a negative relationship between the isoform half-life and isoform length (Figure 3.6d). In agreement with the observations made above, faster degrading isoforms tend to display a smaller sequence length.

Evolutionarily Younger Exons Tend to be Outlier Exons

Previous studies have demonstrated the importance of exon shuffling, the formation of new exons, and alternative splicing [66], [117]. One such study developed an exon size conservation database using 76 vertebrate sequence alignments [66]. We combined our exon degradation dataset with the exon size conservation database to investigate how sequence conservation (phyloP) impacts mRNA stability. The phyloP scores in the exon size database allowed us to assign a sequence conservation score, with phyloP values >3 indicating high exon sequence conservation [66]. The “Ultra-In” score defines exon size conservation across all species tested. As such, an “Ultra-In” score <10 indicates low exon length conservation, a score between 10 – 40 represents moderate exon length conservation, and a score of >40 represents high exon length conservation. The more stable exon bin (4.75-24 hours) displays a significantly higher phyloP and Ultra-In score when compared to less stable exon bin (0-3.25 hours) (Figure 3.7a&b). These results suggests that the more stable exons tend to have high sequence conservation, a feature common to housekeeping genes [118].

Next, we carried out the conservation analysis for outlier exons. Outlier exons are characterized by significantly lower phyloP score and exon size conservation scores when compared to standard exons. However, the exon size conservation score is considered high for both groups as defined in Movassat et al. [66] (Figure 3.7c&d). This trend is also observed when

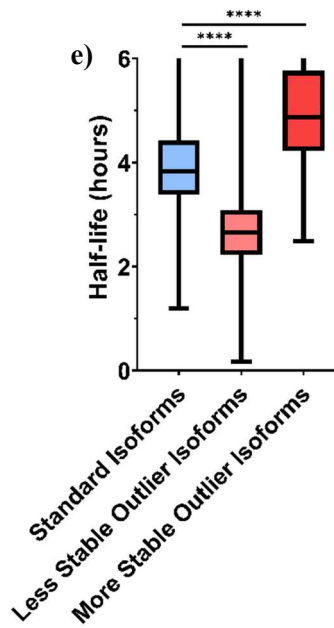
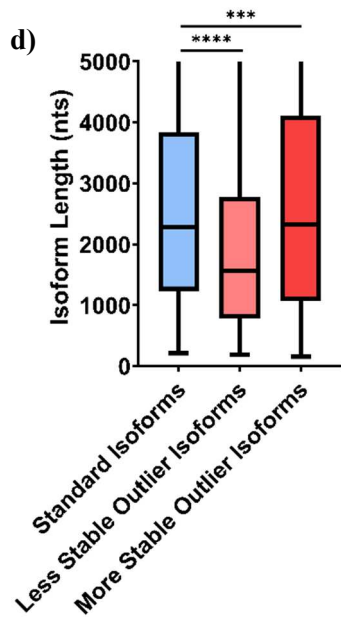
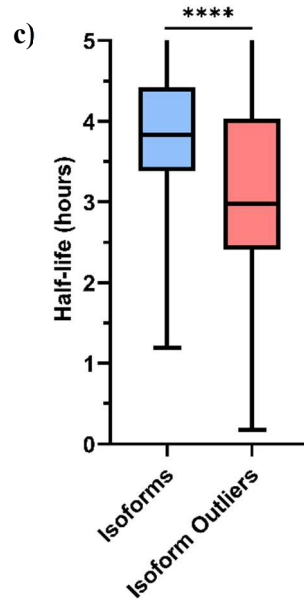
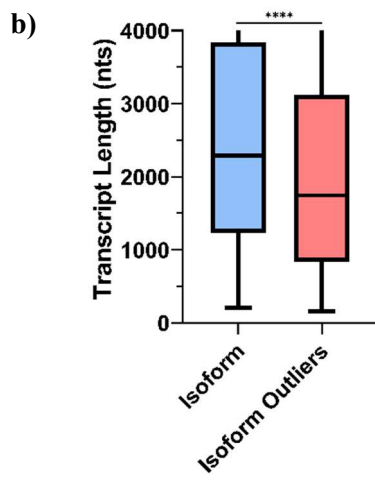
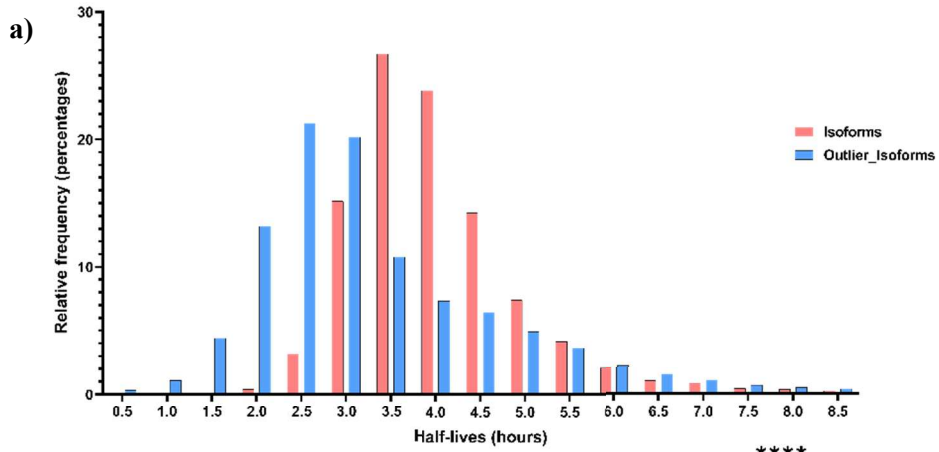


Figure 3.6 Correlation between mRNA isoforms and RNA length.

(a) Standard isoform and outlier isoform half-life histogram. (b) Boxplot analysis correlating standard and outlier isoforms with transcript length. (c) Boxplot analysis correlating standard and outlier isoform with half-life. (d) Boxplot analysis correlating standard, less stable outlier, and more stable outlier isoforms with transcript length. (e) Boxplot analysis correlating standard, less stable outlier, and more stable outlier isoforms with transcript half-life. (**** $P \leq 0.0001$, *** $P \leq 0.001$)

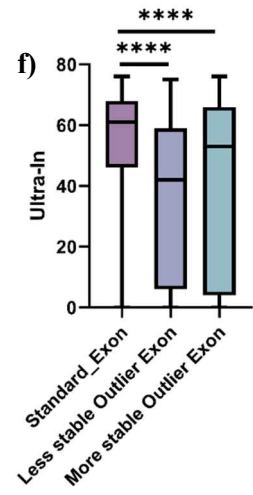
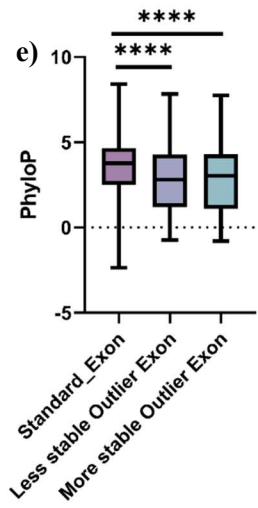
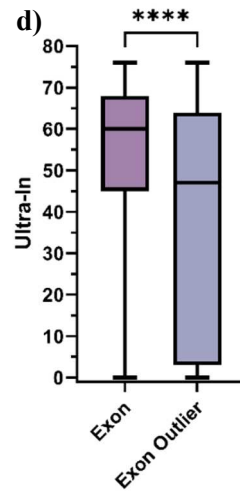
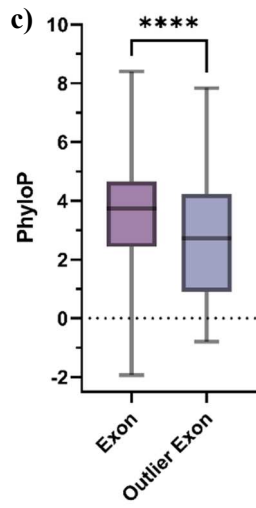
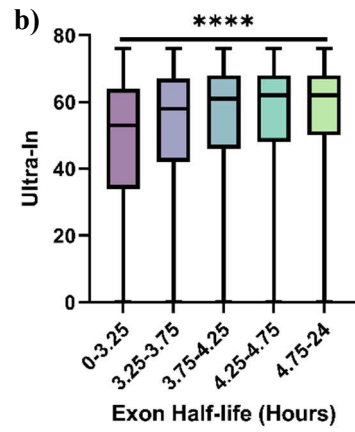
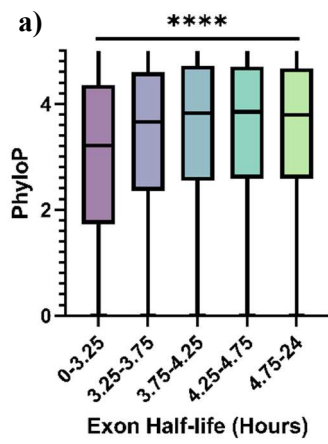


Figure 3.7 Correlation between exon degradation kinetics and sequence conservation.

(a) Bar plot analysis correlating PhyloP across 5 different exon half-life bins. (b) Bar plot analysis correlating exon size conservation (Ultra-in) across 5 different exon half-life bins vs. (c) Bar plot analysis correlating PhyloP across standard exon and outlier exons. (d) Bar plot analysis correlating exon size conservation (Ultra-in) across standard exon and outlier exons. (e) Bar plot analysis correlating PhyloP across standard, less stable outlier, and more stable outlier exons. (f) Bar plot analysis correlating exon size conservation (Ultra-in) across standard, less stable outlier, and more stable outlier exons. (**** $P \leq 0.0001$)

the outlier exons are parsed into less or more stable exons. (Figure 3.7e&f) Using the arguments presented in Movassat et al. 2019 [66], these results suggest that exon half-life outliers tend to consist of evolutionarily younger exons.

Discussion

Investigating gene expression through a dynamic lens is crucial to better understand the mechanisms of gene regulation. To accomplish this, gene expression programs need to be studied from RNA synthesis to RNA degradation. Our lab has established a protocol and conducted experiments to study pre-mRNA synthesis and intron removal rates (Garibaldi et al. unpublished). This chapter has focused on establishing a method using pulse chase metabolic labeling (4sU) to investigate mRNA degradation/stability. A 24-hour 4sU pulse was established with the expectation of reaching steady levels of 4sU-labeled RNA expression to directly capture mRNA degradation rates after 4sU withdrawal. We were able to derive half-lives for canonical gene, exon half-lives, and mRNA isoform half-lives. Our findings demonstrated a positive relationship between gene length and mRNA half-life. Furthermore, our exon half-life analysis suggests outlier exons at both extremes of stability tend to have longer exon lengths compared to standard exons. This trend of longer length leads to an outlier half-life is also observed in the isoform analysis. Lastly, we demonstrated the outlier exons have characteristics of being evolutionarily younger exons when compared to standard exons.

The gene half-life dataset demonstrated that lncRNAs and NMD transcripts are characterized by a lower mean half-life. This agrees with the findings of published studies focusing on lncRNA stability[119]. Unlike our results, a previous study suggested a negative correlation between mRNA length and RNA stability [114]. This discrepancy may be the consequence of the different assays used to derive RNA half-lives. While the previous datasets

were derived by following mRNA disappearance after general transcription inhibition, our 4sU approach did not manipulate cellular homeostasis as strongly.

It is known that the 5' and 3' untranslated regions harbor binding sites for trans-acting factors that can play critical roles in mediating RNA stability [120]–[122]. This led us to analyze the exon half-life data to identify differences in observed half-lives based on exon position (first, internal, or last). Interestingly, first exons display a lower half-life when compared to the internal and terminal exons, perhaps a reflection of major mRNA degradation pathways that moves 5' to 3'. Our exon type results are consistent with previously published observations [69]. Clearly, the exon positional analysis highlighted the importance of the 5' & 3' UTR length in determining the half-life of an RNA. Importantly, our observations lend further support to the notion that longer UTRs provide more binding sites for destabilizing or stabilizing mRNA factors.

With our analysis, we introduced the ability to analyze an mRNA's half-life through an exonic point of view. We were able to identify outlier exons whose half-life differs from their respective canonical gene half-life. Once again, we observed a positive relationship between RNA half-life and exon length. Interestingly, the comparison between standard exons and the less or more stable outlier exons revealed that outlier exons are always longer. (Table 3.1). This correlation applies even when the exon data was parsed based on exon position. It has been proposed that the UTRs house landing hubs for RNA-binding proteins (RBPs) and microRNAs [123]. RBPs have been shown to affect alternative splicing and mRNA stability. In light of this, it is reasonable to speculate that larger UTRs create more potential binding sites, thus leading to outlier exons and consequently outlier isoforms. Keeping within the realm of alternative splicing, some trans-acting factors have been known to act in their capacity as a splicing factor and as a de/stabilizing RBP [19], [20], [21]. Both roles can impact mRNA stability directly and indirectly

[125]. It is also worth stating the optimal exon length for efficient splicing is proposed to be between 50 and 250 nts [53], [56], [66]. In addition to lower expression levels relative to standard exons, the outlier exons tend to fall outside this optimal exon length range.

Creating isoform datasets from small RNA-seq is a novel approach to study the relationship between mRNA processing and mRNA stability. The SALMON tool used in our study is a bioinformatic tool with its limitations as it is considered an indirect route of calculating isoform counts[127]. An alternative and more direct route would be to perform a long read direct sequencing of the mRNA isoforms present [128]. This would provide the most accurate isoform profiling.

To determine the impact of sequence conservation and length conservation we combined our exon data with our previously published exon size database [66]. Our analysis showed that the more stable exons have a more conserved length and sequence relative to the faster degrading exons (Figure 3.7a&b). These are features that are common to housekeeping genes. Housekeeping genes have been known to show very strict conservation in the evolutionary process [118]. We also demonstrated the exon outliers having a significantly lower sequence and length conservation when compared to the standard exons. This suggests the outlier exons are younger exons. One question that would be interesting to explore is if the emergence of these younger outlier exons contributed to the development of a more stable dynamic transcriptome in mammals. This could be explored by carrying out a similar exon half-life analysis across all eukaryotes along with a size and sequence conservation analysis.

There are many mRNA stability determinants that have been identified in mammalian cells. In recent years several papers described mechanisms of regulating mRNA levels through the control of degradation. One such example is codon optimality-mediated mRNA degradation

through translational elongation in mammalian cells [72], [129]. Other regulatory mechanisms take advantage of altered RNA modifications which impact mRNA structure as to elicit changes in mRNA stability [130]. Known for several years is the importance of the 3' poly(A)-tail for effective translation efficiency, ultimately impacting RNA stability [131]. It is important to uncover additional mechanistic facets regulating mRNA stability, as this knowledge can be the key to developing new therapeutics. While it is exciting to learn more about mRNA stability, many of the determinants controlling this important step in gene expression have yet to be confirmed in human cells. Our 4sU pulse chase approach would be best suited to explore expression program differences between a wild-type cell and an impaired cell line or a stem cell vs fully differentiated cell. Coupled with direct mRNA long read technologies metabolic labeling would allow to specifically and directly explore the relationship between alternative splicing and mRNA isoform stability. The mRNA long-read technology would provide not only reliable isoform quantification but also unique mRNA modification information. The resulting data coupled with ENCODE's functional map of human RNA-binding proteins could open a world of possibilities for identifying stabilizing and destabilizing RBPs to be explored [132].

Materials and Methods

Cell Culture and 4sU Metabolic Labeling

HepG2 cells were grown in High Glucose DMEM (HyClone, SH30022.01) with 10% FBS at 37°C in 5% CO₂ in a 15cm dish. For labeling of mRNA, cultured cells at 50% confluence were treated with 40 μM 4sU for 24 hours. After a 24-hour pulse period the cells were washed with PBS and the cell media was replaced with 20mM Uridine new DMEM media. Time points were collected at 0,1,3,6,9,12, and 24hrs after the 4sU pulse period. The cells from the timepoints were resuspended in TRIzol reagent, flash frozen, and stored overnight at -80°C. Cell

lysates were chloroform extracted once, and total RNA was extracted following RNA precipitation as described in Garibaldi et al. 2017 [133].

Purification of 4sU-Labeled RNA

Biotinylation and 4sU-RNA enrichment with HPDP-biotin were carried out based on protocols adapted from Garibaldi et al. (2017) using 80 ug total RNA. A 4sU Act Yeast mRNA spike in was introduced to each RNA timepoint at the same concentration prior to the streptavidin pull down. This 4sU Yeast spike-in was used for a pull-down efficiency normalization.

4sU-Seq Library Preparation and Sequencing

1 ul of 1:100 dilution of ERCC spike-in was added to each 4sU enriched sample prior to library preparation. cDNA libraries were prepared with a polyA selection using Illumina TruSeq mRNA standard protocol. 100 bp paired-end reads were sequenced on the Illumina HiSeq 4000 platform. The samples were run on two lanes with each time point producing between ~ 40,000,000 to 104,000,000 uniquely mapped reads.

Bioinformatic Analysis

Reads were aligned with STAR aligner using the 2-pass mode, to a customized UCSC annotation known canonical genes GENCODE v39. Another alignment was done using GENCODE's v39 annotation file with all known exons to generate the exon dataset. The consequent read count table was normalized to the 4sU Act Yeast spike in, in such a way where the spike-in was equally represented in all timepoints. Only genes or exons with a minimal count of ≥ 50 normalized counts in the 1-hour time point were used. The time dependent series of read counts was used to create a degradation profile based on first-order decay kinetics according to

$[mRNA] = [mRNA]_0 e^{-kt}$, where $[mRNA]_0$ is the initial abundance of the mRNA and k is the observed rate of mRNA degradation. Corresponding half-lives were calculated using $t_{1/2} = \ln(2)/k$. A regression analysis was performed and only those timeseries with a R squared value ≥ 0.60 were used. R scripts and excel were used to analyze data and combine the HepG2 degradation data with the exon size database. Prism was used to carry out statistical tests and to illustrate figures.

CHAPTER 4

Nutritional Control of Splicing Fidelity Contributes to Methionine Dependent Proliferation Defects in Cancer Cells

Summary

Many cancer cells depend on exogenous methionine for proliferation, whereas non-tumorigenic cells can divide in media supplemented with the metabolic precursor homocysteine. This phenomenon is known as methionine dependence of cancer, or the “Hoffman effect.” The underlying mechanisms for this cancer-specific metabolic addiction are unknown. Using a splicing analysis of the methionine-dependent triple negative breast cell line MB468 and its revertant methionine-independent R8 cell line, we find that methionine dependence is associated with severe dysregulation of pre-mRNA splicing. When cultured in homocysteine medium, cancer cells failed to efficiently methylate the spliceosomal snRNP component SmD1, which resulted in reduced binding to the Survival-of-Motor-Neuron (SMN) protein leading to aberrant splicing. These effects were specific to cancer cells as neither Sm protein methylation nor splicing fidelity was affected when non-tumorigenic cells were cultured in homocysteine medium. Sm protein methylation is catalyzed by Protein Arginine Methyl Transferase 5 (PRMT5) and reducing methionine concentrations in the culture medium sensitized cancer cells to PRMT5 inhibition. These results mechanistically connect splicing fidelity to nutrient availability in cancer cells.

Introduction

Recent research suggests that the reprogramming of cellular metabolism is a hallmark of cancer [85]. Cancer metabolic reprogramming is seen as a selected feature for the promotion of tumorigenesis. The most common and best studied example of this energy reprogramming in cancer is the “Warburg effect,” which describes the increased use of inefficient glycolysis for ATP production despite being in the presence of an aerobic environment [84], [85]. Thus, cancer cells require an increased glucose uptake. A less studied cancer-specific metabolic reprogramming is the phenomenon known as the “Hoffman effect,” which describes the dependency of cancer cells on exogenous methionine for cellular proliferation. It was first reported in 1959 when tumor-ridden rats were fed a methionine restricted diet. The altered diet greatly impacted tumor growth [134]. Methionine-dependence has been observed repeatedly in many different cancer types [86]. Methionine-addicted cancer cells can only grow in the presence of exogenous methionine, but not with the introduction of its precursor homocysteine. Importantly, normal cells can easily adjust and proliferate under methionine-depleted (MET-) and homocysteine-supplemented (HCY+) media. Previous research focusing on uncovering the mechanism behind the Hoffman effect revealed that the supplementation of exogenous S-adenosyl methionine (SAM) in MET- HCY+ media restored the proliferation of methionine-dependent cells [90]. In the cell, SAM is the primary methyl donor and metabolite of methionine. These observations suggested that cancer cells have a reduced capacity to generate SAM and highlighted the importance of SAM concentrations for cellular growth. Borrego et al demonstrated that the replacement of methionine with homocysteine resulted in a rewiring of metabolic pathways, with homocysteine being redirected towards glutathione rather than methionine synthesis [89]. This ultimately results in a lower SAM/SAH (S-

adenosylmethionine/S-adenosylhomocysteine) ratio, a proposed indicator for the cellular methylation potential. As a metabolite of SAM, SAH is a competitive inhibitor of SAM-binding methyltransferases. The SAM concentration also plays a major role in epigenetic regulation, nucleotide biosynthesis, and membrane lipid homeostasis [135]. While the methionine cycle itself is directly tied to cancer cell proliferation, it is also connected to several pathways vital for cell proliferation [135], [136]. However, the mechanisms responsible for cell arrest and, ultimately, cell death when cancer cells are starved of methionine are not well understood.

Protein arginine methyltransferases (PRMTs) act as writers of arginine methylation in histones and non-histone proteins by transferring methyl groups from SAM to a guanidine nitrogen of protein arginine. This reaction results in the methylarginine and SAH [91]. Several papers have presented evidence connecting the activities of PRMTs to the regulation of RNA splicing and, consequently, linking proper methylation of splicing factors to correct RNA splicing [80]. Other work has shown that aberrant RNA splicing is linked to the development of cancer [78]. The RNA splicing/cancer connection is so prevalent that it is argued to be a hallmark of cancer [73], [74], [78]. For example, a study of ~9000 cancer patients representing 32 different tumor types revealed many cancer-specific alterations in alternative splicing [81]. Often, these cancer-specific mis-splicing events are associated with mutations/alterations in RNA splicing factors[137].

In this chapter we use a methionine-dependent cell line MDA-MB-468 (MB468) along with its methionine-independent clone MDA-MB-468res-R8 (R8) to study gene expression differences in the context of methionine dependency in cancer cells. We analyze a time course experiment of MB468 or R8 cells incubated in methionine-depleted medium to explore whether differential gene expression provides novel insights into the biological processes involved in

methionine dependency. We also explored the potential connection between methionine stress in cancer and pre-mRNA splicing dysregulation. Our splicing analysis highlighted the importance of methionine availability for favorable splicing activity to support cancer cell growth. We show that the failed methylation of the splicing factor SmD1, and subsequent splicing dysregulation in methionine-depleted conditions, was unique to cancer cells. Our results also highlight the importance of PRMT5 function in the methylation of splicing factor SmD1 at methionine-depleted conditions. This work provides a first link between the Hoffman effect and RNA splicing as a cellular pathway that contributes to cancer cell proliferation.

Results

Expression profile changes in response to methionine stress

Our collaborator, the Kaiser lab at the University of California Irvine, has previously developed a method to derive methionine-independent clones from a methionine-dependent cell line. They achieved this using a methionine-dependent triple-negative breast cancer cell line, MDA-MB-468, to develop MDA-MB-468res-R8 [89], [138] (Figure 4.1A). Using these cell lines the Kaiser lab performed a time course experiment over 12 hours in MET⁻ HCY⁺ medium, carried out in triplicates, to examine and identify molecular characteristics that may be associated with methionine dependency of cancer cells. The time course experiment started with transferring MB468 cells and R8 cells from MET⁺ medium to MET⁻ HCY⁺ medium, followed up by capturing timepoints at 30 and 720 mins into methionine depletion. These time points allow us to evaluate any initial transcriptional changes (30 min) after the cell lines were introduced to the MET⁻ environment or steady-state expression changes due to prolonged methionine starvation (720 min). Cells were harvested at each time point and RNA-seq was carried out from total RNA.

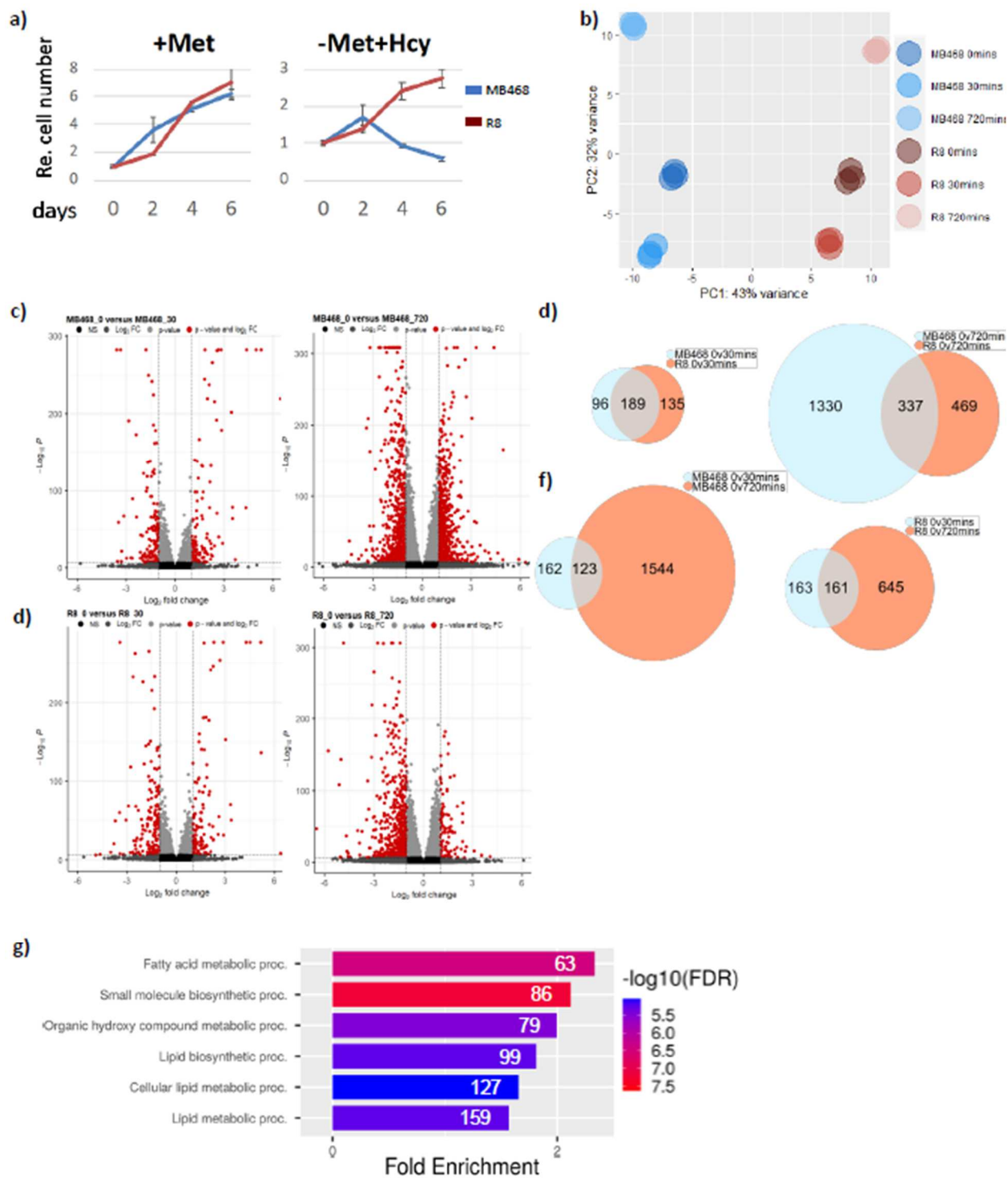


Figure 4.1 Methionine-dependent cells (MB468) experience the highest impact of methionine stress at 720 min in MET- HCY+ medium.

(a) MB468 and R8 cells were cultured in MET+ or MET- HCY+ medium for 6 days and cell proliferation was measured. (b) A principal component analysis plot indicates MB468 and R8 cells lines have distinct expression profiles but respond in a similar way to methionine stress. (c) Volcano plots of differential expression analysis of MB468 0min vs 30 or 720 min into methionine stress medium conditions. The highest-level differential expression is seen at the 720 min timepoint. (d) Volcano plots of differential expression analysis of R8 0min vs 30 or 720 min into methionine stress medium conditions. MB468 experiences the most gene differential expression. (e) Ven diagrams illustrating common differentially expressed genes across cell lines at the same time point into methionine stress. (f) Ven diagrams illustrating common differentially expressed genes within cell lines at different time points into methionine stress. MB468 720 min into MET- HCY+ medium results in the highest number of unique differentially expressed genes. (g) Bar plot displaying GO analysis of differentially expressed genes in MB468 720 min after methionine withdrawal.

A principal component analysis (PCA) demonstrated little to no variability between the replicates (Figure 4.1b) and only minor differences in gene expression between the 0 and 30 min timepoints. A more drastic change in gene expression is observed at the 720 min timepoint, as is highlighted by volcano plots (Fig 4.1c, d). After 30 mins into methionine starvation 285 genes change expression in MB468 cells. After 720 min this number increases to 1667 differential expressed genes (Figure 4.1c). A similar trend is seen for the methionine independent R8 cell line, however, to a lesser extent with 324 differential expressed genes at the 30 min timepoint and 806 genes at the 720 min timepoint (Figure 4.1d). A gene overlap analysis was carried out to identify genes that change expression upon MET- HCY+ medium shift in both cell lines (Figure 4.1e&f). The initial comparison of 0 vs 30mins post methionine withdrawal displayed the highest relative frequency of gene overlap between the two cell lines (Figure 4.1e). Gene ontology of differentially expressed genes in the two cell lines did not reveal any distinct biological processes responsible for the difference in metabolic needs at the early 30 min timepoint. When analyzing the late timepoint, gene ontology analyses linked uniquely differentially expressed genes in MB468 to fatty acid biosynthetic processes (Figure 4.1g). This is consistent with a system-wide lipid profiling study that associated changes in lipid metabolism to ER stress and the Hoffman effect [138]. At late time points of methionine withdrawal, the overlap between MB468 and R8 differential expressed genes is relatively small (20%). These results demonstrate that methionine dependent MB468's gene expression is more impacted by exogenous methionine availability, with 1667 differential expressed genes, of which 80% are uniquely expressed.

Methionine stress impacts splicing fidelity in cancer cells

The differential gene expression analysis did provide insights into the biological processes that are impacted by methionine stress. However, no obvious link to cell proliferation and, ultimately,

a molecular mechanism to methionine dependency was identified. A literature review revealed the importance of protein arginine methyltransferases in splicing, more specifically the methylation of spliceosome proteins by PRMT5 and PRMT7 [91]. For example, PRMT5 inhibition has been shown to disrupt splicing and stemness in glioblastoma [139]. Given that PRMTs use SAM to carry out protein methylation, we explored whether methionine dependency of cancer cells is linked to changes in pre-mRNA splicing.

To determine if the shift from MET⁺ to MET⁻ HCY⁺ media resulted in alternative splicing we performed an rMATs analysis of the methionine depletion time course. rMATs is a computational tool designed to detect differential alternative splicing events from bulk RNA-seq data. Given that the largest number of differentially expressed genes in both cell lines occurred 720 min into methionine depletion, our splicing analysis focused only on differences between the 0 and 720 min timepoints (for nomenclature referred to as the ‘MET⁻’ analysis). The MET⁻ analysis identified 2,429 AS events in MB468 cells and 1,069 AS events in R8 cells (Figure 4.2a). Interestingly, alternative exon inclusion is the most common MB468 MET⁻ AS event (1,616, FDR <0.05, inclusion level difference +0.10), of which ~75% are associated with reduced exon inclusion (Figure 4.2b). These observations suggest that upon methionine stress the recognition of alternative exons is reduced in methionine-dependent cancer cells. The analogous analysis of methionine independent R8 cells identified 773 statistically significant exon inclusion events (Figure 4.2c). The R8 MET⁻ analysis demonstrated that alternative exon inclusion is more evenly distributed between (58% exon skipping, 42% exon inclusion). To determine the splicing changes related to methionine dependency we compared splicing patterns between methionine-dependent MB468 and methionine-independent R8 cells after prolonged methionine withdrawal. Consistent with the notion that the withdrawal of methionine leads to exon skipping, 60% of the

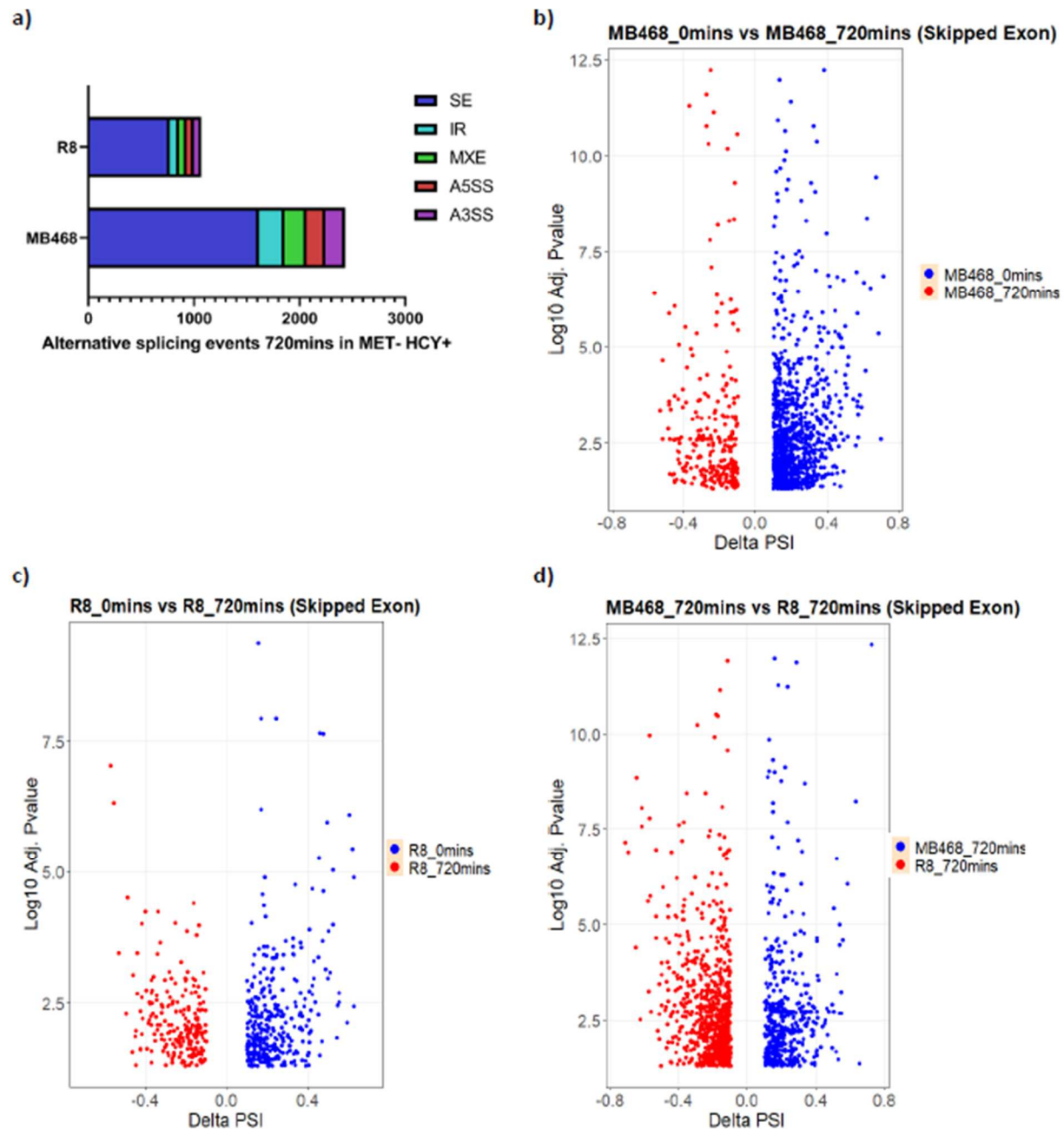


Figure 4.2 Methionine stress in MB468 impacts splicing fidelity. (a) A stacked bar graph displaying MDA-MB468 and R8 alternative splicing event distributions 12 hours into M-H+ medium shift. (b and c) Volcano plots displaying the difference in percent spliced in (PSI) of skipped exon events in MDA-MB468 (b) or R8 (c) cells upon methionine restriction. Blue indicates higher exon inclusion levels before methionine restriction. Red indicates higher inclusion levels after methionine restriction. (d) Volcano plot displaying PSI of skipped exon events in MDA-MB468 and R8 cells after 12 hours of methionine restriction.

differential splicing events detected are characterized by higher exon skipping in MB468 when compared to R8 (Figure 4.2d).

ShinyGO was used to perform a GO analysis to determine what biological process might be impacted by MB468 methionine dependence [140]. Genes characterized by exon skipping upon methionine withdrawal in MB468 fall into general categories of cell cycle, mitotic cell cycle, regulation of cell cycle, DNA repair, and positive regulation of DNA metabolic processes (Figure 4.3a). These terms indicate an association between pre-mRNA splicing changes and cell division. Interestingly, genes characterized by increased exon inclusion during methionine withdrawal are categorized into RNA splicing and regulation of RNA splicing processes (Figure 4.3b). These observations suggest that alternative splicing of RNA processing-related factors mediate reduced inclusion of exons within cell cycle control genes, thereby inhibiting proliferation.

If methionine dependence is a major driving force in splicing differences, it is likely that specific exons skipped in MB468 are retained at higher rates in R8 cells upon methionine stress. This is indeed the case. Figure 4.4a illustrates the inclusion level difference of two rMATS comparisons merged on top of each other based on shared differentially included exon events. The MET-analysis for MB468 ('MB468 0 mins vs 720 mins') and methionine withdrawal analysis across both cell lines ('MB468 720 mins vs R8 720 mins') displayed 645 shared differential exon splicing events. Interestingly, the same events that resulted in reduced exon inclusion in MB468's MET- analysis resulted in increased exon retention in the methionine withdrawal across both cell lines (Figure 4.4a). The inverse correlation is also observed for exons that displayed increased exon inclusion levels in MB468 upon methionine removal.

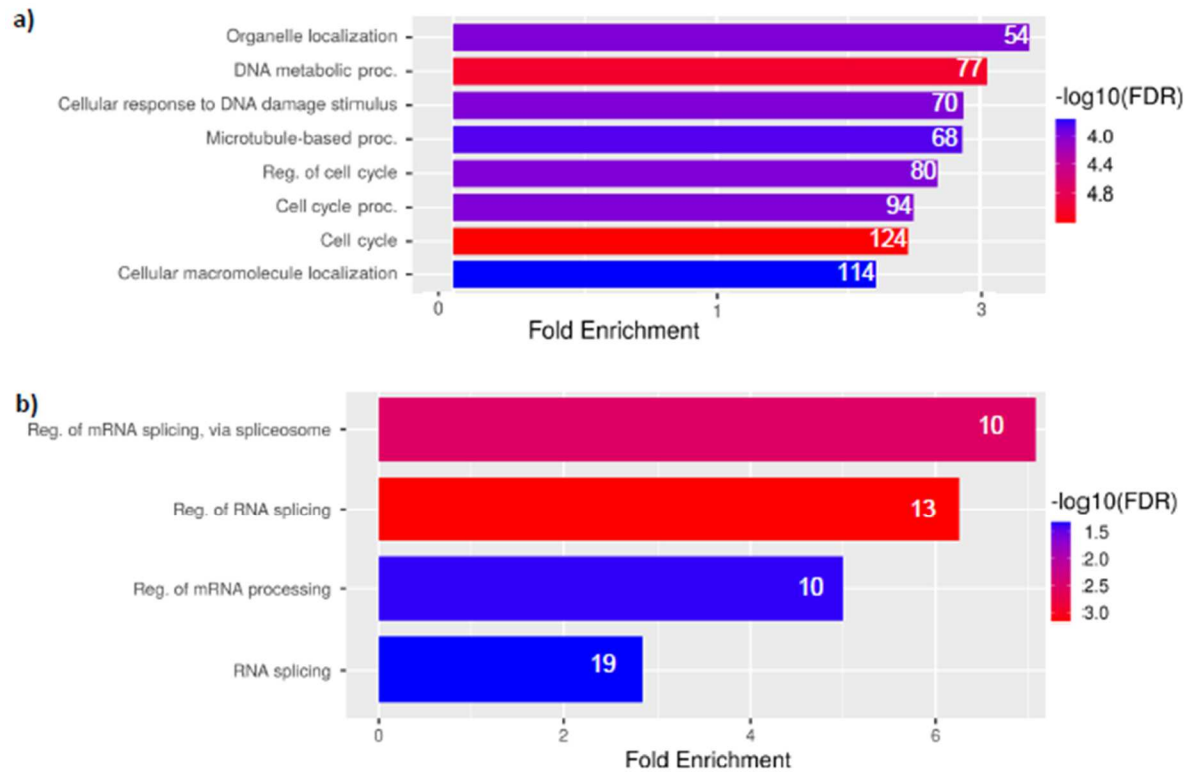


Figure 4.3 Gene ontology analysis of alternatively spliced genes upon methionine stress in MB468 cells.

(a) Significantly enriched biological processes of genes associated with increased exon skipping upon methionine withdrawal in MDA-MB468 cells. (b) Significantly enriched biological processes of genes associated with decreased exon skipping upon methionine withdrawal in MDA-MB468 cells. The numbers in the bars represent the number of genes hits in each category. The colors indicate statistical significance.

A gene ontology analysis of these overlap events focusing on increased skipped exons in MB468 linked the affected genes to RNA processing (Figure 4.4c) while the overlap events focusing on increased exon retention were linked to cell cycle processes (Figure 4.4d). These results provide evidence to suggest that methionine stress in MB468 results in different mRNA isoform expression of genes that regulate pre-mRNA splicing and genes associated with cell division.

A complementary rMATs analysis was carried out analyzing intron retention events in MB468 upon methionine withdrawal, also highlighting genes associated with RNA processing, RNA splicing, and the regulation of RNA processing. These observations further support the notion that methionine stress leads to AS in genes involved in pre-mRNA splicing, potentially contributing to aberrant splicing antagonistic to cell proliferation. Remarkably, as was observed in the exon skipping analysis, splicing events with increased intron retention upon MB468 methionine withdrawal overlapped with decreased intron retention events in the methionine withdrawal analysis across both cell lines. Thus, the same splicing events that result in increased intron retention post methionine withdrawal in MB468 cells are more efficient in methionine-independent R8 cells post methionine withdrawal. The inverse relationship is also observed with decreased intron retention events (Figure 4.4b). Gene ontology analyses of overlapping intron splicing events link associated genes with RNA splicing, as was seen in the exon splicing overlap analysis. Similar relationships were observed to a lesser extent in analyses carried out on overlapping alternative 5'ss, alternative 3'ss, and mutually exclusive splicing events. The results from the overlap comparisons provide evidence that methionine dependency in cancer cells leads to pre-mRNA dysregulation, which may be detrimental for cell proliferation.

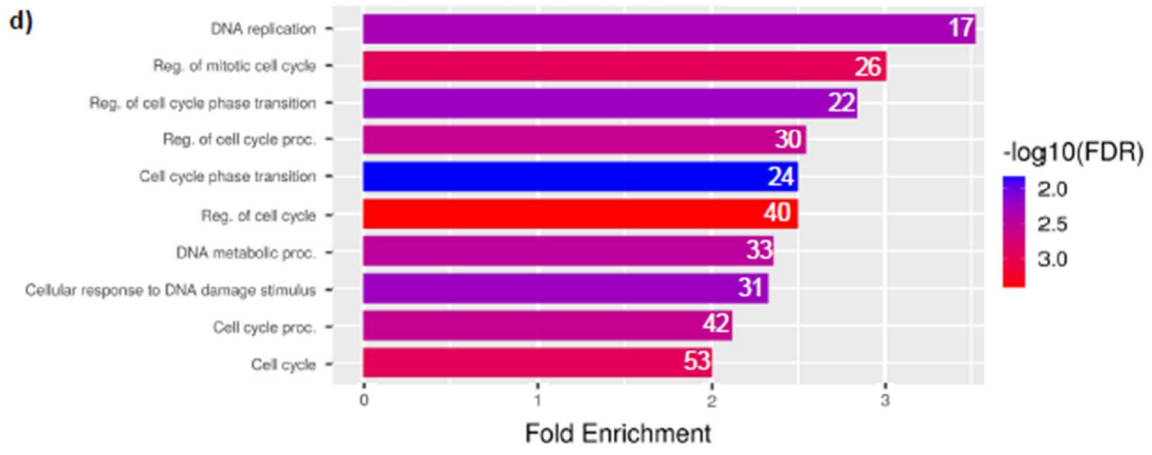
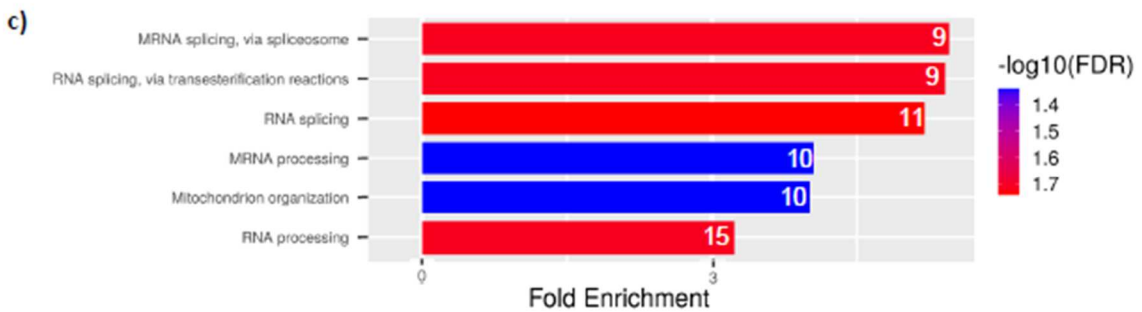
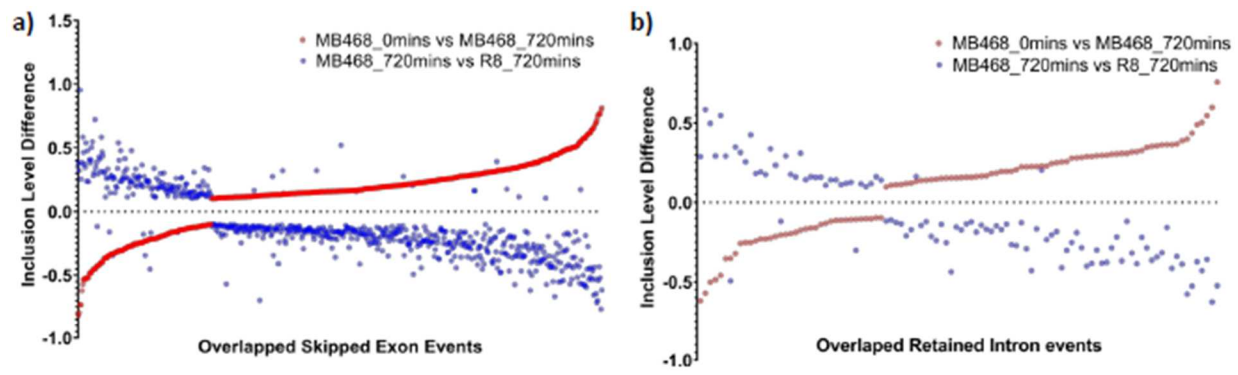


Figure 4.4 Inverse relationship between overlapping skipped exon and intron retention events in MB468 and R8 cells.

(a) Each dot in the graph represents an overlapping exon inclusion or exclusion event, organized by the inclusion difference observed in the MET- analysis (MB468 0 min - MB468 720 min) in MB468 cells (red dots). The blue dots represent the difference observed in identical exon inclusion events when MB468 are compared with R8 cells after prolonged methionine withdrawal (MB468 720 min - R8 720 min). The y-axis displays the event inclusion level difference or. The x-axis displays an overlapping differential alternative splicing event. (b) Plot displaying overlapping alternatively retained intron events for the same datasets as described in (a). (c) Gene ontology analysis of overlapping alternative splicing events (a) that display increased exon skipping in MB468 post methionine withdrawal. (d) Gene ontology analysis of alternative splicing events (a) that display increased exon retention in MB468 post methionine withdrawal. The numbers displayed in the bar graphs represent the number of gene hits in each category. The colors indicate statistical significance.

Methionine-dependent cancer cells fail to efficiently methylate Sm proteins.

The splicing analysis of methionine-dependent MB468 cells and methionine-independent R8 cells showed that the splicing fidelity is sensitive to changes in exogenous methionine supply. This sensitivity was more apparent in MB468 cells post methionine withdrawal (Figure 4.2a). Our collaborators have previously shown that a key aspect of the Hoffman effect is the stability of the cellular methylation potential during metabolic changes [89], [90]. The methylation potential is measured by the SAM/SAH ratio [86], [135]. Metabolomic profiling demonstrated a significant reduction in the SAM/SAH ratio in MB468 but not in R8 cells when presented with methionine withdrawal [89]. PRMT5 and PRMT7 play important roles in catalyzing arginine methylation of splicing factors [91], [139], [141]. To probe for changes in the symmetric dimethylarginine (SDMA) modification during methionine stress, the Kaiser lab used a PAN dimethyl-arginine antibody. Relatively few changes were observed in the dimethyl-arginine methylation patterns upon shifting cells into MET- HCY+ medium. However, one prominent band corresponding to the methylated form of the spliceosomal core component SmD1 disappeared in methionine-dependent MB468, but not in methionine-independent R8 cells (Figure 4.5a). Overexpression by PRMT5 delayed the demethylation of SmD1 in MB468 cells, consistent with SmD1 being a substrate of PRMT5 (Figure 4.5a) [141]. Dimethylation of SmD1 catalyzed by PRMT5 increases SmD1 binding to the Survival of Motor Neuron (SMN) complex. The SMN complex functions as a facilitator of snRNP biogenesis by promoting interactions between Sm proteins, like SmD1, and the snRNAs [86]. Our collaborators therefore tested whether methionine stress in MB468 and R8 cells leads to decreased binding between SMN and SmD1 upon methionine stress. A SMN immunoprecipitation demonstrated decreased interactions between SMN and SmD1 in MB468 upon a shift to MET- HCY+ medium while

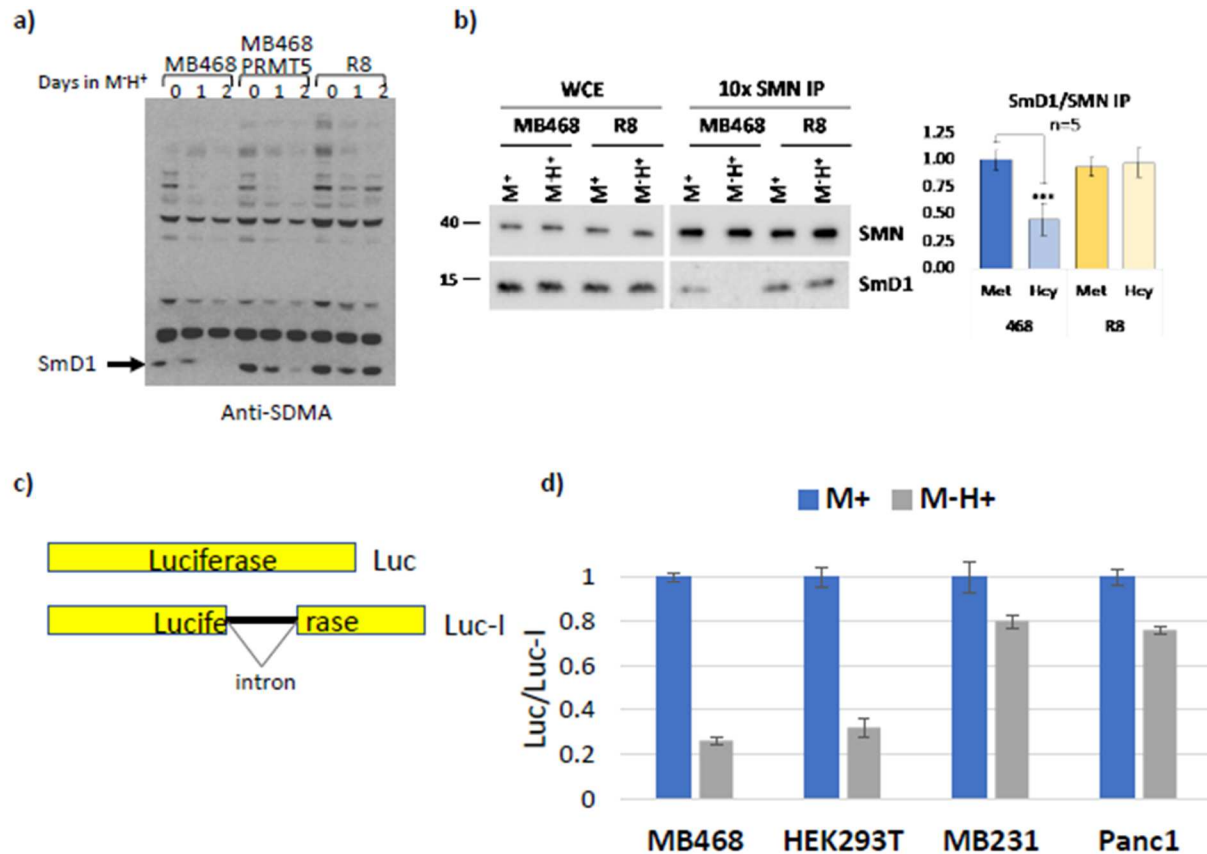


Figure 4.5 Methionine stress in MB468 cells leads to loss in SmD1 methylation and loss of splicing fidelity.

(a) Western blot using anti-SDMA (symmetrical dimethylarginines) antibodies (left) in MB468, MB468 with PRMT5 overexpression, and R8 cell lysate post methionine depletion (M⁻) and homocysteine supplemented (H⁺) medium shift. (b) SMN immunoprecipitation of SmD1 in MB468, R8 in methionine supplemented (M⁺), and M⁻ H⁺ media. The bar graph (right) represents quantification of SmD1 band intensities in MB468, R8 in M⁺, and M⁻ H⁺ media (***)P ≤ 0.001). (c) Luciferase splicing reporter scheme. Luc refers to the intronless luciferase reporter and Luc-I refers to the intron-containing reporter. (d) A bar graph displaying relative luciferase activity in different media conditions in cell lines defined by the x-axis.

methionine-independent R8 cells displayed a continued SMN SmD1 interaction despite the media shift (Figure 4.5b). These results provide further evidence that methionine stress impacts factors important for splicing fidelity.

Methionine stress results in reduced splicing activity

To measure the effect of exogenous methionine on the general splicing activity the Kaiser lab used a luciferase splicing reporter (Figure 4.5c) [142]. Methionine-dependent cell lines (MB468 and HEK293T) and -independent cell lines (MDA-MB231 and PANC1) were shifted from MET+ HCY- to MET- HCY + medium and the splicing efficiency was monitored by luciferase production, normalized to the same cells expressing intronless luciferase. Upon methionine stress methionine-dependent cells displayed a significant reduction in completely spliced luciferase (Figure 4.5d) while the production of methionine remained unchanged for methionine-independent cells. These results suggest that in methionine-dependent cells the splicing activity is directly affected by the availability of exogenous methionine.

Reduced methylation activity of PRMT5 contributes to methionine-dependence of cancer

Our results link a cancer-specific requirement for exogenous methionine with the modulation of spliceosome activity through PRMT5-mediated methylation of SmD1. Whether this nutritional effect on splicing efficiency contributes to the Hoffman effect remains unclear. This is an important question as other metabolic effects that are closely associated with methionine dependence did not contribute to cell apoptosis seen in methionine-dependent cells [89], [138]. To test whether PRMT5 effects on cell proliferation defects associates with the Hoffman effect our collaborators used the PRMT5 inhibitor EPZ015666 [143].

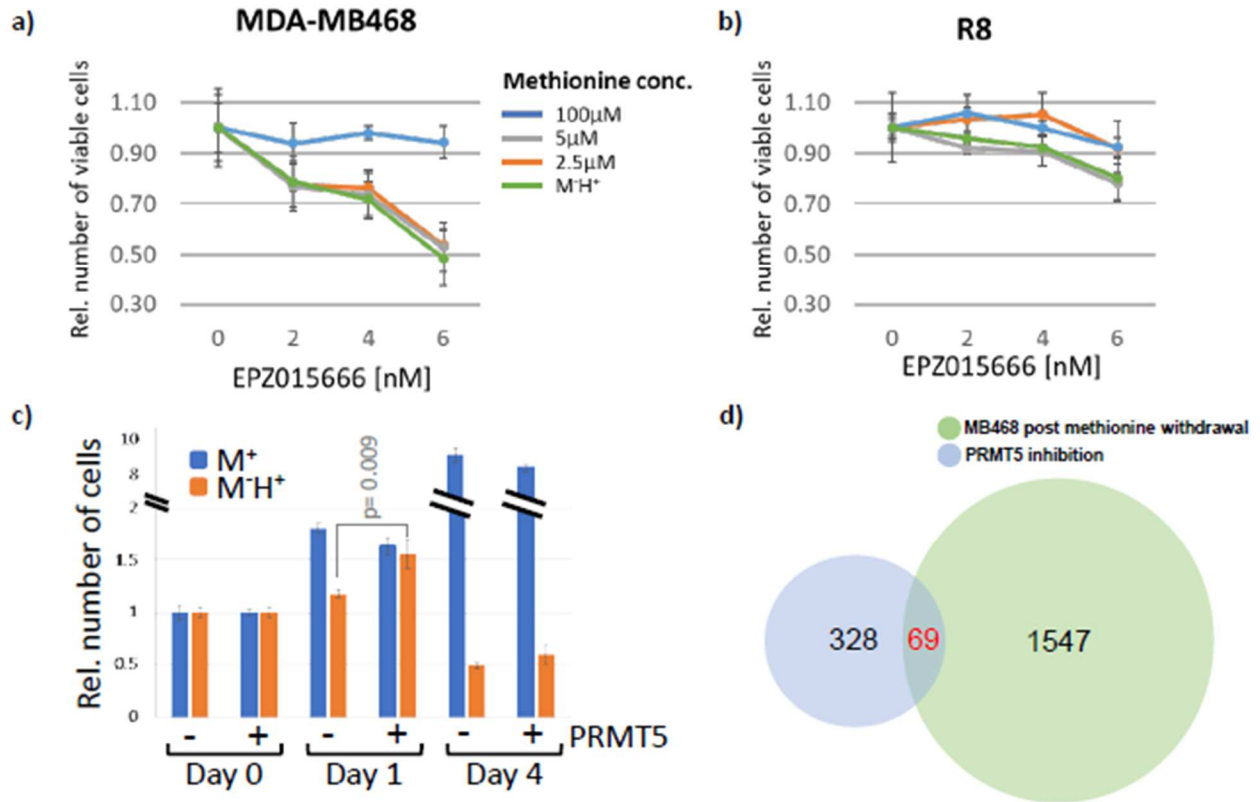


Figure 4.6 Decreased methionine availability coupled with PRMT5 inhibition promotes cell death.

(a) Relative number of viable MB468 cells upon treatment with increased PRMT5 inhibitor (EPZ015666) concentrations, coupled with variable concentrations of exogenous methionine. (b) Same as in (a) except the cell line R8 was used. (c) Cell proliferation assay in MB468 and MB468 overexpressing PRMT5 in MET+(M⁺) and MET⁻ HCY+ (M⁻H⁺) media. (d) Alternative splicing event overlap between MB468 MET⁻ analysis and an alternative splicing analysis of patient-derived glioblastoma cancer stem cell lines treated with PRMT5 inhibitor [139].

Synergistic effects of PRMT5 inhibition and the availability of exogenous methionine in methionine-dependent cells would strongly suggest a link between PRMT5 and the Hoffman effect. Interestingly, restricting exogenous methionine to 5 and 2.5 μ M sensitized MB468 cells to PRMT5 inhibition and ultimately led to cell proliferation defects (Figure 4.6a). By contrast, R8 cells were largely unaffected by the decreasing concentration of exogenous methionine coupled with increasing concentration of PRMT5 inhibitor (Figure 4.6b). These results support the notion that nutrient-impaired methionine metabolism contributes to cell proliferation defects that cause methionine dependency. Overexpression of PRMT5 in MB468 cells initially suppressed the Hoffman effect proliferation defects (day 1 into the media shift, Figure 4.6c). However, cell proliferation in PRMT5-overexpressing cell could not be sustained, indicating that PRMT5 overexpression is not sufficient to overcome the Hoffman effect (Figure 4.6c). The initial suppression of the Hoffman effect is consistent with the sustained SmD1 methylation seen in the overexpressed PRMT5 cells and SmD1/SMN co-immunoprecipitation results (Figure 4.5a), further highlighting the importance of splicing fidelity in the Hoffman effect.

If decreased exogenous methionine concentration leads to higher sensitivity towards PRMT5 inhibition and cell proliferation defects, it is possible that the alternative splicing events triggered by methionine stress are the same alternative splicing events observed upon PRMT5 inhibition. Taking advantage of a previously published RNA-seq dataset evaluating the splicing outcome of PRMT5 inhibition in a patient-derived glioblastoma cell culture system [139], we carried out a splicing event overlap analyses with our RNA-seq datasets ('PRMT5 inhibition' and 'MB468 -MET' analysis) [139]. Despite the differences in cell lines, growth conditions, and library generation, a statistically significant overlap of 69 events between the two datasets was observed (P value < 3.80E-33) (Figure 4.6d). This overlap is also seen in 'PRMT5 inhibition'

analysis and the methionine withdrawal analysis across both cell lines, with 64 overlapping events (P value $< 9.60E-31$). The highly significant alternative splicing event overlap between the datasets suggests that PRMT5 inhibition and methionine withdrawal impact the same gene pathways and splicing factors that promote cell proliferation defects.

Discussion

Methionine dependence of cancer has been known for over 40 years, yet the cellular pathways or molecular mechanisms by which cell proliferation defects occur under methionine stress have not been identified[86]. Metabolic profiling and metabolite supplementation experiments suggested that methionine-dependent cancer cells experience reduced methylation potential due to decreased SAM/SAH ratios when cultured in MET- HCY+ medium [89], [90]. Tracer experiments with labeled homocysteine revealed a redirect of homocysteine toward the transsulfuration pathway and away from the methionine cycle where SAM formation occurs [89]. This study also demonstrated that a reduction in SAM/SAH ratios correlated well with a cell's methionine dependence while cells with continued growth in HCY+ medium maintained unchanged SAM/SAH levels [89]. Furthermore, SAM supplementation was shown to be enough to overcome methionine dependence in MET- HCY+ media [90]. These experiments suggest that the methionine dependence of cancer is caused by the dysregulation of cellular pathways that depend on efficient methylation steps. Our alternative splicing analysis revealed that the splicing fidelity is severely affected when methionine-dependent cancer cells are cultured in homocysteine medium, whereas methionine-independent cells maintain their faithful splicing pattern. Efficient spliceosome assembly, and more specifically snRNP biogenesis, depend on the methylation of Sm proteins. We found that unlike most other arginine methylation events, Sm methylation is hypersensitive to the relatively small reduction in SAM/SAH ratios that are

associated with the Hoffman effect. In agreement with the notion that the loss of Sm methylation reduces the efficiency of the splicing reaction, we observed reduced inclusion of alternative exons in methionine-dependent MB468 cells. This reduced splicing efficiency in methionine-dependent cells was further demonstrated using a luciferase splicing reporter. Interestingly, 5'-deoxy-5'-methylthioadenosine phosphorylase (MTAP), a critical enzyme in the methionine salvage pathway, is hemizygotously co-deleted in 80-90% of cancer cells. MTAP leads to the regeneration of methionine by cleaving 5'-methylthioadenosine (MTA), a polyamine synthesis byproduct and a methyl transferase antagonist [144]. The loss of MTAP triggers a buildup of MTA, which has been shown to impact the ability of PRMT5 to carry out symmetric dimethylarginine (SDMA) methylation [144], as is characteristic for Sm methylation. However, there is no correlation between methionine dependence of cancer and the MTAP status [86], and MTAP deletion has no significant effect on intracellular SAM concentrations [144].

Nevertheless, MTAP deletion can indirectly affect the cellular methylation potential through its MTA inhibitor properties. MTAP deletion renders cancer cells vulnerable to PRMT5 inhibition by synergizing with genetic mutations or small molecule inhibitors [145]. Similarly, we demonstrated that methionine-dependent cells are hypersensitive to PRMT5 inhibition. Our results show that the PRMT5 inhibitor EPZ01566 in the context of decreasing exogenous methionine reduced MB468 growth and induced cell death. This sensitization to the PRMT5 inhibitor was not observed for the methionine-independent cell line R8. The importance of PRMT5 in cancer proliferation was recently highlighted in a study that deciphered the methylome of PRMT4/5/7 and their influences on RNA splicing and cancer growth [10]. This study also demonstrated the importance of methylating hnRNPA1, a snRNP and splicing

regulator. Similarly, our study demonstrated the importance of PRMT5-driven SmD1 methylation in snRNP biogenesis.

Establishing a relationship between exogenous methionine starvation, splicing dysregulation, and defects in cancer cell proliferation is important to understand the mechanisms involved in the Hoffman effect. The mechanistic insights gained from such studies could also be exploited to formulate alternative cancer treatments. Current therapeutic approaches already utilize a low-methionine diet to improve the efficacy of chemotherapy. However, this relationship could be further exploited to develop a more targeted treatment. A recent study showed that the pharmacological inhibition of type 1 PRMTs led to the generation of splicing-derived neoepitopes [146]. Because methionine stress can also induce drastic changes in splicing fidelity, it is possible that novel tumor neoantigens can arise from reduced methionine diets, thus potentially enhancing anti-tumor immunity. If so, methionine starvation might allow for a more targeted cancer treatment [146].

Methods

Bioinformatic Analysis

Gene expression analysis was done following a similar bioinformatic pipeline to Borrego 2021 et. al. [138]. Raw reads were aligned to a custom human genome, GRCh38/hg38, using the UCSC Genome Browser and the ERCC spikein sequences (<http://tools.invitrogen.com/downloads/ERCC92.fa>) using HISAT2 and STAR alignment software [147], [148]. Number of reads mapped to each gene feature was quantified by featureCounts in the Rsubread package, and unwanted sample variation was determined by RUVSeq [149], [150]. Differential gene expression analysis was performed using DESeq2 [151].

Pathway enrichment analysis was conducted using the Shiny GO Enrichment analysis tool [140]. PCA plots and venn diagrams were made using a custom R script. Alternative splicing analysis was carried out using rMATS [152]. Volcano plots were conducted using the R tool “EnhancedVolcano” (<https://github.com/kevinblighe/EnhancedVolcano>).

CHAPTER 5

Perspectives

Introduction

The excision of introns followed by the ligation of exons in pre-mRNA, commonly referred to as pre-mRNA splicing, is a central feature of gene expression in all eukaryotes. This RNA processing event has been shown to be fundamental to organismal complexity characteristic for higher eukaryotes by promoting proteome expansion and by regulating gene expression [16]. This expansion by alternative splicing is highlighted through the human genome's ability to express >90,000 different proteins from only ~25,000 protein-coding genes [3]. Therefore, splicing fidelity is important for cell proliferation, tissue identity, organ development and cellular homeostasis [16]. Additionally, many studies highlight the importance of alternative splicing in various diseases [153].

Alternative splicing decisions are impacted by different modes of exon recognition

The results from Chapter 2 highlight the importance of the intron and exon definition modes during splice site recognition. It has been shown repeatedly that splice site recognition is a combinatorial process influenced by many exon recognition determinants [9], [93]. These determinants include splice site strength, exonic and intronic splicing regulatory sequences, RNA secondary structures, pre-mRNA synthesis, and the exon/intron architecture. Chapter 2 focuses on exon/intron architecture and splice site strength and expands on previous landmark studies. One of these previous studies demonstrated that the proximity of competing splice sites across the intron dictates splice site usage, when exons are recognized by intron definition [94]. This established what the field refers to as the intronic-centric proximity rule. Chapter 2 provides

further evidence for the existence of exon definition and ultimately demonstrates that alternative splicing decisions are influenced by both modes of splice site recognition.

Creating and using the ALTssDB we can confirm that the human genome primarily consists of exons flanked by two long (>250 nts) introns. The splice site selection for these exons is thought to occur via exon definition. Using genome-wide approaches we demonstrated adherence to the intronic centric proximity rule for small introns, where the 5'ss most proximal to its intronic 3'ss or the downstream 5'ss is chosen over the distal 5'ss. Thus, the building block of the spliceosome forms across the intron. This preference was most strongly observed for exons flanked by short introns. However, proximity preferences switched in favor of the upstream 5'ss for exons flanked by long introns when differences in splice site strength between competing 5'ss are mitigated. These data highlight the importance of splice site strength and the exon/intron architecture during splice site selection. These *in-silico* results were further corroborated with an *in-cellulo* experiment using minigenes splicing constructs in HeLa cells. The results provided further evidence for the exon centric proximity rule when an exon of varying length is flanked by large introns. Lastly, we also demonstrated that the intron architecture also impacts 3'ss selection but to a lesser extent. This is probably due to the more complex nature of 3'ss recognition that includes features from the polypyrimidine tract and the branch point. These results from Chapter 2 provide the basis for a unifying model of splice site proximity based on intron and exon definition modes of splice site recognition (Figure 2.5).

The findings of an exon defined proximity rule in this chapter highlight the complexity of splice recognition and ultimate alternative splicing. They provide additional evidence in support of Sterner and Berget's groundbreaking study introducing the concepts of intron and exon definition, and they add context to the Reed and Maniatis study introducing the concept of splice

site proximity preferences [53], [94]. While previous studies have provided insights into the intron and exon definition modes of splice site recognition, they were limited by the use of designer minigenes in cell free experiments [14], [54], [55], [154], making it hard to extrapolate splicing trends to the entire transcriptome. Chapter 2 tackles this limitation by using NGS genome-wide data to evaluate the splicing behavior of the whole transcriptome. It also sets the stage to determine how well this model fares in lower eukaryotes whose genomes primarily consist of small introns, thus presumably operating in the intron definition mode of splice site selection. Equivalent analyses in lower eukaryotes will in turn provide more insights into the role of intron architecture when splicing occurs much quicker, thus, potentially uncovering the hierarchy of importance of splice site selection determinants. Together, these insights have provided new information on what splicing determinants to consider when developing new splicing computational models or splicing predictors.

The influence of alternative splicing on mRNA degradation

An often-overlooked aspect of gene expression studies is the fact that RNA is constantly synthesized, processed, and degraded. Studying gene expression with a dynamic viewpoint allows one to better understand how expression profiles are established in organismal development, tissue identity, and diseased states. In doing so we may be able to leverage the lessons learned from such dynamic expression studies to develop new therapeutics. Our lab has previously established a method to study the birth of mRNA by measuring the rates of RNA transcription and pre-mRNA splicing using the uridine analog 4sU (Garibaldi et al. unpublished). In Chapter 3 we established a method to study the turnover of an mRNA by measuring the rate of degradation and the impact alternative splicing may have on mRNA stability. We carried out a 24-hour pulse-chase experiment using HepG2 cells that permitted us to obtain transcriptome-

wide mRNA half-lives and individual exon half-lives. Our analysis demonstrated a positive relationship between sequence length and mRNA stability. The analysis of exon half-lives demonstrated the importance of the 5' and 3' UTR, which also displayed a positive relationship with exon length. Analyzing mRNA degradation at the canonical exon level is novel. Interestingly, our dataset contained many exons displaying degradation kinetics significantly different from mRNA degradation kinetics, a group of exons we identify as outlier exons. This kind of analysis can infer the impact alternative splicing can have on mRNA stability. Outlier exons are significantly larger than the standard exon, regardless of whether they were first, internal, or last exons. Using the computational tool SALMON we were also able to capture mRNA isoform half-lives [127]. From an evolutionary point of view, we determined that the fastest degrading exons are the most conserved. Perhaps less surprising is the finding that outlier exons are less conserved than standard exons, suggesting that outlier exons are mostly like newer exons on an evolutionary time scale. The rise of these newer or less conserved exons often translates into new alternative splicing events, potentially giving rise to new protein isoforms that contribute to organismal phenotypic diversity and evolutionary fitness. Overall, this chapter establishes a methodology to study mRNA degradation and the impact of alternative splicing on mRNA degradation. It also sets up the stage for future studies. Coupling our methodology with long read direct RNA sequencing could generate dataset to be mined for a wealth of knowledge. For example, recent studies have shown that mRNA degradation can be driven by m6A modification differences [14]. One could adapt our methodology to directly capture mRNA isoform half-lives to evaluate degradation differences that may arise from mRNA methylation modification. One could also study the impact of splicing factors on mRNA stability. This could be done via a knockdown experiment of a splicing factor like SRSF1. SRSF1 is a splicing factor

and known protooncogene, known to influence RNA transcription, nuclear export, miRNA processing, NMD, and mRNA translation. In other words, SRSF1 plays a role in mRNA birth, life, and death. Using the methods and analysis established in this chapter in the context of SRSF1 knockdown can provide crucial insights into an important proto-oncogene, the influence of alternative splicing on mRNA kinetics, and how cellular homeostasis is established.

Splicing fidelity is linked to nutrient availability in cancer cells and contributes to methionine dependence in cancer.

In Chapter 4 we demonstrated the impact of exogenous methionine on splicing fidelity in cancer cells. Like the Warburg effect (increased glucose uptake in cancer cells), cancer cells have been shown to have a metabolic dependency of exogenous methionine. This phenomenon, known as the Hoffman effect, demonstrates that most cancer cells can only proliferate in the presence of exogenous methionine. Insights into molecular mechanisms tying methionine dependence to proliferation in cancer cells is limited. Aberrant pre-mRNA splicing has increasingly been linked to cancer biology [78], [155]. It could even be considered a new hallmark of cancer [73], [74]. In chapter 4 we present results that connect changes to pre-mRNA splicing fidelity to exogenous methionine dependence. The Kaiser lab used a methionine-dependent breast cancer cell line (MDA-MB-468) along with its revertant methionine-independent clone (MDA-MB-468res-R8) to study the molecular differences that arise when exogenous methionine is depleted [138]. A time course RNAseq experiment allowed us to capture gene expression changes upon methionine withdrawal. Our results showed the highest expression difference between the two cell lines 720 mins after methionine depletion, with the MB468 experiencing the highest impact of methionine stress. Performing a genome-wide alternative splicing analysis on the datasets allowed us to demonstrate that methionine stress in MB468 severely impacts splicing fidelity. GO analyses revealed that most of the genes impacted

by this splicing dysregulation are connected to the cell cycle and RNA processing pathways. We also observed an inverse relationship between overlapping exon skipping and intron retention events when compared between the two cell lines. These observations provided evidence that methionine dependency in cancer cells leads to splicing dysregulation which may be detrimental to cell proliferation. To measure the effect of exogenous methionine depletion on general splicing the Kaiser lab used a luciferase splicing reporter. This assay demonstrated methionine stress indeed results in a significant reduction of splicing fidelity across different cancer cell lines. Lastly, using a PRMT5 inhibitor, an important methylation factor for the SmD1 splicing factor, the Kaiser lab demonstrated methionine depletion coupled with PRMT5 inhibition leads to cell death in methionine-dependent cancer cell lines. Together, and described in Chapter 4, these findings link a molecular mechanism to the Hoffman effect. They also set up the stage for future studies. We need to further explore the enhanced effect of cancer cell death that occurs when a PRMT5 inhibitor is coupled with methionine depletion. A potential mouse study would involve feeding a cancer ridden mouse with a methionine restricted diet coupled with PRMT5 inhibitor drugs. This could potentially lead to new and improved cancer therapeutics. In addition, recent findings indicate splicing alterations and PRMT inhibition lead to the formation of neoantigens [146]. One could implement a methionine restricted diet coupled with PRMT inhibition to induce cancer specific neoantigens to develop targeted cancer treatments.

REFERENCES

- [1] M. J. Moore, C. C. Query, and P. A. Sharp, "Splicing of Precursors to mRNA by the Spliceosome," 1993, [Online]. Available: www.cshlpress.com/copyright.
- [2] S. Shukla and S. Oberdoerffer, "Co-transcriptional regulation of alternative pre-mRNA splicing," *Biochimica et biophysica acta*, vol. 1819, no. 7, pp. 673–683, Jul. 2012, doi: 10.1016/J.BBAGRM.2012.01.014.
- [3] T. W. Nilsen and B. R. Graveley, "Expansion of the eukaryotic proteome by alternative splicing," *Nature*, vol. 463, no. 7280, pp. 457–463, Jan. 2010, doi: 10.1038/nature08909.
- [4] Y. Lee and D. C. Rio, "Mechanisms and Regulation of Alternative Pre-mRNA Splicing," 2015, doi: 10.1146/annurev-biochem-060614-034316.
- [5] J. Ule and B. J. Blencowe, "Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution," *Molecular Cell*, vol. 76, no. 2, pp. 329–345, Oct. 2019, doi: 10.1016/j.molcel.2019.09.017.
- [6] L. Chen, S. J. Bush, J. M. Tovar-Corona, A. Castillo-Morales, and A. O. Urrutia, "Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity," *Mol Biol Evol*, vol. 31, no. 6, pp. 1402–1413, Jun. 2014, doi: 10.1093/molbev/msu083.
- [7] W. Jiang and L. Chen, "Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing," *Comput Struct Biotechnol J*, vol. 19, pp. 183–195, 2021, doi: 10.1016/j.csbj.2020.12.009.
- [8] H. Nguyen, U. Das, B. Wang, and J. Xie, "The matrices and constraints of GT/AG splice sites of more than 1000 species/lineages," *Gene*, vol. 660, pp. 92–101, Jun. 2018, doi: 10.1016/j.gene.2018.03.031.
- [9] H. Shenasa and K. J. Hertel, "Combinatorial regulation of alternative splicing," *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, vol. 1862, no. 11–12, pp. 194392–194392, Nov. 2019, doi: 10.1016/J.BBAGRM.2019.06.003.
- [10] D. L. Black, "Mechanisms of alternative pre-messenger RNA splicing," *Annu Rev Biochem*, vol. 72, pp. 291–336, 2003, doi: 10.1146/annurev.biochem.72.121801.161720.
- [11] M. J. Hicks, W. F. Mueller, P. J. Shepard, and K. J. Hertel, "Competing Upstream 5 Splice Sites Enhance the Rate of Proximal Splicing Downloaded from," *MOLECULAR AND CELLULAR BIOLOGY*, vol. 30, no. 8, pp. 1878–1886, 2010, doi: 10.1128/MCB.01071-09.
- [12] X. Zhan, C. Yan, X. Zhang, J. Lei, and Y. Shi, "Structures of the human pre-catalytic spliceosome and its precursor spliceosome," *Cell Res*, vol. 28, no. 12, Art. no. 12, Dec. 2018, doi: 10.1038/s41422-018-0094-7.
- [13] M. C. Wahl, C. L. Will, and R. Lührmann, "The Spliceosome: Design Principles of a Dynamic RNP Machine," *Cell*, vol. 136, no. 4, pp. 701–718, Feb. 2009, doi: 10.1016/j.cell.2009.02.009.
- [14] L. De Conti, M. Baralle, and E. Buratti, "Exon and intron definition in pre-mRNA splicing," *Wiley Interdisciplinary Reviews: RNA*, vol. 4, no. 1, pp. 49–60, Jan. 2013, doi: 10.1002/wrna.1140.
- [15] S. R. Lim and K. J. Hertel, "Commitment to Splice Site Pairing Coincides with A Complex Formation," *Molecular Cell*, vol. 15, no. 3, pp. 477–483, Aug. 2004, doi: 10.1016/j.molcel.2004.06.025.
- [16] F. E. Baralle and J. Giudice, "Alternative splicing as a regulator of development and tissue identity," *Nat Rev Mol Cell Biol*, vol. 18, no. 7, Art. no. 7, Jul. 2017, doi: 10.1038/nrm.2017.27.
- [17] P. Yang, D. Wang, and L. Kang, "Alternative splicing level related to intron size and organism complexity," *BMC Genomics*, vol. 22, p. 853, Nov. 2021, doi: 10.1186/s12864-021-08172-2.

- [18] R. J. Weatheritt, T. Sterne-Weiler, and B. J. Blencowe, "The ribosome-engaged landscape of alternative splicing," *Nat Struct Mol Biol*, vol. 23, no. 12, Art. no. 12, Dec. 2016, doi: 10.1038/nsmb.3317.
- [19] Y. Liu *et al.*, "Impact of Alternative Splicing on the Human Proteome," *Cell Rep*, vol. 20, no. 5, pp. 1229–1241, Aug. 2017, doi: 10.1016/j.celrep.2017.07.025.
- [20] S. Chaudhary *et al.*, "Alternative splicing and protein diversity: Plants versus animals," *Frontiers in Plant Science*, vol. 10, pp. 708–708, May 2019, doi: 10.3389/FPLS.2019.00708/BIBTEX.
- [21] O. Jaillon *et al.*, "Translational control of intron splicing in eukaryotes," *Nature*, vol. 451, no. 7176, Art. no. 7176, Jan. 2008, doi: 10.1038/nature06495.
- [22] L. E. Maquat, "Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics," *Nat Rev Mol Cell Biol*, vol. 5, no. 2, Art. no. 2, Feb. 2004, doi: 10.1038/nrm1310.
- [23] K. J. Hertel, "Combinatorial Control of Exon Recognition*," *Journal of Biological Chemistry*, vol. 283, no. 3, pp. 1211–1215, Jan. 2008, doi: 10.1074/jbc.R700035200.
- [24] K. Saha, W. England, M. M. Fernandez, T. Biswas, R. C. Spitale, and G. Ghosh, "Structural disruption of exonic stem-loops immediately upstream of the intron regulates mammalian splicing," *Nucleic Acids Research*, vol. 48, no. 11, pp. 6294–6309, Jun. 2020, doi: 10.1093/nar/gkaa358.
- [25] M. S. Wong, J. B. Kinney, and A. R. Krainer, "Quantitative Activity Profile and Context Dependence of All Human 5' Splice Sites," *Molecular Cell*, vol. 71, no. 6, pp. 1012–1026.e3, Sep. 2018, doi: 10.1016/j.molcel.2018.07.033.
- [26] X. Roca, R. Sachidanandam, and A. R. Krainer, "Determinants of the inherent strength of human 5 splice sites," 2005, doi: 10.1261/rna.2040605.
- [27] C. W. Smith, T. T. Chu, and B. Nadal-Ginard, "Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns," *Mol Cell Biol*, vol. 13, no. 8, pp. 4939–4952, Aug. 1993, doi: 10.1128/mcb.13.8.4939-4952.1993.
- [28] Y. Dou, K. L. Fox-Walsh, P. F. Baldi, and K. J. Hertel, "Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site," *RNA*, vol. 12, no. 12, pp. 2047–2056, Dec. 2006, doi: 10.1261/RNA.151106.
- [29] S. Wu, C. M. Romfo, T. W. Nilsen, and M. R. Green, "Functional recognition of the 3' splice site AG by the splicing factor U2AF35," *Nature*, vol. 402, no. 6763, Art. no. 6763, Dec. 1999, doi: 10.1038/45590.
- [30] G. Yeo and C. B. Burge, "Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals," 2004. [Online]. Available: www.liebertpub.com
- [31] P. J. Shepard, E.-A. Choi, A. Busch, and K. J. Hertel, "Efficient internal exon recognition depends on near equal contributions from the 3' and 5' splice sites," *Nucleic Acids Research*, vol. 39, no. 20, pp. 8928–8937, Nov. 2011, doi: 10.1093/NAR/GKR481.
- [32] C. Ha, J.-W. Kim, and J.-H. Jang, "Performance Evaluation of SpliceAI for the Prediction of Splicing of NF1 Variants," *Genes*, vol. 12, no. 9, Art. no. 9, Sep. 2021, doi: 10.3390/genes12091308.
- [33] T. D. Schaal and T. Maniatis, "Multiple Distinct Splicing Enhancers in the Protein-Coding Sequences of a Constitutively Spliced Pre-mRNA," *Molecular and Cellular Biology*, vol. 19, no. 1, pp. 261–273, Jan. 1999, doi: 10.1128/MCB.19.1.261.
- [34] P. J. Shepard and K. J. Hertel, "The SR protein family," *Genome Biology*, vol. 10, no. 10, p. 242, Oct. 2009, doi: 10.1186/gb-2009-10-10-242.
- [35] M. D. Chiara, O. Gozani, M. Bennett, P. Champion-Arnaud, L. Palandjian, and R. Reed, "Identification of proteins that interact with exon sequences, splice sites, and the branchpoint sequence during each stage of spliceosome assembly," *Mol Cell Biol*, vol. 16, no. 7, pp. 3317–3326, Jul. 1996, doi: 10.1128/MCB.16.7.3317.

- [36] X. D. Fu, A. Mayeda, T. Maniatis, and A. R. Krainer, "General splicing factors SF2 and SC35 have equivalent activities in vitro, and both affect alternative 5' and 3' splice site selection," *Proc Natl Acad Sci U S A*, vol. 89, no. 23, pp. 11224–11228, Dec. 1992, doi: 10.1073/pnas.89.23.11224.
- [37] Z. Wang and C. B. Burge, "Splicing regulation: From a parts list of regulatory elements to an integrated splicing code," *RNA*, vol. 14, no. 5, pp. 802–813, May 2008, doi: 10.1261/rna.876308.
- [38] X. D. Fu and M. Ares, "Context-dependent control of alternative splicing by RNA-binding proteins," *Nature Reviews Genetics*, vol. 15, no. 10, pp. 689–701, Oct. 2014, doi: 10.1038/nrg3778.
- [39] A. Busch and K. J. Hertel, "Evolution of SR protein and hnRNP splicing regulatory factors," *WIREs RNA*, vol. 3, no. 1, pp. 1–12, 2012, doi: 10.1002/wrna.100.
- [40] A. Mayeda and A. R. Krainer, "Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicing factor SF2," *Cell*, vol. 68, no. 2, pp. 365–375, Jan. 1992, doi: 10.1016/0092-8674(92)90477-T.
- [41] A. Mayeda, S. H. Munroe, J. F. Cáceres, and A. R. Krainer, "Function of conserved domains of hnRNP A1 and other hnRNP A/B proteins," *EMBO J*, vol. 13, no. 22, pp. 5483–5495, Nov. 1994, Accessed: May 19, 2022. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC395506/>
- [42] T. O. Tange, C. K. Damgaard, S. Guth, J. Valcárcel, and J. Kjems, "The hnRNP A1 protein regulates HIV-1 tat splicing via a novel intron silencer element," *EMBO J*, vol. 20, no. 20, pp. 5748–5758, Oct. 2001, doi: 10.1093/emboj/20.20.5748.
- [43] A. E. House and K. W. Lynch, "An exonic splicing silencer represses spliceosome assembly after ATP-dependent exon recognition," *Nat Struct Mol Biol*, vol. 13, no. 10, Art. no. 10, Oct. 2006, doi: 10.1038/nsmb1149.
- [44] S. Erkelenz *et al.*, "Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms," *RNA*, vol. 19, no. 1, pp. 96–102, Jan. 2013, doi: 10.1261/rna.037044.112.
- [45] J. M. Taliaferro *et al.*, "RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation," *Mol Cell*, vol. 64, no. 2, pp. 294–306, Oct. 2016, doi: 10.1016/j.molcel.2016.08.035.
- [46] P. J. Shepard and K. J. Hertel, "Conserved RNA secondary structures promote alternative splicing," *RNA*, vol. 14, no. 8, pp. 1463–1469, Aug. 2008, doi: 10.1261/rna.1069408.
- [47] T. Saldi, K. Riemondy, B. Erickson, and D. L. Bentley, "Alternative RNA structures formed during transcription depend on elongation rate and modify RNA processing," *Molecular Cell*, vol. 81, no. 8, pp. 1789–1801.e5, Apr. 2021, doi: 10.1016/j.molcel.2021.01.040.
- [48] E. Buratti and F. E. Baralle, "Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process DO PRE-mRNAS PRESENT SECONDARY STRUCTURE IN VIVO?," *MOLECULAR AND CELLULAR BIOLOGY*, vol. 24, no. 24, pp. 10505–10514, 2004, doi: 10.1128/MCB.24.24.10505-10514.2004.
- [49] F. Carrillo Oesterreich, L. Herzel, K. Straube, K. Hujer, J. Howard, and K. M. Neugebauer, "Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II," *Cell*, vol. 165, no. 2, pp. 372–381, Apr. 2016, doi: 10.1016/J.CELL.2016.02.045.
- [50] K. M. Neugebauer, "Nascent RNA and the Coordination of Splicing with Transcription," *Cold Spring Harb Perspect Biol*, vol. 11, no. 8, p. a032227, Aug. 2019, doi: 10.1101/cshperspect.a032227.
- [51] S. Kadener *et al.*, "Antagonistic effects of T-Ag and VP16 reveal a role for RNA pol II elongation on alternative splicing," *EMBO J*, vol. 20, no. 20, pp. 5759–5768, Oct. 2001, doi: 10.1093/emboj/20.20.5759.

- [52] N. Fong *et al.*, “Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate,” *Genes Dev.*, vol. 28, no. 23, pp. 2663–2676, Dec. 2014, doi: 10.1101/gad.252106.114.
- [53] D. A. Sterner, T. Carlo, S. M. Berget, and M. McLean, “Architectural limits on split genes,” 1996. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC26359/pdf/pq015081.pdf>
- [54] B. L. Robberson, G. J. Cote, S. M. Berget, and M. Mcclean, “Exon definition may facilitate splice site selection in RNAs with multiple exons,” *Molecular and Cellular Biology*, vol. 10, no. 1, pp. 84–94, Jan. 1990, doi: 10.1128/MCB.10.1.84-94.1990.
- [55] M. Talerico and S. M. Berget, “Intron definition in splicing of small *Drosophila* introns,” *Molecular and Cellular Biology*, vol. 14, no. 5, pp. 3434–3445, May 1994, doi: 10.1128/mcb.14.5.3434-3445.1994.
- [56] S. M. Berget, “Exon recognition in vertebrate splicing,” *Journal of Biological Chemistry*, vol. 270, no. 6, pp. 2411–2414, 1995, doi: 10.1074/jbc.270.6.2411.
- [57] M. Deutsch and M. Long, “Intron-exon structures of eukaryotic model organisms,” *Nucleic Acids Res*, vol. 27, no. 15, pp. 3219–3228, Aug. 1999, Accessed: May 02, 2022. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC148551/>
- [58] T. E. Koralewski and K. V. Krutovsky, “Evolution of Exon-Intron Structure and Alternative Splicing,” *PLOS ONE*, vol. 6, no. 3, p. e18055, Mar. 2011, doi: 10.1371/journal.pone.0018055.
- [59] K. L. Fox-Walsh, Y. Dou, B. J. Lam, S.-P. Hung, P. F. Baldi, and K. J. Hertel, “The architecture of pre-mRNAs affects mechanisms of splice-site pairing,” 2005. [Online]. Available: www.pnas.org/cgi/doi/10.1073/pnas.0508489102
- [60] M. Enculescu *et al.*, “Exon Definition Facilitates Reliable Control of Alternative Splicing in the RON Proto-Oncogene,” *Biophys J*, vol. 118, no. 8, pp. 2027–2041, Apr. 2020, doi: 10.1016/j.bpj.2020.02.022.
- [61] M. A. Arias, A. Lubkin, and L. A. Chasin, “Splicing of designer exons informs a biophysical model for exon definition,” *RNA*, vol. 21, no. 2, pp. 213–229, Feb. 2015, doi: 10.1261/rna.048009.114.
- [62] M. Schneider, C. L. Will, M. Anokhina, J. Tazi, H. Urlaub, and R. Lührmann, “Exon Definition Complexes Contain the Tri-snRNP and Can Be Directly Converted into B-like Precatalytic Splicing Complexes,” *Molecular Cell*, vol. 38, no. 2, pp. 223–235, Apr. 2010, doi: 10.1016/j.molcel.2010.02.027.
- [63] S. Sharma, L. A. Kohlstaedt, A. Damianov, D. C. Rio, and D. L. Black, “Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome,” *NATURE STRUCTURAL & MOLECULAR BIOLOGY*, vol. 15, 2008, doi: 10.1038/nsmb.1375.
- [64] X. Li *et al.*, “A unified mechanism for intron and exon definition and back-splicing,” *Nature*, vol. 573, no. 7774, Art. no. 7774, Sep. 2019, doi: 10.1038/s41586-019-1523-6.
- [65] S. Gelfman *et al.*, “Changes in exon–intron structure during vertebrate evolution affect the splicing pattern of exons,” *Genome Research*, vol. 22, no. 1, pp. 35–50, Jan. 2012, doi: 10.1101/GR.119834.110.
- [66] M. Movassat, E. Forouzmard, F. Reese, and K. J. Hertel, “Exon size and sequence conservation improves identification of splice-altering nucleotides,” *RNA*, vol. 25, no. 12, pp. 1793–1805, Dec. 2019, doi: 10.1261/rna.070987.119.
- [67] T. I. Lee and R. A. Young, “Transcriptional Regulation and Its Misregulation in Disease,” *Cell*, vol. 152, no. 6, pp. 1237–1251, Mar. 2013, doi: 10.1016/j.cell.2013.02.014.
- [68] T. Sanda and W. Z. Leong, “TAL1 as a master oncogenic transcription factor in T-cell acute lymphoblastic leukemia,” *Exp Hematol*, vol. 53, pp. 7–15, Sep. 2017, doi: 10.1016/j.exphem.2017.06.001.
- [69] V. K. Nagarajan, C. I. Jones, S. F. Newbury, and P. J. Green, “XRN 5′→3′ exoribonucleases: structure, mechanisms and functions,” *Biochim Biophys Acta*, vol. 1829, no. 6–7, pp. 590–603, Jul. 2013, doi: 10.1016/j.bbagr.2013.03.005.

- [70] J. Houseley and D. Tollervy, "The many pathways of RNA degradation," *Cell*, vol. 136, no. 4, pp. 763–776, Feb. 2009, doi: 10.1016/j.cell.2009.01.019.
- [71] A. Navickas, S. Chamois, R. Saint-Fort, J. Henri, C. Torchet, and L. Benard, "No-Go Decay mRNA cleavage in the ribosome exit tunnel produces 5'-OH ends phosphorylated by Trl1," *Nat Commun*, vol. 11, no. 1, Art. no. 1, Jan. 2020, doi: 10.1038/s41467-019-13991-9.
- [72] H. Bae and J. Collier, "Codon optimality-mediated mRNA degradation: Linking translational elongation to mRNA stability," *Molecular Cell*, vol. 82, no. 8, pp. 1467–1476, Apr. 2022, doi: 10.1016/j.molcel.2022.03.032.
- [73] S. Oltean and D. O. Bates, "Hallmarks of alternative splicing in cancer," *Oncogene*, vol. 33, no. 46, Art. no. 46, Nov. 2014, doi: 10.1038/onc.2013.533.
- [74] M. Lodomery, "Aberrant Alternative Splicing Is Another Hallmark of Cancer," *International Journal of Cell Biology*, vol. 2013, p. e463786, Sep. 2013, doi: 10.1155/2013/463786.
- [75] L. Wang *et al.*, "SF3B1 and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia," *New England Journal of Medicine*, vol. 365, no. 26, pp. 2497–2506, Dec. 2011, doi: 10.1056/NEJMoa1109016.
- [76] K. Yoshida *et al.*, "Frequent pathway mutations of splicing machinery in myelodysplasia," *Nature*, vol. 478, no. 7367, Art. no. 7367, Oct. 2011, doi: 10.1038/nature10496.
- [77] M. A. Rahman, A. R. Krainer, and O. Abdel-Wahab, "SnapShot: Splicing Alterations in Cancer," *Cell*, vol. 180, no. 1, pp. 208–208.e1, Jan. 2020, doi: 10.1016/j.cell.2019.12.011.
- [78] S. C. Bonnal, I. López-Oreja, and J. Valcárcel, "Roles and mechanisms of alternative splicing in cancer — implications for care," *Nat Rev Clin Oncol*, vol. 17, no. 8, Art. no. 8, Aug. 2020, doi: 10.1038/s41571-020-0350-x.
- [79] G. Biamonti, M. Catillo, D. Pignataro, A. Montecucco, and C. Ghigna, "The alternative splicing side of cancer," *Semin Cell Dev Biol*, vol. 32, pp. 30–36, Aug. 2014, doi: 10.1016/j.semcdb.2014.03.016.
- [80] W. Li *et al.*, "Profiling PRMT methylome reveals roles of hnRNPA1 arginine methylation in RNA splicing and cell growth," *Nat Commun*, vol. 12, no. 1, Art. no. 1, Mar. 2021, doi: 10.1038/s41467-021-21963-1.
- [81] A. Kahles *et al.*, "Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients," *Cancer Cell*, vol. 34, no. 2, pp. 211–224.e6, Aug. 2018, doi: 10.1016/j.ccell.2018.07.001.
- [82] S. Shuai *et al.*, "The U1 spliceosomal RNA is recurrently mutated in multiple cancers," *Nature*, vol. 574, no. 7780, Art. no. 7780, Oct. 2019, doi: 10.1038/s41586-019-1651-z.
- [83] M. Seiler *et al.*, "Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types," *Cell Reports*, vol. 23, no. 1, pp. 282–296.e4, Apr. 2018, doi: 10.1016/j.celrep.2018.01.088.
- [84] M. V. Liberti and J. W. Locasale, "The Warburg Effect: How Does it Benefit Cancer Cells?," *Trends in Biochemical Sciences*, vol. 41, no. 3, pp. 211–218, Mar. 2016, doi: 10.1016/j.tibs.2015.12.001.
- [85] D. Hanahan and R. A. Weinberg, "Hallmarks of Cancer: The Next Generation," *Cell*, vol. 144, no. 5, pp. 646–674, Mar. 2011, doi: 10.1016/j.cell.2011.02.013.
- [86] P. Kaiser, "Methionine Dependence of Cancer," *Biomolecules*, vol. 10, no. 4, Art. no. 4, Apr. 2020, doi: 10.3390/biom10040568.
- [87] R. M. Hoffman, "Is the Hoffman Effect for Methionine Overuse Analogous to the Warburg Effect for Glucose Overuse in Cancer?," in *Methionine Dependence of Cancer and Aging: Methods and Protocols*, R. M. Hoffman, Ed. New York, NY: Springer, 2019, pp. 273–278. doi: 10.1007/978-1-4939-8796-2_21.
- [88] A. A. Parkhitko, P. Jouandin, S. E. Mohr, and N. Perrimon, "Methionine metabolism and methyltransferases in the regulation of aging and lifespan extension across species," *Aging Cell*, vol. 18, no. 6, p. e13034, Dec. 2019, doi: 10.1111/acel.13034.

- [89] S. L. Borrego *et al.*, “Metabolic changes associated with methionine stress sensitivity in MDA-MB-468 breast cancer cells,” *Cancer Metab*, vol. 4, p. 9, 2016, doi: 10.1186/s40170-016-0148-6.
- [90] K. Booher, D.-W. Lin, S. L. Borrego, and P. Kaiser, “Downregulation of Cdc6 and pre-replication complexes in response to methionine stress in breast cancer cells,” *Cell Cycle*, vol. 11, no. 23, pp. 4414–4423, Dec. 2012, doi: 10.4161/cc.22767.
- [91] N. Stopa, J. E. Krebs, and D. Shechter, “The PRMT5 arginine methyltransferase: many roles in development, cancer and beyond,” *Cell Mol Life Sci*, vol. 72, no. 11, pp. 2041–2059, Jun. 2015, doi: 10.1007/s00018-015-1847-9.
- [92] J. Li *et al.*, “Roles of alternative splicing in modulating transcriptional regulation,” *BMC Systems Biology*, vol. 11, no. 5, p. 89, Oct. 2017, doi: 10.1186/s12918-017-0465-6.
- [93] A. Busch and K. J. Hertel, “Splicing predictions reliably classify different types of alternative splicing,” *RNA*, vol. 21, no. 5, pp. 813–823, May 2015, doi: 10.1261/RNA.048769.114.
- [94] R. Reed and T. Maniatis, “A role for exon sequences and splice-site proximity in splice-site selection,” *Cell*, vol. 46, no. 5, pp. 681–690, Aug. 1986, doi: 10.1016/0092-8674(86)90343-0.
- [95] A. Busch and K. J. Hertel, “HEXEvent: a database of Human EXon splicing Events,” *Nucleic Acids Research*, vol. 41, no. D1, pp. D118–D124, Oct. 2012, doi: 10.1093/nar/gks969.
- [96] D. Wang, “IntronDB: a database for eukaryotic intron features,” *Bioinformatics*, vol. 35, no. 21, pp. 4400–4401, Nov. 2019, doi: 10.1093/bioinformatics/btz242.
- [97] A. Piovesan, M. Caracausi, and F. Antonaros, “GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics,” vol. 2016, 2016, doi: 10.1093/database/baw153.
- [98] K. M. Lang and R. A. Spritz, “RNA Splice Site Selection: Evidence for a 5′ → 3′ Scanning Model,” *Science*, vol. 220, no. 4604, pp. 1351–1355, Jun. 1983, doi: 10.1126/science.6304877.
- [99] A. A. Pai, T. Henriques, K. McCue, A. Burkholder, K. Adelman, and C. B. Burge, “The kinetics of pre-mRNA splicing in the *Drosophila* genome and the influence of gene architecture,” *eLife*, vol. 6, Dec. 2017, doi: 10.7554/eLife.32537.
- [100] H. L. Drexler, K. Choquet, and L. S. Churchman, “Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores,” *Molecular Cell*, vol. 77, no. 5, pp. 985–998.e8, Mar. 2020, doi: 10.1016/j.molcel.2019.11.017.
- [101] J. Singh and R. A. Padgett, “Rates of in situ transcription and splicing in large human genes,” *Nat Struct Mol Biol*, vol. 16, no. 11, Art. no. 11, Nov. 2009, doi: 10.1038/nsmb.1666.
- [102] M. Rabani *et al.*, “High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies,” *Cell*, vol. 159, no. 7, pp. 1698–1710, Dec. 2014, doi: 10.1016/j.cell.2014.11.015.
- [103] L. Wachutka, L. Caizzi, J. Gagneur, and P. Cramer, “Global donor and acceptor splicing site kinetics in human cells,” *Elife*, vol. 8, p. e45056, Apr. 2019, doi: 10.7554/eLife.45056.
- [104] M. Amit *et al.*, “Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition,” *Cell Reports*, vol. 1, no. 5, pp. 543–556, May 2012, doi: 10.1016/j.celrep.2012.03.013.
- [105] L. Tammer *et al.*, “Gene architecture directs splicing outcome in separate nuclear spatial regions,” *Molecular Cell*, vol. 82, no. 5, pp. 1021–1034.e8, Mar. 2022, doi: 10.1016/j.molcel.2022.02.001.
- [106] M. Freund *et al.*, “A novel approach to describe a U1 snRNA binding site,” *Nucleic Acids Research*, vol. 31, no. 23, pp. 6963–6975, Dec. 2003, doi: 10.1093/nar/gkg901.
- [107] A. M. Olthof, K. C. Hyatt, and R. N. Kanadia, “Minor intron splicing revisited: identification of new minor intron-containing genes and tissue-dependent retention and alternative splicing of minor introns,” *BMC Genomics*, vol. 20, no. 1, p. 686, Aug. 2019, doi: 10.1186/s12864-019-6046-x.

- [108] J. V. Geisberg, Z. Moqtaderi, X. Fan, F. Ozsolak, and K. Struhl, "Global Analysis of mRNA Isoform Half-Lives Reveals Stabilizing and Destabilizing Elements in Yeast," *Cell*, vol. 156, no. 4, pp. 812–824, Feb. 2014, doi: 10.1016/j.cell.2013.12.026.
- [109] H. Tani and N. Akimitsu, "Genome-wide technology for determining RNA stability in mammalian cells: historical perspective and recent advantages based on modified nucleotide labeling," *RNA Biol*, vol. 9, no. 10, pp. 1233–1238, Oct. 2012, doi: 10.4161/rna.22036.
- [110] E. E. Duffy, J. A. Schofield, and M. D. Simon, "Gaining insight into transcriptome-wide RNA population dynamics through the chemistry of 4-thiouridine," *WIREs RNA*, vol. 10, no. 1, p. e1513, 2019, doi: 10.1002/wrna.1513.
- [111] K. Burger *et al.*, "4-thiouridine inhibits rRNA synthesis and causes a nucleolar stress response," *RNA Biology*, vol. 10, no. 10, pp. 1623–1630, Oct. 2013, doi: 10.4161/rna.26214.
- [112] J. A. Schofield, E. E. Duffy, L. Kiefer, M. C. Sullivan, and M. D. Simon, "TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding," *Nat Methods*, vol. 15, no. 3, Art. no. 3, Mar. 2018, doi: 10.1038/nmeth.4582.
- [113] A. Frankish *et al.*, "GENCODE reference annotation for the human and mouse genomes," *Nucleic Acids Research*, vol. 47, no. D1, pp. D766–D773, Jan. 2019, doi: 10.1093/nar/gky955.
- [114] L. Feng and D.-K. Niu, "Relationship Between mRNA Stability and Length: An Old Question with a New Twist," *Biochem Genet*, vol. 45, no. 1, pp. 131–137, Feb. 2007, doi: 10.1007/s10528-006-9059-5.
- [115] D. Dominguez *et al.*, "Sequence, Structure, and Context Preferences of Human RNA Binding Proteins," *Molecular Cell*, vol. 70, no. 5, pp. 854–867.e9, Jun. 2018, doi: 10.1016/j.molcel.2018.05.001.
- [116] K. D. Meyer, Y. Saletore, P. Zumbo, O. Elemento, C. E. Mason, and S. R. Jaffrey, "Comprehensive Analysis of mRNA Methylation Reveals Enrichment in 3' UTRs and near Stop Codons," *Cell*, vol. 149, no. 7, pp. 1635–1646, Jun. 2012, doi: 10.1016/j.cell.2012.05.003.
- [117] H. Keren, G. Lev-Maor, and G. Ast, "Alternative splicing and evolution: diversification, exon definition and function," *Nat Rev Genet*, vol. 11, no. 5, Art. no. 5, May 2010, doi: 10.1038/nrg2776.
- [118] K. Wei, T. Zhang, and L. Ma, "Divergent and convergent evolution of housekeeping genes in human–pig lineage," *PeerJ*, vol. 6, p. e4840, May 2018, doi: 10.7717/peerj.4840.
- [119] M. B. Clark *et al.*, "Genome-wide analysis of long noncoding RNA stability," *Genome Res*, vol. 22, no. 5, pp. 885–898, May 2012, doi: 10.1101/gr.131037.111.
- [120] P. Zuccotti, D. Peroni, V. Potrich, A. Quattrone, and E. Dassi, "Hyperconserved Elements in Human 5'UTRs Shape Essential Post-transcriptional Regulatory Networks," *Frontiers in Molecular Biosciences*, vol. 7, 2020, Accessed: Dec. 09, 2022. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fmolb.2020.00220>
- [121] K. Leppek, R. Das, and M. Barna, "Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them," *Nat Rev Mol Cell Biol*, vol. 19, no. 3, Art. no. 3, Mar. 2018, doi: 10.1038/nrm.2017.103.
- [122] C. E. Vejnar *et al.*, "Genome wide analysis of 3' UTR sequence elements and proteins regulating mRNA stability during maternal-to-zygotic transition in zebrafish," *Genome Res.*, vol. 29, no. 7, pp. 1100–1114, Jul. 2019, doi: 10.1101/gr.245159.118.
- [123] B. Bae and P. Miura, "Emerging Roles for 3' UTRs in Neurons," *Int J Mol Sci*, vol. 21, no. 10, p. 3413, May 2020, doi: 10.3390/ijms21103413.
- [124] A. E. Brinegar and T. A. Cooper, "Roles for RNA-binding proteins in development and disease," *Brain Research*, vol. 1647, pp. 1–8, Sep. 2016, doi: 10.1016/j.brainres.2016.02.050.

- [125] O. D. Schwich *et al.*, “SRSF3 and SRSF7 modulate 3’UTR length through suppression or activation of proximal polyadenylation sites and regulation of CFIm levels,” *Genome Biology*, vol. 22, no. 1, p. 82, Mar. 2021, doi: 10.1186/s13059-021-02298-y.
- [126] I. Aznarez *et al.*, “Mechanism of Nonsense-Mediated mRNA Decay Stimulation by Splicing Factor SRSF1,” *Cell Rep*, vol. 23, no. 7, pp. 2186–2198, May 2018, doi: 10.1016/j.celrep.2018.04.039.
- [127] D. Sarantopoulou, T. G. Brooks, S. Nayak, A. Mrčela, N. F. Lahens, and G. R. Grant, “Comparative evaluation of full-length isoform quantification from RNA-Seq,” *BMC Bioinformatics*, vol. 22, no. 1, p. 266, May 2021, doi: 10.1186/s12859-021-04198-1.
- [128] R. De Paoli-Iseppi, J. Gleeson, and M. B. Clark, “Isoform Age - Splice Isoform Profiling Using Long-Read Technologies,” *Frontiers in Molecular Biosciences*, vol. 8, 2021, Accessed: Dec. 09, 2022. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fmolb.2021.711733>
- [129] G. Hanson and J. Collier, “Codon optimality, bias and usage in translation and mRNA decay,” *Nat Rev Mol Cell Biol*, vol. 19, no. 1, Art. no. 1, Jan. 2018, doi: 10.1038/nrm.2017.91.
- [130] S. H. Boo and Y. K. Kim, “The emerging role of RNA modifications in the regulation of mRNA stability,” *Exp Mol Med*, vol. 52, no. 3, Art. no. 3, Mar. 2020, doi: 10.1038/s12276-020-0407-z.
- [131] K. Xiang and D. P. Bartel, “The molecular basis of coupling between poly(A)-tail length and translational efficiency,” *Elife*, vol. 10, p. e66493, Jul. 2021, doi: 10.7554/eLife.66493.
- [132] E. L. Van Nostrand *et al.*, “A large-scale binding and functional map of human RNA-binding proteins,” *Nature*, vol. 583, no. 7818, Art. no. 7818, Jul. 2020, doi: 10.1038/s41586-020-2077-3.
- [133] A. Garibaldi, F. Carranza, and K. J. Hertel, “Isolation of Newly Transcribed RNA Using the Metabolic Label 4-Thiouridine,” *Methods Mol Biol*, vol. 1648, pp. 169–176, 2017, doi: 10.1007/978-1-4939-7204-3_13.
- [134] T. Sugimura, S. M. Birnbaum, M. Winitz, and J. P. Greenstein, “Quantitative nutritional studies with water-soluble, chemically defined diets. IX. Further studies on d-glucosaminecontaining diets,” *Arch Biochem Biophys*, vol. 83, pp. 521–527, Aug. 1959, doi: 10.1016/0003-9861(59)90060-8.
- [135] L. Lauinger and P. Kaiser, “Sensing and Signaling of Methionine Metabolism,” *Metabolites*, vol. 11, no. 2, Art. no. 2, Feb. 2021, doi: 10.3390/metabo11020083.
- [136] S. M. Sanderson, X. Gao, Z. Dai, and J. W. Locasale, “Methionine metabolism in health and cancer: a nexus of diet and precision medicine,” *Nat Rev Cancer*, vol. 19, no. 11, Art. no. 11, Nov. 2019, doi: 10.1038/s41568-019-0187-8.
- [137] H. Dvinge, E. Kim, O. Abdel-Wahab, and R. K. Bradley, “RNA splicing factors as oncoproteins and tumour suppressors,” *Nat Rev Cancer*, vol. 16, no. 7, Art. no. 7, Jul. 2016, doi: 10.1038/nrc.2016.51.
- [138] S. L. Borrego *et al.*, “Lipid remodeling in response to methionine stress in MDA-MBA-468 triple-negative breast cancer cells,” *Journal of Lipid Research*, vol. 62, Jan. 2021, doi: 10.1016/j.jlr.2021.100056.
- [139] P. Sachamitr *et al.*, “PRMT5 inhibition disrupts splicing and stemness in glioblastoma,” *Nat Commun*, vol. 12, no. 1, Art. no. 1, Feb. 2021, doi: 10.1038/s41467-021-21204-5.
- [140] S. X. Ge, D. Jung, and R. Yao, “ShinyGO: a graphical gene-set enrichment tool for animals and plants,” *Bioinformatics*, vol. 36, no. 8, pp. 2628–2629, Apr. 2020, doi: 10.1093/bioinformatics/btz931.
- [141] G. Meister, C. Eggert, D. Bühler, H. Brahms, C. Kambach, and U. Fischer, “Methylation of Sm proteins by a complex containing PRMT5 and the putative U snRNP assembly factor pICln,” *Curr Biol*, vol. 11, no. 24, pp. 1990–1994, Dec. 2001, doi: 10.1016/s0960-9822(01)00592-9.
- [142] I. Younis, M. Berg, D. Kaida, K. Dittmar, C. Wang, and G. Dreyfuss, “Rapid-Response Splicing Reporter Screens Identify Differential Regulators of Constitutive and Alternative Splicing,”

- Molecular and Cellular Biology*, vol. 30, no. 7, pp. 1718–1728, Apr. 2010, doi: 10.1128/MCB.01301-09.
- [143] M. Vinet *et al.*, “Protein arginine methyltransferase 5: A novel therapeutic target for triple-negative breast cancers,” *Cancer Medicine*, vol. 8, no. 5, pp. 2414–2428, 2019, doi: 10.1002/cam4.2114.
- [144] K. Marjon, P. Kalev, and K. Marks, “Cancer Dependencies: PRMT5 and MAT2A in MTAP/p16-Deleted Cancers,” *Annual Review of Cancer Biology*, vol. 5, no. 1, pp. 371–390, 2021, doi: 10.1146/annurev-cancerbio-030419-033444.
- [145] P. Kalev *et al.*, “MAT2A Inhibition Blocks the Growth of MTAP-Deleted Cancer Cells by Reducing PRMT5-Dependent mRNA Splicing and Inducing DNA Damage,” *Cancer Cell*, vol. 39, no. 2, pp. 209–224.e11, Feb. 2021, doi: 10.1016/j.ccell.2020.12.010.
- [146] S. X. Lu *et al.*, “Pharmacologic modulation of RNA splicing enhances anti-tumor immunity,” *Cell*, vol. 184, no. 15, pp. 4032–4047.e31, Jul. 2021, doi: 10.1016/j.cell.2021.05.038.
- [147] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype,” *Nat Biotechnol*, vol. 37, no. 8, Art. no. 8, Aug. 2019, doi: 10.1038/s41587-019-0201-4.
- [148] A. Dobin *et al.*, “STAR: ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013, doi: 10.1093/bioinformatics/bts635.
- [149] Y. Liao, G. K. Smyth, and W. Shi, “The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote,” *Nucleic Acids Res*, vol. 41, no. 10, p. e108, May 2013, doi: 10.1093/nar/gkt214.
- [150] D. Risso, J. Ngai, T. P. Speed, and S. Dudoit, “Normalization of RNA-seq data using factor analysis of control genes or samples,” *Nat Biotechnol*, vol. 32, no. 9, pp. 896–902, Sep. 2014, doi: 10.1038/nbt.2931.
- [151] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biology*, vol. 15, no. 12, p. 550, Dec. 2014, doi: 10.1186/s13059-014-0550-8.
- [152] S. Shen *et al.*, “rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 51, pp. E5593–E5601, Dec. 2014, doi: 10.1073/pnas.1419161111.
- [153] Q. Liu, L. Fang, and C. Wu, “Alternative Splicing and Isoforms: From Mechanisms to Diseases,” *Genes*, vol. 13, no. 3, Art. no. 3, Mar. 2022, doi: 10.3390/genes13030401.
- [154] K. L. Fox-Walsh and K. J. Hertel, “Splice-site pairing is an intrinsically high fidelity process,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 6, pp. 1766–1771, Feb. 2009, doi: 10.1073/pnas.0813128106.
- [155] M. Blijlevens, J. Li, and V. W. van Beusechem, “Biology of the mRNA Splicing Machinery and Its Dysregulation in Cancer Providing Therapeutic Opportunities,” *International Journal of Molecular Sciences*, vol. 22, no. 10, Art. no. 10, Jan. 2021, doi: 10.3390/ijms22105110.

APPENDIX A

Isolation of Newly Transcribed RNA Using the Metabolic Label 4-thiouridine

Chapter 13

Isolation of Newly Transcribed RNA Using the Metabolic Label 4-Thiouridine

Angela Garibaldi, Francisco Carranza, and Klemens J. Hertel

Abstract

Isolation of newly transcribed RNA is an invaluable approach that can be used to study the dynamic life of RNA *in cellulo*. Traditional methods of whole-cell RNA extraction limit subsequent gene expression analyses to the steady-state levels of RNA abundance, which often masks changes in RNA synthesis and processing. This chapter describes a methodology with low cytotoxicity that permits the labeling and isolation of nascent pre-mRNA in cell culture. The resulting isolate is suitable for use in a series of downstream applications aimed at studying changes in RNA synthesis, processing, or stability.

Key words 4sU, 4sU-seq, Mammalian cells, Nascent RNA, Decay, Transcription, Nascent pre-mRNA, mRNA processing, Metabolic labeling, 4-Thiouridine

1 Introduction

The majority of gene expression research focuses on RNA transcript abundance at a steady-state level, providing only a snapshot of the cellular state. This glimpse of transcript abundance in the cell limits the understanding of regulation to whether a gene is generally up or down regulated. This obscures whether a change in gene expression is due to differences in the rate of transcription, the rate of degradation, or both. Previous approaches aimed at elucidating the dynamics of cotranscriptional pre-mRNA processing focused on a variety of immunoprecipitation and cell fractionation techniques following a chosen pathway induction (LPS stimulation) [1–3]. Likewise, pulse-chase experiments using well-known transcription inhibitors such as Actinomycin D have been frequently used to measure mRNA stability and degradation [2, 4]. While providing critical advances to the fundamental understanding

Angela Garibaldi and Francisco Carranza contributed equally to this work.

Yongsheng Shi (ed.), *mRNA Processing: Methods and Protocols*, Methods in Molecular Biology, vol. 1648, DOI 10.1007/978-1-4939-7204-3_13, © Springer Science+Business Media LLC 2017

Table 1
Recommended 4sU concentrations [5]

Duration of labeling [min]	Recommended 4sU concentration [μ M]
120	100–200
60	200–500
15–30	500–1000
<10	500–20,000

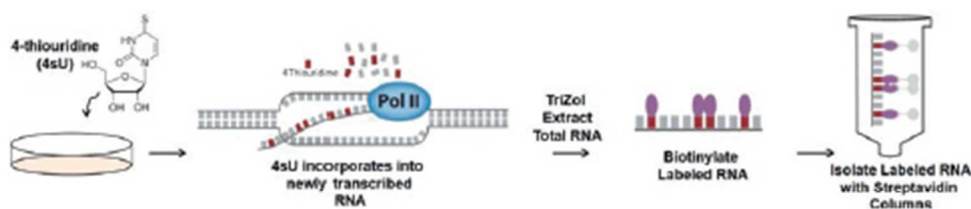


Fig. 1 Conceptual workflow of 4sU labeling and isolation of newly transcribed RNA

of RNA dynamics, these methods are limited by cytotoxicity and a lack of kinetic resolution [4]. The emergence of next-generation sequencing technologies, in conjunction with the uridine analog 4-thiouridine (4sU) as a metabolic label, has opened up an exciting avenue to studying genome-wide RNA kinetics at high resolution [5–7].

4sU can be used to metabolically label and track an RNA from synthesis to degradation by simply adding 4sU to mammalian cell culture media. 4sU is immediately taken up by cells, phosphorylated, and incorporated into any newly transcribed RNA. 4sU-labeled RNA can be tracked and isolated to study nascent RNA behavior using RNA sequencing. Alternatively, media replacement after a longer incubation period in 4sU media can be used to study the half-life and degradation of an RNA. Depending on the experimental approach taken, appropriate concentrations of 4sU should be selected for various cell types and incubation times to minimize off-target effects (*see* Table 1) [8].

Once whole-cell RNA is extracted, the 4sU-labeled RNA can be biotinylated via its sulfhydryl group and selectively isolated using streptavidin-coated magnetic beads. Given the strong biotin/streptavidin interaction, 4sU labeled RNA can be stringently washed. Eluted 4sU labeled RNA can then be used in subsequent qRT-PCR and RNA-seq experiments, with or without ribosomal RNA depletion (*see* Fig. 1). Given the fact that 4sU can be used to mark nascent transcripts, the use of the 4sU labeling protocol can lead to a wealth of new findings that directly relate to immediate changes in gene expression.

2 Materials

All materials must be sterile, RNase-free, molecular biology grade. Large quantities of TRIzol reagent may be used for experiments. In case of contact with skin/eyes, have a polyethylene glycol 300 or 400 in industrial methylated spirits (70:30) solution prepared before proceeding.

2.1 4sU Labeling of Cells

1. 4-Thiouridine dissolved in sterile RNase-free water to 50 mM. Store in small aliquots at -20°C , thawing only once.

2.2 Total RNA Extraction

1. TRIzol.
2. 75% EtOH (ethanol).
3. RNase-free water.
4. RNA Precipitation Solution: 0.8 M NaCl, 1.2 M NaCitrate.
5. (*Optional*) TE: 10 mM Tris, 1 mM EDTA.

2.3 Biotinylation of 4sU-Labeled RNA

1. EZ-Link Biotin-HPDP. Make stock aliquots 1 mg/mL dissolved in Dimethylformamide (*see Note 1*) and store at 4°C .
2. 10 \times Biotinylation Buffer: 100 mM Tris pH 7.4, 10 mM EDTA. Store in aliquots of ~ 1 mL at 4°C .
3. 5 M NaCl.
4. 75% EtOH.
5. (*Optional*) Phase Lock Gel Heavy Tubes (2.0 mL).

2.4 Separation of Labeled and Unlabeled RNA Using Streptavidin-Coated Magnetic Beads

1. μ Macs Streptavidin Kit (*see Note 2*).
2. 1 \times Washing Buffer: 100 mM Tris pH 7.5, 10 mM EDTA, 1 M NaCl, 0.1% Tween20.
3. 100 mM Dithiothreitol (DTT) in RNase-free water.
4. Magnetic Separator and Stand (2 each). Alternatively, one of each is included in the starter kit.
5. (*Optional*) RNeasy MinElute Cleanup Kit.

2.5 Recovery of Unlabeled, Unbound RNA

1. Phenol/chloroform pH 6.7.
2. Isopropanol.
3. EtOH.

3 Methods

3.1 4sU Labeling of Nascent RNA

1. Plate a number of cells of the desired cell type in either a 10 cm or 15 cm tissue culture plate that will reach 70–80% confluency after 24 h. For a 10 cm plate, use at least 10 mL of culture medium. For a 15 cm plate, use at least 20 mL of culture medium.

2. For a 10 cm dish, a minimum of 5 mL of culture medium containing 4sU is needed. For a 15 cm dish, a minimum of 10 mL of culture medium containing 4sU is needed.
3. Once cells reach 70–80% confluency, transfer 5 mL or 10 mL of culture medium from the plate to a clean 15 mL conical tube.
4. Add 4sU to the culture medium in the conical tube and pipette up and down with a serological pipette to mix thoroughly. Refer to Table 1 for general guidelines for 4sU concentrations (*see Note 3*).
5. Aspirate the remaining unlabeled culture medium from the plate. Add the culture medium containing 4sU to the cells (*see Note 4*).
6. Incubate cells with 4sU culture medium for the desired amount of time. A longer incubation period is recommended for RNA decay studies (*see Note 5*).
7. Quench the reaction by quickly aspirating the 4sU culture medium and adding 3 mL of TRIzol for 10 cm plate, or 5 mL of TRIzol for 15 cm dishes.
8. Ensure the entire plate is covered by TRIzol and allow it to sit for 2–5 min for complete cell lysis.
9. Pipette the cell/TRIzol lysate to homogenize the cells and get all cells off the plate. Transfer the lysate to a 15 mL conical tube.
10. Immediately extract total RNA from TRIzol samples, or store at -80°C between 6 months and 1 year (*see Note 6*).

3.2 Total RNA Extraction

1. Transfer 1 mL of TRIzol sample to each of (3) 1.5 mL microcentrifuge tubes.
2. Add 0.2 mL chloroform per mL TRIzol and shake vigorously for 15 s.
3. Incubate at room temperature for 2–3 min.
4. Centrifuge at $20,000 \times g$ for 15 min at 4°C (*see Note 7*).
5. Transfer aqueous upper phase (containing the RNA) to a new tube.
6. Add $\frac{1}{2}$ the reaction volume of both RNA precipitation buffer and isopropanol (e.g., to 3 mL of supernatant add 1.5 mL RNA Precipitation Solution and 1.5 mL isopropanol).
7. Invert to mix well.
8. Incubate at room temperature for 10 min.
9. Centrifuge at $20,000 \times g$ for 10 min at 4°C .
10. Immediately remove the supernatant.
11. Wash with an equal volume of 75% EtOH.

12. Centrifuge at $20,000 \times g$ for 10 min at 4 °C.
13. Immediately remove the supernatant.
14. Centrifuge again briefly to spin down remaining EtOH.
15. Remove remaining ethanol by pipetting using 200 μ L pipette. Repeat step using 20 μ L pipette (*see Note 8*).
16. Add 100 μ L of 1 \times TE or RNase-free water per 100 μ g expected RNA yield.
17. If needed, dissolve RNA by heating to 65 °C for 10 min.
18. Use a NanoDrop spectrophotometer to measure RNA yield. This RNA can be stored at -80 °C for at least 3 months with minimal freeze-thaws.

3.3 Biotinylation of 4sU-Labeled RNA

1. Labeling Reaction (use 60–100 μ g total RNA):
 - 2 μ L Biotin-HPDP (1 mg/mL DMF) per 1 μ g RNA.
 - 1 μ L 10 \times Biotinylation Buffer per 1 μ g RNA.
 - Bring up to 7 μ L with RNase-free water per 1 μ g RNA.
2. Rotate at room temperature in the dark for at least 1.5 h (*see Note 9*).
3. Add an equal volume of Phenol/Chloroform pH 6.7.
4. Mix vigorously by vortex or by manually shaking.
5. Incubate for 2–3 min at room temperature until phases begin to separate and bubbles start to disappear.
6. Centrifuge at full speed ($20,000 \times g$) for 5 min at 4 °C.
7. Carefully transfer upper phase into new tubes (*see Note 10*).
8. *RNA precipitation*: Add 1/10 the reaction volume of 5 M NaCl.
9. Add an equal volume of isopropanol, invert to mix well.
10. Centrifuge at $20,000 \times g$ for 20 min at 4 °C.
11. Remove the supernatant. Add an equal volume of 75% EtOH.
12. Centrifuge at $20,000 \times g$ for 10 min at 4 °C.
13. Remove EtOH completely and resuspend the RNA pellet at approximately 1 μ g/ μ L with RNase-free water or TE.

3.4 Separation of Labeled and Unlabeled RNA Using Streptavidin-Coated Magnetic Beads

1. Heat biotinylated RNA samples to 65 °C for 10 min and immediately place on ice for 5 min.
2. Add up to 100 μ g (max. 100 μ L) of biotinylated RNA to 100 μ L of streptavidin beads (*see Note 11*).
3. Incubate at room temperature with rotation for 15 min.
4. Place μ Macs columns into magnetic stand. Process no more than eight samples at a time (*see Note 12*).

5. Add 0.9 mL of washing buffer to columns to prerun and equilibrate (*see Note 13*).
6. Apply bead-bound RNA to the columns.
7. For recovery of unlabeled/unbound RNA, collect this flow-through and *see* Subheading 3.5. Otherwise, discard the flow-through.
8. Place tubes or alternative collection apparatus underneath columns to catch the wash flow-through.
9. Wash 3× with 0.9 mL 65 °C washing buffer. *Optional*: For recovery of unlabeled RNA, collect the first wash and *see* Subheading 3.5.
10. Wash 3× with 0.9 mL room temperature washing buffer.
11. Elute the labeled RNA by placing the 1.5 mL microcentrifuge tubes underneath the columns and adding 100 µL 100 mM DTT to the columns (*see Note 14*).
12. Perform a second DTT elution into the same tubes 3–5 min later.
13. Immediately perform EtOH precipitation with 2.5 V 100% EtOH and 10 µg glycogen.
14. Precipitate overnight at –20 °C.
15. Spin at 20,000 × *g* for 15 min at 4 °C.
16. Wash with 75% EtOH.
17. Spin at 20,000 × *g* for 5 min at 4 °C.
18. Remove all EtOH using technique used in Subheading 3.2, **step 15** and resuspend in ~30 µL RNase-free water.
19. Spec labeled RNA with NanoDrop (*see Note 15*).

3.5 Recovery of Unlabeled, Unbound RNA (Optional)

1. For recovery of >90% of unbound RNA, collect the flow-through and the first wash for subsequent precipitation.
2. Combine the two fractions and recover the unbound RNA by isopropanol/EtOH precipitation as performed after the biotinylation reaction (*see* Subheading 3.3). Omit the addition of NaCl; the washing buffer has sufficient NaCl.

3.6 Validation

1. Validate with RT-PCR/qPCR by comparing labeled RNA to total RNA or unlabeled RNA for genes/transcripts of interest.

4 Notes

1. Gentle warming will ensure complete solubilization. Store aliquots at 4 °C. Alternatively, store 20 mg/mL at –20 °C. Do not use any polystyrene serological pipette in this process as DMF will degrade the plastic, leading to plastic residues that may inhibit biotinylation.

2. Per conversations with Miltenyi tech support, the beads are subject to the expiration date on the box. Columns, however, are good for 3 years. At the time of publication, beads are not sold separately.
3. Thaw 4sU only once, and just before use. Concentrations should be optimized based on cell line and desired labeling time to balance incorporation efficiency and possible inhibition of rRNA synthesis [8].
4. Handle labeled cells at room temperature as quickly as possible. Note that 4sU has crosslinking ability at 365 nm wavelength. Avoid light sources that may mimic this wavelength.
5. To study RNA decay you can perform a pulse chase experiment in which the duration of the 4sU labeling is increased and chased with cell media absent of 4sU. Timepoints can then be taken during the chase period to determine decay rates.
6. TRIzol samples may be freeze-thawed at least twice, thus allowing for two pull-down reactions on different dates from a single 15 cm plate depending on cell type. Otherwise, freeze in two aliquots to reduce freeze-thaws.
7. While not “best practice,” centrifugation at room temperature will not cause failure.
8. After these two steps, no further drying of the pellet is required. Over drying of pellet may risk making it difficult to dissolve, even with heating.
9. Rotation has been done under general lab lighting with success.
10. Alternatively, this step can be done using phase lock gel heavy tubes to avoid both the loss of material and phenol carry-over.
11. 80 μ L of beads for 80 μ g RNA reaction is also sufficient.
12. When processing replicate samples, we find increased variability when the pulldown is done in different rounds. Therefore, it is recommended to perform the pulldown on replicates in the same round.
13. To initiate the flow through the column you can gently press on the top of the column with your gloved finger.
14. Here you have the option to finish the remainder of this section by eluting directly into 700 μ L RLT buffer and complete RNA isolation/cleanup using RNeasy MinElute cleanup kit. However, residual kit buffer in the RNA may skew NanoDrop OD readings.
15. For very short time points, this may be very low or unreliable detection.

Acknowledgments

Research in the Hertel laboratory is supported by NIH (GM062287, GM110244 and F31CA17179). Special thanks to Nate Hoverter for contribution of key graphics in Fig. 1.

References

- Pandya-Jones A, Black DL (2009) Co-transcriptional splicing of constitutive and alternative exons. *RNA* 15(10):1896–1908
- Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* 322(5909):1845–1848
- Brody Y, Neufeld N, Bieberstein N, Causse SZ, Böhnlein E-M, Neugebauer KM et al (2011) The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing. *PLoS Biol* 9(1):e1000573
- Tani H, Akimitsu N (2012) Genome-wide technology for determining RNA stability in mammalian cells. Historical perspective and recent advantages based on modified nucleotide labeling. *RNA Biol* 9(10):1233–1238
- Rädle B, Rutkowski AJ, Ruzsics Z, Friedel CC, Koszinowski UH, Dölken L (2013) Metabolic labeling of newly transcribed RNA for high resolution gene expression profiling of RNA synthesis, processing and decay in cell culture. *J Vis Exp* 78:e50195
- Rabani M, Levin JZ, Fan L, Adiconis X, Raychowdhury R, Garber M et al (2011) Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat Biotechnol* 29(5):436–442
- Barrass JD, Reid JEA, Huang Y, Hector RD, Sanguinetti G, Beggs JD et al (2015) Transcriptome-wide RNA processing kinetics revealed using extremely short 4tU labeling. *Genome Biol* 16:282
- Burger K, Mühl B, Kellner M, Rohmoser M, Gruber-Eber A, Windhager L et al (2013) 4-thiouridine inhibits rRNA synthesis and causes a nucleolar stress response. *RNA Biol* 10(10):1623–1630

APPENDIX B

Splice site proximity influences alternative exon definition

RESEARCH PAPER

 OPEN ACCESS  Check for updates

Splice site proximity Influences alternative exon definition

Francisco Carranza*, Hossein Shenasa*, and Klemens J. Hertel 

Department of Microbiology and Molecular Genetics, University of California Irvine, Irvine, California, USA

ABSTRACT

Alternative splicing enables higher eukaryotes to expand mRNA diversity from a finite number of genes through highly combinatorial splice site selection mechanisms that are influenced by the sequence of competing splice sites, cis-regulatory elements binding trans-acting factors, the length of exons and introns harbouring alternative splice sites and RNA secondary structures at putative splice junctions. To test the hypothesis that the intron definition or exon definition modes of splice site recognition direct the selection of alternative splice patterns, we created a database of alternative splice site usage (ALTssDB). When alternative splice sites are embedded within short introns (intron definition), the 5' and 3' splice sites closest to each other across the intron preferentially pair, consistent with previous observations. However, when alternative splice sites are embedded within large flanking introns (exon definition), the 5' and 3' splice sites closest to each other across the exon are preferentially selected. Thus, alternative splicing decisions are influenced by the intron and exon definition modes of splice site recognition. The results demonstrate that the spliceosome pairs splice sites that are closest in proximity within the unit of initial splice site selection.

ARTICLE HISTORY

Received 10 February 2022

Revised 07 May 2022

Accepted 09 June 2022

KEYWORDS

Alternative splicing; exon definition; intron-exon architecture; splice site selection; bioinformatics; molecular biology

Introduction



Pre-mRNA splicing is an essential step in eukaryotic gene expression that involves the excision of intronic sequences and the transesterification of exonic sequences by the spliceosome to generate protein coding mRNAs. Alternative exon inclusion is possible through a process known as alternative splicing. At least 95% of human genes undergo alternative splicing in response to cell cycle, developmental, tissue-specific or signalling cues. Alternative splicing increases proteomic diversity from a limited genome in a regulated fashion [1]. Thus, pre-mRNA splicing impacts gene expression [2].

The recognition of splice junctions by the spliceosome initiates the splicing reaction. The 5' splice site (5'ss) is defined by a nine-nucleotide consensus sequence that spans the exon/intron junction at the 5' end of each intron. The 3' splice site (3'ss) includes three sequence elements found within an approximately 40 nucleotides (nts) stretch, upstream of the 3' intron/exon junction. These include the intron/exon junction sequence, which contains the essential AG dinucleotide at the 3' end of the intronic sequence, the polypyrimidine tract (PPT), a region containing 15–20 pyrimidines located upstream of the intron/exon junction and the branch point sequence, a highly degenerate sequence that contains a conserved adenosine located upstream of the PPT.


Exon recognition is a highly combinatorial process that is known to be influenced by many cis- and trans-acting features. These include splicing enhancers, silencers, RNA secondary structure, the intron-exon architecture and the

sequence context of splice junctions [3–5]. The strength of splice sites is determined by how well they conform to consensus splice junction motifs that function in recruiting U1 snRNP to the 5'ss and U2AF to the 3'ss. Consensus similarity scores, derived from the modelling of short sequence motifs using the maximum-entropy principle (MaxEnt), define splice site strength numerically [6]. Splice sites are known to act synergistically and combined 5' and 3'ss scores are a much better predictor for exon inclusion than either splice site score alone [7]. Importantly, the ability of an exon to undergo various forms of alternative splicing is heavily influenced by the strength of its splice sites [8].

Another crucial factor in splice site selection is the genomic architecture [9–12]. The genomes in lower eukaryotes are characterized almost exclusively by the presence of short introns (<250 nts). By contrast, human genes harbour long introns, with >87% of introns longer than 250 nts [10]. This different genomic architecture has been shown to contribute significantly to the manner in which spliceosomal assembly occurs. The two proposed mechanisms through which splice sites are recognized are referred to as the exon or intron definition mode of splice site recognition (Figure 1(a)). During intron definition, the spliceosome assembles across the intron that will be excised. Under conditions that promote exon definition, initial splice site recognition is postulated to occur across the exon. This initial recognition is predicted to be followed by an additional splice site juxtapositioning step to induce intron excision. *In vitro* splicing

CONTACT Klemens J. Hertel  khertel@uci.edu  Department of Microbiology and Molecular Genetics, University of California Irvine, Irvine, California 92697, USA

*These authors contributed equally.

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/15476286.2022.2089478>

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

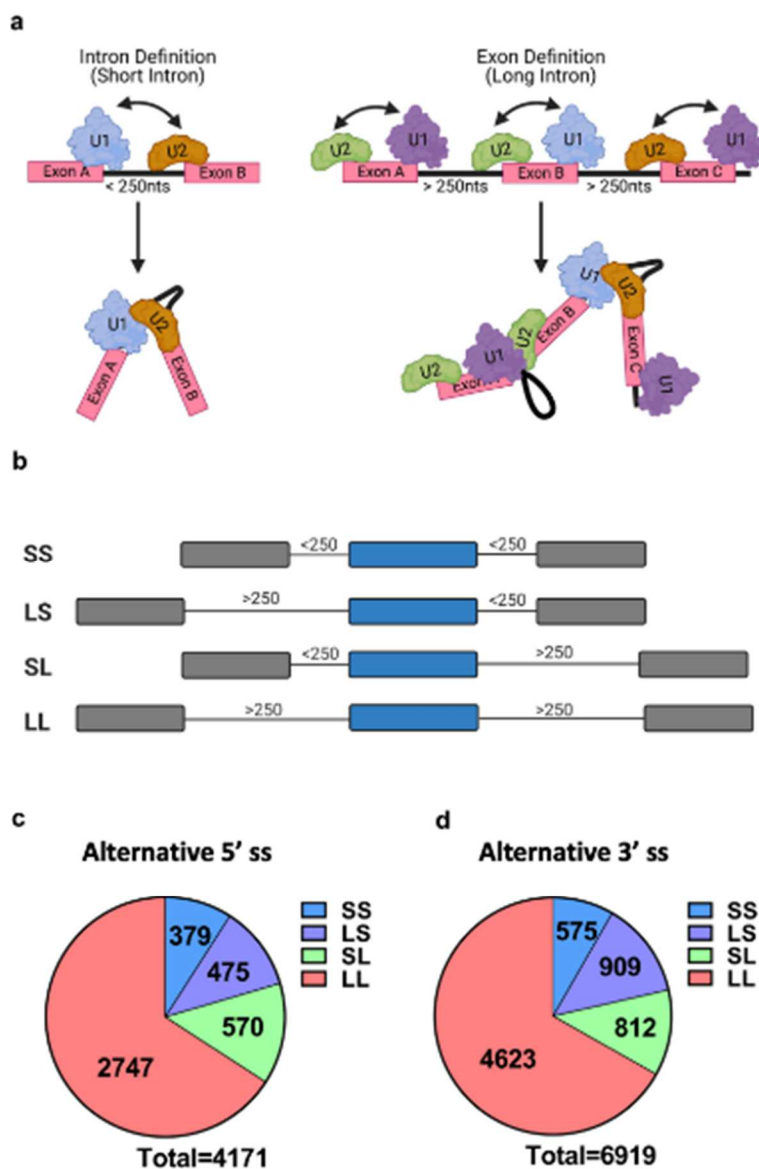


Figure 1. Gene architecture and database (a) The two proposed modes of splice site recognition. During intron definition splice sites are recognized across the intron (left). Under exon definition (right) splice sites are initially recognized across the exon, followed by splice site juxtaposition. (b) ALTssDB categories of internal exons as defined by flanking intron size. S stands for short (less than 250 nts), L stands for long (greater than 250 nts). (c) and (d) Distribution of ALTssDB internal exon categories for alternative 5' (c) and alternative 3' (d) splice site events.

and transfection experiments of designer minigenes demonstrated that the transition between intron and exon definition occurs at an intron length of approximately 250 nts [10]. Thus, splice sites that are flanked by large introns (>250 nts) are recognized through exon definition, while intron-defined splice sites are associated with small flanking

introns (<250 nts). It is currently unknown how exon and intron definition influence alternative splice site selection.

Understanding the relationship between the splice site strength and intron-exon architecture splicing determinants has been a longstanding goal in deciphering the splicing code. The mechanisms utilized by the spliceosome to select the

correct splice site in the presence of multiple nearby cryptic or alternative splice sites are still not completely understood. Differences in intron-exon architecture and splice site strength are known to be important in mediating alternative splice site selection [8]. A series of classical experiments demonstrated that the proximity between the 5' and 3' splice sites, across the intron, plays a crucial role in splice site preference [11]. Reed and Maniatis showed that the splice site closest to its intronic splicing partner was favoured over a distal competing splice site [11]. Thus, in the case of competing alternative 5' splice sites, the downstream 5'ss was preferred because it was more proximal to the pairing 3'ss. Similarly, between competing 3' splice sites, the upstream 3'ss was chosen. These observations suggest that in the absence of confounding factors, shorter distances between splice sites are favoured during intron-defined splicing. This may be because splice site pairing is more efficient across shorter distances. These experiments established a splice site selection proximity rule (for clarity referred to as the intron-centric proximity rule); however, it is unclear how dominant it is within the hierarchical nature of known splicing determinants.

In this study, we carried out computational analyses to assess the impact of the intron-centric proximity rule. We demonstrate that the intron-centric proximity rule is generally applicable for the intron definition mode of splice site definition. For the exon definition mode of splice site definition, we observe an exon-centric proximity rule that deviates from the classical intron-centric proximity rule. The 5' and 3' splice sites closest to each other across the exon are preferentially selected. Thus, when the unit of splice site definition is across the intron (intron definition), the 5' and 3' splice sites closest to each other across the intron preferentially pair. When the unit of splice site recognition is the exon (exon definition), the 5' and 3' splice sites closest to each other across the exon are preferentially selected. Our results provide evidence that alternative splicing decisions are influenced by the intron and exon definition modes of splice site recognition.

Results

The Influence of Intron-exon architecture on 5' splice site selection

To determine the impact of the intron-exon architecture and splice site strength on splice site selection, we created a database of alternative splice sites (ALTssDB) using the Human Exon Splicing Event Database HEXEvent [13], the Intron DB [14] and GeneBase [15]. MaxEntScan, a computational tool, was used to assign splice site scores [6]. To minimize variability, we focused on competing alternative 5' or 3' splice site pairs of internal exons with only one alternative splice pattern. Thus, ALTssDB catalogs pairs of alternative 5' splice sites competing for a common 3'ss or pairs of alternative 3' splice sites competing for a common 5'ss. ALTssDB reports the location of the major splice site and its competing alternative 3' or 5' splice site, corresponding exon sizes, usage levels, splice site scores and flanking intron lengths. Using these filters, ALTssDB captures 4,171 human 5'

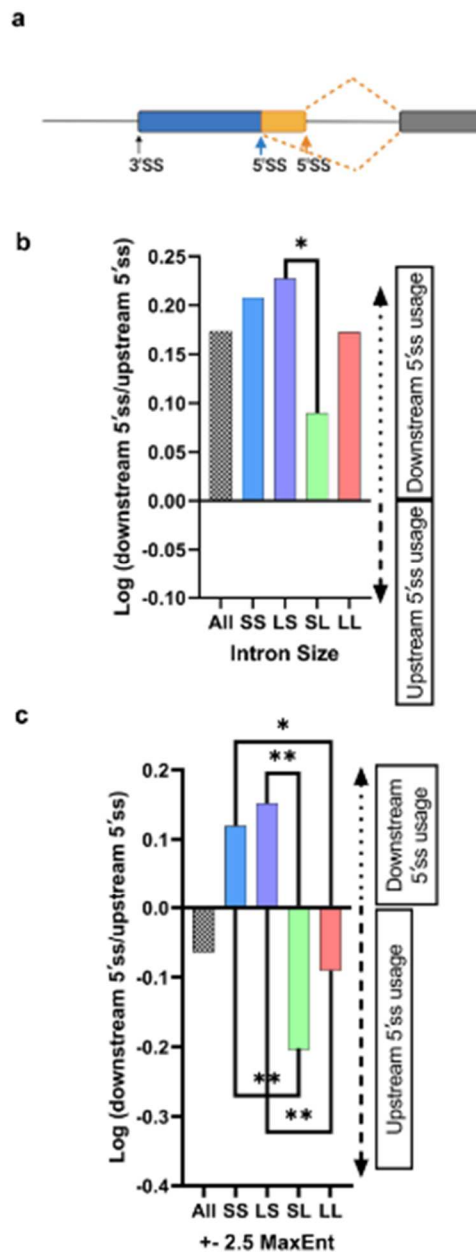


Figure 2. 5' ss selection preference for different internal exon categories. (a) Model depicting alternative 5'ss patterns. (b) Bar graph depicting the preference for downstream or upstream 5'ss selection for different internal exon categories (sample size = 379 SS, 2,747 LL, 570 LS, 475 SL). A positive log ratio represents downstream 5' ss preference, a negative log ratio represents upstream 5' ss preference. (c) Splice site selection preference for alternative 5' splicing events with near equal splice-site strength scores ($\Delta \pm 2.5$ MaxEnt, sample size = 88 SS, 725 LL, 104 LS, 143 SL). Fisher's exact test was performed. (b, c), ** $p < 0.01$, * $p < 0.05$.

ss competition events and 6,919 human 3'ss competition events (Figure 1(b-d)).

We first tested whether the intron-centric proximity rule holds true when evaluating all alternative 5'ss events transcriptome-wide (Figure 2(a)). In agreement with the intron-centric proximity rule expectation that the downstream 5'ss should be selected over a competing upstream 5'ss, we observed a preference for downstream 5'ss selection in ~60% (2,497) of the alternative 5'ss splicing events (Figure 2(b), left bar).

To evaluate whether the 'intron definition' or 'exon definition' mode of splice site selection influence adherence to the intron-centric proximity rule, we parsed the 5'ss dataset into intron definition events (379 SS), exon definition events (2,747 LL), and hybrid events (570 LS, 475 SL) (Figure 1(b and c)). For the purpose of alternative 5'ss selection analysis, the hybrid architectural class LS was categorized as intron defined because the 5'ss is adjacent to a short intron and U1 snRNP binding to the 5'ss at the exon/intron junction initiates early spliceosome formation [16]. By analogy, the architectural class SL was considered exon defined because the 5'ss is contained within a long intron. Surprisingly, in all four intron architecture classes, the majority of events still displayed a preference for the downstream 5'ss, consistent with the intron-centric proximity rule, albeit to varying degrees (Figure 2(b)). For example, the downstream 5'ss is selected more frequently for intron definition events (represented by SS, LS) when compared to exon definition events (represented by LL, SL). These varying degrees of preference suggest that the intron definition mode of splice site selection adheres more stringently to the intron-centric proximity rule.

The influence of intron-exon architecture on 5'ss selection in the absence of splice site strength differences

One important determinant that may mask the influence of splice site proximity is the difference in the splice site strength of competing splice sites. To determine the impact of splice site strength on alternative 5'ss selection, we compared the splice strength of the major 5'ss versus the alternative 5'ss. In 86% of the events evaluated the 5'ss with a higher predicted splice strength was the dominant 5'ss, irrespective of whether the exon was predicted to be recognized through exon definition (LL, SL) (85%) or intron definition (SS, LS) (90%) events. These results support the notion that splice site strength is a strong determinant in alternative 5'ss selection.

To determine how the exon and intron definition modes of splice site selection influence alternative splicing the impact of splice site strength differences was minimized computationally. This was achieved by isolating 5'ss competition events

with near equal splice site scores ($\Delta\text{MaxEnt} = \pm 2.5$), resulting in 88 SS, 725 LL, 104 LS, and 143 SL events. Interestingly, when this splice site strength filter was applied, we observed that the upstream 5'ss is preferentially selected in 60% of competition events, inconsistent with the expectations of the intron-centric proximity rule (Figure 2(c), left bar). Strict intron definition events (SS category) display a downstream 5'ss selection preference, consistent with the intron-centric proximity rule, while strict exon definition events (LL) display a preference for the upstream 5' splice site (Figure 2(c)). The upstream preference under exon definition is inconsistent with the intron-centric proximity rule but consistent with an exon-centric proximity rule. These biases are heightened in the hybrid categories SL (upstream preference) and LS (downstream preference) (Figure 2(c)). These results suggest that for exon definition events the upstream 5'ss, which is proximal across the exon to the upstream 3'ss, is favoured. By contrast, for intron definition events, the 5'ss proximal across the intron to the downstream 3'ss is favoured.

The influence of exon size on 5'ss selection

It is known that exon size can influence splice site selection [9,10]. To determine the influence of exon size on splice site selection, we compared splice patterns between three different exon size groups, exons smaller than 50 nts, exons between 50–250 nts in length, and exon longer than 250 nts. These cut-offs were chosen based on natural exon size distributions. We then calculated how frequently the major isoform contains the stronger 5'ss for the three different exon size classes (Table 1). When the major and the alternative exons are smaller than 50 nts, splice preference is driven almost exclusively by the stronger splice site score (Table 1). This preference weakens when the usage of the alternative 5'ss generates an exon greater than 50 nts. Thus, differences in exon size contribute to splice site selection, with a preference for generating shorter exons. A similar trend is observed for alternative patterns of major exons within the 50–250 nts range. The selection of alternative exons larger than 250 nts is much less likely to be driven by splice site differences. These data provide evidence that exon size contributes to splice site selection with a preference for defining smaller exons.

Experimental verification of genome-wide computational analysis

To test whether the proposed exon-centric proximity rule can be confirmed experimentally, we tested five minigenes that contain an internal exon with two competing 5' splice sites of identical strength (MaxEnt 10.9, CAG/guaagu) and one 3'

Table 1. Alternative 5'ss selection and resulting exon length correlation. The table reports how frequently the major isoform contains the stronger splice site when the alternative splice site lies within one of three different exon size classes. ^{a,b}Within a column, means without a common superscript differ ($p < 0.05$) between size categories in each column. ¹29% preference for the upstream 5'ss. ²42% preference for the upstream 5'ss. ³44% preference for the upstream 5'ss.

Exon size generated with minor splice site usage	Exon size generated with major splice site usage			
	Ex ≤50 nts	50 < Ex ≤250 nts	Ex >250 nts	
Ex ≤50 nts	98% ^a	86% ^a	100% ^a	
50 < Ex ≤250 nts	73% ^b	87% ^{a2}	88% ^a	
Ex >250 nts	75% ^b	58% ^b	77% ^{b1}	

splice site with a MaxEnt of 12.56 (uguccuuuuuuuccacag/CUG) (Figure 3(a)). All minigenes were designed to be recognized through exon definition (flanking intron size of 365 nts) and differ only in the resulting internal exon size. Cell transfection experiments demonstrated that for all constructs tested the upstream 5' splice site was chosen exclusively (Figure 3(b)), consistent with the computational analysis demonstrating that upstream 5' splice sites are favoured under exon definition. To test if this splice site preference is altered when both competing splice sites are weakened, we mutated both 5' splice sites to have a MaxEnt score of -0.5 (GAG/guguca). In the larger exon constructs (L and XL), this resulted in preferential internal exon skipping. In the M and S constructs, the upstream 5' splice site maintained its preference (Figure 3(c)). These results demonstrate that in an isogenic exon definition context the 5' splice site most proximal to the upstream 3' splice site is favoured, supporting the computational analysis of an exon definition 'cross-exon proximity' preference.

The Influence of Intron Architecture on 3' splice site selection

To investigate the impact of intron size and splice site strength on 3' splice site selection we built a 3' splice site dataset analogous to the 5' splice site dataset described above (Figure 4(a)). For our analysis, we took into consideration that the 3' splice site is recognized during the first and the second steps of splicing. Prior to the first step of splicing, the polypyrimidine tract is bound U2AF, which subsequently recruits U2 snRNP to the branch point. After the first step of splicing, the 3' splice junction YAG/N is selected before the exons are ligated via a transesterification reaction. It has been demonstrated that competing 3' splice sites in close proximity (up to 9 nts) are selected during the second step of splicing after identical first step definition [17]. Alternative 3' splice sites further apart (greater than 12–20 nts) are typically defined during initial splice site recognition using different polypyrimidine tract and branch points. Thus, we split the 3' splice site dataset into 'first step recognition' (≥ 20 nts apart from one another, 3839 events) and 'second step recognition' events (≤ 9 nts apart from one another, 2317 events). Both 3' splice site event groups show a preference for upstream 3' splice site usage, consistent with the intron-centric proximity rule (Figure 4(b)). This preference is particularly strong for the second step alternative 3' splice sites. Filtering to obtain competing 3' splice site pairs with comparable strengths and categorizing these events into intron (S/S, 69 events) or exon definition (L/L, 664 events) again demonstrated the influence of the intron architecture on 3' splice site selection (Figure 4(c)). The strong upstream 3' splice site preference observed for intron defined events (S/S) is significantly reduced when splice sites are selected in the exon definition mode (L/L). Consistent with our 5' splice site analysis, the hybrid classes (SL, 88 events and LS, 147 events) display more extreme splice site preferences relative to the SS and LL classes, with SL mimicking intron definition and LS mimicking exon definition behaviour.

Together, our transcriptome-wide analyses demonstrate that the mode of splice site selection critically influences splice site choice. For intron definition, splice sites closest across the intron are preferentially selected. Under exon definition, the selection of splice sites closest across the internal exon are

favoured. These results suggest that the gene architecture influences alternative splicing by promoting splice site recognition via the intron or exon definition pathway.

Discussion

The regulation of pre-mRNA splicing is a combinatorial process that is controlled by splice site sequences, cis-regulatory elements binding trans-acting factors, the intron-exon architecture, and RNA secondary structure among other features [16]. Two mechanisms of splice site recognition have been proposed within the broader concept of intron-exon architecture. It has been postulated that under the intron definition splice sites are recognized across the intron, making the intron the initial unit recognized by the spliceosome. In an alternative mode of splice site recognition, splice sites are postulated to be initially recognized across the exon in a process called exon definition. Once the exon is defined as the initial unit of splice site recognition, subsequent structural rearrangements are predicted to recognize and pair the upstream and downstream splice sites across flanking introns [18]. The mechanisms of intron and exon definition have been studied in the field for almost 30 years [9,10,18–23].

Early evidence that the length of introns and exons is important came from size constraints on exon inclusion from minigenes that were transfected in cell culture. Large exons were efficiently spliced when flanked by short introns, consistent with an intron definition mechanism. However, when intron lengths were increased exons were only included efficiently if they were relatively short, less than ~ 500 nts long. The latter observation suggests that the early spliceosome has a limited 'wing-span' when the exon is the unit of initial splice site recognition. Subsequently, biochemical studies demonstrated that intron definition is more efficient and that the rate of splicing for exon defined substrates is considerably slower. This study identified intron length as the primary determinant in the mode of splice site recognition employed by the early spliceosome and placed the transition from intron definition to exon-definition at the point when flanking introns become longer than 200–250 nts [10].

Another classical study used *in vitro* splicing assays to demonstrate that alternative splice site choice is influenced by the proximity between the pairing splice sites. When two splice sites are in competition, the splice site proximal to the intron is preferred. As a result, this proximity bias induces the preferential excision of the smaller intron [11]. This study and the pioneering study from Sterner and Bergt when analysed together suggest that in the context of splice site competition, selection of proximal splice sites across an intron may allow the intron to be recognized through intron definition, while the selection of the distal splice site may lead to a larger unit of initial splice site recognition that may change the mode of splice site recognition all together [9,11]. In broader terms, the findings by Reed and Maniatis [11] indicated that perhaps proximity across the initial unit of splice site recognition would drive splice site selection and influence alternative splicing. We set out to determine whether the proximity of splice sites across the proposed initial unit of splice site recognition may provide genome-wide evidence for the two

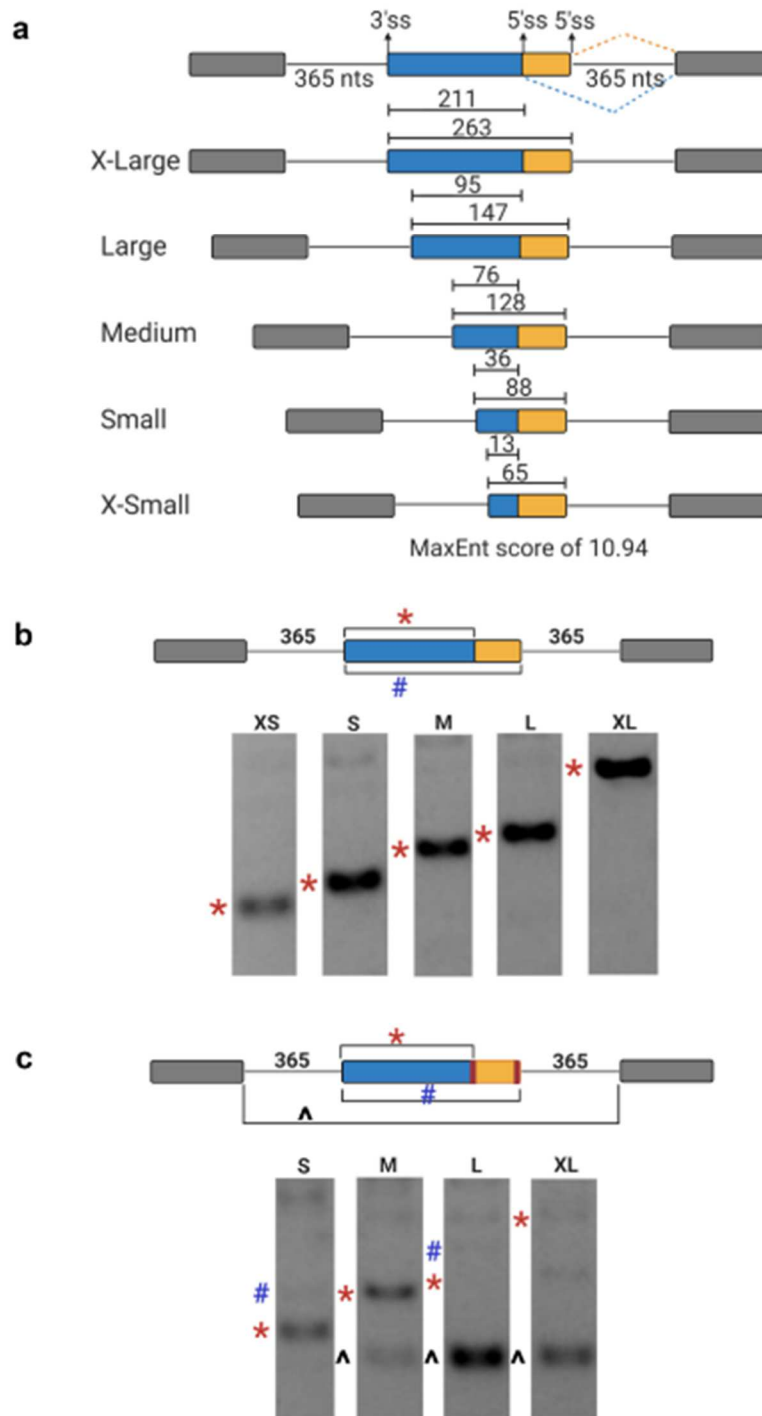


Figure 3. Cross-exon selection of alternative 5' alternative splice sites. (a) Schematic of exon-defined mini-gene constructs with identical splice site strength (CAG/guaagu, MaxEnt = 10.9) used in transfection experiments. The size of the resulting internal exon is indicated for upstream (blue) and downstream 5' ss selection. (b) Representative image of ethidium bromide stained agarose gel splicing analysis. Bands denoting upstream (red symbol) or downstream (blue symbol) 5' ss usage are marked to the left of the image. (c) Splicing outcome of minigene constructs with identical but weakened competing 5' ss (CAG/guguca, MaxEnt = -0.5). Bands denoting upstream 5' ss usage (red symbol), downstream (blue symbol) 5' ss usage, or exon skipping (black symbol) are marked to the left of the image.

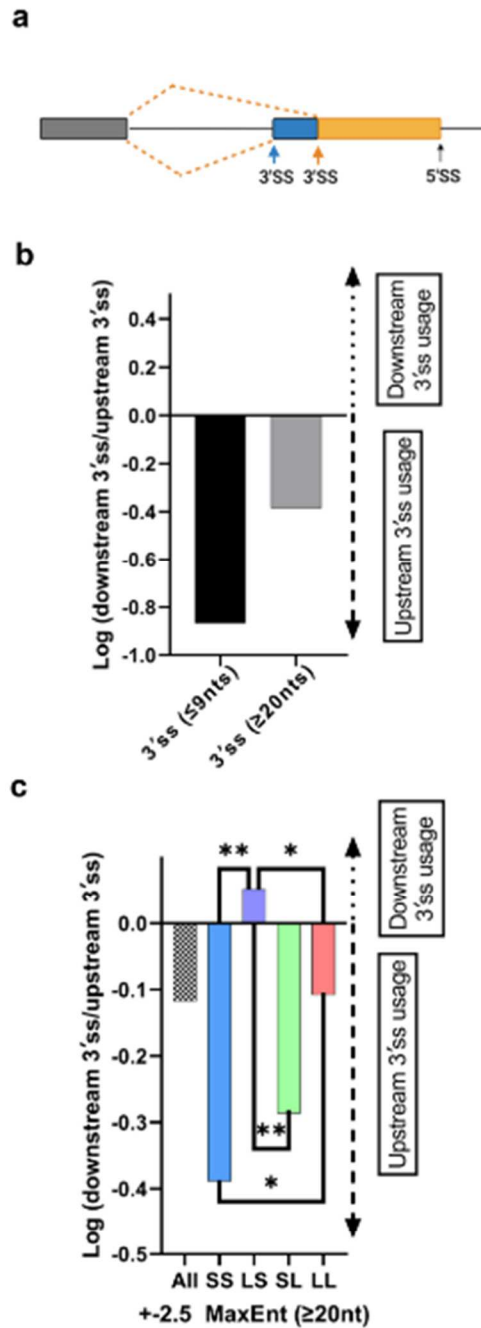


Figure 4. 3'ss selection preference for different internal exon categories. (a) Model depicting alternative 3'ss patterns. (b) Bar graph displaying the 3'ss preference for first (>20 nts distance between competing 3'ssplice sites, 2317 events) or second (<9 nts distance between competing splice sites, 3839 events) step selection. A positive log ratio represents downstream 3'ss preference, a negative log ratio represents upstream 3'ss preference. (c) Bar graph depicting the preference for downstream or upstream 3'ss selection with near equal splice-site strength scores for different internal exon categories ($\Delta \pm 2.5$ MaxEnt; sample size = 69SS, 664 LL, 88 SL, 147 LS). Fisher's exact test was performed. (b, c), ** $p < 0.01$, * $p < 0.05$.

modes of splice site recognition and elucidate their roles in alternative splicing.

Our analysis of the alternative splicing events captured in ALTssDB permitted the derivation of several important conclusions. First, the intron-centric proximity rule observed by Reed and Maniatis is maintained within the context of the intron definition mode of splice site recognition [11]. In the context of exon definition, we observe an exon-centric proximity rule, where the proximity between 5' and 3' splice sites across the exon dictates splice site preference. Alternative exons subject to the intron-centric proximity rule undergo removal of the smaller intron and selection of the larger exon. Conversely, alternative exons subject to the exon-centric proximity rule undergo removal of the larger intron and selection of the smaller exon. Initially, these observations may appear inconsistent with each other, yet they highlight a commonality of spliceosomal assembly across the smallest unit of initial splice site recognition. For the intron definition mode of splice site recognition this unit is the intron, meaning the spliceosome assembles around the 5' and 3' splice sites that define the intron to be excised (Figure 5, top cartoon). For the exon definition mode of splice site recognition, the unit of recognition is the exon, meaning that initial splice site recognition by the spliceosome occurs across the exon (Figure 5, bottom cartoon). In both modes of splice site recognition, a preference for splice site selection that promotes the definition of the smaller initial recognition unit (as defined by the number of nucleotides) is observed. Thus, the proximity of 5'

and 3' splice sites within the unit of initial recognition determines preferential splice site selection (Figure 5). We therefore conclude that an additional mechanism of alternative splicing can be the proximity of splice sites across the initial unit of definition.

Since the initial concepts of intron and exon definition were introduced, generating supporting evidence for the existence of these two proposed modes of splice site recognition has been challenging. Initial studies were limited to select cases where insights were gained from transfected designer minigenes or *in vitro* transcribed RNAs spliced using the nuclear extract system [9,10,19,21–23]. These studies, while mechanistically enlightening, could not be extrapolated to the entire transcriptome.

Recent analyses of *in vivo* splicing kinetics offer more comprehensive insights into the mechanisms of exon recognition. These studies lend support to the notion that exon and intron definition events display different global splicing kinetics. They also raise questions about the generality of exon definition and intron definition [24–26]. A single molecule intron tracking technique was used to determine the amount of splicing as a function of RNA polymerase II position along the gene. This technique and an orthogonal nanopore-based variation found splicing rates to be strikingly fast in *Saccharomyces cerevisiae* [25] demonstrating that 50% of splicing can be completed 1.4 seconds after 3'ss synthesis for the genes studied. The onset of splicing for a subset of the analysed genes was detected only 26 nucleotides after

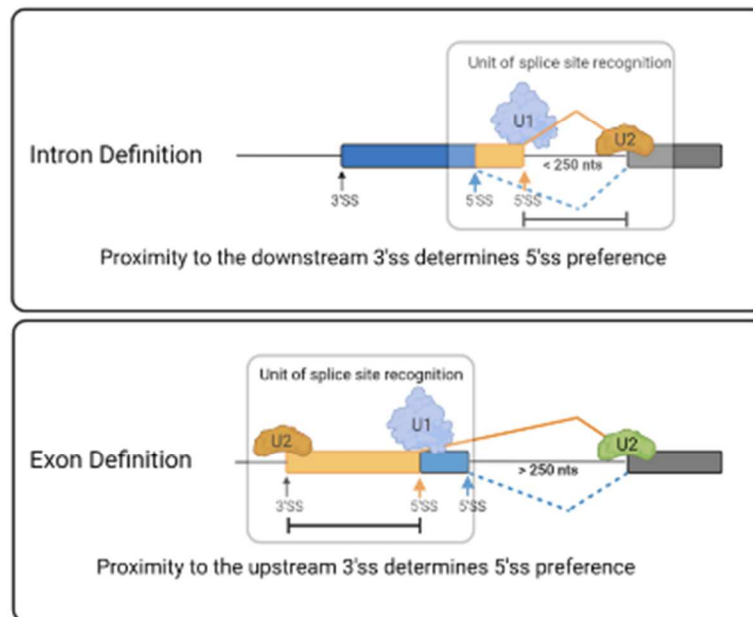


Figure 5. Unifying model for the influence of splice site proximity in alternative exon definition. Depending on the size of flanking introns the splice sites of internal exons are initially recognized across the intron (top – intron definition) or across the exon (bottom – exon definition). In both scenarios, the 5' and 3' splice sites closest to each other across the unit of initial splice site recognition are preferentially selected. Thus, in intron definition 5' and 3' splice sites across the intron are preferentially selected. In exon definition, 5' and 3' splice sites across the exon are preferentially selected.

transcription of the 3' ss. The observation that splicing can be completed before the entire exon is transcribed is consistent with an intron definition mechanism in *Saccharomyces cerevisiae*, but begs the question is exon definition possible in lower eukaryote? The average *Saccharomyces cerevisiae* exon is ~1400 bases suggesting that exon definition would be highly unlikely for those genes where splicing rates were calculated to occur on the order of several seconds [27]. However, a recently proposed unifying model provides evidence for exon definition in *Saccharomyces cerevisiae* [28]. Electron microscopy analyses suggest that the splicing factor Prp40 can bridge the 5' ss bound U1 snRNP and branch point sequence bound BBP/Mud2 (SF1/U2AF65 homologs) either across the intron or across the exon to define E-complex. Structural evidence for exon definition in *Saccharomyces cerevisiae* was supported by genetic and biochemical analysis, which included the circularization of single exon constructs in yeast splicing extracts. The latter study provides strong structural, biochemical, and *in vivo* evidence for exon definition, even in *Saccharomyces cerevisiae*, where most splice sites would be expected to be recognized through intron definition [28].

Regarding the intron-exon architecture of higher eukaryotes, ligation of 3' adapters and long read nanopore sequencing of nascent RNA were used to determine the splicing rates in *Drosophila* and human cells [26]. The nano-COP method determined that the majority of splicing in *Drosophila* occurs within 2 kilobases, once the 3' ss has been transcribed. This is in contrast to human cells where the majority of splicing is completed ~4 kilobases past the 3' ss [26]. The rate of splicing calculated from nano-COP is consistent with previous $t_{1/2}$ measurements that are ~2 minutes for *Drosophila* and 7–14 minutes for mammalian cells [24,29–31]. Interestingly, nano-COP found that *Drosophila* introns less than 100 nts in length were spliced more quickly than introns greater than 300 nts, suggesting that intron definition is more efficient than exon definition [26]. These results are supported by an earlier study that used progressive metabolic labelling and also found a local maximum of splicing rates for introns that were 60–70 nts long [24]. However, the latter study also found that a subset of very long introns (>2944 nts) was spliced even more quickly with a $t_{1/2}$ of ~1.5 minutes suggesting gene level and pathway-specific splicing programmes may have evolved to utilize the rapid splicing that very long exon-defined introns undergo. Taken together these kinetic measurements suggest that while exon definition is broadly less efficient and intron definition is broadly more efficient as was first shown by Fox-Walsh and Hertel [10], exceptions do exist.

Recent investigations provide further support that both intron definition and exon definition occur *in vivo* [32,33]. However, these studies present evidence that the mechanism by which splice sites are initially recognized is dictated by the difference in GC content, referred to as GC differential, between the exon and the flanking introns. Specifically, two architectures are described, referred to as the 'differential architecture' and the 'leveled architecture' [32,33]. 'Differential architecture' exons have a low GC content, and their flanking introns have an even lower GC content. The 'leveled architecture' exons are characterized by a high GC

content, less difference in the GC content of flanking introns and short introns. The former class of exons was demonstrated to be localized to the nuclear periphery and recognized through exon definition while the latter was shown to be localized to the nuclear centre and recognized through intron definition. Altering the GC content between exon and the downstream intron can be used to alter the mode splice site recognition, without changing the length of the intron. These observations suggest that intron length may not be the determining factor in the mode of splice site recognition for exon definition [32,33]. It will be important for future exon definition studies to consider the GC content across the exon and flanking introns. For example, a recent analysis of high-throughput mutagenesis data for an alternatively spliced exon in the proto-oncogene *RON* demonstrated that the alternatively spliced exon is recognized through exon definition, even though it is flanked by short introns on either side (87 and 80 nts) [34]. Thus, splice sites of short introns can be recognized through exon definition, perhaps because the unique GC content that typifies exon definition splice sites.

Finally, a recent transcriptome-wide study demonstrated that introns that undergo efficient co-transcriptional splicing have sharp structural transitions across the intron-exon boundary [35]. These introns display a peak of RNA structure downstream of the 5' ss and upstream of the 3' ss. Furthermore, some introns displayed enhanced co-transcriptional splicing under conditions where the elongation rate of RNA polymerase II was slowed down genome-wide, a process that promotes increased RNA folding. The latter group of introns had significantly steeper structural transitions when transcription was slow [35]. GC content is an indicator of the potential to form RNA secondary structures [36]. Thus, it may be the case that the differential architecture associated with exon definition is driven partially by the propensity for RNA secondary structure formation that can help delineate the intron-exon boundary.

We set out to determine the degree of agreement between the intron length-dependent definitions of 'intron-defined' and 'exon-defined' splice sites with the 'leveled' and 'differential' architecture. We calculated GC content differentials between the LL and SS architectural classes of alternatively spliced 5' and 3' ss exons. Remarkably, the intron length-defined LL and SS categories closely resemble the 'differential' and 'leveled' architectures respectively [32,33] (Supplemental Figure 2). Thus, the GC content, as defined the Amit et al. [32], of long introns (>250 nts) differs significantly from the GC content of short introns (<250 nts), suggesting that GC content or intron size definitions are comparable approaches to define exon and intron definition modes of splice site recognition. This notion is supported by evolutionary analyses that show the emergence of a distinct differential GC architecture as intron lengths increased through vertebrate evolution [37]. Thus, the emergence of the 'differential architecture' may be a co-evolutionary adaptation to define exons in the context of expanding introns. To evaluate whether the use of proximal or distal splice sites changes 'leveled' and 'differential' architecture designations, we calculated GC content for alternatively spliced exons captured by ALTsDB. Interestingly, the resulting GC differential does not change

significantly (Supplemental Figure 2), suggesting that alternative splice site selection is not dependent on differential GC content but contingent on defining the smallest unit of initial recognition.

Collectively, the results of our transcriptome-wide analysis of alternative splice site usage provide evidence that exon definition and intron definition do occur transcriptome-wide. When exons are flanked by long introns, the spliceosome tends to favour splice sites located internally within the exon being defined. By contrast, the spliceosome tends to move into the intron for splice site definition for exons flanked by short introns. These observations suggest that the spliceosome can define the exon and the intron independently.

Our computational analysis of alternative 3'ss events permitted an evaluation of alternative 3'ss selection in the context of first or second step recognition. Initial 3'ss selection is mainly driven by the strength of the polypyrimidine tract and the presence of a consensus branch point. Upon recruitment of U2 snRNP to the branch point and tri-snRNP incorporation, the first step of the splicing reaction is initiated without engaging the 3'ss junction. After spliceosome rearrangements, the 3'ss junction is selected during the second step of splicing as the spliceosome aligns the AG/N intron/exon junction into the active site. It is well established that competing 3'AGs in close proximity (less than 9 nts) use the same upstream polypyrimidine tract and branch point and that their selection is directed during the second step of splicing [17]. Interestingly, our analysis of alternative 3'ss selection in close proximity demonstrated that the upstream AG/N junction is almost exclusively chosen over the downstream AG/N. Thus, it appears that aligning the closest AG/N 3'ss junction is the default pathway of second step splice junction selection (Figure 4(b)).

The intron-exon architecture of genes is a major driver of splice site selection. Since the initial postulation of these two modes of splice site recognition, various forms of evidence have been presented, often in form of kinetic principles supporting intron or exon definition. However, measurements of splicing rates as a function of intron length do not constitute direct evidence of alternative spliceosomal assembly pathways. The ability of yeast E-complex to assemble across the intron or the exon is perhaps the strongest evidence yet for exon definition. Our study provides support for exon definition by demonstrating the spliceosome favours internal splice sites within exons when the splice site strengths of competing sites are comparable. This suggests that the exon is being defined and not the intron. This study provides a unifying model for splice site selection, whereby the spliceosome assembles across the smallest unit of initial splice site recognition. In the case of intron definition, this entails removal of smaller introns and inclusion of larger exons. Indeed, studying the evolutionary trends in intron-exon architecture, lower eukaryotes tend to have larger exons and smaller introns. Upon intron expansion and a gradual shift towards exon defined gene architecture, the initial unit of splice site recognition often tends to be the exons. This may be due to the increased number of decoy splice signals associated with larger genome sizes. It would therefore be expected that the

smaller exons would be favoured in higher eukaryotes. This trend is also broadly observed from yeast to humans. It is possible that the exon-centric proximity rule is an evolutionary adaptation to accurately recognize exons surrounded by long stretches of intronic sequence. Our results not only provide *in vivo* and transcriptome-wide evidence for exon definition, they also demonstrate that exon and intron definition influence alternative splicing in the context of alternative 5' or 3' splice site competition.

Methods

Construction of ALTssDB

ALTssDB was created using EST data from the Human Exon Splicing Events (HEXEvent) database [13]. HEXEvent contains information regarding the location of competing splice sites, the resulting exon sizes, alternative splice site usage levels and the gene associated with each mRNA. The HEXEvent data was filtered to obtain a dataset comprising of only pairs of competing 5' and 3' splice sites separately. This database was subsequently modified to include splice site junction information and splice site strength scores using MaxEntScan [6]. Although other approaches exist to evaluate the strength of 5 splice sites [38,39], MaxEntScan is the preferred tool as it also permits comparable splice site score derivation for 3 splice sites. Using an R script and IntronDB dataset, (a database detailing eukaryotic intron features) flanking intron lengths were added to the database [14]. Alternative splicing events were further filtered to include only events that have 10 or more EST counts. The data was filtered into four categories according to intron length and included: both flanking introns around the exon of interest being short (<250 nts, SS), both flanking introns being long (>250 nts, LL), the upstream intron being short and downstream intron being long (SL) or the upstream intron being long and the downstream intron being short (LS). ALTssDB does not differentiate between canonical U2 introns and U12-type introns. Given their rarity and limited involvement in alternative splicing beyond intron retention, it is anticipated that U12-type introns are not well represented in ALTssDB [40].

ALTssDB does not distinguish between isoforms that originate from a tissue specific splice switch or disease comparison. It lists all known splice patterns for a particular exon, independent of origin. EST data was used to build ALTssDB to obtain high enough numbers of alternative splice site choices within the human genome to carry out all analyses. While datasets for tissue-specific splicing are available, the quantity of significant alternative splice site events is limiting.

Plasmid design

Five minigene constructs were designed containing three exons and two introns. The plasmid design was based primarily on previously validated constructs used to study splice site strength [7]. The internal exon was designed to contain two functional competing 5' splice sites (CAG/guaagu), with equal MaxEnt scores MES of 10.9, separated by 52 nucleotides. The

sequence preceding the upstream 5' splice site was progressively shortened (Figure 1(b)). Additional constructs were created where the MES of both competing 5' splice sites were changed from 10.9 to -0.5 (GAG/guuga) for S, M, L, and XL plasmid. Lastly, the upstream 5' splice sites were changed from MaxEnt = 10.9 to MaxEnt = -5.2 (UCG/gucgau) for the S, M, and XL to show that the downstream 5'ss is viable (Supplementary Figure. 1).

Cloning protocols to change splice site strength sequences

To linearize the plasmids, 10 nanograms (ng) of plasmid DNA obtained by midiprep was amplified using divergent primers. PCR reactions were carried out with NEB[®] Phusion[®] polymerase in 50 μ L according to NEB protocols. DH5a E.coli midiprep derived plasmids in the PCR reaction were digested with 40 units of DpnI according to NEB protocols. Plasmids were purified with Zymo DNA clean and concentrator[™] kit and DNA concentrations were obtained using a nanodrop 2000 instrument. For each construct, 0.03 picomoles of linearized plasmid DNA was mixed with a 10X molar ratio of phosphorylated double stranded DNA inserts, purchased from IDT, in 20 μ L ligation reactions. Synthetic inserts were cloned into linearized vectors using T4 ligase according to the standard NEB protocol and 10 μ L of the ligation reaction was transformed using in house DH5a E.coli cells. Colonies were screened using PCR to detect the correct size insert. Colonies with the correct size insert were grown from 3 mL cultures to 20 mL cultures and underwent midiprep DNA extraction. Plasmid DNA from each colony was sequenced to ensure the correct orientation of inserts.

Cell transfections and RT-PCR Analysis

Transfection experiments were performed in triplicate using HeLa cells. 1 mL of 0.1×10^6 cells/mL was plated into each well of 12 well plates. Cell confluency was checked 24 hours later and 1 μ g of plasmid DNA was transfected according to Bioland Scientific's BioT protocol. Cells were harvested 48 hours post-transfection. Each well was washed two times with phosphate buffered saline (PBS) and subsequently RNA was extracted with the standard Trizol[™] protocol. RNA pellets were resuspended in 50 μ L water and put through ZYMO RNA Clean and Concentrator[™] columns. Sample volumes were adjusted to 80 μ L, yielding RNA concentrations of ≤ 200 ng/ μ L. DNase digestion was performed with Turbo[™] DNase (Ambion[®]) according to Ambion's protocol in 100 μ L reactions. RNA was subsequently extracted with 100 μ L phenol: chloroform and the aqueous phase was put through ZYMO RNA Clean and Concentrator[™] columns. DNase digested and purified RNA samples were resuspended in 25 μ L. A nanodrop 2000 instrument was used to obtain RNA concentrations. Reverse transcription reactions were carried out in 20 μ L using 250 ng of total RNA and 200 ng of OligodT18 primer according to SuperScript[™] III protocol. PCR primers are as follows: first exon forward primer (5'cgtctgctcactctctc3') and third exon reverse primer (5'agatcccaaggactcaaga3').

PCR primers were designed that bound the flanking exons and thus would detect upstream, proximal or downstream, distal 5' splice site usage. PCR reactions contained 5 μ L cDNA (10% vol:vol), 0.2 mM dNTPs, 0.2 μ M of each primer, 1.5 mM MgCl₂ and 0.25 units taq polymerase (Apex Bioresearch). Semi-quantitative PCR using long extension times to limit PCR product size bias was carried out to demonstrate that the ratio of upstream and downstream splice site usage, or the alternative exon skipping pattern, remained constant throughout the dynamic linear range of the amplification reaction (data not shown). Based on these results 25 cycles of PCR were performed for each sample and 5 μ L was subsequently loaded onto a 2% agarose gel and stained with ethidium bromide. Agarose gels were run at 150 V for 1 hour in 1X Tris-Borate EDTA (TBE).

Calculating splice site selection preference

5'ss selection preference was determined by calculating the log ratio of the number of splice site events preferring the downstream 5'ss over the upstream 5'ss. 3' splice site selection preference was determined by calculating the log ratio of the number of splice site events preferring the downstream 3'ss over the upstream 3'ss.

Acknowledgments

This work was supported by grants from the NSF (DGE-1321846 to F.C.) and the NIH (R01 GM062287 to K.J.I).

Data Availability Statement

The data that support the findings of this study are openly available in Dryad at (https://datadryad.org/stash/share/kVSpUjBvhjulYsl.WLgG38JigYJy_UdcpODKw6DXs8).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Science Foundation [DGE-1321846]; National Institutes of Health [GM062287].

ORCID

Klemens J. Hertel  <http://orcid.org/0000-0002-7560-9529>

References

- [1] Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. *Nature*. 2010 Jan;463(7280):457–463.
- [2] Li J, Wang Y, Rao X, et al. Roles of alternative splicing in modulating transcriptional regulation. *BMC Syst Biol*. 2017 Oct;11(5):89.
- [3] Fu XD, Ares M. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet*. 2014 Oct;15(10):689–701.
- [4] Hertel KJ. Combinatorial control of exon recognition. *J Biol Chem*. 2008 Jan;283(3):1211–1215.

- [5] Eirkelenz S, Mueller WF, Evans MS, et al. Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *RNA*. 2013 Jan;19(1):96–102.
- [6] Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* 2004;11(2–3):377–394.
- [7] Shepard PJ, Choi E-A, Busch A, et al. Efficient internal exon recognition depends on near equal contributions from the 3' and 5' splice sites. *Nucleic Acids Res.* 2011 Nov;39(20):8928–8937.
- [8] Busch A, Hertel KJ. Splicing predictions reliably classify different types of alternative splicing. *RNA*. 2015 May;21(5):813–823.
- [9] Sterner DA, Carlo T, Berget SM, et al. Architectural limits on split genes. 1996;[Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC26359/pdf/pq015081.pdf>
- [10] Fox-Walsh KL, Dou Y, Lam BJ, et al. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. 2005;[Online]. Available: www.pnas.org/cgi/doi/10.1073/pnas.0508489102
- [11] Reed R, Maniatis T. A role for exon sequences and splice-site proximity in splice-site selection. *Cell*. 1986 Aug;46(5):681–690.
- [12] Movassat M, Forouzmand E, Reese F, et al. Exon size and sequence conservation improves identification of splice-altering nucleotides. *RNA*. 2019 Dec;25(12):1793–1805.
- [13] Busch A, Hertel KJ. HEXEvent: a database of human exon splicing events. *Nucleic Acids Res.* 2012 Oct;41(D1):D118–D124.
- [14] Wang D. IntronDB: a database for eukaryotic intron features. *Bioinformatics*. 2019 Nov;35(21):4400–4401.
- [15] Piovesan A, Caracausi M, Antonaros F. GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. *Database (Oxford)*. Vol. 2016, 2016. DOI:10.1093/database/baw153.
- [16] Shenasa H, Hertel KJ. Combinatorial regulation of alternative splicing. *Biochim Biophys Acta Gene Regul Mech.* 2019 Nov;1862(11–12):194392–194392. DOI:10.1016/j.bbagr.2019.06.003.
- [17] Dou Y, Fox-Walsh KL, Baldi PF, et al. Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. *RNA*. 2006 Dec;12(12):2047–2056.
- [18] Berget SM. Exon recognition in vertebrate splicing. *J Biol Chem.* 1995;270(6):2411–2414.
- [19] Robberson BL, Cote GJ, Berget SM, et al. Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol Cell Biol.* 1990 Jan;10(1):84–94.
- [20] Schneider M, Will CJ, Anokhina M, et al. Exon definition complexes contain the Tri-snRNP and can be directly converted into B-like pre-catalytic splicing complexes. *Mol Cell.* 2010 Apr;38(2):223–235.
- [21] Talerico M, Berget SM. Intron definition in splicing of small *Drosophila* introns. *Mol Cell Biol.* 1994 May;14(5):3434–3445.
- [22] De Conti L, Baralle M, Buratti E. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA.* 2013 Jan;4(1):49–60.
- [23] Lang KM, Spritz RA. RNA splice site selection: evidence for a 5' → 3' scanning model. *Science*. 1983 Jun;220(4604):1351–1355.
- [24] Pai AA, Henriques T, McCue K, et al. The kinetics of pre-mRNA splicing in the *Drosophila* genome and the influence of gene architecture. *eLife*. 2017 Dec;6: DOI:10.7554/eLife.32537
- [25] Carrillo Oesterreich F, Hertel L, Straube K, et al. Splicing of nascent RNA coincides with intron exit from RNA polymerase II. *Cell*. 2016 Apr;165(2):372–381.
- [26] Drexler HL, Choquet K, Churchman LS. Splicing kinetics and coordination revealed by direct nascent RNA sequencing through nanopores. *Mol Cell.* 2020 Mar;77(5):985–998.e8.
- [27] Koralewski TE, Krutovsky KV. Evolution of exon-intron structure and alternative splicing. *PLoS ONE*. 2011 Mar;6(3):e18055.
- [28] Li X, Liu S, Zhang L, et al. A unified mechanism for intron and exon definition and back-splicing. *Nature*. 2019 Sep;573(7774, Art. no. 7774):375–380.
- [29] Singh J, Padgett RA. Rates of in situ transcription and splicing in large human genes. *Nat Struct Mol Biol.* 2009 Nov;16(11, Art. no. 11):1128–1133.
- [30] Rabani M, Raychowdhury R, Jovanovic M, et al. High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell*. 2014 Dec;159(7):1698–1710.
- [31] Wachutka L, CaiZZi L, Gagneur J, et al. Global donor and acceptor splicing site kinetics in human cells. *eLife*. 2019 Apr;8:e45056.
- [32] Amit M, Donyo M, Hollander D, et al. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep.* 2012 May;1(5):543–556.
- [33] Tammer L, Hameiri O, Keydar I, et al. Gene architecture directs splicing outcome in separate nuclear spatial regions. *Mol Cell.* 2022 Mar;82(5):1021–1034.e8.
- [34] Enculescu M, Braun S, Thonta Setty S, et al. Exon definition facilitates reliable control of alternative splicing in the RON proto-oncogene. *Biophys J.* 2020 Apr;118(8):2027–2041.
- [35] Saldi T, Riemondy K, Erickson B, et al. Alternative RNA structures formed during transcription depend on elongation rate and modify RNA processing. *Mol Cell.* 2021 Apr;81(8):1789–1801.e5.
- [36] Shepard PJ, Hertel KJ. Conserved RNA secondary structures promote alternative splicing. *RNA*. 2008 Aug;14(8):1463–1469.
- [37] Gelfman S, Burstein D, Penn O, et al. Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. *Genome Res.* 2012 Jan;22(1):35–50.
- [38] Wong MS, Kinney JB, Krainer AR. Quantitative activity profile and context dependence of all human 5' splice sites. *Mol Cell.* 2018 Sep;71(6):1012–1026.e3.
- [39] Freund M, et al. A novel approach to describe a U1 snRNA binding site. *Nucleic Acids Res.* 2003 Dec;31(23):6963–6975.
- [40] Olthof AM, Hyatt KC, Kanadia RN. Minor intron splicing revisited: identification of new minor intron-containing genes and tissue-dependent retention and alternative splicing of minor introns. *BMC Genomics.* 2019 Aug;20(1):686.