Title
The Kinetic Scope of Alternatively Spliced pre-mRNA

Author
Garibaldi, Angela

Publication Date
2018

UNIVERSITY OF CALIFORNIA,
IRVINE

# The Kinetic Scope of Alternatively Spliced pre-mRNA

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biomedical Sciences

by

Angela A. Garibaldi

Dissertation Committee:
Prof. Klemens J. Hertel, Ph.D., Chair
Professor Bert L. Semler, Ph.D.
Professor Marian L. Waterman, Ph.D.

2018

# DEDICATION

To

My biggest cheerleaders since Day 1
KJO & LLO

# TABLE OF CONTENTS

Page

LIST OF FIGURES v

LIST OF TABLES vii

ACKNOWLEDGMENTS viii

CURRICULUM VITAE x

ABSTRACT OF THE DISSERTATION xiii

CHAPTER 1: Introduction 1
- Pre-mRNA Splicing 1
- Regulation of Splicing 4
- Intron/Exon Definition 5
- Splicing Regulatory Elements 7
- SRSF1: A Classical SR Protein 8
- Co-transcriptional pre-mRNA Splicing 10
- Summary 13

CHAPTER 2: Large Scale Analysis of Splicing Kinetics Reveals That Alternative Splicing is Promoted by Slow Intron Removal Rates 14
- Summary 14
- Introduction 14
- Results 16
- Discussion 38
- Methods and Materials 40

CHAPTER 3: Reduction of Splicing Factor SRSF1 Increases Alternative Splicing Through the Modulation of RNA Processing Kinetics
- Summary 42
- Introduction 43
- Results 48
- Discussion 64
- Methods and Materials 65

CHAPTER 4: Perspectives 69

REFERENCES 77

APPENDIX A: Isolation of Newly Transcribed RNA Using the Metabolic Label 4-Thiouridine 85

APPENDIX B: The TCF C-Clamp DNA Binding Domain Expands the Wnt Transcriptome via Alternative Target Recognition 94

APPENDIX C: The Silent Sway of Splicing by Synonymous Substitutions 113

# LIST OF FIGURES

Page

Figure 1.1 Sequence elements for splice site recognition 3

Figure 1.2 Splice site recognition by distance constraint models 6

Figure 1.3 Splicing regulatory components 12

Figure 2.1 Workflow of 4sU labeling and determination of rates 17

Figure 2.2 Progressive 4sU labeling permits the tracking of synthesis and splicing 20

Figure 2.3 Global scope of intron dynamics 22

Figure 2.4 Intron retention relationship of splicing rates and synthesis 24

Figure 2.5 Long introns are removed quickly 27

Figure 2.6 First and last introns have distinct characteristics 28

Figure 2.7 Alternative exon inclusion is captured by consecutive intermediates kinetics 30

Figure 2.8 Exons prone to being skipped fit the model 32

Figure 2.9 Order of intron removal in LAPTM4A 36

Figure 3.1 Western blot analysis demonstrates SRSF1 knockdown 45

Figure 3.2 Isogenic sample replicates 46

Figure 3.3 Intron retention is increased by SRSF1 knockdown 49

Figure 3.4 SRSF1 promotes efficient removal of long internal introns 51

Figure 3.5 Intron order: internal intron kinetics influenced by SRSF1 53

Figure 3.6 Introns that are differentially spliced upon SRSF1 knockdown 55

Figure 3.7 Comparisons of exon inclusion levels based on exon length and processing dynamics 57

Figure 3.8 Overview of eCLIP analysis of differentially included exons 59

Figure 3.9 Processing dynamics of exons that switch kinetic modes 62

# LIST OF TABLES

Page

Table 2.1 Determination of the order of intron removal for LAPTM4A 37

Table 3.1 eCLIP analysis of SRSF1 binding on differentially included exons 61

# ACKNOWLEDGMENTS

I would like to thank people in chronological order. First, I want to thank my family for being there for me from day one. Thank you for your love, faith, and support through the good times and the bad. Special thanks to my mom and my grandpa for believing in me and somehow knowing just what to say to get me back on track every step of the way.

Next, I want to acknowledge Joy Brittain for the amazing opportunities she works hard to provide young people. Being in the Math Science Upward Bound program was a life altering and inspiring experience. It showed me that science was a lot more fun than just reading books and memorizing facts. I would also like to thank my first PI, Dr. Michael Danciger for demonstrating that science is fluid and meaningful. For six years he was my example of how hard work and passion could drive R1 science at a PUI. Additionally, I am grateful to Dr. Kam Dahlquist for teaching me the broader social implications of science, and for roping me into applying for PhD programs at the last possible moment.

This brings me to UC Irvine. I'd like to thank Dr. Klemens (Dicki) Hertel for staying optimistic when I just "knew" the data was bad, showing me how fun conferences are, and always caring. I especially want to thank you for your support in going to the NCGR at the least ideal time. I will forever appreciate your mentorship, your jokes, and your prowess in making science come to life through artful storytelling.

I would like to thank my committee member and collaborator, Dr. Marian Waterman, for always having thoughtful questions and creative interpretations of scientific data. I admire your ability to constantly think outside of the box to drive new directions and collaborative projects. Thank you to you and Nate Hoverter for being enthusiastic and committed collaborators. I really enjoyed the comradery and teamwork. Next, I would like to thank my committee member, Dr. Bert Semler for your humor, pragmatic suggestions, and periodic reality checks.

Thank you to my lab family past and present for making all of the failures bearable. Thank you to Will for being so positive and mentoring me in my early days in the lab. You always made late nights in lab seem normal and made it fun. Anke, thanks for always being a good friend who proved you could be awesome at science and still dance with the cool kids after work. Special thanks to my infirmary crew, Maliheh and Wendy. Your friendship and our conversations have kept me sane and hopeful through even the darkest times. Your advice and perspectives on both science and life have been priceless. Even though Bert calls us the "old ladies", I hope that we can still laugh together when we truly fit that title. To the fellas, Francisco and Hossein and Elmira, thank you for breathing new life into the lab. I hope we can continue to meet up in our Ugg's for a cup of Pete's and some tarof.

I thank the *Methods and Protocols-Springer Nature, Journal of Biological Chemistry,* and *Nucleic Acids Research* for permission to include previously published material in my dissertation. Financial support was provided by the NIH, including NIH F31 CA171791 NRSA.

Last, but not least, I would like to thank Atet. You have always been there to show me the light at the end of the tunnel. Thank you for making me laugh when I didn't think I could. Thank you for always encouraging me to step out of my comfort zone to learn new skills and try new things. And finally, thank you for helping to push this project forward when no one else could.

# CURRICULUM VITAE

## Angela A. Garibaldi

## EDUCATION

University of California, Irvine, CA 2018
PhD in Biomedical Sciences

Loyola Marymount University, Los Angeles, CA 2008
Bachelors of Science in Biology

## RESEARCH/PROFESSIONAL EXPERIENCE

Ph.D. Research, Dr. Klemens J. Hertel, Ph.D. 2010-2018
Department of Microbiology and Molecular Genetics
University of California, Irvine

Intern 2016
Bioinformatics
National Center for Genomics Research, Santa Fe, NM

Research Associate III 2010
Research and Development (LABType)
One Lambda Inc (Thermo Fisher Scientific), Canoga Park, CA

Research Technician
Retinal Genetics Laboratory, Dr. Michael Danciger, Ph.D. 2008-2010
Loyola Marymount University, Los Angeles, CA

Laboratory Assistant
Retinal Genetics Laboratory, Dr. Michael Danciger, Ph.D. 2004-2008
Loyola Marymount University, Los Angeles, CA

## AWARDS, FELLOWSHIPS, GRANTS

F31 CA171791 NRSA Individual Fellowship 2012-2016
"Characterizing aberrant alternative-splicing in breast cancer"

National Center for Genomics Research Bioinformatics Internship 2016
Funded by INBRE-NM via Institutional Development Award
NIH P20GM103451

UC Irvine Faculty Mentor Program, Honorable Mention 2011

## TEACHING EXPERIENCE

| | |
|---|---|
| Teaching Assistant<br>Data Science Initiative: Intro to R<br>University of California, Irvine | 2017 |
| Teaching Assistant<br>Cell Biology D103<br>University of California, Irvine | 2015 |
| Teaching Assistant<br>Dev Cell Lab 111<br>University of California, Irvine | 2013 |

## SELECT ABSTRACTS/ORAL PRESENTATIONS

"Large Scale Analysis of Splicing Kinetics Reveals That Alternative Splicing is Promoted by Slow Intron Removal Rates" RNA-SIG Meeting
(2017, Prague, Czech Republic)

"Large Scale Analysis of Splicing Kinetics Reveals That Alternative Splicing is Promoted by Slow Intron Removal Rates" Cold Spring Harbor Laboratories Eukaryotic mRNA Processing Meeting
(2015, Cold Spring Harbor, NY)

## PUBLICATIONS

Garibaldi A., Kao A., Busch A., Hertel K.J. (2018) Large Scale Analysis of Splicing Kinetics Reveals That Alternative Splicing is Promoted by Slow Intron Removal Rates" (in preparation)

Garibaldi A., Carranza F., Hertel K.J. (2017) Isolation of Newly Transcribed RNA Using the Metabolic Label 4-Thiouridine. In: Shi Y. (eds) mRNA Processing. Methods in Molecular Biology, vol 1648. Humana Press, New York, NY

Mueller, W.F., Larsen, L.S., Garibaldi, A., Hatfield, G.W., Hertel, K.J. (2015).The Silent Sway of Splicing by Synonymous Substitutions. Journal of Biological Chemistry,290(46), 27700-27711.

Hoverter, N. P., Zeller, M. D., McQuade, M. M., Garibaldi, A., Busch, A., Selwan, E. M., … Waterman, M. L. (2014). The TCF C-clamp DNA binding domain expands the Wnt transcriptome via alternative target recognition. Nucleic Acids Research, 42(22), 13615–13632.

## OUTREACH

Co-founder, Bioinformatics Support Group 2018
University of California, Irvine

Founder, Cool Kids Computational 2017
University of California, Irvine

Relay for Life, Team Captain 2016
Chao Family Comprehensive Cancer Center
American Cancer Society, Irvine, CA

Relay for Life, Team member 2014-2015
American Cancer Society, Irvine, CA

# ABSTRACT OF THE DISSERTATION

## The Kinetic Scope of Alternatively Spliced pre-mRNA

By

Angela A. Garibaldi

Doctor of Philosophy in Biomedical Sciences

University of California, Irvine, 2018

Professor Klemens J. Hertel, Chair

Eukaryotic gene expression is coordinated through a series of processes from which RNA is transcribed, processed, and translated into the proteins that serve as the functional building blocks of complex cellular organisms. These steps are highly integrated, often occurring in the same spatial and temporal space. Although this co-transcriptional connection is well described, it remains unclear how the concerted rates of global RNA processing steps affect the final mRNA isoform. Here we disrupt steady-state RNA levels using 4-thiouridine (4sU) metabolic labeling and utilize high-throughput sequencing to determine the global rates of pre-mRNA processing. We find that introns that display higher retention are subject to slower splicing kinetics, with longer introns being removed quicker. Exon skipping is subject to competing splice site pairing kinetics and size constraints that highlight optimal exon recognition features by the spliceosome. Integration of this information permits the determination of the order of intron removal across entire genes, thus producing detailed gene RNA processing maps.

Intron retention and exon skipping (cassette exons) are two types of alternative splicing that can be regulated by SR proteins by strengthening the recognition of introns and exons that are otherwise prone to alternative splicing. To test the hypothesis that SR proteins modulate alternative splicing through changes in splicing kinetics, we depleted SRSF1 in human hepatocellular carcinoma cells and derived RNA processing rates. Loss of SRSF1 leads to higher intron retention and more exon skipping. This is primarily achieved through changes in splicing rates and Pol II density, with a strong dependence on optimal feature length constraints. eCLIP data further demonstrates that SRSF1 binds preferentially to weaker exons that are prone to being skipped.

Together these data suggest that alternatively spliced introns and exons have distinct kinetic profiles, constrained by lengths that favor exon definition. The splicing factor SRSF1 acts primarily as an activator, promoting the constitutive splicing of exons and introns through the modulation of Pol II density and subsequent splicing kinetics.

# CHAPTER 1

# Introduction

Gene expression is coordinated through a series of processes from which RNA is transcribed, processed, and translated into the proteins that serve as the functional building blocks of complex cellular organisms. Importantly, each of these processes serve as points of regulation used by the cell to efficiently modify its gene expression profile to accommodate both reversible and irreversible changes and adaptations. In recent years it has been increasingly appreciated that most of these processes are dynamic and co-dependent. This chapter will introduce the cascade of steps that are required for pre-mRNA processing to occur, the influence of co-dependent steps on one another, and the contribution of regulatory elements on RNA processing dynamics.

## Pre-mRNA Splicing

Pre-mRNA splicing is an essential process required for the expression of genes in metazoans. Splicing describes the processing of nascent pre-mRNAs into protein-coding mRNA via the excision of non-coding intronic regions and the subsequent ligation of the flanking exonic sequences. Pre-mRNA splicing is carried out by a large ribonucleic protein complex comprised of over 300 proteins. Of these proteins, the uridine-rich small nuclear ribonucleoproteins (snRNPs) U1, U2, U4, U5, and U6 play primary roles in splice site recognition, selection, and excision of introns [1, 2] .

The first step of splicing requires the recognition of the 5' and 3' splice sites (ss). The 5'ss is defined by U1 snRNP binding the 9 nucleotide consensus sequence, YAG/guragu (where Y is a pyrimidine and R is a purine), located at the junction between the 3' end of an exon and the 5' end of the downstream intron [3]. The 3'ss is defined through the contributions of three sequence elements located at the 3' intron/exon junction. Spliceosome component U2AF binds to the polypyrimidine tract (PPT), a region containing 15-20 pyrimidines (C or U), located approximately 18-40 nucleotides upstream of the 3'ss (Figure 1.1) [4]. The U2AF subunit U2AF35 associates with the 3'ss [5]. The branch point sequence (BPS) is a highly degenerate sequence flanking a conserved branch point adenosine that serves as a binding site for SF1 [6, 7]. U2AF65 and U2AF35 work in concert to recruit U2 snRNP to replace SF1 at the BPS [8]. These combined associations define an ATP-independent step early in the process of spliceosome assembly (E Complex). After stable binding of U2 snRNP to the BPS, ATP hydrolysis solidifies the spliceosome's commitment to the defined splice site pair (A Complex) – technically, U2 stably associates with A complex formation, ie after ATP hydrolysis [9]. The B Complex and catalytically-active C Complex are formed after subsequent recruitment and rearrangement of the U4/U5/U6 tri-snRNP. Finally, two trans-esterification reactions occur in which the 5'ss phosphate of the exonic junction nucleotide is attacked by the 2'OH of the branchpoint adenosine, followed by the ligation of the 5' and 3 exons and the excision of an intron lariat [10].

Alternative splicing is a process whereby different splice sites are selected, ultimately leading to the generation of multiple mRNA isoforms from single genes. These mRNA isoforms can employ similar, different, or opposing functions through the resulting protein

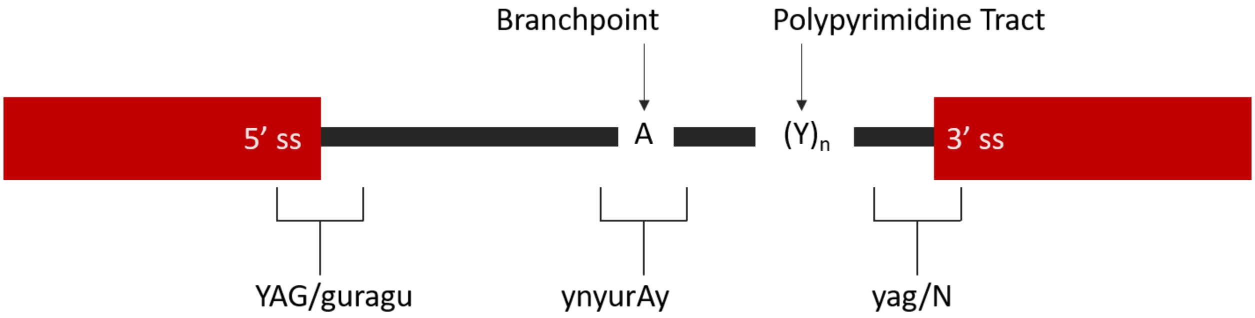


**Figure 1.1. Sequence Elements for Splice Site Recognition**
Schematic of location and consensus sequences for the sequence elements required for spliceosomal recognition: exon/intron junction (5'ss), intron/exon (3'ss) junction, Branchpoint Sequence (BPS), and Polypyrimidine Tract (PPT). Y refers to a pyrimidine (C or U nucleotide), R refers to a purine (A or G nucleotide), N refers to any nucleotide, and "/" denotes a junction. The intron is represented as a thick black line, and exons as the flanking rectangles.

isoforms [11]. With the advent of genome-wide and proteome-wide studies, approximately 20,000 protein-coding genes have been identified [12]. In agreement with the over 66,000 alternatively spliced isoforms annotated by the UCSC Genome Institute, previous studies show that alternative splicing occurs in approximately 95% of multi-exon genes [12, 13]. Different types of alternative splicing have been characterized: alternative 5' splice site selection (5'ss), alternative 3' splice site selection (3'ss), skipped exons (SE), retained introns (RI), and mutually exclusive exons (MXE). These modes of alternative splicing are responsible for the rearrangements of coding sequence that either increase proteomic diversity, or reveal pre-mature stop codons that trigger degradation of the offending mRNA by surveillance machineries, effectively silencing production of the intended protein by that mRNA [14].

While the predominant form of alternative splicing found in various cancers is exon skipping (50-60%), intron retention has also been identified as significantly more frequent compared to normal cells (20%) [15]. Alternative inclusion or exclusion of an intron or exon suggests an inefficiency of splice site recognition by the spliceosome, likely due to the abundance of splice site sequences that stray from the consensus. However, the number of pseudo exons/splice sites contained in introns vastly outnumbers the quantity of real splice sites [16]. This implies that additional regulatory factors and features are employed to help guide the spliceosome to authentic splice sites.

## Regulation of Splicing

Considering the role of a multitude of splicing factors that must bind to a range of degenerate and yet specific sequences, it is important to consider distance constraints in the context of protein binding and interactions.  Intron/exon architecture refers to the length of exons and introns and the influence this feature has on the approach the spliceosome takes to recognize splice sites. *In vitro*- and EST-based studies have demonstrated that recognition of introns and exons larger than 300 nt is inefficient, except where flanking exons/introns are much shorter [17, 18]. Additionally, exon skipping has been shown to be more common when the exons are flanked by long introns. These data support the hypothesis that there is an optimal distance across which splice site recognition occurs. Indeed, splice site recognition can be accomplished either through intron definition or exon definition (Figure 1.2).

### ***Intron/Exon Definition***

Intron definition is used to recognize 5'ss and 3'ss across an intron when introns are shorter than 250 nucleotides. Exon definition is utilized for the recognition of splice sites across a short exon when flanking introns are longer than 300 nucleotides. The average human intron length is 3.4kb, yet 55% of introns in gene-coding regions range between 100bp-2000bp in length. More than 80% of all annotated exons are less than 200bp in length [19, 20]. Considering the genomic bias toward short exons and long introns, it is not surprising that human splice sites are typically recognized via exon definition [21]. To summarize, the spliceosome employs the method of recognition that best optimizes distance parameters.

In addition to optimal distance, efficient splice site recognition also depends on sequence complementarity. Multiple sequences are responsible for the effective recognition of the 5' and 3' splice sites by U1 snRNP and U2AF respectively. A strong 5' splice site is defined by high sequence complementarity to U1 snRNP. A strong 3' splice site is determined by a

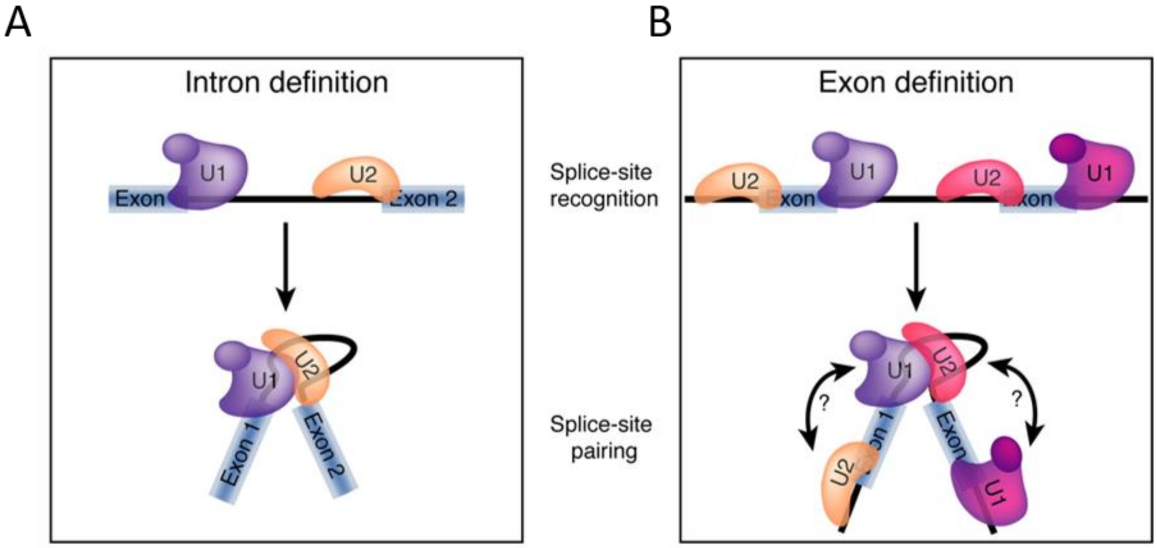


**Figure 1.2. Splice site recognition by distance constraint models**

A. Intron definition is the mode of recognition when introns are short (less than ~250 nts). Splice site pairing occurs across the intron.

B. Exon definition is the mode of splice site recognition when introns are long (greater than~250 nts). Splice sites are paired across the exon.

contribution of three sequence elements: the PPT, BPS, and the 3' intron/exon junction splice site [3].

Popular Maximum Entropy modeling utilizes these 9nt and 23nt sequence features to determine a splice site strength score which can distinguish splice sites from decoy sites [22]. Combining these scores can be used to distinguish constitutively spliced exons from alternatively spliced exons [23]. However, additional sequence elements have been shown to recruit non-spliceosome proteins that can alter splice site recognition otherwise determined by complementarity, context, and architecture.

### *Splicing Regulatory Elements*

Splicing regulatory elements (SRE) are non-splice site sequence elements that serve as binding sites for regulatory proteins that can improve or reduce spliceosome recruitment to nearby splice sites. SREs that are located within exons are considered ESE (exon splicing enhancers) or ESS (exon splicing silencers) depending on whether the regulatory protein that binds increases or decreases inclusion of the exon. Similarly, SREs that are located within introns are called ISE (intron splicing enhancers) or ISS (intron splicing silencers) [24]. The primary classes of proteins that bind to SREs are serine/arginine rich (SR), known for being splicing activators, and heterogenous nuclear ribonucleoproteins (hnRNPs), known for being splicing repressors.

SR proteins contain a serine/arginine rich C-terminal domain (RS domain) and an RNA recognition domain at the N-terminus (RRM domain). ESE-dependent SR protein binding sites have been identified in both alternatively spliced and constitutively spliced exons. SR proteins can induce the inclusion of an alternative exon through increasing the recognition of weak splice sites. Importantly, the regulatory effect that SR and hnRNP family proteins play are position and context dependent [25]. For example, SR proteins bound in the exon act as activators, but intron-bound SR proteins may repress [26]. Moreover, SREs in close proximity that recruit SR proteins or hnRNPs to locations that direct opposing influences are additive and may neutralize or exacerbate activation or repression [27].

### SRSF1: A Classical SR Protein

SRSF1, previously known as SF2/ASF, was one of the first SR proteins to be identified and characterized [28, 29]. Its original depiction defined its role in promoting spliceosome assembly to preserve constitutive pre-mRNA splicing, and to regulate alternative splicing [29]. Its primary identification as a splicing factor has been superseded by the discovery of its additional functions in regulating: mRNA transcription, stability, nuclear export, NMD, translation, and protein sumoylation [30-33]. With roles in practically every step of gene expression, it is not surprising that it has also been designated as a proto-oncogene [34, 35].

Crosslinking immunoprecipitation and high-throughput sequencing (CLIP-seq) experiments revealed widespread binding of SRSF1, primarily to exonic regions [27]. Such extensive binding profiles permitted the identification of GGAGA as its consensus motif [36]. SRSF1 recognizes ESE sequence elements on its target pre-mRNA and promotes exon definition by facilitating early spliceosome recognition of the proximal alternative 5'ss or 3'ss [37]. Deletion of the RS domain has shown that is it not essential for SRSF1's splicing activity [38] . However, the RS domain is required for its export to the cytoplasm, and regulation of its subnuclear localization [39]. The RRM domain mediates interactions with the U1-70K subunit of U1 snRNP, with the RS domain contributing regulatory functions [40].

Extensive phosphorylation and dephosphorylation of the RS domain serine residues is a critical mechanism for controlling SRSF1's interaction with other proteins, ability to bind RNA targets, and subcellular localization [41, 42] . Partial dephosphorylation of the RS domain permits SRSF1 binding to the TAP/NXF1 receptor whereby it serves as an export adaptor to facilitate the nuclear export of bound mRNAs [43, 44] . Therefore, differential phosphorylation of SRSF1 is a mechanism for regulating SRSF1's role in nuclear export.

SRSF1 is present in the cytoplasm, found to associate with polyribosomes in cytoplasmic extracts [45] . It enhances cap-dependent translation initiation through its activation of the mTORC1 signaling pathway through two mechanisms. First, it drives the expression of the MNK2b isoform, which phosphorylates translation initiation factor eIF4E. Second, it modulates the expression of S6 kinase 1 short mRNA isoforms, which bind to mTORC and enhance 4EBP1 phosphorylation [46, 47]. High-throughput sequencing of polysome fractions in cells overexpressing SRSF1 identified ~1500 mRNAs that are translational targets of SRSF1. The identified mRNAs encode proteins involved in cell mitosis, possibly explaining why cells with reduced levels of SRSF1 fail to divide properly [48].

Consistent with the critical processes it regulates, knock out of *SRSF1* is embryonic lethal in mice [49]. Yet, even slight overexpression of *SRSF1* in human mammary epithelial cells leads to oncogenic transformation, forming cancerous tumors upon transplantation into mouse models. Overexpression of SRSF1 in lung adenocarcinoma cells causes even a more aggressive phenotype, conferring resistance to carboplatin and paclitaxel [50] .

While many mechanisms have been elucidated, SRSF1's interactions with well-known cancer promoting pathways are at the forefront. SRSF1 is directly targeted and positively regulated by the potent oncogene *MYC* via two non-canonical E-boxes in its promoter [51].

Likewise, SRSF1 promotes Wnt signaling-mediated tumorigenesis by enhancing the translation of β-catenin mRNA through activation of the mTOR pathway [52]. Cell motility and invasion are amplified through the production of a protein isoform of the macrophage stimulating protein tyrosine kinase receptor RON, produced by splicing changes that SRSF1 triggers [35]. In response to these attacks on cellular regulation, p53 induction in response to *SRSF1* overexpression in primary human cells leads to premature cellular senescence. Therefore, SRSF1-mediated oncogenesis appears to rely on the inactivation of the p53 tumor suppressor pathway [53].

## Co-transcriptional pre-mRNA Processing

Eukaryotic gene expression progresses through a series of steps, starting with RNA Polymerase II (Pol II) transcription of template DNA into nascent pre-mRNA and a coordinated addition of a 5' 7-methylguanylate cap [54]. Immunofluorescence and chromatin immunoprecipitation (ChIP) experiments tracking U1 snRNP binding show that spliceosome components assemble around splice sites on the nascent pre-mRNA immediately after they are transcribed [55, 56]. Chromatin-RNA immunoprecipitation experiments show that RNA is physically linked by transcribing Poll II to chromatin recruited splicing factors U2AF65, U1, and U5 snRNPs to intron containing genes. That is, spliceosome assembly largely occurs co-transcriptionally. When Pol II elongation is stalled with camptothecin, an increase in splicing factor accumulation and splicing occurs, substantiating a kinetic link between transcription, spliceosome assembly, and splicing catalysis [57, 58].

Previous studies also indicated that a decrease in transcription rate, often caused by transcriptional pause sites, allows more time for upstream splicing to occur, thereby increasing alternative exon inclusion [59, 60]. In fact, work from the Hertel lab demonstrated that delaying the synthesis of downstream exons increased upstream exon inclusion [61]. Furthermore, the relative rate at which competing exons are produced can bring about changes in splicing patterns through a kinetic advantage to otherwise weaker splice sites [2, 62]. The time element stipulated through co-transcriptional spliceosomal assembly may also create a window of opportunity for both positive and negative splicing regulators to recognize their binding sites, providing prospects to induce alternative splicing (Figure 1.3) [25].

Studies in yeast, *Drosophila*, and human cells using chromatin isolation, RT-PCR, and cell fractionation approaches demonstrate that splicing can occur co-transcriptionally; that is, splice site recognition and intron excision can be completed while the nascent RNA is still connected to chromatin by elongating Pol II [63]. Mounting evidence suggests that as much as 50-95% of pre-mRNA processing occurs co-transcriptionally [59, 64]. While this coupling with transcription has been shown to affect splicing on a gene-by-gene basis in yeast, human, and drosophila, a genome-wide understanding of this connection has yet to be elucidated in human.

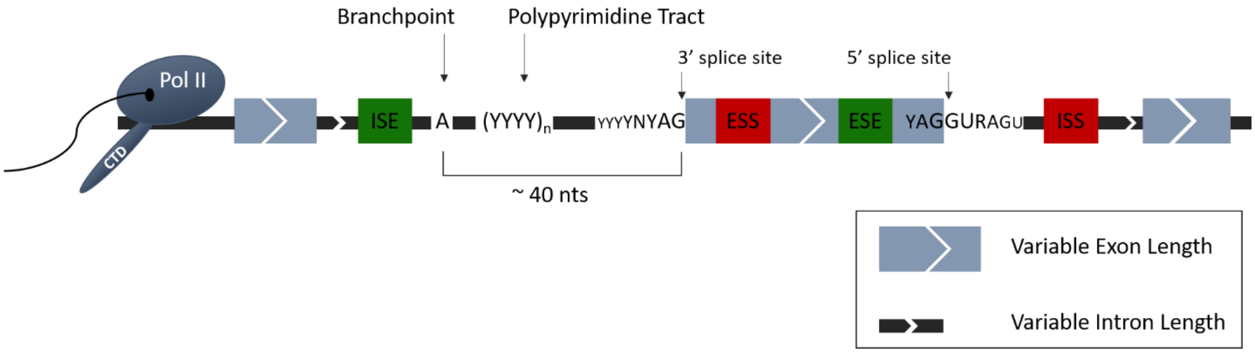


**Figure 1.3. Splicing regulatory components.**
Depiction of parameters that influence the affinity of spliceosomal components to the exon in higher eukaryotes. These include splice site strength, exon/intron architecture, the presence or absence of splicing enhancers/silencers, and the rate of transcription by RNA Polymerase II. CTD represents the C-terminal domain.

## Summary

Pre-mRNA processing is a complex and dynamic process that is influenced by multiple elements and factors simultaneously. Not surprisingly, the mechanistic and temporal link between transcription and splicing further convolute the decision tree behind alternative splicing outcomes. Historically, gene expression studies were limited to the analysis of steady-state levels of expressed mRNAs, ultimately negating rapid adaptations in processing steps by the cell. Recent studies have employed low-throughput methods that are limited to investigating a handful of genes at once and lack the ability to take measurements at precise time intervals. Larger scale approaches have been limited to lower eukaryotes, neglecting human-specific genome-wide splicing kinetics.

The following work aims to close the gap in knowledge of global pre-mRNA processing dynamics using a nucleotide resolution approach. Chapter 2 describes a genome-wide determination of splicing kinetics by metabolically labeling human cells with the uridine analog 4-thiouridine. Isolation of nascent pre-mRNAs over a time course enables the tracking of pre-mRNA as they are synthesized, processed through intron removal, ultimately resulting in mRNAs of variable intron and exon retention levels. The cell based kinetic approach demonstrates that alternative splicing is mainly driven by slow splicing kinetics and that splice site strength and the intron/exon architecture are the foundation for variable intron excision speeds.

Chapter 3 discusses the influence of an essential SR protein, SRSF1, on splicing kinetics. Knockdown of SRSF1 in HepG2 cells paired with an expanded time course and replicate samples provide high resolution representations of how a renowned oncoprotein and splicing regulator directly drives the switch from exon inclusion to exclusion.

# CHAPTER 2

# Large Scale Analysis of Splicing Kinetics Reveals That Alternative Splicing is Promoted by Slow Intron Removal Rates

## Summary

Each step of eukaryotic RNA processing can serve as a point of regulation for the purpose of modifying gene expression. These steps are highly intertwined both physically and temporally. However, it is unclear how the rates of global RNA processing events affect the final mRNA isoform. Methods to assess how RNA processing affects alternative splicing have been historically limited to steady-state levels; preventing the assessment of changes in these steps in real-time. Here we use 4-thiorudine (4sU) metabolic labeling to globally determine the rates of RNA processing steps. Through this analysis, we find that alternatively spliced introns are subject to slower splicing kinetics, with longer introns being removed more rapidly. Exon inclusion levels are subject to competing splice site pairing kinetics and size constraints that highlight optimal exon recognition features by the spliceosome.

## Introduction

Eukaryotic gene expression is a dynamic process regulated at multiple points in its genomic workflow. Aberrant regulation of any one of these steps can be deleterious to the cells ability to adapt effectively. RNA Polymerase II (Pol II) is responsible for synthesizing the majority of protein-coding genes in human cells via transcription of template DNA. 92% of human genes contain introns that must be removed through the recognition of splice sites and neighboring sequence elements by the spliceosome to produce a mature RNA (mRNA) that is translation competent [65]. Spliceosome components are recruited to the nascent RNA prior to a gene being completely transcribed [66]. The majority of intron removal has been shown to occur co-transcriptionally in the lower eukaryotes drosophila and yeast [67].

Previous studies of gene expression were limited to analyzing whole cell RNA, representing the cumulative, steady-state effect of each dynamic step in processing. To unmask the individual contributions of each step, the field has developed experimental approaches using external cellular stimuli paired with cellular fractionation, RNA Immuno-Precipitation (RIP), U1 snRNP ChIP, live cell imaging, and targeted RT-PCR technologies to interrogate the state of intron-containing pre-mRNAs. Live cell imaging and RT-PCR approaches are limited to assessing only a handful of genes at a time [68-70]. Extracting RNA bound to chromatin requires cell fractionation methods, which have inherently low time resolution due to the lack of a fast quench to the time course reaction [64, 71]. More recent studies have made progress in utilizing 4-thiouridine (4sU) metabolic labeling for well-defined time point experiments. However, they required creative strategies to overcome sampling of only a few time points, or the lack of replicates [72, 73]. This ultimately reduces the ability to derive true rates of processing for a large number of introns in organisms where most genes have multiple long introns, as is the case for humans.

Here we use high-throughput sequencing of 4sU labeled pre-mRNAs to generate a large-scale human dataset that offers unprecedented sampling depth and that is unique in its extensive time series and the number of replicates. This metabolic labeling dataset permits a high-resolution determination of RNA synthesis, rates of pre-mRNA processing, and projections of steady-state intron and exon retention levels. The analysis demonstrates that slow splicing kinetics are a hallmark of alternative splicing.

## Results

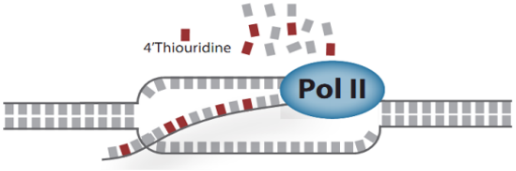
### *Approach to determine pre-mRNA splicing rates*

To determine intron removal rates at a genome-wide level, we labeled HepG2 cells transduced with a non-target shRNA with the uridine analog 4-thiouridine (4sU) for an extensive time series (0, 2, 5, 10, 20, 30, 50, 60, 70, 90, 100, 120 minutes). Following isolation of nascent pre-mRNAs transcribed with 4-thiouridine, high throughput-sequencing was performed. Reads were normalized to an ERCC spike-in to control for global changes in signal across each time point. Transcripts per Million (TPM) were calculated for each expressed intron and exon to further normalize for library size and feature length [74]. TPMs were used to calculate a "Fraction of Intron Inclusion", by which the TPM of an intron is divided by the mean of its flanking exons.  To identify exons that may be alternatively skipped, the same calculation was performed for an exon relative to its flanking exons.

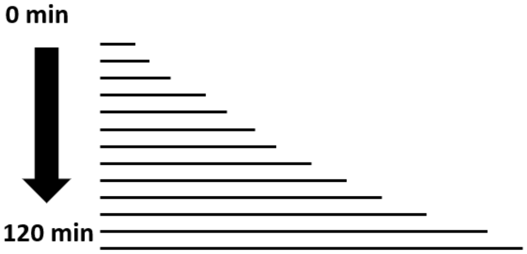
This approach provides an internally controlled measure of intron appearance upon synthesis in relation to its flanking exons, and the subsequent disappearance of the intron relative to constitutively included exons (Figure 2.1 (1-7)). These "Fractions of Intron Inclusion" are tracked over time and modeled to fit a consecutive intermediates model, a kinetic equation that describes the synthesis (appearance) of the intron followed by removal through spliceosomal excision (disappearance) (Figure 2.1 (8)).

**Figure 2.1. Workflow of 4sU labeling and determination of rates.** 1) Uridine analog 4-thiouridine (4sU) is added to cells and incorporates into all newly transcribed RNA. 2) Incubation with 4sU is done for a series of time points: 0, 2, 5, 10, 20, 30, 40, 50, 60, 70, 90, 100, 120min, and quenched by TRIzol extraction. 3). Labeled nascent RNA is biotinylated by the sulfyl hydryl group of 4sU, and isolated with streptavidin columns. 4) RNA-seq libraries are prepared with the Illumina TruSeq stranded mRNA protocol, without polyA selection. 5. Libraries are sequenced for 100 cycle single end reads. 6) TPMCalculator is used derive reads for each intron and exon. TPMs are calculated after normalization to ERCC. 7) Intron TPM is divided by the mean TPMs of its flanking exons to determine the Fraction of Intron Retention. 8) Fractions of each time point are used to model RNA dynamics with a kinetic equation that describes the synthesis (k1), removal (k2), and steady state levels (k3).

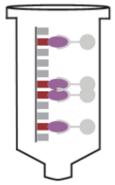
1 . Pulse 4-thiouridine into cells time series



2. Quench time point with TRIzol extraction



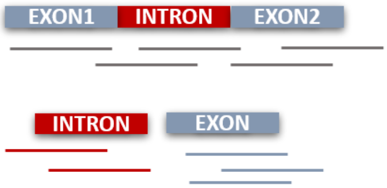
3. Biotinylate and isolate labeled RNA with streptavidin columns



4. Prepare RNA-seq libraries (Not Poly A  Selected)

5. Sequence Illumina 100 SE reads

6. Calculate expression value (TPM) for each exon and intron

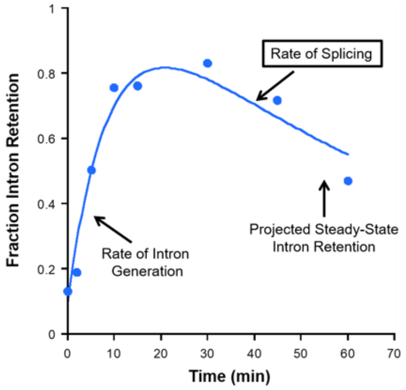


7. Calculate the fraction of inclusion

$$\frac{\text{INTRON}}{\dfrac{(\text{EXON} + \text{EXON})}{2}}$$

8. Model fractions to kinetic equation to determine RNA synthesis and splicing rates

$$\text{DNA} \underset{\text{RNA synthesis}}{\xrightarrow{k_1}} \text{pre-mRNA} \underset{\text{Intron removal}}{\xrightarrow{k_2}} \overset{k_3}{\text{mRNA}}$$



$$[\text{intron}] = k_1/(k_2-k_1)\ (e^{-k_1 t} - e^{-k_2 t}) + k_3$$

### *Global features of intron removal dynamics*

We analyzed intron and exon behavior of 30,698 canonical and non-canonical genes that displayed sufficient coverage and expression level (see Methods and Materials). From this pool 80,840 introns and 13,332 exons were selected to be further analyzed based on their fit to the kinetic model (see Materials and Methods). These modeled introns and exons represent 13,245 genes. Exons that fit well to the consecutive intermediate model behave kinetically like introns. They appear at similar representation to flanking exons in early time points, but they become more underrepresented at later times during the labeling experiment. The intron-like behavior of exons strongly suggests that they are alternatively skipped.

A regression approach was used to find the optimal fit to the consecutive intermediate equation using three variables. The first variable, k1, refers to the rate at which exons and introns are generated. The second variable, k2, describes the rate of intron disappearance, which is a measure for the rate of pre-mRNA splicing. The third variable represents an approximation for the steady-state levels of intron retention. Early time points are enriched for intron containing transcripts, thus allowing the tracking of the overall generation and removal of these gene segments (Figure 2.2).

The rate of intron/exon generation does not reflect the speed of Poll II transcription elongation because Pol II was not synchronized with the inhibitor 5,6-Dichloro-1-β-D-ribofuranosylbenzimidazole (DRB) [68]. In the experiments described here, 4sU disrupts the steady-state production of RNA, injecting itself into the crowd of genes being actively transcribed in the nucleus. No matter which position Pol II is occupying at the time of 4sU addition, 4sU is incorporated and the entire RNA can theoretically be pulled down in

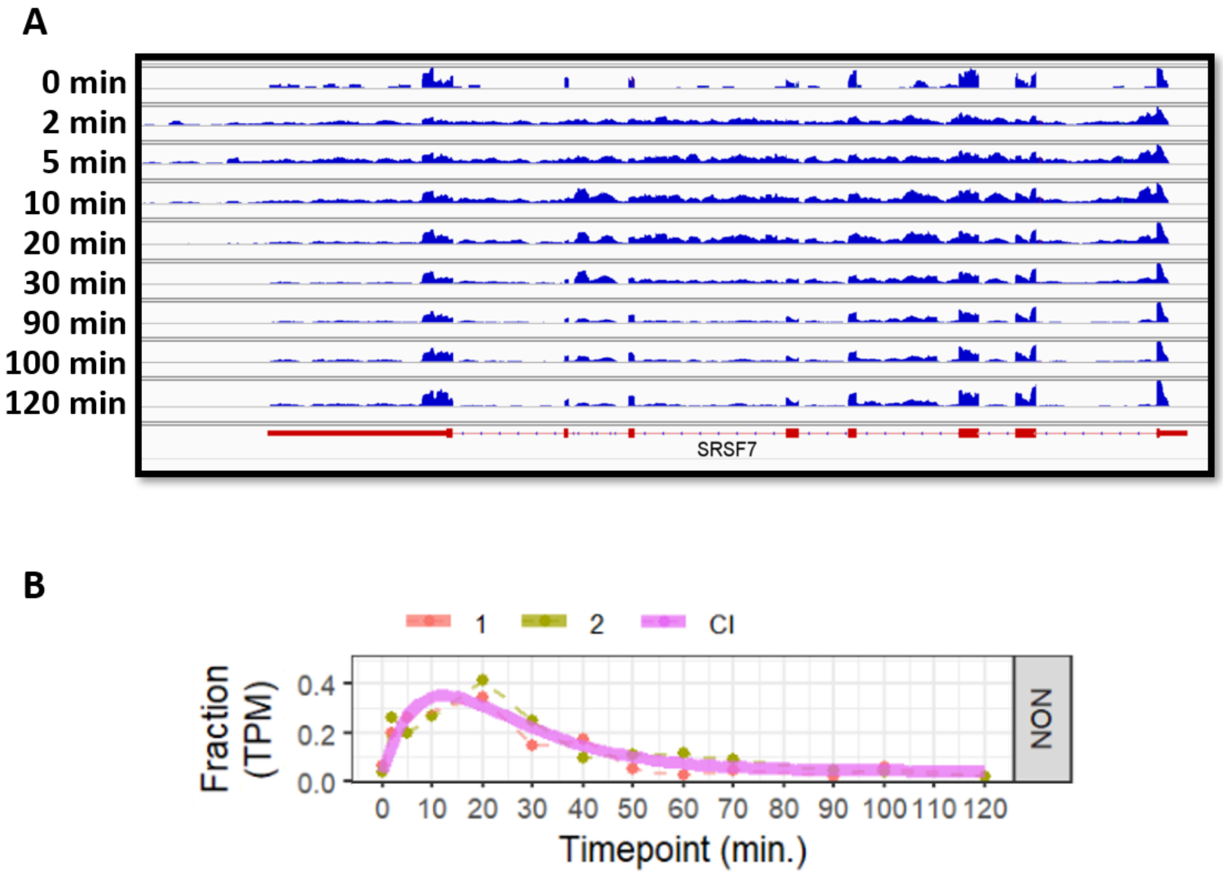


**Figure 2.2. Progressive 4sUlabeling permits the tracking of synthesis and splicing.** A) Read coverage of *SRSF7* RNA levels at 0, 2, 5, 10, 20, 30, 90, 100, 120 min. 40, 50, 60, 70 min are not show. RNA production can be seen across the gene in earlier time points. The removal status of introns can be seen at later time points. B) Intron from *ABCC1* shown as example of replicate data plotted and fit to the model. Fraction of intron retention relative to flanking exons are plotted in Transcripts Per Million (TPM) mapped reads. Timepoint in minutes. Red dotted line shows replicate 1, green dotted line shows replicate 2, and pink solid line shows the model fit through the data points.

subsequent isolation steps. Thus, the more appropriate interpretation of the RNA synthesis is that it represents the density of Pol II on a defined mRNA segment. The rate reflects how frequently a gene segment is generated during a defined time period. Overall, populations of introns have a median synthesis rate of 0.098 min$^{-1}$. 85% of introns have synthesis rate less than 0.2 min$^{-1}$, suggesting a narrow range of Pol II density along a given transcript (Figure 2.3A).

Splicing rates determined by tracking the disappearance of an intron relative to the immediately upstream and downstream exons were converted into a half-life estimation using $t_{1/2}$ calculation. Half-life distributions convey 73% of the introns analyzed have a half-life of under 8 min, suggesting that they are efficiently removed (Figure 2.3B). These intron removal kinetics are in agreement with previous PCR-based rate determinations for intron of constitutively spliced genes [75].

Projected levels of intron retention at steady-state levels are modeled from the time course largely relying on later time points when intron removal is approximating completion. 12% of introns have a projected retention level of >0.2 (20%) (Figure 2.3C). Constitutively spliced introns are expected to be predominately removed at steady-state with small contributions of expression from newly transcribed copies as labeling times progress. Based on the intron retention distribution, we hypothesize that introns with predicted retention levels greater than 0.2, or 20%, are likely to be retained with distinct processing kinetics. Thus, the bulk of lower intron retention dynamics are likely to represent ongoing nuclear processing.

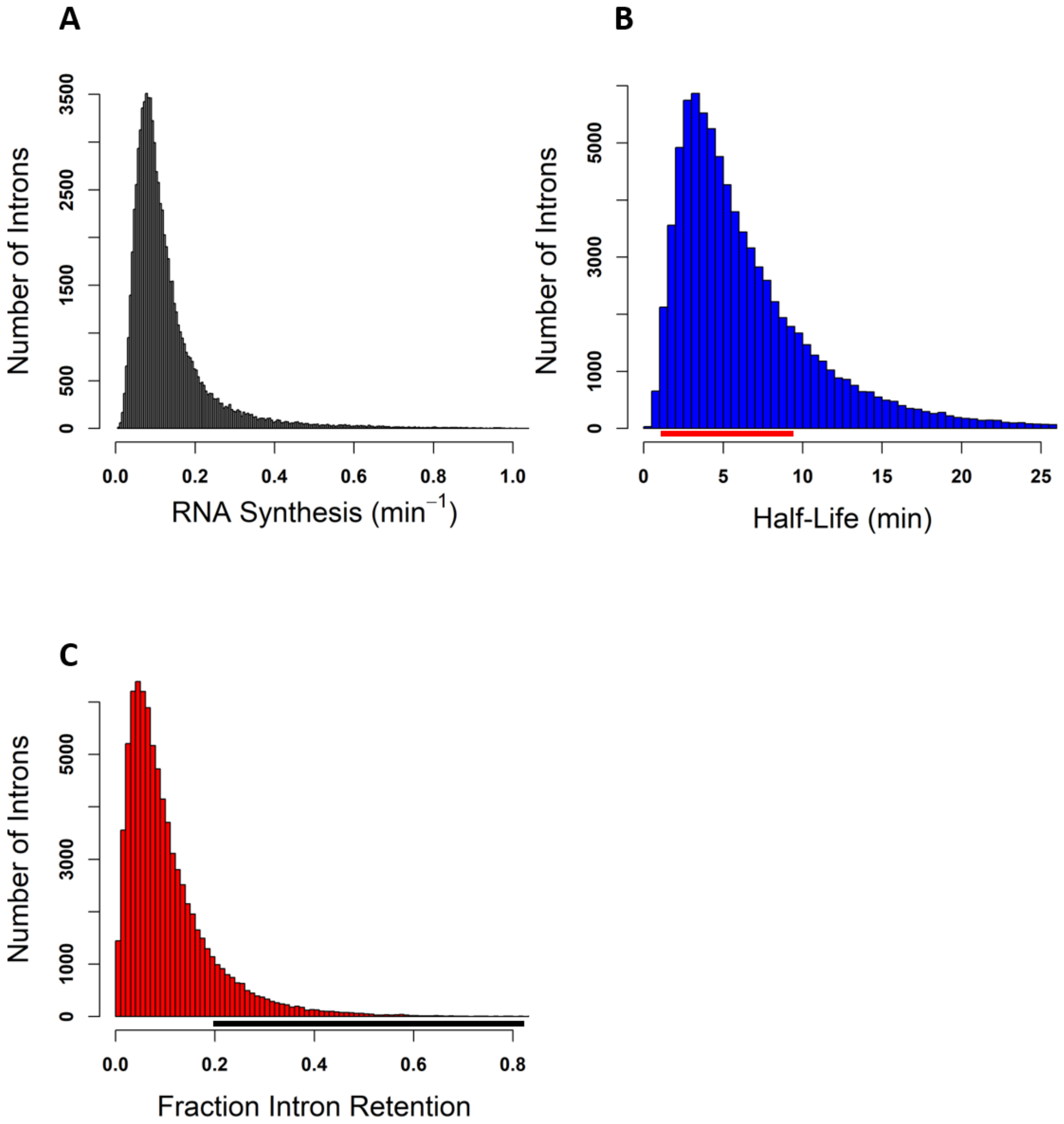


**Figure 2.3. Global scope of intron dynamics.** A) Distribution of RNA synthesis levels across all modeled introns. B) Distribution of half-life values for all modeled introns. Red line marks half-life values that confer constitutive splicing (between 0.4-8 min). C) Distribution of predicted steady-state intron retention levels of all modeled introns. Black line denotes introns with higher retention than constitutive introns (greater than 0.2 or 20% retention).

### *Higher intron retention is promoted by slower splicing kinetics*

To test the hypothesis that intron retention levels correlate directly with intron removal kinetics, we carried out a retention distribution analysis relative to corresponding intron removal half-lives. Interestingly, slower splicing kinetics strongly correlate with higher retention levels (p value 2.2e-16) (Figure 2.4A). We conclude that slower slicing kinetics result in increased intron retention.

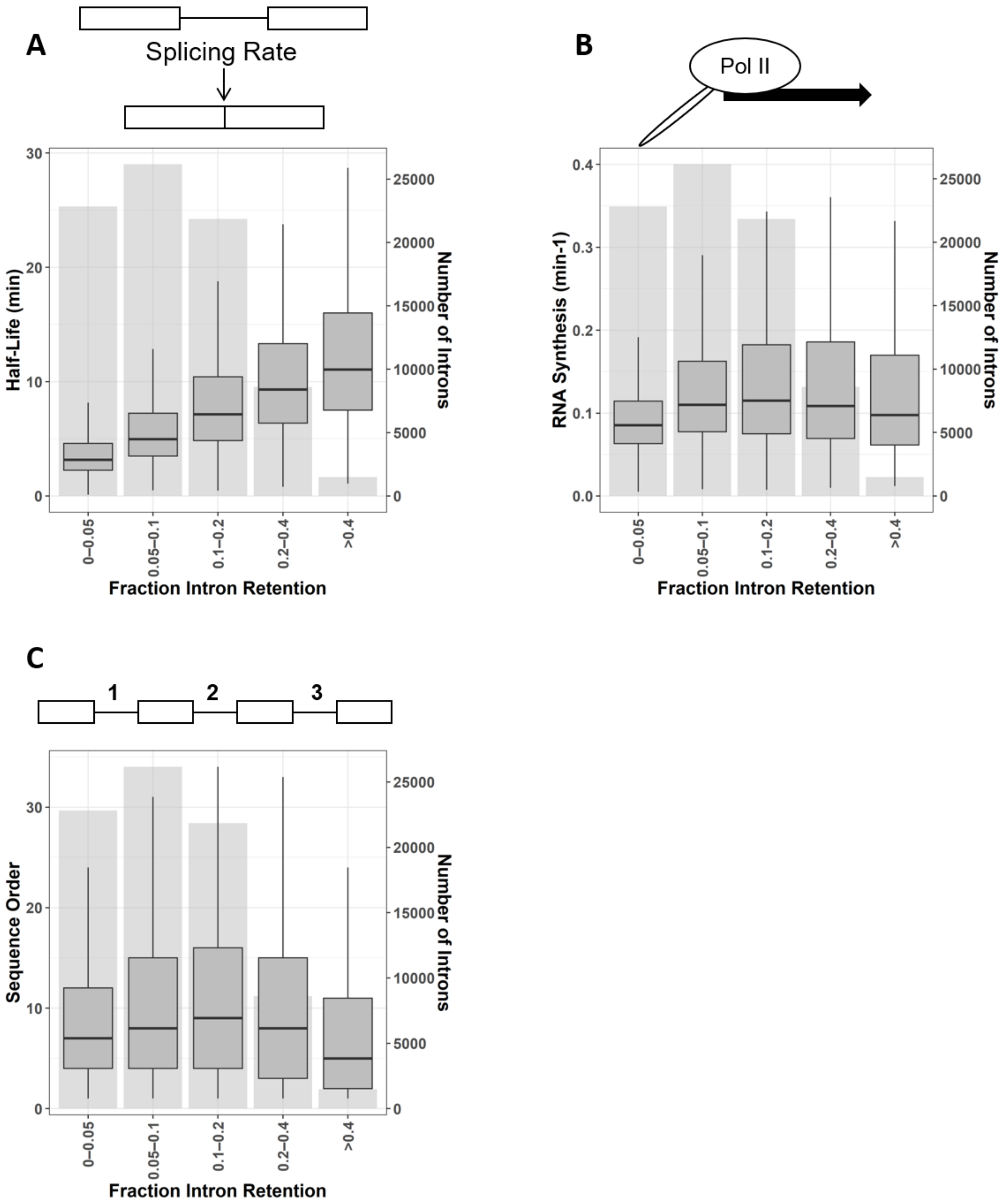
To test whether the rate of intron synthesis, or the Pol II density across the intron, influences intron retention an analogous correlation was generated using RNA synthesis as a readout. While RNA synthesis appears slightly reduced for introns that have very low retention levels (Figure 2.4B), the observed differences are less striking.

Fragmenting genes into exon and intron segments provides the opportunity to determine whether the genomic location of an intron influences its speed of removal or its propensity to remain retained. Interestingly, our data indicates that the intron order within a gene influences its removal efficiency (Figure 2.4C). Highly retained introns are more likely to be located at the 5' end of a gene.

### *Splicing kinetics are influenced by intron architecture*

Previous *in vitro* splicing and reporter assays demonstrated that the length of flanking introns can influence splicing kinetics and alternative splicing pattern. It was argued that the intron definition mode of splice site recognition is more efficient than the exon definition mode due to differences in spliceosomal assembly mechanisms. To test whether these *in vitro* observations are also observed when evaluating endogenous genes, a correlation between intron size and splicing kinetics was carried out. Surprisingly, the

**Figure 2.4. Intron retention relationship of splicing rates and synthesis.** Box plots depict relationship. Bar plot shows the number of introns for each bin. A) Boxplot shows the distribution of predicted intron retention values compared to the half-life (min). B) Boxplot depicts the distribution of predicted intron retention values compared to the RNA synthesis estimate. C) Boxplot shows the relationship of intron retention with the sequence (intron) position within the gene.


A
Splicing Rate
Half-Life (min)
0
10
20
30
Number of Introns
0
5000
10000
15000
20000
25000
0–0.05
0.05–0.1
0.1–0.2
0.2–0.4
>0.4
Fraction Intron Retention
B
Pol II
RNA Synthesis (min-1)
0.0
0.1
0.2
0.3
0.4
Number of Introns
0
5000
10000
15000
20000
25000
0–0.05
0.05–0.1
0.1–0.2
0.2–0.4
>0.4
Fraction Intron Retention
C
1
2
3
Sequence Order
0
10
20
30
Number of Introns
0
5000
10000
15000
20000
25000
0–0.05
0.05–0.1
0.1–0.2
0.2–0.4
>0.4
Fraction Intron Retention

global analysis of endogenous intron splicing clearly shows that intron excision kinetics increase as intron length increases (Figure 2.5A). These results hold up even if intron size bins below and above the 250 nt size transition between intron and exon definition are evaluated (data not shown). We conclude that human intron splicing has successfully adapted mechanisms to compensate for the challenges of pairing splice sites far apart. Such mechanisms could include exon tethering to elongating Pol II as recently suggested [76].

An evaluation of combined splice site scores across introns demonstrates that introns with the shortest half-life display slightly higher median MaxEnt splice site scores (MES) than all other bins, consistent with the idea that stronger splice sites increase splicing efficiency. However, the combined splice site scores across all half-life bins are robust and fail to show significant differences between most half-life bins (Figure 2.5B).

To investigate if the intronic location within a gene influences intron removal kinetics we analyzed 1,970 genes that contained at least 6 introns and for which we had reliable rate determinations. Interestingly, this analysis revealed that first introns are spliced significantly more slowly compared to internal introns. Similarly, last introns are also removed at a slower rate (Figure 2.6A). These results demonstrate that terminal intron removal is unique. Flanked by a capped upstream exon the first intron may be recognized through different mechanisms when compared to internal introns. The same reasoning can be proposed for last introns, which are flanked by longer terminal exons that may or may not yet be polyadenylated. The interplay between the capping or polyadenylation machinery with spliceosomal factors at first and last introns is likely to significantly influence their excision rates. As is expected from slower intron removal kinetics, first and

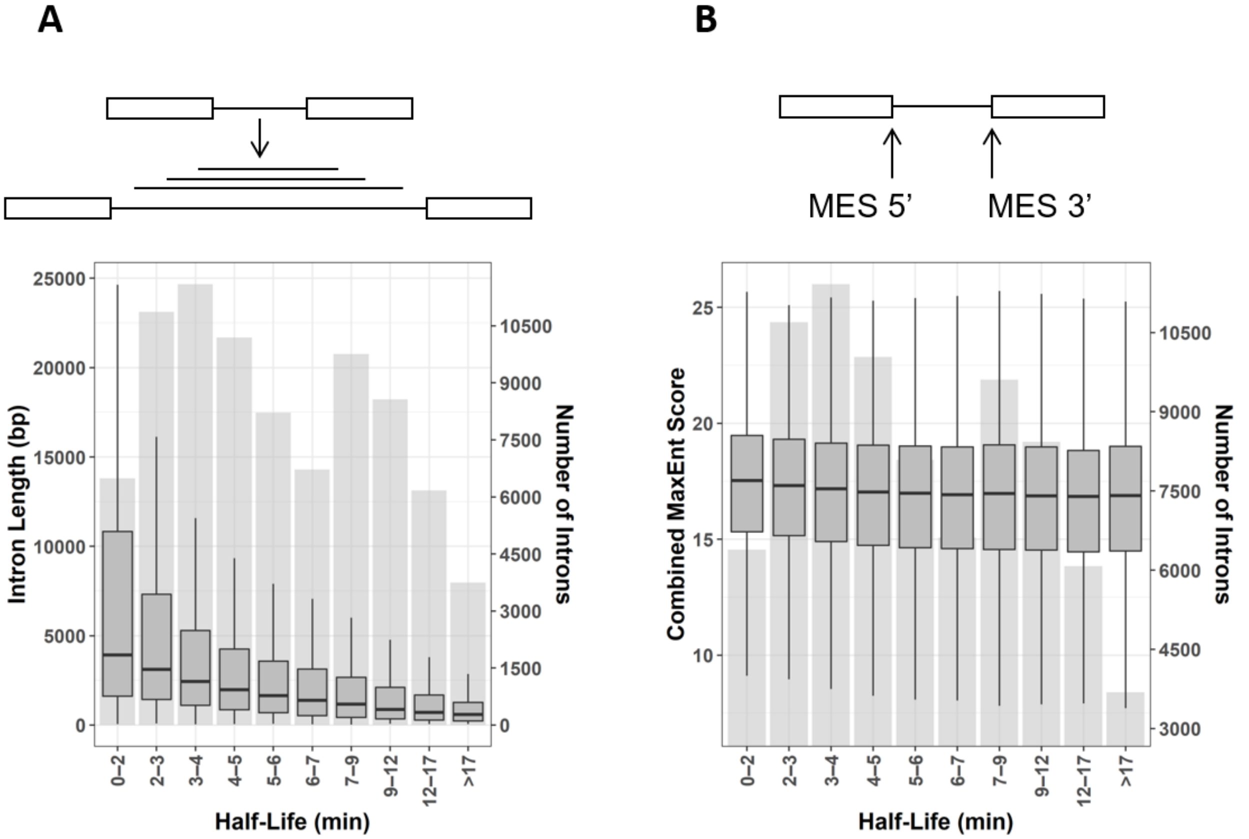


**Figure 2.5. Long introns are removed quickly.** Boxplots depict the relationship in question. Transparent bar plot represents the number of introns within each half-life bin. A) Intron length in bp is considered in relation to intron half-life (min). B) The combined (5'ss + 3'ss) MaxEnt Score (MES), representing splice site strength, is considered in relation to intron half-life.

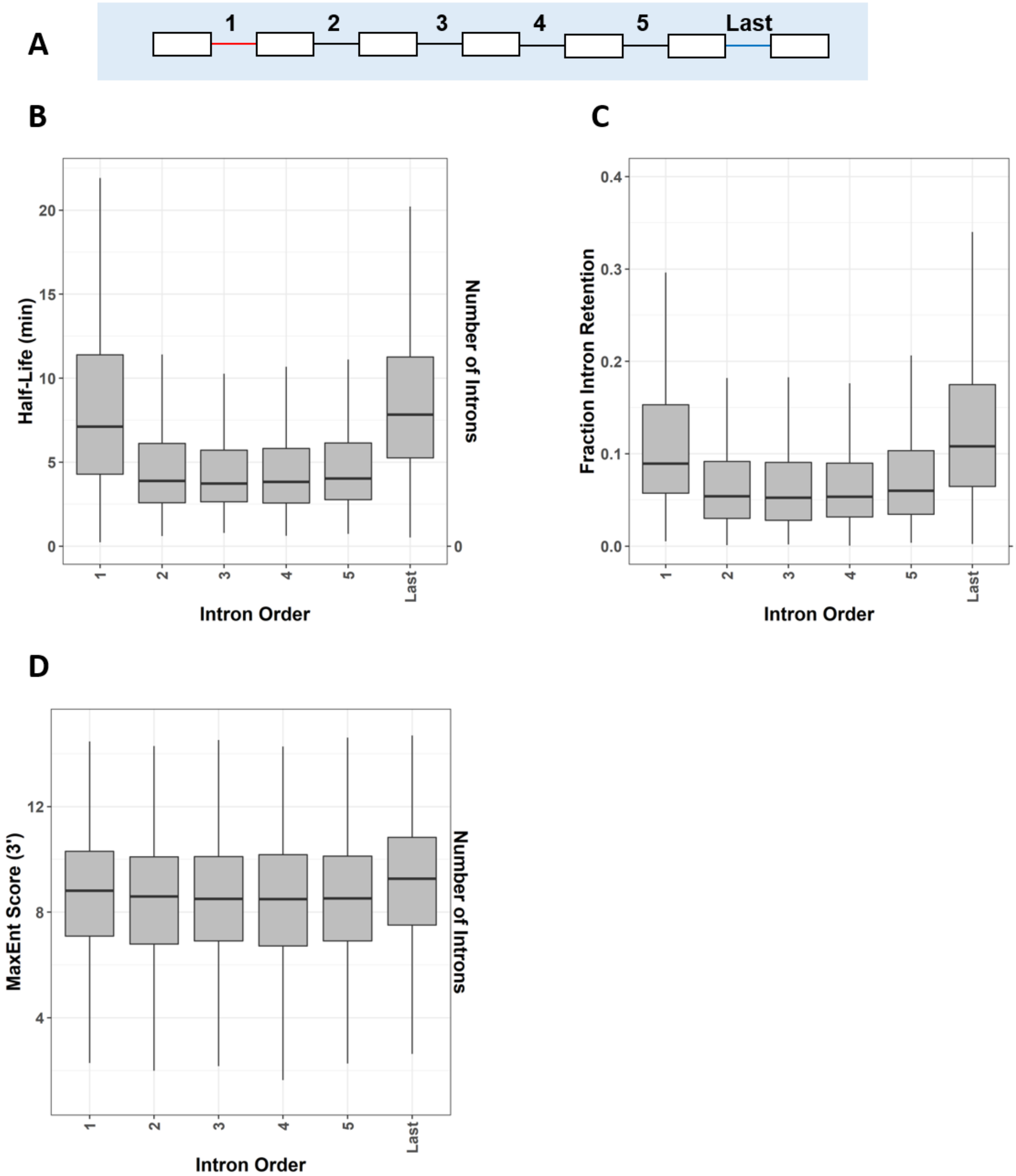


**Figure 2.6. First and last introns have distinct characteristics.** A) Only genes with at least 6 introns were analyzed. B) Half-life (min) based on intron order within a gene. For example. 1 represents the first intron with the gene.  B) Predicted intron retention level based on intron order within a gene. C) 3'ss splice site strength (MaxEnt Score) is shown based on intron order.

last introns display an increase in median intron retention levels compared to internal introns (Figure 2.6B). This difference is likely to reflect nuclear events, and in the case of last introns it may even represent an enriched fraction of introns that are processed uncoupled from transcription.

In *Drosophila*, last introns have been shown to have a stronger 3'ss. To determine if human terminal introns display a similar preference, 5' and 3'ss MaxEnt scores were calculated using MaxEntScan [22] and compared with those of internal introns. While the 5'ss does not display significant differences between terminal and internal introns, the 3'ss strength is notably stronger for last introns, especially when considering that MaxEnt scores are based on a log-scale (Figure 2.6C). These data show that longer introns display faster kinetics, yet first and last introns are distinctly prone to slower splicing, and consequently higher retention levels.

### ***Alternatively skipped exons***

Exon skipping is the most common form of alternative splicing. In this case, the entire exon becomes part of a larger intron and is removed from the pre-mRNA transcript. While various levels of exon inclusion are observed within the transcriptome, exons tend to be mainly included or mainly excluded. Thus, intermediate levels of exon inclusion are less abundant. If an exon is alternatively excluded, by definition it will behave like an intron. Therefore, alternatively excluded exons can be identified if their representation to their flanking exons decreases over the period of the time course (Figure 2.7). Exons that fit to the kinetic model of consecutive intermediates, the rate description described above to determine the processing of intron intermediates, are exons that are excluded from the

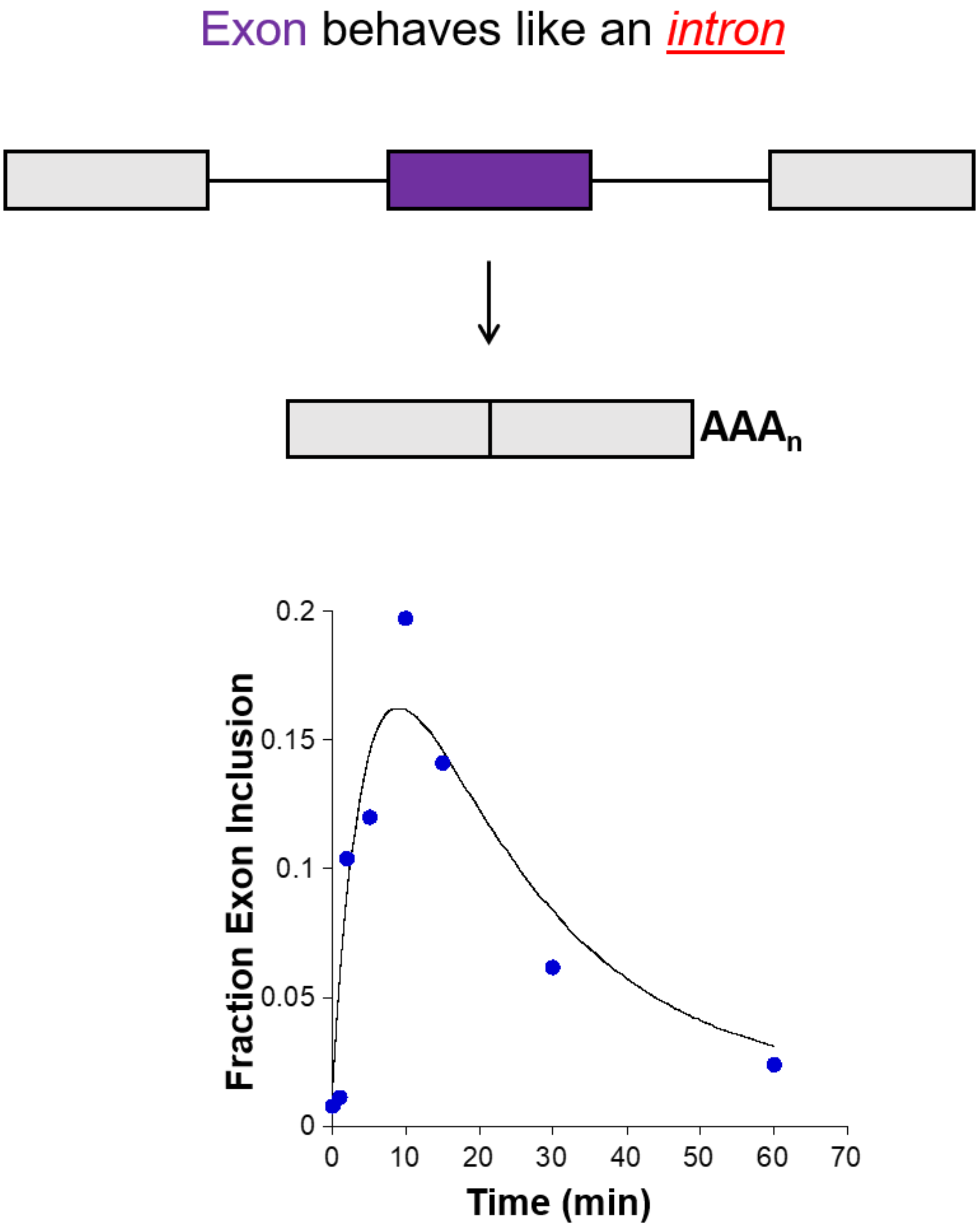
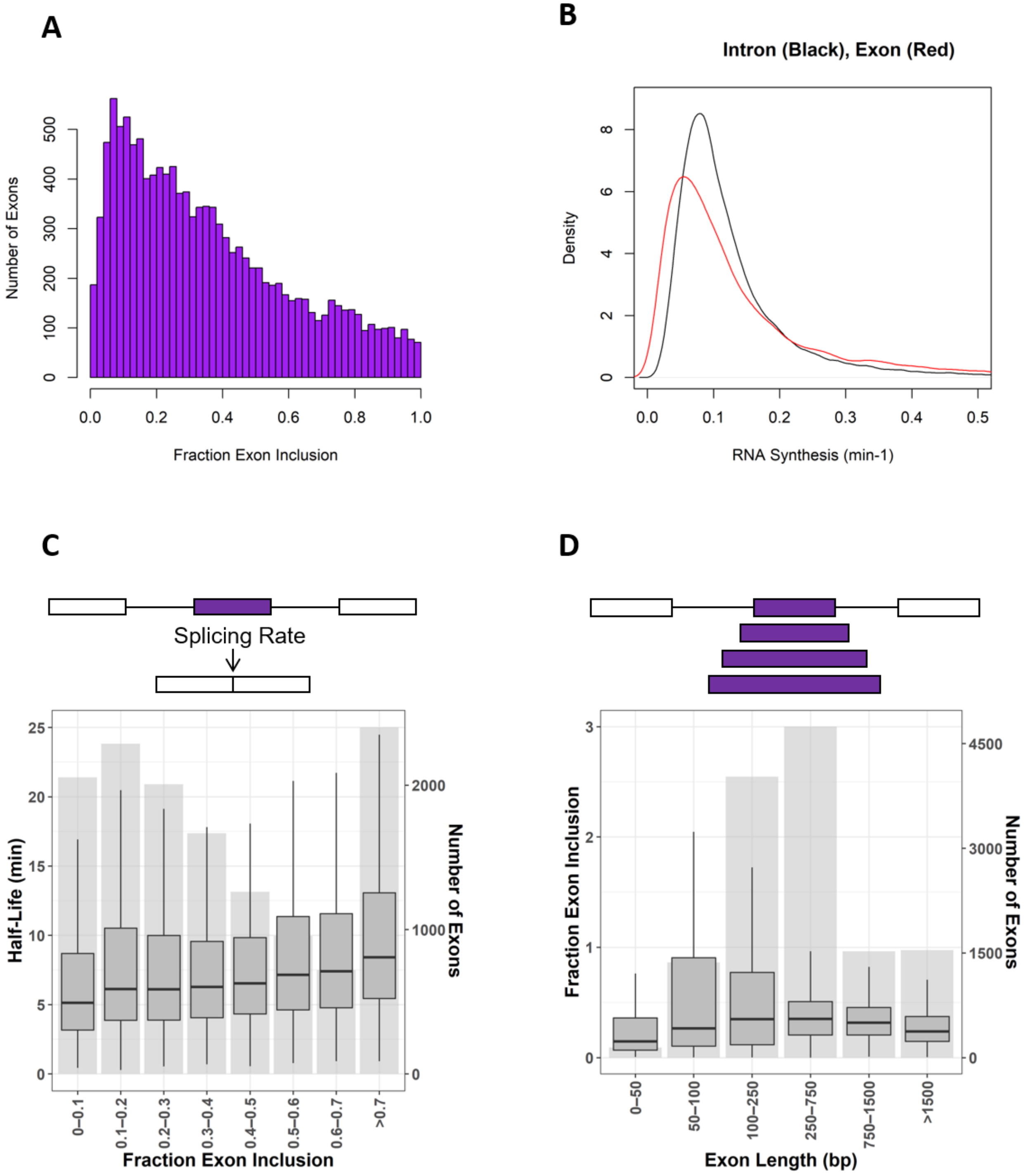


**Figure 2.7. Alternative exon inclusion is captured by consecutive intermediates kinetics.** Exons that fit well to the consecutive intermediates model act like introns. These exons are more prone to being alternatively spliced.

mature transcript. Of the 239,536 expressed exons, 5.6% fit to the intron-like profile of a skipped exon. The exon inclusion distribution of these “weak” exons shows that the majority have an inclusion level of 60% or less (Figure 2.8A). These data demonstrate that a simple analysis of kinetic behaviors between flanking exons can be used to identify exons that are alternatively spliced. Furthermore, the kinetic approach predicts a steady-state exon inclusion level that is likely to reflect cytoplasmic exon representation.

To investigate plausible connections between alternative exon exclusion and splicing kinetics or gene architecture, exon exclusion levels were correlated with half-life measurements or exon size distributions. Low exon inclusion levels are supported by faster removal rates (Figure 2.8C). This observed rate represents the dynamics of the competing splicing pathway that is using a downstream splice site. The faster this rate is when compared to the removal kinetics of the upstream or downstream flanking exons, the more likely it is that the exon in question is excluded from the final mRNA isoform. Therefore, faster exon removal kinetics imply higher exon skipping frequencies. The exon definition model suggests that very short exons (<50 nts) and very long exons (>>250 nt) may be more prone to exon skipping. This is because the assembly of spliceosomal factors across the exon is most efficient across an optimal size. A prediction of this model is that exons that fall outside of this optimal size range are less efficiently recognized and should be more excluded. Consistent with this model prediction, exons that fall within the optimal length range of 100-250 bp display higher inclusion levels (Figure 2.8D). Smaller exons are less efficiently recognized and removed more frequently. We conclude that exon size and competing splice site pairing kinetics determine exon inclusion levels.

**Figure 2.8. Exons prone to being skipped fit the model.** A) Histogram of the distribution of exon inclusion levels of introns that fit the model. B) Density plot of RNA synthesis (min$^{-1}$) modeled introns (black) and exons (red). C) Box plot depicts the relationship between fraction exon inclusion distribution relationship to half-life (min). Transparent bar plot represents the number of exons within each inclusion bin. D) Boxplot shows exon length (bp) relationship with exon inclusion. Bar plot shows number of exons per length bin.


A
Number of Exons
0
100
200
300
400
500
0.0
0.2
0.4
0.6
0.8
1.0
Fraction Exon Inclusion
B
Intron (Black), Exon (Red)
Density
0
2
4
6
8
0.0
0.1
0.2
0.3
0.4
0.5
RNA Synthesis (min-1)
C
Splicing Rate
Half-Life (min)
0
5
10
15
20
25
Number of Exons
0
1000
2000
0–0.1
0.1–0.2
0.2–0.3
0.3–0.4
0.4–0.5
0.5–0.6
0.6–0.7
>0.7
Fraction Exon Inclusion
D
Fraction Exon Inclusion
0
1
2
3
Number of Exons
0
1500
3000
4500
0–50
50–100
100–250
250–750
750–1500
>1500
Exon Length (bp)

### *Comparison of intron and exon synthesis rates*

Recent Pol II Chip analyses have suggested that the Pol II density across exons is different from the Pol II density across intron [77]. These observations suggest that Pol II displays variable processivities depending on whether intron or exon segments of a gene are transcribed. The data further suggested the presence of Pol II pile-ups as it approaches exon junctions [78]. While the metabolic labeling approach described here does not directly measure the elongation rate of Pol II, the observed synthesis rate is a reflection of Pol II density. Consistent with the notion that exon and intron generation is characterized by different Pol II densities, the synthesis measurements of exons are notably slower than that of the introns (Figure 2.8B). These observations are in agreement with the model of variable Pol II processivity, which is dictated by the process of splicing. Thus, the co-transcriptional nature of pre-mRNA splicing is likely a significant contributor to Pol II processivity.

### *The order of intron removal*

The data collectively permits the evaluation of intron excision orders for many genes. While not all individual splicing rates are available for all genes, precise predictions can be made for 5,116 genes. For example, the gene *LAPTM4A* contains 7 exons and 6 introns. All introns fit the consecutive intermediate rate description to allow full kinetic profiling. In addition, one of the exons, exon 5, displays intron behavior kinetics, implying it is an alternatively excluded exon. Steady-state projections suggest low exon 5 inclusion at 34%. Based on the reasonable assumption that transcription proceeds at an average speed of 50 nt/sec, the entire gene transcription is completed within approximately 6 min (Figure 2.9). Comparing the half-life splicing rates for each annotated intron (Table 2.1), excluding the larger intron that would be generated through exon 5 skipping, intron 2 would be expected to be removed first before the first intron is removed. The terminal intron would be expected to be removed before introns 3, 4, and 5. Based on these considerations, some intron excision will be completed after completion of transcription and presumably polyadenylation. Thus, while initial spliceosome assembly is likely to occur co-transcriptionally for most splice sites, completion of several intron removal events is expected to be uncoupled from transcription.

Introducing the alternative skipping of exon 5 rearranges the order of intron removal. Interestingly, the intron defined between exon 4 and 6, thus containing exon 5, displays the fastest removal kinetics. In the modeling, this exon 5 skipping intron is expected to be removed first within the transcript and likely completed before polyadenylation and transcription termination. With the loss of intron 4 and 5, intron 3 will take the place of the last intron to be removed for the exon 5 skipping mRNA isoform.

An additional test to determine the accuracy of the measured kinetics and projected exon retention levels is to evaluate the kinetic ratio of the alternative splicing pathways. Exon 5 inclusion is designated by the faster of its flanking intron removal events, here intron 5 with a half-life of 6.1 min. This rate is 6 times slower than the competing skipping event, which boasts a half-life of 1.1 min. Simple rate ratio calculation then predict that the inclusion level of exon 5 should be around 20% (1.1/6.1), in remarkable agreement with the projected exon 5 retention level of 34% (Table 2.1). The examples discussed here demonstrate the utility of comparing rate profiles across an entire gene. Interestingly, the

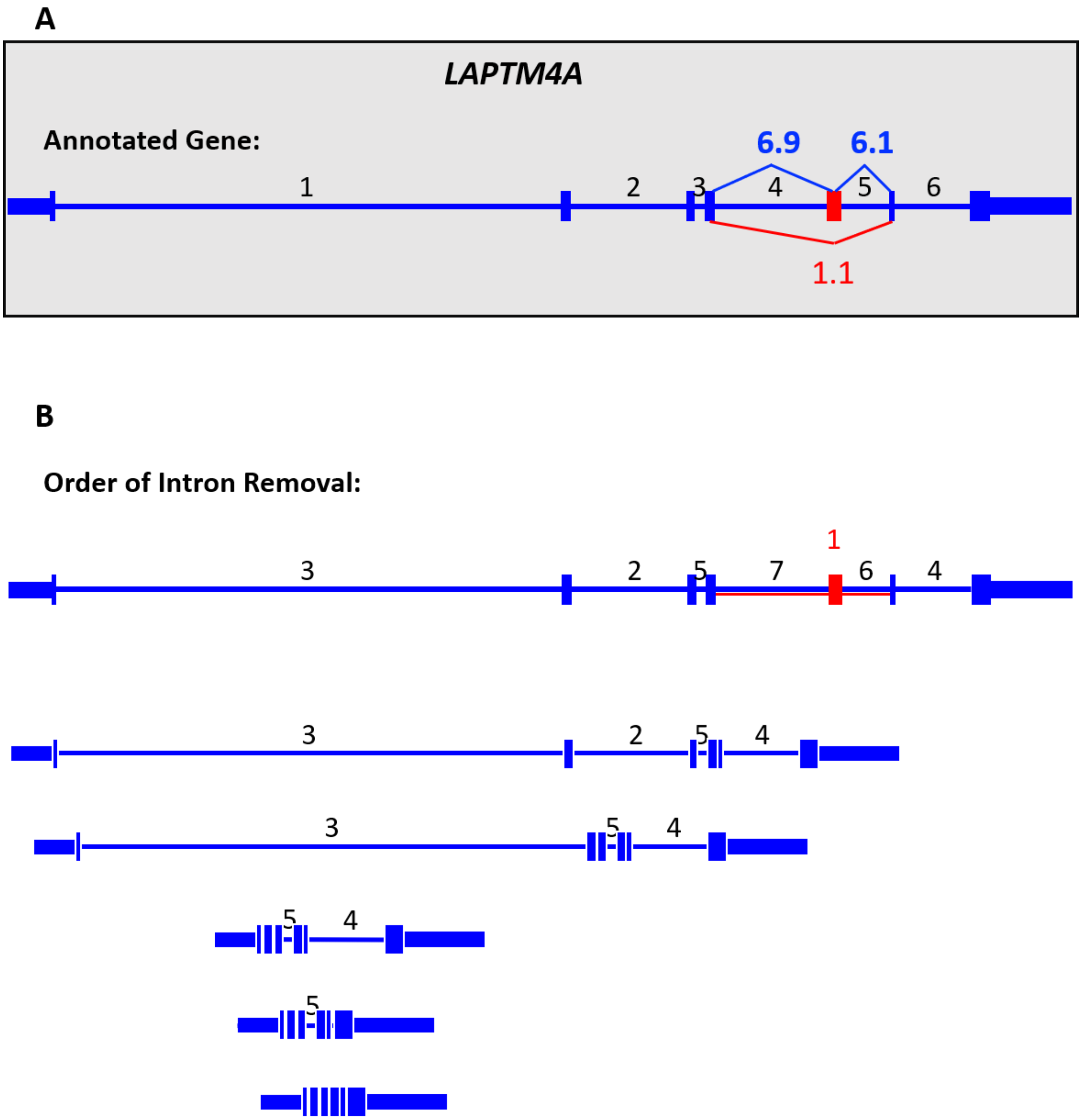
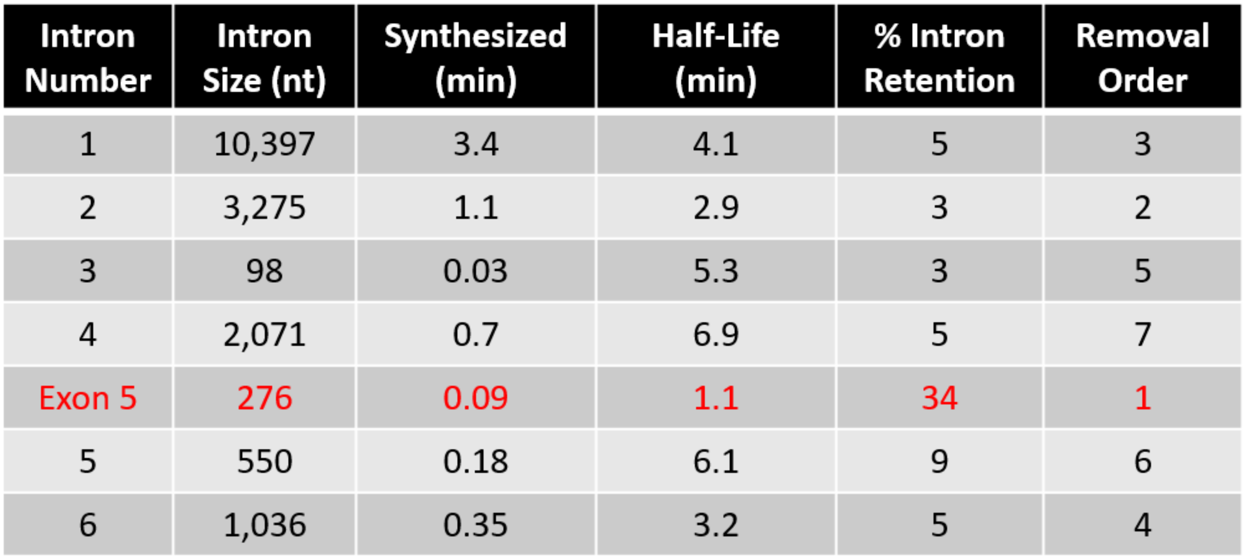


**Figure 2.9. Order of intron removal in *LAPTM4A*.** Annotated gene structure of *LAPTM4A*. Black numbers represent the intron order as annotated. Blue boxes depict exons. Thick blue lines depict introns. A) Numbers in blue are half-lives for introns 4 and 5. Numbers in red are the half-life of exon 5. B) Gene structure of *LAPTM4A* showing order of intron removal as determined by processing kinetics (see Table 2.1). Numbers in black depict order of removal for introns. Numbers in red represent exon 5.

**Table 2.1. Determination of the order of intron removal for *LAPTM4A*.** Intron number refers to the annotated position of the introns, and an exon that displays alternative splicing. Synthesized (min) is the time required to complete transcription of the intron or exon based on the intron size (nt) and an assumed transcription rate of 50 nt/second. Half-life and % intron retention are determined by kinetic modeling to consecutive intermediate model. This information was used to determine the removal order (see Figure 2.9).



| Intron Number | Intron Size (nt) | Synthesized (min) | Half-Life (min) | % Intron Retention | Removal Order |
|---|---|---|---|---|---|
| 1 | 10,397 | 3.4 | 4.1 | 5 | 3 |
| 2 | 3,275 | 1.1 | 2.9 | 3 | 2 |
| 3 | 98 | 0.03 | 5.3 | 3 | 5 |
| 4 | 2,071 | 0.7 | 6.9 | 5 | 7 |
| Exon 5 | 276 | 0.09 | 1.1 | 34 | 1 |
| 5 | 550 | 0.18 | 6.1 | 9 | 6 |
| 6 | 1,036 | 0.35 | 3.2 | 5 | 4 |

event of alternative splicing can influence the order of intron excision, which could affect other nearby splicing events, other RNA processing events or mRNA stability [79] . Deriving additional examples of the order of intron/exon removal events will allow the derivation of detailed gene RNA processing maps.

## Discussion

In this chapter we have presented an unparalleled analysis of human RNA processing dynamics. Previous studies have been limited by time resolution and depth, restricting the analysis of processing kinetics to a small number of genes. The co-transcriptional nature of splicing is well appreciated in higher eukaryotes, especially in humans. The timing of transcription has been linked to changes in alternative splicing, but the influence of the rate of splicing has been largely under-studied [80]. Through the attachment of multiple fluorophores to GFP reporter genes, real-time visualization of intron removal in yeast and *Drosophila* cells raised the question of whether kinetics distinguish constitutive from regulated splicing [81]. Still, genome-wide studies in human are required to fully answer this question. Here, we find that the kinetics of splicing itself have a large influence on inclusion levels, whereby splicing kinetics are a hallmark of alternative splicing in introns.

The most striking finding is contrary to the notion that shorter introns are more efficiently spliced by intron-definition than longer introns are by exon-definition [17]. In our analysis, longer introns are spliced faster, with shorter introns displaying higher retention levels. This finding has been corroborated by smaller-scale analysis in human dendritic cells after LPS stimulation and in *Drosophila* [72, 82]. This suggests that a switch to the exon-definition mode of splice site recognition, which likely requires additional help to explain such efficient removal of large human introns. One possible method of facilitating efficient splicing is recursive splicing, where long introns (>10kb) are removed piecewise such that two or more adjacent intron sections are excised as discrete splicing reactions, each producing their own lariat [83, 84]. Over 342 RS-sites (recursive splicing sites) have been found to be utilized co-transcriptionally as a supplement to canonical splicing to enhance the efficiency of long intron removal in human cells [85].

Still, this does not explain why first introns, which are known to be long and more conserved, are spliced more slowly and have higher retention rates [86]. Conserved sequences in the first intron are correlated with higher enrichment for several chromatin marks, indicative of active regulatory regions compared to other conserved intronic sequences within the same gene [87]. Together, this points to a balance between maintaining active transcription of a gene and preventing splicing from being a rate limiting step too early.

Our approach also permits the remarkable ability to simultaneously identify alternative splicing events in both introns and exons. The observation that exon exclusion levels correlate with faster removal kinetics coincides with the finding that exon size and competing splice site pairing kinetics dictate inclusion levels. By combining the ability to describe the removal of every intron within a gene, including the identification of cassette exons, we are able to determine the order of intron removal across entire genes. Previously, this was only possible with site directed RT-PCR for a few genes after external stimulus, thus restricting the gene pathways that could be analyzed [68]. In a field flooded with genome-wide data pertaining to regulatory RNA binding proteins, epigenetics, and steady-state RNA-seq, it can often be difficult to narrow down the interplay with splicing events of interest.

The ability to determine the order of intron removal highlights the ability to pinpoint when and how splicing decisions are made. Such RNA processing maps narrow the window researchers may look into to clarify what roles corresponding regulatory factors may play in generating the final protein coding mRNA. Thus, the demonstrated findings provide a profound understanding of the processing kinetics that modulate alternative splicing of human introns and exons globally.

## Methods and Materials

### *Cell Culture*

HepG2 cells were generously gift from Brent Gravely. HepG2 cells were grown in High Glucose DMEM (HyClone, SH30022.01) with 10% FBS at 37°C in 5% $C0_2$.

### *High Throughput Sequencing*

1 µL of 1:100 dilution of ERCC spike-in was added to each sample prior to library preparation. cDNA libraries were prepared without polyA selection using Illumina TruSeq mRNA stranded protocol. 100bp single-end reads were sequenced on the Illumina HiSeq 4000 platform. 3 isogenic replicates were sequenced, but only 2 were available at the time of analysis. Between 18-65million uniquely mapped reads per time point, per replicate, were analyzed.

### *Computational Analysis*

Reads were aligned with STAR aligner using the 2-pass mode, to UCSC annotation known gene hg38 Gencode V24. A merged annotation was created to generate exon coordinates that encompass all alternative exons, leaving pure introns. The coverage of uniquely mapped reads was used to assign reads and portions of reads to the respective exons and introns defined in the merged annotation using TPMCalculator [88]. Canonical genes and pseudo genes were included.  Only genes with an average read count of 10 across the gene were used. Transcripts per Million (TPM) was calculated:

*(Mapped feature reads /Feature Length in kb) = Feature Reads Per Kilobase (RPK)*

$$\frac{\sum \textit{(all sample RPKs)}}{\textit{1,000,000}}$$
$$\textit{= Scaling Factor}$$

*TPM = Feature RPK/Scaling Factor*

TPM values were used to calculate the fraction of inclusion for each intron by dividing the TPM of an intron, divided by the average TPM of its flanking exons. This was calculated for a given intron for each time point. These values were subjected to a non-linear regression analysis using a least-squares method that fit the data points to predict the rates of intron generation (synthesis), intron removal (splicing), and intron retention (steady-state).

Extensive filtering was performed to ensure only those with high quality fits to the model were characterized. First, exons or introns that resulted in negative synthesis, splicing, and retention estimates were removed as these had high standard error values and were products of low sequencing coverage, indicating a poor fit to the model. Next, we kept the top 90% based on combined splice site scores. This effectively removed abnormally low scores that represented false splice sites that were artifacts of using a merged exon definition.  Junction read alignments corroborated that these were not true splice junctions.

# CHAPTER 3

# Reduction of Splicing Factor SRSF1 Increases Alternative Splicing Through the Modulation of RNA Processing Kinetics

## Summary

Alternative splicing expands genomic diversity through the generation of multiple mRNA isoforms and thereby, protein isoforms. Two such types of alternative splicing are intron retention and exon skipping (cassette exons). The splicing factor SR-protein SRSF1 is known to regulate these specific types of alternative splicing, typically by strengthening the recognition of introns and exons that are otherwise prone to alternative splicing. While the detection of individual alternative events is easily studied, it is unclear how the rates of alternative pre-mRNA processing are influenced by such a potent splicing regulator. 4sU metabolic labeling of human cells was used to determine rates of RNA processing steps in cells depleted of SRSF1. High-throughput sequencing of isogenic replicates across high-resolution time series demonstrates that SRSF1 is responsible for regulating retention levels of introns and exons susceptible to alternative splicing. This is primarily achieved through changes in splicing rates and Pol II density, with a strong dependence on optimal feature length constraints. eCLIP data further shows that SRSF1 binds preferentially to weaker exons that are prone to being skipped.

## Introduction

Alternative splicing is used in as much as 95% of multi-exon genes to expand the human proteome [24]. Two types of alternative splicing are exon skipping and intron retention. Intron retention (IR) describes the inclusion of an intronic sequence in the mature RNA, a process that has the potential to expand the number of protein isoforms by 5,044-12,612, as supported by cDNA and expressed sequence tag (EST) data [89]. While intron retention occurs only 2-5% of the time, it plays pivotal roles in fine tuning protein production in dendritic cells and neurons, and in genes involved in the immune response [90-92]. Recent genomic studies in over 2,573 human tissue samples show that as many as 80% of coding genes can be affected by IR. Interestingly, retained introns are enriched for containing putative RNA binding protein sites, suggesting that intronic binding of RNA binding proteins modulate the level of intron retention [93].

Exon skipping, or cassette exon, is the most prevalent type of alternative splicing, accounting for 50-60% of alternative splicing events [15]. It describes a splicing event where an internal exon is skipped, thereby being removed as part of the intron. Internal exons greater than 300 bp are more prone to being skipped [94, 95]. Changes in cassette exon splicing are associated with human disease, such as renal cancer or Duchenne Muscular Dystrophy [96, 97]. In concordance with the premise that the majority of splicing occurs co-transcriptionally, the decision to include or skip an exon is frequently made during transcription [71].

Introns and exons that are prone to alternative retention and skipping are often characterized by weak splice sites, abnormal length constraints and the presence of binding sites for splicing regulators [93, 98]. One of the best studied of these RNA binding regulators is SRSF1. It has been described as a splicing factor with roles in transcription, splicing, translation, and cancer progression [32]. SRSF1 is best known for its ability as a splicing activator that binds to exonic splicing enhancer (ESE) sequences, promoting the retention of exons that are otherwise deemed too weak to be recognized by the spliceosome. Even though most SR proteins are associated with splicing activation, they have been shown to be capable of repressing splicing in a position-dependent manner [26].

The regulation of intron retention and exon skipping events by RNA binding proteins has been extensively studied using *in vitro* and genome-wide approaches [99, 100]. RNA binding proteins are recruited co-transcriptionally, in line with the proposed timing when alternative splicing decisions are made [101]. However, the global effects of SRSF1 on splicing kinetics and exon/intron inclusion levels are unknown. Given the finding that exon/intron length and their genomic position coincide with kinetic profiles specific to alternative splicing events (see Chapter 2), we aimed to test the hypothesis that the reduction of the splicing factor SRSF1 triggers changes in splicing rates that affect exons and introns with suboptimal features for splice site recognition. To test this hypothesis, lentiviral knockdown of *SRSF1* was paired with 4sU-seq experiments utilizing the same high-resolution time point series as described in Chapter 2. A 63% reduction of SRSF1 protein levels was determined by Western blot analysis (Figure 3.1). Replicates showed remarkable reproducibility and associated by condition in PCA analyses (Figure 3.2 A, B). Comparison of RNA processing dynamics and their respective intron/exon architecture reveals that specific kinetic profiles are dependent on feature length.

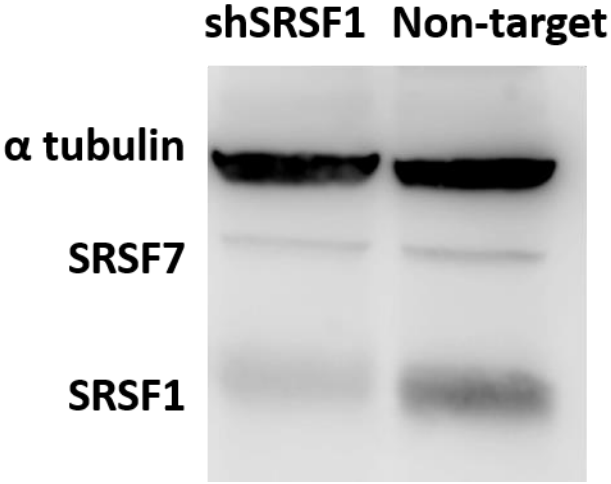
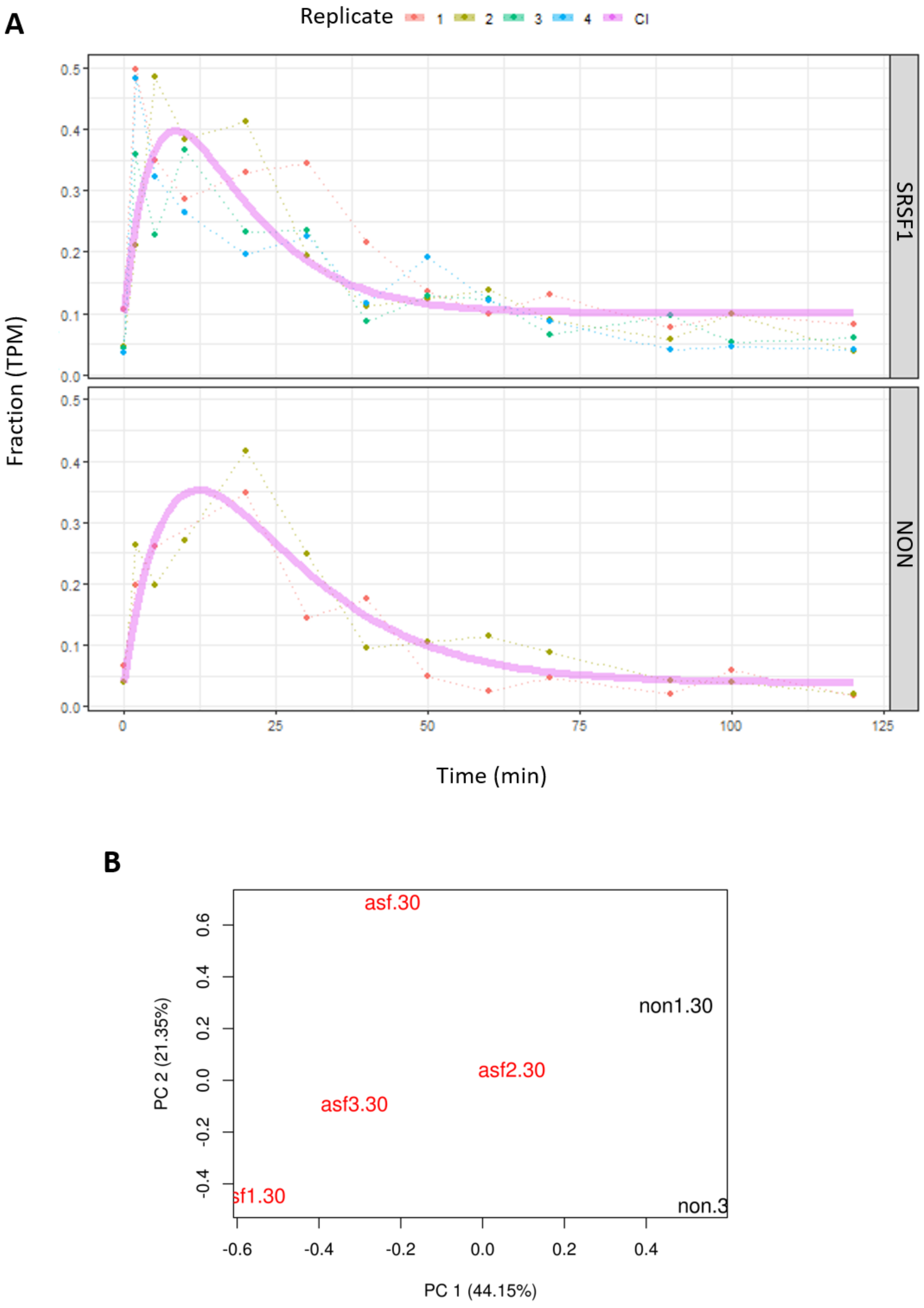


**Figure 3.1. Western blot analysis demonstrates SRSF1 knockdown.** α tubulin is used as a loading control. Antibody probing SRSF7 is used as a control to demonstrate no off-target effects.

**Figure 3.2. Isogenic sample replicates.** A) Intron from *ABCC1* shown as example of replicate data plotted and fit to the consecutive intermediate kinetic model. Fraction of intron retention relative to flanking exons are plotted in Transcripts Per Million (TPM) mapped reads. Timepoints are shown in minutes. Red dotted line shows replicate 1, light green dotted line shows replicate 2, green dotted line shows replicate 3, blue dotted line shows replicate 4, and pink solid line shows the model fit through the data points. B) PCA plot of replicates. SRSF1 replicates are in red. NON replicates are shown in black.


A
Replicate
1
2
3
4
CI
Fraction (TPM)
0.0
0.1
0.2
0.3
0.4
0.5
SRSF1
NON
0
25
50
75
100
125
Time (min)
B
asf.30
non1.30
asf2.30
asf3.30
sf1.30
non.3
PC 2 (21.35%)
-0.4
-0.2
0.0
0.2
0.4
0.6
-0.6
PC 1 (44.15%)

## Results

### *Reduction of SRSF1 leads to increased intron retention and reduced Pol II density*

SRSF1 is best known as a splicing activator that promotes the removal of weak introns. To test whether the loss of SRSF1 affects alternative splicing of introns through changes in splicing rates, a comparative analysis was carried out using 80,840 introns that had a good fit to the consecutive intermediate kinetic model in both the control group (NON) and the SRSF1 knockdown group (SRSF1) (see Materials and Methods). Upon reducing SRSF1 levels, introns display a redistribution into higher retention levels (0.2+) when compared to the control group. Conversely, fewer introns in SRSF1 knockdown conditions have low intron retention levels of 0-0.05 (Figure 3.3A). Consistent with the results described in Chapter 2, introns with slower splicing kinetics tend to be retained more often in both the control and the SRSF1 knockdown condition. However, a direct comparison of the experimental conditions within the high retention bins demonstrates that more introns are retained upon SRSF1 knockdown despite faster intron removal kinetics (Figure 3.3A, bin >0.4). These observations suggest that intron retention levels are not solely dictated by observed splicing kinetics. Presumably, the loss of SRSF1 influences additional aspects of the splicing reaction that results in elevated intron retention levels.

Pol II density and the resulting effects on transcription elongation have been shown to influence splice site selection [102]. Hyperphosphorylated SRSF1 moves from nuclear speckles to active transcription sites to promote splicing [32]. As described in Chapter 2 the measured RNA synthesis rates reflect an approximation of Pol II density on a gene segment. A comparison of the RNA synthesis rate between the control and the SRSF1 knockdown groups demonstrates that SRSF1 reduction results in decreased RNA synthesis rates across

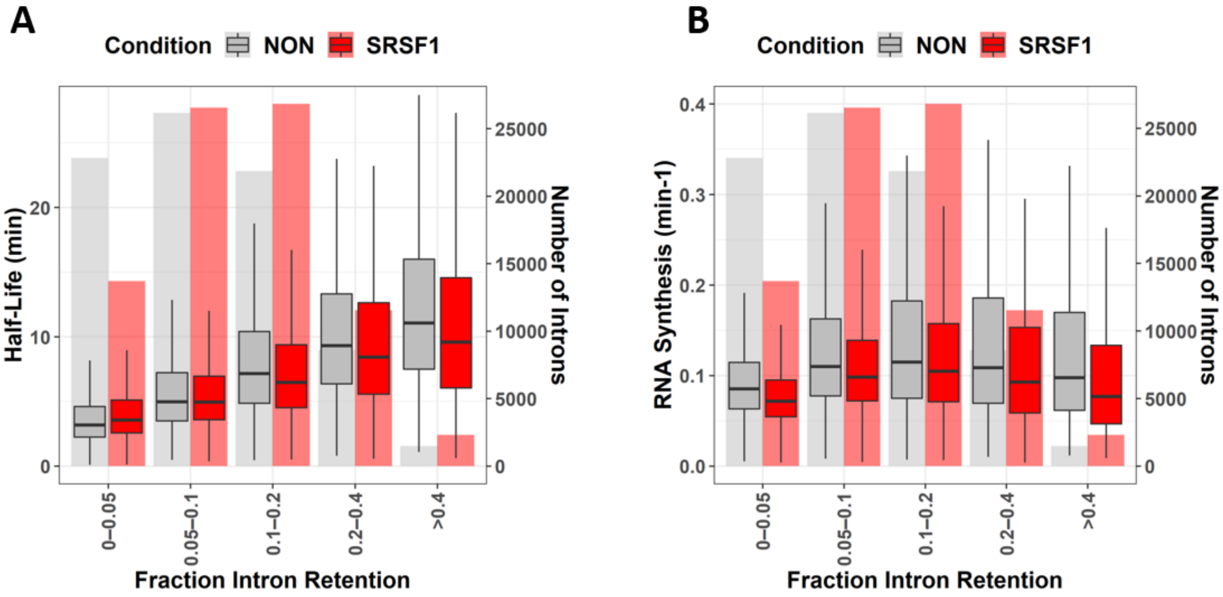


**Figure 3.3. Intron retention is increased by SRSF1 knockdown.** Each boxplot depicts the distribution of the relationship analyzed. The bar plot in the background of each boxplot represents the number of introns in each bin. Red is SRSF1 knockdown, grey is non-target control (NON). A) Distribution of "Fraction of Intron Retention" relative to intron removal half-lives (min). B) Distribution of "Fraction of Intron Retention" relative to "RNA-Synthesis (min$^{-1}$)".

all retention levels evaluated (Figure 3.3B). Although introns with either higher retention levels (>0.2) appear to have a lower Pol II density readout, it is unclear whether this is significantly greater than other retention bins. Given the lack of clear correlation between the RNA synthesis rates and intron retention levels (Chapter 2, Figure 2.4B), it is currently unclear what functional consequence the increased rate of RNA synthesis, a proxy for Pol II density, has on intron splicing that is influenced by SRSF1. Regardless, the results suggest that SRSF1 knockdown directly or indirectly influences Pol II transcription efficiency.

### ***SRSF1 modulation of intron removal is constrained by intron length***

Length constraints dictate mechanisms of intron removal by creating a switch of spliceosome assembly across an intron when the distance is under 300 nts (intron definition) to assembly across an exon when introns are greater than 300 nts [7, 103]. SR proteins aid in the recruitment of core spliceosome components, thus promoting splice site recognition of its target [38]. Thus, it is possible that intron splicing kinetics may be altered by the depletion of SRSF1 in an intron length-dependent manner. Indeed, the ability to remove larger introns quickly appears diminished in SRSF1 depleted cells. The median length of introns with splicing half-lives between 0-2 min is 1,514 bp for SRSF1 knockdown conditions and 3,918 bp for the control condition, with similar ratios between 2-3 min half-life bin (Figure 3.4A).

To analyze the effects of SRSF1 knockdown at the intron definition/exon definition size transition, a distribution of intron length for every 25 bp between 50 and 500 bp was created. In agreement with spliceosome constraints, SRSF1 knockdown reduces the spliceosome's ability to promote efficient removal of introns of length 50-300 bp. The

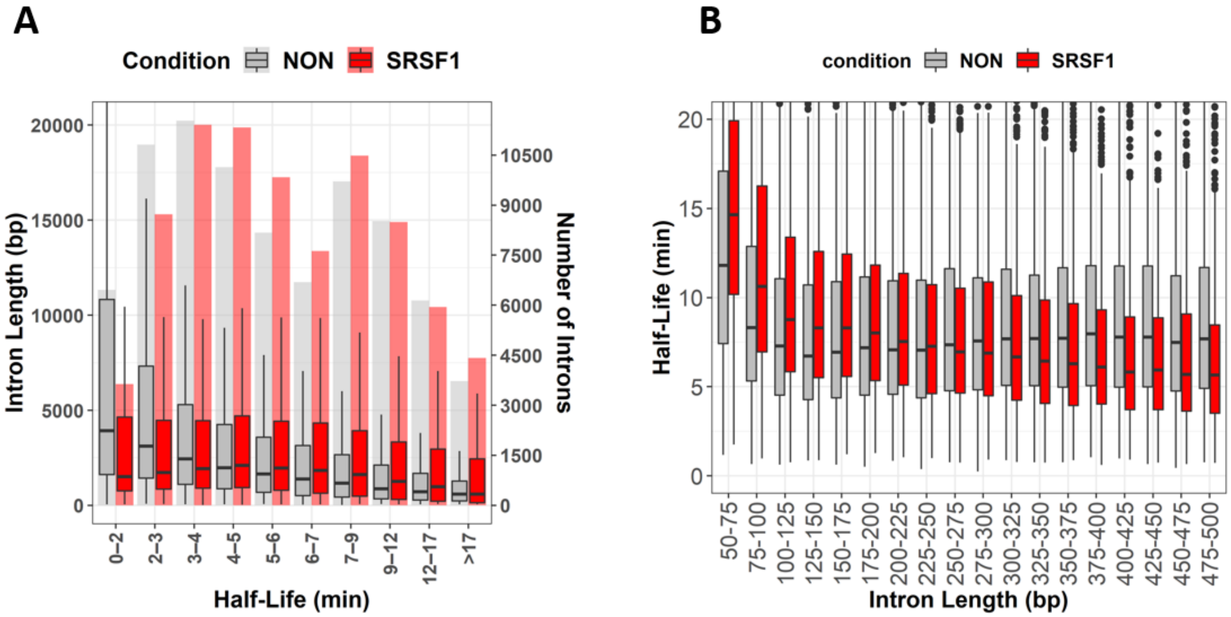


**Figure 3.4. SRSF1 promotes efficient removal of long internal introns.** Non-target (NON) dynamics in gray, and SRSF1 knockdown dynamics in red. Box plots depict the distributions of the relationships being assessed. Transparent box plots represent the number of introns analyzed for each bin. A) Intron length (bp) is plotted in relation to the distribution of half-lives of introns common in NON and SRSF1. B) Intron Length (bp) distribution from 50-500 bp in relation to half-lives (min). Bins are every 25 bp.

differences between control and knockdown splicing rates are statistically significant for each 25 bp increment within this size range using a paired Wilcoxon signed rank test (p-values range from 2.2e-16 to 0.044 in ascending length order) (Figure 3.4B). The majority of human introns are longer than 500 bp.  Additional distribution analyses with 50 bp resolution up to 5,000 bp intron length demonstrated that SRSF1 reduction increases splicing rates between 300-1400, suggesting that in wild-type conditions SRSF1 may purposefully reduce the efficiency of splicing of introns of this length (data not shown).  No difference was detected in half-lives of introns between 1,400 and 2,000 bps in length. Half-lives of introns >2,000 bp corroborate the overall loss of ability for SRSF1 knockdown cells to efficiently remove long introns as seen in Figure 3.4A.

Splice site score correlations were also evaluated. Interestingly, when comparing SRSF1 knockdown with the control group, no detectable splicing kinetics or intron retention differences were observed for introns with variable splice sites (data not shown).

The splicing efficiency of introns located in a more proximal 5' position within the gene may influence the splicing of downstream introns and the analysis described in Chapter 2 demonstrated that more proximal introns are retained at higher levels. To investigate if SRSF1 preferentially regulates the removal of more 5' introns, a correlation between intron position and intron retention was carried out. The analysis demonstrates that SRSF1 knockdown affects introns at all gene locations nearly equally (Figure 3.5A). Thus, SRSF1 knockdown influences intron splicing independent of gene position.

First and last introns are affected differently than internal introns, most likely due to the influence of alternative promoters, alternative terminal exons, and interactions between the splicing and capping or polyadenylation machineries that assemble on terminal exons

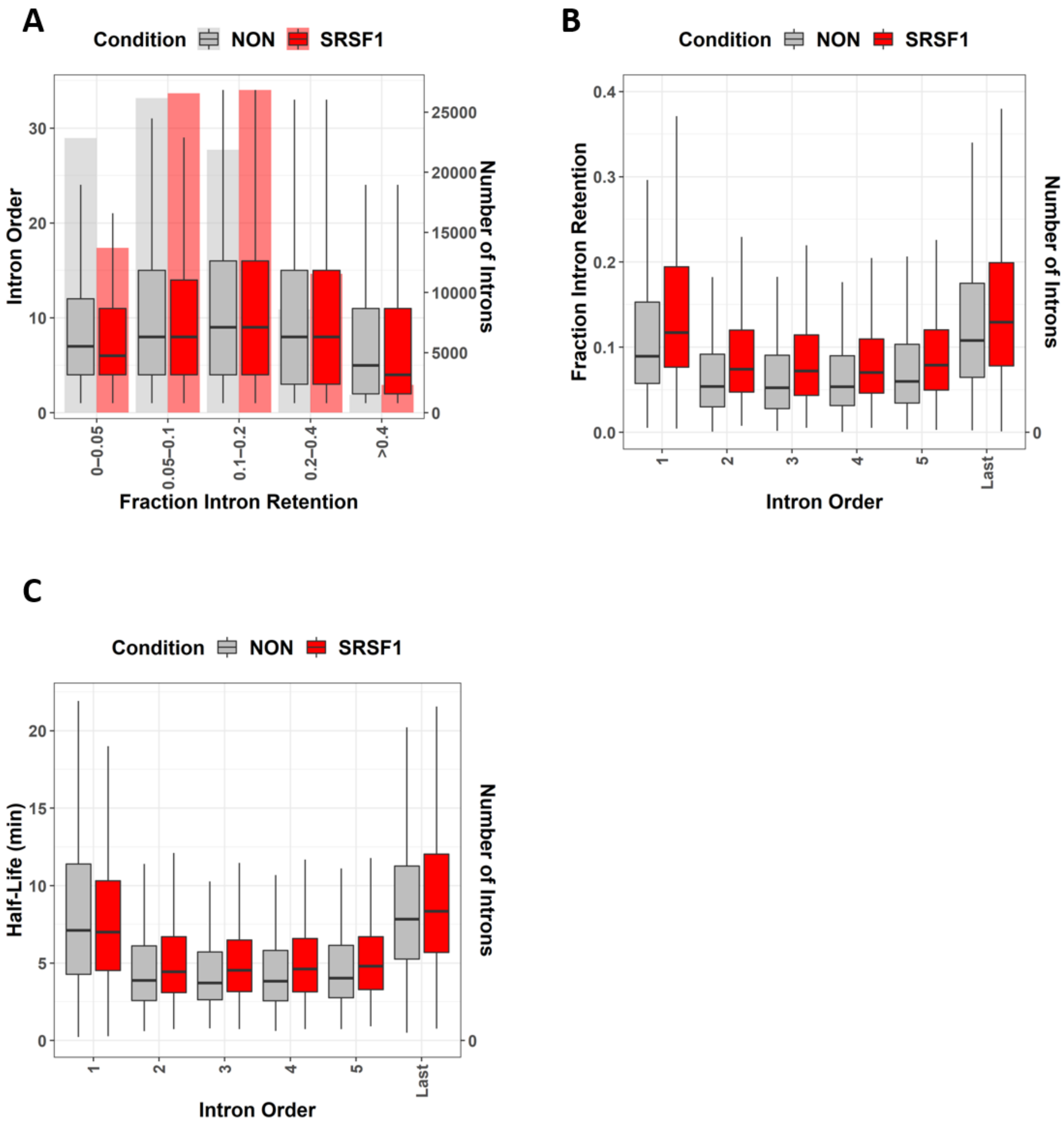


**Figure 3.5. Intron order: internal intron kinetics influenced by SRSF1.** Non-target (NON) dynamics in gray, and SRSF1 knockdown dynamics in red. Box plots depict the distributions of the relationships being assessed. Transparent bar plots represent the number of introns analyzed for each bin. A) Distribution of "Fraction of Intron Retention" levels relative to an intron's order within a gene. B) For genes containing at least 6 introns, the order of the intron within the gene is assessed based on predicted retention levels and C) half-life (min).

[104-106]. To ask the question whether SRSF1 knockdown affects terminal introns to similar levels as internal introns, genes with at least 6 introns were analyzed for changes in splicing rates and retention levels. This analysis permits the comparison of the same number of introns across all conditions. In agreement with our intron retention distribution findings, retention levels were higher regardless of intron position within its gene (Figure 3.5B). However, SRSF1's influence on intron removal rates is limited to the internal introns, with a minor influence on last introns (Figure 3.5C). These observations suggest that for first introns SRSF1's contribution towards efficient removal goes beyond modulating splicing kinetics. It is tempting to speculate that SRSF1 may play an additional important role in mediating interactions between the capping and the splicing machinery. Similar arguments can be made for last intron removal efficiencies.

To investigate the differences in splicing dynamics of introns for which SRSF1 typically maintains efficient splicing, compared to those for which SRSF1 may negatively regulate its removal, introns were grouped based on whether or not their retention levels resulted in a switch from being constitutively removed in one condition and retained in the other. These intron classes (switch introns) were compared to introns that did not have a significant switch in retention levels. A switch cutoff of 0.2 was used based on the intron retention distribution (Figure 2.3C). For the purpose of this analysis, introns with retention <0.2 were considered constitutively spliced, and introns with retention >0.2 were considered retained.  RNA synthesis measures do not significantly influence an intron's tendency toward being alternatively spliced when SRSF1 levels are perturbed (Figure 3.6A). Introns that switch from constitutively spliced in the control group to retained in SRSF1 knockdown condition are much more abundant than introns that switch from retained in

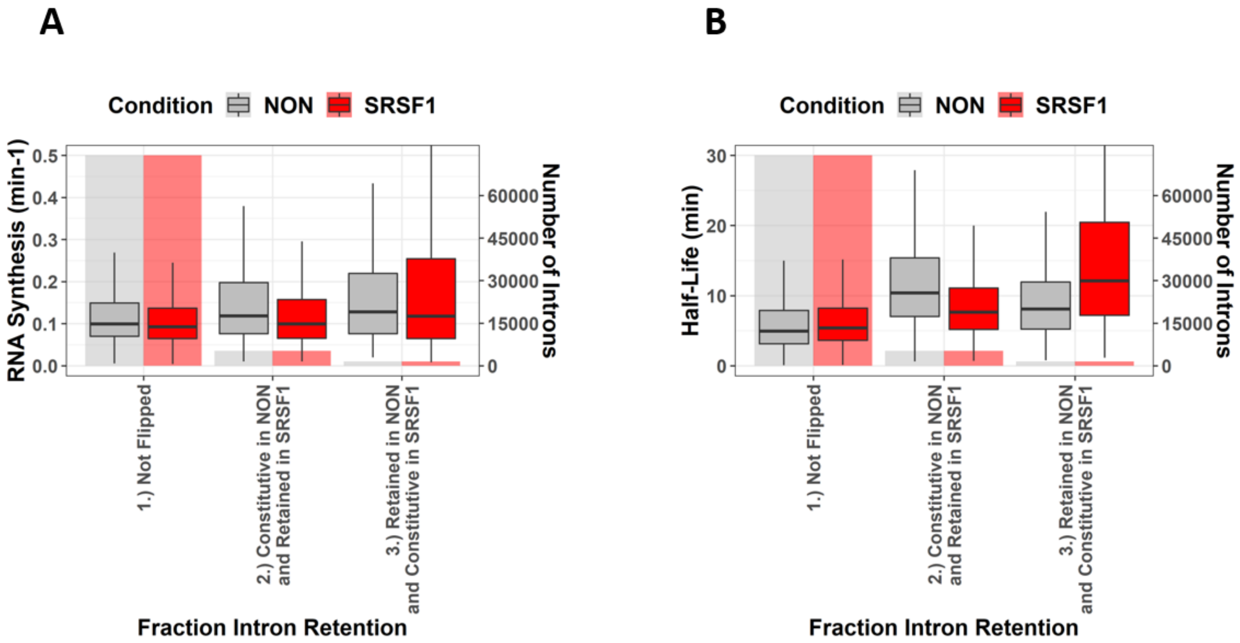


**Figure 3.6. Introns that are differentially spliced upon SRSF1 knockdown.** SRSF1 knockdown (SRSF1) is shown in red and the control group (NON) is in grey. Boxplots depict the relationship. Bar plot shows the number introns in each category. The categories are 1.) Introns that have consistent retention levels between both conditions (Not Flipped). 2.) Introns that have retention <0.2 (constitutive) in NON and >0.2 (retained) in SRSF1 condition. 3.) Introns that have retention >0.2 (retained) in NON and <0.2 (constitutive) in SRSF1 condition. A) RNA Synthesis (min$^{-1}$) for each category. B) Half-life (min) for each category.

NON to constitutive in SRSF1 (Figure 3.6A, B). Interestingly, introns that switch from constitutive to retained in SRSF1 knockdown conditions show a faster rate of splicing, despite being more retained (Figure 3.6B). Collectively, these data support the previously held view that SRSF1 primarily acts as an activator of intron removal.

### *Cassette exon inclusion levels under the influence of SRSF1*

SRSF1 is best known as an exonic splicing activator that assists in the recognition of exons. As outlined in Chapter 2, the time course datasets can also be used to evaluate the levels of exon inclusion, thus assisting in the identification of exons that are alternatively spliced (Figure 2.7). The comparison between SRSF1 knockdown and control conditions should therefore identify exons that are directly or indirectly influenced by the activities of SRSF1. Several comparative analyses were carried out to test this notion. A first set of comparisons analyzes relationships between exons that fit the consecutive intermediate rate description in both conditions, control and SRSF1 knockdown. As was argued in Chapter 2, exons that fit this rate description are considered weak exons that are prone to undergo alternative exon skipping. Thus, the first set of analyses investigates SRSF1's influence on weaker exons.

In accordance with this assumption, for the wild-type condition it was argued that exon skipping was promoted by faster removal rates (Chapter 2, Figure 2.8C). Intriguingly, the loss of SRSF1 neutralizes these kinetics and exon exclusion correlation and highly skipped exons in SRSF1 knockdown conditions display slower removal rates (Figure 3.7A). As was argued above, these observations suggest that exon skipping of weak exons induced by the loss of SRSF1 is not primarily driven by absolute changes in the rate of exon skipping.

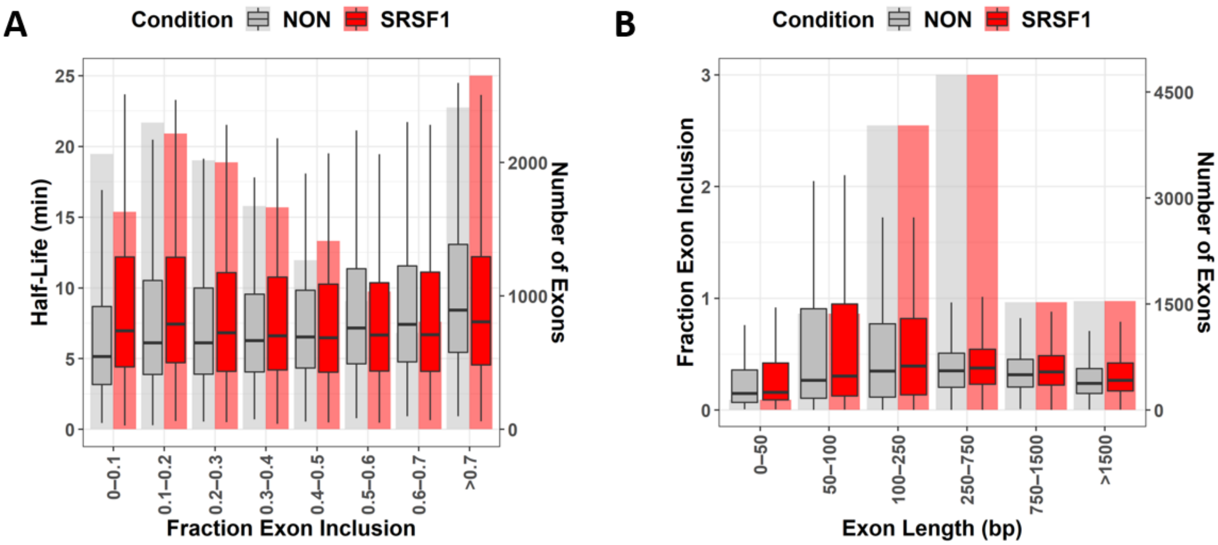


**Figure 3.7. Comparisons of exon inclusion levels based on exon length and processing dynamics.** A) Distribution of exon inclusion levels relative to half-lives. B) Exon length (bp) comparison to exon inclusion levels.

The majority of alternative exon lengths falls within the previously established optimal range of 100-300 bp to support the exon definition mode of splice site recognition. Similar to introns, SRSF1 regulation of exons is promoted by this optimal length, as loss of SRSF1 causes increased exon inclusion levels, primarily for this range (Figure 3.7B).

### ***SRSF1 preferentially binds to exons that are prone to being skipped***

SRSF1 binding to exons was determined using eCLIP data with 2 replicates from the ENCODE project [107]. eCLIP is an immune precipitation method that maps the binding sites of RNA binding proteins (RBPs) on their target RNAs using a modified nucleotide resolution CLIP (iCLIP) protocol [108]. Similar to ChIP-seq, peaks representing binding of the precipitated protein are normalized to input signal to determine whether direct RNA interactions of a protein of interest are enriched along specific segments of the genome. Given SRSF1's canonical role as an exonic splicing enhancer, it is expected that exons that are skipped more often upon knockdown of SRSF1 will be enriched for SRSF1 eCLIP-seq reads.

Using the ENCODE data, eCLIP peaks with significant enrichment (p-value <0.05) on exons compared to input were identified genome-wide. This exon eCLIP peak dataset was then cross-referenced with three different pools of alternatively spliced exons that were identified using our consecutive intermediate kinetic analysis. The first exon category represents exons where SRSF1 knockdown decreases its inclusion level by more than 10%. The second exon pool contains exons where SRSF1 knockdown increases its inclusion level by more than 10%. The third pool contains exons that exhibit less than 0.5% change between conditions. This third group was defined as the control group (Figure 3.8). If SRSF1 is directly

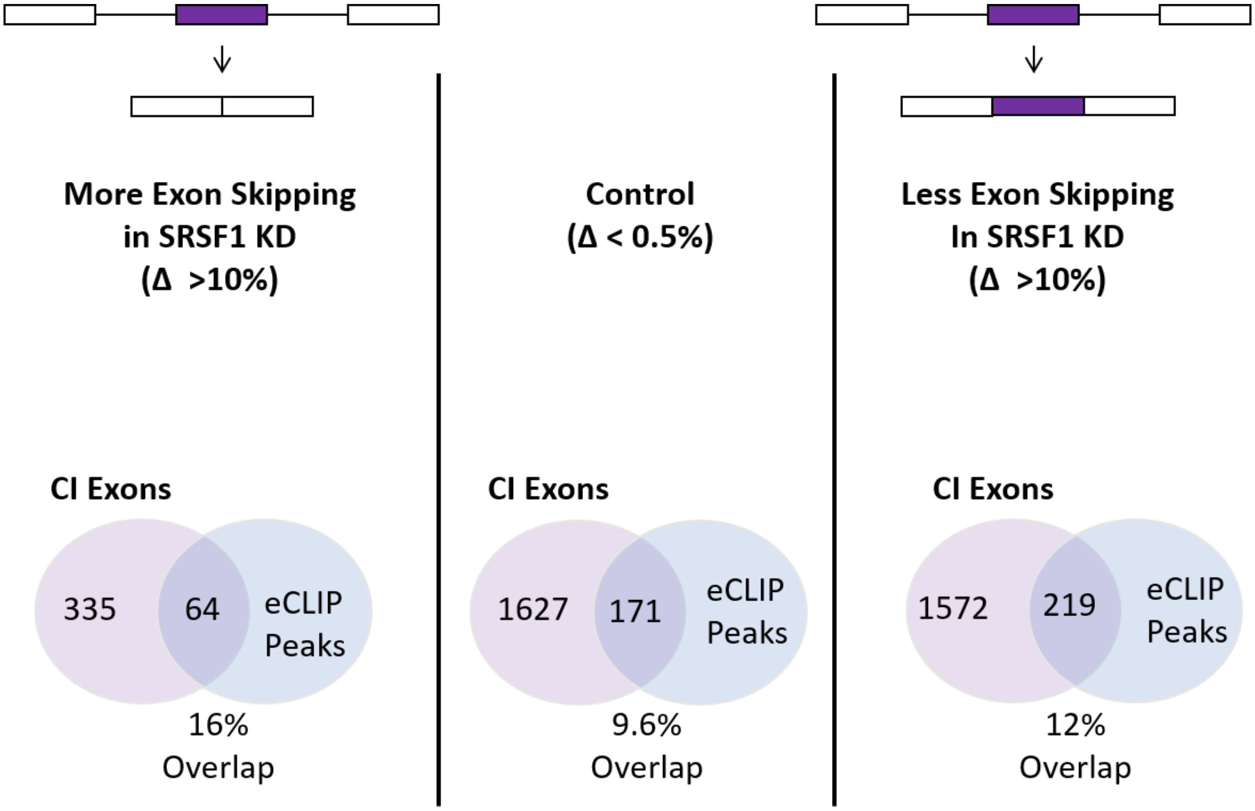
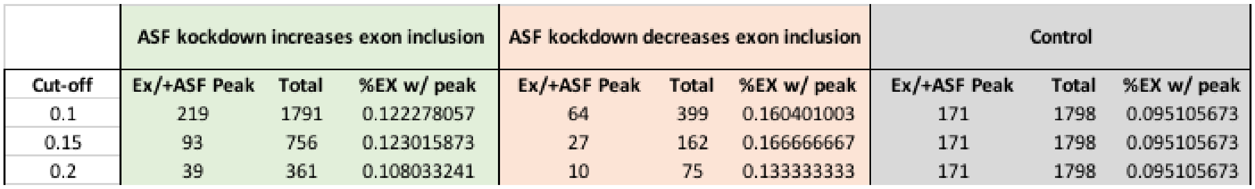


**Figure 3.8. Overview of eCLIP analysis of differentially included exons.** Exons that fit the consecutive intermediate rate model are colored in purple. Venn diagrams depict the overlap of exons that fit the consecutive intermediates (CI) kinetic model, with the SRSF1 binding beaks on the exon.

involved in promoting exon inclusion, it is expected that SRSF1 eCLIP peaks are enriched in exon pool 1, where SRSF1 knockdown results in increased exon skipping. Exons in each pool were cross-referenced with eCLIP exons to determine relative eCLIP representation. Overlaying exons with the binding peaks of SRSF1 shows a 16% overlap for category 1, a 12% overlap for category 2, and a 9.6% overlap for the control category. The differences observed between categories 1 and the control group are highly significant with a p-value of 2.2e-19. Similarly, the differences observed between pools 1 and 2 are statistically significant (p-value 1.7e-6) (Table 3.1). Increasing the stringency of exon pool selection by demanding greater changes in exclusion/inclusion differentials (from 10% to 15% or 20%) did not change the pool 1 SRSF1 eCLIP overrepresentation (Table 3.1). This SRSF1 binding/activity correlation is consistent with a preferred role for SRSF1 as an exonic splicing enhancer. However, the data also shows that SRSF1 knockdown elicits increased exon inclusion, perhaps through SRSF1's position-dependent activities, or through indirect effects that this analysis cannot decipher.

The comparative exon inclusion analysis may be limited by the number of exons that can be analyzed, because only those exons that display intron-like behavior were assessed. To test the effect of SRSF1 knockdown on a broader group of exons, an additional population of exons was investigated. A set of exons was identified that did not fit the consecutive intermediate rate model in the control condition (ie exons that do not follow intron-like behavior and are thus considered constitutively included), yet they do fit the rate model in SRSF1 knockdown conditions. These exons are examples of exons that switch from constitutive inclusion to some form of exon exclusion in SRSF1 knockdown conditions (Figure 3.9 A, B Group 3.)). A second population of exons was identified that displayed the

**Table 3.1. eCLIP analysis of SRSF1 binding on differentially included exons. Cut-off denotes the change in inclusion levels between SRSF1 and NON conditions.** Ex/+ASF peaks are the number of putative cassette exons identified through a fit to the consecutive intermediates model (CI exons) that have SRSF1 binding peaks in the eCLIP data. %EX w/peak is the percentage of CI exons that contain binding peaks in the eCLIP data.



| | ASF kockdown increases exon inclusion | | | ASF kockdown decreases exon inclusion | | | Control | | |
|---|---|---|---|---|---|---|---|---|---|
| Cut-off | Ex/+ASF Peak | Total | %EX w/ peak | Ex/+ASF Peak | Total | %EX w/ peak | Ex/+ASF Peak | Total | %EX w/ peak |
| 0.1 | 219 | 1791 | 0.122278057 | 64 | 399 | 0.160401003 | 171 | 1798 | 0.095105673 |
| 0.15 | 93 | 756 | 0.123015873 | 27 | 162 | 0.166666667 | 171 | 1798 | 0.095105673 |
| 0.2 | 39 | 361 | 0.108033241 | 10 | 75 | 0.133333333 | 171 | 1798 | 0.095105673 |

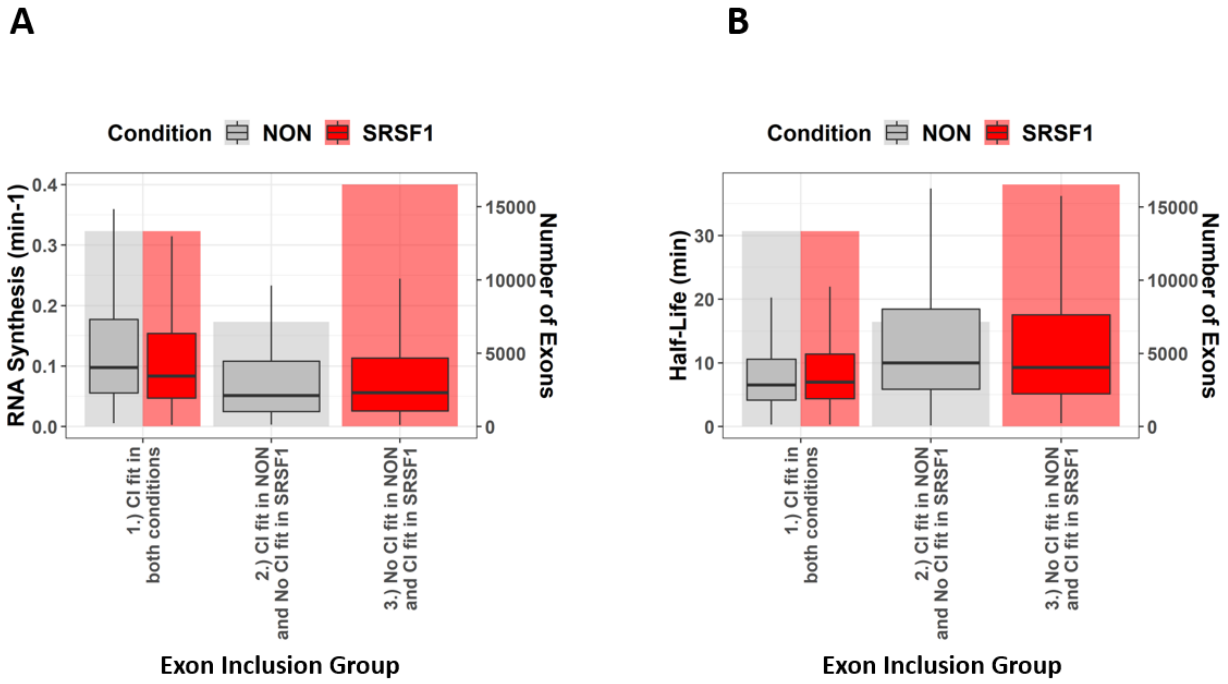


**Figure 3.9. Processing dynamics of exons that switch kinetic modes.** Exons that fit consecutive intermediate model (CI) have skipping tendencies. Those that do not fit are presumed constitutive. Bar plots represent the number of exons for each exon inclusion group. For inclusion group 1.) Exons fit CI model in both SRSF1 and NON. SRSF1 half-lives are in red and NON in grey. 2.) Exons fit CI model in NON but not in SRSF1. 3.) Exons do not fit CI model in NON but do fit CI model in SRSF1. Grey and red represent the same exons for both NON and SRSF1 for groups 2.) and 3.) A) RNA Synthesis (min$^{-1}$) for each inclusion group. B) Half-lives (min) for each inclusion group.

mirror image behavior, considered alternatively excluded in the control NON condition, but constitutively included upon SRSF1 knockdown (Figure 3.9 A, B Group 2.)). As controls, two additional categories were evaluated. First, exons that fit the consecutive intermediate model in both NON and SRSF1 knockdown conditions (Figure 3.9 A, B Group 1.)) and second, exons that did not fit the intron-like kinetic model in either condition (data not shown). Exons that do not fit are assumed to represent exons that are constitutively included relative to their neighboring exons. Thus, their inclusion may be least likely to depend on SRSF1 interactions. The first valuable information from this analysis is obtained when comparing the number of events in each of the categories. The most abundant class of switch exons are those that display alternative exclusion behavior upon SRSF1 knockdown (~16,000) (Figure 3.9 A, B Group 3.)). The least abundant group displays constitutive inclusion behavior upon SRSF1 knockdown (~7,000) (Figure 3.9 A, B Group 2.)). Additionally, exons that are prone to switching, ie switched in one condition (Figure 3.9 A, B Groups 2.), 3.)), display lower Pol II density and slower splicing compared to exons that are committed to a skipping profile in both control and SRSF1 depleted conditions (Figure 3.9 A, B Group 1.)). Thus, SRSF1 has a much stronger effect on modifying the processing kinetics to maintain the inclusion of constitutive exons.

These data underscore the notion that SRSF1 is more involved in mediating exon inclusion. Further support is provided by eCLIP peak cross-referencing as described above. The control group that does not exhibit any deviation from constitutive inclusion in both experimental conditions has an SRSF1 eCLIP overlap of 6.4%, below the by chance rate of 7.25%. All other categories of exons that do display SRSF1-associated inclusion changes have 12% overlap with eCLIP peaks. In summary, these observations suggest that SRSF1 predominantly acts as a splicing enhancer and that its enhancing activity is partially mediated by direct exonic binding.

## Discussion

We tested the hypothesis that SRSF1 preferentially regulates introns and exons that are prone to alternative splicing. We found that lower levels of this splicing factor lead to a general increase in intron inclusion and exon skipping. Our results reveal that SRSF1 is not only responsible for regulating the steady-state retention levels of introns and exons, but that it achieves these effects through the enhancement or diminution of processing rates. The mode of kinetic modulation is primarily dependent on the length of the particular gene segment being acted upon. For example, introns with lengths within the optimal 50-300 bp range or greater than 2,000 bp are removed faster in the presence of SRSF1. Reflecting on the mounting evidence that shorter introns are more prone to retention, and that long introns are subject to the slower splice site recognition accomplished by exon definition, SRSF1 is likely an activator of introns prone to inefficient recognition by the spliceosome due to length constraints [109].

Long introns (>2,000bp) are common in human genes. SRSF1 may alter the secondary structure of the RNA to bring splice sites into closer proximity for recognition, thus explaining how long introns recognized by exon definition can still be removed efficiently. Splice site strength has been established as an important factor in effective exon recognition. Weaker splice sites often correlate with poor recognition and higher tendencies for alternative splicing. Introns in our analysis that were identified as being prone to alternative splicing failed to show a significant difference in overall splice site strength. Still, length of each feature type must fall under a size that is spatially adhered to spliceosome component interactions. Our data suggest that a switch from intron definition to exon definition could be a key determinant for SRSF1's functional contributions.

SRSF1 may have additional control over splicing rates through its effects on transcription. In histone-depleted cells, chromatin accessibility and elongation rates increased on several genes and many transcripts had elevated intron retention and altered alternative splicing [110]. Moreover, SRSF1 can be recruited not only by the CTD of Pol II, but also by heterochromatin protein HP1-gamma [111]. Our findings showed that the loss of SRSF1 reduced Pol II density on all gene segments, possibly indicating that SRSF1 could contribute to regulating alternative splicing by altering co-transcriptional connections. To conclude, these data further validate previously known roles for SRSF1. We find evidence that SRSF1's primary role in maintaining constitutive splicing is mediated through its influence on transcription. As a result, the rate of splicing dictates the spliceosomal recognition of splice sites within optimal size ranges.

## Methods and Materials

### *Cell Culture*

293T cells were generously gifted by Brent Gravely. 293 T cells (catalog number: CRL-11268, ATCC) in (2) 15 cm tissue culture plates with 10 % FBS (catalog number: 30-2020, ATCC) DMEM (catalog number: 11995-065, Life technologies) medium without penicillin and streptomycin. This was done in order to scale up production of virus such that the same viral batch could be used for many time course experiments, thereby limiting experimental variability. HepG2 cells were grown as described in Chapter 2 Methods and Materials.

### *Producing shRNA Lentiviral Particles*

All plasmids were generously gifted by Brent Gravely including: pLKO-shRNA (SRSF1), psPAX2 Packaging DNA, PMD2.G Envelope DNA, pLKO-shRNA (NON-target). Production was scaled up from original estimations for 6-well plates to 15 cm plates to result in a large batch of virus. A cocktail of the above plasmids for shSRSF1 and shNON in serum free medium was transfected using FuGENE HD Transfection reagent (Cat: E2311). Incubate at room temperature for 20 min before adding DNA mix dropwise to the cells. Incubate at 37°C for 12-15 hr. Aspirate media containing transfection cocktail and wash with PBS before replenishing with fresh growth media. The next day, harvest media from the cells and store at 4°C. Add fresh growth media to the cells and incubate for another day. Next harvest the media and pool it with the media collected the previous day. Spin at 1250rpm for 5 min to remove the cells. In the same day, perform qPCR Lentivirus titration assay using the kid from Applied Biological Materials Inc (Cat LV900). Once titer has been determined, make aliquots contain the volume needed for the desired MOI for transduction experiments.

### *Lentiviral Transduction and Nascent pre-mRNA Isolation*

4sU was used to label 15 cm tissue culture plates of HEPG2 cells that were transduced with a scramble (non-target) shRNA, or shRNA targeting SRSF1. Lentiviral transduction was performed at 60% cell confluency overnight. Media was replenished with puromycin for 92 hours of selection. Cell propagation and maintenance throughout this selection resulted in 10 cm tissue culture plates with 70 % confluency. 500μM of 4sU was added for 0, 2, 5, 10, 20, 30, 40, 50, 60, 70, 90, 100, 120 minutes where “0 min” as the unlabeled control. Isolation of 4sU labeled RNA was completed as per the protocol in Appendix A.

### *High-throughput sequencing*

(same as Chapter 2 Methods and Materials)

### *Computational Analysis*

For control (NON) and SRSF1 knockdown (SRSF1) comparisons, only exons and introns that fit the consecutive intermediates model in both are used unless otherwise described as in exon Figure 3.8. Otherwise, analysis was completed as described in Chapter 2 Methods and Materials.

### *Western analysis of SRSF1 knockdown*

Protein was isolated from 2-4 million cells using RIPA buffer. Protein concentrations were determined using a BCA assay. 60 μg of protein was run on a 12% SDS-PAGE for separation. Primary antibodies were used to probe for anti-mouse SRSF1 (LifeTechnologies #32-4500), anti-rabbit SRSF7 (MBL #RN079PW) and anti-mouse α tubulin (Calbiochem #DM1A). SRSF7 was probed to confirm a lack of off-target effects of knockdown. α tubulin was probed as a loading control. Background-subtracted band density of SRSF1 protein levels for both conditions were normalized to their respective α-tubulin signal. Normalized values were used to calculate the percentage of change for SRSF1 knockdown compared to the control (NON).

### *eCLIP Analysis*

Replicate eCLIP data for SRSF1 was downloaded from encodeproject.org. Accession numbers: ENCFF522HEA, ENCFF937LBT**.** These narrowPeak files were already aligned to hg38 genome assembly/annotation and normalized to the input control by Gene Yeo's group. Peaks that had significant ($p$-value < 0.05) enrichment compared to the control input in either replicate were used, resulting in 30k significant peaks.

# CHAPTER 4

# Perspectives

The discovery that RNA processing steps such as splicing and polyadenylation are often completed co-transcriptionally has been substantiated over the past years with confidence by many different methods and groups [77, 81-83]. Understanding how RNA processing steps are spatially and temporally linked has been a scientific goal that many have strived to attain. Coincidentally, this endeavor has connected scientists with interests in eukaryotic transcription, splicing, decay, epigenetics, translation, and bioinformatics in collaborative and multidisciplinary studies in attempts to determine these connections on a genome-wide scale.

## *The Approach to Determining Global RNA Processing Kinetics*

The use of 4-thiouridine (4sU) as an approach to isolate newly synthesized transcripts has been increasingly popular and powerful in assessing each of these gene expression steps [84-88]. However, the cleavable HPDP biotin that is used to subsequently capture the labeled RNA using streptavidin columns is a low efficiency process [84]. This methodological limitation, paired with the biological reality that nascent RNA will represent only 3-5% of whole cell RNA, results in low sample yields when short 4sU labeling time points are taken. As a consequence, the shortest time points reported in published human cell-based experiments is 5 min [88-90]. For the time course series described in this thesis, a 1 min 4sU labeling burst was attempted but did not yield sufficient enrichment for nascent

RNA based on measured intron inclusion levels above the unlabeled (0 min) background control. Fortunately, 2 minutes of labeling was sufficient to significantly enrich nascent RNA, and in the end, this early time point was found to be critical in the derivation of RNA synthesis information for thousands of transcripts. This is despite the fact that shorter labeling time points (2 min, 5 min) yield as much as 30-fold less RNA than longer 4sU incubations (100 min, 120 min).

In recent years, improvements in nature of the biotin reagents and associated approaches have been made. MTS-biotin accommodates lower amounts of 4sU labeled input RNA for reliable isolation, mainly due to a more efficient biotinylation reaction [112]. For this study, our recent testing of MTS-biotin indeed resulted in much higher RNA pull-down efficiency. Unfortunately, it also consistently showed much higher background levels as was demonstrated by qRT-PCR experiments where renilla-luciferase RNA spike-in RNAs were tracked prior and post-4sU isolation (data not shown). Other groups have reported similar experiences [113]. Thus, it was impossible to input equal quantities of RNA for sequencing library preparations as is typically recommended. To mitigate this, an exogenous mixture of 92 RNA transcripts known as the ERCC spike-in, was included just prior to the library preparation for sequencing as a method for normalizing RNA quantity and transcript diversity across the samples time points [114].

Recent studies have focused primarily on the use of intron/exon junction reads to track the presence of introns relative to their flanking exons [90] or using alternative splicing algorithms such as Mixture of Isoforms (MISO) to determine levels of intron inclusion relative to flanking exons [73, 115]. By focusing only on junction reads, the downstream analysis was limited to only ~5,500 *Drosophila* genes as junction reads only account for a fraction of all uniquely sequenced reads from a sequencing lane. MISO analyses are also restricted by the depth of sequencing and the length of the reads, as determining alternative isoforms is a feat in itself and new algorithms are at the forefront of many bioinformatics lab projects. Most importantly, our method of using the fraction of intron or exon inclusion relative to the flanking exons permitted us to perform more extensive determinations of rates in human.

By using a method which utilizes both junction and non-junction reads, it is possible to uncover additional classes of RNAs that may be at work. For example, intron retention levels described in this thesis could contain contributions from reads that reflect enrichment in parts of longer introns that can only be detected in later time points. Such events could reflect a putative recursive splicing event in which multiple lariats are formed and then removed piece by piece [116]. Another scenario exists in which enriched intronic signals in later time points are not supported as a classic intron retention event due to the lack of exon-intron junction reads. One prospect for shorter introns that fall within this scenario, and for which we see higher retention levels, is the formation of a class of non-coding RNAs called circular RNAs (cRNA). One type of cRNAs is purely intronic (ciRNAs), derived from introns of 100-3,000 bps [91]. ciRNAs are formed through the escape of debranching, and they depend on consensus RNA elements near the 5' splice site and the branchpoint for proper processing. Interestingly, at least one example exists where a processed ciRNA may associate with the elongating Pol II complex at their parent gene loci to enhance transcription activity [91]. Thus, by identifying intronic reads and kinetics that are indicative of a stable intron, our genome-wide metabolic labeling approach is capable of identifying ciRNAs that may exert transcriptional influence on splicing rates. These events could be validated by RT-PCR using divergent primers that target the introns of interest. Product sizes will indicate the length of a circular construct, whereas a lack of signal would be evidence for linearity.

### *RNA Processing Kinetics*

In this study, RNA synthesis (Pol II density), splicing and retention levels were assessed. Next, the dynamics of each of these steps in relation to one another and to additional features were investigated in unprecedented scale and resolution for the human transcriptome. Time course representations of introns were fit to a consecutive intermediates kinetic model that describes their generation, processing (removal), and steady-state levels. Readouts of RNA synthesis (Pol II density), splicing rates in terms of half-lives, and retention levels were used to better understand the global dynamics of introns.

In Chapter 2, half-lives of over 80,000 modeled introns confirmed elaborate *in vitro* reports that the majority of splicing is constitutive with half-lives in the 0.4-7 min range [82]. In contrast, a study that determined splicing rates in budding yeast argues that splicing occurs much faster, where 50% of splicing is complete within ~1.4 seconds after 3'ss synthesis [67]. However, this scenario is highly unlikely in human as rates this fast would be prohibitive to alternative splicing; an important source of proteomic diversity in mammals. Predicted intron retention levels suggest that ~12% of introns have a retention of >0.2. This is in line with a genome-wide search for intron retention events, suggesting a frequency of 14.8% in intron containing genes [92]. The same kinetic rate model applied to analyze intron removal was used to identify exons that displayed intron-like behavior, that is, exons that fit to the model with inclusion levels less than 60% were likely to be excluded. In support of this interpretation, splicing rates were observed to be much faster for exons with lower inclusion rates. Furthermore, inclusion levels were much higher for exons that fall within 100-250 bp range, a size that has been well-defined as optimal for splice site recognition by the spliceosome [117] .

The ability to model introns and a subset of alternatively spliced exons permits the determination of the order of intron removal. This type of analysis allows a much better understanding as to how the rates of neighboring introns and exons could influence decisions. One potential limitation in this analysis is that it is based on the assumption of a ubiquitous transcription rate of 50 nt/second [68, 118, 119], a reasonable expectation based on previous experiments. However, small molecule (DRB) synchronization of transcription and fluorescent imaging experiments have demonstrated that elongation rates vary across a given gene depending on factors such as chromatin structure, GC content, exons/introns and pause sites [93]. Future determinations might benefit from using gene-specific elongation rates, if available, to ensure accurate estimations of the order of intron removal.

Introns with higher levels of retention displayed slower rates of splicing. Shorter introns were even more prone to slow removal and higher retention compared to longer introns. To some degree, this observation contradicts the notion that shorter introns are more efficiently removed by the spliceosome. It was argued that spliceosomal assembly across the intron is intrinsically more efficient because it constitutes a single-step identification of the splice sites to be paired [17]. While previously thought to be less efficient, it is possible that the human cell has developed additional ways to facilitate the efficiency of recognition across the exon and to adequately deal with extreme exon/intron architectures.

### *Regulation of Splicing by SRSF1*

SR proteins are best known to act as splicing activators for the recognition of exons and introns that are prone to be alternatively spliced [120, 121]. In addition, SR proteins may be involved in other gene expression steps. For example, SRSF1 has been shown to have roles in transcription, splicing, and translation [27, 32, 48]. Thus, SR proteins could influence splicing rates through different mechanisms. This is because Pol II elongation rates influence splicing outcomes, as demonstrated in the case of pause sites [63, 94]. Experiments that induce either very fast elongation or very slow elongation rates can induce exon skipping and intron retention depending on the flanking sequence lengths, suggesting a required "just right" speed of elongation to maintain constitutive splicing [95]. Upon knockdown of SRSF1, we observed an overall reduction in RNA synthesis (Pol II density) levels across all retention groups assessed. The loss of SRSF1 also induced a distance/length driven switch in the rate of splicing, by which both introns and exons are constrained. Thus, it is possible that SRSF1 influences transcription dynamics, which in turn change the splicing outcome in a distance-dependent manner. Indeed, this hypothesis may have merit when considering one of SRSF1's described roles as a facilitator for the association of several RNA processing factors to unique chromatin loci [122]. In agreement with our findings, depletion of SRSF1 was shown to decrease RNA pol II–mediated transcription and aberrant recruitment of transcription factors [96].

Length and half-life analyses demonstrated that SRSF1 is important for the efficient removal of long introns (>2,000 bp) and very short introns (<300 bp). Similar to introns, the majority of alternatively spliced exons lengths fall within the 100-300 bp range, supporting the exon definition mode of splice-site recognition. Such strong length correlations with

SRSF1 activity provide compelling evidence for SRSF1's role in enforcing efficient exon definition. Recently, Melissa Moore's group used fluorescently tagged spliceosomal subcomplexes to investigate how cross-intron and cross-exon processes cooperatively promote pre-spliceosome assembly. They found that the flanking upstream and downstream 5'ss work synergistically to efficiently recruit a U2 subcomplex to the 3' ss for internal introns and exons [97]. However, upon SRSF1 knockdown, half-lives of only the internal introns were affected, with a reduction in rates compared to wild-type. This observation suggests that while pre-spliceosome assembly occurs synergistically, SRSF1 plays a role in the actual execution of splicing, possibly through promoting the transition from cross-exon pre-spliceosome composition to cross-intron pre-spliceosome assembly [98]. To further investigate this hypothesis, upstream and downstream splice sites strength and intron/exon length would need to be considered when investigating the role of SRSF1.

Conversely, the changes in splicing rates due to SRSF1 depletion do not appear to directly drive the frequency of exon skipping. A comparison of presumably constitutive exons and skipped exon events with eCLIP data for SRSF1 binding demonstrated that SRSF1 preferentially binds to exons that are prone to being skipped. In other words, exons that fit the consecutive intermediates model upon knockdown of SRSF1 showed enriched binding of SRSF1 in wild-type conditions. These observations support the notion that SRSF1 primarily acts as a splicing activator and that its enhancing activity is partially mediated by direct exonic binding.

Collectively, these data demonstrate that splicing kinetics drive the decision as to whether an intron or an exon are alternatively spliced. Slow splicing rates were associated with alternatively spliced introns, whereas faster rates were characteristic of exon skipping.

SRSF1's role as an activator is corroborated through the enriched binding of exons that are conversely skipped upon SRSF1 depletion. For introns, its role as an activator is strongly driven by length constraints, as it aids in the efficient removal of very short (<300 bp) and long (>2,000 bp) introns.  However, this may only explain the larger picture with an undertone of lower Pol II density, suggesting an intricate network of changes in transcription to possibly further influence splicing decisions.

Most importantly, the methodology to achieve these findings is an achievement of its own. Creation of this high-quality data set has permitted the detailed analysis of multiple steps of gene expression. We deliberately used the HepG2 cell line because it is one of the cell lines investigated in detail by the ENCODE project, which includes data sets ranging from RNA protein binding profiles, chromatin modifications and structure, CRISPR knockouts, to variations of library preparations. As the regulation of gene expression is revealed to be ever more interconnected, it is a critical asset to be able to integrate data sets from the ENCODE community to further assess how other processes may be influencing splicing decisions, such as we did with the eCLIP analysis described in Chapter 3.  In an era with an increasing need for interdisciplinary work, the ability to connect multiple experiments to ask new questions is the future.

# REFERENCES

1. Jurica, M.S. and M.J. Moore, Pre-mRNA splicing: awash in a sea of proteins. Mol Cell, 2003. 12(1): p. 5-14.
2. Kaida, D., et al., U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature. 468(7324): p. 664-668.
3. Hertel, K.J., Combinatorial control of exon recognition. Journal of Biological Chemistry, 2008. 283(3): p. 1211.
4. Guth, S. and J. Valcarcel, Kinetic role for mammalian SF1/BBP in spliceosome assembly and function after polypyrimidine tract recognition by U2AF. J Biol Chem, 2000. 275(48): p. 38059-66.
5. Wu, S., et al., Functional recognition of the 3' splice site AG by the splicing factor U2AF35. Nature, 1999. 402(6763): p. 832-5.
6. Gao, K., et al., Human branch point consensus sequence is yUnAy. Nucleic acids research, 2008. 36(7): p. 2257-2267.
7. Reed, R., Initial splice-site recognition and pairing during pre-mRNA splicing. Curr Opin Genet Dev, 1996. 6(2): p. 215-20.
8. Abovich, N. and M. Rosbash, Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals. Cell, 1997. 89(3): p. 403-12.
9. Lim, S.R. and K.J. Hertel, Commitment to splice site pairing coincides with A complex formation. Mol Cell, 2004. 15(3): p. 477-83.
10. Moore, M.J. and P.A. Sharp, Evidence for two active sites in the spliceosome provided by stereochemistry of pre-mRNA splicing. Nature, 1993. 365(6444): p. 364-8.
11. Dredge, B.K., A.D. Polydorides, and R.B. Darnell, The splice of life: alternative splicing and neurological disease. Nature Reviews Neuroscience, 2001. 2(1): p. 43-50.
12. Harrow, J., et al., GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res, 2012. 22(9): p. 1760-74.
13. Pan, Q., et al., Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet, 2008. 40(12): p. 1413-1415.
14. Lewis, B.P., R.E. Green, and S.E. Brenner, Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A, 2003. 100(1): p. 189-92.
15. Dvinge, H. and R.K. Bradley, Widespread intron retention diversifies most cancer transcriptomes. Genome Med, 2015. 7(1): p. 45.
16. Zhang, X.H., C.S. Leslie, and L.A. Chasin, Computational searches for splicing signals. Methods, 2005. 37(4): p. 292-305.

17. Fox-Walsh, K.L., et al., The architecture of pre-mRNAs affects mechanisms of splice-site pairing. Proceedings of the National Academy of Sciences of the United States of America, 2005. 102(45): p. 16176-16181.

18. Sterner, D.A., T. Carlo, and S.M. Berget, Architectural limits on split genes. Proc Natl Acad Sci U S A, 1996. 93(26): p. 15081-5.

19. Sakharkar, M.K., V.T. Chow, and P. Kangueane, Distributions of exons and introns in the human genome. In Silico Biol, 2004. 4(4): p. 387-93.

20. Lander, E.S., et al., Initial sequencing and analysis of the human genome. Nature, 2001. 409(6822): p. 860-921.

21. Sakharkar, M.K., et al., An analysis on gene architecture in human and mouse genomes. In Silico Biol, 2005. 5(4): p. 347-65.

22. Yeo, G. and C.B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol, 2004. 11(2-3): p. 377-94.

23. Shepard, P.J., et al., Efficient internal exon recognition depends on near equal contributions from the 3' and 5' splice sites. Nucleic Acids Res, 2011. 39(20): p. 8928-37.

24. Black, D.L., Mechanisms of alternative pre-messenger RNA splicing. Annual review of biochemistry, 2003. 72(1): p. 291-336.

25. Fu, X.D. and M. Ares, Jr., Context-dependent control of alternative splicing by RNA-binding proteins. Nat Rev Genet, 2014. 15(10): p. 689-701.

26. Erkelenz, S., et al., Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. RNA, 2013. 19(1): p. 96-102.

27. Pandit, S., et al., Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. Mol Cell, 2013. 50(2): p. 223-35.

28. Ge, H., P. Zuo, and J.L. Manley, Primary structure of the human splicing factor ASF reveals similarities with Drosophila regulators. Cell, 1991. 66(2): p. 373-82.

29. Krainer, A.R., G.C. Conway, and D. Kozak, Purification and characterization of pre-mRNA splicing factor SF2 from HeLa cells. Genes Dev, 1990. 4(7): p. 1158-71.

30. Zhong, X.Y., et al., SR proteins in vertical integration of gene expression from transcription to RNA processing to translation. Mol Cell, 2009. 35(1): p. 1-10.

31. Zhang, Z. and A.R. Krainer, Involvement of SR proteins in mRNA surveillance. Mol Cell, 2004. 16(4): p. 597-607.

32. Das, S. and A.R. Krainer, Emerging functions of SRSF1, splicing factor and oncoprotein, in RNA metabolism and cancer. Mol Cancer Res, 2014. 12(9): p. 1195-204.

33. Pelisch, F., et al., The serine/arginine-rich protein SF2/ASF regulates protein sumoylation. Proc Natl Acad Sci U S A, 2010. 107(37): p. 16119-24.

34. Anczukow, O., et al., The splicing factor SRSF1 regulates apoptosis and proliferation to promote mammary epithelial cell transformation. Nat Struct Mol Biol, 2012. 19(2): p. 220-8.

35. Ghigna, C., et al., Cell motility is controlled by SF2/ASF through alternative splicing of the Ron protooncogene. Mol Cell, 2005. 20(6): p. 881-90.

36. Wang, X., et al., Predicting sequence and structural specificities of RNA binding regions recognized by splicing factor SRSF1. BMC Genomics, 2011. 12 Suppl 5: p. S8.

37. Krainer, A.R., G.C. Conway, and D. Kozak, The essential pre-mRNA splicing factor SF2 influences 5' splice site selection by activating proximal sites. Cell, 1990. 62(1): p. 35-42.

38. Zhu, J. and A.R. Krainer, Pre-mRNA splicing in the absence of an SR protein RS domain. Genes Dev, 2000. 14(24): p. 3166-78.

39. Caceres, J.F., et al., Role of the modular domains of SR proteins in subnuclear localization and alternative splicing specificity. J Cell Biol, 1997. 138(2): p. 225-38.

40. Cho, S., et al., Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and U1-70K snRNP protein determines early spliceosome assembly. Proceedings of the National Academy of Sciences, 2011. 108(20): p. 8233.

41. Gui, J.F., W.S. Lane, and X.D. Fu, A serine kinase regulates intracellular localization of splicing factors in the cell cycle. Nature, 1994. 369(6482): p. 678-82.

42. Rossi, F., et al., Specific phosphorylation of SR proteins by mammalian DNA topoisomerase I. Nature, 1996. 381(6577): p. 80-2.

43. Huang, Y., et al., SR splicing factors serve as adapter proteins for TAP-dependent mRNA export. Mol Cell, 2003. 11(3): p. 837-43.

44. Lai, M.C. and W.Y. Tarn, Hypophosphorylated ASF/SF2 binds TAP and is present in messenger ribonucleoproteins. J Biol Chem, 2004. 279(30): p. 31745-9.

45. Sanford, J.R., et al., A novel role for shuttling SR proteins in mRNA translation. Genes Dev, 2004. 18(7): p. 755-68.

46. Michlewski, G., J.R. Sanford, and J.F. Caceres, The splicing factor SF2/ASF regulates translation initiation by enhancing phosphorylation of 4E-BP1. Mol Cell, 2008. 30(2): p. 179-89.

47. Karni, R., et al., The splicing-factor oncoprotein SF2/ASF activates mTORC1. Proc Natl Acad Sci U S A, 2008. 105(40): p. 15323-7.

48. Maslon, M.M., et al., The translational landscape of the splicing factor SRSF1 and its role in mitosis. Elife, 2014: p. e02028.

49. Xu, X., et al., ASF/SF2-regulated CaMKIIdelta alternative splicing temporally reprograms excitation-contraction coupling in cardiac muscle. Cell, 2005. 120(1): p. 59-72.

50. Gout, S., et al., Abnormal expression of the pre-mRNA splicing regulators SRSF1, SRSF2, SRPK1 and SRPK2 in non small cell lung carcinoma. PLoS One, 2012. 7(10): p. e46539.

51. Das, S., et al., Oncogenic splicing factor SRSF1 is a critical transcriptional target of MYC. Cell Rep, 2012. 1(2): p. 110-7.

52. Fu, Y., et al., SRSF1 and SRSF9 RNA binding proteins promote Wnt signalling-mediated tumorigenesis by enhancing beta-catenin biosynthesis. EMBO Mol Med, 2013. 5(5): p. 737-50.

53. Das, S., O.I. Fregoso, and A.R. Krainer, A new path to oncogene-induced senescence: at the crossroads of splicing and translation. Cell Cycle, 2013. 12(10): p. 1477-9.

54. Shatkin, A.J., Capping of eucaryotic mRNAs. Cell, 1976. 9(4 PT 2): p. 645-53.

55. Kotovic, K.M., et al., Cotranscriptional recruitment of the U1 snRNP to intron-containing genes in yeast. Mol Cell Biol, 2003. 23(16): p. 5768-79.

56. Misteli, T., J.F. Caceres, and D.L. Spector, The dynamics of a pre-mRNA splicing factor in living cells. Nature, 1997. 387(6632): p. 523-7.

57. Listerman, I., A.K. Sapra, and K.M. Neugebauer, Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. Nat Struct Mol Biol, 2006. 13(9): p. 815-22.

58. Huang, S. and D.L. Spector, Nascent pre-mRNA transcripts are associated with nuclear regions enriched in splicing factors. Genes Dev, 1991. 5(12A): p. 2288-302.

59. Schmidt, U., et al., Real-time imaging of cotranscriptional splicing reveals a kinetic model that reduces noise: implications for alternative splicing regulation. The Journal of cell biology, 2011. 193(5): p. 819.

60. de la Mata, M., et al., A Slow RNA Polymerase II Affects Alternative Splicing In Vivo. Molecular Cell, 2003. 12(2): p. 525-532.

61. Hicks, M.J., B.J. Lam, and K.J. Hertel, Analyzing mechanisms of alternative pre-mRNA splicing using in vitro splicing assays. Methods, 2005. 37(4): p. 306-313.

62. HOWE, K.J., C.M. KANE, and M. ARES, Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae. Rna, 2003. 9(8): p. 993-1006.

63. Listerman, I., A.K. Sapra, and K.M. Neugebauer, Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. NATURE STRUCTURAL AND MOLECULAR BIOLOGY, 2006. 13(9): p. 815.

64. Khodor, Y.L., et al., Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. Genes & Development, 2011. 25(23): p. 2502-2512.

65. Shepard, S., M. McCreary, and A. Fedorov, The peculiarities of large intron splicing in animals. PLoS One, 2009. 4(11): p. e7853.

66. Görnemann, J., et al., Cotranscriptional Spliceosome Assembly Occurs in a Stepwise Fashion and Requires the Cap Binding Complex. Molecular Cell, 2005. 19(1): p. 53-63.

67. Oesterreich, F.C., et al., Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. Cell, 2016. 165(2): p. 372-381.

68. Singh, J. and R.A. Padgett, Rates of in situ transcription and splicing in large human genes. nAture structurAl & moleculAr biology, 2009. 16(11): p. 1128-1133.

69. Brody, Y., et al., The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing. PLoS Biology, 2011. 9(1): p. e1000573.

70. Larson, D.R., et al., Real-time observation of transcription initiation and elongation on an endogenous yeast gene. Science, 2011. 332(6028): p. 475-8.

71. Pandya-Jones, A. and D.L. Black, Co-transcriptional splicing of constitutive and alternative exons. Rna, 2009. 15(10): p. 1896.

72. Pai, A.A., et al., The kinetics of pre-mRNA splicing in the Drosophila genome and the influence of gene architecture. Elife, 2017. 6.

73. Windhager, L., et al., Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. Genome research, 2012. 22(10): p. 2031-2042.

74. Li, B., et al., RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics, 2010. 26(4): p. 493-500.

75. Audibert, A., D. Weil, and F. Dautry, In Vivo Kinetics of mRNA Splicing and Transport in Mammalian Cells. Molecular and Cellular Biology, 2002. 22(19): p. 6706-6718.

76. Dye, M.J., N. Gromak, and N.J. Proudfoot, Exon tethering in transcription by RNA polymerase II. Mol Cell, 2006. 21(6): p. 849-59.

77. Wilhelm, B.T., et al., Differential patterns of intronic and exonic DNA regions with respect to RNA polymerase II occupancy, nucleosome density and H3K36me3 marking in fission yeast. Genome Biol, 2011. 12(8): p. R82.

78. Kwak, H., et al., Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science, 2013. 339(6122): p. 950-3.

79. Crabb, T.L., B.J. Lam, and K.J. Hertel, Retention of spliceosomal components along ligated exons ensures efficient removal of multiple introns. Rna, 2010. 16(9): p. 1786.

80. Oesterreich, F.C., N. Bieberstein, and K.M. Neugebauer, Pause locally, splice globally. Trends in cell biology, 2011. 21(6): p. 328-335.

81. Vargas, D.Y., et al., Single-Molecule Imaging of Transcriptionally Coupled and Uncoupled Splicing. Cell, 2011. 147(5): p. 1054-1065.

82. Rabani, M., et al., High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. Cell, 2014. 159(7): p. 1698-710.

83. Hatton, A.R., V. Subramaniam, and A.J. Lopez, Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. Mol Cell, 1998. 2(6): p. 787-96.

84. Conklin, J.F., A. Goldman, and A.J. Lopez, Stabilization and analysis of intron lariats in vivo. Methods, 2005. 37(4): p. 368-75.

85. Zhang, X.O., et al., The temporal landscape of recursive splicing during Pol II transcription elongation in human cells. PLoS Genet, 2018. 14(8): p. e1007579.

86. Bradnam, K.R. and I. Korf, Longer first introns are a general property of eukaryotic gene structure. PLoS One, 2008. 3(8): p. e3093.

87. Park, S.G., S. Hannenhalli, and S.S. Choi, Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals. BMC Genomics, 2014. 15: p. 526.

88. Vera Alvarez, R., et al., TPMCalculator: one-step software to quantify mRNA abundance of genomic features. Bioinformatics, 2018.

89. Hube, F. and C. Francastel, Mammalian introns: when the junk generates molecular diversity. Int J Mol Sci, 2015. 16(3): p. 4429-52.

90. Glanzer, J., et al., RNA splicing capability of live neuronal dendrites. Proc Natl Acad Sci U S A, 2005. 102(46): p. 16859-64.

91. Racca, C., et al., The Neuronal Splicing Factor Nova Co-Localizes with Target RNAs in the Dendrite. Front Neural Circuits, 2010. 4: p. 5.

92. Wong, J.J., et al., Orchestrated intron retention regulates normal granulocyte differentiation. Cell, 2013. 154(3): p. 583-95.

93. Middleton, R., et al., IRFinder: assessing the impact of intron retention on mammalian gene expression. Genome Biol, 2017. 18(1): p. 51.

94. Berget, S.M., Exon recognition in vertebrate splicing. J Biol Chem, 1995. 270(6): p. 2411-4.

95. Bembich, S., et al., Predominance of spliceosomal complex formation over polyadenylation site selection in TDP-43 autoregulation. Nucleic acids research, 2013: p. gkt1343.

96. Christinat, Y., R. Pawlowski, and W. Krek, jSplice: a high-performance method for accurate prediction of alternative splicing events and its application to large-scale renal cancer transcriptome data. Bioinformatics, 2016. 32(14): p. 2111-9.

97. Veitenhansl, M., et al., 40(th) EASD Annual Meeting of the European Association for the Study of Diabetes : Munich, Germany, 5-9 September 2004. Diabetologia, 2004. 47(Suppl 1): p. A1-A464.

98. Cui, Y., M. Cai, and H.E. Stanley, Comparative Analysis and Classification of Cassette Exons and Constitutive Exons. Biomed Res Int, 2017. 2017: p. 7323508.

99. Miro, J., et al., Identification of Splicing Factors Involved in DMD Exon Skipping Events Using an In Vitro RNA Binding Assay. Methods Mol Biol, 2018. 1687: p. 157-169.

100. Fredericks, A.M., et al., RNA-Binding Proteins: Splicing Factors and Disease. Biomolecules, 2015. 5(2): p. 893-909.

101. Hurt, E., et al., Cotranscriptional recruitment of the serine-arginine-rich (SR)-like proteins Gbp2 and Hrb1 to nascent mRNA via the TREX complex. Proc Natl Acad Sci U S A, 2004. 101(7): p. 1858-62.

102. Ip, J.Y., et al., Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. Genome research, 2011. 21(3): p. 390.

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

104. Misra, A. and M.R. Green, From polyadenylation to splicing: Dual role for mRNA 3' end formation factors. RNA Biol, 2016. 13(3): p. 259-64.

105. Davuluri, R.V., et al., The functional consequences of alternative promoter use in mammalian genomes. Trends Genet, 2008. 24(4): p. 167-77.

106. Ramanathan, A., G.B. Robb, and S.H. Chan, mRNA capping: biological functions and applications. Nucleic Acids Res, 2016. 44(16): p. 7511-26.

107. Consortium, E.P., An integrated encyclopedia of DNA elements in the human genome. Nature, 2012. 489(7414): p. 57-74.

108. Van Nostrand, E.L., et al., Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods, 2016. 13(6): p. 508-14.

109. Rabani, M., et al., Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. Nature biotechnology, 2011. 29(5): p. 436-442.

110. Prado, F., S. Jimeno-Gonzalez, and J.C. Reyes, Histone availability as a strategy to control gene expression. RNA Biol, 2017. 14(3): p. 281-286.

111. Salton, M., T.C. Voss, and T. Misteli, Identification by high-throughput imaging of the histone methyltransferase EHMT2 as an epigenetic regulator of VEGFA alternative splicing. Nucleic Acids Res, 2014. 42(22): p. 13662-73.

112. Duffy, E.E., et al., Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry. Mol Cell, 2015. 59(5): p. 858-66.

113. Marzi, M.J. and F. Nicassio, Uncovering the Stability of Mature miRNAs by 4-Thio-Uridine Metabolic Labeling. Methods Mol Biol, 2018. 1823: p. 141-152.

114. Jiang, L., et al., Synthetic spike-in standards for RNA-seq experiments. Genome research, 2011. 21(9): p. 1543-1551.

115. Katz, Y., et al., Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods, 2010. 7(12): p. 1009-15.

116. Sibley, C.R., et al., Recursive splicing in long vertebrate genes. Nature, 2015. 521(7552): p. 371-375.

117. Will, C.L. and R. Luhrmann, Spliceosome structure and function. Cold Spring Harb Perspect Biol, 2011. 3(7).

118. Gotta, S.L., O.L. Miller, Jr., and S.L. French, rRNA transcription rate in Escherichia coli. J Bacteriol, 1991. 173(20): p. 6647-9.

119. Darzacq, X., et al., In vivo dynamics of RNA polymerase II transcription. Nat Struct Mol Biol, 2007. 14(9): p. 796-806.

120. Graveley, B.R. and T. Maniatis, Arginine/serine-rich domains of SR proteins can function as activators of pre-mRNA splicing. Mol Cell, 1998. 1(5): p. 765-71.

121. Bradley, T., M.E. Cook, and M. Blanchette, SR proteins control a complex network of RNA-processing events. RNA, 2015. 21(1): p. 75-92.

122. Tripathi, V., et al., SRSF1 regulates the assembly of pre-mRNA processing factors in nuclear speckles. Mol Biol Cell, 2012. 23(18): p. 3694-706.

# 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 [μM] |
|---|---|
| 120 | 100–200 |
| 60 | 200–500 |
| 15–30 | 500–1000 |
| <10 | 500–20,000 |

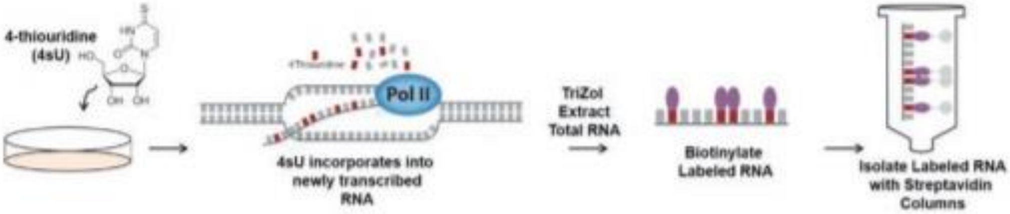


**Fig. 1** Conceptual workflow of 4sU labeling and isolation of newly transcribed RNA

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 sulfylhydryl 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× 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. μMacs Streptavidin Kit (*see* **Note 2**).
2. 1× Washing Buffer: 100 mM Tris pH 7.5, 10 mM EDTA, 1 M NaCl, 0.1% Tween20.
3. 100 mM Dithriothreitol (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

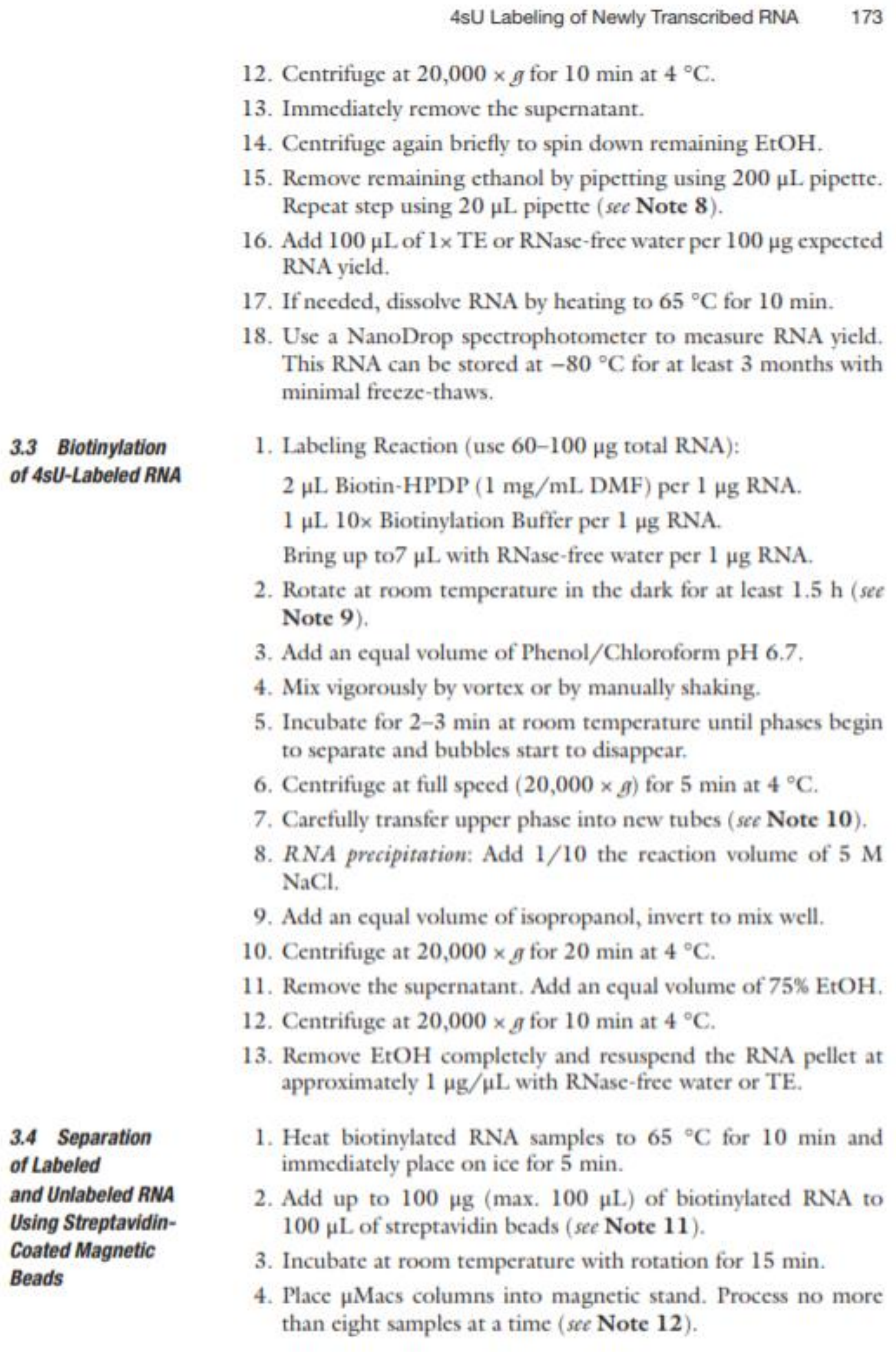
### 3.1 4sU Labeling of Nascent RNA

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 × *g* for 15 min at 4 °C (*see* **Note 7**).
5. Transfer aqueous upper phase (containing the RNA) to a new tube.
6. Add ½ 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 × *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 × *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 μL pipette. Repeat step using 20 μL pipette (*see* **Note 8**).
16. Add 100 μL of 1× TE or RNase-free water per 100 μg 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 μg total RNA):

   2 μL Biotin-HPDP (1 mg/mL DMF) per 1 μg RNA.

   1 μL 10× Biotinylation Buffer per 1 μg RNA.

   Bring up to7 μL with RNase-free water per 1 μg 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 × *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 × *g* for 20 min at 4 °C.
11. Remove the supernatant. Add an equal volume of 75% EtOH.
12. Centrifuge at 20,000 × *g* for 10 min at 4 °C.
13. Remove EtOH completely and resuspend the RNA pellet at approximately 1 μg/μL 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 μg (max. 100 μL) of biotinylated RNA to 100 μL of streptavidin beads (*see* **Note 11**).
3. Incubate at room temperature with rotation for 15 min.
4. Place μMacs 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 μL of beads for 80 μg 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 μL 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

1. Pandya-Jones A, Black DL (2009) Co-transcriptional splicing of constitutive and alternative exons. RNA 15(10):1896–1908
2. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322(5909):1845–1848
3. 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
4. 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
5. 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
6. 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
7. 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
8. Burger K, Mühl B, Kellner M, Rohrmoser 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

# The TCF C-Clamp DNA Binding Domain Expands the Wnt Transcriptome via Alternative Target Recognition

*Published online 20 November 2014*

*Nucleic Acids Research, 2014, Vol. 42, No. 22* **13615–13632**
*doi: 10.1093/nar/gku1186*

# The TCF C-clamp DNA binding domain expands the Wnt transcriptome via alternative target recognition

**Nate P. Hoverter[1,†], Michael D. Zeller[2,†], Miriam M. McQuade[1], Angela Garibaldi[1], Anke Busch[1], Elizabeth M. Selwan[1], Klemens J. Hertel[1], Pierre Baldi[2] and Marian L. Waterman[1,*]**

[1]Department of Microbiology and Molecular Genetics, University of California, Irvine, Irvine, CA 92697, USA and [2]Department of Information and Computer Science, University of California, Irvine, Irvine, CA 92697, USA

Received March 11, 2014; Revised November 01, 2014; Accepted November 04, 2014

## ABSTRACT

**LEF/TCFs direct the final step in Wnt/β-catenin signalling by recruiting β-catenin to genes for activation of transcription. Ancient, non-vertebrate TCFs contain two DNA binding domains, a High Mobility Group box for recognition of the Wnt Response Element (WRE; 5′-CTTTGWWS-3′) and the C-clamp domain for recognition of the GC-rich Helper motif (5′-RCCGCC-3′). Two vertebrate TCFs (TCF-1/TCF7 and TCF-4/TCF7L2) use the C-clamp as an alternatively spliced domain to regulate cell-cycle progression, but how the C-clamp influences TCF binding and activity genome-wide is not known. Here, we used a doxycycline inducible system with ChIP-seq to assess how the C-clamp influences human TCF1 binding genome-wide. Metabolic pulse-labeling of nascent RNA with 4′Thiouridine was used with RNA-seq to connect binding to the Wnt transcriptome. We find that the C-clamp enables targeting to a greater number of gene loci for stronger occupancy and transcription regulation. The C-clamp uses Helper sites concurrently with WREs for gene targeting, but it also targets TCF1 to sites that do not have readily identifiable canonical WREs. The coupled ChIP-seq/4′Thiouridine-seq analysis identified new Wnt target genes, including additional regulators of cell proliferation. Thus, C-clamp containing isoforms of TCFs are potent transcriptional regulators with an expanded transcriptome directed by C-clamp-Helper site interactions.**

## INTRODUCTION

The Wnt signaling pathway is one of several vital developmental pathways conserved in all phyla of the animal kingdom (1,2). Wnt proteins are secreted morphogens that bind to their cognate transmembrane receptor, a complex of Frizzled and LRP family proteins (3). Binding leads to the release of the central messenger β-catenin from a destruction complex so it can translocate to the nucleus (4). Nuclear β-catenin then complexes with a member of the Lymphoid Enhancer Factor/T Cell Factor (LEF/TCF) family of DNA binding proteins, and in turn, recruits chromatin modifiers and components of the general transcription machinery to activate a Wnt target gene expression program (5–7). Depending on the developmental context and the gene programs that β-catenin-LEF/TCF complexes activate, cells are directed to proliferate, self-renew or differentiate toward specific cell fates. In abnormal settings, such as when mutations in Wnt pathway components cause overactive signaling, gene expression patterns of proliferation are unbalanced. For example, early development of the majority (80%) of colon cancer cases are driven by overactive Wnt/β-catenin signaling (8,9). The targets and gene programs that are misregulated in these cells are specified by the DNA binding specificities of the LEF/TCFs.

Genetic and biochemical studies have implicated the LEF/TCF family to be the primary sequence-specific transcription factors that mediate WNT target gene activation (5,10–12). For model systems, such as *Drosophila* and *Caenorhabditis elegans*, a single TCF carries this responsibility (dTCF/pangolin and POP-1, respectively). In vertebrate systems, the LEF/TCF family has expanded in both the number of family members as well as the diversity of isoforms produced from each gene. Diversification and subfunctionalization is due in part to alternative splicing and alternative promoter usage, which interestingly does not introduce newly evolved domains but instead has made ancient, existing domains, such as the N-terminus (β-catenin

[*]To whom correspondence should be addressed. Tel: +1 949 824 2885; Fax: +1 949 824 8598; Email: marian.waterman@uci.edu

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Present address: Marian Waterman, Department of Microbiology and Molecular Genetics, University of California, Irvine, Irvine, CA 92697, USA.

binding) and C-terminus (E-tail; C-clamp) an alternative choice (13,14). In mammals, the LEF/TCF family consists of four members: TCF1 (TCF7 gene), TCF4 (TCF7L2), LEF1 (LEF1) and TCF3 (TCF7L1). Family members are often co-expressed as sets of alternatively spliced isoforms and knockout studies indicate important non-redundant functions (14).

All LEF/TCF isoforms contain a highly conserved sequence-specific High Mobility Group (HMG) DNA binding domain (DBD) that binds to a eight nucleotide DNA sequence motif frequently called a Wnt Response Element (WRE; 5′-CTTTGWWS-3′), but other domains can greatly influence DNA binding. For example, alternative splicing of the C-terminus of TCF1 and TCF4 produces an isoform (named an 'E-tail' isoform) that is particularly potent in its ability to regulate transcription (15–17). TCF1E has been shown to be the only LEF/TCF isoform that can strongly regulate the SP5, CDX1 and LEF1 Wnt-target promoters (15), while TCF4E was shown to regulate the CDX1 promoter (17,18). Global gene expression analysis as well as proliferation and cell-cycle analysis in colon cancer cells showed that E-tail isoforms are potent regulators of cell-cycle progression through the G1 phase of the cell cycle (15,19), whereas other isoforms of LEF/TCFs are not effective. Interestingly, the predominant form of TCF1 in colon cancer is TCF1B, a form that does not contain the E-tail. This is in contrast to its expression in normal colon cells where the E-tail form is more prevalent and is expressed as a dominant negative isoform (dnTCF1E) that does not contain the N-terminal β-catenin binding domain (16).

The E-tail isoforms of TCF1 and TCF4 have been well studied because they include a second ancient DBD called the C-clamp and because they are predominant isoforms in the intestinal epithelium (16,20–22). The C-clamp is a zinc binding domain that recognizes a GC-rich sequence motif ('Helper site'; 5′-RCCGCCR-3′) in combination with the HMG-WRE interaction *in vivo* and *in vitro* (15,23,24). Binding of the C-clamp to Helper sites is thought to be largely responsible for observations that the E-tail isoforms of LEF/TCFs are transcriptionally potent isoforms. There are features of HMG and C-clamp cooperation that are unique, including an unusual degree of flexibility in the spacing and orientation between the WRE and the Helper site; Helper sites can be located 5′ or 3′ of WREs and the spacing between motifs can vary from 1 nucleotide to 11 nucleotides (20,23). While *in vitro* studies and transient transfection experiments confirm that this degree of flexibility can be tolerated, the extent to which the C-clamp-Helper site interaction contributes to genome-wide binding and transcriptional regulation of Wnt target genes is unknown. Additionally, the pattern of co-occurring WREs and Helper sites in the context of chromatin has not been studied and therefore the constraints on the allowable distance and orientation between Helper and WRE sites for functional synergy are poorly understood. Finally, since only a few target genes have been identified where the C-clamp makes an essential contribution, a genome-wide analysis is needed to reveal not only patterns of binding, but to identify gene programs that are directly targeted by these ancient isoforms.

We performed ChIP-seq experiments to determine the binding profile of a wild-type and C-clamp mutant version of TCF1 in DLD-1 colon cancer cells. Our results indicate that the C-clamp-Helper site interaction widely contributes to binding site selection and strength. We find that wild-type TCF1 bound to a greater number of gene bodies and promoters compared to mutant TCF1, suggesting that the C-clamp helps position TCF1 for transcriptional regulation. Our bioinformatics analysis also indicates that the C-clamp forms of TCF1 can utilize Helper site(s) for transcription regulation independent of any obvious canonical WRE motif. We also assessed early changes in gene expression in tandem with ChIP-seq experiments using a metabolic labeling and high-throughput RNA-seq technique called 4′Thiouridine-seq. This technique selectively labels actively transcribed nascent RNAs, and by comparing 4′Thiouridine-seq transcript changes with ChIP-seq data, we identified new C-clamp-dependent Wnt target genes that showed direct regulation by TCF1, including histone genes, and other genes with high relevance to cell proliferation. We conclude that the C-clamp enables gene regulation independent of canonical WRE interactions in modulating the DNA activities and transcriptional output of TCF1E. However, our results also indicate that the DNA binding specificities of the HMG box and C-clamp synergize to control gene expression of a subset of Wnt target genes.

## MATERIALS AND METHODS

### Establishment of inducible DLD-1 colon cancer cells

The establishment of inducible dnTCF1EWT and dnTCF1Emut DLD-1 cell lines has been described previously (15).

### ChIP-Seq and ChIP validation

See Supplementary Methods. ChIP-seq data set was deposited to GEO with accession GSE53536.

### 4′Thiouridine-seq

See Supplementary Methods.

### Plasmids

Construction of TCF1EWT, TCF1Emut, expression plasmids was described previously (25,26).

### Luciferase assay

COS-1 cells were transiently transfected with BioT transfection reagent according to the manufacturer's protocol (Bioland Scientific LLC). COS-1 cells were plated at a density of 200 000 cells/well in 6-well plates 20 h before transfection. Luciferase reporter constructs (0.4 ug) were cotransfected with β-catenin (0.4 ug), β-galactosidase (0.1 ug) and a LEF/TCF expression vector (0.1 ug). Cells were harvested after 20 h, and luciferase and β-galactosidase activities were determined as described (25).

### Electrophoretic Mobility Shift Assay (EMSA)

EMSAs were carried out with 1 ng (~200 cps) of radioactive oligonucleotide in a final reaction volume of 20 ul containing 10 mM HEPES (pH 8.0), 2.5 mM ethylenediaminetetraacetic acid (EDTA), 10% glycerol, 20 mM KCl, 5 mM MgCl2, 0.024 ug/ul salmon sperm DNA and 20 mM dithiothreitol (DTT). COS-1 cells were transiently transfected with EVR2 and expression vectors for full-length human TCF1EWT. COS-1 cells were prepared 48 h after transfection by swelling cells on ice, immersing them for 15 min in hypotonic lysis buffer (10 mM Tris pH 7.9, 50 mM KCl, 10 mM MgCl2, 0.01 mM EDTA, 1 mM DTT, 0.01 mM EGTA, 1 mM phenylmethylsulfonyl fluoride, protease inhibitor cocktail), and douncing. The methylated probe was ordered from Fisher Scientific and contains four methylated CpGs (two on each strand in the Helper sites).

## RESULTS

### ChIP-seq of dnTCF1EWT and dnTCF1Emut

Most ChIP-seq studies are descriptive studies that focus on establishing the binding profiles of one or several DNA binding proteins. However, ChIP-seq experiments are well suited to testing the functional contribution of individual protein domains within a DNA binding protein. For example, it was shown that the genome-wide binding profile of the E2F transcription factor in MCF-7 breast cancer cells is not mediated through protein–protein interaction domains, but rather almost exclusively through the E2F DBD (27). Interestingly, while there are a significant number of transcription regulators that have more than one DBD (28,29), to our knowledge there are no studies that examine the relative contribution of one DBD versus another in terms of genome-wide binding patterns. Furthermore, the coupling of this analysis to high-throughput detection of nascent transcripts is an emerging approach with few, if any, published examples. While this is a similar approach to integrating ChIP-seq with microarray or RNA-seq data sets, this technique allows for more sensitive detection of newly transcribed RNA (30).

To determine the contribution of the C-clamp domain to the genome-wide binding pattern of TCF1 binding, ChIP-seq experiments were performed with a well established, doxycycline-inducible system in DLD-1 colon cancer cells (15,19,31). DLD-1 cells have high levels of endogenous Wnt signaling and therefore constitutive occupancy of Wnt target sites by endogenous β-catenin-LEF/TCF complexes. The doxycycline inducible system takes advantage of a dominant negative form of TCF1E (dnTCF1E) which when expressed in colon cancer cells, can interfere with endogenous β-catenin-LEF/TCF complexes by competing for occupancy of binding sites throughout the genome (Figure 1A) (31). Competition results in a downregulation of Wnt target gene expression because β-catenin and its transcriptional co-activators are displaced from regulatory sites (15,19) (Figure 1A). This system was first established by van de Wetering *et al.* to discover that either dnTCF1E or dnTCF4E expression causes a stall in the G1 phase of the cell cycle (19). We subsequently used this system to show that the stall requires the C-clamp domain in the E-tail (15). A microarray analysis of gene expression in parallel cell lines with either induced wild-type dnTCF1E (dnTCF1EWT) or a mutant version which has a five amino acid substitution in the C-clamp rendering it null for DNA binding (dnTCF1Emut), identified C-clamp-dependent changes in gene expression, changes that were connected to cell-cycle progression and proliferation (15). dnTCF1Emut is equally capable as dnTCF1EWT of regulating WRE-driven Wnt reporter constructs, such as TOPTK ((15), unpublished observation), so to what extent these changes reflect differences in dnTCF1E occupancy of target genes or secondary and tertiary effects not related to binding is not fully known (20). Also, whether the C-clamp defines new patterns of sequence specificity and a new subclass of target gene, or whether its function is strictly auxiliary to classic WRE-dependent gene targeting is not known. We therefore used ChIP-seq comparisons of parallel doxycycline induction of dnTCF1EWT and dnTCF1Emut to assess the role of the C-clamp in the genome-wide binding patterns of TCF1.

dnTCF1EWT and dnTCF1Emut (C-clamp mutant) were induced with doxycycline in duplicate cultures for 1 h (undetectable expression), 2 h (low amount of expression) or 9 h (higher amount of expression) (Figures 1B, C and 2A) and protein-DNA interactions were stabilized by crosslinking with formaldehyde. For each cell line, an untreated control was also included in the analysis (no doxycycline, 0 h). The early 2 h time point was included in the ChIP-seq analysis because little is known about the genome-wide binding patterns of newly translated TCFs or induced DNA binding proteins in general. In addition, the levels of immunoprecipitated wild-type and mutant protein were very similar at 2 h (Figure 2A). Much of the subsequent analysis in this study derives from this early time point. Immunoprecipitations were performed with FLAG-antibody-conjugated magnetic beads because dnTCF1EWT and dnTCF1Emut both have an N-terminal FLAG tag (a tag which we have previously shown is neutral for dnTCF1E activities (15)). Western blot analysis demonstrated that no detectable dnTCF1E was pulled down in untreated (0 h) cells and that similar levels of dnTCF1EWT and dnTCF1Emut were pulled down after 2 and 9 h of doxycycline treatment (Figure 2A). ChIP-seq was then performed in duplicate for each condition to identify regions of occupancy (see Materials and Methods). A comparison of the peaks called at each time point exhibited significant overlap for each of the dnTCF1EWT and dnTCF1Emut biological replicates. For example, the reciprocal overlap of the top 20% of dnTCF1EWT 2 h peaks was 61% and 54% and for dnTCF1Emut 2 h it was 80% and 86%. We further assessed reproducibility using scatter plot and Irreproducible Discovery Rate analysis (IDR; (32,33)). IDR analysis is a stringent rank-order approach that compares the rank order *P*-values of peaks between biological replicates and assigns a value to the concordance (or discordance) much like a false discovery rate value. IDR analysis showed that the top 1000 peaks from 2 h of Doxycline induction had a reproducibility index of 0.1 or better for the dnTCF1EWT replicates and 0.04 for the dnTCF1Emut replicates (Supplementary Figure S1D). Following the ENCODE protocol for ChIP-seq (32), we therefore pooled the biological replicates for each time point (see Supplementary

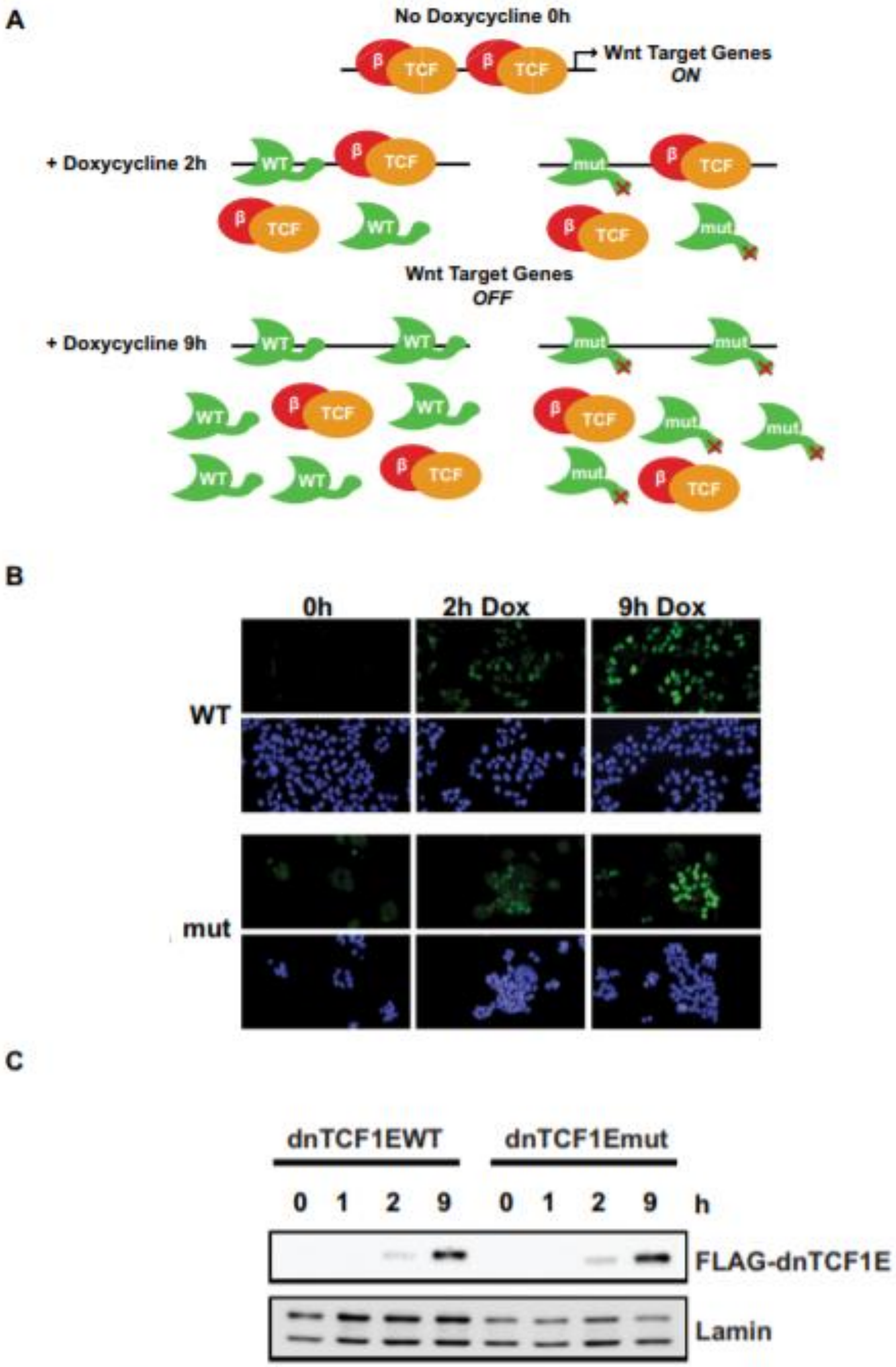


**Figure 1.** dnTCF1EWT and dnTCF1Emut doxycycline inducible system in DLD-1 colon cancer cells. (A) In the absence of doxycycline, dnTCF1E is not produced and high levels of β-catenin (red) signaling activate Wnt target gene expression. At 2 h of doxycycline induction, low amounts of dnTCF1EWT and dnTCF1Emut (green) are produced, which compete with endogenous LEF/TCF factors (orange) for binding sites and cause a downregulation of Wnt target gene expression. At 9 h post-induction, there are higher levels of dnTCF1E induced, which cause increased repression of Wnt target gene expression. (B) Immunofluorescence images of dnTCF1EWT and dnTCF1Emut expression in DLD-1 cells. dnTCF1EWT and dnTCF1Emut (green) are found exclusively in the nucleus (stained blue with DAPI). (C) Western blot analysis of a Doxycycline induction time course of FLAG-dnTCF1EWT and FLAG-dnTCF1Emut expression.

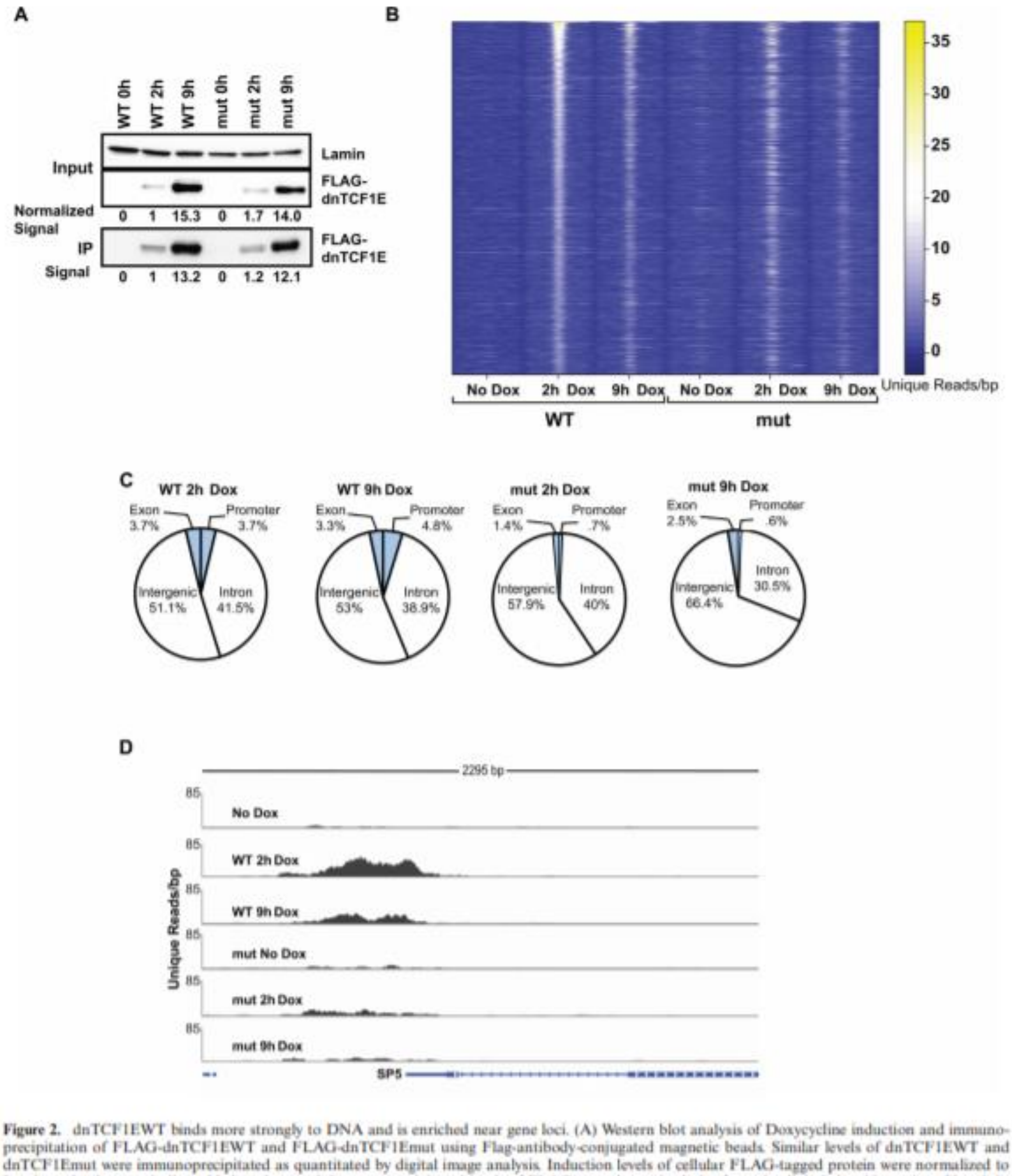


**Figure 2.** dnTCF1EWT binds more strongly to DNA and is enriched near gene loci. (A) Western blot analysis of Doxycycline induction and immunoprecipitation of FLAG-dnTCF1EWT and FLAG-dnTCF1Emut using Flag-antibody-conjugated magnetic beads. Similar levels of dnTCF1EWT and dnTCF1Emut were immunoprecipitated as quantitated by digital image analysis. Induction levels of cellular FLAG-tagged protein were normalized to Lamin and values are indicated below the panel. Immunoprecipitated levels of FLAG-tagged protein derive from the digital signal intensity of bands in the bottom panel. (B) The top scoring 1000 peaks from the dnTCF1EWT 2 h analysis were sorted by read intensity and displayed in a heat map. Read intensities (number of reads per 10 base pairs) for these regions were also displayed for the other ChIP-seq samples. The window length for each sample is 2000 bp, centered on the peak of dnTCF1E occupancy. dnTCF1E peaks were stronger at 2 h induction than 9 h induction. dnTCF1Emut was enriched at most, but not all regions that dnTCF1EWT bound. However, dnTCF1Emut enrichment at these regions was mostly weaker. (C) Distribution of the top 1000 dnTCF1EWT and dnTCF1Emut peaks. dnTCF1EWT bound to a greater number of promoter (defined as 4 kb centered over the transcription start site) and exonic regions than dnTCF1Emut. dnTCF1EWT also bound to a greater number of untramslated regions, exons and transcription termination sites, which together encompass the 'Other' category. (D) Example of a ChIP-seq peak at the promoter of the Wnt target gene SP5. dnTCF1EWT showed stronger binding at the SP5 promoter, consistent with its known status as a C-clamp-preferred target gene.

Figure S1 and Supplementary Methods), and for much of the analysis that follows, only the top 1000 peaks from each of the 2 h time points was used.

A heat map representation of the distribution of read counts for the top 1000 dnTCF1EWT peaks at 2 h (out of 5580 total peaks) across all samples from this pooled data set shows sharp peaks of occupancy (Figure 2B) and known high affinity binding sites close to the peak centers (Supplementary Figure S2). The fainter heatmap signals for dnTCF1Emut at these sites indicate weaker binding, a general pattern also reflected in the fewer number of total peaks called for dnTCF1Emut (1863 total peaks) and the overall lower scores for those peaks (Supplementary Figure S2A and B). dnTCF1EWT bound to the same regions at 9 h post-induction as 2 h post-induction, however, binding was weaker at 9 h, despite greater amounts of dnTCF1E in the nucleus and immunoprecipitate at the later time point (Figures 1B and C and 2A and B). dnTCF1Emut bound to most of the same regions as dnTCF1EWT at 2 h post-induction, but binding to these regions was generally weaker, as indicated by fewer total unique reads within 500 bp of the peak centers (269.0 versus 314.2; $P = 4.5E\text{-}16$; paired *t*-test). Binding by dnTCF1Emut was also weaker at 9 h post-induction compared to 2 h post-induction, suggesting that the decrease in binding seen at later time points (even with greater amounts of dnTCF1E in the nucleus) is due to a C-clamp-independent phenomenon. To test whether dnTCF1E was inducing a repressive chromatin state that reduces binding at 9 h post-induction, ChIP-quantitative polymerase chain reaction (qPCR) experiments were performed on H3K9acetyl and H3K9me3 chromatin marks (Supplementary Figure S3A and B). H3K9me3 is associated with a closed chromatin state and decreased access of transcription factor binding (34), while H3K9acetyl is associated with an open chromatin state (35). Induction of dnTCF1EWT did not cause an increase in H3K9me3 (Supplementary Figure S3A) and unexpectedly caused an increase in H3K9acetyl by 9 h (Supplementary Figure S3B). Therefore, it is unlikely that dnTCF1E is reducing its own binding at 9 h post-induction through a chromatin-mediated mechanism. Taken together, dnTCF1EWT shows a global pattern of rapid, strong and focused binding to specific sites.

### The C-clamp binds Helper sites on a genome-wide scale and targets dnTCF1EWT to gene loci

Sites of dnTCF1EWT and dnTCF1Emut occupancy were chiefly within intergenic and intronic regions, a pattern of binding that has been observed for TCF4, a LEF/TCF family member that is co-expressed with TCF1 in colon epithelial cells (Figure 2C) (36). However, dnTCF1EWT bound more than a 5-fold greater number of promoter regions than dnTCF1mut (promoter defined as a 4 kb region centered on the transcription start site (TSS)), and 2-fold greater number of exons, indicating that the C-clamp is important for targeting dnTCF1E to regulatory sites near gene bodies. An example of a promoter-bound region is shown in Figure 2D. We have previously shown that the promoter of the Wnt target gene SP5 is strongly regulated by the C-clamp in a Helper site-dependent manner and therefore as expected, dnTCF1EWT showed strong binding to the SP5 promoter, whereas dnTCF1Emut showed weak, barely significant binding to the promoter (Figure 2D) (15).

Although the C-clamp has been shown to interact with select GC-rich Helper sites in mammals (15) and *Drosophila* (23), its role in the genome-wide binding of C-clamp isoforms of TCFs is unknown. We have previously shown that the human C-clamp interacts with a short Helper site (5′-RCCG-3′) with an unusual degree of flexibility in that the site can be recognized on the 5′ or 3′ side of a WRE with a tolerance for varied spacing between the elements (15). We also determined that the C-clamp recognizes a 7 nucleotide extended Helper site (5′-GCCGCCR-3′), a motif first identified in *Drosophila* as occurring adjacent to WREs for dTCF/pangolin recognition (23). We therefore searched for enrichment of a slightly shorter version of the extended Helper site (5′-RCCGCC-3′) in our ChIP-seq peaks because the C-clamp is highly conserved between humans and *Drosophila* and because the short 4 nucleotide Helper site occurs too frequently for meaningful searches. The ChIP-seq peaks occupied by dnTCF1EWT showed a greater total number of extended Helper sites compared to dnTCF1Emut, confirming that the C-clamp interacts with Helper sites on a genome-wide scale (Figure 3A). In contrast, dnTCF1EWT and dnTCF1Emut showed similar enrichment of the WRE in ChIP-seq peaks (Figure 3A). This suggests that the C-clamp does not interfere with the ability of the HMG DBD to bind to the WRE. To account for the issue that promoters and gene bodies tend to be more GC-rich than intergenic regions, we evaluated the occurrence of the Helper site in peaks within promoter regions (defined as -1 kb to +100 bp from the TSS) for both wild type and mutant as compared with the frequency of the Helper site in all human RNA polymerase II promoters. We used the incidence of (5′-RCCGCC-3′) per base pair to determine if the Helper site occurrence is significantly higher than expected. Despite a higher background GC content within promoters, the incidence of RCCGCC per nucleotide is still significantly greater within promoters bound by dnTCF1EWT (Figure 3B; Mann–Whitney–Wilcoxon test; $P = 2.3E\text{-}12$). For comparison, the incidence of the Helper site in promoters occupied by dnTCF1Emut is not significantly greater than the genome background promoters ($P < 0.05$; Figure 3B). We also observed a greater percentage of dnTCF1EWT ChIP-seq peaks that have at least one occurrence of the Helper site compared to dnTCF1Emut, whereas the percentage of peaks containing a WRE was similar between the two (Figure 3C). Thus, ChIP-seq analysis reveals a genome-wide association between the C-clamp and the Helper site.

For an unbiased analysis we used multiple *de novo* motif enrichment approaches including Regulatory Sequences Analysis Tools (RSAT) (37) and Hypergeometric Optimization of Motif Enrichment (HOMER) (38). These analyses demonstrated that the WRE was the most consistently enriched motif in both dnTCF1EWT and dnTCF1Emut peaks at 2 h, a result consistent with the HMG box being the dominant sequence specific DBD in dnTCF1E (Figure 3D). For the 9 h time point, the WRE continued to be highly enriched (Supplementary Figure S4), but HOMER also detected additional enriched motifs including MEF-2C, LRX, SREBP-1, -2, ATF-1 and CRX (Supplementary

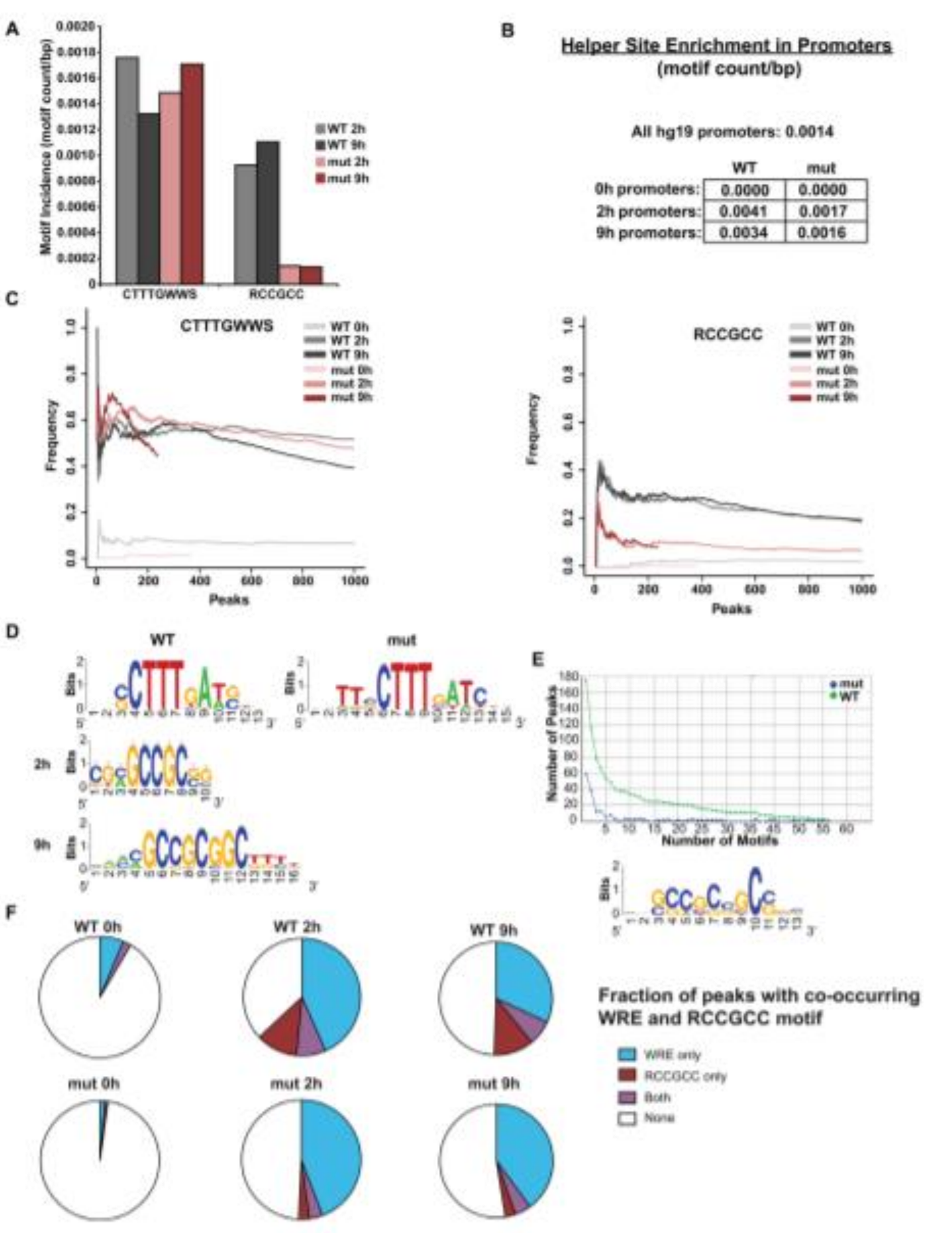


**Figure 3.** The C-clamp binds to Helper sites on a genome-wide scale. (A) The average motif incidence of the WRE (5′-CTTTGWWS-3′) and the Helper (5′-RCCGCC-3′) in the top 1000 ChIP-seq peaks was calculated by counting the incidence of the motif and dividing by the length of the peak. dnTCF1EWT and dnTCF1Emut had similar enrichment of the WRE. However, dnTCF1EWT had significant enrichment of the Helper site compared to dnTCF1Emut at 2 h (Wilcoxon–Mann–Whitney, $P = 8.6E\text{-}23$) and 9 h ($P = 1.9E\text{-}10$). (B) Helper site enrichment in promoters. The Helper site is found on average 1.4 times per 1000 bp in all human promoters (-1000 bp upstream to 100 bp downstream the TSS). Promoter regions bound by dnTCF1EWT had an enrichment of the Helper site over all human promoters ($P = 2.3E\text{-}12$), whereas dnTCF1Emut-bound promoter regions did not ($P = 0.053$). (C) Motif frequency of the WRE (left) and the Helper site (right) in dnTCF1EWT and dnTCF1Emut peaks. dnTCF1EWT and dnTCF1Emut peaks had a similar frequency of the WRE, but the Helper site was found in a greater frequency of dnTCF1EWT peaks. (D) RSAT *de novo* motif enrichment performed on dnTCF1EWT and dnTCF1Emut peaks. The WRE was the top motif found in both dnTCF1EWT and dnTCF1Emut peaks. The Helper site was the fifth most enriched motif at 2 h post-induction and the second most enriched motif at 9 h post-induction. (E) Top graph, the GC-rich motif is enriched in both frequency and copy number of dnTCF1EWT peaks compared to dnTCF1Emut peaks. A full display of *de novo* motif enrichment results is found in Supplementary Figures S4–S6. Bottom panel, the most enriched motif from RSAT differential *de novo* motif analysis which uses dnTCF1Emut ChIP-seq peaks as a background to find motifs that are comparably enriched in dnTCF1EWT peaks. (F) Fraction of the top 1000 ChIP-seq peaks with the indicated motifs.

Figure S5A). Importantly, the Helper site was enriched in dnTCF1EWT peaks at 2 and 9 h post-induction (Figure 3D, Supplementary Figure S4). In contrast, there was no enrichment of Helper-like motifs in dnTCF1Emut peaks (Figure 3D, Supplementary Figures S4–S6) or in the 0 h controls (data not shown). To directly compare the differences between dnTCF1EWT and dnTCF1Emut we used differential *de novo* motif analysis, an approach that searches for motifs enriched in one ChIP-seq data set compared to a second data set. We determined that all motifs preferentially enriched in dnTCF1EWT versus dnTCF1Emut ChIP-seq peaks were GC-rich Helper-containing motifs (Figure 3E, Supplementary Figure S6). Interestingly, peaks with Helper sites very often did not have an identifiable WRE motif (5′-CTTTGWWS-3′; Figure 3F), suggesting that the C-clamp-Helper site interaction frequently occurs independent of canonical WRE recognition. Collectively, these different motif analyses suggest that the primary difference between dnTCF1EWT and dnTCF1Emut genome-wide binding site selection is that dnTCF1EWT utilized the C-clamp to bind regions with an enrichment of the Helper site.

### Co-occurrence of the WRE and Helper site

Previous work has established that the C-clamp binds to Helper-sites as an auxiliary domain reliant on a HMG box-WRE interaction (15,23). We therefore determined if the Helper site was enriched in dnTCF1EWT peaks that also contained a WRE (CTTTGWWS; Table 1). Both dnTCF1EWT and dnTCF1Emut contained an enrichment of peaks with 1 Helper site and 1 WRE over what was expected by chance alone (approximated using a binomial distribution, see Materials and Methods), indicating co-evolution of these two motifs. However, out of the top 1000 peaks, dnTCF1EWT had 27 peaks with 2 Helper sites and 1 WRE ($P = 1.6E\text{-}11$), while dnTCF1Emut had only 4 peaks with 2 Helper sites and 1 WRE ($P = 0.021$). dnTCF1EWT had 18 peaks with 3 or more Helper sites and 1 WRE, whereas dnTCF1Emut had 0 peaks with 3 or more Helper sites and 1 WRE. This suggests that the C-clamp can utilize multiple Helper sites to contribute to binding as has been previously suggested (20). It has also been previously suggested that despite the flexibility in orientation and spacing for the C-clamp-Helper interaction, there is a preferred orientation relative to the WRE and a selection for close proximity to the WRE (usually within 10 bp). However, an analysis of the distance between Helper sites and WREs in dnTCF1EWT ChIP-seq peaks revealed no clear pattern of orientation and distance constraints of the Helper site relative to the WRE (Supplementary Figure S7A). This again suggests that the C-clamp-Helper site interaction may occur independently of a concurrent HMG box-WRE interaction. The role of the Helper site may be to serve as a sink to bring dnTCF1EWT in closer proximity to the more transcriptionally potent WRE.

### The C-clamp-Helper site contributes to ChIP-seq peak strength

We detected 510 genomic regions that were bound both by dnTCF1EWT and dnTCF1Emut at 2 h post-induction. However, even though these regions were occupied by wild-type and mutant protein, dnTCF1EWT generally bound more strongly than dnTCF1Emut (strength defined as enrichment or peak score from MACS; $P = 2.6E\text{-}10$; Figures 2B and 4A). This difference in strength is not due to different protein levels as induction and immunoprecipitation of dnTCF1Emut was equivalent to dnTCF1EWT (Figures 1C and 2A). Motif analysis of the set of peaks bound more strongly by dnTCF1EWT revealed that the short Helper site 5′-RCCG-3′ was significantly enriched (Figure 4B). This enrichment suggests that the C-clamp-Helper site interaction contributes to dnTCF1EWT peak strength, even in regions that are bound by dnTCF1Emut and therefore presumably have strong HMG-WRE interactions. The difference in peak strengths between wild-type and mutant dnTCF1E was even more pronounced when multiple copies of the short Helper site are present (Figure 4C), suggesting that Helper site copy number makes a contribution to peak strength. The short Helper site was found near the center of wild-type but not mutant peaks (Figure 4D), consistent with the C-clamp being able to make contacts with this short motif. As previously discussed, dnTCF1EWT is generally bound to genomic regions more strongly at 2 h than it is at 9 h post-induction (Figures 2B and 4A). Interestingly, the peaks that went against this trend, that is, regions that continued to be strongly occupied by dnTCF1EWT at 9 h were peaks that had significant enrichment of the extended Helper site (Figure 4F). dnTCF1Emut also showed stronger binding at 2 h versus 9 h, but unlike wild-type protein, very few peaks remained strongly occupied at 9 h and these regions were not enriched for Helper sites (Supplementary Figure S7B). These results are consistent with the C-clamp being a DBD of moderate strength (kd = 18 nM) (20), whereas the HMG box is a stronger DBD (kd = 1 nM). At 2 h post-induction, low levels of dnTCF1EWT favor use of the HMG-WRE interaction, although the C-clamp Helper site interaction still makes an important contribution to binding (Figures 2B, 3 and 4). However, at 9 h, greater amounts of dnTCF1EWT in the nucleus allow the C-clamp-Helper site interaction to contribute more to peak strength. Taken together these analyses clearly show that the Helper site contribution to ChIP-seq peak strength is C-clamp specific.

### 4′Thiouridine-seq identifies responsive target genes of dnTCF1EWT

To assess the consequences of dnTCF1EWT binding on changes in gene expression, and to identify the most likely direct, regulated target genes of TCF1E, we used a metabolic labeling and high-throughput RNA sequencing technique called 4′Thiouridine-seq. 4′Thiouridine is a nucleic acid precursor that can be incorporated into nascent transcribed RNA without impairing transcription or translation (30,39). To perform 4′Thiouridine-seq, dnTCF1EWT protein was induced for 2 or 9 h and cells were pulsed with 4′Thiouridine for 30 min, allowing incorporation of the nucleotide analogue into actively transcribed nascent RNA species (Figure 5A). The 30-min labeling time was started at the same time that ChIP-seq samples were formaldehyde cross-linked so that the effects of

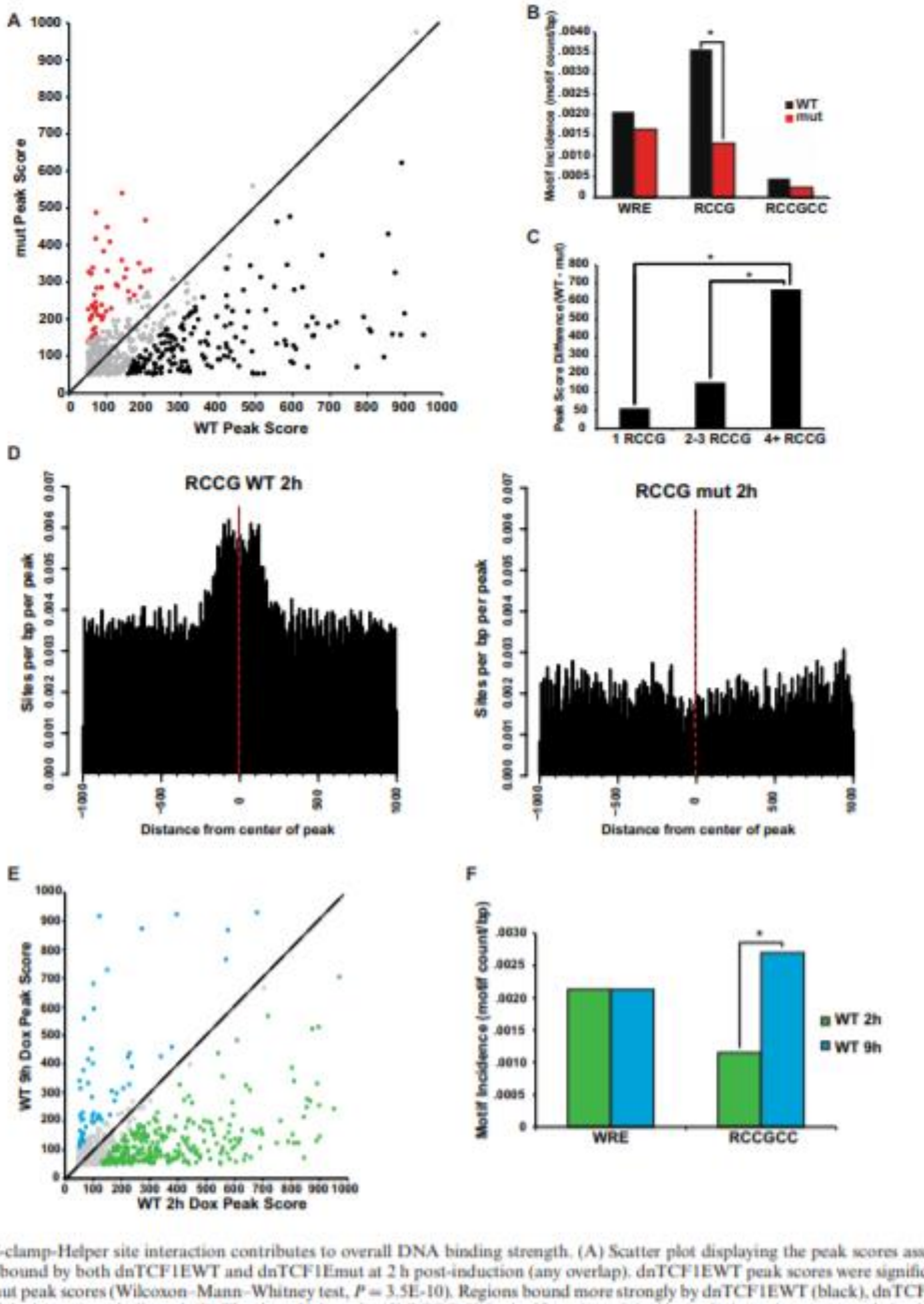


**Figure 4.** The C-clamp-Helper site interaction contributes to overall DNA binding strength. (A) Scatter plot displaying the peak scores associated with genomic regions bound by both dnTCF1EWT and dnTCF1Emut at 2 h post-induction (any overlap). dnTCF1EWT peak scores were significantly higher than dnTCF1Emut peak scores (Wilcoxon–Mann–Whitney test, $P = 3.5E\text{-}10$). Regions bound more strongly by dnTCF1EWT (black), dnTCF1Emut (red) and similarly enriched (gray) are indicated. (B) The short Helper site (5′-RCCG-3′) is significantly enriched in peaks bound more strongly by dnTCF1EWT (Wilcoxon–Mann–Whitney test, $P = 1.3E\text{-}5$), whereas the WRE is not significantly enriched in regions bound more strongly by dnTCF1EWT and dnTCF1Emut ($P = .10$). There were very few Helper sites (5′-RCCGCC″-3) bound by dnTCF1Emut and therefore a significant difference between peaks more strongly bound by wild-type and mutant was not seen ($P = 0.97$). (C) Increased copy number of the short Helper site is associated with greater differences in peak scores between dnTCF1EWT and dnTCF1Emut (1 RCCG compared to 4+ RCCG: $P = 1.0E\text{-}7$, 2–3 RCCG compared to 4+RCCG: $P = 8.0E\text{-}6$). (D) Histogram showing that the short Helper site is enriched near the center of dnTCF1EWT but not dnTCF1Emut peaks. (E) Scatter plot displaying peak scores associated with genomic regions bound by dnTCF1EWT at both 2 and 9 h of induction. Peak scores were significantly higher for dnTCF1EWT at 2 h post-induction (Wilcoxon–Mann–Whitney, $P = 1.7E\text{-}15$). (F) Regions bound more strongly by dnTCF1EWT at 9 h have an enrichment of the Helper site ($P = 0.0008$), but not the WRE (Wilcoxon–Mann–Whitney test, $P = 0.97$).

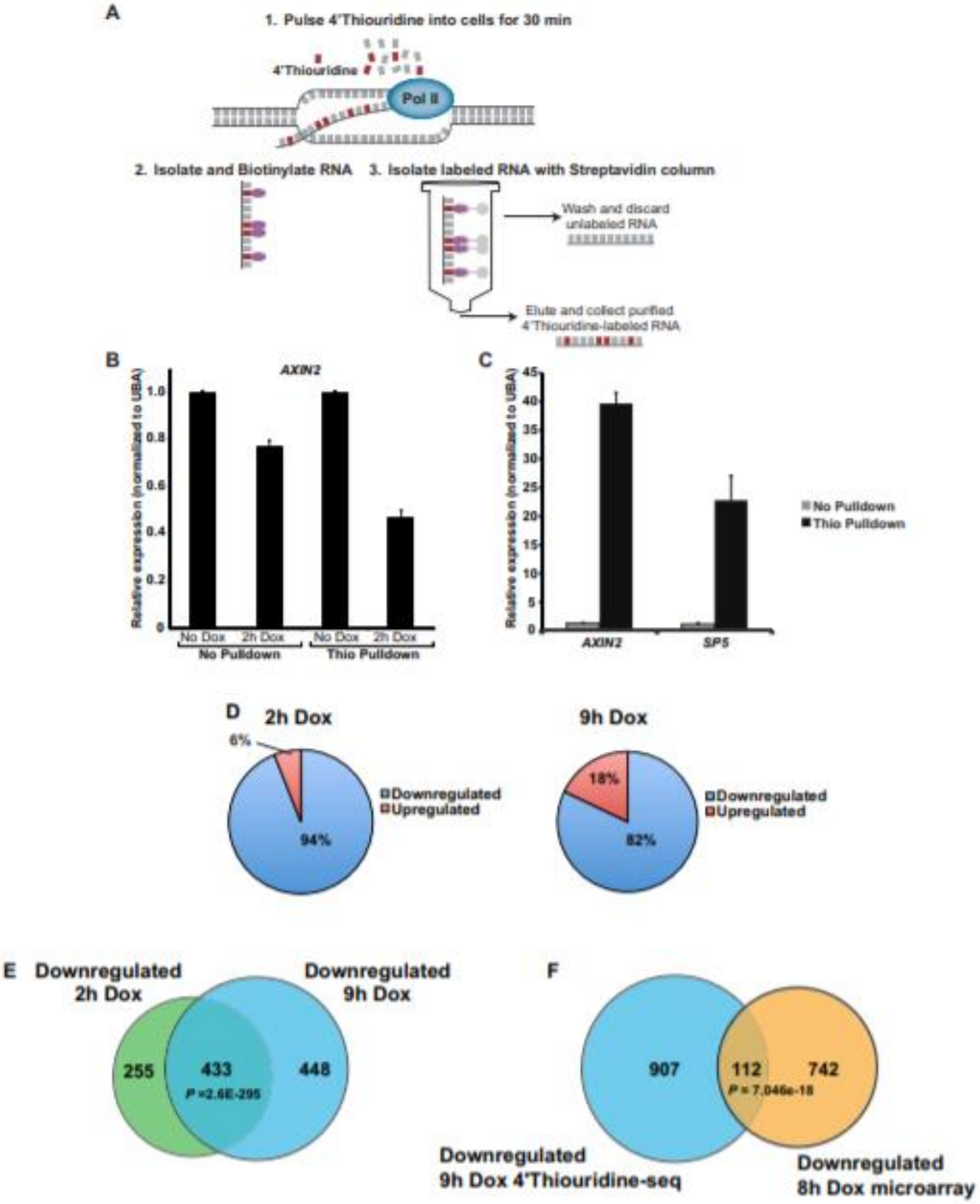


**Figure 5.** 4'Thiouridine-seq identifies genes repressed by dnTCF1EWT. (A) 4'Thiouridine-seq involves addition of 4'Thiouridine into cells for 30 min to label nascent RNA transcripts. 4'Thiouridine-labeled RNA is isolated and sequenced, allowing for snapshots of transcription at any given time point of dnTCF1EWT induction. (B) Comparison of AXIN2 levels as assessed by normal RT-PCR (before pulldown) versus 4'Thiouridine-RT-PCR (same sample after pulldown), $n = 3$. Pulldown of 4'Thiouridine-incorporated transcripts causes a more dramatic decrease in AXIN2 in response to dnTCF1E WT induction to be detected ($n = 3$). (C) Wnt target genes are greatly enriched after 4'Thiouridine pulldown when compared to the housekeeping gene UBA as assessed by RT-PCR. (D) The majority of genes that changed expression after 2 h of dnTCF1EWT induction ($P < 0.02$) and 9 h of dnTCF1EWT induction ($P < 0.012$) were downregulated. (E) The overlap of genes downregulated at 2 h post-induction and 9 h post-induction was highly significant (*P*-value denotes the result of a hypergeometric test). (F) The overlap of genes downregulated at 9 h post-induction in the 4'Thiouridine-seq and at 8 h post-induction in a previous microarray experiment in a different clonal DLD-1 cell line was highly significant (*P*-value denotes the result of a hypergeometric test).

**Table 1.** Observed and expected number of the top 1000 peaks that contain multiple number of Helper sites and at least one WRE

| | Number of peaks with 1+ WRE | Helper sites per peak (*N*) | Number of peaks with N+ Helper sites | Expected number of peaks with both[a] | Observed number of peaks with both | *P*-value |
|---|---|---|---|---|---|---|
| **WT** | 506 | 1 | 193 | 33 | 80 | 2.6E-12 |
| **WT** | 506 | 2 | 185 | 4 | 27 | 1.6E-11 |
| **WT** | 506 | 3 | 37 | 0 | 11 | 3.3E-07 |
| **WT** | 506 | 4 | 27 | 0 | 4 | 0.021 |
| **WT** | 506 | 5 | 21 | 0 | 3 | 0.081 |
| **mut** | 423 | 1 | 54 | 1 | 32 | 6.6E-29 |
| **mut** | 423 | 2 | 7 | 0 | 4 | 0.020 |
| **mut** | 423 | 3 | 1 | 0 | 0 | - |
| **mut** | 423 | 4 | 1 | 0 | 0 | - |
| **mut** | 423 | 5 | 0 | 0 | 0 | - |

[a] See Materials and Methods.

binding on changes in gene expression could be directly correlated. Like with ChIP-seq, 4′Thiouridine-seq was performed in duplicate, and a control was included (no doxycycline). Labeled RNA was selectively purified from cell lysates, allowing for a snapshot of transcription to be taken at the 2-h time point. A distinct advantage of 4′Thiouridine-seq is the ability to detect rapid changes in transcription even if the mRNA species are stable or overwhelmingly abundant. For example, in the case of the well established Wnt target gene AXIN2, total AXIN2 mRNA levels show a modest decrease in expression after 2 h of dnTCF1E induction (Figure 5B). This is due to a lag time between changes in transcription rate and changes in total transcript levels, with RNA stability contributing to the lag time before the full magnitude of transcriptional suppression is detected. However, since the 4′Thiouridine pulldown detects nascent RNA species, it serves as a direct measure of RNA polymerase II activity during the pulse period, and in the case of AXIN2, reports a more responsive and pronounced decrease in AXIN2 expression (Figure 5B). This confirms the suitability of the 4′Thiouridine labeling and isolation scheme for RNA-seq and in particular for the matched 2 h ChIP-seq condition because stronger decreases in gene expression are more easily detected by RNA-seq. A caveat of the 4′Thiouridine-based pulldown method is that actively transcribed genes (high density of RNA polymerase and nascent RNA) will show greater overall enrichment than genes that show low transcription rates (low number of transcribing polymerases). For example, the Wnt target genes AXIN2 and SP5 are presumed to be highly transcribed in DLD-1 cells due to Wnt-activating mutations in the APC gene. Consistent with this, 4′Thiouridine labeling and pulldown caused a greater enrichment of these transcripts relative to the housekeeping gene UBA, which is not a Wnt target gene (Figure 5C). Induction of dnTCF1EWT caused a decrease in the transcription of these known Wnt target genes (AXIN2, SP5 and TNFRSF19) as assessed by 4′Thiouridine-reverse transcriptase-PCR (RT-PCR) (Supplementary Figure S8A), and transcription rates continued to decrease after 9 h of induction. Induction of dnTCF1Emut caused a similar initial decrease in transcription of AXIN2, but this was quickly followed by a recovery at 9 h post-induction (Supplementary Figure S8A). By contrast, induction of dnTCF1Emut did not significantly decrease SP5 transcription, as predicted from the weaker ChIP-seq occupancy of the SP5 promoter and our previous finding that SP5 is a C-clamp-dependent Wnt target gene (30) (Supplementary Figure S8B). Interestingly, there was still compensation in transcription rate, even an overcompensation, and as for AXIN2, this recovery was evident at the 9 h time point. These data demonstrate that the strong versus weak binding patterns of dnTCF1EWT and dnTCF1Emut connect to strong and weak effects on transcription.

### Dynamic patterns of occupancy at Wnt target genes

Purified 4′Thiouridine-labeled RNA was submitted for RNA-seq for a global analysis of effects on transcription. The great majority (94%) of the 733 genes that changed expression after 2 h of induction of dnTCF1EWT showed a decrease in transcription rate (Figure 5D), confirming that dnTCF1E is a transcriptional repressor. There was also a high degree of concordance between the two time points, as 433 genes that were downregulated at 2 h continued to be repressed at 9 h (Figure 5E), including many previously validated Wnt target genes (Table 2). Additionally, a significant overlap was observed between the 9 h downregulated genes and a previous microarray experiment at 8 h post-induction (Figure 5F) (15). Many of the ChIP-seq peaks associated with known Wnt targets were located more than 100 kb from the TSS, illustrating the difficulty in making conclusions about distal peaks and their involvement in the regulation of Wnt target gene expression. Nearly all of the known Wnt target genes that were downregulated contained at least one annotated peak from the 2 and 9 h ChIP-seq analysis (Table 2). Interestingly, a close look at Table 2 reveals that in some cases different binding sites were utilized at 2 and 9 h post-induction. For example, for AXIN2, 3 binding sites were utilized at 2 h post-induction, whereas 9 binding sites were utilized at 9 h post-induction. These 9 binding sites included the 3 that were bound at 2 h post-induction. In the case of SGK1 one binding site was located 12 239 bp from the TSS at 2 h, but at 9 h, three different sites were occupied at 16 390, 17 705 and 31 467 bp from the TSS suggesting frequent and active use of distal enhancers. This suggests that Wnt target genes contain redundant TCF binding sites and that different binding sites may be utilized depending on the chromatin conformation or the concentration of TCF.

**Table 2.** Known Wnt target genes that were downregulated and annotated ChIP-seq peaks

| Wnt target gene | Fold change: 2 h, 9 h | 2 h peak distance from TSS (bp) | 9 h peak distance from TSS (bp) |
|---|---|---|---|
| *ASCL2* | -2.4, -3.0 | - | - |
| *AXIN2* | -1.7, -2.8 | 483, 4234, 127 488 | -122 214, -31 782, -4127, 550, 4401, 7366, 64 102, 127 472, 167 808 |
| *BMP4* | -1.6, -2.6 | -215 569, -152 132, -145 446 | -42 955 |
| *CDX2* | -1.9, -2.6 | 2396, 3221 | 2362, 3257 |
| *DKK1* | -5.5, -6.7 | 39 631, 138 559 | - |
| *FZD7* | -2.4, -2.6 | -40 047, 34 670, 35 777 | 34 525, 93 752 |
| *JAG1* | -1.4, -1.9 | -32154 | -45 733, -34 836 |
| *LGR5* | N/A, -1.8 | 931 | 990, 29 670, 54 199 |
| *MYC* | -2.4, -3.1 | 7022 | -104 616 |
| *PITX2* | -1.5, -1.7 | -284 838, -219 399, -183 246 | -422 541, -419 957, -200 616 |
| *PPIF* | -1.4, -1.7 | - | - |
| *SGK1* | -2.2, -2.5 | 12 239 | 16 390, 17 705, 31 467 |
| *SOX4* | -2.4, -3.1 | -208 433, -31 951, 11 965 | -527 001, -174 480, -90 544, -37 651, 1316 |
| *SOX9* | -1.6, -1.9 | -160 304 | 746 421, -602 259, -336 893, -258 580, -199 838, -11 372 |
| *SP5* | -2.2, -2.9 | -176 | -162 |
| *TBX3* | -1.8, -1.8 | -693 657, -503 686, -97 493, -84 395, 27 773, 97 991 | -315 734, -283 946, -138 493, -137 130 |
| *TCF4* | -1.3, -2.2 | 31 461, 134 605, 159 291 | 3716, 24 346, 31 485 |

Column titles

We assessed the overall degree in which transcription was repressed for genes downregulated at both time points. There was a significant decrease in relative transcription at 9 h versus 2 h post-induction (Supplementary Figure S9A). This same pattern was seen with Wnt target genes (Supplementary Figure S9B) and is consistent with the 4'Thio-RT-PCR results (Supplementary Figure S8). Interestingly, this trend is directly opposite from the changes in dnTCF1E peak strength (peaks were strongest at 2 h post-induction and declining by 9 h). To investigate this seemingly contradictory result, we counted the number of peaks within 50 kb of the TSS of downregulated genes for the two ChIP-seq time points (Supplementary Figure S9C). We hypothesized that decreased transcription at the 9 h time point was due to a greater number of occupied sites surrounding the gene. This was confirmed for one-third of the downregulated genes where ChIP-seq peaks were easily linked to nearby target genes. This analysis does not take into account the actions of even longer range enhancers that might be involved in the downregulation of additional genes. It is therefore possible that the stable repression of transcription derives from increased involvement of multiple regulatory sites. Additionally, there may be a lag between the initial protein binding event at 2 h post-induction and the greater changes in gene expression seen at 9 h post-induction.

The degree to which dnTCF1EWT binding sites were linked to downregulated genes was assessed by comparing the ChIP-seq and 4'Thiouridine-seq datasets (Figure 6A). dnTCF1EWT ChIP-seq peaks were linked to 85 downregulated gene loci at 2 h post-induction (Figure 6A). In contrast, the background ChIP-seq peaks from untreated cells were associated with only seven downregulated genes. This demonstrates that dnTCF1EWT is likely involved in the direct downregulation of the majority of the 85 downregulated genes associated with a ChIP-seq peak at 2 h post-induction. Half of the peaks occupied by dnTCF-1EWT are more than 30 kb away from the TSS, suggesting that there may be widespread use of long-range chromatin loops by TCF/ß-catenin complexes (Figure 6A). Forty-two of the downregulated genes contained a ChIP-seq peak within 30 kb of their TSS (Table 3), and most of these are not currently known as Wnt target genes. We also assessed the peak distribution of peak scores of the peaks associated with the 42 downregulated genes, and performed gene set enrichment analysis (GSEA) and found a significant enrichment of strong peak scores for the peaks within 30 kb of these genes as compared to the max peak score for all genes with at least one peak within 30 kb of the TSS (Figure 6B). These represent new possible direct Wnt target genes, an illustration of the distinct advantage of combining ChIP-seq with 4'Thio-seq.

To test whether the ChIP-seq peaks from these potential new target genes are functional and capable of conferring β-catenin transcription regulation, we used an established transient transfection assay in COS-1 cells where transiently expressed TCF1E/β-catenin complexes specifically drive luciferase reporter gene expression when *bona fide* WREs are present in the reporter plasmid (Figure 6C; (15)). For our validation test, we subcloned nine dnTCF1EWT peaks into luciferase reporter plasmids: three regions are located within introns (GADD45B, AXIN2, TGIF), one is in an exon (HIST2H2AC), one is located at a transcription termination site (HIST2H4B), one is in an intergenic region (BAMBI) and three are within basal promoters that encompass the TSS (HIST1H4K, TMEM177, ZBED5). All nine of these peaks were specifically occupied by dnTCF1EWT, not the mutant protein, hinting that these could be new C-clamp-dependent regulatory sites. We observed that full-length TCF1EWT/β-catenin activated luciferase expression from 4 of the 9 regions (those associated with GADD45B, AXIN2, TGIF and HIST2H4B) while TCF1Emut/β-catenin could not (Figure 6C; see Supplementary Figure S10 for nucleotide sequences of the four, regulated regions). In contrast, TCF1EWT/β-catenin

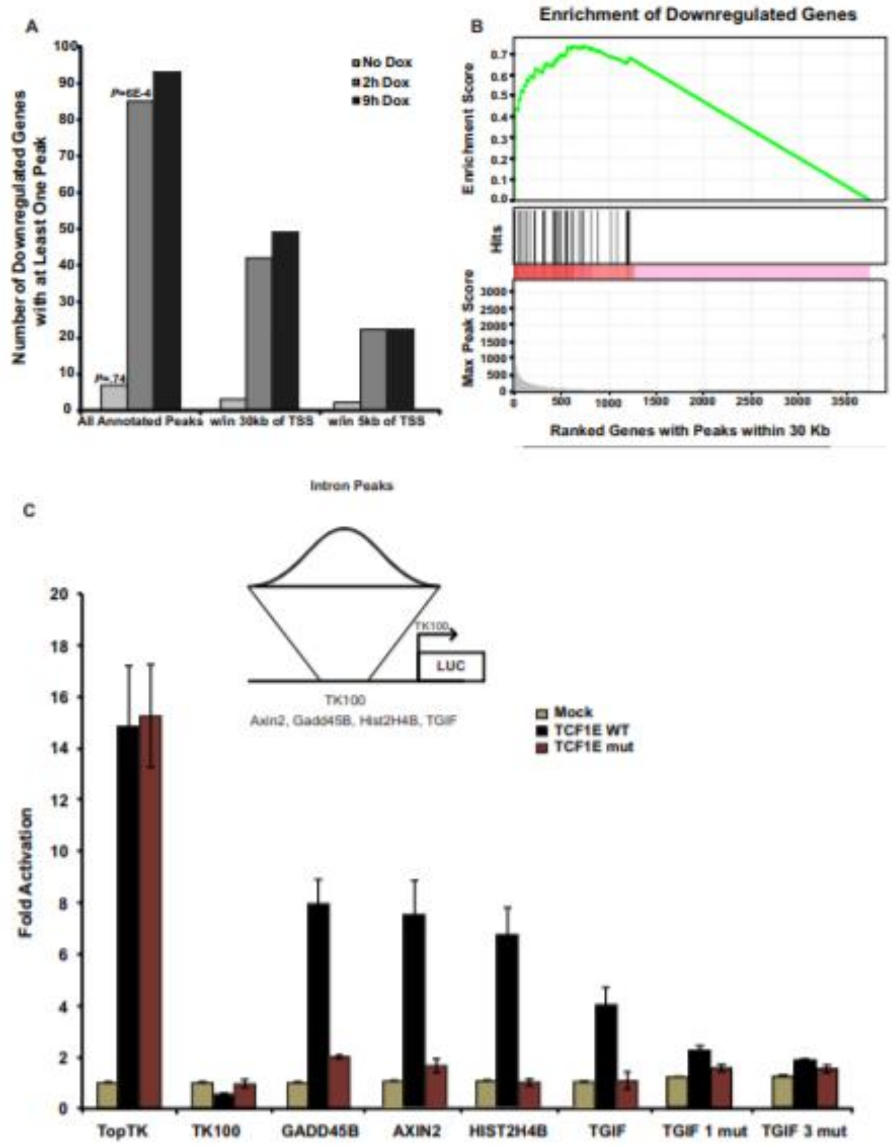


**Figure 6.** Identification of new Wnt target genes. (A) Downregulated genes with an annotated ChIP-seq peak. Note that 733 genes that had an expression change after 2 h of dnTCF1EWT induction ($P < 0.02$) were assessed for the presence of an annotated peak from the pooled analysis (peak score >67). A significant number of genes contain at least one annotated peak from the 2 h ChIP-seq time point, but not the 0 h (untreated) ChIP-seq sample. There are also many more genes that have a peak associated with the 9 h ChIP-seq than the 0 h ChIP-seq. (B) GSEA was performed on the 42 genes with at least one peak within 30 kb of the TSS as compared to the set of all genes with at least one peak within 30 kb of the TSS at 2 h post-induction and a significant enrichment ($P < 0.05$) of high scoring peaks was found. (C) Wild-type bound regions from part (A) were cloned into the Thymidine Kinase 100 (TK100) luciferase reporter. Luciferase assays were carried out in COS-1 cells with transient transfection of expression plasmids for β-catenin and full-length TCF1EWT, or TCF1Emut. Mock refers to co-transfection of an empty expression plasmid to control for β-catenin expression. Three biological replicates were included for each luciferase construct tested and luciferase/light unit values were normalized to β-galactosidase levels before comparing to normalized Mock values as described in (42). The TopTK reporter, which contains three multimerized WREs without Helper sites, was activated equally by TCF1EWT/β-catenin and TCF1Emut/β-catenin. The TK100 backbone was not activated by TCF1EWT/β-catenin or TCF1Emut/β-catenin. Helper site-containing ChIP-seq regions with one WRE (GADD45B, AXIN2) conferred activation of the reporter by TCF1EWT but not TCF1Emut. Helper site-containing regions that do not have an identifiable match to a degenerate, shorter WRE (5′-CTTTGW-3′; TGIF, HIST2H4B) also enabled activation by TCF1EWT but not TCF1Emut. Mutation of the three Helper site sequences in the TGIF region (TGIF1mut; TGIF3mut) eliminated regulation (sequences for TOPtk, GADD45B, AXIN2, HIST2H4B and TGIF and its corresponding Helper site mutations are presented in Supplementary Figure S10).

**Table 3.** Proposed[a] (bold) and known Wnt target genes

| Gene Symbol | WRE (CTTTGWWS) | RCCG | Helper (RCCGCC) | Distance to TSS (bp) |
|---|---|---|---|---|
| ***ASPN*** | 0 | 1 | 0 | -14987 |
| ***ASPN*** | 0 | 1 | 0 | -12286 |
| *AXIN2* | 1 | 14 | 3 | 483 |
| ***BTG1*** | 0 | 4 | 0 | -14715 |
| *CDX2* | 0 | 12 | 1 | 2396 |
| ***CEBPD*** | 0 | 3 | 0 | -27765 |
| ***CXCL5*** | 2 | 0 | 0 | -6836 |
| ***DUSP5*** | 0 | 0 | 0 | -13975 |
| ***DUSP6*** | 1 | 0 | 0 | -18481 |
| ***FAM198B*** | 0 | 1 | 1 | 16517 |
| ***HIST1H2BK*** | 0 | 3 | 0 | 232 |
| ***HIST1H2BO*** | 0 | 3 | 0 | 268 |
| ***HIST1H4J*** | 0 | 7 | 1 | -111 |
| ***HIST1H4K*** | 0 | 7 | 1 | -125 |
| ***HIST3H2A*** | 0 | 1 | 0 | 9194 |
| ***HMGCR*** | 1 | 0 | 0 | -11577 |
| ***HOXA9*** | 0 | 4 | 0 | -136 |
| ***IER3*** | 1 | 2 | 0 | -25238 |
| ***IER5*** | 1 | 2 | 0 | 15769 |
| ***LOC729678*** | 7 | 5 | 0 | -1620 |
| ***LOC84931*** | 0 | 5 | 0 | 130 |
| ***MIR3143*** | 0 | 1 | 0 | 28366 |
| *MYC* | 1 | 1 | 1 | 7022 |
| ***NEDD9*** | 1 | 0 | 0 | -22066 |
| ***PIM1*** | 1 | 3 | 0 | 2023 |
| ***PKDCC*** | 0 | 0 | 0 | 15835 |
| ***PLK2*** | 1 | 3 | 0 | 1062 |
| ***PMAIP1*** | 0 | 1 | 0 | 23456 |
| ***RHOB*** | 2 | 0 | 0 | 3365 |
| ***SESTD1*** | 0 | 0 | 0 | 29682 |
| ***SMAD7*** | 2 | 1 | 0 | -1573 |
| *SP5* | 3 | 18 | 5 | -176 |
| *TBX3* | 0 | 2 | 0 | 27773 |
| ***TERC*** | 0 | 2 | 0 | 1951 |
| ***TGIF1*** | 0 | 10 | 3 | 820 |
| ***TMEM177*** | 5 | 9 | 1 | -564 |
| ***TNFAIP3*** | 1 | 0 | 0 | -29256 |
| ***WDR74*** | 0 | 6 | 1 | -1561 |
| ***ZBED5*** | 0 | 7 | 4 | 65 |
| ***ZBTB10*** | 0 | 0 | 0 | 1684 |
| ***ZIC2*** | 0 | 12 | 2 | -519 |

[a] Genes that were downregulated at both 2 and 9 hours and contain a ChIP-seq peak within 30kb of the TSS at 2 hours post-induction (see Figure 6A).

and TCF1Emut/β-catenin equally activated a Wnt reporter plasmid that contains multimerized copies of the WRE, but no Helper sites (see TOPTK in Figure 6C). The remaining five regions were either non-responsive, or as independent promoters, they were not active suggesting that the subcloning disrupted either the full regulatory region or chromosomal context. Similar results were recently reported where 50% of high confidence ChiP-seq peaks for β-catenin conferred Wnt regulation in a transient transfection assay (31). The positive results are also consistent with a microarray experiment that we previously reported showing that two of these four genes (GADD45B and TGIF) were regulated in a C-clamp-dependent manner (15). HIST2H4B mRNA was not detected in the microarray, likely because mature histone mRNA is not polyadenylated and is not commonly enriched for library construction and microarray analysis. Our data therefore highlights a potential new class of Wnt target gene, as the HIST2H4B region was as responsive to TCF1E/β-catenin as regions from well known, validated Wnt targets. AXIN2 is one of the well known Wnt target genes and our published microarray analysis had identified it as independent of C-clamp regulation (15). However, since there are multiple ChIP-seq peaks associated with AXIN2, it is likely that one or more of these peaks are C-clamp-dependent including the region analyzed in this study. Most surprising was the fact that two of the subcloned regions (those associated with TGIF and HIST2H4B) contain only Helper sites and no significant match to a canonical WRE motif. To exclude the possibility that WRE sites were missed by the use of a too stringent canonical motif in our sequence search (CTTTGWWS), we used the exact WRE motif developed using the WT 2 h peak data to generate a positional weight matrix with more sequence flexibility (Supplementary Figure S5C). Using this more flexible motif we repeated the search and identified a second WRE site in the AXIN2 insert. However, we found no additional matches to the more degenerate WRE in TGIF or HIST2H4B sequences. Instead, two sequences with key, important mismatches to the canonical WRE were present. Since both of these regions drove C-clamp-dependent activation of reporter gene expression, we hypothesized that the C-clamp-Helper site interactions are critical transcription regulatory motifs in regions that do not contain identifiable matches to WREs. To test this hypothesis, we introduced site-specific mutations in the three Helper sites in the TGIF peak (see Supplementary Figure S10 for the sequence and mutations). Mutation of one Helper site greatly reduced regulation β-catenin, and mutation of all three Helper sites eliminated regulation (Figure 6C).

To address whether Helper sites are sufficient to regulate transcription on a global level, we divided peaks within 30 kb of the TSS of genes detected in 4'Thiouridine-seq into several categories: peaks with no WREs or Helper sites, those with WREs, those with Helper sites and those with both Helper sites and WREs. As expected, peaks with one or more WREs were significantly associated with downregulated genes compared with peaks with no WREs (Supplementary Figure S11). In contrast, peaks with Helper sites alone were not significantly associated with downregulated genes, suggesting that WRE-independent Helper site-mediated transcriptional regulation is likely a gene-specific rather than a globalized mode of regulation. However, peaks with two or more Helper sites and a WRE were significantly associated with greater downregulation of genes compared with peaks with only WREs. Taken together, our results indicate that the C-clamp Helper site interaction synergizes with the HMG-WRE interaction for strong transcriptional control of target genes and in special cases, the C-clamp can participate in direct transcriptional regulation independent of a WRE.

## DISCUSSION

We have used ChIP-seq experiments with dnTCF1EWT and dnTCF1Emut to discover that the C-clamp-Helper site interaction plays an important role in dnTCF1E binding and target gene regulation across the genome. Several of the most notable effects of the C-clamp were its role in the targeting of dnTCF1E to promoters of polymerase II genes, the rapid and stable association of dnTCF1E with those genomic locations and the significant variation in the relative positioning of the WRE and Helper sites. The Helper site was significantly enriched in dnTCF1E peaks, but the number, orientation and the distance of those sites relative to WRE motifs were highly variable. These observations connect to *in vitro* selection experiments with oligonucleotides where the relative positions of WRE and Helper site were also variable (15). The use of 4'Thiouridine-seq confirmed that the C-clamp-Helper site interaction plays an important role in transcriptional regulation of target genes, especially when a WRE is present in the vicinity (within the same ChIP-seq peak) of the Helper site (Supplementary Figure S11). dnTCF1EWT showed stronger binding than dnTCF1Emut at most co-occupied genomic regions (Figure 4A). This occurred at many regions without an obvious Helper site, suggesting that the C-clamp strengthens the overall binding of dnTCF1EWT to DNA, or at least its interaction with chromatin. dnTCF1EWT also bound to many more promoters and gene loci than dnTCF1Emut (more than 3-fold; Figure 2C). Therefore, there seems to be at least two contributions of the C-clamp. One is to recruit dnTCF1E to CpG islands that contain copies of the Helper site and another is to increase the overall affinity of dnTCF1E for DNA.

### The C-clamp connects to Helper sites genome-wide

While the C-clamp exhibits specific DNA binding *in vitro*, our analysis is the first to test whether this specificity holds for binding to sites genome-wide in living cells. Our data demonstrate a strong correlation between the C-clamp and Helper sites. First, Helper sites were present in 30% of the top 500 strongest dnTCF1EWT peaks but only 10% of the top 500 dnTCF1Emut peaks (Figure 3C). Second, the number of Helper sites per region was much greater in dnTCF1EWT peaks than dnTCF1Emut peaks. This suggests that the C-clamp utilizes multiple copies of the Helper site for binding (Figure 3A and E), an observation consistent with other studies reporting that clusters of Helper in the vicinity of WREs was an effective criteria for *in silico* searches for Wnt target genes (22,23). Third, *de novo* motif enrichment revealed that the Helper site was specific for dnTCF1EWT peaks and not dnTCF1Emut peaks (Figure 3D and E, Supplementary Figures S4–S6). Finally, the presence of the short Helper site (5'-RCCG-3') in regions co-occupied by dnTCF1EWT and dnTCF1Emut was associated with stronger binding by dnTCF1EWT (Figure 4B and C). Co-occurrence statistics suggest that there is significant co-enrichment of the WRE (5'-CTTTGWWS-3') and Helper sites (5'- RCCGCC-3') in dnTCF1EWT-bound regions, especially in those regions with multiple copies of the Helper site (Table 1). Co-occurrence of WREs and Helper sites is conserved (Supplemental Figure S7A), suggesting that the 10% of dnTCF1Emut peaks that contain both elements might derive from the co-evolution of strong WREs with Helper sites.

Motif analysis also identified an unexpected feature, which was the functional targeting of TCFs to regulatory regions without a canonical WRE. Greater than half of the dnTCF1EWT peaks with Helper sites did not contain a match to a canonical WRE, even with sequence-degenerate flexibility at the last three positions of the WRE motif (5'-CTTTGWWS-3'; Figure 3F, Supplementary File S1). This continued to be the case when the criteria was further relaxed (5'-CTTTGW-3'; Figure 6C). This was surprising given the initial reports that identified functional associations of the Helper site with the WRE in several target genes (20,22,23). A purified C-clamp protein fragment has been shown to bind independently and specifically to the Helper site *in vitro*, suggesting that the C-clamp can function as an autonomous sequence specific DBD (23,24). However, functional experiments with *Drosophila* TCF (which always contains a C-clamp) showed that a luciferase reporter plasmid with multimerized Helper sites was not responsive, whereas the Helper site greatly augmented activation of a reporter with multimerized WRE motifs (23). Nevertheless, our combined ChIP-seq/4'Thiouridine-seq analysis identified 27 candidate C-clamp-dependent genes that contain a ChIP-seq peak within 30 kb of the TSS and were downregulated selectively by dnTCF1EWT. The occupied peaks surrounding most of these genes do not contain obvious canonical WREs, whereas they have at least one or more Helper sites (Supplementary File S3). Our luciferase validation experiments demonstrated that at least two of these peaks (from the TGIF and HIST2H4B genes) conferred transcriptional regulation by full-length TCF1E/β-catenin complexes (Figure 6C). These results suggest that the C-clamp-Helper site interaction enables biologically relevant regulation of target genes in the absence of an identifiable WRE. Therefore, the difference between multimerized Helper sites and a native region of the genome in conferring C-clamp-dependent regulation may have to do with the fact that native binding sites in the context of chromatin convey additional regulatory information. For example, it is important to note that the absence of an obvious WRE does not mean that the HMG DBD is not participating in binding to other native sequences. As a class of DBD, HMG boxes are flexible in sequence recognition, showing at most a 40-fold difference between specific and non-specific binding (40) and the ability to recognize bent DNA structures. This relaxed DNA binding property implies that the HMG DBD is likely to be a constant contributor to DNA binding, just not always to recognizable WREs. Interestingly, *de novo* motif enrichment analysis of the ChIP-seq peaks shows that Helper sites are often flanked by A-T tracks, a sequence prone to bending and known to be recognized by HMG box DBDs (Supplementary Figure S6) (41). Therefore, the HMG box and the C-clamp likely can bind coordinately to DNA, but the presence of the C-clamp may enable HMG recognition of highly degenerate sequences possibly including T-tracts that do not function as independent, functional WREs on their own (15). An alternative explanation is that the C-clamp engages in protein–protein interactions for recruitment to some of the sites that do not have WREs (42). Our data do not rule out this possibility, however, given that purified C-clamp protein fragments bind specifically to the same Helper site sequence motif that we find enriched in ChIP-seq peaks, we propose that we have identified a non-WRE motif responsible for some of the direct genome-wide binding patterns of TCF. This has implications for the discovery of new Wnt target genes, which are often determined partly through searches for high scoring canonical WREs in regulatory regions.

### TCFs exhibit dynamic binding to target sites

Our analysis also identified a dynamic pattern of genome occupancy where both dnTCF1EWT and dnTCF1Emut binding decreased at 9 h post-induction (Figure 2B). Given that all transcription factors exhibit a dynamic on/off rate of binding to target sites, we hypothesized that dnTCF1E might induce a change in chromatin conformation which restricts binding at later time points. However, ChIP-qPCR analysis of repressive chromatin marks do not support this hypothesis (Supplementary Figure S3A and B). Another possible source for this dynamic shift could be a compensatory response where increased DNA binding of other LEF/TCF family members competes with the induced dnTCF1E for binding at the later time points. While this is certainly possible, western blot analysis does not show any increases in the expression of other LEF/TCF family members after induction of dnTCF1E (data not shown). Another possible source for this phenomenon could derive from the experimental design where the increased number of genomic binding sites at later time points (∼5700 sites at 2 h versus ∼13 000 sites at 9 h) means that the early binding sites are proportionally less represented in the sequencing analysis of the later 9 h time point. However, it is unclear why this result is also evident in our ChIP-qPCR validation experiments because qPCR utilizes primers to survey specific enriched regions (Supplementary Figure 3C–F). Another possible source for the decreased binding at 9 h is post-translational modification of dnTCF1E. For example, TAK/NLK proteins have been shown to phosphorylate LEF/TCFs, causing their disassociation from the DNA template (43). Therefore, the decrease in binding at later time points could be due to a stress response, where high levels of dnTCF1E cause the upregulation of a kinase or other regulatory partner that is capable of removing dnTCF1E from the DNA template.

In addition to the dynamic temporal shifts of occupancy, the DNA sequence of the Helper site suggests another way in which TCF binding can be modulated. Embedded in each DNA strand of the Helper site sequence is one CpG dinucleotide, whereas the WRE does not have any CpG pairs. It is possible that the portion of the Wnt transcriptome that depends on C-clamp activity might be rendered inaccessible through DNA methylation. A preliminary test of this notion was carried out by an EMSA of TCF-1EWT binding to an oligonucleotide probe encoding two WREs and two Helper sites (Supplementary Figure S12). A mutation in the Helper site reduced binding to the oligonucleotide, but not as much as a version of the EMSA probe that contained a methylated CpG dinucleotide in both strands of the Helper site. Binding was strongly reduced even though the consensus, canonical WRE remained the same in all the probes. Thus, targeted methylation of Helper sites in the genome might be one way in which the Wnt transcriptome can be shaped by epigenetic modification.

### Linking ChIP-seq and 4'Thio-seq identifies direct target genes

4'Thiouridine-seq was a useful technique for identifying immediate decreases in transcription in response to induction of dnTCF1EWT. Combining this data with our ChIP-seq results enabled us to link hundreds of dnTCF1EWT occupied regions to stably downregulated genes and therefore prompted predictions of additional direct target genes (Supplementary File S4, and also see Table 3). 4'Thiouridine-seq was particularly well suited for detecting the consequence of dnTCF1EWT binding at 2 h post-induction when induced protein was barely detectable (Figure 6A). Many genes were downregulated by dnTCF1EWT at both 2 and 9 h post-induction including many known Wnt target genes (Table 2). However, genes were downregulated more strongly at 9 h post-induction than 2 h (Supplementary Figure S9A and B), a surprising result given that peaks of dnTCF1EWT occupancy were generally strongest at the 2 h time point. The reason for the apparent discrepancy between occupancy and transcription is not fully understood, however, we note that the number of distinct genomic regions occupied by dnTCF1EWT that are within a distance cut-off of 50 kb relative to target genes increases at 9 h post-induction, a pattern likely due to increased concentration of dnTCF1EWT protein in the nucleus (Figure 1C and Supplementary Figure S9C). Thus, Wnt target genes might be regulated from multiple redundant binding sites with the concentration of LEF/TCFs playing a role in which sites will be utilized for regulation.

The great majority of the downregulated genes are not known Wnt target genes. This is even true after 2 h post-induction when indirect effects have not had as much time to accumulate and therefore direct gene expression changes should be the most enriched. The identification of downregulated genes that contain closely linked ChIP-seq peaks at 2 h post-induction is strong evidence that our analysis identified new Wnt target genes (see Table 3). Database for Annotation, Visualization and Integrated Discovery analysis of these genes revealed an ontology enrichment of genes that are involved in regulation of cell proliferation: cell cycle, Wnt signalling, TGFβ/BMP signaling and embryonic/morphogenesis processes (Supplementary File S5), consistent with the known role of Wnt signaling in proliferation and differentiation. We previously showed that the C-clamp controls a gene expression program that is critical to the G1-S phase transition in DLD-1 colon cancer cells through regulation of the cyclin-dependent kinase inhibitor p21 (15,19). That our analysis identified multiple histone genes hints that Wnt signaling may also target this class of gene to prepare cells for the S phase of the cell cycle when a newly replicated daughter genome must be packaged with nucleosomal histones. Our validation focused on one ChIP-seq peak associated with one histone gene. However, transcription of over 20 different histone genes was inhibited by dnTCF1EWT hinting that regulation may be more widespread, a new finding that has not been previously reported. Histone mRNA ends in a stem-loop sequence and, therefore, this class of Wnt target gene may have been largely undetected in prior transcriptome studies that enrich for polyadenylated, RNA polymerase II transcripts. Taken together our discovery highlights a distinct advantage of the 4'thiouridine-seq method, which enriches for nascently transcribed RNA regardless of polyadenylation status.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of the UCI Institute for Genomics and Bioinformatics, the UCI Genomics High-Throughput Facility and the Chao Family Comprehensive Cancer Center for support, and NVIDIA for donated hardware. Additional support of our computational infrastructure was provided by Yuzo Kanomata. The authors would also like to thank Dr. William Gordon and Rachel Herndon for valuable discussions about ChIP-seq data, and members of the Waterman and Peter Donovan laboratory for valuable critique of the Wnt transcriptome strategy and analysis.

## FUNDING

National Institutes of Health (NIH) [CA096878, CA108697 to M.L.W. and [CA177651 to K.J.H. and M.L.W.]; T15 LM07443 to M.Z. and P.B.; LM010235 to P.B.]; National Science Foundation [IIS-0513376 to P.B.]; National Cancer Institute [P30CA062203] to the Chao Family Comprehensive Cancer Center and its support for the Genomics High Throughput Facility and Optical Biology Center Shared Resource. Funding for open access charge: NIH [RO1 CA108697 and RO1 CA096878].

*Conflict of interest statement.* None declared.

## REFERENCES

1. Holstein,T.W. (2012) The Evolution of the Wnt Pathway. *Cold Spring Harb. Perspect. Biol.*, **4**, a007922.
2. Srivastava,M., Begovic,E., Chapman,J., Putnam,N.H., Hellsten,U., Kawashima,T., Kuo,A., Mitros,T., Salamov,A., Carpenter,M.L. *et al.* (2008) The Trichoplax genome and the nature of placozoans. *Nature*, **454**, 955–960.
3. Bhanot,P., Brink,M., Samos,C.H., Hsieh,J.C., Wang,Y., Macke,J.P., Andrew,D., Nathans,J. and Nusse,R. (1996) A new member of the frizzled family from Drosophila functions as a Wingless receptor. *Nature*, **382**, 225–230.
4. Yost,C., Torres,M., Miller,J.R., Huang,E., Kimelman,D. and Moon,R.T. (1996) The axis-inducing activity, stability, and subcellular distribution of beta-catenin is regulated in Xenopus embryos by glycogen synthase kinase 3. *Genes Dev.*, **10**, 1443–1454.
5. Molenaar,M., van de Wetering,M., Oosterwegel,M., Peterson-Maduro,J., Godsave,S., Korinek,V., Roose,J., Destrée,O. and Clevers,H. (1996) XTcf-3 transcription factor mediates beta-catenin-induced axis formation in Xenopus embryos. *Cell*, **86**, 391–399.
6. Mosimann,C., Hausmann,G. and Basler,K. (2009) Beta-catenin hits chromatin: regulation of Wnt target gene activation. *Nat. Rev. Mol. Cell Biol.*, **10**, 276–286.
7. Van der Heyden,M.A., Rook,M.B., Hermans,M.M., Rijksen,G., Boonstra,J., Defize,L.H. and Destrée,O.H. (1998) Identification of connexin43 as a functional target for Wnt signalling. *J. Cell Sci.*, **111**(Pt 1), 1741–1749.
8. Morin,P.J. (1997) Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta -catenin or APC. *Science*, **275**, 1787–1790.
9. Muzny,D.M., Bainbridge,M.N., Chang,K., Dinh,H.H., Drummond,J.A., Fowler,G., Kovar,C.L., Lewis,L.R., Morgan,M.B. and Newsham,I.F., et al. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330–337.
10. Riese,J., Yu,X., Munnerlyn,A., Eresh,S., Hsu,S.C., Grosschedl,R. and Bienz,M. (1997) LEF-1, a nuclear factor coordinating signaling inputs from wingless and decapentaplegic. *Cell*, **88**, 777–787.
11. Rocheleau,C.E., Downs,W.D., Lin,R., Wittmann,C., Bei,Y., Cha,Y.H., Ali,M., Priess,J.R. and Mello,C.C. (1997) Wnt signaling

and an APC-related gene specify endoderm in early C. elegans embryos. *Cell*, **90**, 707–716.

12. Bottomly,D., Kyler,S.L., McWeeney,S.K. and Yochum,G.S. (2010) Identification of {beta}-catenin binding regions in colon cancer cells using ChIP-Seq. *Nucleic Acids Res.*, **38**, 5735–5745.
13. Klingel,S., Morath,I., Strietz,J., Menzel,K., Holstein,T.W. and Gradl,D. (2012) Subfunctionalization and neofunctionalization of vertebrate Lef/Tcf transcription factors. *Dev. Biol.*, **368**, 44–53.
14. Cadigan,K.M. and Waterman,M.L. (2012) TCF/LEFs and Wnt signaling in the nucleus. *Cold Spring Harb. Perspect. Biol.*, **4**, pii:a007906.
15. Hoverter,N.P., Ting,J.-H., Sundaresh,S., Waterman,M.L. and Baldi,P. (2012) A WNT/p21 circuit directed by the C-clamp, a sequence-specific DNA binding domain in TCFs. *Mol. Cell. Biol.*, **32**, 3648–3662.
16. Najdi,R., Syed,A., Arce,L., Theisen,H., Ting,J.-H., Atcha,F., Nguyen,A.V., Martinez,M., Holcombe,R.F., Edwards,R.A. *et al.* (2009) A Wnt kinase network alters nuclear localization of TCF-1 in colon cancer. *Oncogene*, **28**, 4133–4146.
17. Weise,A., Bruser,K., Elfert,S., Wallmen,B., Wittel,Y., Wöhrle,S. and Hecht,A. (2010) Alternative splicing of Tcf7l2 transcripts generates protein variants with differential promoter-binding and transcriptional activation properties at Wnt/beta-catenin targets. *Nucleic Acids Res.*, **38**, 1964–1981.
18. Wallmen,B., Schrempp,M. and Hecht,A. (2012) Intrinsic properties of Tcf1 and Tcf4 splice variants determine cell-type-specific Wnt/β-catenin target gene expression. *Nucleic Acids Res.*, **40**, 9455–9469.
19. Van de Wetering,M., Sancho,E., Verweij,C., de Lau,W., Oving,I., Hurlstone,A., van der Horn,K., Batlle,E., Coudreuse,D., Haramis,A.P. *et al.* (2002) The beta-catenin/TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells. *Cell*, **111**, 241–250.
20. Atcha,F.A., Syed,A., Wu,B., Hoverter,N.P., Yokoyama,N.N., Ting,J.-H.T., Munguia,J.E., Mangalam,H.J., Marsh,J.L. and Waterman,M.L. (2007) A unique DNA binding domain converts T-cell factors into strong Wnt effectors. *Mol. Cell. Biol.*, **27**, 8352–8363.
21. Koo,B.-K., Robine,S., van den Born,M., Itzkovitz,S., Korving,J., van Es,J.H., Boj,S.F., Haegebarth,A., van Oudenaarden,A., Kujala,P. *et al.* (2012) A critical role for the Wnt effector Tcf4 in adult intestinal homeostatic self-renewal. *Mol. Cell. Biol.*, **32**, 1918–1927.
22. Bhambhani,C., Ravindranath,A.J., Mentink,R.A., Chang,M.V., Betist,M.C., Yang,Y.X., Koushika,S.P., Korswagen,H.C. and Cadigan,K.M. (2014) Distinct DNA binding sites contribute to the TCF transcriptional switch in C. elegans and Drosophila. *PLoS Genet.*, **10**, e1004133.
23. Chang,M. V, Chang,J.L., Gangopadhyay,A., Shearer,A. and Cadigan,K.M. (2008) Activation of wingless targets requires bipartite recognition of DNA by TCF. *Curr. Biol.*, **18**, 1877–1881.
24. Ravindranath,A. and Cadigan,K.M. (2014) Structure-function analysis of the C-clamp of TCF/Pangolin in Wnt/ß-catenin signaling. *PLoS One*, **9**, e86180.
25. Atcha,F.A., Munguia,J.E., Li,T.W.H., Hovanes,K. and Waterman,M.L. (2003) A new beta-catenin-dependent activation domain in T cell factor. *J. Biol. Chem.*, **278**, 16169–16175.
26. Higuchi,R., Krummel,B. and Saiki,R.K. (1988) A general method of in vitro preparation and specific mutagenesis of DNA fragments: study of protein and DNA interactions. *Nucleic Acids Res.*, **16**, 7351–7367.
27. Cao,A.R., Rabinovich,R., Xu,M., Xu,X., Jin,V.X. and Farnham,P.J. (2011) Genome-wide analysis of transcription factor E2F1 mutant proteins reveals that N- and C-terminal protein interaction domains do not participate in targeting E2F1 to the human genome. *J. Biol. Chem.*, **286**, 11985–11996.
28. Emerson,R.O. and Thomas,J.H. (2009) Adaptive evolution in zinc finger transcription factors. *PLoS Genet.*, **5**, e1000325.
29. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.*, **79**, 233–269.
30. Dölken,L., Ruzsics,Z., Rädle,B., Friedel,C.C., Zimmer,R., Mages,J., Hoffmann,R., Dickinson,P., Forster,T., Ghazal,P. *et al.* (2008) High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA*, **14**, 1959–1972.
31. Schuijers,J., Mokry,M., Hatzis,P., Cuppen,E. and Clevers,H. (2014) Wnt-induced transcriptional activation is exclusively mediated by TCF/LEF. *EMBO J.*, **33**, 146–156.
32. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
33. Li,Q., Brown,J.B., Huang,H. and Bickel,P.J. (2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5**, 1752–1779.
34. Barski,A., Cuddapah,S., Cui,K., Roh,T.-Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
35. Koch,C.M., Andrews,R.M., Flicek,P., Dillon,S.C., Karaöz,U., Clelland,G.K., Wilcox,S., Beare,D.M., Fowler,J.C., Couttet,P. *et al.* (2007) The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res.*, **17**, 691–707.
36. Mokry,M., Hatzis,P., de Bruijn,E., Koster,J., Versteeg,R., Schuijers,J., van de Wetering,M., Guryev,V., Clevers,H. and Cuppen,E. (2010) Efficient double fragmentation ChIP-seq provides nucleotide resolution protein-DNA binding profiles. *PLoS One*, **5**, e15092.
37. Thomas-Chollier,M., Darbo,E., Thieffry,D., van Helden,J., Defrance,M. and Herrmann,C. (2012) A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nat. Protoc.*, **7**, 1551–1568.
38. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
39. Melvin,W.T., Milne,H.B., Slater,A.A., Allen,H.J. and Keir,H.M. (1978) Incorporation of 6-thioguanosine and 4-thiouridine into RNA. Application to isolation of newly synthesised RNA by affinity chromatography. *Eur. J. Biochem.*, **92**, 373–379.
40. Giese,K., Amsterdam,A. and Grosschedl,R. (1991) DNA-binding properties of the HMG domain of the lymphoid-specific transcriptional regulator LEF-1. *Genes Dev.*, **5**, 2567–2578.
41. Giese,K., Pagel,J. and Grosschedl,R. (1997) Functional analysis of DNA bending and unwinding by the high mobility group domain of LEF-1. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 12845–12850.
42. Hecht,A. and Stemmler,M.P. (2003) Identification of a promoter-specific transcriptional activation domain at the C terminus of the Wnt effector protein T-cell factor 4. *J. Biol. Chem.*, **278**, 3776–3785.
43. Ishitani,T., Ninomiya-tsuji,J., Nagai,S., Nishita,M., Meneghini,M., Barker,N., Waterman,M., Bowerman,B., Clevers,H., Shibuya,H. *et al.* (1999) The TAK1 ± NLK ± MAPK- related pathway antagonizes signalling between b -catenin and transcription factor TCF. *Nature*, **399**, 798–802.

# APPENDIX C

## The Silent Sway of Splicing by Synonymous Substitutions

THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 290, NO. 46, pp. 27700–27711, November 13, 2015

# The Silent Sway of Splicing by Synonymous Substitutions*[S]

Received for publication, August 5, 2015, and in revised form, September 29, 2015 Published, JBC Papers in Press, September 30, 2015, DOI 10.1074/jbc.M115.684035

**William F. Mueller[‡], Liza S. Z. Larsen[§], Angela Garibaldi[‡], G. Wesley Hatfield[‡§], and Klemens J. Hertel[‡§1]**

*From the [‡]Department of Microbiology and Molecular Genetics and the [§]Institute for Genomics and Bioinformatics, University of California Irvine, Irvine, California 92619*

**Background:** The effects of silent mutations on pre-mRNA splicing are poorly understood.
**Results:** Silent mutations can significantly influence exon inclusion and are under purifying selection.
**Conclusion:** Splicing and coding pressures have co-evolved to maintain sufficient exon inclusion levels.
**Significance:** Modified species alignment approaches can be used to identify silent mutations that may alter exon inclusion.

**Alternative splicing diversifies mRNA transcripts in human cells. This sequence-driven process can be influenced greatly by mutations, even those that do not change the protein coding potential of the transcript. Synonymous mutations have been shown to alter gene expression through modulation of splicing, mRNA stability, and translation. Using a synonymous position mutation library in *SMN1* exon 7, we show that 23% of synonymous mutations across the exon decrease exon inclusion, suggesting that nucleotide identity across the entire exon has been evolutionarily optimized to support a particular exon inclusion level. Although phylogenetic conservation scores are insufficient to identify synonymous positions important for exon inclusion, an alignment of organisms filtered based on similar exon/intron architecture is highly successful. Although many of the splicing neutral mutations are observed to occur, none of the exon inclusion reducing mutants was found in the filtered alignment. Using the modified phylogenetic comparison as an approach to evaluate the impact on pre-mRNA splicing suggests that up to 45% of synonymous SNPs are likely to alter pre-mRNA splicing. These results demonstrate that coding and pre-mRNA splicing pressures co-evolve and that a modified phylogenetic comparison based on the exon/intron architecture is a useful tool in identifying splice altering SNPs.**

Pre-mRNA splicing is an essential process necessary for both proper gene expression and the generation of transcript diversity throughout metazoans (1). Intron removal, directed by sequence signals within the pre-mRNA, is catalyzed by the spliceosome, a large ribonuclear protein complex (2). The interaction of splicing regulatory elements and their trans-acting binding partners (hnRNPs, SR proteins, and small nuclear ribonucleoproteins)[2] determines where and how splicing takes place (3). Mutations within binding sites of these trans-acting factors will likely alter the identity of the resulting mRNA (4).

The presence of sequence elements that influence splicing throughout the transcript suggests that organisms that rely on high fidelity splicing will be under evolutionary pressure to maintain optimal splice signals within the pre-mRNA molecule (5). Indeed, it has previously been shown that splicing regulatory elements exhibit positive selection, that intron/exon boundaries have a decreased frequency of single nucleotide polymorphisms (SNPs), and that certain codons are generally preferred around intron/exon junctions (6–8). These observations suggest that the spliceosome has to recognize exons by using sequences that are co-evolving with the amino acid sequence code to generate an RNA molecule that both translates and properly splices. However, it is difficult to uncouple sequence requirements imposed by evolutionary coding pressures from sequence requirements necessary to generate the appropriate mRNA through splicing.

Splicing requires sequence elements from both the intron and the exon for proper intron excision. Within the exon, those sequence elements are restricted in sequence by the need for correct protein coding. Given the diversity of the genetic code in the third or wobble position, it has been possible in a few cases to separate splicing constraints from coding constraints (9, 10). Here, we characterize the contribution of splicing evolutionary pressures through the identification of synonymous mutations that specifically alter the efficiency of intron removal. We created a library of synonymous mutations across exon 7 of the *SMN1* gene to identify positions that influence exon inclusion using deep-sequencing approaches. Phylogenetic comparisons with organisms that have a similar *SMN1* exon 7 architecture demonstrated a selection against exon inclusion-reducing mutants, clearly demonstrating that evolutionary pressures exist at wobble positions to uphold efficient splicing of this crucial exon. A survey of human synonymous SNPs showed that 45% contain potentially splice altering nucleotide substitutions. These observations suggest that a large proportion of synonymous SNPs can cause defects in splicing. Thus, filtering multiple species alignments by exon architectural similarity represents a novel strategy to identify exonic splicing regulatory elements and synonymous mutations that disrupt proper pre-mRNA splicing.

* This work was supported, in whole or in part, by National Institutes of Health Grant R01 GM62287 (to K. J. H.) and F31 Ca171791 (to A. G.). We declare no competing interests or conflicts of interest in the writing or publishing of this paper.

[S] This article contains supplemental Table S1.

[1] To whom correspondence should be addressed. Tel.: 949-824-2127; Fax: 949-824-8598; E-mail: khertel@uci.edu.

[2] The abbreviations used are: hnRNP, heterogeneous nuclear ribonucleoprotein; SMN, survival of motor neuron; SMA, spinal muscular atrophy; nt, nucleotide; PhyloP, part of the PHAST package (PHylogenetic Analysis with Space/Time models) used through UCSC Genome Browser to evaluate phylogenetic conservation.

## Experimental Procedures

*Cell Culture and Transfection of HeLa Cells*—HeLa cells used in this work were maintained at 37 °C in a monolayer in Dulbecco's high glucose modified Eagle's medium (Invitrogen) supplemented with 10% fetal bovine serum, 4 mM L-glutamate, and 1 mM Na-pyruvate. Their cell confluence was kept to ~80% or less before splitting cells. Cells were transfected according to the manufacturer's specifications for Lipofectamine 2000 (Invitrogen) for plate sizes of 10-cm 6-well plates with 3-cm wells, or 15-cm plates.

*Creation/Transfection of SMN1 Exon 7 Library Mutations*—We used previously described SMN mini-genes containing exons 6, 7, and 8 with shortened introns between each exon (11–13). To generate a library of synonymous mutations across exon 7, we created a series of mutagenesis primers to make a site-directed saturation mutagenesis library of exon 7. This library is comprised of synonymous site-directed mutations spanning the 54-nucleotide exon 7 region of the *SMN1* gene. All possible combinations of synonymous mutations were generated within a sliding 2-codon window such that the gene amino acid sequence remained the same and all possible synonymous mutations were created.

Sets of oligonucleotide primers were used to generate the mutations to each synonymous position in exon 7. To generate the library with fewer per-base errors, oligonucleotide sets used were kept as short as possible and the mutagenesis reactions were split into three separate pooled reactions (sequences of oligonucleotides used are available upon request). The first reaction contained a pool of 8 forward and 8 reverse oligonucleotides (oligos 1–8), the second reaction contained a pool of 10 forward and 10 reverse oligonucleotides (oligos 9–18), and the third reaction contained a pool of 12 forward and 12 reverse oligonucleotides (oligos 19–30), all of which contain a double randomized NNN NNN codon at its center such that the two amino acids are simultaneously mutated (but only to synonymous mutations).

PCR amplifications for each mutagenesis pool of *SMN1* exon 7 were performed in 50-μl reactions containing: 50 ng of the plasmid based pCi *SMN1* mini-gene template, 5 units of Pfu Ultra-High Fidelity DNA polymerase (Stratagene); 400 μM dNTPs; 1× Pfu Ultra-High Fidelity reaction buffer; and 0.03 μM of each complementary mutant primer pair. Primer extension and PCR amplification reactions were carried out by: 10 min denaturation at 95 °C, followed by 16 cycles of 15 s at 95 °C, 40 s at 55 °C, 3 min at 72 °C, and a final step of 10 min at 72 °C. The 50-μl PCR products were digested with 30 units of DpnI for 3 h at 37 °C to remove the methylated template plasmid. 5 μl of the digestion reaction was used for transformation of DH5α cells and plated onto LB-agar plates (100 μg/ml of ampicillin). At this point, a plasmid library containing saturated site-directed mutated target regions is generated. Colonies were selected and suspended in 5 ml of LB medium (100 μg/ml of ampicillin) and incubated at 37 °C overnight. 15% glycerol stocks were created, and larger 200-ml flasks of LB medium (100 μg/ml of ampicillin) were inoculated and incubated at 37 °C overnight. Plasmid libraries were purified from the large cultures using the Nucleobond Midi-prep Purification system. Plasmid libraries were then combined in equal molar amounts to make one master library.

*Transfection and Sequencing Library Preparation*—The approach for the generation of sequencing libraries is outlined in Fig. 1. To generate a library of spliced mRNAs, HeLa cells at ~80% confluence in 10-cm plates were transfected with 22 μg of the master plasmid library using Lipofectamine 2000 for 6 h in serum-free medium at 37 °C. After 6 h, the Lipofectamine medium was removed and medium-containing serum was added. The cells were incubated another 18 h at 37 °C. Total RNA was purified from the cells using TRIzol reagent (Ambion) according to the manufacturer's recommendations and treated with DNase (Invitrogen). Reverse transcription using SuperScript II reverse polymerase (Invitrogen) was carried out using library plasmid-specific primers (pCI backbone forward primer: GCTAACGCAGTCAGTGCTTC; pCI backbone reverse primer: GTATCTTATCATGTCTGCTCG) and 4 μg of RNA. The cDNA reaction was then cleaned up using a Qiagen PCR purification kit according to the manufacturer's recommendations. Approximately 0.5 μg of purified cDNA was then amplified using primers specific to exon 6 and exon 8 (output cDNA forward primer: CCCTACACGACGCTCTTCCGATCTCATGAGTGGCTATCATACTGGC; output cDNA reverse primer: CCTGCTGAACCGCTCTTCCGATCTGTCATTTAGTGCTGCTCTATGC; Seq. tail forward primer: AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT; Seq. tail reverse primer: CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT) that also had tails containing the initial segment of Illumina-specific sequencing primers in a 50-μl reaction containing 400 nM dNTPs using proofreading DNA polymerase. PCR amplification reactions were carried out by: 5 min denaturation at 95 °C, 3 cycles of: 15 s denaturing at 95 °C, 30 s primer annealing at 55 °C, 20 s of extension at 72 °C, and a final extension of 5 min at 72 °C. This initial PCR library amplification was then run out on a 1.5% agarose gel to remove exon exclusion events and gel purified using a Qiagen Gel extraction kit. The purified PCR product was then amplified again for 12 cycles using the same PCR cycle conditions as described above, except using a full-length Illumina sequencing primer. The products of this reaction were separated by 1.5% agarose gel electrophoresis and the PCR product band was purified using the Qiagen gel extraction kit according to the manufacturer's recommendations. This final library was analyzed for quality on an Agilent Bioanalyzer 2100, quantified using a Nanodrop (ND-1000 Thermo), and then diluted to 10 nM and submitted for sequencing on the Illumina Hi-seq platform. The spliced mRNA pool is referred to as the "output pool."

To generate an equivalent sequencing library of the expression plasmid pool, the same procedure (minus transfection and RNA purification) was carried on the plasmid pool using primers specific to introns 6 and 7 (input DNA forward primer: CCCTACACGACGCTCTTCCGATCTGCTATTTTTTTTTAACTTCCTTTATTTTC; input DNA reverse primer: CCTGCTGAACCGCTCTTCCGATCTGTAAGATTCACTTTCATAATGCTG). The expression plasmid pool is referred to as the "input pool." PCR analysis of SMN exon 7 mutants for valida-

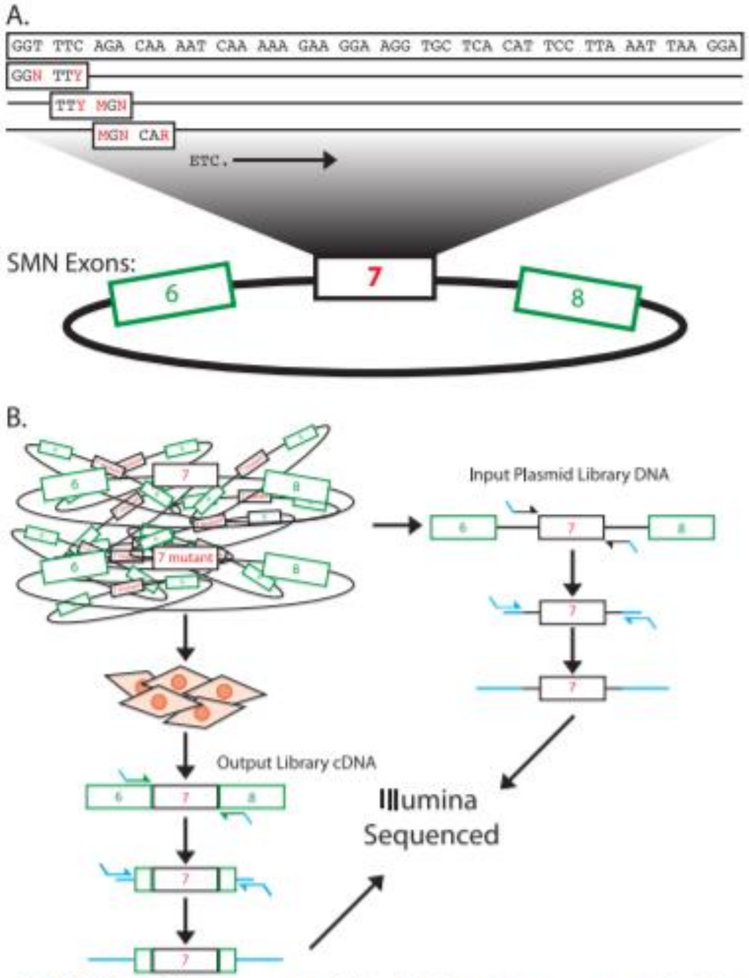

A.
GGT TTC AGA CAA AAT CAA AAA GAA GGA AGG TGC TCA CAT TCC TTA AAT TAA GGA
GGN TTY
TTY MGN
MGN CAR
ETC.
SMN Exons:
6
7
8
B.
7
7 mutant
6
8
Input Plasmid Library DNA
Output Library cDNA
Illumina
Sequenced

FIGURE 1. **Mutation scheme for the *SMN1* exon 7 library preparation.** *A*, all possible silent mutations were generated per in-frame hexamer across the exon. For example, the first two codons depicted are GGT TTC. All three mutations were made in GGT resulting in GGN and combined with all silent mutations in TTC (TTT), resulting in eight combinations including the wild-type sequence. *B*, the library of mutations were sequenced and transfected into HeLa cells. The RNA resulting from the transfection was purified and sequenced. The relative abundance of each mutation was compared between the two pools to generate fitness index values.

tion and comparison of our sequencing experiment was carried out using primers as previously described (11, 17).

*Bioinformatic Analysis of SMN1 Exon 7 Library Mutations*—We received 54,780,073 single-end reads of 100 nt from our sequencing run. Using Bowtie these reads were aligned to a custom index made from the genomic segment of *SMN1* spanning exons 6 to 8 (14), from the spliced mRNA sequence spanning exons 6, 7, and 8, and from all library mutants made. Thus, the alignment index contained an entry for each library mutation to control for possible mutations or sequencing errors. Only reads that contained the wild-type sequence or library mutations were aligned. Approximately three fourths of the reads were aligned to the custom index. Another 8% of the reads corresponded to exon exclusion reads, and the remaining reads did not align.

The total read count for each library mutation was used to determine the relative representation of each mutation within in the sequenced input and output pools. To evaluate changes in exon 7 inclusion, we calculated WT-normalized differences in mutant output and input ratios according to "fitness index" = (output/WT output)/(input/WT input). These ratios are defined as the fitness index. The average fitness index for a position is the average of the fitness indices for all mutations at that position.

For statistical analysis we compared the output and input reads for each mutation to the non-mutated reads using a Fisher's Exact test and determined the significance using a Bonferroni correction for comparison between the 138 mutations. In addition to the Fisher's Exact test, we applied a biological cutoff for significance of the fitness index value. Twenty percent exon

7 inclusion is consistent with mRNA expression reports for SMA type I. An observable difference in the exon inclusion phenotypes between SMA types I and III has been reported (12, 15–17). In SMA type III there is ~70% inclusion of exon 7 compared with WT levels, *versus* the 20% observed in type I. Therefore, we set our biological significance cutoff at this 70% value.

Our analysis relied on the circumstantially validated assumptions that all plasmids transfect and transcribe with the same efficiency and that the resulting RNAs are similarly stable in HeLa cells. A decrease in transcription or transcriptional pausing causing a loss of reads is unlikely, as significant numbers of reads were generated and analyzed corresponding to each mutation (average cDNA reads 209,891/mutation, median: 39,288). This suggests that each mutation was represented within the input plasmid pool, successfully transfected, and successfully expressed. The lowest read count observed for a single mutant was 30 output reads for the mutations in positions 24 and 27. This position is known for being a splicing regulator (Tra2-$\beta$1/SRSF10) binding site, suggesting that the loss in reads was due to the loss of an enhancer binding sequence (18). Consistently, low output reads correlated with a loss of a known positive splicing regulator, the gain of a negative splicing regulator, or increasing known local RNA secondary structures. It is not known whether mRNAs generated from the reporter plasmids are translated, and therefore are potential NMD targets. However, because all exon 7 mutations are located within 50 nt of the last exon-exon junction, it is unlikely that the resulting mRNA would be targeted by NMD (19). It is also unknown whether the stability of exon 7 included mRNAs generated from the mini-gene constructs is different from mRNAs lacking exon 7.

*Comparative Alignment Analysis of SMN1 Exon 7 Mutations and Synonymous SNPs*—The position of *SMN1* exon 7 was identified through the UCSC Genome Browser. The UCSC Table Browser was then used to download the PhyloP Score by position across the exon (position chr5: 70,247,768 to 70,247,821) using the Comparative Genomics Group and the Conservation Track in the Vertebrate 46 way alignment table *SMN1*. PhyloP conservation scores allow a measure of conservation by position. Multiple alignment data were generated by taking this same genomic position into the Esembl! database (GRCh37) and selecting the comparative genomics (text option) tool for all 36 eutherian mammals in the Ensembl! Genome Sequence Viewer. This was also used to validate intron/exon architecture. Intron/exon architecture similarity was determined through sequence analysis using the comparative genomics tool. Organisms with similar exon length (±6 nt) and splice site sequences were compared. Species that exhibit similar exon/intron architecture were kept in the alignment. Species exhibiting differing exon/intron architecture were excluded from the alignment. In the case of the SMN alignment, all primates were included in the intron/exon architecture filter, however, this excluded multiple species from rodents to the alpaca. The same approach was used for filtering SNP-based species alignments.

Synonymous SNPs associated with human disease were taken from the literature and from the Entrez dbSNP web browser and added to synonymous human SNPs downloaded using the Ensembl! PerlAPI (20). The 40 disease associated SNPs and 40 Ensembl! SNPs were randomly selected and analyzed using the Ensembl! genome browser comparative genomics alignment tools as was done for the *SMN* exon analysis. Eighty more human synonymous SNPs from non-coding exons, plus an additional 82 synonymous SNPs with a single exon were also acquired using the Ensembl! PerlAPI and analyzed in the same manner (supplemental Table S1). Filtered organisms were compared by sequence conservation in an 11-nucleotide window around the SNP position. The numbers of the classified putative splice altering SNPs and neutral SNPs were compared using a $\chi^2$ test of independence.

## Results

*Analysis of Library Mutations*—To determine the influence of synonymous mutations on exon inclusion, we used the well studied *SMN1* mini-gene, spanning exons 6–8 (13, 17, 21) in which exon 7 can be included or excluded depending on the splicing signals within the pre-mRNA. Each set of two neighboring codons in exon 7 was mutated to every possible combination of silent mutations within the context of a hexamer, a minimal binding platform for splicing regulatory proteins (Fig. 1*A*) (22). The resulting library of plasmids was transfected into HeLa cells and plasmid-specific mRNAs were analyzed by deep sequencing (Fig. 1*B*). Relative exon inclusion indexes were determined by calculating the abundance of mutations present in the exon 7-included mRNA population, normalized to the abundance in the input plasmid control. Presumably, differences in inclusion indexes reflect differences in exon 7 inclusion levels (Table 1). To test this assumption a small number of mutants (Table 2 and Fig. 2) were tested individually for exon inclusion levels using standard RT-PCR approaches. The near linear relationship between the different experimental approaches demonstrate that changes in the fitness indexes are largely driven by altered pre-mRNA splicing efficiencies. If a mutation was observed more frequently in the spliced mRNA pool than the input plasmid pool, that mutation was considered to be beneficial to exon inclusion. If a mutation was observed less frequently in the spliced mRNA pool compared with the input plasmid pool, that mutation was considered to be inhibitory to exon inclusion.

The statistical significance of an observed change in relative mutant abundance was compared with the wild-type (WT) representation using a Fisher's Exact test with Bonferroni correction. In addition a biological filter was imposed based on previous observations that mRNA levels of SMA type III patients display exon 7 inclusion levels at ~70% of that observed for healthy individuals (12, 15). All mutants that caused a decrease in the fitness index to 70% or less of the unchanged WT value were considered to have a negative effect on exon inclusion. Of 138 mutations tested, 32 met the criteria for significant changes in the fitness index (Table 1).

The fact that only 32 of the 138 mutations evaluated caused significant fitness index reductions demonstrates that the majority of the mutations tested did not appear to cause striking changes in exon inclusion. Most mutations that caused significant changes negatively affected the fitness index. Interest-

**TABLE 1**

**All mutations analyzed**

Summary of results obtained for all library mutations analyzed. The mutation positions are from the first position of *SMN1* exon 7. The wild-type and mutant nucleotides are separated by ">". Read counts and exon inclusion ratios were found and calculated as described under "Experimental Procedures". Mutants highlighted in grey were found to cause a significant reduction in the fitness index value.

| Mutation and Position | Input Read Count | Output Read Count | Inclusion Index Value | Log Inclusion Index | Mutation and Position | Input Read Count | Output Read Count | Inclusion Index Value | Log Inclusion Index |
|---|---|---|---|---|---|---|---|---|---|
| Wildtype, No Mutations | 2969387 | 6464990 | 1.00 | 0.00 | 34:T>A, 35:C>G, 36:A>T* | 92751 | 133695 | 0.66 | -0.18 |
| 3:T>A* | 65023 | 174024 | 1.23 | 0.09 | 34:T>A, 35:C>G, 36:A>C | 127089 | 274765 | 0.99 | 0.00 |
| 3:T>C* | 72883 | 213356 | 1.34 | 0.13 | 34:T>A, 35:C>G, 36:A>T, 39:T>C* | 17411 | 233 | 0.01 | -2.21 |
| 3:T>G | 104141 | 213785 | 0.94 | -0.03 | 34:T>A, 35:C>G, 36:A>C, 39:T>C* | 53816 | 1956 | 0.02 | -1.78 |
| 3:T>A, 6:C>T* | 28855 | 13109 | 0.21 | -0.68 | 36:A>C* | 63524 | 178900 | 1.29 | 0.11 |
| 3:T>C, 6:C>T* | 47776 | 27248 | 0.26 | -0.58 | 36:A>G* | 105766 | 299431 | 1.30 | 0.11 |
| 3:T>G, 6:C>T* | 68754 | 19519 | 0.13 | -0.88 | 36:A>T* | 219252 | 506661 | 1.06 | 0.03 |
| 6:C>T* | 231830 | 195369 | 0.39 | -0.41 | 36:A>T, 39:T>C* | 10722 | 2188 | 0.09 | -1.03 |
| 6:C>T, 7:A>C* | 63130 | 163781 | 1.19 | 0.08 | 36:A>G, 39:T>C* | 22456 | 8618 | 0.18 | -0.75 |
| 6:C>T, 7:A>C, 9:A>T* | 71615 | 185308 | 1.19 | 0.07 | 36:A>C, 39:T>C* | 41583 | 25908 | 0.29 | -0.54 |
| 6:C>T, 7:A>C, 9:A>C* | 87558 | 237283 | 1.24 | 0.10 | 39:T>C* | 56096 | 18783 | 0.15 | -0.81 |
| 6:C>T, 7:A>C, 9:A>G* | 89105 | 263489 | 1.36 | 0.13 | 39:T>C, 40:T>A, 41:C>G* | 7176 | 34778 | 2.23 | 0.35 |
| 6:C>T, 9:A>G* | 202960 | 95344 | 0.22 | -0.67 | 39:T>C, 40:T>A, 41:C>G, 42:C>T* | 7581 | 22007 | 1.33 | 0.12 |
| 7:A>C | 215137 | 697820 | 1.49 | 0.17 | 39:T>C, 42:C>T* | 3822 | 11742 | 1.38 | 0.14 |
| 7:A>C, 9:A>T* | 239937 | 633391 | 1.21 | 0.08 | 39:T>C, 42:C>A* | 4682 | 16120 | 1.58 | 0.20 |
| 7:A>C, 9:A>G* | 252521 | 882474 | 1.61 | 0.21 | 39:T>C, 42:C>G* | 4901 | 19685 | 1.84 | 0.27 |
| 7:A>C, 9:A>C* | 308880 | 910247 | 1.35 | 0.13 | 40:T>A, 41:C>G* | 14984 | 19322 | 0.59 | -0.23 |
| 7:A>C, 9:A>T, 12:A>G* | 110159 | 320179 | 1.33 | 0.13 | 40:T>A, 41:C>G, 42:C>T* | 8090 | 4891 | 0.28 | -0.56 |
| 7:A>C, 9:A>C, 12:A>G* | 114606 | 337835 | 1.35 | 0.13 | 40:T>A, 41:C>G, 42:C>T, 43:T>C | 1286 | 2975 | 1.06 | 0.03 |
| 7:A>C, 9:A>G, 12:A>G* | 142308 | 458433 | 1.48 | 0.17 | 40:T>A, 41:C>G, 42:C>T, 43:T>C, 45:A>T | 1206 | 1980 | 0.75 | -0.12 |
| 7:A>C, 12:A>G* | 106081 | 305749 | 1.32 | 0.12 | 40:T>A, 41:C>G, 42:C>T, 43:T>C, 45:A>C | 3094 | 10505 | 1.56 | 0.19 |
| 9:A>G* | 518071 | 1048763 | 0.93 | -0.03 | 40:T>A, 41:C>G, 42:C>T, 43:T>C, 45:A>G | 3510 | 11954 | 1.56 | 0.19 |
| 9:A>G, 12:A>G* | 258109 | 526741 | 0.94 | -0.03 | 40:T>A, 41:C>G, 42:C>T, 45:A>G | 8597 | 17255 | 0.92 | -0.04 |
| 12:A>G* | 479922 | 999508 | 0.96 | -0.02 | 40:T>A, 41:C>G, 43:T>C | 7745 | 18558 | 1.10 | 0.04 |
| 12:A>G, 15:T>C* | 225435 | 542544 | 1.11 | 0.04 | 40:T>A, 41:C>G, 43:T>C, 45:A>T | 1627 | 4423 | 1.25 | 0.10 |
| 15:T>C* | 462016 | 1074101 | 1.07 | 0.03 | 40:T>A, 41:C>G, 43:T>C, 45:A>G* | 3808 | 17866 | 2.15 | 0.33 |
| 15:T>C, 18:A>G* | 110917 | 288864 | 1.20 | 0.08 | 40:T>A, 41:C>G, 43:T>C, 45:A>C* | 3863 | 17870 | 2.07 | 0.32 |
| 18:A>G | 377050 | 802943 | 0.98 | -0.01 | 40:T>A, 41:C>G, 45:A>G | 3724 | 11039 | 1.36 | 0.13 |
| 18:A>G, 21:A>G* | 272585 | 706508 | 1.19 | 0.08 | 42:C>A* | 5994 | 18705 | 1.43 | 0.16 |
| 21:A>G* | 314749 | 774770 | 1.13 | 0.05 | 42:C>T* | 7417 | 22495 | 1.39 | 0.14 |
| 21:A>G, 24:A>G* | 58143 | 146114 | 1.15 | 0.06 | 42:C>G* | 13930 | 52007 | 1.71 | 0.23 |
| 24:A>G* | 179832 | 290267 | 0.74 | -0.13 | 42:C>A, 43:T>C | 536 | 2379 | 2.04 | 0.31 |
| 24:A>G, 27:A>T* | 18298 | 30 | 0.00 | -3.12 | 42:C>T, 43:T>C | 574 | 2342 | 1.87 | 0.27 |
| 24:A>G, 27:A>C* | 31048 | 3203 | 0.05 | -1.32 | 42:C>G, 43:T>C* | 3213 | 16193 | 2.31 | 0.36 |
| 24:A>G, 27:A>G* | 71168 | 75298 | 0.49 | -0.31 | 42:C>A, 43:T>C, 45:A>T | 902 | 3068 | 1.56 | 0.19 |
| 27:A>C* | 147704 | 39043 | 0.12 | -0.92 | 42:C>T, 43:T>C, 45:A>T | 1001 | 3806 | 1.75 | 0.24 |
| 27:A>T* | 150992 | 1333 | 0.00 | -2.39 | 42:C>T, 43:T>C, 45:A>G* | 1701 | 7735 | 2.09 | 0.32 |
| 27:A>G* | 218079 | 398116 | 0.84 | -0.08 | 42:C>G, 43:T>C, 45:A>G* | 1784 | 10870 | 2.80 | 0.45 |
| 27:A>C, 28:A>C* | 59948 | 107050 | 0.82 | -0.09 | 42:C>T, 43:T>C, 45:A>C* | 2102 | 9964 | 2.18 | 0.34 |
| 27:A>T, 28:A>C* | 87645 | 153342 | 0.80 | -0.09 | 42:C>G, 43:T>C, 45:A>C* | 2176 | 11188 | 2.36 | 0.37 |
| 27:A>G, 28:A>C* | 125619 | 228546 | 0.84 | -0.08 | 42:C>A, 43:T>C, 45:A>G* | 2522 | 11896 | 2.17 | 0.34 |
| 27:A>T, 28:A>C, 30:G>C | 637 | 1407 | 1.01 | 0.01 | 42:C>G, 43:T>C, 45:A>T* | 4712 | 22728 | 2.22 | 0.35 |
| 27:A>T, 28:A>C, 30:G>A | 1146 | 2638 | 1.06 | 0.02 | 42:C>A, 43:T>C, 45:A>C* | 9964 | 45096 | 2.08 | 0.32 |
| 27:A>G, 28:A>C, 30:G>C | 2369 | 4298 | 0.83 | -0.08 | 42:C>T, 45:A>G* | 4759 | 19918 | 1.92 | 0.28 |
| 27:A>G, 28:A>C, 30:G>A | 2932 | 6053 | 0.95 | -0.02 | 42:C>A, 45:A>G* | 9597 | 43639 | 2.09 | 0.32 |
| 27:A>T, 28:A>C, 30:G>T* | 5086 | 6751 | 0.61 | -0.21 | 42:C>G, 45:A>G* | 9696 | 49135 | 2.33 | 0.37 |
| 27:A>C, 28:A>C, 30:G>T | 5450 | 12751 | 1.07 | 0.03 | 43:T>C* | 9503 | 39639 | 1.92 | 0.28 |
| 27:A>G, 28:A>C, 30:G>T* | 5463 | 6631 | 0.56 | -0.25 | 43:T>C, 45:A>G* | 7342 | 39288 | 2.46 | 0.39 |
| 27:A>C, 28:A>C, 30:G>A | 6967 | 15411 | 1.02 | 0.01 | 43:T>C, 45:A>C* | 8064 | 41261 | 2.35 | 0.37 |
| 27:A>C, 28:A>C, 30:G>C* | 16372 | 28212 | 0.79 | -0.10 | 43:T>C, 45:A>T* | 11361 | 44770 | 1.81 | 0.26 |
| 27:A>G, 30:G>A | 18656 | 46030 | 1.14 | 0.06 | 43:T>C, 45:A>T, 48:T>C* | 2525 | 11256 | 2.05 | 0.31 |
| 27:A>T, 30:G>A* | 46870 | 27557 | 0.27 | -0.57 | 43:T>C, 45:A>G, 48:T>C* | 7494 | 39959 | 2.45 | 0.39 |
| 27:A>C, 30:G>A* | 47164 | 12476 | 0.12 | -0.92 | 43:T>C, 45:A>C, 48:T>C* | 12183 | 62275 | 2.35 | 0.37 |
| 28:A>C* | 156401 | 520880 | 1.53 | 0.18 | 43:T>C, 48:T>C* | 2715 | 13250 | 2.24 | 0.35 |
| 28:A>C, 30:G>A* | 59716 | 195669 | 1.50 | 0.18 | 45:A>G* | 11015 | 49451 | 2.06 | 0.31 |
| 28:A>C, 30:G>T* | 86092 | 225256 | 1.20 | 0.08 | 45:A>G*, 48:T>C* | 31178 | 138530 | 2.04 | 0.31 |
| 28:A>C, 30:G>C* | 87453 | 260649 | 1.37 | 0.14 | 48:T>C* | 34094 | 128744 | 1.73 | 0.24 |
| 28:A>C, 30:G>T, 33:C>T | 35627 | 85210 | 1.10 | 0.04 | 48:T>C, 50:A>G* | 25972 | 100932 | 1.78 | 0.25 |
| 28:A>C, 30:G>A, 33:C>T* | 43630 | 134625 | 1.42 | 0.15 | 48:T>C, 51:A>G* | 7375 | 656 | 0.04 | -1.39 |
| 28:A>C, 30:G>C, 33:C>T* | 55156 | 152696 | 1.27 | 0.10 | 50:A>G* | 42309 | 144136 | 1.56 | 0.19 |
| 28:A>C, 33:C>T* | 105855 | 314480 | 1.36 | 0.13 | 50:A>G, 54:A>T | 3657 | 6954 | 0.87 | -0.06 |
| 30:G>A* | 448725 | 1381967 | 1.41 | 0.15 | 50:A>G, 54:A>C* | 11668 | 41602 | 1.64 | 0.21 |
| 30:G>A, 33:C>T* | 176490 | 498528 | 1.30 | 0.11 | 50:A>G, 54:A>G* | 11677 | 75474 | 2.97 | 0.47 |
| 33:C>T* | 347154 | 651408 | 0.86 | -0.06 | 51:A>G* | 14915 | 856 | 0.03 | -1.56 |
| 33:C>T, 34:T>A, 35:C>G, 36:A>C* | 17143 | 12875 | 0.34 | -0.47 | 51:A>G, 54:A>G* | 590 | 3204 | 2.49 | 0.40 |
| 33:C>T, 34:T>A, 35:C>G, 36:A>T* | 44240 | 11285 | 0.12 | -0.93 | 51:A>G, 54:A>C* | 4612 | 842 | 0.06 | -1.03 |
| 33:C>T, 36:A>T | 10175 | 19445 | 0.88 | -0.06 | 51:A>G, 54:A>T* | 7808 | 577 | 0.03 | -1.47 |
| 33:C>T, 36:A>C | 60262 | 128984 | 0.98 | -0.01 | 54:A>G* | 1751 | 12496 | 3.28 | 0.52 |
| 33:C>T, 36:A>G* | 108350 | 293254 | 1.24 | 0.09 | 54:A>C* | 9417 | 35763 | 1.74 | 0.24 |
| *-significant by Fisher's Exact Test; Greyed rows values are 70% or less than WT | | | | | 54:A>T | 9963 | 18727 | 0.86 | -0.06 |

ingly, the fitness index-altering mutations were observed across the entire SMN exon 7. This observation suggests that optimization of exon inclusion levels for SMN was achieved at multiple exonic locations, providing a sufficient degree of flexibility to respond to localized coding pressures.

*Single Library Mutations*—When single nucleotide mutations were analyzed we observed both increased and decreased splicing efficiencies (Fig. 2*B*). Of the 32 mutations that caused significant decreases in the fitness index value, 6 of them were single nucleotide mutations (Table 1). Several of the single point mutations that caused the largest decreases in the fitness index have previously been reported as prominent splice altering mutations. Of 8 mutations previously evaluated in the literature, all are faithfully reproduced using our pooled deep

**TABLE 2**

**Validated mutations**

The mutations in this table have been previously reported in the literature or were validated from individual clones isolated from our library pools. Comparison to our data set shows that previous results are replicated faithfully (compare the similarity in changes between fitness index value or log transform with the previously published inclusion values).

| Mutation and position | Inclusion index value | Log inclusion index value | Previously analyzed % inclusion for exon 7 | Citations |
|---|---|---|---|---|
| Wild-type, no mutations | 1 | 0 | 75–100% | 11, 12, 20, 23, 31, 32, see Fig. 2 |
| 3:T>G | 0.94 | −0.03 | 96% | 31 |
| 3:T>G, 6:C>T | 0.13 | −0.88 | 6% | 31 |
| 6:C>T | 0.39 | −0.41 | 26% | 23 |
| 6:C>T, 7:A>C | 1.19 | 0.08 | 100% | 26, 32 |
| 7:A>C | 1.49 | 0.17 | 100% | 26, 32 |
| 27:A>C | 0.12 | −0.92 | 46% | See Fig. 2 |
| 27:A>T | 0.00 | −2.40 | 11% | See Fig. 2 |
| 27:A>G | 0.84 | −0.08 | 89% | See Fig. 2 |
| 27:A>T, 30:G>A | 0.27 | −0.57 | 76% | See Fig. 2 |
| 27:A>G, 30:G>A | 1.14 | 0.06 | 93% | See Fig. 2 |
| 39:T>C | 0.15 | −0.81 | 33% | See Fig. 2 |
| 39:T>C | 0.15 | −0.81 | 24% | 20 |
| 39:>C, 42:C>A | 1.58 | 0.20 | 91% | See Fig. 2 |
| 39:T>C, 42:C>G | 1.85 | 0.27 | 85% | See Fig. 2 |
| 42:C>A | 1.43 | 0.16 | 84% | See Fig. 2 |
| 42:C>G | 1.71 | 0.23 | 87% | See Fig. 2 |
| 45:A>G | 2.06 | 0.31 | 93% | 31 |
| 54:A>G | 3.28 | 0.52 | 100% | 20 |

sequencing approach (Table 2). Similarly, some of the mutations in our data that increased the fitness index value were previously reported in the literature to either destabilize an inhibitory RNA secondary structures or to increase the strength of the splice sites, *i.e.* alter the sequence to be closer to the splicing consensus sequence as measured by maximum entropy score (Fig. 3) (23). This agreement with previously published work further demonstrates that our deep sequencing approach reliably evaluates changes in splicing efficiencies. For example, when the highly conserved residue at position 6 was mutated to a synonymous codon (cytidine to thymidine), thus emulating the known difference in sequence between *SMN1* and *SMN2*, a 2.6-fold decrease in the fitness index was observed, in agreement with exon inclusion differences observed for *SMN1* and *SMN2* (4, 12, 24–27) (Table 1). Increases in fitness index values due to mutations may also be due to pre-mRNA or mature mRNA stability issues.

Furthermore, single nucleotide mutant influences on the fitness index values are nucleotide specific. For example, changing nucleotide position 27 from an adenosine to thymidine or cytidine caused a ~200- or ~8-fold decrease in exon inclusion, respectively (Fig. 2*B*, *inset*). However, when mutated to guanosine no significant change in the fitness index was observed.

To test the hypothesis that impermissible nucleotide changes at synonymous positions correlate with evolutionary conservation, we compared the mutated positions with their conservation scores as measured by PhyloP (28, 29). PhyloP uses a 46-vertebrate species alignment to assign positive values to positions that are conserved and gives lower or negative values to swiftly evolving or less conserved positions. Presumably, if a synonymous position is important for splicing, it is expected that its nucleotide identity would be conserved. This expectation is met when evaluating nucleotides at positions 6 and 51 (Fig. 2*C*). Based on the PhyloP score both positions are highly conserved, which is in agreement with the hypothesis that the nucleotide identity may be evolutionarily fixed due to selective pressures for correct splicing. However, as data from mutations at positions 27 and 39 indicate, this splicing correlated conservation is not typical in our dataset (Fig. 2*C*). Several of the synonymous mutation positions that had significant fitness index changes exhibit very low PhyloP scores. We conclude that the PhyloP score at a given wobble position does not directly predict the influence of the nucleotide on pre-mRNA splicing.

One potential explanation for the lack of a direct correlation between the PhyloP score and the fitness index value at synonymous positions is the fact that the exon/intron architecture may vary between species (Fig. 4*A*). As such, the evolutionary pressures to maintain a sequence for a particular splicing pattern can be different between species, eliminating positional conservation restrictions imposed by splicing. Our analysis clearly shows that different nucleotide changes at the same position can range from neutral to high impact, demonstrating that some nucleotide changes are permissible whereas others are not (Fig. 2*B*, *inset*, and Table 1). As PhyloP only measures the frequency of a nucleotide substitution irrespective of patterns or trends of mutations at a specific position, it is expected that in some cases the conservation score cannot adequately reflect splicing pressures. To circumvent these PhyloP limitations we filtered species alignments based on the requirement of similar intron/exon architecture around SMN1 exon 7. Using the Ensembl! genomic alignment tool, we hand filtered species alignments based on SMN exon 7 architectural conservation. This filtering approach reduced the number of aligned species to 17, demonstrating that even among the more closely related placental vertebrates large differences within the exon/intron architecture exist. The filtered alignments were then evaluated for the presence or absence of synonymous mutations that were introduced by our high-throughput approach (Fig. 4*B*). Interestingly, none of the synonymous mutations that caused significant decreases in the fitness index were found in this alignment data, whereas synonymous mutants that do not decrease the fitness index value were admissible (Fig. 4*B*). This perfect correlation between admissible/inadmissible synonymous mutations and their fitness index values suggests that the exon/intron architecture filtered conservation evaluation at synonymous positions can be utilized to determine exon posi-

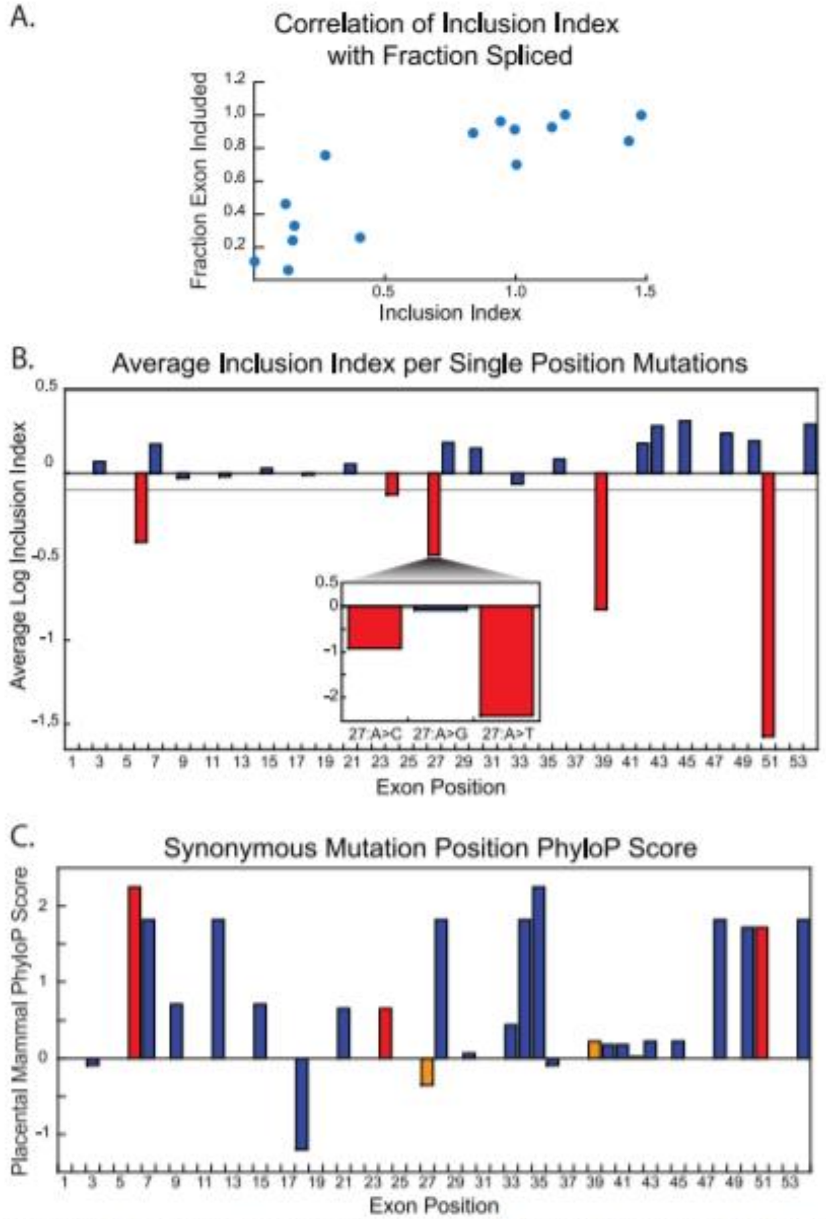
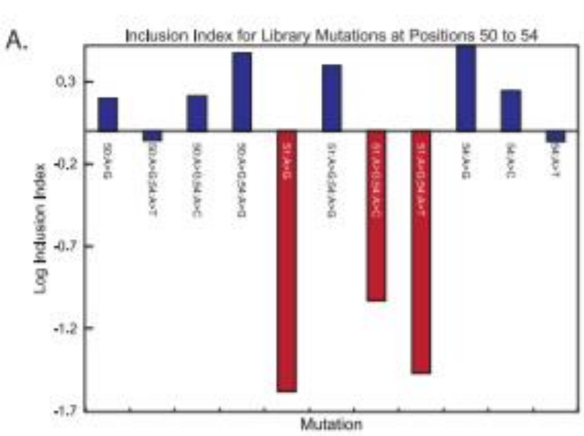


FIGURE 2. **Analysis of library mutants with a single mutation.** *A*, correlation between fitness index values derived from the high-throughput analysis and exon inclusion levels of individual mutants as determined by PCR analysis. *B*, average fitness index values for library mutants with mutations at a single hexamer position. The *gray line* represents the significance cutoff. Mutations that caused a significant decrease in the fitness index value are highlighted in *red*. All other single mutant positions are in blue. The *inset* is an example of the nucleotide-specific mutational effects observed. Specific mutations at a given position can display large differences in the fitness index as illustrated by position 27. *C*, conservation scores at synonymous mutation positions across *SMN1* exon 7 based on placental mammal PhyloP scores. *Bars in red* denote high conservation at the same position where mutations cause significant decreases in the fitness index. *Bars in orange* show the positions where conservation is low and where mutations cause significant decreases in the fitness index.

tions important for efficient splicing, at least in the context of SMN exon 7.

*Hexamer Library Mutations*—Mutations of more than one nucleotide position within a hexamer appear to either rescue or to exacerbate the fitness index changes seen in the single mutant variations (Fig. 5 and Table 1). By themselves, mutations 24:A>G, and 27:A>C/T cause significant decreases in the fitness index value, but surrounding positions (21, 28, and 30) do not (Fig. 5*A*). When mutated in combination, synergistic decreases in the fitness index are observed, even in the case of the mutation 24:A>G with 27:A>G (not significant on its own, but significant with position 24). Conversely, when paired with



B.

MES of mutants at 54:

| | | |
|---|---|---|
| WT | (GGA/GAAATG) | MAXENT: -10.98 |
| A>G | (GGG/GAAATG) | MAXENT: -3.34 |
| A>C | (GGC/GAAATG) | MAXENT: -10.02 |
| A>T | (GGT/GAAATG) | MAXENT: -11.85 |

FIGURE 3. **Analysis of library mutants at the 5′ splice site of *SMN1* exon 7/intron 7.** *A*, fitness index values for mutants at the last positions of exon 7. Mutations that caused a significant decrease in the fitness index value are highlighted in *red*. All other mutations are in *blue*. *B*, maximum entropy scores for the wild-type (WT) 5′ splice site and the mutations at position 54, which is the last exonic nucleotide. Greater maximum entropy score correlate with more efficient splice site usage. The *slash* in the sequences denotes the exon-intron junction.

mutations at positions 28, 29, or 30, the affects of position 27 are decreased, although not always to the point of being not significant (except in the case of 27:A>C, 30:G>A where there was no appreciable change). These results suggest that neighboring mutations within the context of a hexamer can influence exon inclusion in multiple ways. These data and others (Figs. 3 and 5) suggest that specific internal sequences in the exon are vital for optimal exon inclusion in a hexamer specific context.

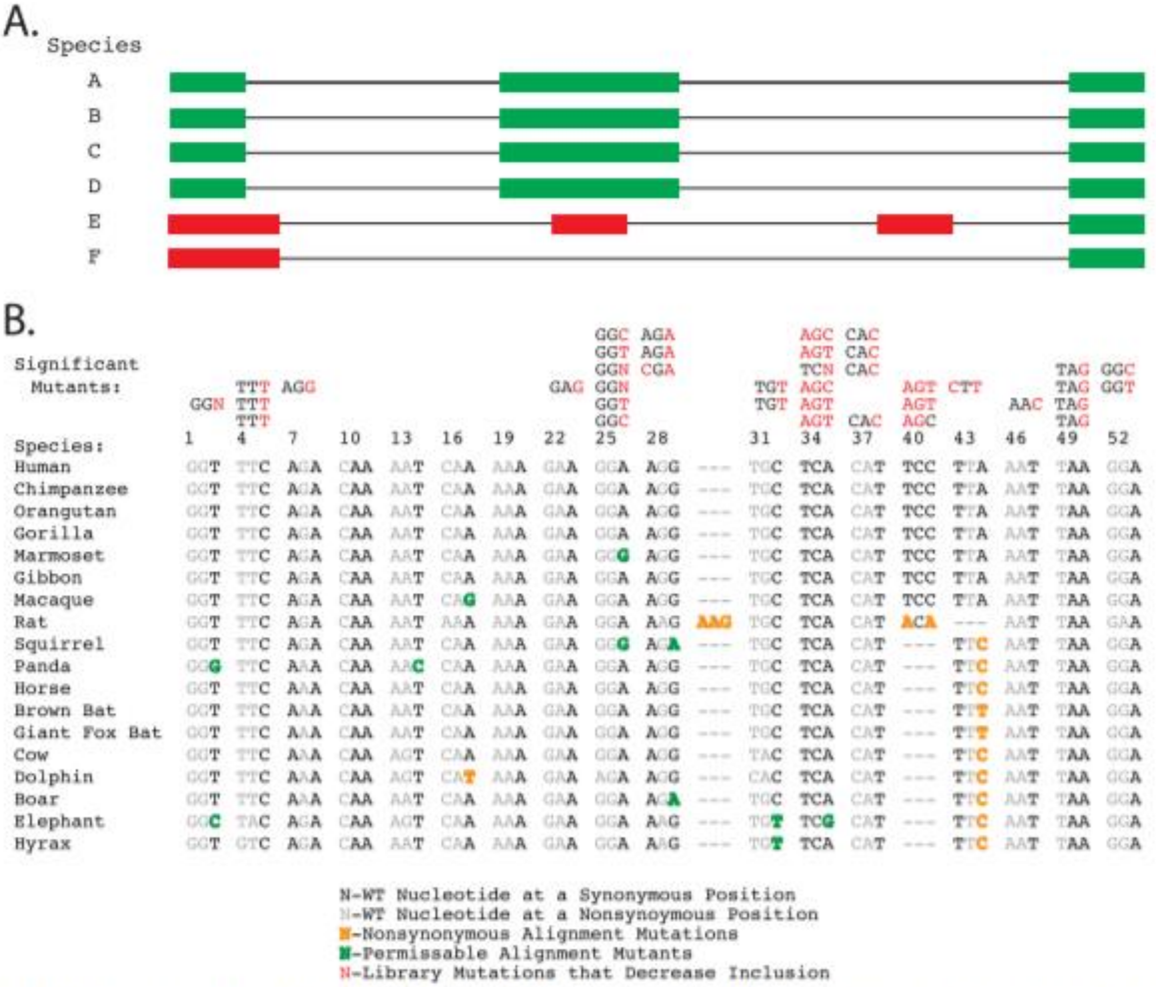
The extensive amount of study devoted to the *SMN1* gene allowed multiple built-in controls for our analysis by correlating changes in regions that are known to be important binding sites for splicing regulatory elements with changes observed in the fitness index (Fig. 5*C*) (4, 13, 24–27, 30–32). At position 3, none of the 3 possible single silent mutations alter the fitness index appreciably (Fig. 5*B*). However, when combined with the C to T mutation at position 6 exon skipping is exasperated, consistent with the notion that the binding site for splicing regulators are disrupted or that more stable RNA secondary structures are formed (4, 13, 24, 25, 27, 30–32). The same synergy is also observed when mutations in positions 6 and 9 are combined. However, if position 7 is mutated along with positions 6 or 9, the reduction in exon inclusion is no longer significant. This result reproduces the finding that the A to C mutation at position 7 rescues the C to T mutation at position 6 either through inhibiting the binding of Sam68, or by preventing the binding of another inhibitory protein (33, 34). Similar examples of synergistic or compensatory mutations are observed to occur at the Tra2-β1 (SRSF10) binding site (positions 18–28), at the 3′ inhibitory secondary structure (positions 33–42), and at the 5′ splice site (21, 23, 32, 35) (Figs. 2, 3, and 5, Table 1). These mutations all involve sequence-specific compensatory mutations that can relieve the decrease in the fitness index caused by other mutations.

When multiple mutations within a hexamer are analyzed using the intron/exon architecture filtered alignment method described above, none of the mutations that cause negative changes in the fitness index were represented in organisms with similar intron/exon architecture. The alignment showed 11 possible synonymous mutations, all of which appeared in our analysis as non-significant changes to the fitness index (Fig. 4*B*). This observation supports the conclusion that splicing-detrimental hexamer mutants are selected against due to splicing-related sequence constraints.

*Analysis of Synonymous SNPs*—The filtered phylogenetic comparison approach may allow the identification of genetic mutations that have a high probability to alter pre-mRNA splicing. To test this hypothesis we applied the modified alignment analysis (Fig. 4*A*) to evaluate whether disease-associated synonymous SNPs are likely to be associated with inducing changes in pre-mRNA splicing. Forty human disease-associated synonymous SNPs and 40 randomly selected synonymous SNPs from the Ensembl! Perl API were combined as a representation of genomic SNPs in coding regions (20, 36). The genomic sequence within a ±5 nt window around the SNP was compared with the homologous sequence of aligned organisms containing similar exon/intron architecture. This sequence window was analyzed for the occurrence of the human SNP nucleotide and surrounding nucleotide changes (supplemental Table S1).

The SNPs were divided into two categories based on the appearance of the evaluated SNP: neutral SNPs and putative splice altering SNPs. There were also non-conserved SNPs that had fewer than 5 organisms in the species alignment after the exon/intron architecture filtering. These cases were not considered further due to limited alignment information. Neutral SNPs were defined as SNPs that were found in the alignment one or more times without additional nucleotide changes in the surrounding sequence. The presence of sequences identical to such human SNPs in other species suggests that they are not selected against and, thus, are likely to be splicing neutral. Putative splice altering SNPs were defined as SNP sequences that did not occur in the alignment or only occurred in conjunction with other mutations in the surrounding hexamer sequence. As was argued by our experimental analysis, such a representation of human SNP sequence across species suggests that nucleotide changes at this position may alter exon inclusion.

In 15 of 80 SNPs there were fewer than 5 organisms that maintained intron/exon architecture conservation (Table 3). Another 36 alignments had multiple occurrences of the human SNP sequence without mutations in the surrounding hexamers, suggesting that these synonymous position mutations were not selected against (Table 3). In 10 of 80 cases the synonymous SNP was not observed in any aligned organism, providing circumstantial evidence that the SNP may alter splicing efficiencies, *i.e.* the SNP is selected against for that particular intron/exon architecture. In 19 cases the SNP mutation was observed in the aligned species, however, only in conjunction with additional nucleotide changes within the 5-nucleotide flanking window (Table 3). Thus, 45% of the SNPs surveyed are potentially


A.
Species
A
B
C
D
E
F
B.
N-WT Nucleotide at a Synonymous Position
N-WT Nucleotide at a Nonsynoymous Position
N-Nonsynonymous Alignment Mutations
N-Permissable Alignment Mutants
N-Library Mutations that Decrease Inclusion

| Species: | 1 | 4 | 7 | 10 | 13 | 16 | 19 | 22 | 25 | 28 | | 31 | 34 | 37 | 40 | 43 | 46 | 49 | 52 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Significant Mutants: | GGN | TTT<br>TTT<br>TTT | AGG | | | | | GAG | GGC<br>GGT<br>GGN<br>GGN<br>GGT<br>GGC | AGA<br>AGA<br>CGA | | TGT<br>TGT | AGC<br>AGT<br>TCN<br>AGC<br>AGT<br>AGT | CAC<br>CAC<br>CAC<br>CAC | AGT<br>AGT<br>AGC | CTT | AAC | TAG<br>TAG<br>TAG<br>TAG | GGC<br>GGT |
| Human | GGT | TTC | AGA | CAA | AAT | CAA | AAA | GAA | GGA | AGG | --- | TGC | TCA | CAT | TCC | TTA | AAT | TAA | GGA |
| Chimpanzee | GGT | TTC | AGA | CAA | AAT | CAA | AAA | GAA | GGA | AGG | --- | TGC | TCA | CAT | TCC | TTA | AAT | TAA | GGA |
| Orangutan | GGT | TTC | AGA | CAA | AAT | CAA | AAA | GAA | GGA | AGG | --- | TGC | TCA | CAT | TCC | TTA | AAT | TAA | GGA |
| Gorilla | GGT | TTC | AGA | CAA | AAT | CAA | AAA | GAA | GGA | AGG | --- | TGC | TCA | CAT | TCC | TTA | AAT | TAA | GGA |
| Marmoset | GGT | TTC | AGA | CAA | AAT | CAA | AAA | GAA | GGG | AGG | --- | TGC | TCA | CAT | TCC | TTA | AAT | TAA | GGA |
| Gibbon | GGT | TTC | AGA | CAA | AAT | CAA | AAA | GAA | GGA | AGG | --- | TGC | TCA | CAT | TCC | TTA | AAT | TAA | GGA |
| Macaque | GGT | TTC | AGA | CAA | AAT | CAG | AAA | GAA | GGA | AGG | --- | TGC | TCA | CAT | TCC | TTA | AAT | TAA | GGA |
| Rat | GGT | TTC | AGA | CAA | AAT | AAA | AAA | GAA | GGA | AAG | AAG | TGC | TCA | CAT | ACA | --- | AAT | TAA | GAA |
| Squirrel | GGT | TTC | AGA | CAA | AAT | CAA | AAA | GAA | GGG | AGA | --- | TGC | TCA | CAT | --- | TTC | AAT | TAA | GGA |
| Panda | GGG | TTC | AAA | CAA | AAC | CAA | AAA | GAA | GGA | AGG | --- | TGC | TCA | CAT | --- | TTC | AAT | TAA | GGA |
| Horse | GGT | TTC | AAA | CAA | AAT | CAA | AAA | GAA | GGA | AGG | --- | TGC | TCA | CAT | --- | TTC | AAT | TAA | GGA |
| Brown Bat | GGT | TTC | AAA | CAA | AAT | CAA | AAA | GAA | GGA | AGG | --- | TGC | TCA | CAT | --- | TTT | AAT | TAA | GGA |
| Giant Fox Bat | GGT | TTC | AAA | CAA | AAT | CAA | AAA | GAA | GGA | AGG | --- | TGC | TCA | CAT | --- | TTT | AAT | TAA | GGA |
| Cow | GGT | TTC | AAA | CAA | AGT | CAA | AAA | GAA | GGA | AGG | --- | TAC | TCA | CAT | --- | TTC | AAT | TAA | GGA |
| Dolphin | GGT | TTC | AAA | CAA | AGT | CAT | AAA | GAA | AGA | AGG | --- | CAC | TCA | CAT | --- | TTC | AAT | TAA | GGA |
| Boar | GGT | TTC | AAA | CAA | AAT | CAA | AAA | GAA | GGA | AGA | --- | TGC | TCA | CAT | --- | TTC | AAT | TAA | GGA |
| Elephant | GGC | TAC | AGA | CAA | AGT | CAA | AAA | GAA | GGA | AAG | --- | TGT | TCG | CAT | --- | TTC | AAT | TAA | GGA |
| Hyrax | GGT | GTC | AGA | CAA | AAT | CAA | AAA | GAA | GGA | AAG | --- | TGT | TCA | CAT | --- | TTC | AAT | TAA | GGA |

FIGURE 4. **Sequence alignment of *SMN1* exon 7 with 17 animals with similar intron/exon architecture.** *A*, a graphical representation of the method used to filter species alignments based on conservation of intron/exon architecture. Species E and F would not be included in an architecture-filtered alignment. *B*, the alignment of similar mammals that pass the intron/exon architecture filter to human *SMN1* exon 7. The mutations that caused significant decreases in the fitness index value are listed *above* their position in the alignment with the specific hexamer mutations in *red*. The alignment below the hexamers shows the silent mutation positions that diverge from the human sequence. The mutations that did not cause changes in the fitness index value are highlighted in *green*. The color-coding used to differentiate the effects of the observed mutants is shown in the key on the figure bottom. The category "nonsynonymous alignment mutations" refers to positions in the multispecies alignment where observed nucleotide substitutions would be considered nonsynonymous mutations in the human exon.

splice altering SNPs. These observations suggest that in many cases of sequence evolution, nucleotide changes within the hexamer context are driven by the need to uphold splicing efficiency.

To determine whether synonymous SNPs in coding exons exhibit unique selection pressures, we prepared a dataset of 162 SNPs from noncoding exons or from single exon genes (supplemental Table S1). These exons should either only be under evolutionary selection for correct splicing or, in the case of single exon genes, should only be under selection for optimal protein coding. Because of their lack of protein coding constraints, noncoding exons would be expected to have more freedom to mutate nucleotides throughout the exon, not just in silent positions. For these exons it is expected that the selection on synonymous substitution positions would be reduced. Single exon genes, on the other hand, should have a different selective pressure on their synonymous mutation positions that are strictly based on protein coding within the reading frame. We analyzed these 162 control SNPs in the same manner as described above and found that single exon and non-coding genes had a significantly lower number of putative splice altering SNPs at synonymous positions, as expected from the relaxation of evolutionary pressures (Table 3). We conclude that applying the exon/intron architecture filtered approach of phylogeny aids in the identification of exonic sequence that may be necessary for exon inclusion.

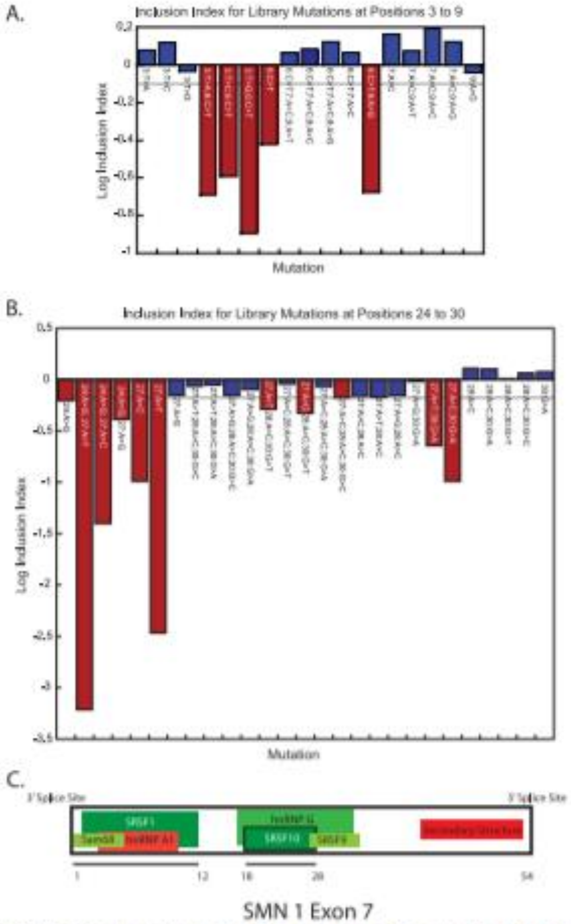


FIGURE 5. **Compensatory, additive, and synergistic mutational effects within hexamers.** *A*, the *bars* represent the fitness index values at positions 3 through 9. Mutations with significant fitness index changes are highlighted in *red*, whereas all others are in *blue*. The *gray line* represents the significance cutoff. *B*, the *bars* represent the fitness index values for mutations made between position 24 and 30. Color-coding follows the same convention as in *A*. *C*, model of known protein interactions across *SMN* exon 7. *Boxes* here depict a schematic of *SMN* exon 7 and proteins that have been shown to interact with the exon and with each other to modulate exon 7 inclusion. Elements that are assumed to increase exon inclusion have been shaded in *green*, whereas those that decrease inclusion are shown in *red* (22, 30).

## Discussion

The work described above outlines a unique approach to identify exonic positions that influence alternative pre-mRNA splicing. Through the creation of a library of synonymous mutations impacting splicing of a single exon coupled with deep sequencing, we found that 23% of those synonymous mutations affected exon inclusion significantly in SMN exon 7. Several of the *SMN1* exon 7 mutations evaluated behaved as expected from previous work, highlighting the reliability of our experimental approach. We expected that synonymous positions that significantly alter exon inclusion when mutated would be conserved. However, such positions were not well conserved by measure of PhyloP scores across 46 vertebrates. This prompted us to filter the species used for the evolutionary comparison based on similar intron/exon architecture. With this filter in place, none of the exon inclusion reducing mutations was found in the species alignment. Additionally, mutations that decreased exon inclusion were often rescued back to wild-type inclusion levels by other mutations within a 5-nt radius, suggesting an ability to modulate exon inclusion through the use of compensatory mutations within a hexameric sequence space. The selection of the hexamer as a sequence space for testing combinatorial mutations was based on the size of known RNA-binding protein footprints (22). Although this framework may have been limiting, its use here clearly demonstrates the importance of incorporating analyses of local RNA sequence contexts when evaluating exon inclusion efficiencies and evolutionary flexibility.

The modified alignment approach was used to demonstrate that up to 45% of verified synonymous SNPs might be splice altering. We propose that the exon/intron architecture-based method of species alignment is much more likely to attribute functional significance to sequence elements involved in splicing.

*Evolutionary Selection Against Splicing Mutants*—Our systematic approach to determine the effect of synonymous mutations on exon inclusion allowed us to comparatively interrogate alignment data in a splicing centric context. The finding that mutations that cause significant reduction of the fitness index were not represented in multiple comparative alignments points to an influence of nucleotide selection by splicing. The presence of splicing regulatory elements at many of these conserved synonymous positions further suggests that there has been positive selection throughout evolution for specific splicing events. Positive selection due to splicing within splicing regulatory elements and around intron/exon junctions has been previously suggested and documented, but only recently experimentally analyzed by testing for missense and nonsense SNPs associated with disease (4, 7, 37–39). However, positive selection is also at work in the genetic code, masking changes that may have occurred due to splicing. Starting with a library of mutants at synonymous positions and filtering phylogenetic data based on the exon/intron architecture allowed us to determine the interesting compensatory nature of evolution to favor a specific functional transcript and to find a way to add information through small nucleotide changes.

Loss of full-length SMN transcript has been shown to be problematic for multiple organisms and its splicing regulation seems to be relatively consistent across species (31). Many examples of mutations affecting splicing and causing disease have been described (4, 20, 37, 40, 41). Despite this, SNPs are poorly characterized when it comes to splicing. They are frequently misattributed as missense or frameshift mutations due to the SNP not being found in close proximity to an intron/exon junction (4, 40, 42). Additionally, synonymous SNPs have only recently been studied in earnest as causes of disease after being found to follow non-neutral evolution (8–10, 37, 41). Our work demonstrated that 23% of synonymous mutation positions in the evaluated coding exon play a role in their inclusion, comparable with previous work (38, 43). The ability to predict how a synonymous SNP may alter pre-mRNA maturation could

**TABLE 3**

**Summary of SNPs analyzed by filtering phylogenetic alignments for similar intron/exon architecture**

"Splice altering SNPs" represent SNPs that did not occur in the alignment of organisms with similar intron/exon architecture or, if they occurred, only occurred alongside other mutations within a 6-mer (possible compensatory mutations). "Neutral SNPs" were SNPs that had SNP occurrences without mutations in the surrounding hexamer. The classes of coding exons and exons derived from non-coding genes or single exon genes' exons were found to be independent by a $\chi^2$ test ($p < 0.05$). The SNP identities for all groups are listed in supplemental Table S1.

| Data set | Splice altering SNPs | Neutral SNPs |
|---|---|---|
| Disease associated and random synonymous SNPs | 45% (29) | 55% (36) |
| Randomly selected noncoding exonic or single exon SNPs | 30% (34) | 70% (80) |
| $p < 0.05$ between the two SNP datasets | | |

prove invaluable in the determination of molecular causes of genetic disease. Our method using splicing centric filtered alignments helps to assign the importance of a particular SNP in changing pre-mRNA splicing.

*Functional Filtering of Alignments to Enrich for Splicing Related Information*—Previous discovery and analysis of splicing regulatory elements and splice altering SNPs relied on conservation data. Our analysis suggests that the limitations of this approach will cause important splicing regulatory positions to be missed. An initial comparison between our mutation outcomes and PhyloP conservation scores did not support the notion that fixation of synonymous exon positions are driven by splicing constraints. In part, this observation is explained by the fact that PhyloP scoring is not able to predict the nucleotide-specific alterations in exon inclusion that occurred in our dataset. Our analysis shows that the use of sequence identity analysis instead of divergence from a specific nucleotide is important when evaluating splicing. Limiting our comparative alignment to similar intron/exon architectures presumably enriched our search space for organisms that regulate exons in a similar fashion, a notion supported by our SNP analysis, which demonstrated that up to 45% of exonic synonymous SNPs are likely to be splice-altering SNPs.

Here, we evaluated the influence of nucleotide identities at synonymous positions on the overall splicing efficiency within the context of the 54-nt long SMN exon 7. We chose this exon mainly because its disease association sparked extensive studies that identified many exonic positions as regulator binding sites that have been shown to be important for splicing, thus providing a knowledge base essential to explain any observed fitness index differences (4, 13, 24–27, 30–32). At 54 nucleotides in length, SMN exon 7 is significantly smaller than the average length of a human exon (~120 nt). Thus, it is possible that the density of splicing regulatory sites embedded within SMN exon 7 is greater than for human exons of average size. Given these considerations, it is possible that average sized human exons may be exposed to different purifying selection pressures.

Synonymous mutations can alter the ability of an mRNA to code for protein through creation of regulatory defects. They have been shown to alter protein-folding abilities, to change RNA stability, and to induce exon loss (8, 41, 44). These observations suggest that selective pressures to maintain efficient splicing work in tandem with selective pressures to uphold all aspects of the genetic code. In our attempt to uncouple these influences we evaluated the influence of synonymous mutations on pre-mRNA splicing. It is also appreciated that a codon usage bias exists. Lower organisms experience a correlation between codon usage and tRNA copy number, suggesting an optimization for particular codons for efficient translation (45). Interestingly, codon usage in mammals does not correlate with tRNA abundance or tRNA gene copy number, an observation distinctly different from other taxa (8, 45). This deviation, along with this work, suggest that whereas codon bias may influence the final sequence of an exon, it is part of a larger evolutionary picture that includes influences from other processes, such as pre-mRNA splicing, to properly produce functional proteins.

*Author Contributions*—Experimental design was conceived by K. J. H. and W. F. M. Cloning and construction of plasmids was carried out by L. L. and W. H. Library prep and data analysis was done by W. F. M and A. G. Manuscript preparation was completed by W. F. M. and K. J. H.

*Acknowledgments—We thank W. Augustine Dunn III and Anke Busch for their discussions and advice, and Tom Caldwell for technical assistance.*

## References

1. Nilsen, T. W., and Graveley, B. R. (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463,** 457–463
2. Black, D. L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72,** 291–336
3. Hertel, K. J. (2008) Combinatorial control of exon recognition. *J. Biol. Chem.* **283,** 1211–1215
4. Cartegni, L., Chew, S. L., and Krainer, A. R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.* **3,** 285–298
5. Fox-Walsh, K. L., and Hertel, K. J. (2009) Splice-site pairing is an intrinsically high fidelity process. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 1766–1771
6. Willie, E., and Majewski, J. (2004) Evidence for codon bias selection at the pre-mRNA level in eukaryotes. *Trends Genet.* **20,** 534–538
7. Parmley, J. L., Chamary, J. V., and Hurst, L. D. (2006) Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol. Biol. Evol.* **23,** 301–309
8. Chamary, J. V., Parmley, J. L., and Hurst, L. D. (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* **7,** 98–108
9. Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T., and Ast, G. (2006) Comparative analysis identifies exonic splicing regulatory sequences: the complex definition of enhancers and silencers. *Mol. Cell* **22,** 769–781
10. Pagani, F., Raponi, M., and Baralle, F. E. (2005) Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 6368–6372
11. Geib, T., and Hertel, K. J. (2009) Restoration of full-length SMN promoted by adenoviral vectors expressing RNA antisense oligonucleotides embedded in U7 snRNAs. *PLoS ONE* **4,** e8204
12. Madocsai, C., Lim, S. R., Geib, T., Lam, B. J., and Hertel, K. J. (2005) Correction of SMN2 Pre-mRNA splicing by antisense U7 small nuclear RNAs. *Mol. Ther.* **12,** 1013–1022
13. Lim, S. R., and Hertel, K. J. (2001) Modulation of survival motor neuron

pre-mRNA splicing by inhibition of alternative 3' splice site pairing. *J. Biol. Chem.* **276,** 45476–45483

14. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25
15. Mailman, M. D., Heinz, J. W., Papp, A. C., Snyder, P. J., Sedra, M. S., Wirth, B., Burghes, A. H., and Prior, T. W. (2002) Molecular analysis of spinal muscular atrophy and modification of the phenotype by SMN2. *Genet. Med.* **4,** 20–26
16. Monani, U. R., Lorson, C. L., Parsons, D. W., Prior, T. W., Androphy, E. J., Burghes, A. H., and McPherson, J. D. (1999) A single nucleotide difference that alters splicing patterns distinguishes the SMA gene SMN1 from the copy gene SMN2. *Hum. Mol. Genet.* **8,** 1177–1183
17. Lorson, C. L., Hahnen, E., Androphy, E. J., and Wirth, B. (1999) A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proc. Natl. Acad. Sci. U.S.A.* **96,** 6307–6311
18. Hofmann, Y., Lorson, C. L., Stamm, S., Androphy, E. J., and Wirth, B. (2000) Htra2-β 1 stimulates an exonic splicing enhancer and can restore full-length SMN expression to survival motor neuron 2 (SMN2). *Proc. Natl. Acad. Sci. U.S.A.* **97,** 9618–9623
19. Maquat, L. E. (2004) Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat. Rev. Mol. Cell Biol.* **5,** 89–99
20. Sauna, Z. E., and Kimchi-Sarfaty, C. (2011) Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.* **12,** 683–691
21. Singh, N. N., Singh, R. N., and Androphy, E. J. (2007) Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes. *Nucleic Acids Res.* **35,** 371–389
22. Graveley, B. R., and Hertel, K. J. (2005) SR Proteins. in *Encyclopedia of Life Sciences*, John Wiley and Sons, Ltd., Chichester
23. Singh, R. N. (2007) Evolving concepts on human SMN Pre-mRNA splicing. *RNA Biol.* **4,** 7–10
24. Lorson, C. L., and Androphy, E. J. (2000) An exonic enhancer is required for inclusion of an essential exon in the SMA-determining gene SMN. *Hum. Mol. Genet.* **9,** 259–265
25. Miyaso, H., Okumura, M., Kondo, S., Higashide, S., Miyajima, H., and Imaizumi, K. (2003) An intronic splicing enhancer element in survival motor neuron (SMN) pre-mRNA. *J. Biol. Chem.* **278,** 15825–15831
26. Lefebvre, S., Bürglen, L., Reboullet, S., Clermont, O., Burlet, P., Viollet, L., Benichou, B., Cruaud, C., Millasseau, P., and Zeviani, M. (1995) Identification and characterization of a spinal muscular atrophy-determining gene (see comments). *Cell* **80,** 155–165
27. Kashima, T., and Manley, J. L. (2003) A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. *Nat. Genet.* **34,** 460–463
28. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20,** 110–121
29. Pollard, K. S., Salama, S. R., King, B., Kern, A. D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J. S., Bejerano, G., Baertsch, R., Rosenbloom, K. R., Kent, J., and Haussler, D. (2006) Forces shaping the fastest evolving regions in the human genome. *PLoS Genet.* **2,** e168
30. Lorson, C. L., Rindt, H., and Shababi, M. (2010) Spinal muscular atrophy: mechanisms and therapeutic strategies. *Hum. Mol. Genet.* **19,** R111–118
31. Singh, N. N., Androphy, E. J., and Singh, R. N. (2004) The regulation and regulatory activities of alternative splicing of the SMN gene. *Crit. Rev. Eukaryot. Gene Expr.* **14,** 271–285
32. Singh, N. N., Androphy, E. J., and Singh, R. N. (2004) In vivo selection reveals combinatorial controls that define a critical exon in the spinal muscular atrophy genes. *RNA* **10,** 1291–1305
33. Cartegni, L., Hastings, M. L., Calarco, J. A., de Stanchina, E., and Krainer, A. R. (2006) Determinants of exon 7 splicing in the spinal muscular atrophy genes, SMN1 and SMN2. *Am. J. Hum. Genet.* **78,** 63–77
34. Pedrotti, S., Bielli, P., Paronetto, M. P., Ciccosanti, F., Fimia, G. M., Stamm, S., Manley, J. L., and Sette, C. (2010) The splicing regulator Sam68 binds to a novel exonic splicing silencer and functions in SMN2 alternative splicing in spinal muscular atrophy. *EMBO J.* **29,** 1235–1247
35. Hoffman, B. E., and Grabowski, P. J. (1992) U1 snRNP targets an essential splicing factor, U2AF65, to the 3' splice site by a network of interactions spanning the exon. *Genes Dev.* **6,** 2554–2568
36. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311
37. Teng, M., Wang, Y., Wang, G., Jung, J., Edenberg, H. J., Sanford, J. R., and Liu, Y. (2011) Prioritizing single-nucleotide variations that potentially regulate alternative splicing. *BMC Proc.* **5,** S40
38. Sterne-Weiler, T., Howard, J., Mort, M., Cooper, D. N., and Sanford, J. R. (2011) Loss of exon identity is a common mechanism of human inherited disease. *Genome Res.* **21,** 1563–1571
39. Ke, S., Zhang, X. H., and Chasin, L. A. (2008) Positive selection acting on splicing motifs reflects compensatory evolution. *Genome Res.* **18,** 533–543
40. Krawczak, M., Reiss, J., and Cooper, D. N. (1992) The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet.* **90,** 41–54
41. Sauna, Z. E., Kimchi-Sarfaty, C., Ambudkar, S. V., and Gottesman, M. M. (2007) The sounds of silence: synonymous mutations affect function. *Pharmacogenomics* **8,** 527–532
42. Faustino, N. A., and Cooper, T. A. (2003) Pre-mRNA splicing and human disease. *Genes Dev.* **17,** 419–437
43. Chen, R., Davydov, E. V., Sirota, M., and Butte, A. J. (2010) Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PLoS ONE* **5,** e13574
44. Kimchi-Sarfaty, C., Oh, J. M., Kim, I. W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V., and Gottesman, M. M. (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* **315,** 525–528
45. Plotkin, J. B., and Kudla, G. (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* **12,** 32–42