Abstract

Coordination Between Pre-mRNA Splicing and Cleavage in Budding Yeast

Tara Alpert

2020

Splicing is an essential step in gene expression, and modulators often target the process of splicing to regulate sets of genes. Furthermore, splicing is intricately coordinated with other co-transcriptional processes including transcription elongation, cleavage, and termination. These processes are, in turn, dependent on splicing as evident by the increased transcriptional output of a gene when an intron is inserted. My PhD work has focused on investigating the coordination between splicing and transcription from several different angles. To do so, I use advanced sequencing technologies to capture single molecules of RNA while they are actively being synthesized (nascent RNA). These technologies include 1) Illumina short-read sequencing which uses a fluorescent sequencing-by-synthesis approach and 2) Oxford Nanopore Technologies long-read sequencing which directly identifies the oligonucleotide base by detecting fluctuations in electric in current as the oligonucleotide passes through a nanopore. Methods development was a large component of my thesis work, as both techniques that I employed required optimization.

Previously, our lab identified extensive heterogeneity of splicing kinetics in budding yeast. My investigation of how splicing is coordinated with other transcriptional processes capitalized on this heterogeneity to train a computational model for the identification of gene-specific splicing regulators. Splicing kinetics were first measured with great precision by two previous graduate students from our lab, who developed a new technique calling

Single Molecule Intron Tracking or SMIT. This technique uses paired-end Illumina sequencing to sequence both ends of nascent RNA, recording both the splicing status of the molecule and the position of RNA Polymerase II (Pol II). They then computed the fraction spliced of each intron as a function of Pol II position and observed the nascent RNA progress from unspliced to spliced as Pol II elongates. The observed heterogeneity in splicing kinetics exceeded expectations of the relatively simple yeast genome. A machine learning model was trained to use gene-specific features present at each locus (such as RNA binding proteins, histone modifications, Pol II density, etc.) to predict splicing kinetic parameters. A number of features were identified as strong predictors of co-transcriptional splicing. I disrupted these features using either genetic deletions or depletions and performed SMIT in this context to observe what effect the protein was having on splicing kinetics. Many of the predicted proteins had no specific effect on splicing kinetics, pointing to the resilient nature of splicing. Intriguingly, a conserved polyA binding protein, Nab2, was found to play a previously undocumented role in proper cleavage and termination of nascent RNA transcripts, and we show how perturbance of this 3′ end processing results in changes to co-transcriptional splicing of nearby genes.

Furthermore, I discovered that failure to splice co-transcriptionally is tightly correlated with transcriptional readthrough of the same gene in budding yeast. During the course of the SMIT assays, a strong correlation was observed between the failure of an RNA molecule to splice and the failure to cleave at the 3′ end of the gene. RNA cleavage promotes proper Pol II termination at the polyadenylation site, and when this cleavage does not occur the polymerase continues transcribing downstream of the gene and can continue reading through downstream genes. I became interested in determining whether failure to

splice or cleave was determinant of the other. Is failing to splice sufficient to cause failure of cleavage? I targeted several components of both the splicing and cleavage machinery in yeast for degradation using an auxin-mediated degron tag. After depleting my targets, I sequenced the nascent RNA on the Nanopore long read sequencing platform. I found that inhibition of splicing does indeed inhibit cleavage, a finding that will alter our understanding of how co-transcriptional splicing impacts gene expression. Additionally, I discovered that failure to cleave does not inhibit splicing of the upstream intron, but an overall splicing defect is observed in cleavage-deficient strains because introns present downstream of the failed cleavage site often go unspliced more than their properly initiated counterparts, agreeing with my earlier reports of Nab2.

Overall, my thesis has addressed the coordination between transcription in splicing using both an unbiased machine-learning based approach and a targeted, observation driven strategy. I found Nab2 to be a key regulator of splicing and cleavage, and I discovered that splicing can directly impact the fidelity of 3′ end processing. I have identified new ways in which splicing and transcription are intimately associated and my findings will help further our understanding of how gene expression is achieved.

Coordination Between Pre-mRNA Splicing and Cleavage in Budding Yeast

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
In Candidacy for the Degree of
Doctor of Philosophy

by
Tara Alpert

Dissertation Director: Karla Neugebauer

May 2020

Table of Contents

List of Figures

List of Tables

Acknowledgments

This thesis would not have been possible without the tremendous support of my mentors, my colleagues, my department, my friends, and my family. I would like to thank everybody who contributed to my well-being during this time from the bottom of my heart. I would not have made it through without your scientific guidance, your sympathetic ear, your enthusiastic encouragement, and your unwavering faith in me.

Karla, you have shown me what it is to be a scientist. You have taught me how to navigate the academic world as a woman, and lifted me up when I doubted myself. You showed me that we are all human in a career that demands perfection. Most importantly, you truly cared about my success and well-being, and have done everything in your power to support my dreams. Thank you.

Dr. Joan Steitz and Dr. Mark Hochstrasser have been instrumental in the development of these experiments which have progressed from loose ideas to concrete conclusions under their guidance as my thesis committee. Mark generously provided me with the gene deletion and auxin-inducible degron yeast strains that were used in both chapters of this thesis, and continued to advise me while I learned about using yeast as a model organism. Joan has been awarded some of the highest honors available in our field for her outstanding scientific achievements and her dedication to training young scientists, so it comes as no surprise that I look up to her as an example. As her teaching assistant, I was inspired by her dedication to teaching and learned her strategies for engaging students in discussion and guiding them through material. Her thoughtful questions and attention to detail in my presentations and committee meetings were instrumental to my scientific development. I

am very grateful to have had both Mark and Joan as scientific mentors through my dissertation work.

My Neugebauer lab mates have made coming into lab every day a true joy. They have been my family here in Connecticut. Their scientific input, their understanding and support, and their laughter have not only been vital to finishing this degree, but to thriving in all areas of my life. I would like to say a special thank you to Korinna Straube who contributed an enormous amount of data to this thesis, and became one of my best friends along the way. Dahyana Arias Escayola, thank you for listening to me without judgment and always being there with positivity and understanding. Edward Courchaine, when you go to battle alongside someone it creates bonds that can never be broken, and I cannot imagine anybody better to be in the trenches with. We did this together. David Phizicky, your drive and love for science set a new bar for what I could see myself achieving through your example. Thank you to splicing group past and present, Kirsten, Tucker, and Lydia for the suggestions and critiques that pushed my project in the right directions. You are all geniuses.

Thank you to my fellow MB&B graduate students for remaining a source of support, advice, and fun memories. You pushed me to expand my horizons through camping trips, actin seminars, and enlightened scientific and philosophical conversations. You have been a central element of my graduate school experience that I will always cherish.

I would like thank my undergraduate mentor, Joseph Jez, for giving me independence on a project early on, pushing me to apply for fellowships that I didn't even know existed, solving our structures before we even finished beamtime but letting us solve it ourselves anyways, and doing it all with a huge smile. Your push has gotten me where I am.

I have had two spectacular roommates in New Haven that have contributed to my personal development in monumental ways. Thank you to both Estee and Arya for always hearing about my day, for the late-night study sessions, marathons of The Office, and philosophical discussions.

Thank you to my fantastic friends Rosemary, Lauren, and Callie for the long phone calls, messages, and conversations. Each one of you has inspired me with your perseverance and passion, and I'm excited to embark on my next chapter with your support behind me. Whether birth, high school, or college, you have all been there for me since the beginning.

Finally, I want to thank my family for everything and more. They believe in me so strongly that they don't even need to understand what experiments I'm trying to do to know that I can succeed at them. Mom and Dad and Brian, you have supported me the entire way through this long journey, and I absolutely could not have done it without your love and encouraging words. Finally, I'd like to thank my grandparents who are the heart of our family. Grandpa Lester and Grandma Ruth helped inspire and pay for my education at Wash. U. Despite living far away, my grandma followed and documented my every move in meticulous detail from graduating college, working abroad in Israel, and completing my dissertation research, she insisted I describe exactly what I was doing and why while she took notes on pen and paper. They are both dearly missed. Traveling home to the excited hugs from Grandma Sandy and Grandpa Walter was exactly what I needed to recover from long periods in the lab. Hearing stories about pursuing their passion for education, even when times were hard, has absolutely shaped my outlook on life and steered me in the direction of a PhD. I am so grateful for all the love they have shown me.

Publications

My research during this thesis has contributed to the following publications:

Alpert T, Straube K, Carrillo Oesterreich F, Neugebauer KM (submitted) Widespread transcriptional readthrough leads to splicing defects upon Nab2 depletion.

Alpert T, Reimer K, Straube K, Neugebauer KM (accepted) Long read sequencing of nascent RNA from budding and fission yeasts. *Methods in Molecular Biology*.

Herzel L, Ottoz DSM, **Alpert T**, Neugebauer KM (2017) Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. *Nature Rev Mol Cell Biol* 18, 637-650.

Alpert T, Herzel L, Neugebauer KM (2016) Perfect timing: splicing and transcription rates in living cells. *Wiley Interdiscip Rev RNA.*

1. Introduction

Portions of this chapter have been modified from previously published works:

Alpert T, Herzel L, Neugebauer KM (2016) Perfect timing: splicing and transcription rates in living cells. *Wiley Interdiscip Rev RNA.*

Herzel L, Ottoz DSM, **Alpert T**, Neugebauer KM (2017) Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. *Nature Rev Mol Cell Biol* 18, 637-650.

Efficiency of cellular processes is key to survival in an ever-changing environment. For a cell to survive and proliferate, it must coordinate the production of RNAs and proteins that are relevant in that moment of time. Producing a translation-competent RNA involves accessing the DNA, initiating transcription of a given gene, editing that transcript through splicing, cleaving the nascent RNA, terminating transcription, adding untemplated adenosine nucleotides to the 3′ end, and exporting it to the cytoplasm. This is a large energetic undertaking for the cell and mistakes in any process lead to waste and potentially disease. Thus, regulation and coordination of these processes must be tightly controlled.

Pre-mRNA splicing plays a central role in gene expression. Splicing is the removal of non-coding intronic sequences from pre-mRNA and the subsequent ligation of the coding exons that surround it. This two-step transesterification reaction is carried out by a megadalton RNA-based machine called the spliceosome (Plaschka et al., 2019; Wahl et al., 2009). The spliceosome is comprised of five small nuclear RNAs (snRNAs) each bound to a set of proteins ultimately forming small nuclear ribonucleoproteins (snRNPs). Ordered and step-wise assembly of snRNPs must occur de novo on every intron and is a significant

commitment of energy and resources for the cell. My thesis is dedicated to investigating the coordination between pre-mRNA splicing and other nuclear gene expression pathways. Before this investigation can begin, it is crucial to understand how quickly splicing occurs after RNA synthesis because this is the window in which regulation of splicing can occur. This topic has been the subject of immense study for decades in the field of splicing. The majority of splicing occurs simultaneously with transcription, or co-transcriptionally. 75% of introns in *Saccharomyces cerevisiae* are removed co-transcriptionally, and the values are similar in other organisms such as fly (83%) and human (74-85%) (Ameur et al., 2011; Carrillo Oesterreich et al., 2010; Khodor et al., 2011, 2012; Tilgner et al., 2012). The following section will show evidence that the spliceosome is physically close to RNA polymerase II when splicing occurs, highlighting that these processes are closely related.

1.1 **Measuring the rate of splicing**

An important step towards understanding gene regulation is measuring the time necessary for the completion of individual steps. Measurement of reaction rates can reveal potential nodes for regulation. Since the 1980s, numerous model systems and approaches have been used to determine the precise timing of splicing *in vivo*. Because splicing can be co-transcriptional, the position of Pol II when splicing is detected has been used as a proxy for time by some groups, including ours. In addition to these "distance-based" measurements, "time-based" measurements have been possible through live cell imaging, metabolic labeling of RNA, and gene induction. Yet splicing rates can be include the time it takes for transcription, spliceosome assembly and spliceosome disassembly. The variety of assays and systems used has, perhaps not surprisingly, led to reports of widely differing splicing rates *in vivo*. In this section, I will summarize the monumental effort made by the

field over the previous decades to answer the pressing question of when does splicing happen.

1.1.1. **Classic distance-based measures of splicing progression**

Historically, co-transcriptional splicing was discovered and measured by analyzing electron micrographs of chromatin spreads (Osheim et al., 1985). Osheim et al observed two particular species of RNP particles on nascent transcripts, that appeared at predictable sites of early embryo genes from *Drosophila melanogaster.* These particles were predicted to represent subunits of the spliceosome based on several observations including RNA looping between the two particles and eventual disappearance of the particle after loop removal. By measuring the DNA distance in micrometers (μm) from known locations on the gene (Figure 1.1A) and assuming a maximum chromatin compaction of 4.8 kb/μm, this study measured one of the first instances of *in vivo* splicing occurring Pol II was 4.5 kb downstream of the 3′ splice site (Figure 1.2) and began a decades-long investigation (Beyer and Osheim, 1988).

A more modern and versatile tool for acquiring information about co-transcriptional spliceosome assembly is Chromatin Immunoprecipitation (ChIP). Rosbash and colleagues utilized HZ18 reporter genes, which harbor MS2 RNA stem loops in either the intron or expressed as two halves of the MS2 RNA stem loop in its exons, such that the loop forms after splicing (Lacadie et al., 2006). MS2-coat protein binds the stem loop and serves as a target in ChIP experiments that aimed to detect the Pol II position along the gene the moment splicing occurred. MS2 ChIP signal 1.5 kb downstream of the intron suggested that splicing takes place long after intron synthesis (Lacadie et al., 2006). Moreover, these studies employed an intronic

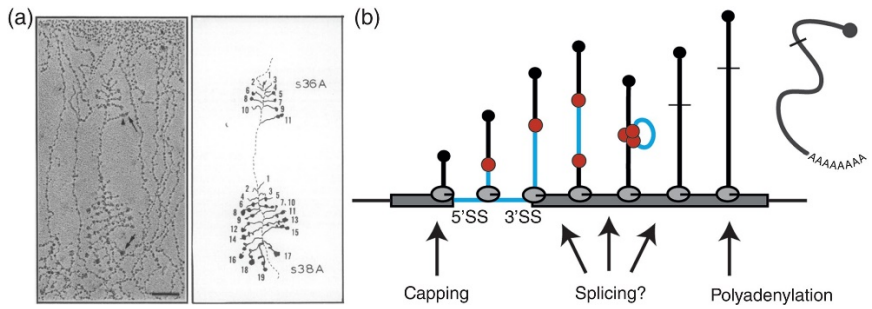

Figure 1.1 Observation of co-transcriptionally spliced transcripts leads to distance-based view of splicing kinetics in vivo
A. Electron micrograph of chromatin spreads from *D. melanogaster* (left panel) visualize electron dense spliceosomes as they assemble near the 5′ ends of nascent transcripts; shortening of transcript 10 is indicative of intron removal. Camera lucida drawing of the chromatin spread is shown in the right panel (Osheim et al., 1985). B. Diagram of a simple gene with a single intron undergoing transcription by several active Pol II molecules. The 5′ methyl cap (black ball) is added shortly after transcription of the 5′ end of the RNA transcript. Spliceosomal components (red balls) bind the 5′ SS and 3′ SS of the pre-mRNA co-transcriptionally.

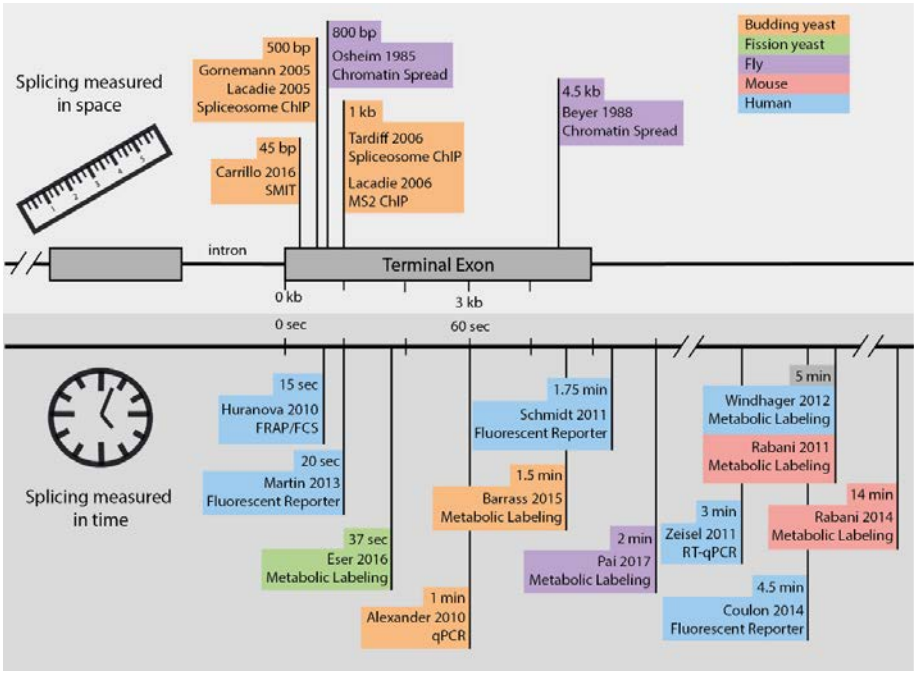

Figure 1.2 Summary of distance- and time-based splicing measurements *in vivo*
Above the gene diagram are flags marking where along the exon each distance-based study found ~50% of splicing had occurred (or other similar measure if unavailable). Below the diagram are time-based measurements. Flags are colored by organism indicated in legend on top right.

ribozyme sequence to measure splicing in a more direct way, in which RT-qPCR was used to determine the relative fraction of RNAs that had been self-cleaved by the ribozyme as opposed to those species that co-transcriptionally spliced the ribozyme before cleavage. These assays identified that splicing occurs when Pol II has traveled 500 and 1500 bp downstream of the 3′ SS. Taking an average measured transcription elongation rate of 1500 bp/min in budding yeast into account (Mason and Struhl, 2005), these authors estimated that splicing occurs within 20-60 seconds *in vivo*. Because the majority of genes in budding yeast have terminal exons that are shorter than 1500 bp (median 434 bp), Rosbash and colleagues concluded that most splicing must be post-transcriptional in yeast (Tardiff et al., 2006). A study from our lab directly tested co-transcriptional splicing frequency in yeast and refuted this conclusion by showing that the majority of introns are removed in nascent RNA (Carrillo Oesterreich et al., 2010).

1.1.2. **Time-based measures of splicing rate**

The development of alternative approaches that measure splicing in terms of time compensates for the limitations of ChIP-based approaches, which are unavoidably indirect. These methods include live cell imaging to quantify intron lifetimes and snRNP dynamics and metabolic labeling to quantify intron lifetimes and the emergence of spliced RNAs (Figure 1.3). Several studies have set out to determine when splicing occurs by quantitatively measuring the amount of spliced and unspliced RNA transcripts using RT-qPCR or RNA-seq. Some time-based studies have implemented a system of inducible transcription that allows tracking of pre-mRNA intermediates from a given gene over a time course. To date, the highest time resolution achieved with RT-qPCR analysis is 30 seconds, during which an integrated, tetracycline-inducible reporter gene in budding yeast

yielded observed spliced transcripts 60 seconds after pre-mRNA transcription was induced (Alexander et al., 2010). In mammalian cells, where genes and introns are very large, longer time points are needed to sample the transcription and splicing of induced transcripts. The Padgett group set out to measure transcription and splicing kinetics for very long genes and introns to determine whether special regulation accounts for the expression of such genes (Singh and Padgett, 2009). To achieve this, they developed a system to observe splicing of endogenous human genes *in vivo*. Using treatment and then wash out of a reversible inhibitor of Pol II elongation, 5,6-dichloro-1-β-D-ribofuranosylbenzimidazole (DRB), the authors collected time-point samples of newly synthesized transcripts that were then quantified using RT-qPCR. They found that despite the incredible length of introns assayed (>100 kb) all introns were spliced in the first or second time point (5 and 10 min), which thereby delimits a maximum time window for splicing in human Tet21 cells. A related study conducted in human mammary epithelial MCF10A cells stimulated with EGF measured pre-mRNA half-life of 2-3 minutes by RT-qPCR (Zeisel et al., 2011), placing splicing within the same time window in another cell type. Interestingly, the Padgett study also quantitated Pol II elongation rates over these long introns at 3.8 kb/min, a relatively high value that agrees with the report of faster elongation rates along introns (Jonkers et al., 2014; Veloso et al., 2014). The Padgett study was therefore critical in confirming that splicing is relatively quick and often co-transcriptional, even in endogenous human genes with extremely long introns.

Studies utilizing live cell imaging for the analysis of splicing kinetics rely primarily on reporter genes in human cells. Results from these experiments vary widely, even when using a similar reporter element. For example, two studies used stably integrated β-globin

reporter genes with MS2 or PP7 stem loops inserted into intronic or exonic sequences to track pre-mRNA (Coulon et al., 2014; Martin et al., 2013) (Figure 1.3). The first study reported splicing in HEK293 cells within 20 and 30 seconds for the first and second introns, respectively. The second study observed splicing 267 seconds after transcription of the 3′ SS of the terminal intron in U2OS cells. Analyzing live cell imaging data is challenging, however, and varying analysis methods along with different cell types likely contribute to the discrepancy between the studies.

The attraction of live-cell imaging of fluorescent reporters, such as those described in the above experiments, is that these measurements in real time provide higher time resolution than metabolic labeling or time points taken for RT-qPCR after gene induction. If splicing is or can be a very fast reaction, 5- and 10-minute time points are not sufficient to deepen our understanding of splicing kinetics. Additionally, single cell information can provide insight into cell-to-cell variation and compensate for the lack of synchronization in a large cell population. However, chromosomally integrated reporter genes will never report the diversity of kinetics available in endogenous genomes. Splicing rates likely differ among genes as well as introns and exons within each gene; reporter genes miss out on this source of information. To understand splicing kinetics in complex cellular systems, it is prudent to look at a sampling of endogenous genes. One study took a global approach by measuring the residency times of fluorescently-labeled spliceosomal snRNPs on pre-mRNA transcripts using FRAP and FCS in HeLa cells (Huranová et al., 2010). While U1 and U4 snRNPs that transiently associate with the assembling spliceosome display shorter residency times, subunits of the active spliceosome – U2 and U5 snRNPs – reside on pre-mRNA for 15-30 seconds. This indicates that the average splicing duration at steady state

in HeLa cells lies within a 30 second window. Finally, it is important to note that measurements of intron half-lives as well as snRNP dynamics will encompass remaining transcription of intron and exon elements, spliceosome assembly, splicing, spliceosome disassembly and/or intron release, intron debranching and degradation. Thus, the overall range of times observed – 0.5-3 minutes – could reflect differences in the rates of any one or more of these processes.

Metabolic labeling also affords time-based measurements by feeding cells modified nucleic acids (4-thiouridine or 4sU) which are incorporated into newly synthesized RNA. Thiol-specific purification enriches for new RNA which can be assessed by sequencing or qPCR. At least six studies detect substantial amounts of pre-mRNA splicing at the very early labeling time points: 1.5 minutes in *S. cerevisiae* (Barrass et al., 2015), 2 minutes in *S. pombe* with a median intron splicing time of 37 seconds (Eser et al., 2016), 2 minutes in *Drosophila melanogaster* S2 cells (Pai et al., 2017), 5 minutes in human B cells (Windhager et al., 2012), and 10 minutes in LPS-stimulated mouse dendritic cells with a median intron splicing time of 14 minutes (Rabani et al., 2014). It is also important to note that these time windows reflect the convolution of transcription, processing and degradation, as is the case for live-cell imaging approaches addressed above. Nevertheless, the global nature of metabolic labeling has enabled the discovery of gene architecture and sequence motif correlations linked to synthesis, degradation and splicing kinetics (Barrass et al., 2015; Eser et al., 2016; Rabani et al., 2011; Windhager et al., 2012), which drive hypotheses for future investigation.

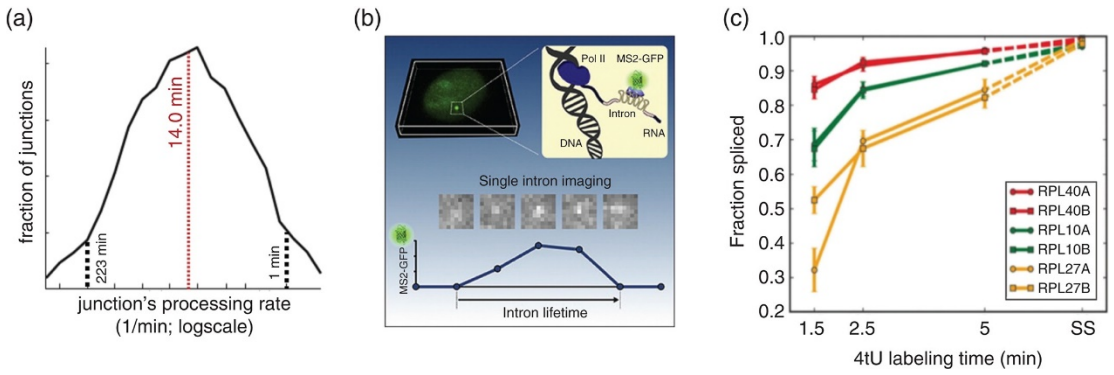

Figure 1.3 Time-based measurements vary widely in methodology and results
A. The median number of exon-intron junctions are processed in 14 min according to RNA-seq reads of metabolically labeled mouse RNA (Rabani et al., 2014). B. Use of fluorescent reporter genes permits imaging of introns and quantitation of their half-lives (Martin et al., 2013). C. Fraction spliced values from a metabolic labeling experiment are plotted at different time points for three pairs of gene paralogs in yeast. Paralogs have identical exonic sequences and different intronic sequences. Splicing values appear highly similar between paralogs, yet different between genes (Barrass et al., 2015).

1.1.3. **Single Molecule Intron Tracking and long read sequencing of nascent RNA**

The final methods to measure splicing rates discussed in this section are our own, Single Molecule Intron Tracking (SMIT) and long read sequencing of nascent RNA (Figure 1.4A&B). Both strategies use 3′ end linker ligation and single molecule RNA-seq to identify the position of Pol II along the length of genes at the moment introns are removed (Carrillo Oesterreich et al., 2016). SMIT uses paired-end sequencing to measure the fraction of transcripts spliced at each Pol II position with approximately 300 observations per nucleotide. Direct, long read sequencing of nascent transcripts with Pol II positions marked by their 3′ ends provide images reminiscent of chromatin spreads. SMIT analysis of 87 endogenous genes and long read sequencing of nascent RNA from *S. cerevisiae* and *S. pombe* reveal exon-exon ligation detectable when Pol II is 26 and 36 nt downstream of the 3′ SS, respectively. These findings indicate that the active spliceosome is physically very close to Pol II and revitalize models that consider direct interactions

between Pol II and spliceosome components (Saldi et al., 2016). The mammalian U2AF and FUS proteins, which interact directly with Pol II and components of the splicing machinery, are examples of factors that bridge the two machineries (Újvári and Luse, 2004; Yu and Reed, 2015). Given the observed association of particular Pol II CTD phosphorylation states with spliceosome assembly and splicing (Harlen et al., 2016; Nojima et al., 2015), it is intriguing to consider the regulation of such interactions during the elongation process. SMIT precisely measures the occurrence of the second step of splicing – exon-exon ligation – in distance. Evidence that Pol II did not pause at or near 3′ SSs of the 87 genes analyzed permits an estimation of splicing rate in time; SMIT curves reached saturation at a median Pol II position of 129 nt downstream of 3′ SSs, indicating that splicing completes within ~5 seconds if transcription elongation proceeds at 1.5 kb/min (Carrillo Oesterreich et al., 2016; Mason and Struhl, 2005). Pausing has been reported in yeast and human cells under various conditions and may impact the time estimate derived from Pol II position (Alexander et al., 2010; Chathoth et al., 2014; Harlen et al., 2016; Kwak et al., 2013; Mayer et al., 2015; Nojima et al., 2015). Changes in splicing kinetics measured with distance-based methods would be expected to reflect such changes in transcription elongation. Indeed, a mutation in Pol II that causes faster transcription elongation, leads to lengthening of the distance measurement by SMIT and an increase in levels of unspliced transcripts (Braberg et al., 2013; Carrillo Oesterreich et al., 2016). Thus, faster Pol II elongation rates result in greater physical separation between the spliceosome and Pol II, while slower elongation would be expected to bring the two machines closer together, consistent with the observation that slow Pol II mutants correlate with increased levels of spliced transcripts (Braberg et al., 2013). Together with recent findings that

mammalian transcription and splicing rates in vivo are optimized (Fong et al., 2014), the data suggest that the two rates have co-evolved. The matching of the reaction rates of splicing and transcription elongation rates observed in budding and fission yeast suggest perfect timing in the coordination of these two macromolecular machines.

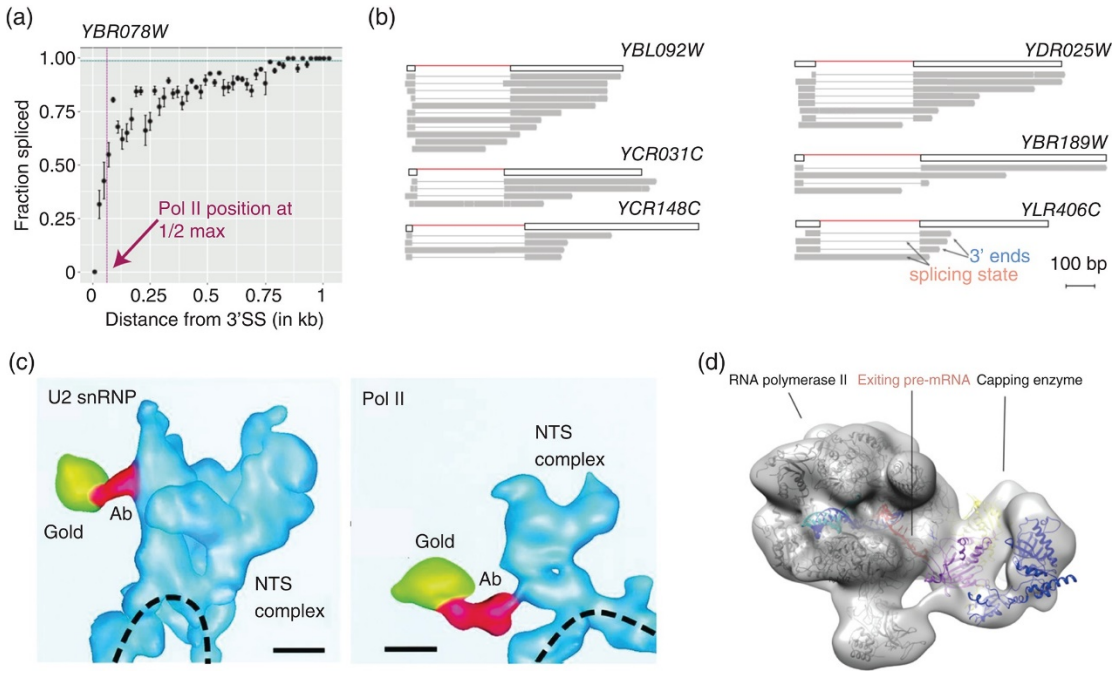

Figure 1.4 Recent distance-based techniques predict that the spliceosome is adjacent to polymerase during splicing

A. Representative SMIT trace from budding yeast shows splicing reaches half-maximum levels at approximately 62 bp past the 3′ SS (Carrillo Oesterreich et al., 2016). B. Long-read sequencing of budding yeast genes enables splicing analysis of single molecules (Carrillo Oesterreich et al., 2016). C. 3D reconstruction of nascent transcription and splicing complex (NTS) bound to Balbiani ring 3 locus on chromatin (dashed line) from *C. tentans* with antibody gold particles against U2 snRNP (left) and Pol II CTD (right) (Wetterberg et al., 2001). D. Crystal structures are docked into electron microscopy density show that capping enzymes bind the RNA exit tunnel of RNA Polymerase II to modify the 5′ end of RNA immediately (Martinez-Rucobo et al., 2015).

1.2 Crosstalk between splicing and gene expression elements

The entire transcription elongation machinery and the nascent RNA itself interact with proteins, forming RNP complexes. Many of these proteins belong to complexes that are involved in mRNA 5′ end capping, splicing, 3′ end processing, editing, folding, nuclear export and decay, and bind to specific transcript regions, such as untranslated regions, introns and exons (Baejen et al., 2014; Müller-McNicoll and Neugebauer, 2013; Singh et al., 2015). The extensive crosstalk between splicing, transcription and other nuclear machineries can be appreciated by considering the multitude of reported genetic and physical interactions between them, summarized in Figure 1.5. In this section, I will discuss in depth the known interactions between splicing and nuclear gene expression and the reciprocal nature of these interactions.

The transcription start site (TSS) marks the 5′ end of the first exon, and the poly(A) site (PAS) marks the 3′ end of the last exon. Exon–intron organization provides important additional landmarks for the alignment of signals and activities, such as Pol II density, chromatin modifications, and RNA sequence and structure elements. Mammalian genes are longer than yeast genes, primarily because they contain more and longer introns (Lander et al., 2001). Nevertheless, several aspects of gene architecture are conserved from yeast to humans. For example, the last exon is almost always the longest. Moreover, intron structure is similar across evolution: the GU and AG dinucleotides, which are contained in short and conserved sequences known as splice sites, define the 5′ and 3′ intron boundaries, respectively. A third sequence, the branchpoint sequence (BPS), is 18–40 nucleotides upstream of the 3′ splice site (3′SS) (Mercer et al., 2015).

The positioning of nucleosomes relative to TSSs, transcription termination sites (TTSs), exons and introns helps to define gene architecture (Beckmann and Trifonov, 1991; Huff et al., 2016; Schwartz et al., 2009; Tilgner et al., 2009). Conversely, nucleosome phasing seems to be facilitated by the inherent exon–intron primary structure (such as elevated GC content in exons), by the sequence elements that are required for pre-mRNA splicing (Beckmann and Trifonov, 1991; Schwartz et al., 2009; Tilgner et al., 2009) and by the local activity of chromatin remodelers (Schwartz et al., 2009; Venkatesh and Workman, 2015). The median length of internal exons in the human genome is 137 bp, which is very close to the 147 bp that are wrapped around a nucleosome (Lander et al., 2001). Nucleosomes can interfere with transcription progression (Churchman and Weissman, 2011; Milligan et al., 2016; Weber et al., 2014). Consistent with nucleosome phasing over exons, slower transcription elongation has been measured over exonic sequences (Kwak et al., 2013; Mayer et al., 2015). The histone tails of nucleosomes that are positioned over exons can be enriched for PTMs that may facilitate exon definition by affecting the recruitment of splicing factors, as well as the transcription process itself (Braunschweig et al., 2013; Hérissant et al., 2014; Kfir et al., 2015). Slow passing of the transcription machinery through nucleosomes may facilitate the relocation of splicing factors and regulators from the chromatin template to Pol II or to the nascent RNA.

The transcription rate can influence splice site identification by the spliceosome. Current models suggest that different local rates of transcription elongation can influence the time frame between the synthesis of sequential splice sites, thereby possibly modulating RNA folding or the interactions with RNA-binding proteins (Naftelberg et al., 2015). The synthesis of RNA by Pol II occurs with an average elongation rate of 1–4 kb per minute

(Jonkers and Lis, 2015; Kwak and Lis, 2013; Veloso et al., 2014). However, Pol II can transiently pause, stall or terminate prematurely (Jonkers and Lis, 2015; Kwak and Lis, 2013). Pol II elongation rate is influenced by a multitude of factors, such as the underlying DNA sequence, nucleosome position, and histone modifications, which affect local chromatin structure, the activity of elongation factors, and the folding and processing of the nascent RNA (Kwak and Lis, 2013; Nedelcheva-Veleva et al., 2013).

Recent studies have implicated the splicing process in transcriptional pausing. For example, pausing at terminal exons was detected in efficiently spliced genes in yeast (Carrillo Oesterreich et al., 2010). Upon splicing inhibition, either by introducing a temperature-sensitive mutant allele of the RNA helicase Prp5, or by introducing mutations in the BPS or the U2 snRNA, the Pol II ChIP signal increases on introns, suggesting the activation of a transcription elongation checkpoint to allow spliceosome assembly (Chathoth et al., 2014). The extent to which splicing-related pausing occurs and a mechanistic understanding of this process are still elusive.

Changes in transcription elongation rates influence nascent RNA folding and so may affect splice site selection (Lai et al., 2013; Warf and Berglund, 2010). The propensity for RNA folding directly depends on sequence-specific folding rates, transcription elongation rates and the rate of proteins binding to the RNA (Liu et al., 2016). RNA secondary structures can conceal or expose the 5′SSs, BPSs and 3′SSs, which are consequently ignored or readily recognized by the splicing machinery (Buratti and Baralle, 2004). By concealing or exposing alternative cis-acting elements, secondary structures may have a role in alternative splicing (Meyer et al., 2011). The splicing machinery may recognize splicing targets on nascent RNA before the RNA has time to fold into a secondary structure

(Warf and Berglund, 2010). Similarly, RNA-binding proteins may influence the transient folding of nascent RNA and, therefore, may modulate the timing of splice site exposure to the splicing machinery (Buratti and Baralle, 2004). For example, hairpins with a small loop readily fold after transcription, thereby concealing the splice site that is contained in their stem. By contrast, the folding of hairpins with bigger loops takes longer, allowing longer splice site exposure to the splicing machinery and/or to regulatory RNA-binding proteins (Eperon et al., 1988).

Changes in post-translational modifications (PTMs) of the Pol II carboxy-terminal domain (CTD) mirror and influence the different phases of transcription and nascent RNA processing, owing to the interaction of the CTD with factors that regulate transcription, mRNA processing, and downstream steps such as mRNA export (Custódio and Carmo-Fonseca, 2016). CTD-modifying enzymes often have additional cellular targets, thereby integrating the CTD into a greater network of gene expression (Zaborowska et al., 2016). The CTD consists of repeats of almost the same seven amino acids Tyr1-Ser2-Pro3- Thr4-Ser5-Pro6-Ser7 (26 repeats in yeast and 52 in humans (Corden et al., 1985)), which are mainly modified by phosphorylation of Ser2, Ser5, Ser7, Thr4 and Tyr1 (Kim et al., 2010). Pol II is differentially modified at the start and the end of transcription units. PTM transitions have recently been mapped to transcription pause positions along yeast gene bodies and, in particular, to 3′SSs, consistent with changes in Pol II elongation rate around intron–exon boundaries (Harlen et al., 2016; Milligan et al., 2016). The phosphorylation levels of Ser5 (Ser5P) are highest at the beginning of transcription units, a link between this PTM and splicing has been found in both yeast and humans (Harlen et al., 2016; Nojima et al., 2016). Pronounced peaks of Ser5P and Pol II levels are observed at the 5′

SSs of alternatively included exons compared with excluded exons (Mayer et al., 2015; Nojima et al., 2016). Nevertheless, the relationship between specific Pol II CTD PTM profiles and nascent RNA processing events is far from being understood.

The complexity of gene architecture varies between phyla (Deutsch and Long, 1999), requiring different mechanisms to identify splice sites. Metazoan splice sites are short and poorly conserved, in contrast to budding yeast splice sites (Will and Lührmann, 2011). In vertebrates, intron length varies from a few hundred nucleotides to several thousand nucleotides, and the median length of internal exons is 137 nucleotides (Hawkin, 1988). Surrounding the internal exon, the 3′ SS of the upstream intron and the 5′ SS of the downstream intron pair across the exon, thereby committing the upstream intron to splicing through an 'exon definition' mechanism (Berget, 1995). By contrast, transcripts in lower eukaryotes usually contain introns that are shorter than 250 nucleotides (Hawkin, 1988). In this case, a 5′ SS pairs with the downstream 3′ SS of the same intron, and splicing is triggered through an 'intron definition' mechanism (Talerico and Berget, 1994). Splicing of the first intron depends on first exon definition. First exon boundaries consist of the 7-methylguanosine (m7G) cap structure at the 5′ end of the transcript and the 5′ SS of the first intron. The capping enzyme adds the m7G cap to the 5′ end of all Pol II-transcribed RNAs when the nascent RNA is less than 20 nucleotides in length (Izaurralde et al., 1994; Martinez-Rucobo et al., 2015). The nuclear cap-binding complex (CBC) serves as a platform for interacting with factors that are involved in RNA processing (Gonatopoulos-Pournatzis and Cowling, 2014), enhancing splicing of the first intron *in vitro* (Berget, 1995; Izaurralde et al., 1994; Konarska et al., 1984), and directly interacting with tri-snRNP protein components *in vivo* (Pabis et al., 2013).

Splicing of the last intron depends on terminal exon definition (Berget, 1995). The 3′ SS of the last intron and the PAS set the terminal exon boundaries (Proudfoot, 2016). The PAS, the nearby AU-rich sequences and other cis-elements on the nascent RNA are bound by the cleavage and polyadenylation complex (CPA). The PAS is required for termination, and its elimination leads to transcription readthrough (Connelly and Manley, 1988). PAS elimination also results in the specific inhibition of last intron splicing, indicating that 3′ end processing contributes to terminal exon definition (Cooke et al., 1999). The U2 snRNP, U2AF65 and cleavage and polyadenylation specificity factor (CPSF; a component of the CPA) functionally and physically interact; this supports a model in which the PAS triggers the removal of the last intron by facilitating spliceosome assembly (Kyburz et al., 2006; Millevoi et al., 2002; Vagner et al., 2000). Splicing and the regulation of 3′ end processing are reciprocal, as the inactivation of the terminal 3′ SS inhibits 3′ end processing and transcription termination (Davidson and West, 2013; Dye and Proudfoot, 1999). The splicing and 3′ end processing machineries seem to serve as recruitment platforms, as the physical presence of the two machineries is sufficient for coupling between splicing and 3′ end processing, and the catalytic activity of neither is required for the regulation of the complementary process (Davidson and West, 2013; Martins et al., 2010; Rigo and Martinson, 2008). Indeed, artificially induced cleavage of the nascent RNA impairs splicing and 3′ end processing *in vitro* (Rigo and Martinson, 2008). *In vivo*, the Pol II CTD stimulates coupling between splicing and 3′ end processing (Bird et al., 2004). Taken together, these observations suggest that coupling between splicing and 3′ end processing can determine whether splicing occurs before transcription termination (Kaida, 2016).

The molecular mechanisms of 3′ end processing involve components of the splicing machinery. The mammalian U1 snRNP component U1 70k directly interacts with poly(A) polymerase, a component of the CPA, and inhibits polyadenylation (Gunderson et al., 1998). *In vivo*, functional inhibition of the U1 snRNP inhibits splicing and causes premature 3′ end processing (Kaida et al., 2010). Interestingly, the U1 snRNP is much more abundant than all other spliceosomal snRNPs and this may reflect its role in protecting nascent RNA from premature 3′ end processing (Baserga and Steitz, 1993). In such a model, the U1 snRNP binds to nascent RNA at frequent cryptic 5′ SSs and suppresses the activity of adjacent cryptic PASs (Kaida et al., 2010). Overall, the PAS and CPA have a major role in defining the last exon and thus the 3′ SS for last intron removal in mammalian systems. In addition, splicing aids 3′ end processing by preventing the recruitment of the CPA to cryptic PASs. It is unclear how coupling of splicing and 3′ end formation may function in a system of intron definition, where the PAS is not required for splicing of the terminal intron.

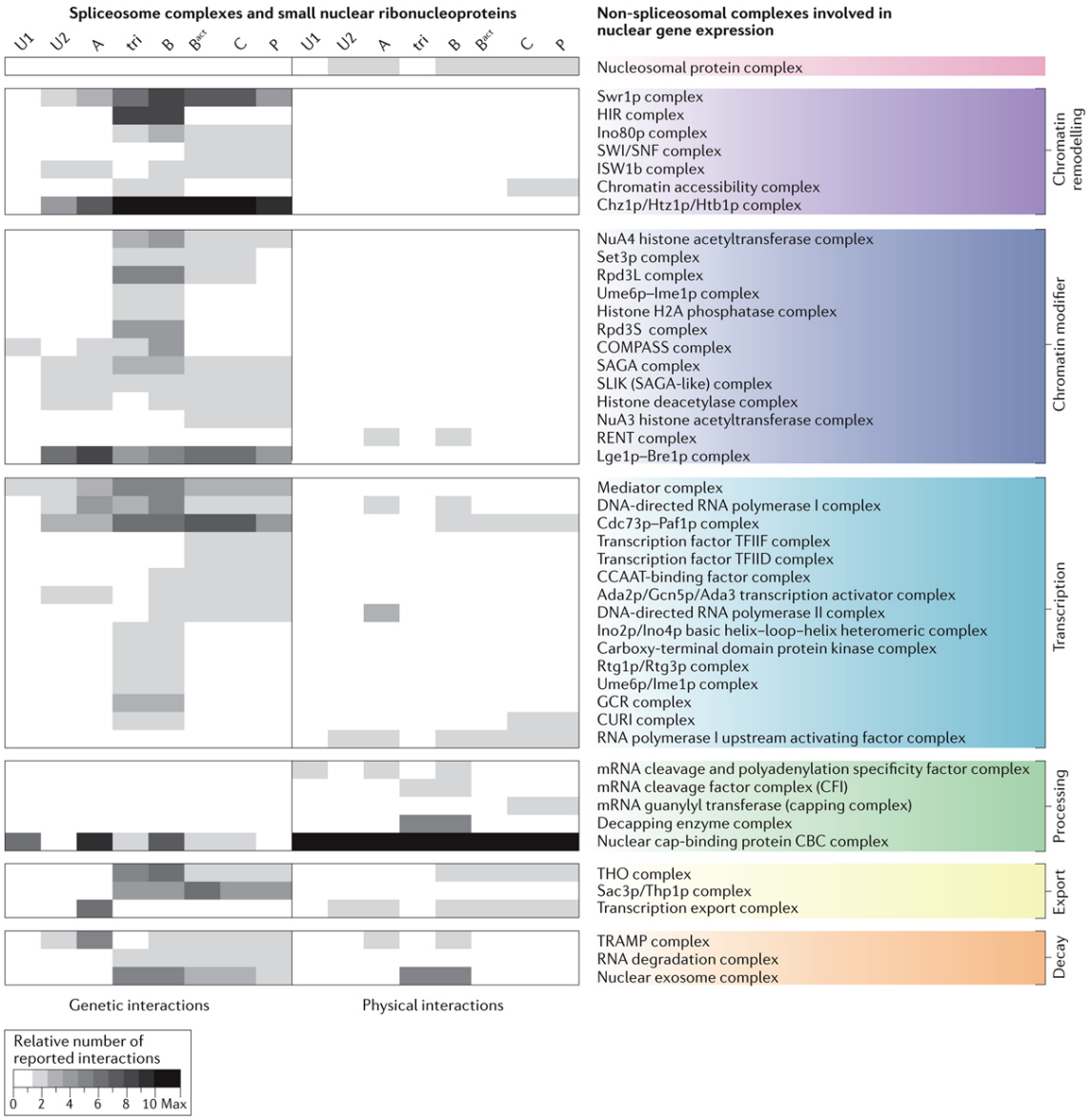

Figure 1.5 Crosstalk of the assembling spliceosome with nuclear gene expression machineries

Genetic and physical interactions that involve core splicing factors of *S. cerevisiae* were obtained from the Biological Repository for Interaction Datasets (BioGRID). The grey scale reflects the number of reported interactions between spliceosomal and non-spliceosomal complex subunits. The number of reported interactions is adjusted to the number of reported non-spliceosomal complex subunits. A minimum of two reports for the same interaction was required. Figure reproduced from Herzel, Ottoz, Alpert, and Neugebauer, 2017.

1.3 Sequencing approaches for the study of co-transcriptional splicing

The development of next-generation sequencing has revolutionized life science and medical fields, with the cost of sequencing entire genomes exponentially declining and requiring only a few days of work. With this unprecedented leap in sequencing technology, science has begun addressing questions that were previously out of reach. In this section I will discuss the three sequencing platforms relevant to this study and how we take advantage of each to address different biological questions.

An outstanding majority of RNA sequencing experiments are done with Illumina sequencing, a sequencing-by-synthesis (SBS) method using bridge-amplification to generate clusters of identical DNA copies on a chip. DNA clusters are required for signal amplification and detection as the sample undergoes iterative cycles of fluorescent base incorporation, washing, imaging, and cleavage. Improper base incorporation dilutes the fluorescent signal until it is completely overtaken by noise, thus read lengths are limited to 250 bp maximum. Classic RNA-seq experiments fragment the transcripts and randomly sequence short reads from this pool to quantify changes in gene expression. SMIT utilizes the same technology to target more specific information from each molecule. Both ends of the DNA amplicon can be sequenced (paired-end sequencing) and the internal sequence inferred from mapping to a genome. SMIT utilizes a primer just upstream of the intron to obtain the splicing status of the read with sequencing of the first end, and a primer complementary to the 3′ end adapter for sequencing of the paired end. The Illumina HiSeq 2500 instrument used in this thesis yields ~250 million reads per lane, and in line with the incredibly fast pace of technology development in the sequencing sphere, new NovaSeq instruments producing upwards of 1 billion reads per lane have already come to market

since these experiments were completed. High read counts lend statistical strength to fraction spliced measurements in SMIT and are necessary to cover every nucleotide position of the exon. The single-nucleotide resolution of SMIT makes it one of the most precise splicing measurements available.

Long read sequencing offers many benefits over Illumina short-read, particularly for nascent RNA with the major drawback being reduced read count. Capturing the entire molecule with long reads describes several stages of the life-cycle of that RNA, including which TSS was used, whether splicing has occurred, and how "old" the molecule is via Pol II position. Previously, our lab has used the Single Molecule Real Time (SMRT) sequencing by Pacific Biosciences (PacBio) to investigate the order of intron removal in *S. pombe* (Herzel et al., 2018). Double-stranded libraries are circularized and serve as the template for addition of fluorescently-labeled nucleotides by a strand-displacing polymerase which processes around the DNA molecule multiple times. These repetitive sequencing events are aligned to form a consensus, increasing the accuracy of PacBio long reads to near 99% to the detriment of read length.

Around the beginning of my dissertation research, a new long read sequencing platform called Oxford Nanopore Technologies was commercialized, launching the era of third-generation sequencing. Nanopore sequencing is distinct from Illumina and PacBio with the direct detection of oligonucleotides as opposed to a sequencing-by-synthesis approach that is so central to the other platforms. Instead, oligonucleotides are threaded through a nanometer-wide pore embedded in synthetic polymer membranes on a flowcell. As an electric current is applied to the membrane, engineered motor proteins ratchet the oligonucleotide through the pore at a consistent rate and the nucleotides block the flow of

electrons through the pore. Current disruptions characteristic to each nucleotide are translated into base sequence with a proprietary Recurrent Neural Network (RNN). Accuracy of single nanopore reads is lower than PacBio (~90-95%) for several reasons. Current fluctuations are incredibly precise, detecting base modifications that slightly differ from that of the original nucleotide, and we can only adapt the model for modifications we are aware of. Detecting base modifications is a major advantage of Nanopore, but deviations from the nucleotide signal that we cannot account for likely arise from unknown base modifications and continue to impede accuracy. Remarkably, improvements to the basecalling RNN have and continue to improve read accuracy, and data from old experiments can be basecalled again with newer algorithms to benefit from the highest accuracy possible (Wick et al., 2019). Nanopore sequencing reads are longer than other platforms, reaching upwards of 2 Mbp, and higher read numbers are achieved per cost of flowcell compared to PacBio. The extreme cost efficiency and improved readcount make Nanopore ideally suited to my experiments, and the results we uncover could not have been attained with short read sequencing.

2. Identifying modulators of co-transcriptional splicing kinetics

2.1 Author contributions

This work was done in collaboration with our lab manager, Korinna Straube, and a previous graduate student, Fernando Carrillo Oesterreich. Fernando developed the SMIT technique (in collaboration with Lydia Herzel) and trained the machine learning algorithm. Korinna assisted in cloning the strains and performed the nascent RNA purification and SMIT sequencing libraries. Together, Korinna and I made decisions about which steps of the protocol to optimize and strategized about which genes to target during optimization to represent future genome-wide samples. I was responsible for optimizing the analysis pipeline and performing all data processing and analysis. I prepared all samples for the long read sequencing of Nab2-AA and processed and analyzed the data.

2.2 Introduction

Proper gene expression is the result of many diverse cellular processes functioning in a concerted manner. RNA must be synthesized, spliced, cleaved, and exported to be competent for translation into protein and precise coordination of these steps is critical for responding to environmental stimuli and preserving efficiency and resources.

Splicing is the removal of intronic sequences from pre-mRNA and is an essential, highly regulated step in gene expression. Despite requiring the de novo assembly of a megadalton spliceosome on every intron, splicing is a very fast process. In fact, the majority of splicing is completed well before transcription terminates, a phenomenon termed co-transcriptional splicing. The simultaneous nature of splicing and transcription

invites a plethora of opportunities to coordinate with one another. Indeed, splicing is coupled to transcription as evident by the enhanced transcriptional output of intron-containing transgenes in mice (Brinster et al., 1988). Conversely, promoter sequences can influence splicing efficiency (Cramer et al., 1997; Moldón et al., 2008; Nissen, 2017). At the other end of the transcript, cleavage and polyadenylation factors define the terminal exon in mammalian cells (Cooke et al., 1999; Fong and Bentley, 2001; Niwa and Berget, 1991; Rigo and Martinson, 2008), and without this definition, splicing of terminal introns is greatly impaired. The relationship between mammalian splicing and 3′ end formation is mutual, as mutations of the 3′ SS similarly hinder RNA cleavage (Cooke et al., 1999; Davidson and West, 2013; Martins et al., 2010). Recent work from our lab discovered that units of Pol II associated with unspliced transcripts in *S. pombe* have a striking tendency to continue transcribing past the polyA site (Herzel et al., 2018).

We set out to identify which factors in budding yeast form the channel of communication between splicing and the other processes involved in RNA synthesis and maturation. To do so, we took advantage of the high-precision splicing kinetic measurements from our previously published Single Molecule Intron Tracking (SMIT) approach (Carrillo Oesterreich et al., 2016). These data were intriguing for the high amount of gene-specific variability present, which we could use to our advantage for identifying factors that influence splicing. A machine learning model was trained to exploit this variation for the prediction of splicing kinetic parameters using quantifications of RNA and DNA binding proteins, Pol II density, and histone modifications, among others (14 genetic and 398 epigenetic features derived from genome annotations and genome-wide experiments respectively). The model successfully predicts splicing kinetics and identifies

21 candidates for splicing regulation involved in diverse cellular processes including elongation, 3′ end processing, and export. We perturbed expression of seven of these candidates and performed SMIT and long read sequencing in the context of factor depletion, identifying new roles for the essential and conserved polyA binding protein, Nab2, in polyA RNA cleavage and termination and its implications for co-transcriptional splicing.

2.3 Results

2.3.1. Predicting modulators of co-transcriptional splicing kinetics

Splicing kinetics in budding yeast are highly variable between genes (Carrillo Oesterreich et al., 2016). Previous work from our lab identified two key parameters to describe splicing kinetics, namely saturation value (the mean fraction spliced of the last 120 bp before the polyA site or before the end of collected data) and ½ max value (Pol II position where half of the saturation value is reached). To obtain mechanistic insights into what influences gene-specific variation, we trained a machine learning model to predict saturation and ½ max using gene-specific quantification of regulators. We then determined the relative importance of each regulator for the model's prediction strength (Table 5.1). Each gene was characterized by 14 genetic and 398 epigenetic features derived from the budding yeast genome and genome-wide experiments (e.g. ChIP and CLIP) respectively. Hierarchical clustering of the features produced 100 feature groups that share similar function and position along the gene. For example, U1, U2, and U5 snRNPs are all prominently detected at the 3′ SS by ChIP (Tardiff and Rosbash, 2006) and compose one feature group.

A Lasso regression model (Tibshirani, 1994) was trained on 80% of the data and was able to predict the remaining 20% of saturation values, validating the model's ability to identify important regulators of splicing (Figure 2.1B). The model selected 21 non-genetic factors (along with 8 additional genetic features; Table 5.1) that contribute to prediction performance, 13 factors positively correlated with splicing (Figure 2.1C above gene diagram) and 8 negatively correlated (Figure 2.1C below gene diagram). A second model was unsuccessful at predicting ½ max values because of the reduced variation in this parameter (Appendix Figure 5.1).

Several identified features agree with previous reports including the correlation of splicing with the U1, U2, and U5 feature group at the 3′ SS (Görnemann et al., 2005; Lacadie and Rosbash, 2005; Lacadie et al., 2006) (Figure 2.1C). Other exciting results included significant correlation of 3′ end processing factors at the 3′ SS (Figure 2.1C). Cleavage and polyadenylation factors are known to play a role in exon-definition of terminal exons in metazoans (Fong and Bentley, 2001; Li et al., 2001; Niwa and Berget, 1991; Rigo and Martinson, 2008), but a direct role for these proteins in the process of splicing upstream of the polyA site has not been found. Pcf11 specifically is one of few cleavage factors with a CTD-interacting domain and is bound to Pol II along the entire gene-body (Baejen et al., 2017; Licatalosi et al., 2002). Thus, Pcf11 is located near the active spliceosome and raises the potential for direct interaction which will be discussed in more depth in later chapters. An alternative histone, H2A.Z, is annotated to promote open chromatin and transcription near the promoters of certain genes and even promote splicing of weak splice sites (Neves et al., 2017; Nissen et al., 2017). Yet, the model utilizes H2A.Z as a very strong negative predictor of splicing ($\beta = -0.68$) (Figure 2.1C).

Intriguingly, the presence of a conserved polyA binding protein, Nab2, at the 5′ and 3′ splice sites was the most positively correlated feature in our model ($\beta = 0.26$) (Figure 2.1C). At the time we identified this putative regulator, no connection to splicing had been observed, however a study was soon published that found Nab2 truncations resulted in splicing defects in yeast and interactions between the full length protein and components of the spliceosome in both yeast and humans (Soucek et al., 2016). This study further confirmed the validity of our model in identifying putative splicing regulators, and encouraged us to investigate the remaining features.

2.3.2. **Splicing is resilient to perturbations of predicted splicing modulators**

SMIT measurements in the context of factor perturbation would reveal whether the targets affect co-transcriptional splicing kinetics. To accomplish this, we made genomic deletions for every factor that would retain cell viability (*rtt103Δ*, *gbp2Δ*, *pub1Δ*, *npl3Δ*, *tho2Δ*, and *htz1Δ* (H2A.Z)) and tagged one essential factor Nab2 with a motif for nuclear depletion (Nab2-AA) using the Anchor-Away approach (Haruki et al., 2008; Schmid et al., 2015). These factors represented a diversity of cellular function as well as novelty; Npl3 is a known splicing modulator (Kress et al., 2008), while Rtt103 and others had never been implicated in splicing.

We sourced deletion strains from the Genome Deletion Project (Winzeler et al., 1999). However, this collection has been shown to harbor frequent secondary mutations (Teng et al., 2013). To ensure that our samples did not harbor undetectable compensatory mutations, we amplified the deletion cassette and retransformed it into a background strain. I then assayed splicing changes of select genes with RT-PCR of nascent RNA from the *htz1Δ* mutant (Appendix Figure 5.6). I found splicing was unaffected by *htz1Δ*, but given

the high degree of gene-specific variation we observed in the original splicing kinetic data, I decided to continue assaying splicing perturbations on a larger scale with SMIT.

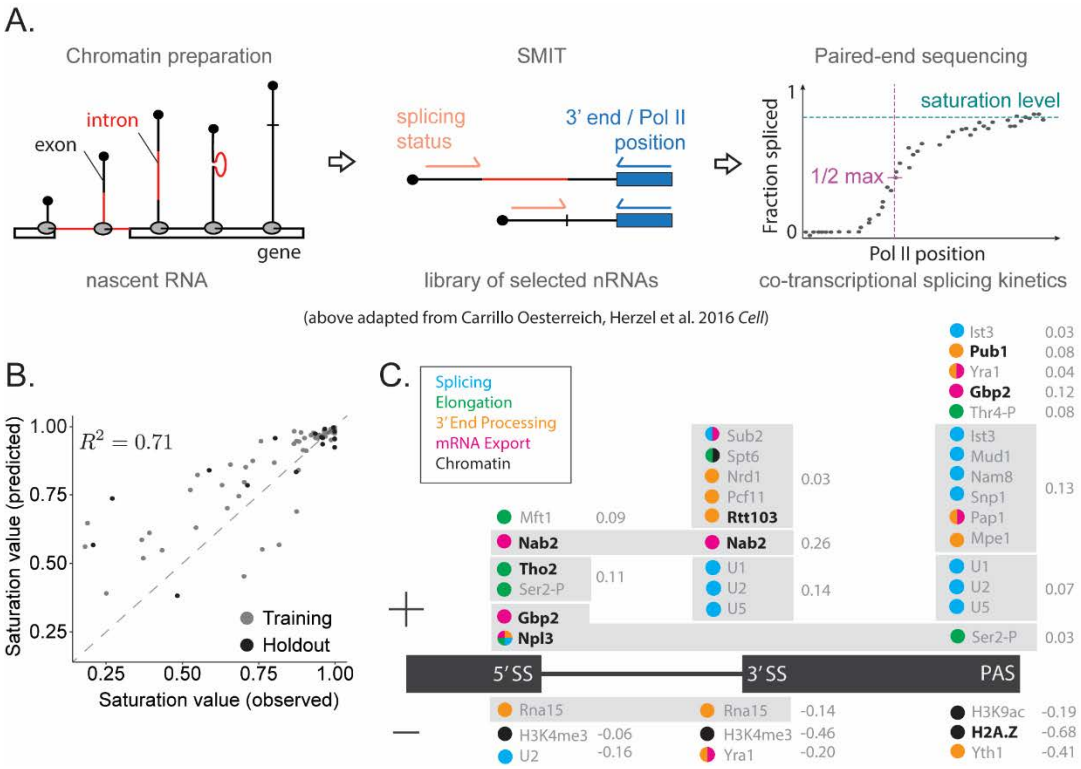

Figure 2.1 Machine learning model identifies putative splicing modulators
A. Schematic of SMIT protocol (adapted from (Carrillo Oesterreich et al., 2016)) describes nascent RNA purification from chromatin (left), targeted amplification of library (center), and analysis of sequencing data (right). SMIT curves were parameterized with saturation level and ½ max (right). B. Observed and predicted saturation values are correlated with the variance explained (R^2) as indicated for training (grey) and holdout (black) data. C. Predictive feature groups (grey boxes) are displayed above (positive correlation) or below (negative correlation) the gene diagram. No box indicates the feature was the sole group member. Factors are positioned along the gene where their presence is correlated with saturation level. Those listed in bold are the factors we target in this study. Learned regression coefficients β are shown to the right of each feature group.

Endogenous heterogeneity among co-transcriptional splicing is what enabled our study; therefore, I expected great heterogeneity in how genes react to different perturbations. It was important to assay as many genes as possible to ensure robust representation of the diverse genes and genomic environments. Before collecting large amounts of SMIT data, considerable time was devoted to optimizing the SMIT protocol to promote reproducibility and reduce length of the protocol (details of optimization can be found in Appendix 5.2). I identified a set of 53 intron-containing genes where I was likely to see effects of the deletions by correlating levels of each factor with splicing saturation. Three intronless genes were included for normalization of amplification bias. Some genes are not suitable for SMIT given low expression, short first exon sequence, or gene duplication events. We sequenced SMIT libraries for all six deletion strains alongside wildtype controls in duplicate as well as 0-, 10-, and 30-minute nuclear depletions of Nab2 and the isogenic control. Surprisingly, we observed that nearly all genes were unaffected by the different deletions we had made (Figure 2.2), suggesting that levels of co-transcriptional splicing are robustly maintained when presented with perturbations to diverse nuclear pathways, with the exception of *UBC4* (discussed below). The lack of effect on splicing was intriguing to us given the observed steady-state splicing changes in *npl3Δ* mutants (Kress et al., 2008). However, measurements of steady-state splicing are confounded by co-transcriptional splicing, post-transcriptional splicing, and RNA stability and are thus may differ from our direct observations of co-transcriptional splicing.

Additionally, there is considerable gene-to-gene diversity in our ability to replicate SMIT results. We encountered several cases where splicing changes were observed in one replicate, but not the other. These results are in agreement with a more in-depth replicate

analysis we pursued, where it appears certain genes have more natural variation than others (Appendix Figure 5.5). It is possible that, with additional replicates of our deletion strains, we could identify more genes with altered splicing kinetics.

The viability of the deletion strains may also contribute to the observed stability of splicing kinetics. If there is little impact of factor deletion on cell viability, the role that factor plays in its respective process (e.g. elongation, export, etc.) cannot be substantial. While *npl3Δ* and *tho2Δ* did grow more slowly, all strains were viable. Accordingly, the only non-viable perturbation we assayed (Nab2-AA) has a more substantial effect on splicing (discussed further in Section 2.3.3). The only exception of note is *UBC4*, an E2 ubiquitin-conjugating enzyme, whose transcript is spliced better than wildtype across all deletion strains, yet unchanged in Nab2-AA (Figure 2.3). Gene architecture (intron and exon length) are completely average for budding yeast genes (~70 bp first exon, ~90 bp intron, ~400 bp terminal exon) and both 5′ and 3′ SSs match the consensus motif, so it is unclear why splicing of *UBC4* is improved in our various deletion conditions. One possibility is that the cells are experiencing a general stress response from the perturbations, and thus increased splicing efficiency may be evidence of upregulation of *UBC4* expression to handle the stress. However, such a response would need to be experimentally determined (O'Duibhir et al., 2014; Teng et al., 2013).

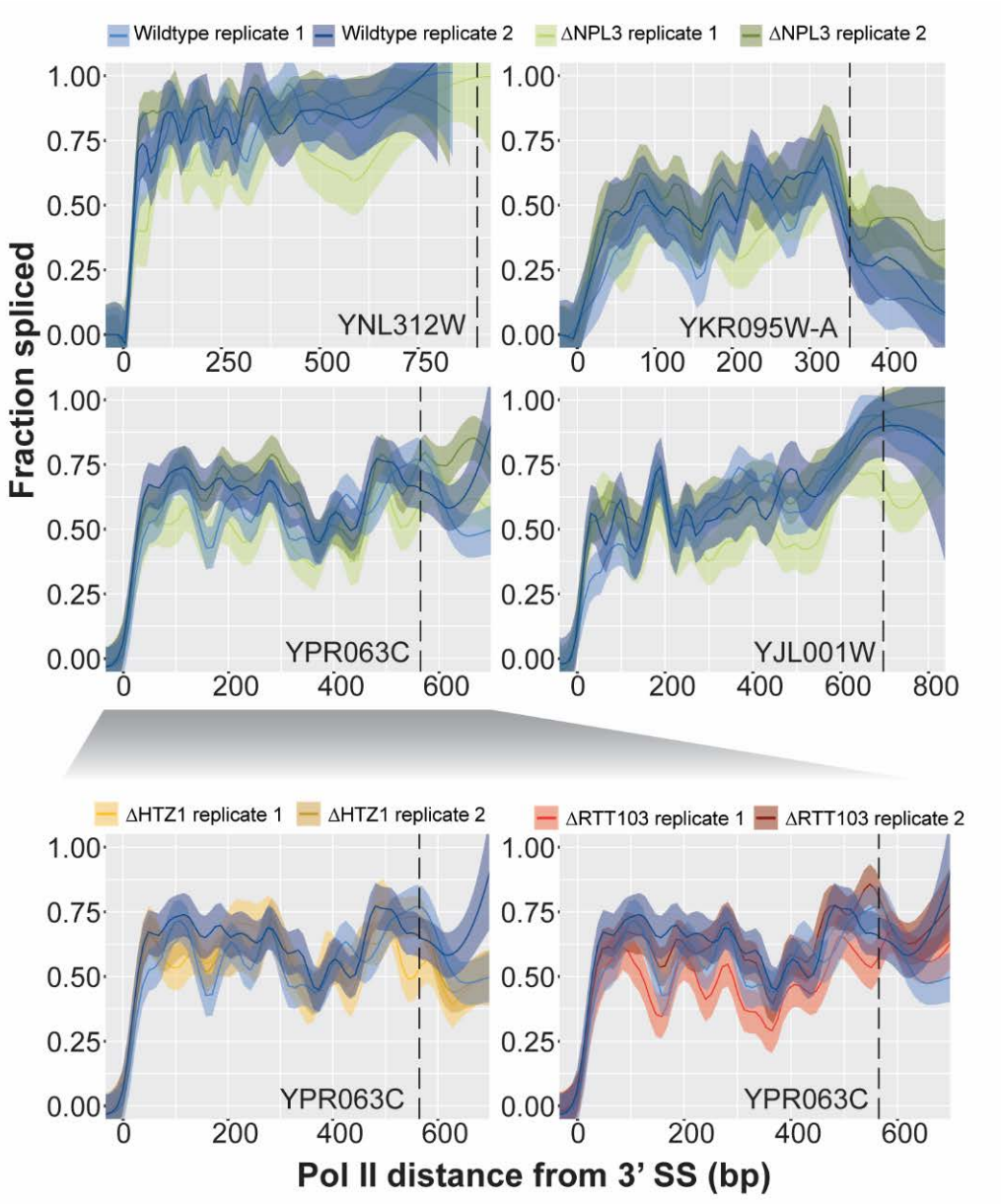

Figure 2.2 Co-transcriptional splicing profiles are robust to deletion of many non-essential factors

Example SMIT curves for four genes from the ΔNPL3 dataset (green) show that splicing is unchanged from wildtype (blue). Additional SMIT curves for YPR063C show that this is true for other datasets as well (ΔHTZ1 in yellow and ΔRTT103 in red).

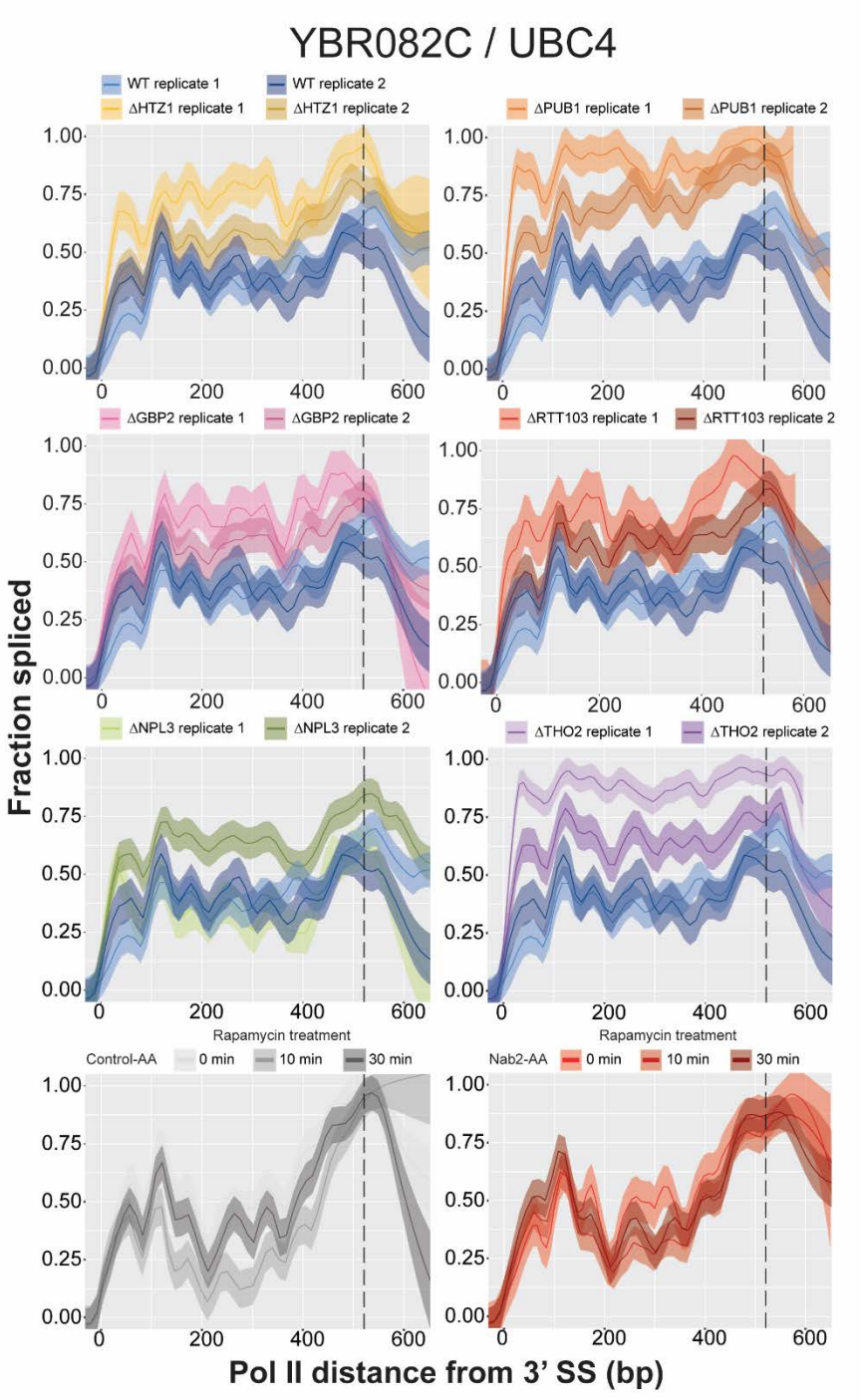

Figure 2.3. UBC4 is spliced better than wildtype in all deletion strains
SMIT curves for UBC4 (YBR082C) shown from all deletion strain datasets as well as the Control-AA and Nab2-AA dataset. Wildtype is displayed in blue, while each deleted factor has a unique color as indicated. PolyA site marked by vertical, dashed line.

2.3.3. **Nab2 depletion results in reduced co-transcriptional splicing**

Our machine learning model identified the evolutionarily conserved polyA binding protein, Nab2, as the factor with the most positive predictive power for co-transcriptional splicing levels. In budding yeast, Nab2 is essential for preventing mRNA decay in the nucleus (Schmid et al., 2015) and mutant Nab2 alleles result in increased intron-retention as observed by microarray (Soucek et al., 2016). Mass spectrometry shows that its vertebrate homolog, ZC3H14, has physical interactions with essential splicing factors such as U2AF2/U2AF65 (Soucek et al., 2016), suggesting that Nab2 plays an important and conserved role at several stages of mRNA maturation.

The Anchor-Away technique is ideal for depleting essential nuclear factors like Nab2 for short time intervals, minimizing secondary effects and preserving viability (Haruki et al., 2008). Nab2 is C-terminally tagged with the FRB domain of human mTOR and heterodimerizes with FKBP12 upon addition of rapamycin. Fusion of FKBP12 with the ribosomal protein RPL13A, which is quickly and efficiently exported to the cytoplasm, brings the bound Nab2-FRB to the cytoplasm and thus depletes it from the nucleus (Schmid et al., 2015). The strain is made resistant to rapamycin-induced secondary effects with the *TOR1-1* mutation. Complete depletion of the Nab2 Anchor-Away construct (Nab2-AA) from the nucleus was observed in as little as 5 minutes of rapamycin treatment (Schmid et al., 2015). We expect to see an immediate effect of Nab2 depletion on our nascent RNA because the nascent RNA is newly synthesized, whereas changes to steady-state RNA-seq measurements would be diluted by older RNA in a short time window.

We performed SMIT at 0-, 10-, and 30-minutes of rapamycin treatment for both Nab2-AA and an isogenic control strain expressing endogenous, untagged Nab2 (Control-AA).

In contrast to the deletion strains, Nab2 depletion has a significant effect on co-transcriptional splicing, generally reducing the fraction spliced for most pre-mRNAs. Examples in Figure 2.4A show the full range of gene-specific responses to Nab2 depletion, including instances of reduced splicing, improved splicing, and unchanged splicing. The Euclidean distance between the 10- and 30-minute treated samples and the 0-minute sample was used to quantify the difference between SMIT curves (ΔSMIT curve or ΔSC), using the first 300 bp binned by 60 bp to minimize the effect of sequencing noise. The density of reads declines exponentially along the length of the terminal exon, limiting the accuracy of values towards the PAS (Appendix Figure 5.4). The distribution of ΔSC values for Nab2-AA are significantly reduced from the Control-AA at 10-minutes (p-value = 4.28e-05) and 30-minutes (p-value = 0.0357) (Mann-Whitney U test) (Figure 2.4B). These results were validated by RT-PCR (Figure 2.4C).

2.3.4. **Depletion of Nab2 induces readthrough transcription**

In order to relate the splicing defect observed upon Nab2 depletion to other RNA processing steps such as transcription initiation, elongation, and RNA cleavage, I performed long read sequencing of nascent RNA in the Nab2-AA and Control-AA strains after 10 minutes of rapamycin treatment. A similar nascent RNA purification strategy to SMIT was used; however, the reverse transcription (RT) step was substituted for a strand-switching RT, adding common sequences to both 5′ and 3′ ends of the new cDNA. Global amplification of cDNA was followed by blunt ligation of Nanopore barcode adapters, size selection, and sequencing on a minION flow cell. Approximately 7 million basecalled reads (12.7 Gb) were generated.

A total count of splicing events revealed fewer instances of splicing in the Nab2-AA sample when normalized for total reads (Figure 2.5A), and a comparison of the fraction spliced for each gene revealed a downward shift in Nab2-AA (Figure 2.5B), in agreement with the reduced splicing observed in SMIT. Remarkably, we observe transcriptional readthrough past the polyA site (PAS) in both Control-AA and Nab2-AA (Figure 2.5C). In the control condition, readthrough transcripts were predominantly unspliced. Coupling between co-transcriptional splicing and cleavage has been observed previously in *Schizosaccharomyces pombe* (Herzel et al., 2018) and *Mus musculus* (Reimer et al.), but this is the first record of this phenomenon in *S. cerevisiae* (Figure 2.5D). Intriguingly, Nab2 depletion induces readthrough of both spliced and unspliced transcripts as seen for several examples (Figure 2.5 & Figure 2.6) and quantified genome-wide (Figure 2.5D), increasing both the quantity and length of readthrough transcripts. The striking effect of Nab2 depletion on cleavage was validated with RT-PCR for several genes (Figure 5.8). Overall, the fraction of reads extending past the PAS is considerably greater for the Nab2 depletion than the control for the majority of genes (Figure 2.5E).

2.3.5. **Splicing defects arise from upstream readthrough transcription**

One consequence of readthrough is that a single transcript can encompass the coding regions of several genes. So, a transcript that initiates in an upstream gene will continue transcribing into downstream genes, complicating the identification of which gene each read belongs to. Fragmentation of such a sample for classic Illumina RNA-seq would yield coverage across all encompassed genes, but the connection between these reads would not be interpreted correctly. With long reads, we can identify which reads initiated in an upstream coding region and readthrough into our gene of interest as opposed to reads that

initiated near the proper transcription start site (TSS) for our gene of interest. By separating these two classes of reads based on read start, it is clear that proper initiation is a key determinant of splicing efficiency (Figure 2.7B). This finding is exemplified by reads aligning to *YDL064W* in Figure 2.8 wherein readthrough from upstream genes impacts the fraction of the downstream intron spliced. We detect reduced splicing upon Nab2 depletion in the SMIT data for this gene as well (Figure 2.8B); however, the long reads provide additional information to interpret the mechanism behind this result. Namely, we are not detecting potentially productive transcripts that are failing to splice, rather we are detecting likely unproductive transcripts intruding into the coding region of the measured intron.

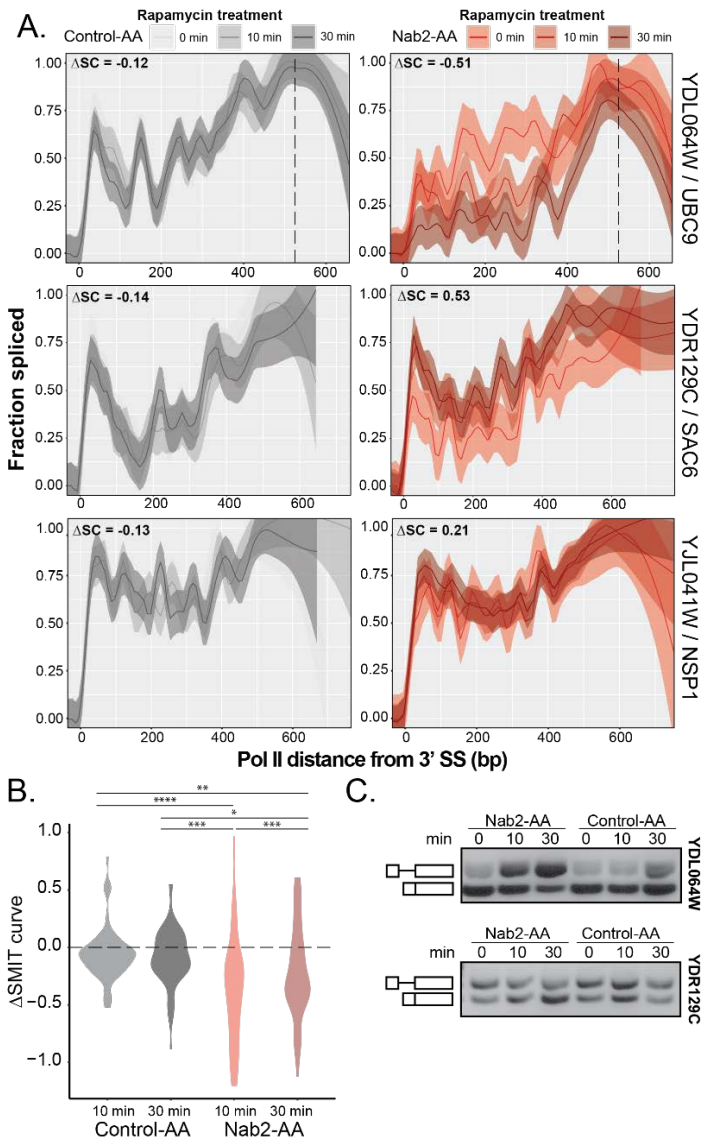

Figure 2.4. Nab2 depletion reduces splicing of most genes

A. Example SMIT curves show that splicing can decrease, increase, or remain the same upon Nab2 depletion. ΔSMIT curve (ΔSC) values are indicated for each curve between the 0- and 10-minute time points. The polyA site is shown as vertical dashed line (if data extend that far). B. Distribution of ΔSC values from the 0-minute time point for all samples are shown with significance (Mann-Whitney U test) as follows: $p \leq 0.05$ (*), $p \leq 0.01$ (**), $p \leq 0.001$ (***), $p \leq 0.0001$ (****). ΔSC was calculated as Euclidean distance from the 0-minute sample of first 300 bp after the 3′ SS was binned by 60 bp. C. RT-PCR validation for two genes from A. Random hexamers were used to reverse transcribe nascent RNA and intron-spanning primers amplified unspliced (top) and spliced (bottom) bands.

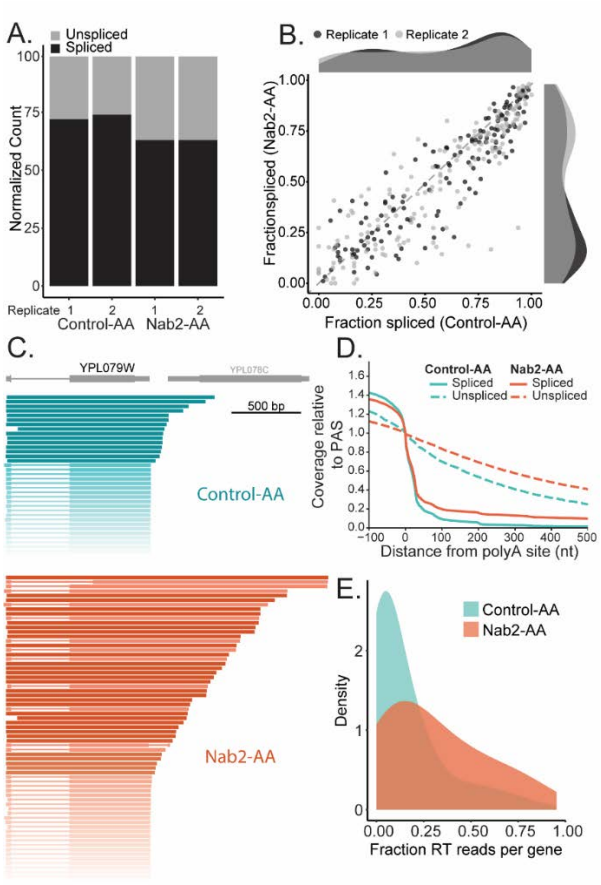

Figure 2.5 Readthrough transcription is elevated during Nab2 depletion
A. Number of spliced (black) and unspliced (grey) reads are normalized to total read count for each dataset. B. Fraction of spliced reads over total reads per gene are plotted. Y = X axis indicated by dashed line. Replicates are shown separately in light and dark grey. Density plots on the perimeter show distribution of points along x and y axis. C. Nanopore long reads aligned to YPL079W for Control-AA (teal) and Nab2-AA (orange). Scale bar = 500 base pairs (bp). Reads are filtered for proper TSS usage and arranged vertically by 3′ end location. Direction of transcription is to the right D. Read coverage normalized to coverage at the polyA site (PAS) is plotted over the region downstream of the polyA site. Control-AA (teal) and Nab2-AA (orange) reads are divided into spliced (solid line) and unspliced (dashed line) groups. E. Distributions of the fraction of readthrough (RT) reads per gene for Control-AA (teal) and Nab2-AA (orange).

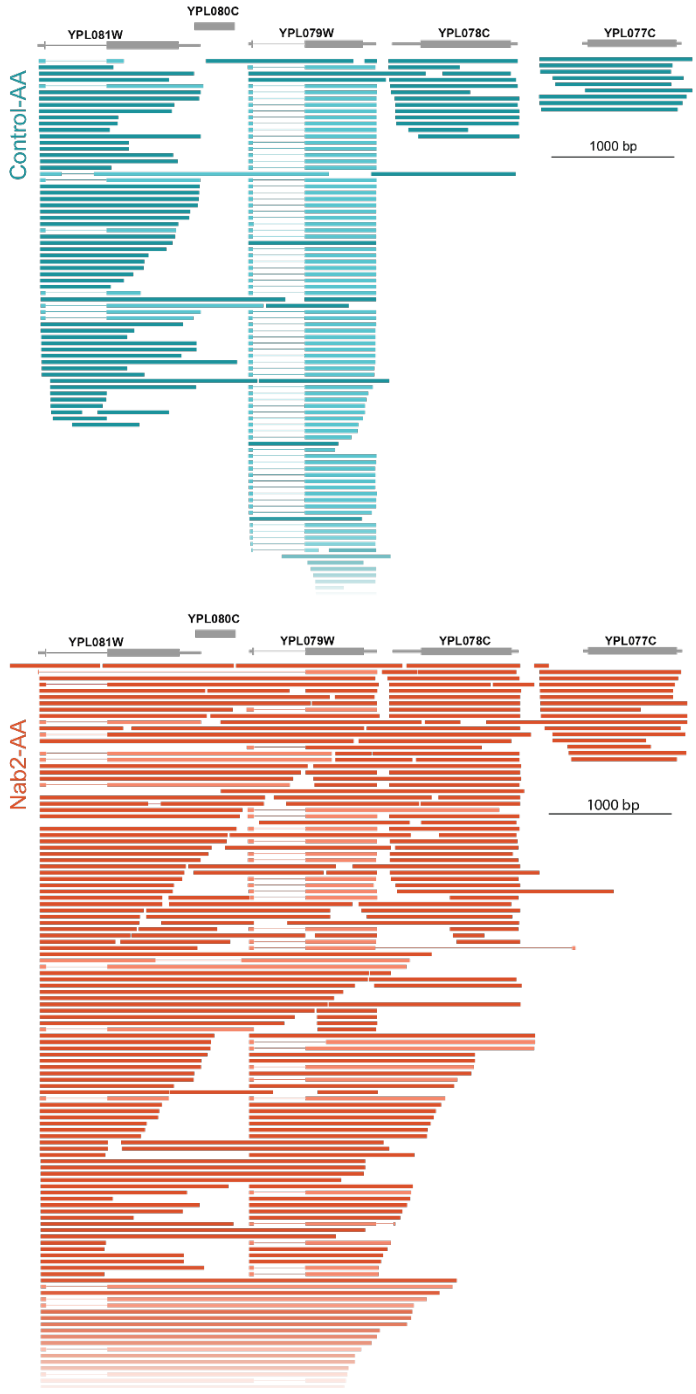

Figure 2.6 Nab2-induced readthrough disrupts the transcriptome
Long read data for Control-AA (teal) and Nab2-AA (orange) after 10 minutes of rapamycin treatment. Gene annotations (grey) are displayed above each bock of reads. Darker shades indicate unspliced reads and lighter shades indicate spliced reads.

2.3.6. **Readthrough is pervasive across entire chromosome**

The majority of intron-containing pre-mRNAs are cleaved efficiently, but a subset shows elevated levels of readthrough even in wild-type conditions (Control-AA) (Figure 2.5E teal). It is unknown whether common factors lead to readthrough. Are these pre-mRNAs inefficiently spliced? Perhaps they have weak PAS signals? One hypothesis is that readthrough is more common in areas of open chromatin where fewer roadblocks inhibit Pol II elongation. I determined the distribution of readthrough transcripts across each chromosome for two examples (Figure 2.9) and found fairly constant levels of readthrough across all binned genomic coordinates for both Control-AA and Nab2-AA, refuting this hypothesis. This figure again shows the elevated level of readthrough in Nab2-AA compared to Control-AA and proves that readthrough is a general phenotype of Nab2 depletion. The strong correlation between splicing and cleavage is conserved between budding yeast, fission yeast, and mouse, suggesting an important role in controlling gene expression. More analyses of these data, along with new experiments (described in Chapter 3) could help infer what drives readthrough.

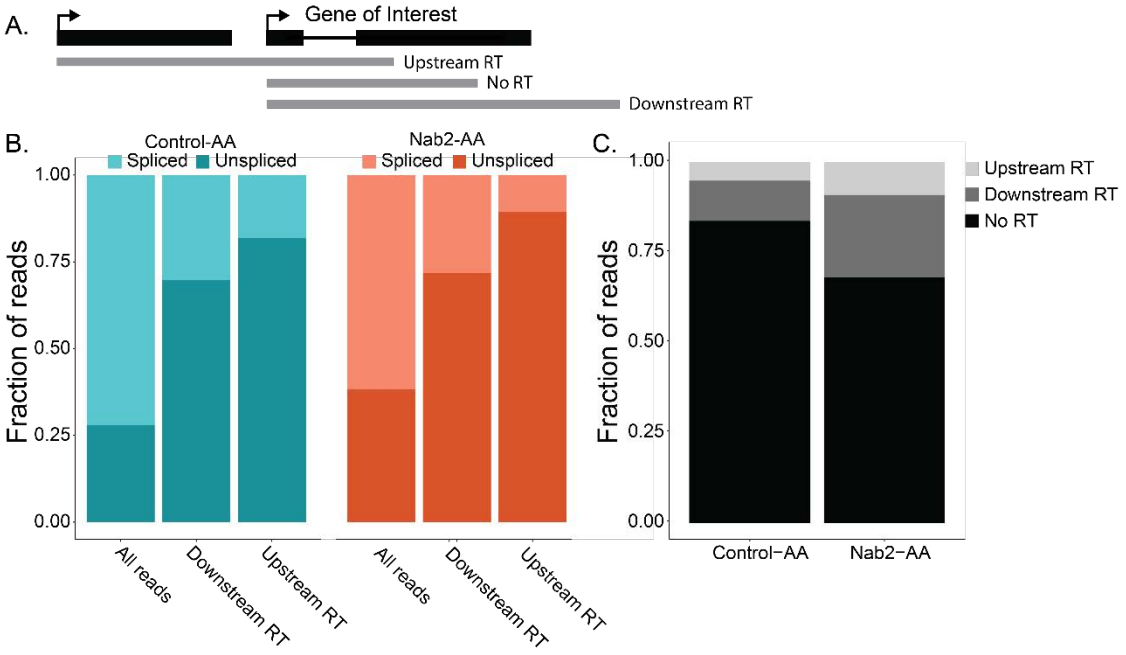

Figure 2.7 Two types of readthrough are both enriched for unspliced reads
A. Diagram of how reads (grey) are classified for each intron-containing gene of interest according to the genome annotation (black). Upstream readthrough (RT) reads initiate at upstream loci and readthrough into the gene of interest. Downstream RT reads initiate within the gene of interest but readthrough the 3′ end. B. Fraction of reads that are spliced/unspliced for subset of all reads, downstream readthrough reads, or upstream readthrough reads are shown for both Control-AA and Nab2-AA datasets. C. Fraction of reads that belong to each RT designation in Control-AA and Nab2-AA datasets.

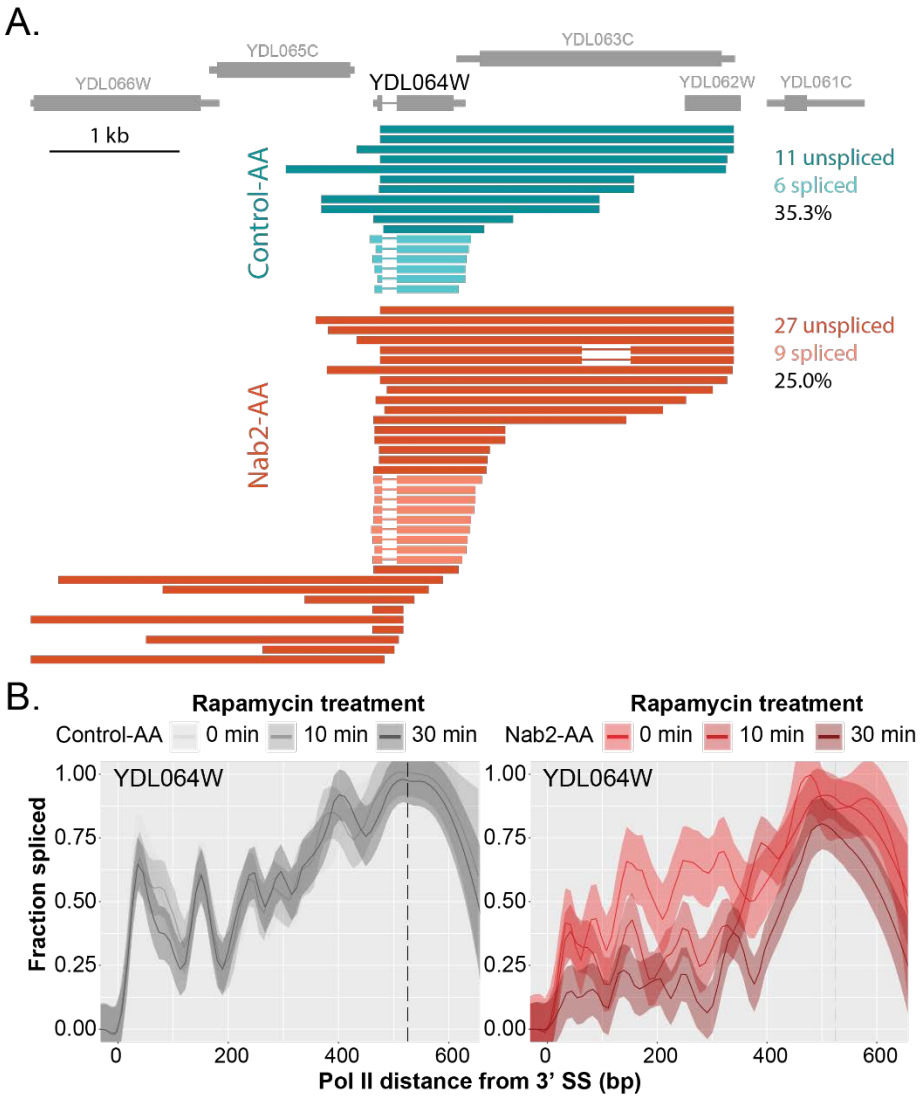

Figure 2.8 Comparison of long read and SMIT data for an example gene
A. Nanopore long reads aligned to the annotation (grey) for YDL064W from Control-AA (teal) and Nab2-AA (orange). Reads with the intronic sequence present (unspliced) are in the darker shade of the respective colors, while spliced reads are shown in lighter colors. Both samples were treated with rapamycin for 10 minutes. B. SMIT curves for YDL064W of the Control-AA (left) and Nab2-AA (right) samples.

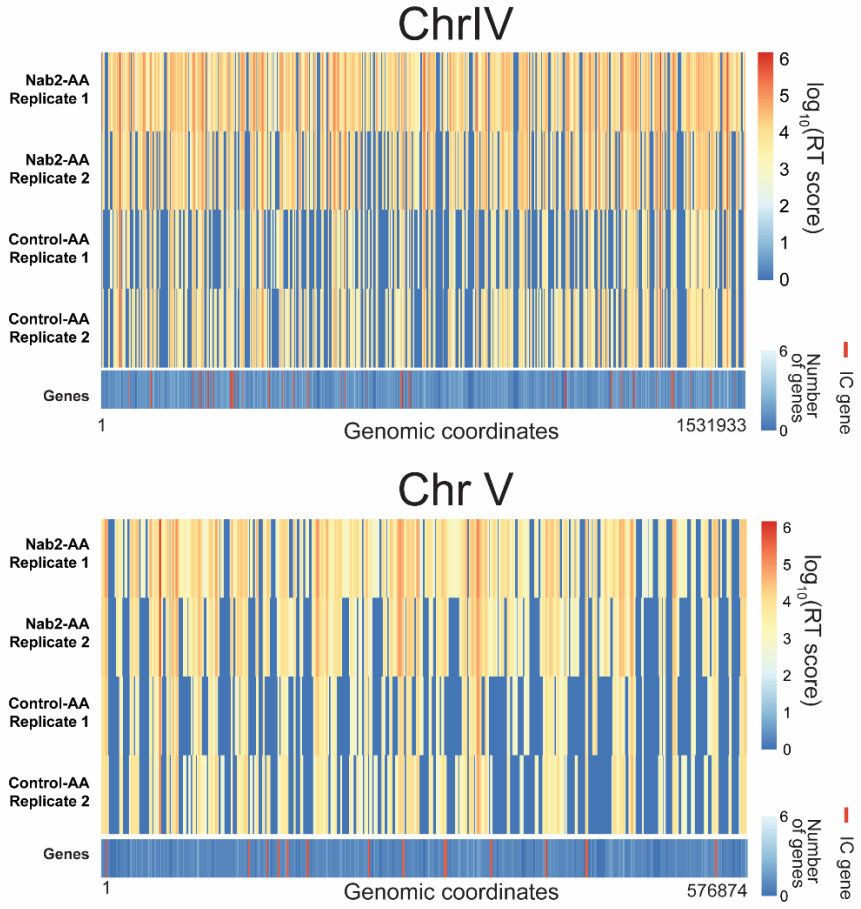

Figure 2.9 Readthrough events are distributed across entire chromosomes
Coverage of readthrough reads at each genomic coordinate was normalized to size of the dataset (RT score). Heatmaps for two example chromosomes are shown with each row corresponding to the indicated sample. Bars are colored to represent the $\log_{10}$ transformation of RT scores for each genomic coordinate bin (1,000 total bins). The bottom track shows the number of genes in each bin. Intron-containing (IC) genes are shown in red along the gene track.

2.4 Discussion

In the previous sections, I have asked the question of how co-transcriptional splicing regulation is achieved in a gene-specific manner. To do so, I combined two single molecule RNA-seq strategies applied on a global scale in budding yeast to generate gene-specific co-transcriptional splicing profiles and later test perturbations of genes identified by machine learning as potential regulators. Nab2 – the yeast homologue of ZC3H14, associated with intellectual disability in flies and humans (Pak et al., 2011) – emerged as an important candidate. As predicted, rapid (10-30 min) Nab2 depletion led to a reduction of splicing in some but not all genes. Unexpectedly, analysis by long read sequencing clearly showed that the predominant co-transcriptional role of Nab2 is on transcription; Nab2 depletion led to abundant readthrough transcription, whereby cleavage of nascent RNA at the polyA site fails and Pol II transcribes long distances downstream. I show that splicing defects are a secondary consequence of readthrough transcription, which causes Pol II to transcribe from upstream genes into intron-containing genes and apparently create a nascent transcript that cannot be spliced. Thus, Nab2 is not having a direct effect on splicing, but splicing is a byproduct of the increase in readthrough. This study therefore highlights the importance of considering the genomic context in which gene expression and RNA processing occur. Previous studies have identified Nab2-dependent splicing changes with microarray and concluded that Nab2 cooperates with the splicing machinery to regulate RNA processing because an increase in intronic signal is observed (Soucek et al., 2016). However, our long read results indicate that consideration of the broader genomic context is critical to understanding the true effect of perturbations such as Nab2

depletion on transcription and splicing, and previous studies have missed these conclusions.

There are several important considerations explaining why most of the model predictions showed no phenotypes with the genomic deletions. First, the factors targeted in this study were all viable deletions, and while some of them (specifically Tho2 and Npl3) did result in growth defects, most seemed unimportant for cell growth and survival. This means that the factors were either not serving a critical function in cell growth or that the cell can compensate for deletion of the factors. Second, due to the scale and cost of the experiment, the number of replicates we performed was limited. Several genes exhibited splicing phenotypes in various deletion strains; however, this effect was not replicated in the second sample. Further replication of this study could identify additional phenotypes that were missed here. Additionally, the machine learning predictions were clustered to enhance performance of the model, but we made individual deletions. Nab2 was alone in a feature group, while others such as Rtt103 were modeled together with several other factors, which could further explain why Nab2 showed a phenotype and the others did not. Additionally, Nab2 perturbation was assayed with a unique strategy from the other factors, and the short-time window of depletion from the nucleus could limit secondary effects that may confound the measured splicing values. Finally, we assayed 56 of the ~300 intron-containing genes in budding yeast, but it is possible that the machine learning model identified certain predictive elements based on a small number of highly regulated genes. For instance, it is possible that Tho2 strongly promotes splicing of genes with strong promoters, and it was by chance that we did not include these genes in our SMIT experiments.

The findings from the Nab2-AA long read data support the idea that transcription initiation and splicing are coordinated events. Indeed, I can see this coordination for certain genes in wildtype budding yeast data as shown in Figure 3.5 in the following chapter. This is further supported by previous work in the field. Transcriptional output is elevated by the insertion of an intron into a transgene in mice (Brinster et al., 1988; Nott et al., 2003), and there is evidence that promoter usage can drive splicing patterns (Cramer et al., 1997; Moldón et al., 2008; Nissen, 2017). Additionally, the C-terminal domain (CTD) of RNA Polymerase II may recruit splicing factors in a phosphorylation-specific manner (Harlen et al., 2016; Nojima et al., 2018). The phosphorylation state of the CTD changes rapidly across the body of the gene and is markedly different during initiation, elongation, and termination. Therefore, the read start position is likely to impact the phosphorylation state of the CTD when it encounters the intron.

Nab2 is an essential, predominantly nuclear protein that has been implicated in multiple steps of mRNA expression. Nab2 is thought to be a functional equivalent of mammalian PABPN1 given its role in defining polyA tail length, mRNA export, and RNA stability (Chan et al., 2011; Schmid et al., 2015). Initially identified as an mRNA export factor (Green et al., 2002), Nab2 is known to interact with proteins that associate with nuclear pores (Aibara et al., 2017; Soucek et al., 2016). Nab2's role in export and stability are likely related to its role in binding to polyA tails, where it appears to assemble into multimers (Aibara et al., 2017; Batisse et al., 2009; Tuck and Tollervey, 2013; Viphakone et al., 2008). Nab2 depletion leads to global loss of polyA+ mRNA, irrespective of whether the gene contains an intron; this effect was attributable to the nuclear exosome, indicating that Nab2's role in binding and/or export prevents decay (Schmid et al., 2015). Additional

reports demonstrate that the mammalian homolog of Nab2 interacts with the THO complex to coordinate RNA processing with nuclear export (Morris and Corbett, 2018). More recently, mutant Nab2 alleles were shown by microarray to increase the retention of a subset of introns in yeast (Soucek et al., 2016); the same study found that the human homologue associates with U2AF2, which recognizes the poly-pyrimidine tract of metazoan 3′ splice sites. Intriguingly, the yeast homolog of U2AF65, Mud2, may not be an orthologue, because Mud2 is non-essential. Yeast introns lack classical polypyrimidine tracts, and Mud2 does not bind specifically near 3′ SSs but rather along entire introns (Baejen et al., 2014). Thus, the mechanism whereby Nab2 could affect splicing in yeast was not apparent.

The results in this chapter demonstrate the power of long read sequencing for identifying coordinated transcription and RNA processing events. Previous studies employing long read sequencing of nascent RNA have noted coordination among multiple introns in the same transcript (Drexler et al., 2019; Herzel et al., 2018; Tilgner et al., 2018). Moreover, the correlation between the failure to splice and readthrough termination was first observed in *S. pombe* (Herzel et al., 2018). The data presented here in wildtype budding yeast show that this relationship is evolutionarily conserved. More broadly speaking, this emerging method is likely to transform how we analyze and draw conclusions about the effects of mutations that impact splicing in cells. Many studies from yeast to humans have perturbed the abundance of regulatory factors and used short-read RNA-seq to quantify the abundance of RNA isoforms. This study shows that the mechanisms underlying those results may be less straightforward than initially assumed. Finally, my findings underscore the importance of 3′ end formation and transcription

termination in ensuring the independent expression of genes. In renal clear cell carcinoma cells, transcriptional readthrough generates aberrant exons resulting in giant fusion transcripts originating from neighboring genes (Grosso et al., 2015). A high proportion of human diseases are associated with mutations in trans-acting splicing factors or cis-acting splicing regulatory elements in genes (Manning and Cooper, 2017), making it important to further investigate the mechanisms underlying splicing changes as well as the downstream consequences of splicing inhibition.

3. Splicing activity determines cleavage at polyA sites

3.1 Author Contributions

I completed all the cloning to tag proteins of interest with the auxin-inducible degron (AID) and validated the degradation time course of each. I prepared all AID nascent RNA samples presented here. I constructed all Nanopore sequencing libraries and ran the flow cells on the minION in our lab. Finally, I performed all data analysis for these experiments. Leonard Schärfen wrote a script for coverage over the region downstream of the polyA site that contributed to Figure 3.1D and Figure 3.3A.

3.2 Introduction

Evolution has crafted a diverse set of gene architectures across species. In higher eukaryotes such as mammals, short exons are interspersed with long introns that can span many kilobases (kb), with genes carrying on average 7-8 introns (Sakharkar et al., 2005). Organisms like yeast carry far fewer introns and many genes are indeed completely intronless. When they do occur in budding yeast, these short introns separate a first exon from a longer second exon. Splicing, or removing intronic sequences, is required for gene expression, but differing architectures require distinct mechanisms of locating the intron within the larger pre-mRNA sequence. Rather than communicate across the long distance of a mammalian intron, mammalian cells identify splice sites (SSs) across the exon in a model of SS recognition termed exon-definition (Berget, 1995). Yeast genes, on the other hand, rely on defining SSs across the intron (intron-definition).

Complications arise during the definition of first and terminal exons in mammals owing to a lack of canonical splice sites at the 5′ and 3′ ends, respectively. Instead, the cap-

binding complex (CBC) recognizes the 7-methyl-guanosine cap that is rapidly added to nascent RNA when RNA Polymerase II (Pol II) has transcribed a mere 20 nucleotides (nt) (Izaurralde et al., 1994, 1995; Listerman et al., 2006; Visa et al., 1996). The CBC aids in defining the first exon by interacting with splicing factors. Similarly, recognition of the polyA site (PAS) at the 3′ end of the terminal exon by the cleavage and polyadenylation factor (CPF) is required for efficient splicing of the final intron. The relationship between splicing and cleavage of the nascent RNA at the PAS is reciprocal such that mutation of the terminal 3′ SS inhibits cleavage at the PAS (Davidson and West, 2013; Dye and Proudfoot, 1999). This suggests a direct relationship between pre-mRNA splicing and cleavage in mammals.

The relationship between splicing and PAS recognition in yeast is far less clear. The spliceosome is a megadalton complex of RNA and proteins that is highly conserved between mammals and yeast. The 3′ end processing machinery, however, exhibits less homology, and numerous human factors have no yeast equivalent (Mandel et al., 2008). The lack of 3′ end processing homology together with an intron-centric approach to splicing leave very little evidence to suggest that splicing and cleavage are coordinated in yeast. Despite this, a previous study from our lab identified a strong correlation between the failure to splice and the failure to cleave in fission yeast (Herzel et al., 2018). There is currently no evidence to suggest how coordination between these processes is achieved in this system. Therefore, I set out to determine the nature of the relationship between pre-mRNA splicing and cleavage, using budding yeast as a model.

3.3 Results

To address the hypothesis that pre-mRNA splicing and cleavage are directly coordinated with one another and not correlated through indirect mechanisms, I targeted core components of splicing and cleavage processes for degradation with an auxin-inducible degron (AID) (Morawska and Ulrich, 2013). In this heterologous system, the degron is ubiquitinated by a constitutively expressed E3 ubiquitin ligase when bound to the plant hormone auxin, resulting in efficient degradation of the targeted protein. Short auxin treatment intervals (30-60 min) reduce the likelihood of secondary effects which become especially meaningful when working with required gene expression pathways. Long read sequencing of nascent RNA after perturbation of either process was used to relate splicing defects to a cleavage-deficient phenotype and vice versa.

Control samples confirm a similar correlation between splicing and cleavage in our budding yeast system (Figure 3.1A), as had been observed previously in fission yeast (Herzel et al., 2018). The control sample (teal) in Figure 3.1A shows an accumulation of unspliced reads with 3′ ends downstream of the PAS (readthrough transcripts), whereas spliced reads promptly disappear soon after the PAS is reached (presumably as they are cleaved). Fission and budding yeast are evolutionarily very distant organisms, and conservation of the connection between unspliced transcripts and disrupted cleavage points to the importance of this phenomenon for cellular function.

3.3.1. Cleavage inhibition has an indirect effect on splicing

Three separate proteins were targeted to disrupt cleavage of nascent RNA at the PAS: the sole endonuclease responsible for cleavage activity (Ysh1), an essential CTD-binding cleavage factor (Pcf11), and an elongation factor with a known cleavage phenotypes (Spt5)

(Baejen et al., 2014; Chan et al., 2011; Dominski, 2010; Ryan et al., 2004). Data from Ysh1 is a work in progress and will not be described in this thesis. Deletion of a traditional elongation factor, Spt5, was recently shown to impede PAS cleavage by 4tU-seq (Baejen et al., 2017). Our nascent RNA analysis, however, did not agree with this published report, which may highlight discrepancies between nascent RNA detection methods as well as protein depletion strategies. This will be further discussed in Chapter 5 (Appendix).

Pcf11 was of great interest to me for two reasons. First, Pcf11 is one of only two cleavage factors that bind the CTD of Pol II along the entire gene body (Ahn et al., 2004; Barillà et al., 2001; Kim et al., 2004; Licatalosi et al., 2002; Sadowski et al., 2003), which places this important cleavage factor near the spliceosome when splicing occurs. Second, Pcf11 was identified by our machine learning model in Section 2.3.1 as a significant predictor of splicing kinetics. Budding yeast cells harboring endogenously tagged PCF11-AID were treated with auxin or ethanol carrier (control) for 60 minutes. Nascent RNA was purified from chromatin and 3′ ends were ligated to an adapter as done for SMIT (Carrillo Oesterreich et al., 2010;Carrillo Oesterreich et al., 2016). Strand-switching reverse transcription (RT) uses the 3′ end adapter as a handle to synthesize full-length double-stranded cDNA. This global RT method does not require a specific forward primer, but rather adds untemplated nucleotides after completing synthesis of the first strand of cDNA, which are then used for priming the second strand. Generic primers complementary to the untemplated sequence additions contain an overhang which can be used for downstream PCR amplification. The library is amplified with less than 20 cycles of PCR and is then ligated to Nanopore barcode adapters, which facilitate sequencing on a minION flow cell. I obtained up to 12 Gb of data per flow cell, or ~7 million reads.

Pcf11-AID degradation results in the failure of 3′ end processing as determined by an increase in the number of reads which continue past the PAS (Figure 3.1 A, C, D), indicating that cleavage has failed to occur and that Pol II has continued transcribing downstream. This confirms that Pcf11 is a necessary component of 3′ end processing, as we had anticipated. The more pertinent question is how does this disruption to cleavage impact splicing? Reads were first classified into categories determined by their readthrough status, including readthrough past the PAS of that gene (Downstream RT), reads that start in an upstream gene and readthrough into the intron-containing gene of interest (Upstream RT), and reads that are completely within the gene body (no RT) (Figure 3.1B diagram). The fraction unspliced of both Downstream and Upstream RT reads is substantially greater than No RT in both the control and auxin-treated samples (Figure 3.1B). Downstream RT reads are better spliced during Pcf11-depletion, possibly because these reads have had more time for catalysis. Next, I calculated the normalized coverage over the region downstream of the PAS and found that coverage is similarly increased for both spliced and unspliced reads when cells are treated with auxin (Figure 3.1D). If impaired cleavage has direct consequences for splicing, we would expect to see increased coverage of unspliced reads specifically. The equivalent change observed for both spliced and unspliced reads shows that cleavage has no impact on splicing of the intron preceding that PAS. However, failure to cleave the transcript of upstream genes does impair splicing of the downstream introns into which the polymerase is intruding (Figure 3.1 B & E). These findings mirror those of Nab2-AA depletion in Section 2.3.5 and expand the findings to encompass a general feature of 3′ end processing inhibition, as opposed to a specific trait of Nab2. Thus, splicing defects arise from inhibition of nascent RNA cleavage and Pol II termination of upstream

PASs, but splicing is unaffected by inhibition of cleavage and termination of the intron-containing gene.

Pcf11 is a multi-functional enzyme whose different domains can function independently of one another (Sadowski et al., 2003). The CTD-interaction domain (CID) is necessary for both terminating transcription and cell viability, but deletion of this domain does not impact cleavage. In fact, cleavage-deficient mutants of Pcf11 can be complemented *in trans* with CID truncations. Our AID strategy targets the entire protein for depletion, impacting both cleavage and termination, and confirms that neither of these processes is sufficient to block splicing of the upstream intron.

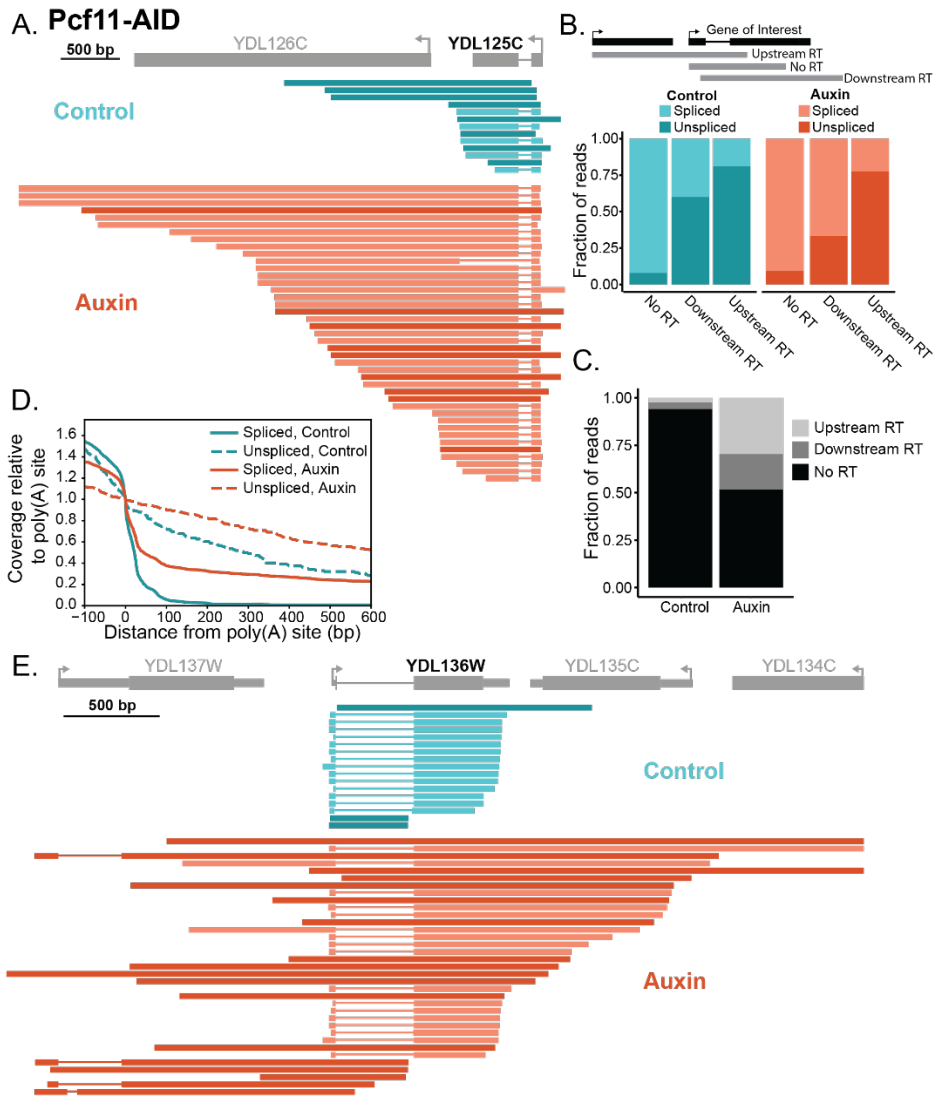

Figure 3.1 Pcf11-AID depletion induces readthrough without a direct effect on splicing

A. Nanopore reads from Pcf11-AID from control (teal) or auxin-treated (orange) samples are aligned to the gene YDL125C (grey) and sorted by 3′ end position. Darker colors represent unspliced reads, light colors are spliced. Scale bar = 500 bp. B. A diagram of readthrough (RT) classifications is shown (top) in reference to an intron-containing gene of interest. Annotated gene bodies are shown in black and example reads in grey. Fraction of reads that are spliced or unspliced in each RT category for the two samples. C. The fraction of total reads that belong to each RT category from the two samples shows an increase in the amount of readthrough in auxin-treated condition. D. Read coverage is computed over the region downstream of the PAS and normalized to the signal at the PAS. Line color and style are indicated in the legend in upper right corner. E. Nanopore reads are aligned to YDL136W and exemplify Upstream RT and its impact on splicing.

3.3.2. **Splicing inhibition has a direct effect on cleavage of a subset of genes**

The same approach was used for depletion of splicing factors Prp2 and Prp9 to induce splicing inhibition and measure their impact on cleavage. Prp2 is a DExD/H-box ATPase required for activation of the spliceosome before the first transesterification reaction. Despite substantial depletion of Prp2-AID within 60 min of auxin treatment (Figure 5.9), preliminary sequencing data revealed high levels of spliced transcripts (Figure 5.11). Given these results, it is possible that Prp2 ATPase activity is not a rate-limiting step of spliceosome assembly and residual protein levels were sufficient to reproduce near-wildtype conditions. Prp9 is a structural component of the early spliceosome complex, and depletion of Prp9-AID was highly effective at suppressing splicing (Figure 3.2B), although splicing was not completely blocked likely due to the stability of snRNPs and possible obstruction of the AID tag inside this complex. While the majority of pre-mRNAs exhibits reduced splicing (Figure 3.2B), the degree of splicing inhibition is gene-specific, suggesting that some transcripts have a different requirement for Prp9 as seen in Figure 3.4. Importantly, splicing inhibition directly impacts cleavage of the nascent RNA at the downstream PAS as seen by increased coverage of unspliced transcripts specifically downstream of the PAS (Figure 3.3A).

Although readthrough transcripts are widely distributed across the genome (Figure 2.9), they are found at relatively low levels with only a few reads per gene for most examples. This low expression may have biological roots, but is likely exaggerated by the inherent 5′ end bias of a nascent RNA sample as well as the PCR amplification, which enriches shorter molecules over longer ones. Direct sequencing of RNA or cDNA without amplification would address this issue. Meanwhile, I observe that the fraction of

readthrough per gene is significantly increased in the Prp9-AID depleted samples (Figure 3.3B), although the effect size is small (Figure 3.3 B & D). This small fold-change in readthrough could be due to the bias against detecting longer transcripts as discussed above, as well as the heterogenous and incomplete splicing inhibition. It is clear, however, that the link between splicing inhibition and readthrough is specific, as shown in Figure 3.3C, where the Downstream RT reads undergo a much larger shift to unspliced than the No RT reads during auxin treatment. This suggests that the splicing status of the nascent RNA molecule is being communicated to the 3′ end processing machinery and that the two processes are co-regulated to promote efficient gene expression.

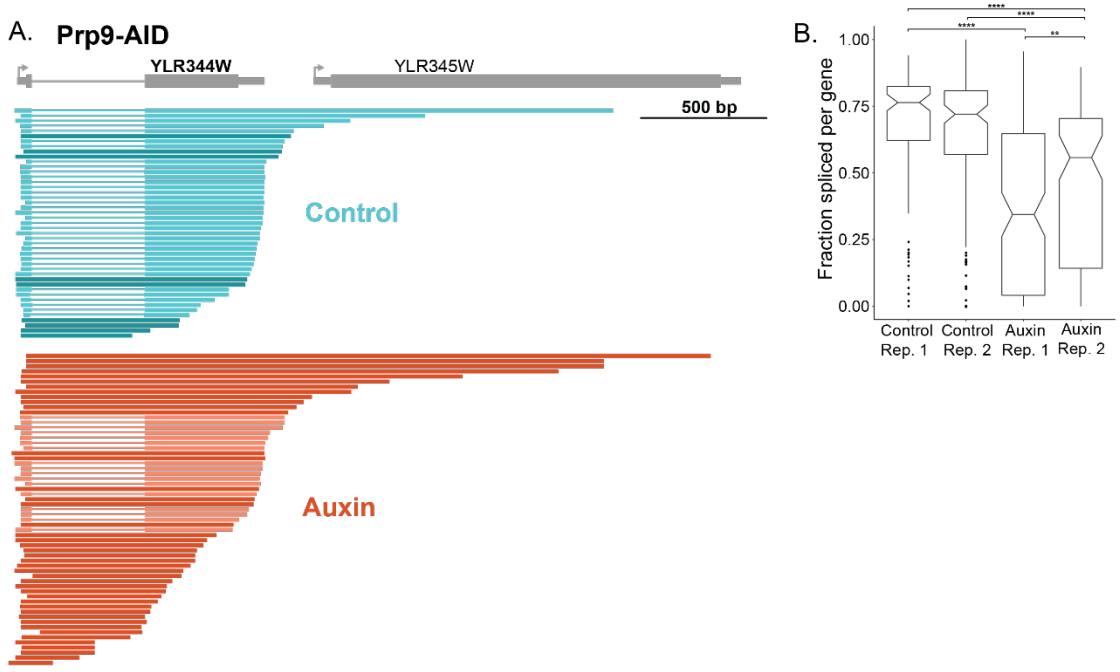

Figure 3.2 Prp9-AID depletion successfully inhibits splicing
Reads from the Prp9-AID control (teal) and auxin-treated (orange) sample are aligned to the annotation for YLR344W (grey). B. Boxplot shows distribution of fraction spliced values per gene for each replicate of control or auxin-treated samples (x-axis). P-values calculated with Mann-Whitney U test indicated as follows: $p \leq 0.05$ (*), $p \leq 0.01$ (**), $p \leq 0.001$ (***), $p \leq 0.0001$ (****).

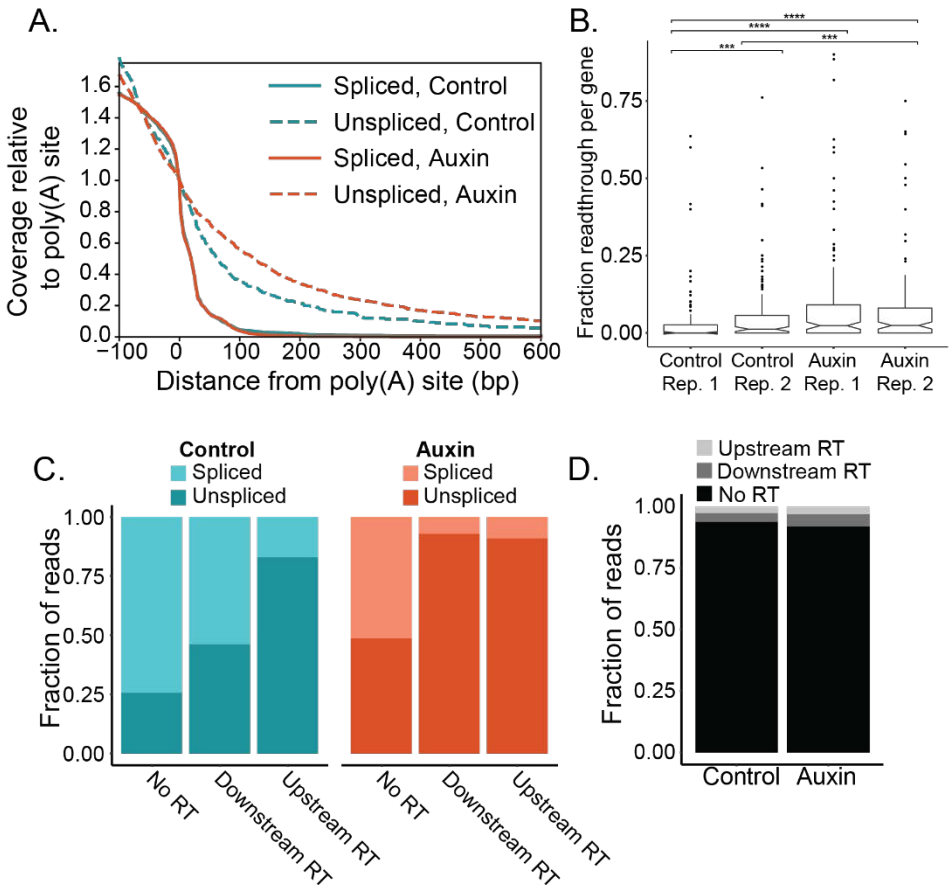

Figure 3.3 Prp9-AID mediated splicing inhibition promotes readthrough
A. Read coverage computed over the region downstream of the PAS is normalized to the signal at the PAS. Line color and style are indicated in the legend. B. Boxplot of the fraction of reads aligned to each gene which exhibit Downstream RT. P-values calculated with Mann-Whitney U test indicated as follows: $p \leq 0.05$ (*), $p \leq 0.01$ (**), $p \leq 0.001$ (***), $p \leq 0.0001$ (****). C. Fraction of reads that are spliced or unspliced in each RT category for the two samples. D. Bar plot shows the fraction of reads belonging to each RT category for the two samples.

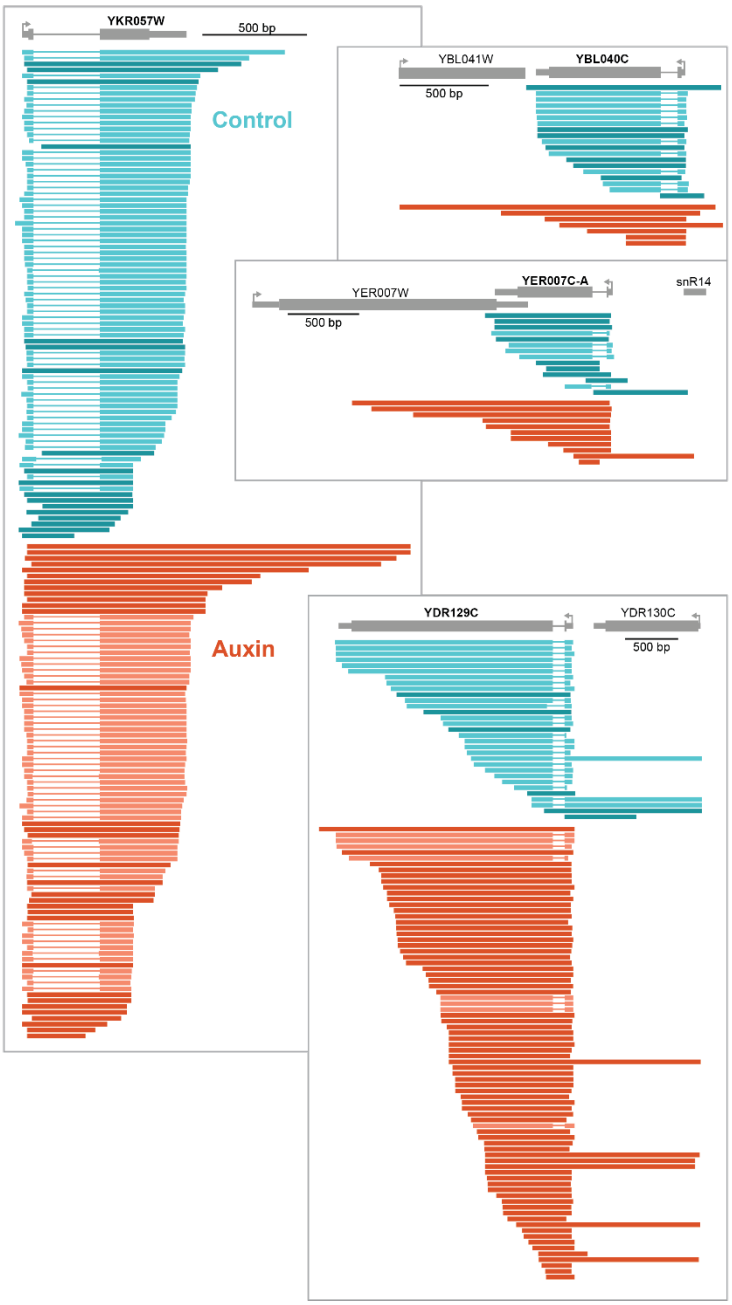

Figure 3.4 Examples of Prp9-AID sequencing reads
Additional examples of nanopore sequencing reads aligning to intron-containing gene annotations (grey). Scale bars all indicate 500 bp. Control sample colored teal and auxin-treated Prp9-AID colored in orange. Darker shades of each color represent unspliced reads, lighter shades represent spliced reads.

3.4 Discussion

Pre-mRNA splicing and endonucleolytic cleavage of nascent RNA at the PAS are coordinated processes in mammalian cells where exon-definition is the primary model of splice site recognition. No clear mechanism exists for such coordination in intron-defined systems like yeast, yet we continue to observe the same trends. In this chapter I have addressed the question of whether the coordination between splicing and cleavage in yeast is an indirect correlation or direct causal relationship. I show that inhibition of splicing has a direct effect on cleavage of those unspliced transcripts, prompting a failure of both cleavage and termination. Cleavage inhibition, on the other hand, does not have a direct impact on splicing of the upstream intron. However, a general splicing deficit in these samples is observed due to multi-gene fusion transcripts that initiate in upstream intronless genes and readthrough into downstream intron-containing genes, which fail to splice.

My results from Prp9-AID have important implications for our understanding of gene expression and the role that co-transcriptional splicing plays in the collaboration of cellular processes. Poorly-spliced genes are more likely to exhibit readthrough, which interrupts the coding regions of downstream genes. My research highlights how co-transcriptional splicing of one gene can affect expression of genes located nearby. Deep understanding is obscured by using short read data that cannot connect these related events. Sequencing of mRNA samples would reveal changes in expression of these downstream genes without the observation that such changes are a secondary effect of splicing regulation of neighboring genes. Readthrough is not restricted to intron-containing genes, however, and the consequences of transcripts from upstream genes that intrude into downstream introns are equally significant. Transcripts which readthrough into downstream introns fail to

splice with an extremely high frequency, and techniques such as microarrays and RNA-seq that investigate introns independently from their genomic context miss these biological conclusions. How often are reported changes to nascent RNA splicing actually a result of impaired 3′ end processing? While the answer to this question may remain obscure, the continued adoption of long read sequencing technology by the field will ensure that future studies take genomic context into account when measuring changes to transcription and splicing.

My findings confirm a direct relationship between splicing and cleavage, but further investigation is needed to determine a mechanism that mediates this coordination. One possibility is that components of a stalled, pre-catalytic spliceosome interact with cleavage factors and inhibit the cleavage process. This hypothesis is supported by the direct physical interaction between mammalian cleavage factor I (CF I_m) and U2AF (Kielkopf et al., 2001; Kyburz et al., 2006; Selenko et al., 2003). Additional support for this hypothesis comes from the finding that functional inhibition of U1 snRNP causes premature 3′ end processing (Kaida et al., 2010). In that model, the authors speculate that U1 binds to cryptic 5′ SSs throughout the gene body and antagonizes use of adjacent cryptic PASs. U1 recognizes the 5′ SS as one of the earliest steps in assembly of the spliceosome, but rearrangement into a catalytically-active spliceosome involves release of U1 as other splicing factors such as the U4/U6•U5 tri-snRNP are recruited (Wahl et al., 2009). Together, this information could lead one to hypothesize that the presence of U1 in stalled early spliceosome complexes would inhibit cleavage of that unspliced transcript. Stalling the spliceosome at stages where U1 has already been released but before catalytic activation would address this hypothesis. Prp3 is a component of the tri-snRNP, and AID depletion of this protein would stall

spliceosome assembly as described, before catalysis and without U1 present. Thus, a long read sequencing library of nascent RNA from Prp3-AID depleted cells, similar to the data shown in this chapter, would address the hypothesis that the U1 snRNP is involved in inhibiting cleavage of unspliced transcripts.

Several possibilities exist to explain why readthrough transcripts from upstream intronless genes fail to splice an intron in the downstream gene. First, there is evidence supporting promoter usage as a driver of splicing patterns, which helps explain tissue-specific alternative isoforms in mammals (Cramer et al., 1997; Moldón et al., 2008; Nissen, 2017). Perhaps intron-containing genes in yeast contain elements that recruit splicing factors to initiating units of Pol II, while Pol II at intronless genes are unprepared to recognize an intron. My own long read data from wildtype budding yeast further support the importance of promoter choice for co-transcriptional splicing; endogenous diversity of TSS usage is highly correlated with splicing of certain genes as shown in Figure 3.5. The data show that two different TSSs are used for the gene YDL125C and that the intron-proximal TSS is spliced much more efficiently. As a result, the majority of mRNA reads for YDL125C (taken from (Garalde et al., 2018)) belong to the TSS associated with better splicing. Second, the phosphorylation state of the Pol II CTD changes across the gene body to help recruit relevant factors. Ser5P and Ser7P are predominant at promoter-proximal regions then gradually disappear as Ser2P, Thr4P, and Tyr1P begin to accumulate over the second exon and reach a peak over the PAS (Herzel et al., 2017). Splicing factors are recruited by phosphorylation of certain CTD residues (Harlen et al., 2016; Nojima et al., 2018). How does the phosphorylation status of the CTD change downstream of the PAS? Does the CTD revert back to an initiation-typical phosphorylation state when it reads

through the PAS and encounters a second TSS? Can the CTD associated with readthrough transcripts recruit splicing factors? The answers to these questions are unknown and will be important to address for a robust characterization of the effect readthrough has on gene expression.

The fate of readthrough transcripts is under active investigation to establish whether they are degraded in the nucleus or stored for post-transcriptional processing. Re-analysis of traditional RNA-seq from *S. pombe* suggests that readthrough transcripts are targeted for degradation by the nuclear exosome (Rrp6) (Herzel et al., 2018), but the possibility remains that some of these molecules could be post-transcriptionally spliced and cleaved before polyadenylation and export. Exonuclease degradation of readthrough transcripts implies that eventually one end of the RNA is made accessible to the enzyme, but it is not clear how eventual cleavage or termination is achieved. Does the Pol II that was previously not termination-competent eventually encounter an even stronger termination signal that enables 3′ end processing? Perhaps prolonged exposure of a PAS to the nucleoplasm enables a chance encounter with cleavage factors despite the lack of proper CPF recruitment by the CTD.

Evolutionary conservation of the coordination between splicing and cleavage among two organisms as distant as budding and fission yeast indicates the cellular importance of this phenomenon. Moreover, mammalian cells also exhibit readthrough of unspliced transcripts despite the added complication of coordinating across multiple introns per gene (Reimer et al.). The 3′ end processing machinery in mammalian cells is generally similar to that of yeast, but has some important differences, including additional factors and documented involvement in defining the terminal exon. These distinctions suggest that the

coordination between splicing and cleavage in mammalian cells works via a different mechanism from that in yeast. If this is true, the coordination we observe must be incredibly important for cellular function for these organisms to have developed separate strategies to accomplish the same goal. Finding ways to uncouple the two processes (e.g. conditions where splicing is inhibited without detriment to cleavage) will be crucial for understanding how the cell uses this coordination to its advantage.

Alternatively, the unspliced readthrough transcripts that we observe could derive from a subset of the cell population in our bulk experiments that are experiencing stress or undergoing mitosis. Unspliced readthrough in this context could be the result of shutting down all gene expression systems or a regulatory response to the stress. Indeed, readthrough is induced by osmotic, oxidative, heat, aging, viral, and replication stress, but the splicing status of these events has not been investigated (Enge et al., 2017; Grosso et al., 2015; Muniz et al., 2017; Rutkowski et al., 2015; Vilborg et al., 2015). Hyperosmotic, heat, and oxidative stress induce expression of a class of long non-coding RNAs called DoG (downstream of gene) RNAs that result from readthrough of protein-coding genes in human cells and extend many kilobases downstream (Vilborg et al., 2015). DoG RNAs are distributed across the genome and remain attached to the chromatin. The authors postulate that DoG RNAs may regulate the functional genome when the nucleus is experiencing shrinkage and chromatin collapse associated with hyperosmotic stress. Furthermore, chimeric readthrough transcripts are elevated in clear cell renal cell carcinoma patient cells and actually correlate with reduced survival (Grosso et al., 2015). The authors of this study identify lowered expression of *SETD2*, a histone methyltransferase, as a putative contributing factor to readthrough. In agreement with this chromatin-centric model,

another study proposes that depletion of H2A.Z during senescence is responsible for readthrough transcription of convergent genes, reflecting the important role chromatin structure plays in intergenic regions (Muniz et al., 2017).

The lack of intergenic space in the yeast genome mandates that gene expression be a highly efficient process. A single instance of failure to properly cleave and terminate a nascent transcript can produce readthrough that disturbs expression of not only that one gene, but intrudes into coding regions of several downstream genes. Readthrough transcripts are observed to include the sequence of up to 5 different genes. This likely has consequences for chromatin accessibility and histone placement, which are tightly coordinated with transcription and gene architecture. Thus, tight regulation and coordination of these co-transcriptional processes are necessary to avoid small errors that propagate into far-reaching consequences for the cellular gene expression.

Figure 3.5 Transcription start site correlates with splicing status for YDL125C
Nascent RNA was reverse transcribed from wildtype budding yeast and amplified for sequencing on minION (left). Reads are separated vertically into populations that roughly align with two different TSSs and each population is sorted according to 3′ end. PolyA+ RNA was sequenced directly on the minION (right) (Garalde et al., 2018). Dark blue represent unspliced reads, light blue reads are spliced. Arrows represent approximate location of the productive TSS associated with splicing (green) and the unproductive TSS associated with unspliced reads (red).

4. Methods and Data Analysis

4.1 Constructing Strains

4.1.1. Constructing linear cassette for deletions

Saccharomyces Genome Deletion Collection strains were a generous gift from the Hochstrasser lab (Winzeler et al., 1999). Each gene locus was substituted with the *KanMX* gene which confers resistance to geneticin or G418, which was used as a selection marker for successful transformants. Deletions were newly made during the beginning of my PhD to limit compensatory mutations that arise in large collections. Genomic DNA was isolated from each deletion strain from 2 ml of saturated overnight yeast culture in YPAD. Cells were resuspended in lysis buffer (final 10 mM Tris-HCl, 1 mM EDTA, 100 mM NaCl, 1% SDS, 2% Triton X-100) with equal volume Phenol:Chloroform pH 8 and zirconia beads for vortexing. Spin at room temperature to separate phenol and collect aqueous layer for ethanol precipitation. Amplification of the KanMX cassette with homology arms for the genomic locus of choice was PCR amplified with primers in Table 5.4. Purified, linear PCR product is used as the linear insert for transformation in Section 4.1.3.

4.1.2. Auxin-inducible degron strains

AID background strains and plasmids were a generous gift from the Hochstrasser lab. PCR was performed with primers containing homology to plasmid bearing the AID*-9myc tag as well as 5′ overhang sequence homologous to the gene locus just upstream (forward primer) and downstream (reverse primer) of the stop codon. PCR purification of this linear amplicon was transformed into the AID background strain (Table 5.9) as described in Section 4.1.3. All strains in this study were prepared with c-Myc tags and kanamycin resistance, however, future experiments will require Prp3 to be tagged with an alternative

tag for verifying depletion via western blot (Prp3 runs at the same size as Tir1). Western blots were performed to validate the extent and timing of depletion. Whole cell extracts were made by vortexing cell pellets with 20% TCA and glass beads. Lysate was pelleted at 14,000 rpm for 10 min at 4°C. Pellet was resuspended in 1 M Tris-HCl pH 8.0 and boiled at 95°C in 30 mM Tris pH 6.8, 1% SDS, 5% glycerol, 0.1% bromophenol blue, and 50 mM DTT (final concentrations). Samples were run on a 4-12% gradient polyacrylamide Bis-Tris gel and transferred onto a nitrocellulose membrane for staining with Anti-c-myc (9E10) antibody (Santa Cruz Biotechnology #sc-40).

4.1.3. **S. cerevisiae transformation**

Yeast cells were grown in 50 ml YPAD complete media at 30°C and shaking at 200 rpm to an OD600 = 0.5 (logarithmic growth phase). Cells were pelleted and washed with sterile water before resuspension in 0.1M LiAc, 10 mM Tris-HCl, 1 mM EDTA, pH 7.4. One µg linear PCR product was added to cells with 10 µl single-stranded carrier DNA (salmon sperm DNA, Invitrogen). LiAc-TE-PEG buffer (1/10 of 10x TE, 1/10 of 1M LiAc, 8/10 of 50% PEG 4000) was added to 6x the volume of the cell mixture. Sample was incubated 30 min at room temp. while rotating on wheel. 70 µl of 100% DMSO (prewarmed) were added before heat shocking the samples for 15 min at 42°C. Cells were pelleted at 1,100 x g for 5 min at room temperature, resuspended in 300 µl YPAD and incubated on a wheel at room temp. for four hours, and plated on YPAD plates containing 350 ug/ml G418. After ~48 hours, single colonies were picked for culture growth and strain validation.

4.2 **Single Molecule Intron Tracking (SMIT)**

SMIT samples were prepared as in Carrillo, Herzel et al., 2016 *Cell* and Alpert, Reimer et al., *Methods in Mol. Biol.* (accepted).

4.2.1. *S. cerevisiae* growth conditions and harvest

Yeast cells were grown in YPAD complete media at 30°C and shaking at 200 rpm. For SMIT, cells were inoculated in 50 ml YPAD from a 5 ml culture and grown over night to an OD600 = 0.6-0.8 (logarithmic growth phase). Cells were transferred to a 50 ml Falcon tube and centrifuged at 4°C, 1100 g for 5 min. Pellets were washed once with cold 1x PBS and then transferred to a 2 ml tube for a last wash at 4°C , 1100 g for 5 min. Pellets were snap frozen in liquid nitrogen and stored at -80°C. For Anchor-Away strains 1 µg/ ml final concentration of rapamycin (Calbiochem, Cat. Number 553211) was added to medium for 10 and 30 minutes of incubation. The same concentration of rapamycin was added to 1xPBS for all washing steps until cells were snap frozen. Anchor-Away strains were a generous gift from Torben Jensen's lab (Schmid et al., 2015).

4.2.2. Nascent RNA preparation from Chromatin

Frozen cell pellets were prepared for cell lysis by resuspension in 1 ml buffer 1 (Table 4.1). Cells were lysed by vortexing with zirconia beads for 5 cycles of 1 minute shaking and 1 minute cool down on ice. Beads were filtered from cell suspension using a custom setup which places a 15 ml falcon tube with punctured bottom (22G BD needle) inside the carved-out lid of a 50 ml falcon tube. The filtering apparatus was centrifuged at 500 x g for 5 min at 4ºC. Supernatant was spun at 2,000 x g for 15 min at 4ºC, and then pellets were resuspended in buffer 2 (Table 4.1) and centrifuged at 20,000 x g for 15 min at 4ºC. Each round of centrifugation was performed twice, introducing clean buffer to ensure purity. Finally, pellets were resuspended in buffer P (Table 4.1) and phenol:chloroform:IAA (pH 6.0). The suspension was incubated with shaking (1150 rpm Thermomixer) at 37ºC for 1 hour. Tubes were spun at 13,000 rpm for 3 min at RT and the aqueous phase was transferred

to a new tube for precipitation with 3M Sodium Acetate pH 5.3 and 100% ethanol and incubation at -80ºC overnight. Samples were centrifuged at 20,000 x g for 30 min at 4ºC followed by a 1 ml cold 75% EtOH wash. Pellets were dried and resuspended in 80 µl of water for DNase treatment.

Table 4.1 Buffers for nascent RNA purification
Buffer 1 and 2 are sterile filtered and stored at 4ºC. Certain ingredients are added to each buffer aliquot immediately before use to avoid freeze/thaw cycles (labeled as fresh). Buffer P is sterile filtered and stored at room temp.

SMIT Buffer 1 components	Stock	Added to 200 ml final volume	Final concentration
HEPES pH 8.0	1 M	4 ml	20 mM
KCl	1 M	12 ml	60 mM
NaCl	5 M	600 µl	15 mM
$MgCl_2$	1 M	1 ml	5 mM
$CaCl_2$	1 M	200 µl	1 mM
Triton X-100	10%	16 ml	0.8%
Sucrose	fresh	17.12 g	0.25 M
DTT (fresh)	1 M	1000x	1 mM
PMSF (fresh)	200 mM	1000x	0.2 mM
Spermidine (fresh)	250 mM	100x	2.5 mM
Spermine (fresh)	500 mM	1000x	0.5 mM
SMIT Buffer 2 components	Stock	Added to 100 ml final volume	Final concentration
HEPES pH 7.6	1 M	2 ml	20 mM
NaCl	5 M	9 ml	450 mM
$MgCl_2$	1 M	750 µl	7.5 mM
EDTA	1 M	4 ml	20 mM
Glycerol	70%	14.3 ml	10%
NP-40	fresh	1 ml	1%
Urea	Fresh	12.01 g	2 M
Sucrose	fresh	17.12 g	0.5 M
DTT (fresh)	1 M	1000x	1 mM
PMSF (fresh)	200 mM	1000x	0.2 mM
SMIT Buffer P components, pH 5.0	Stock	Added to 50 ml final volume	Final concentration
Sodium Acetate	3 M	833 µl	50 mM
NaCl	5 M	500 µl	50 mM
SDS	10%	2.5 ml	1%

4.2.3. **DNase digest**

10 µl 10x buffer and 10 µl TurboDNase (Ambion) were added and solution was incubated at 37ºC for 30 min. The solution was cleaned up using the Zymogen RNA purification kit and eluted in 85 µl of water. 5 µl were saved to check for RNA integrity on an agarose gel and 80 µl were used for another DNase treatment as before. Finally, RNA was eluted from the column with 150 µl water (yield ~2 ug).

4.2.4. **PolyA+ RNA removal**

Removal of polyA+ RNA Oligo-dT coated cellulose was used (MicroPolyA Purist kit, Life technologies) and polyA- RNA was separated from polyA+ RNA following the manufacturer's instructions. In short, RNA samples were denatured for 2 min at 70ºC and then equal amount of Lysis/Binding buffer (150 µl) were added. Solution was transferred to a new tube containing 30 µl of prepared beads by pipetting up and down 10 times. Samples were incubated at room temperature for 5 minutes and then placed on magnetic rack to separate beads and supernatant. Depletion was repeated 2 more times (no more addition of Lysis/Binding buffer needed). After the 3rd round of depletion, supernatant was cleaned up using the Zymogen RNA purification kit and samples were eluted in 11 µl and concentration was measured on the Nanodrop (yields ~ 3-6 µg total).

4.2.5. **Adapter ligation**

3′ end ligation if a DNA adapter was used to label 3′ end of the nascent RNA and to create a universal template for reverse transcription. 600 ng of nascent RNA was used and nuclease free water was added to a total volume of 6 µl in PCR tubes. To each reaction 0.5 µl adapter (50 pmol, see table for sequence) was added and samples were incubated at 65°C (cycler) for 10 min and left on ice ≥ 1 min. For overnight reaction, 2 µl ligation buffer, 10

µl PEG 50%, 1 ul T4 RNA ligase (truncated K227Q, NEB) and 1 µl RNaseOUT were added and mixed well (final 50 mM Tris-HCl, 10 mM MgCl2, 1 mM DTT, pH 7.5, 25% PEG 8000, 40 U RNaseOUT, 200 U T4 RNA ligase II [truncated K227Q, NEB]). Extra samples with (+) and without (-) ligase enzyme were prepared for polyacrylamide gel confirmation of successful adapter ligation and RNA integrity. Reactions were incubated for 12 hours at 16°C. Unligated adapter and enzyme were removed using the Zymogen RNA purification kit. Samples were eluted in 13 µl water which was used as template for the reverse transcription reaction.

4.2.6. **SMIT library preparation**

Adapter ligated nascent RNA served as template for reverse transcription using SuperScript™ III Reverse Transcriptase (Invitrogen) according to the manufacturer's protocol with 0.5 ul SMIT_RT primer (5 µM) (Table 5.3). cDNA samples were diluted 10-fold for input into the first Phusion High-Fidelity (NEB) PCR with all 62 gene-specific forward primers pooled (1 µM each final) (Table 5.2) and adapter-specific reverse primer (SMIT_1st_5N_rev). A second round of PCR incorporates primers with Illumina adapters and barcodes (Table 5.3). Each PCR reaction was amplified for 15 cycles (30 cycles total). Reaction products were cleaned using the MinElute PCR Purification Kit (Qiagen) after each reaction, yielding ~ 10 µg final library.

4.2.7. **SMIT sequencing**

Samples were submitted to the Yale Center for Genome Analysis (YCGA) for size selection (250 bp – 1000 bp) and sequencing on Illumina HiSeq 2500 (High-Output Mode V4, paired-end, 2x75 bp read length). Up to 6 different samples were pooled per lane (~ 50MIO reads/ sample).

4.2.8. **Data analysis**

Fastq files were filtered for read quality with the FASTX toolkit and 3′ end linker sequences were trimmed with cutadapt (Martin, 2011) in forward and reverse reads. PCR duplicates were removed with Prinseq (Schmieder and Edwards, 2011), followed by 5 nt random 3′ barcode removal with the FASTX toolkit. Reads were mapped with paired-end, splicing-sensitive parameters using HISAT2 (Kim et al., 2019) to the *S. cerevisiae* genome. Custom scripts were written in R to extract splicing status and 3′ end position for plotting. Insert length bias correction was performed as described previously (Carrillo Oesterreich et al., 2016).

4.3 **SMIT RT-PCR validation**

Nascent RNA was purified from chromatin as described above and depleted of polyA+ RNA. Samples were reverse transcribed with SuperScript III and random hexamers. Intron-spanning primers amplified spliced and unspliced products which were qualitatively characterized on a 1% TBE-Agarose gel. Validation of Anchor-Away used random hexamer (Roche) RT priming, while deletion strain RT used gene-specific RT primers downstream of the PAS (Figure 5.6).

4.4 **Nanopore sequencing**

4.4.1. **Library preparation and Sequencing**

PolyA+ depleted, nascent RNA was prepared as described above (Sections 4.2.2 -4.2.4). Ribosomal RNA (rRNA) was depleted with up to three rounds of Terminator 5′-phosphate-dependent exonuclease (Lucigen) followed by cleanup with the Zymogen Clean and Concentrator kit. rRNA-depleted nascent RNA was adapter ligated as in Section 4.2.5 and

50-100 ng used as input for a strand-switching reverse transcription with the SMARTer PCR cDNA synthesis kit (Takara) with a PAGE-purified, custom RT primer (899_CDS_RT, below). Double-stranded cDNA was amplified using Primer IIA from the Takara kit and the Advantage2 polymerase (Takara) for 12 cycles. Purified product was then end-prepped with the NEBNext Ultra II repair/dA-tailing module (NEB) and ligated to Nanopore barcode adapters (Oxford Nanopore Technologies, PCR Barcoding Kit, SQK-PBK004) with Blunt TA/Ligase Master Mix (NEB). A second round of PCR using Nanopore barcode primers (ONT PCR Barcoding Kit) was performed with Advantage2 for 8 cycles. AMPure XP (Beckman Coulter) beads were used to clean up the sample between each reaction with a ratio of 2:1 (beads:sample) until the final cleanup where a ratio of 0.6:1 was used for size selection > 250 bp and eluted in 10 mM Tris-HCl pH 8.0 with 50 mM NaCl. Barcoded product was pooled for a total 25 ng and incubated with 1 µl RAP (ONT) for 5 min at room temp. MinION flow cells were brought to room temperature from 4ºC storage and washed with flow cell priming mix as described in ONT protocol. Pooled library was combined with sequencing buffer and library beads per protocol and loaded onto the flow cell and immediately sequenced on the MinION device for 48 hours. Four samples were typically loaded per flow cell to attain ~1-2 million reads per sample.

899_CDS_RT AAGCAGTGGTATCAACGCAGAGTACATTGATGGTGCCTACAG

4.4.2. **Data Processing and Analysis**

Raw fast5 files were basecalled with the high-accuracy model of Guppy 3.3.0 algorithm and demultiplexed with Qcat (https://github.com/nanoporetech/qcat). Barcode adapters were removed with Cutadapt (Martin, 2011) and reads were mapped to the *S. cerevisiae* genome with Minimap2 (Li, 2018). A custom script was written to filter out mapped reads

with soft-clipped polyA stretches. Finally, only reads overlapping intron-containing genes were considered for this study, with a required 50 bp minimum overlap and a read start position no more than 100 bp downstream of the annotated TSS.

Reads were classified into 3 groups regarding readthrough (RT) status. “Downstream RT” reads must start no later than 100 bp downstream of the TSS and terminate more than 150 bp downstream of annotated PAS. “Upstream RT” reads begin more than 100 bp upstream of the TSS. If a read starts both upstream of the TSS and ends past the PAS (as would be the case for a read that covers multiple gene bodies), then it is assigned as “Upstream RT”. All other reads are considered “No RT”.

4.5 Genome annotation used

For all experiments described here I used S. cerevisiae genome version 3 (sacCer3). For accurate representation of untranslated regions (UTRs) I matched experimentally derived UTR (Nagalakshmi et al., 2008) with the genome version used here.

4.6 Machine Learning

Machine learning was performed on previously published SMIT data (Carrillo Oesterreich et al., 2016). Code for the model can be found at: https://github.com/carrillo/SMITproject/blob/master/smit_r/machineLearning.R

5. Appendix

5.1 Machine Learning Model for Splicing Prediction

The model was trained on both ½ max and saturation value parameters from the previously published SMIT data (Carrillo Oesterreich et al., 2016). The saturation value was calculated as the mean fraction spliced of the last 4 bins (30 nt/bin) of available data for each gene. ½ max is the Pol II position at which half of the saturation value is reached. The distribution of saturation values was much wider than that of ½ max, suggesting that most genes have a similar onset of splicing and reach more varying levels of final splicing fractions. Heterogeneity aids model prediction, as it looks to identify trends in the data. Accordingly, the model was unable to predict ½ max value but was able to predict saturation value (Figure 5.1).

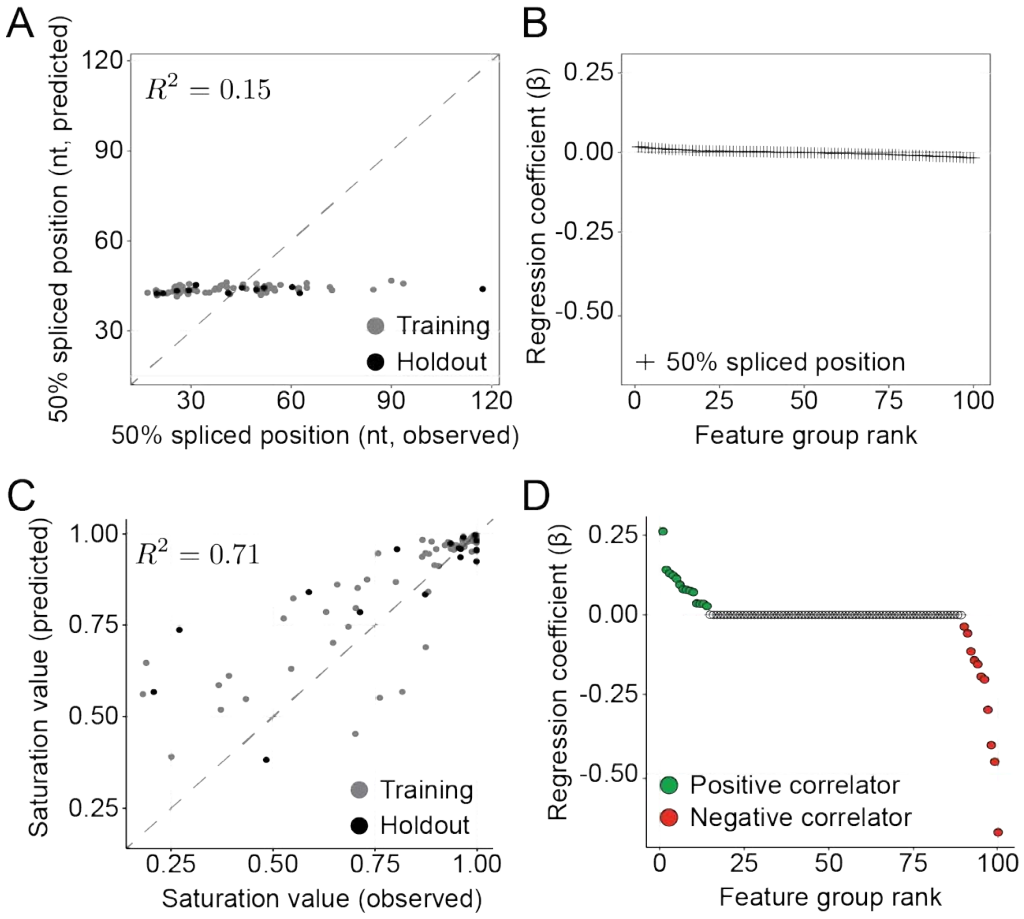

Figure 5.1 Model is unsuccessful at predicting half-max value
A. Observed position at which 50% of saturation value (1/2 max) is reached plotted against predicted 1//2 max position. Data was split into 80% training (grey) and 20% holdout (black). Correlation coefficient (R^2) indicated on plot. B. Learned regression coefficients β are sorted by value and plotted. Feature groups with positive (green), negative (red), and no (grey) correlation are indicated. C. Reproduced from section 2.3.1 with same plotting as in A for saturation value. D. Same as B but for saturation value.

Table 5.1 Machine Learning Predictions

All non-zero model components for saturation prediction. β value indicates relative importance of that feature group to the model's performance. Feature group strings are composed of location, SRA access number, protein identifier, and window around location.

β Value	Feature Group	Position	Annotated Function
0.262013404	fiveSS_SRR1523033_Nab2_50ntDownstream;threeSS_SRR1523033_Nab2_50ntDownstream;threeSS_SRR1523033_Nab2_50ntUpstream	fiveSS threeSS	Nab2: export
0.141637948	threeSS_SC_U1_Scer3_50ntUpstream;threeSS_SC_U2_Scer3_50ntUpstream;threeSS_SC_U5_Scer3_50ntUpstream	threeSS	U1: splicing; U2: splicing; U5: splicing
0.13033394	polyASite_SRR1523026_Ist3_50ntUpstream;polyASite_SRR1523029_Mpe1_50ntUpstream;polyASite_SRR1523031_Mud1_50ntUpstream;polyASite_SRR1523034_Nam8_50ntUpstream;polyASite_SRR1523036_Pap1_50ntUpstream;polyASite_SRR1523039_Snp1_50ntUpstream	polyA	Ist3: splicing; Mpe1 3' end processing; Mud1: splicing; Nam8: splicing; Pap1: 3' end processing; Snp1: splicing
0.123784504	polyASite_IP.Gbp2_Scer3_50ntDownstream;polyASite_IP.Gbp2_Scer3_50ntUpstream	polyA	Gbp2: 3' end processing
0.113969147	fiveSS_IP.TAP.Tho2_Scer3_50ntDownstream;fiveSS_IP.TAP.Tho2_Scer3_50ntUpstream;fiveSS_Ser2P_InputNormalized_Scer3_50ntDownstream;fiveSS_Ser2P_InputNormalized_Scer3_50ntUpstream	fiveSS	Tho2: elongation; Ser2: elongation
0.094677836	fiveSS_IP.Mft1_Scer3_50ntDownstream;fiveSS_IP.Mft1_Scer3_50ntUpstream	fiveSS	Mft1: elongation
0.08044451	polyASite_Thr4P_Scer3_50ntDownstream;polyASite_Thr4P_Scer3_50ntUpstream	polyA	Thr4P: elongation
0.078718025	terminalExonGCContent	other	terminal exon GC content: elongation

0.075647557	polyASite_SRR1523037_Pub1_50ntDownstream;polyASite_SRR1523037_Pub1_50ntUpstream	polyA	Pub1: 3' end processing
0.071139196	polyASite_SC_U1_Scer3_50ntDownstream;polyASite_SC_U2_Scer3_50ntDownstream;polyASite_SC_U5_Scer3_50ntDownstream	polyA	U1: splicing; U2: splicing; U5: splicing
0.036028723	polyASite_SRR1523042_Yra1_50ntDownstream	polyA	Yra1: export
0.034906804	fiveSS_IP.Gbp2_Scer3_50ntDownstream;fiveSS_IP.Gbp2_Scer3_50ntUpstream;fiveSS_IP.Npl3_Scer3_50ntDownstream;fiveSS_IP.Npl3_Scer3_50ntUpstream;polyASite_Ser2P_InputNormalized_Scer3_50ntDownstream;polyASite_Ser2P_InputNormalized_Scer3_50ntUpstream	fiveSS polyA	Gbp2: 3' end processing; Npl3: export; Ser2-P: elongation
0.034672598	threeSS_IP.aSub2_Scer3_50ntDownstream;threeSS_IP.aSub2_Scer3_50ntUpstream;threeSS_Nrd1_Scer3_50ntDownstream;threeSS_Nrd1_Scer3_50ntUpstream;threeSS_Pcf11_InputAndMockNormalized_Scer3_50ntDownstream;threeSS_Pcf11_InputAndMockNormalized_Scer3_50ntUpstream;threeSS_Rtt103_Scer3_50ntDownstream;threeSS_Rtt103_Scer3_50ntUpstream;threeSS_Spt6deltaC_InputNormalization_Scer3_50ntDownstream;threeSS_Spt6deltaC_InputNormalization_Scer3_50ntUpstream	threeSS	Sub2: splicing; Nrd1: 3' end processing; Pcf11: 3' end processing; Rtt103: 3'end processing; Spt6DeltaC: inactive elongation
0.027394976	polyASite_SRR1523026_Ist3_50ntDownstream	polyA	Ist3: splicing
0.000287952	polyASite_SRR343339_mnase_50ntDownstream;polyASite_SRR343339_mnase_50ntUpstream	polyA	Mnase: chromatin
7.51E-06	threeSS_SRR343339_mnase_50ntDownstream;threeSS_SRR343339_mnase_50ntUpstream	threeSS	Mnase: chromatin
-0.000823405	terminalExonLength		

-0.036863396	BPSRank;BPSLevenshteinDistance	other	BPS: splice site strength
-0.058349216	fiveSS_SRR583969_h3k4me3Rep1_50ntUpstream;fiveSS_SRR583970_h3k4me3Rep2_50ntUpstream	fiveSS	H3K4me3: chromatin
-0.114566538	X5SSRank;X5SSLevenshteinDistance	other	5SS: splice site strength
-0.14243186	fiveSS_SRR1523038_Rna15_50ntDownstream;threeSS_SRR1523038_Rna15_50ntDownstream;fiveSS_SRR1523038_Rna15_50ntUpstream;threeSS_SRR1523038_Rna15_50ntUpstream	fiveSS threeSS	Rna15: 3' end processing
-0.155177712	fiveSS_SC_U2_Scer3_50ntDownstream;fiveSS_SC_U2_Scer3_50ntUpstream	fiveSS	U2: splicing
-0.193930073	polyASite_SRR488715_h3k9acRep2_50ntDownstream;polyASite_SRR488715_h3k9acRep2_50ntUpstream	polyA	H3K9ac: chromatin
-0.202983428	threeSS_IP.aYra1_Scer3_50ntDownstream;threeSS_IP.aYra1_Scer3_50ntUpstream	threeSS	Yra1: export
-0.298257312	PolyPyGCContent	other	PolyPyGC: splice site strength
-0.409010528	polyASite_SRR1523043_Yth1_50ntUpstream	polyA	Yth1: 3' end processing
-0.461094358	threeSS_SRR583969_h3k4me3Rep1_50ntUpstream;threeSS_SRR583970_h3k4me3Rep2_50ntUpstream	threeSS	H3K4me3: chromatin
-0.682882832	polyASite_SRR654067_h2az_50ntDownstream;polyASite_SRR654067_h2az_50ntUpstream	polyA	H2Az: chromatin

5.2 SMIT optimization

Before preparing SMIT libraries for the deletion and AA strains, we optimized the protocol and analysis pipeline to increase efficiency and accuracy of the technique. We used a subset of 24 genes for optimization.

5.2.1. Protocol optimization

Two primary goals drove our optimization strategy. First, in the original protocol, ever gene assayed with SMIT was PCR amplified in a separate reaction, which quickly becomes unwieldy to manage experimentally and not cost-effective when performing SMIT for many genes in many samples. Therefore, our first goal was to test whether pooling gene-specific forward primers (Table 5.2) would have an impact on our results and interpretations (Figure 5.2). Second, size bias is introduced into the sample during PCR amplification, which preferentially amplifies shorter molecules, which may lead us to detect more spliced transcripts than unspliced at a given location. This bias is corrected bioinformatically (explained further below), however, we wanted to test ways of reducing this bias biochemically by including a forward only PCR reaction. This unique PCR strategy uses only a forward primer and therefore repeatedly generates only one strand of the DNA molecule, generating a linear increase in product rather than the exponential increase of traditional PCR. We reasoned that inclusion of a linear amplification step could reduce the number of traditional PCR cycles we needed for a library. We found that insert size was unaffected by pooling primers and by amplification under 35 cycles (Figure 5.3), therefore we settled on Protocol 2 (pooled primers and 30 PCR cycles) moving forward.

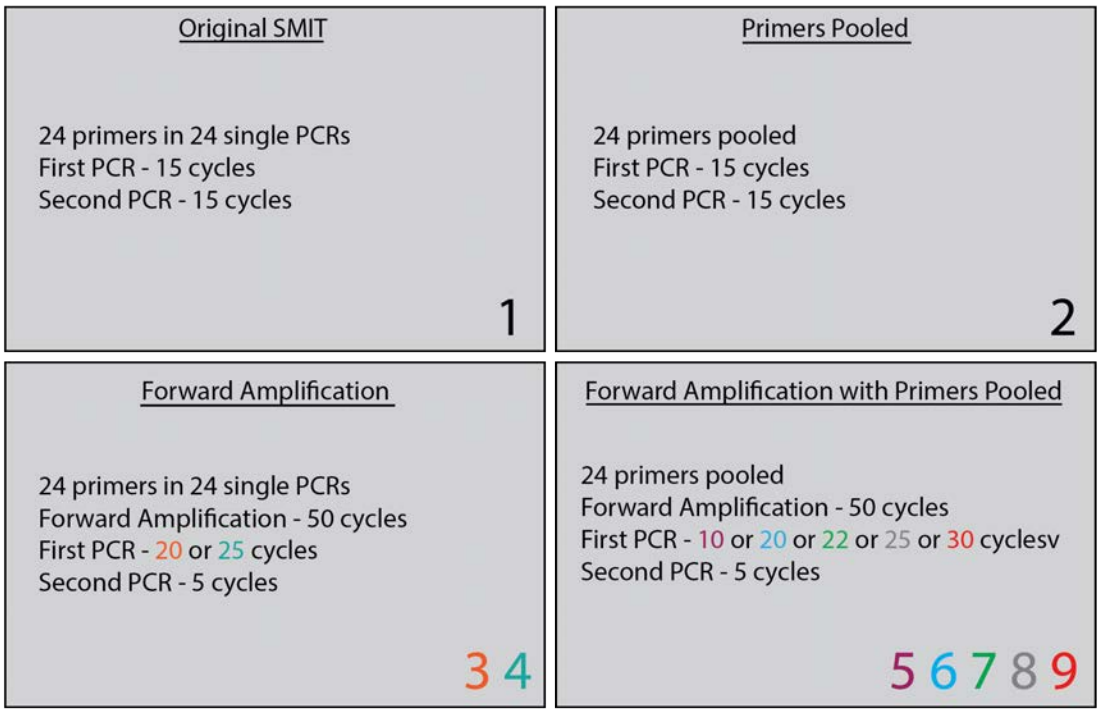

Figure 5.2 SMIT optimization strategy
Each version of the SMIT protocol is described in the boxes above with the sequencing index number colored accordingly.

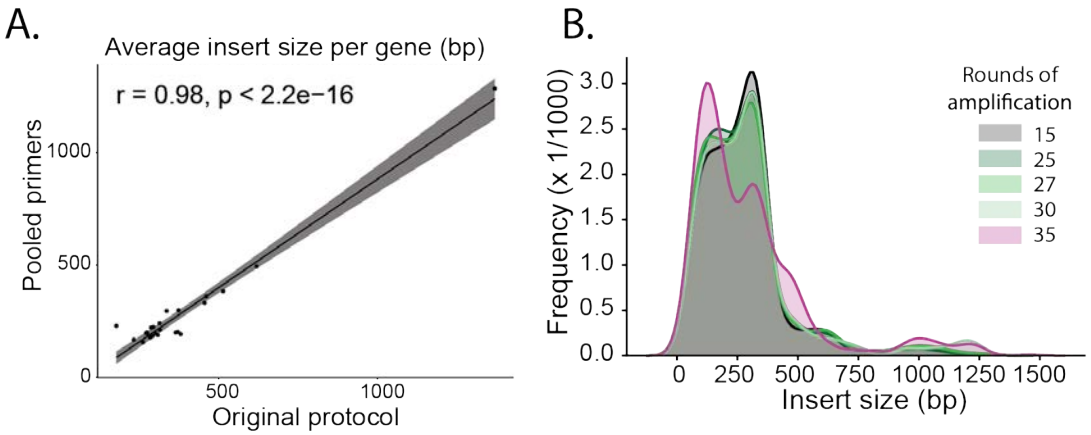

Figure 5.3 Results from optimization trials
A. Mean insert size per gene in the original protocol data set (1) plotted against the pooled primer protocol dataset (2). Linear fit (black line) r- and p-value are indicated. B. Histogram of insert size distribution between 0 and 1500 bp in different protocols. Data sets are color coded by the number of PCR cycles indicated in legend.

Table 5.2 SMIT gene-specific forward primers

SMIT primers all begin with the universal 5' start primer sequence (overhang) followed by a gene-specific sequence. Primer numbers are for internal reference and can be found in -20°C storage. All primers ordered with desalt purification.

Systematic name	Sequence	Primer number	Intron length (bp)
Universal 5′ start	GACGTGTGCTCTTCCGATCT		
YNL265C	CTCCGTCAATGATTCCGTTC	16	105
YGL030W	CAATTAATCAATATACGCAGAGATG	114	230
YDR025W	CCACTGAATTAACCGTTCAATCCGAG	78_new	339
YDR064W	AAATGGGTCGTATGCACAGT	79	539
YHR021C	AAAACACACCAGATAATTAGTGCAT	138	550
YML024W	GTCAAGATCTCGAGACTAGCAATAAC	202_new	398
YFL039C	ATTTACTGAATTAACAATGGATTCTG	109	308
YBR189W	GAAAGCAACAAGACAACTAAGACTAAGC	53_new	413
YFR031C-A	AAGAAACCATTAGATCAATAAGCAA	111	147
YAL003W	GCTGACAAGTCATACATTGAAGG	34	366
YBR082C	CGTATTGCTAAAGAACTAAGTGA TCTAGAAAG	45_new	95
YDL130W	CCGCTGGTGCTAATGTCG	71	301
YBL087C	CGGTGCTCAAGGTACTAAGTTT	41	504
YGR214W	GTTGGCTGCTAACACTCACTTAGG	131_new	455
YDL191W	CCGTAGAATAGGTACAGTGAGACA	75	491
YDL136W	CGATAAAGAACCAAATAGGACTAAAAA	72	405
YML056C	TTTCTCTGGCTTCCCAGTTA	207	408
YMR242C	CAAGTCTTTTAAGGAAAAACAGTGCGG	221_new	477
YPL218W	GTTGGGATATTTTTGGTTGGT	255	139
YJL001W	TGAAAAAGGGCGAAGTCAGT	27	116
YDR129C	CCAAAACACAATGAATATTGTCAAA	82	111
YDL125C	CTTGATGCTGCCTGTATTTTT	70	111
YPR028W	TCACTCTCAAATGAAACAATTCG	31	133
YDL029W	AATGGACCCACATAATCCAA	63	123
YCR028C-A	CAATGAACCTCAAATCAATTTTT	60	83
YER133W	CTAGAGTTAGAAGCCCCAATTAAA	104	525
YLR093C	TTATATTTTTACCAAATGAAACGCT	185	141
YMR225C	ATCCCTTTGGCAAGGAAG	219	147
YDR381W	AGGGACATTAAGCAGGATGC	29	766
YBR078W	GCTATTCTAAGTGCCTCCGC	24	330

YHR001W-A	TTAAAACTAACCTCAATGGCG	135	63
YPR063C	GCCCGACCTTTGTGTTTC	260	86
YDR092W	ATCATTACCCAAGAGAATAATCAAG	80	268
YDL012C	TCTGCCTCCAAACAAAGC	62	86
YBR230C	CAGCATCTCATAATATGTCTGCAA	56	97
YNL112W	GTTGCTACTGATGTGGCCG	28	1002
YBL050W	TGTCAGACCCTGTAGAGTTATTGAA	22	116
YBL040C	GCAATGAATCCGTTTAGAATCTT	38	97
YPR187W	CATGTCAGACTACGAGGAGGC	263	76
YDR139C	TCAACAAAGACTTATATTCCAGGG	83	73
YPL081W	GAAAACTAATACAGCAACAGAAA TACAAAAGTATAC	248_new	501
YGR183C	AACAATAGCAATACGGACTAAAATG	130	213
YJL191W	CAATAACAATTAAGAATGGCTAACG	163	408
YDL219W	CGTCGATTCAAAAGTTATTTCAAG	76	71
YER003C	AGCTGTTCAGGTTAGATGCAG	95	93
YNL312W	CTAGTTTAAGCATATACATAATGGCAA	236	108
YDL064W	TGTGTCTACAGCGTCTTCAGG	30	110
YDR059C	GCCAAGGAATTAAGTGATTTAGGGAG	15_new	90
YJL041W	AGTAATAAGCTCTGATCGTTTTGAA	158	118
YGL137W	GGACACGATGAAGTTGGATATAAA	120	200
YNR053C	TGGTACCTACCTGGGTTGC	237	531
YMR033W	GCTCCATTTAGGCAGGACAG	25	86
YDR005C	CCTAAAGAATCACGACAATGAAA	77	80
YKR095W-A	AAACGGGAAAAGTCACTGGA	8	75
YMR194C-B	TCCATGCCAGAAGGAGGC	217	72
YHR077C	AATACATTGGACAGAAATTATGGAC	141	113
YMR079W	GCCCTAAAACACAATGGTTACA	213	156
YHR101C	TGCAAAACCAGACCAATGTT	3	87
YER093C-A	GGGCCATAAAAGTACGAAAAT	101	75
YOR153W	ACTAGCTACTCCTCCGCGTC	898	NA
YAL012W	CGAACCCATTTCTTTGTCCA	18	NA
YBR152W	AGAGCATCCAGACCAAAACG	19	NA

Table 5.3 SMIT generic primers

SMIT_2ndPCR_fwd_index includes a 6 bp barcode (XXXXXX) that are listed below (SMIT_index1-10). SMIT_1st_5N_rev and SMIT_2ndPCR_rev primers are PAGE purified; all others are HPLC purified.

Name	Sequence
SMIT 3′ end adaptor	/5rApp/NNNNNCTGTAGGCACCATCAAT/3ddC/
SMIT_RT	GATTGATGGTGCCTACAG
SMIT_1st_5N_rev	TTTCCCTACACGACGCTCTTCCGATCTNNNNN GATTGATGGTGCCTACAG
SMIT_2ndPCR_rev	AATGATACGGCGACCACCGAGATCTACACTC TTTCCCTACACGACGCTCTT
SMIT_2ndPCR_fwd_index	CAAGCAGAAGACGGCATACGAGAT-XXXXXXGTGACTGGAGTTC-AGACGTGTGCTCTTCCGATCT
SMIT_index1	CGTGAT
SMIT_index2	ACATCG
SMIT_index3	GCCTAA
SMIT_index4	TGGTCA
SMIT_index5	CACTGT
SMIT_index6	ATTGGC
SMIT_index7	GATCTG
SMIT_index8	TCAAGT
SMIT_index9	CTGATC
SMIT_index10	AAGCTA

PCR is necessary for nearly all sequencing techniques, but amplification can bias the sample for shorter transcriptions. Since intron removal shortens the transcript, this would make it easier to detect spliced reads than unspliced reads. To correct for this bias, we include three intronless controls whose inset size (length of the RNA transcript) distribution we can more easily correct (without the complication of partial intron removal) and apply this same correction to our intron-containing genes. We can see that detection of

insert sizes falls off increasingly towards the end of the gene (Figure 5.4), meaning that the beginning of the curve is more reliable than more distant Pol II positions.

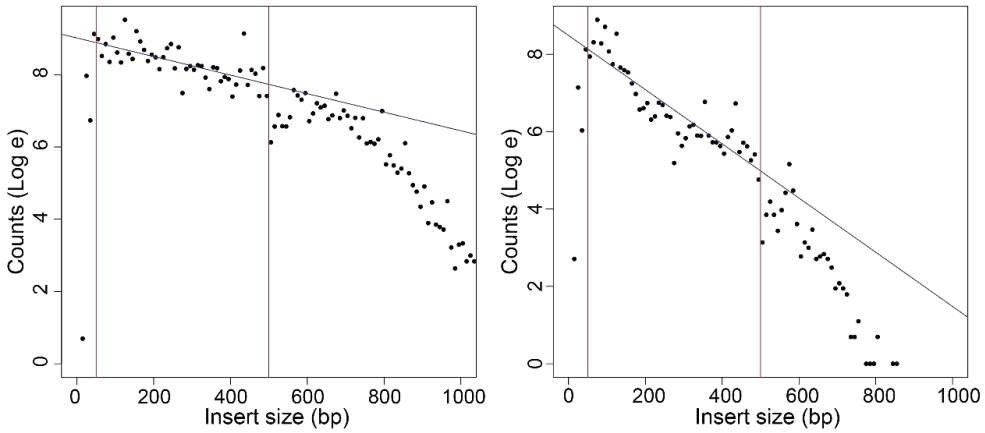

Figure 5.4 Insert size distribution of reads from intronless genes
Two examples of insert size distributions from different SMIT datasets. Counts (y-axis) of insert sizes (x-axis) are shown on a natural log scale. Cutoffs are set manually (red) to define region for calculation of slope from best linear fit (blue).

5.2.2. **Analysis optimization**

We realized that reads with Pol II positions further away from the 3′ SS became increasingly difficult to detect and thus the data was less reliable. So, we shifted our analysis to focus to the beginning of the curve (Figure 5.4). Additional pipeline optimization included updates of mapping software from older algorithms to Hisat2 and plotting of SMIT curves with Loess model fitting and 1 std. dev. confidence intervals.

5.2.3. **SMIT Replicate analysis**

Determining the number of replicates to perform for SMIT experiments is a balance between cost efficiency and accuracy. Six genes were chosen for extensive replication to ascertain how much variation should be expected. We were surprised to find that some genes were highly variable (YBR082C, Figure 5.5), while others were highly reproducible.

All samples were processed side-by-side except for replicate 5 which had to be reprocessed separately and the subsequent increase in variability arising from this sample can be seen in the data for two genes (YPR028W and YJL191W).

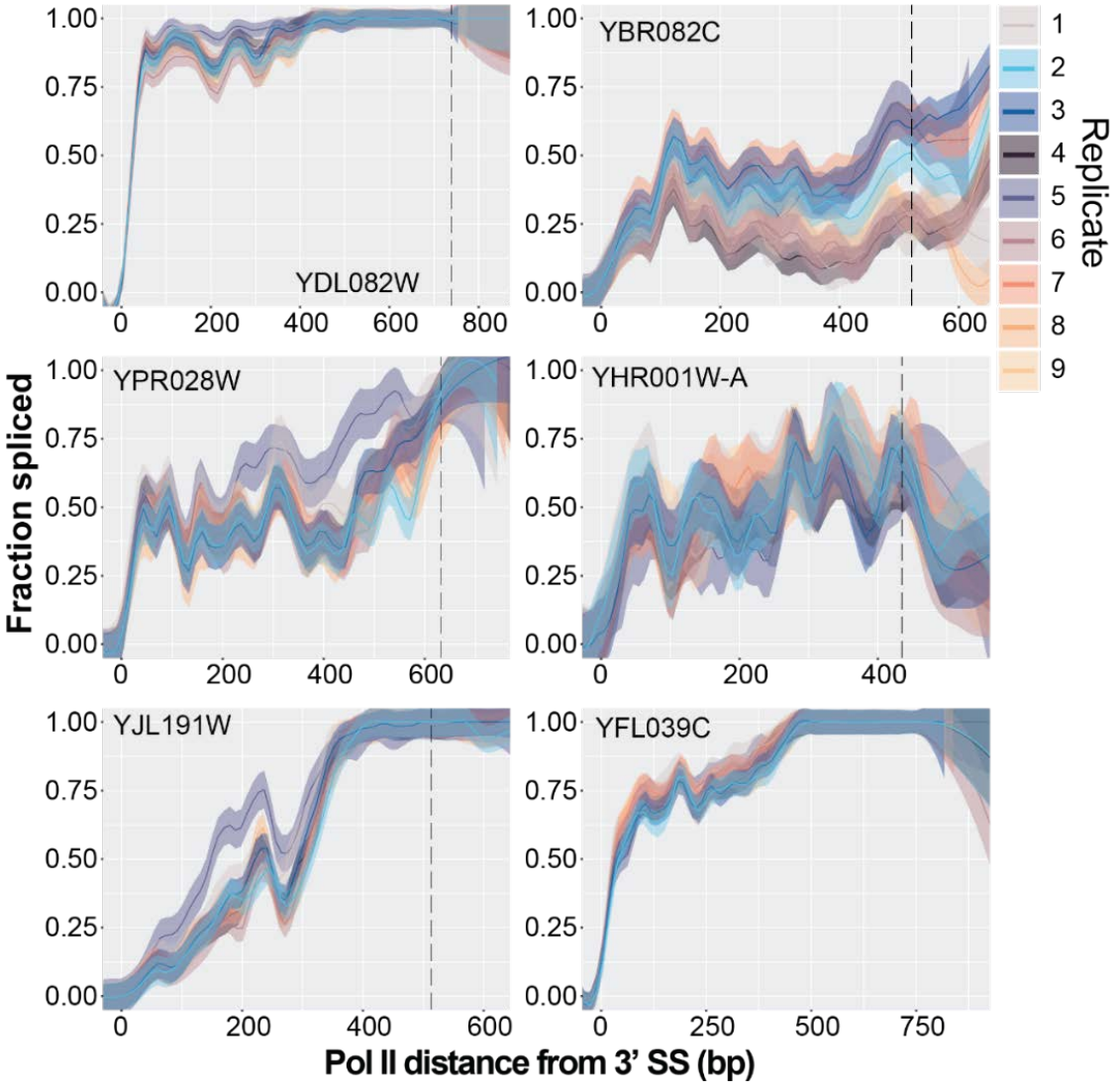

Figure 5.5 SMIT replicates reveal natural variation in some genes
Nine SMIT replicates are shown for six genes. All samples were prepared in parallel with the exception of Replicate 5.

5.3 SMIT experiments

Table 5.4 Deletion strain cloning primers
These primers have the homology arms for the locus in the primer name as well as regions that prime the *KanMX* deletion cassette.

Primer name	Sequence (5′ to 3′)
416_H2Az_UP45L	GGACTCTATTATACTTCAGTCGGAAAAAATATC GTTAAATTCAATTTCGCACTATAGCCGCACGT ATGATACAGACGTTAATATGGATAGAATGAGG ATACAGGAGCAGGGAGAATTACGGGAAATGGG AA
417_H2Az_DN45L	
418_Rtt103_UP45L	AATATAAGTACAAGCGAAGAAGGTTTTAACGA AGAGGTAGAAAATTTGAAGAAAGCAATAATCC A
419_Rtt103_DN45L	TACGAAAAAAGACTGGACTTAAACGGGTTACT ATATATTTGTATAAGTTATCTCCTTGTTTTCTT
422_Pub1_UP45L	TAGAGTTTAATCTTCCCTTCTATCATCTTAATAG CAGGTTCATAATAAGAAGATTACCACATCTA CTTTTTTTATTTTATCTTGGGATTTGTAGGTTGC
423_Pub1_DN45L	CTCTCTTTATTCTTTCTTTTTGTTTCATTCC AATCAATTATATTGAGGATTTCCCATTAGAAAT
424_Gbp2_UP45L	AAGCTATGGGCGAAAAGGAAACAAACATCAGC
425_Gbp2_DN45L	GTTATTTATAACCCGCCCGCTTCCTTATTATTTA TACGTTATCATAAAGTACACAGGTCATGGTT AGAAAAATAATTTCCTCTCTTCTAAATATATAT
426_Npl3_UP45L	ACTTTTGAAGGAATCAAAATTAAGCAATTACG TTCTCAGTCTCATATTTAAGTTTTAAAACAATTC ATATCTTTTGTTAATTTCTCCTTTTTTTTTCTCAA
427_Npl3_DN45L	CTATATAAATGGCTTACGGTGTCGGTCTCGT CAGTTGATACATATTCGCACCAGTATACATTTT CAGGACTTTATGGATGTCCACGAGGTCTCTTTC
874_Tho2_UP45Kan	TATCCGCGATGGCTGGTCGTACGCTGCAGGTCG ACGG
	GTACACGTTAAAATTCAGCTCGGGTATGTTAAG TACTAGTAATTACGGTGTCGGTCTCGTAGAACA
875_Tho2_DN45Kan	TCTCGCAAGGCGGGTAATCGATGAATTCGAGCT CGTTTTCGAC

Table 5.5 Deletion strain validation primers

These primers were used for sanger sequencing validation of proper cassette insertion into the locus of choice. KanB and KanC prime sequences within the cassette out into the surrounding genomic region. All other primers prime regions just outside the homology arm sequence and read into the cassette.

Primer name	Sequence (5′ to 3′)
300_KanB	CTGCAGCGAGGAGCCGTAAT
301_KanC	TGATTTTGATGACGAGCGTAAT
302_H2Az_A	TCCATGCTAGATTAGCACACAGTAA
303_H2Az_B	TATTCCAACACAGCAGTCAAATAAA
304_H2Az_C	GTTAGATTCTTTGATCAGGGCTACA
305_H2Az_D	CTTTTGTAGGTGTCCTTAATTTCCA
308_Rtt103_A	GTTTCTTTTTGATAGGTCCTTCCTC
309_Rtt103_B	AAAATCTTTCAATTTCTGGGGTAAC
310_Rtt103_C	TGAGTTAGATATAGAAGGCCACGTC
311_Rtt103_D	ATAAAAAGTTTAGAAAAGCGCGAAT
317_Pub1_A	ACTCGTTCTCTTTTCATCATTTTGT
833_Pub1_B	TCGGTGATAGCTTTGTCTAGGTTAC
834_Pub1_C	GGTTTACCTCCTCAAGTAAATCCTC
835_Pub1_D	AAAGAAAGCCTTCAGAAAATACGTT
326_Gbp2_A	TATCCTGAAACGACCACTTTTTATC
830_Gbp2_B	TCTCTTTCAACAATTGGACCTAAAG
831_Gbp2_C	CCTTGAAGATACCAGAGGTACTGAA
832_Gbp2_D	ATAAAGACAATAGCACAACCCAGAG
314_Npl3_A	GGCTTATTGATTACAATTGCTTGTT
836_Npl3_B	GGCAATTTAGAGTAAACAACTTCCA
837_Npl3_C	TACGATAGTCCTAGAGGTGGTTACG
838_Npl3_D	AAGGATGTTAAATGTTATCATGGGA
323_Tho2_A	ATTTACATGTTCTGAATGAGAAGGC
839_Tho2_B	GGGAGAATTTTCATCGTTTTTATTT
840_Tho2_C	CGAATAAAAGATTCAAGAAGGATGA
841_Tho2_D	CGATAAAAGAAAGAACGGTTTGTTA

5.4 **RT-PCR validation for deletion strain SMIT**

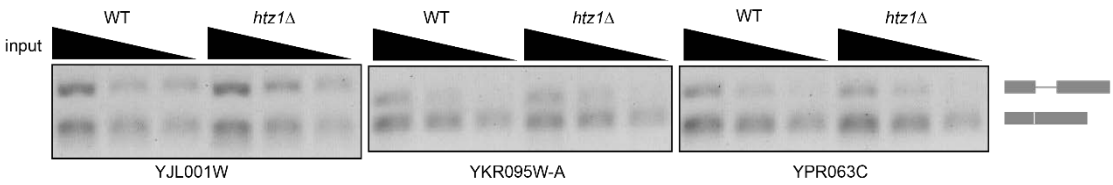

Figure 5.6 RT-PCR detects no change in relative levels of spliced vs unspliced species between WT and *htz1Δ*

1% TBE-agarose gels show RT-PCR products for three genes with high levels of H2A.Z. Upper band is unspliced product and lower band is spliced product. cDNA input was diluted 3x and 9x for triplicate PCR as indicated by black triangle above gels.

5.5 RT-PCR validation for Nab2 Anchor-Away SMIT

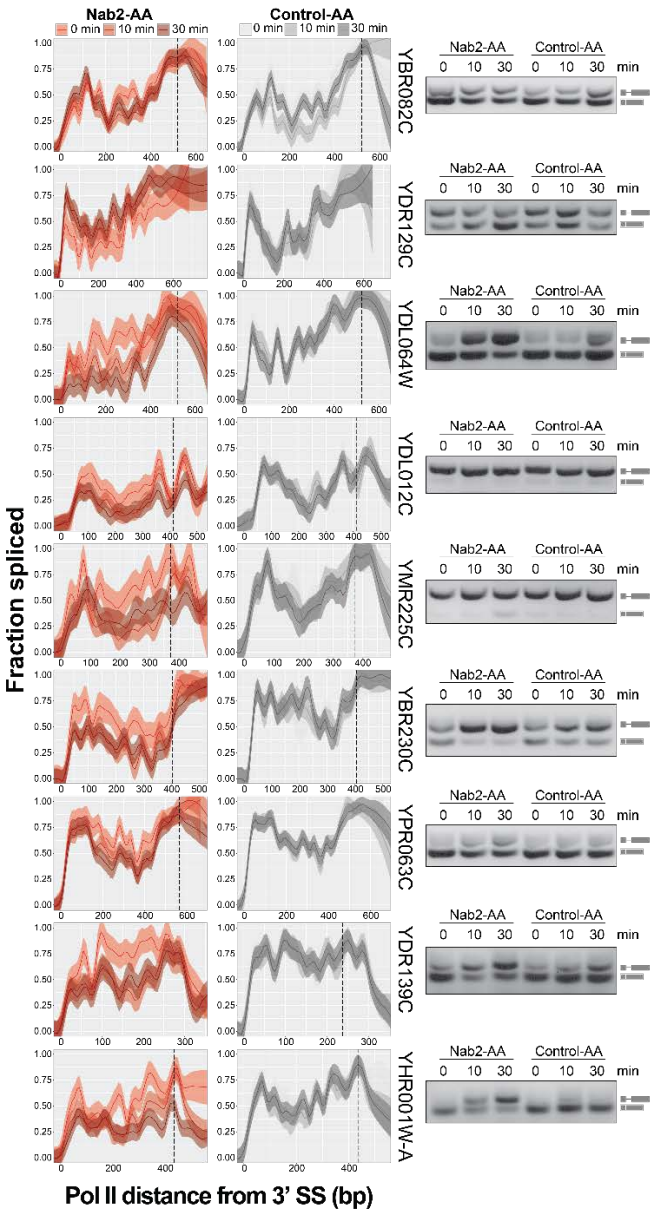

Figure 5.7 RT-PCR validates SMIT results for Nab2-AA
SMIT curves for validated genes are shown on the left with Nab2-AA time points overlaid in red and Control-AA time points overlaid in grey. Legend is above the top curve. Corresponding RT-PCR products on 1% agarose gels (right) show levels of spliced and unspliced product indicated by grey gene annotation on the right. Top gene is an example of no change in splicing, second gene from the top is an example where splicing improves during Nab2-AA depletion, and all others are examples of a reduction in splicing.

Table 5.6 RT-PCR validation primers for AA strains
Forward (Fwd) and reverse (Rev) intron-spanning primers are listed 5′ to 3′ for each gene indicated. Product size for spliced and unspliced are listed to the right.

Gene	Direction	Primer Sequences	Unspliced (bp)	Spliced (bp)
YBR230C	Fwd	CCGACGTAACCCTATTCCAA		
	Rev	GGTGGTGACCGTCTTCAGAG	223	126
YDL064W	Fwd	CACACACTGGCACCATTTTT		
	Rev	CATGGACCCATCAGCTTTCT	248	138
YPR063C	Fwd	TACTCCGCTGCTACCTCCTC		
	Rev	GATATGCTTGGTGTGGCAGA	210	124
YBR082C	Fwd	GGAAAATCACTATCGCCACAA		
	Rev	CCGGCTGAACATGAAGTAGG	245	150
YDR129C	Fwd	GGGGCGAAGTTTATAATGAAGA		
	Rev	TCAATGGCTCTGAACTTTTCAA	327	216
YDL012C	Fwd	ATTTACGCAACCGAAAAGGA		
	Rev	AGACGCCGTTTCATATCCTG	247	161
YDR139C	Fwd	GGGATTCCACCATCTCAACA		
	Rev	AAGTGGAGTTGCATTCCCTCTA	177	104
YMR225C	Fwd	TTCTTCCCTGAACCGTTTTG		
	Rev	GGTGGAATGGCAGCAAGTA	369	222
YHR001W-A	Fwd	CAAAAGCGCAAGTCGAATAA		
	Rev	CCGAAATGTAGACCAGTTTTTG	172	109

5.6 RT-PCR validation of Nab2-induced transcriptional readthrough

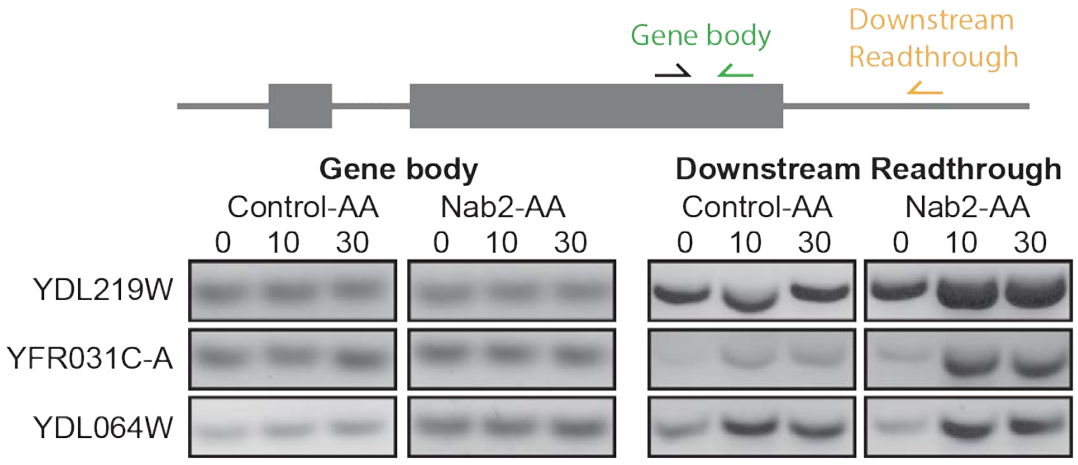

Figure 5.8 Validation of increased readthrough upon Nab depletion
RT-PCR was performed for three genes by reverse transcribing nascent RNA from Control-AA or Nab2-AA cells with random hexamers. PCR was performed with a common forward primer (black) and reverse primers in either the gene body (green) or the region downstream of the polyA site (orange). PCR products are visualized on agarose gels and show increased signal in the lanes corresponding to Downstream Readthrough for the Nab2-AA 10- and 30-minute samples compared to the Control-AA lanes.

5.7 Cloning Auxin-Inducible Degron strains

After sequence validation of the transformants, total protein content was extracted from the yeast by TCA precipitation after 0, 30, 60, and 90 minutes of auxin or ethanol treatment. Auxin is only soluble in ethanol, so ethanol alone serves as the control.

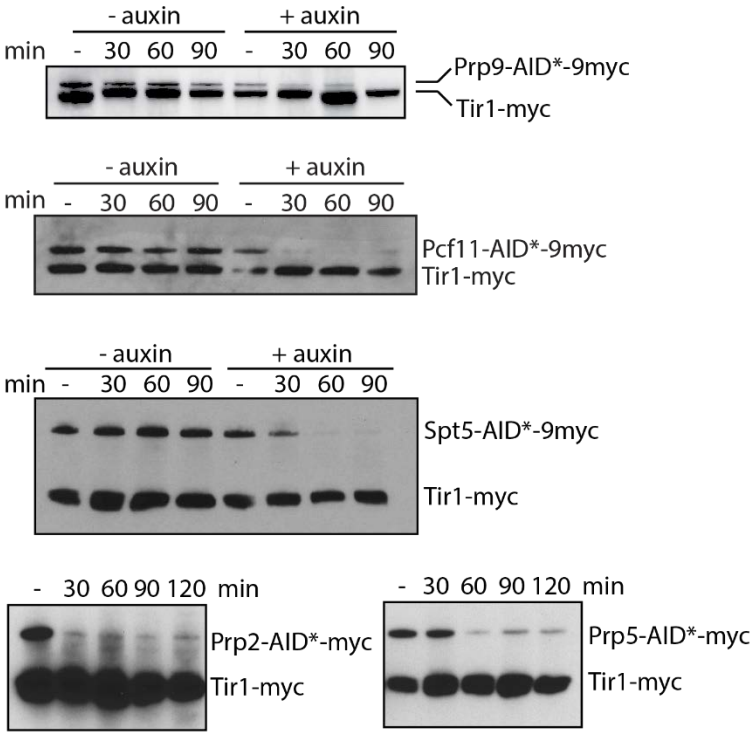

Figure 5.9 Time course of auxin-induced depletion
Cells were grown in YPAD and auxin (+ auxin) or ethanol (- auxin) was added for the indicated time period in minutes. Whole cell extract was run on polyacrylamide gels and transferred onto nitrocellulose membranes for western blotting with an anti-Myc antibody. Constitutively expressed Tir1-myc served as loading control.

Table 5.7 Plasmids for AID strains

Bacterial stocks harboring the following plasmids were generously given to us by Mark Hochstrasser. The ID number is for location in the -80C bacterial stock storage (stored in 75% LB saturated culture, 25% glycerol). Bacterial strain is TOP10F′ and plasmids are all derived from pFA6a with Amp resistance. (Morawska and Ulrich, 2013)

ID number	Name	Yeast drug resistance
638	pNAT-AID*-9Myc	Clonat
639	pHyg-AID*-6HA	HygromycinB
640	pHyg-AID*-6FLAG	HygromycinB
641	pHyg-AID*-GFP	HygromycinB
642	pHyg-AID*-9myc	HygromycinB
643	pKan-AID*-9myc	G418

Table 5.8 Primers used in cloning of AID tagged strains

Capital letters denote genomic sequences while lowercase letters base pair to the plasmid. These primers were used either to amplify linear products for transformation off the plasmids in the previous table or to sequence validate proper transformation. All primers were ordered with desalt purification. I did not end up using all the AID-tagged strains that I generated with these primers here.

Primer Name	Sequence (5′ to 3′)
968_pKan-AID-9myc-fwd	CGTACGCTGCAGGTCGAC
969_pKan-AID-9myc-rev	TCGATGAATTCGAGCTCG
972_ysh1_aid_fwd	GGGTGGAAAGCCTCTTAAATATTGGTGGTAATTTGG TCACACCGCTATGTcctaaagatccagccaaacctccgg
973_ysh1_aid_rev	GGTTTGGTATTACTTCTATAAAGTAGTCTACTTAGT ATGCGTAACTGTTTcagtatagcgaccagcattcacata
974_spt5_aid_fwd	ACCAAGGAAATAAGTCAAACTATGGTGGTAACAGT ACATGGGGAGGTCATcctaaagatccagccaaacctccgg
975_spt5_aid_rev	TTAAAGTCTTTTTTATTGATTTCTTCTTGGGTGATAT TGGTTCTCCTTTTGGTGAcagtatagcgaccagcattcacata
976_prp2_aid_fwd	CACAAATCTTTAAAGATTTAATTGACGATAAAACA AATAGGGGGAGGCGGcctaaagatccagccaaacctccgg

977_prp2_aid_rev	AGAATGGAGCCTGCGTTTCTAGCAATACACATACAC CTGTCAAAAAACCTcagtatagcgaccagcattcacata
978_prp5_aid_fwd	AAGAGGGGGTCGTAAAGGCTGCAAGCTTGTCTTTG AAGAGTACTAAATACcctaaagatccagccaaacctccgg
979_prp5_aid_rev	GAACTAACTACGAAAGTATATAGCACCACGAGTGA GTTAAATTCTAAAAAcagtatagcgaccagcattcacata
980_prp16_aid_fwd	ACGGCAAAGAAAATTCAATGAAACCTTTCAAAAGA AGGAAGCCTTTTTTTcctaaagatccagccaaacctccgg
981_prp16_aid_rev	TATAATAACATATATGAATATTTTGCCTATTAGCAC GCTCTTCCCATAAAcagtatagcgaccagcattcacata
982_prp22_aid_fwd	GACTAAGCTCAATAAGGCAGTCAAGGGAAAGGGCA TTAGGTATCAAGAGGcctaaagatccagccaaacctccgg
983_prp22_aid_rev	ATATAGGTCTATAAAACTCGATAATTATAATGCATA AAAAGCTAACAATGcagtatagcgaccagcattcacata
984_nab2_aid_fwd	CTCCTCCGCAAACCAGTTTTACGCACCAAGAACAA GATACGGAAATGAACcctaaagatccagccaaacctccgg
985_nab2_aid_rev	ATCAAAAGGGTCACAGGAACATGAATTTCGTTCCG TGATTTTAATAGTAAcagtatagcgaccagcattcacata
986_npl3_aid_fwd	GAGATGCATACAGAACCAGAGATGCTCCACGTGAA AGATCACCAACCAGGcctaaagatccagccaaacctccgg
987_npl3_aid_rev	ACAATTCATATCTTTTGTTAATTTCTCCTTTTTTTTTC TCAACTATATAAATGGCcagtatagcgaccagcattcacata
988_rad53_aid_fwd	GGGCAAAATTGGACCAAACCTCAAAAGGCCCCGAG AATTTGCAATTTTCGcctaaagatccagccaaacctccgg
989_rad53_aid_rev	ATCTTCTCTCTTAAAAAGGGGCAGCATTTTCTATGG GTATTTGTCCTTGGcagtatagcgaccagcattcacata
990_rtt103_aid_fwd	CCGGAGGGGTTTCTTCTAGTATACAAGACTTGTTAA GTAAGCTTGCAAATcctaaagatccagccaaacctccgg

991_rtt103_aid_rev	ATATATTTGTATAAGTTATCTCCTTGTTTTCTTTTTA CTCAACCATCATAcagtatagcgaccagcattcacata
992_pcf11_check	ATCTGGTGAATGGGTTTGGA
993_ysh1_check	GGGTGAAGCAAATCTCAAGG
994_prp2_check	CATAAGGAAAAGGCGCAGAG
995_nab2_check	CACCCAATTCAAACGTTCCT
996_prp5_check	ATTGATCGAAGGCCAAGATG
997_prp16_check	AATGATCAGGAGGCCACAAC
998_prp22_check	TCGAAGTAGCGCCTCATTTT
999_rad53_check	GCAAACAGCCGAAGAAAAAG
1000_rtt103_check	CGCTAATGACATACCGGAAAA
1001_spt5_check	TACAAGAGATGGCGGAGCTT
1002_kanR	tcgatagattgtcgcacctg
1003_prp22_check2	GAACGCTGCTAAGCGAGACT
1004_npl3_check2	TGACAATCCTCCACCAATCA
1005_rtt103_check2	TATGAAGTGGGGGATGGAGA
1006_pcf11_check2	AGCATTTGGACTGGCATTTC
1007_nab2_check2	ACCTCCGGTTGAAAAGTCCT
1008_pcf11_fwd2	CAAATCTAATAGTGGCAAGGTCGGTTTGGATGACTT AAAGAAATTGGTCACAAAAcctaaagatccagccaaacctccgg
1009_pcf11_rev2	TAATATAATATATAGTTATTAAATTTAAATGTATAT ATGCAGTTCTGCTCcagtatagcgaccagcattcacata
1010_nab2_fwd	ACCTCCGGTTGAAAAGTCCT
1011_nab2_rev	TCCAATTATGCGATGCATGT
1014_prp3_aid_fwd	GTACGCTGGGTCAGTTTGATTCAGAGCATTTTTATT CACCTGTTCAAACGcctaaagatccagccaaacctccgg
1015_prp3_aid_rev	GTAAAATAATATTTAATATGAAACAAAGCGTATCA TTTTGTAGACACCGATAcagtatagcgaccagcattcacata
1016_prp9_aid_fwd	GTAATGAGTAAGAAGGTCTACGATGAACTTAAGAA GCAAGGTTTGGTGcctaaagatccagccaaacctccgg
1017_prp9_aid_rev	CATACAACTGCTATCTATCAAACAAATATACATATC ACAGAGAGATTcagtatagcgaccagcattcacata
1018_kan_rev_check	gcgccttttagctagtgga
1019_prp9_fwd_check	CGACGCACTTTTGAAAGACA
1020_prp3_fwd_check	AACGTTGAAAAGCCGACAGT

5.8 Reverse transcription for long read sequencing

Processivity of the reverse transcriptase (RT) enzyme is a major limitation for both short- and long-read RNA sequencing strategies, but is especially noticeable among long reads. A striking feature of long read sequencing data is the abundance of reads that do not align to a transcription start site (TSS) but instead originate within a gene body. These partial reads are filtered out for analyses presented here, although the presence of the 3′ end adapter sequence validates a true nascent RNA end and these reads could be used for Pol II density and other such analyses. Whether these reads represent RNA degradation products or incomplete RT is difficult to distinguish, however synthetic spike in controls suggest these reads are the result of RT falloff. Spike ins are *in vitro* transcribed and pooled with the sample near the end of the protocol where they are unlikely to experience breakage from the previous cell lysis and stringent purification steps. Additionally, we know the amount of incomplete RT is substantial for the enzymes commonly used in the field (Zhao et al., 2018), and while we cannot quantify what portion of partial reads can be attributed to this, we know they exist as a large fraction. Recently, Morgan Vanderwall began adapting use of the Marathon Reverse Transcriptase in her rotation for our sequencing library protocols because this RT has significantly reduced falloff (Zhao et al., 2018), however further optimization is still required to ensure the strand-switching activity is robust enough for sequencing. Marathon RT in its current form would be ideal for targeted long read sequencing experiments, however, overcoming the limitations of MMLV-based enzymes such as SuperScript III or IV and SMARTscribe. Alternatively, direct Nanopore sequencing of RNA eliminates the need for the RT step (synthesizing a single cDNA strand

can promote sequencing by eliminating secondary structure, but is not strictly required), however fewer reads are collected (Garalde et al., 2018).

5.9 Spt5-AID long read sequencing of nRNA

Leonard Schärfen prepared nascent RNA from Spt5-AID samples during his rotation, sequenced on the minION, and analyzed the data with my assistance. He also wrote a script used for analyzing the coverage downstream of the PAS (Figure 5.10). We chose to target Spt5 for degradation because of a published report that found a readthrough phenotype associated with Spt5 in 4tU-seq experiments (Figure 5.10) (Baejen et al., 2017). A preliminary sequencing dataset that we generated did not agree with this finding, although this sample produced fewer reads than usual on the minION. Our reports are in agreement with their data that Pcf11 depletion induces readthrough. The Spt5 phenotype is less pronounced than that of Pcf11 or Ysh1, therefore it is possible that the different techniques, time points, and sequencing methods may produce slightly varied results.

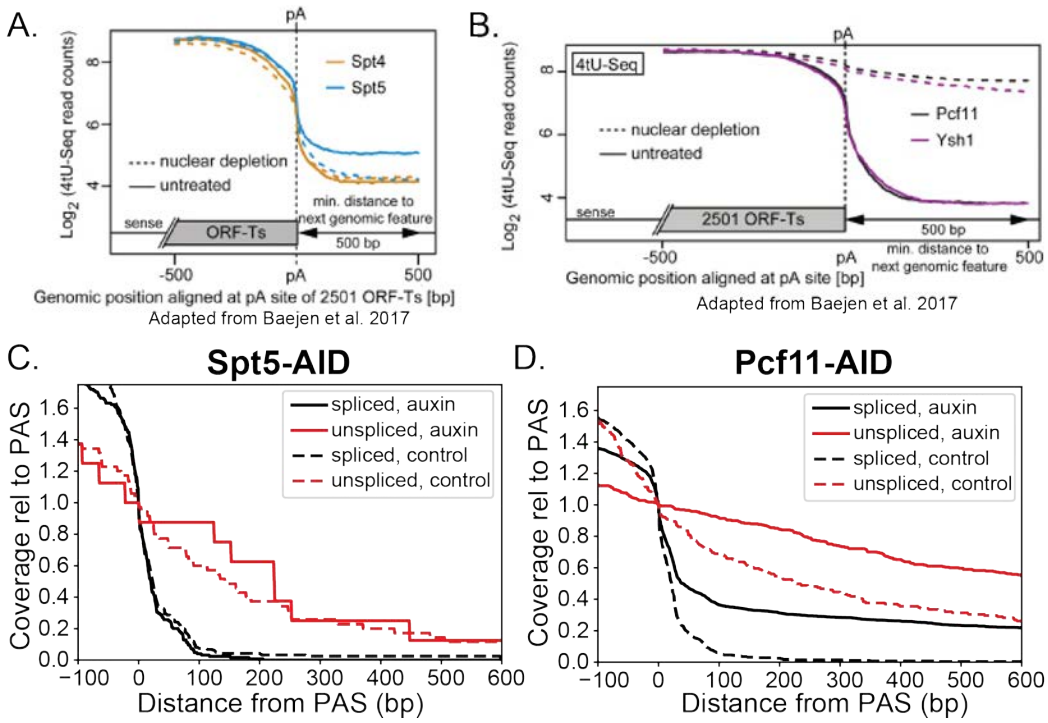

Figure 5.10 Spt5-AID long read sequencing data
Nuclear depletion with Anchor-Away of Spt5, Pcf11, or Ysh1 generates increased coverage of signal downstream of the PAS as measured by short read sequencing of new RNA (4tU-labeled RNA). Our long read nascent RNA sequencing disagrees with this conclusion for Spt5 but is in agreement for Pcf11.

5.10 **Prp2-AID long read sequencing of nRNA**

Splicing is only mildly affected by depletion of Prp2-AID (Figure 5.11), despite efficient degradation of the construct (Figure 5.9). Therefore, we continued tagging other splicing factors (such as Prp9 and Prp3) to achieve more robust inhibition of splicing.

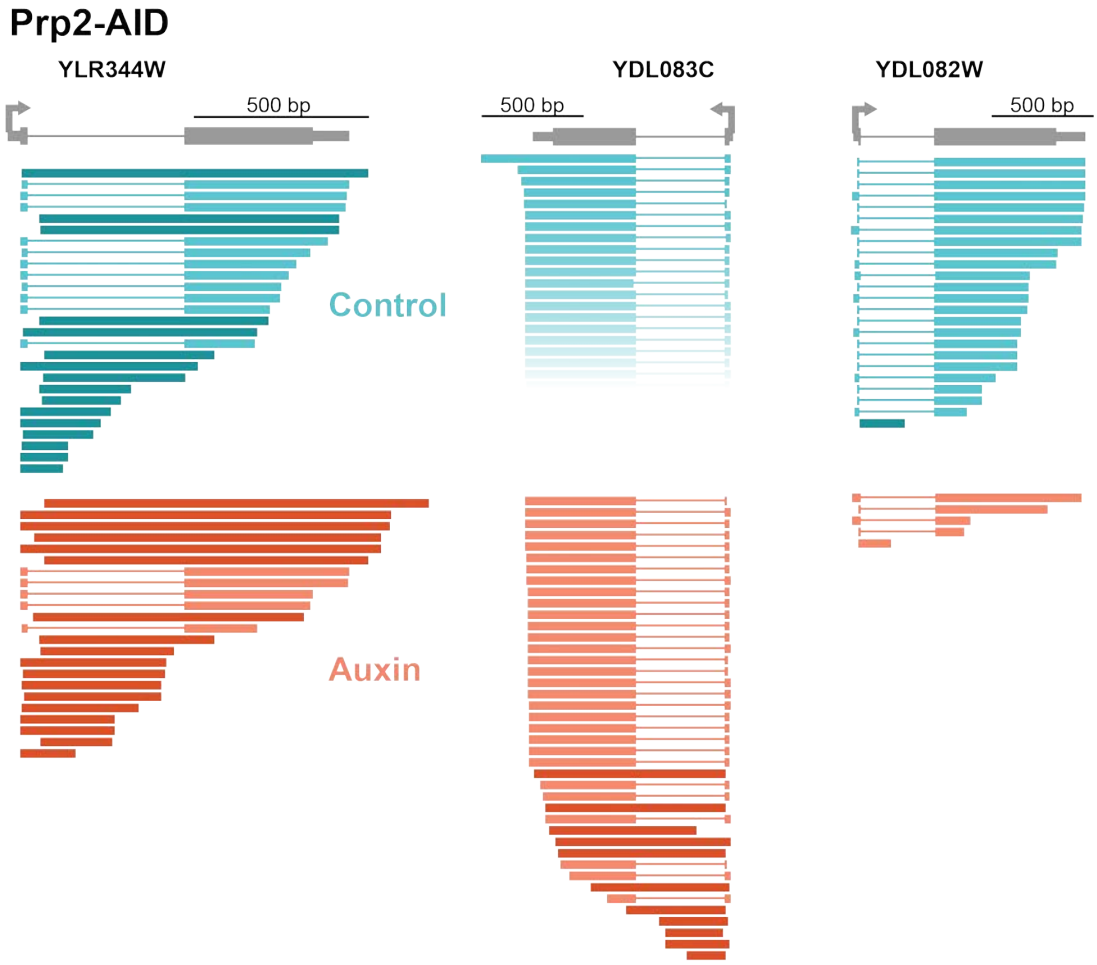

Figure 5.11 Prp2-AID has minimal impact on fraction spliced
Nanopore sequencing reads from control (teal) or auxin-treated (orange) samples of Prp2-AID are aligned to 3 different genes. Thick lines in darker colors are unspliced reads, while lighter colors are spliced reads. Arrows indicate direction of transcription. YDL083C control reads are faded out because there were too many to display.

5.11 **Read strand issue with Nanopore sequencing**

Correct identification of the 3′ end adapter should unambiguously determine which strand the transcript arose from. Strangely, a non-trivial fraction of reads are classified on the unexpected strand (e.g. aligning to the opposite strand of an ORF) (blue; Figure 5.12). I believe this to be a computational aberration because these reads are spliced at sites which do not harbor consensus splice sites on the reverse strand. Consistent with a computational issue, aberrant reads also arise from the *in vitro* transcribed spike in (~7% of reads), eliminating the possibility that these are biologically relevant. For this reason, I avoided analyses that quantified 3′ end (Pol II) density, as this example indicates such an analysis would be unreliable.

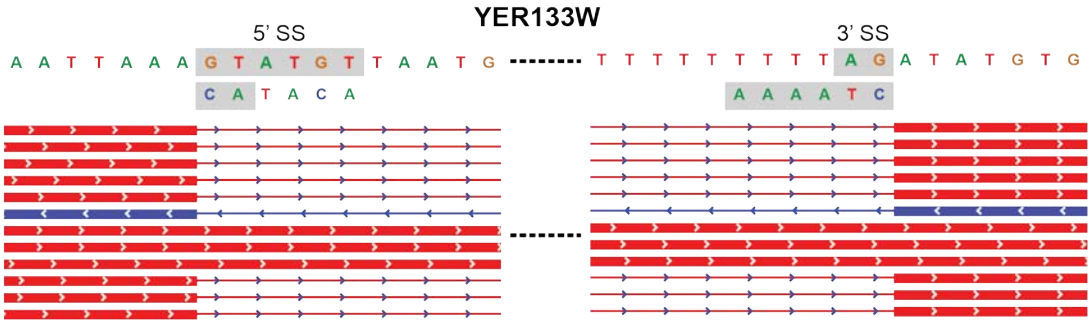

Figure 5.12 Example of reads with incorrect strandedness
Nanopore reads are aligned to genomic sequence (5′ to 3′; top). Arrows indicate read strand, either forward (red) or reverse (blue).

5.12 *S. cerevisiae* strains

Table 5.9 Strains of budding yeast
All yeast stocks are stored in the -80C common tower in 75% YPAD (overnight culture) and 25% glycerol.

ID number	Name	Genotype	Background
	NPL3Δ	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 npl3Δ::KanMX	BY4741
	HTZ1Δ	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 htz1Δ::KanMX	BY4741
	GBP2Δ	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 gbp2Δ::KanMX	BY4741
	THO2Δ	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 tho2Δ::KanMX	BY4741
	RTT103Δ	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 rtt103Δ::KanMX	BY4741
	PUB1Δ	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 pub1Δ::KanMX	BY4741
652	Anchor-Away Control	MATα tor1-1 fpr1::loxP-LEU2-loxP RPL13-2xFKBP12::loxP-TRP1-loxP	W303
653	Nab2-AA	MATα tor1-1 fpr1::loxP-LEU2-loxP RPL13-2xFKBP12::loxP-TRP1-loxP NAB2-FRB::HIS	W303
708	AID background	MATa his3-Δ200 leu2-3,112 lys2-801 trp1-1 URA3::TIR1-9Myc	DF5
709	AID control	MATa his3-Δ200 leu2-3,112 lys2-801 trp1-1 URA3::TIR1-9Myc RFA1-AID*-9Myc-HphMX	DF5
710	PRP2-AID*-9myc	MATa PRP2-AID*-9myc-KanMX URA3::TIR1-9myc his3-Δ200 leu2-3,112 lys2-801 trp1-1	DF5
711	PRP5-AID*-9myc	MATa PRP5-AID*-9myc-KanMX URA3::TIR1-9myc his3-Δ200 leu2-3,112 lys2-801 trp1-1	DF5
712	PRP16-AID*-9myc	MATa PRP16-AID*-9myc-KanMX URA3::TIR1-9myc his3-Δ200 leu2-3,112 lys2-801 trp1-1	DF5
713	PRP22-AID*-9myc	MATa PRP22-AID*-9myc-KanMX URA3::TIR1-9myc his3-Δ200 leu2-3,112 lys2-801 trp1-1	DF5

714	YSH1-AID*-9myc	MATa YSH1-AID*-9myc-KanMX URA3::TIR1-9myc his3-Δ200 leu2-3,112 lys2-801 trp1-1	DF5
715	NPL3-AID*-9myc	MATa NPL3-AID*-9myc-KanMX URA3::TIR1-9myc his3-Δ200 leu2-3,112 lys2-801 trp1-1	DF5
716	RTT103-AID*-9myc	MATa RTT103-AID*-9myc-KanMX URA3::TIR1-9myc his3-Δ200 leu2-3,112 lys2-801 trp1-1	DF5
717	SPT5-AID*-9myc	MATa SPT5-AID*-9myc-KanMX URA3::TIR1-9myc his3-Δ200 leu2-3,112 lys2-801 trp1-1	DF5
718	RAD53-AID*-9myc	MATa RAD53-AID*-9myc-KanMX URA3::TIR1-9myc his3-Δ200 leu2-3,112 lys2-801 trp1-1	DF5
719	PCF11-AID*-9myc	MATa PCF11-AID*-9myc-KanMX URA3::TIR1-9myc his3-Δ200 leu2-3,112 lys2-801 trp1-1	DF5
720	PRP3-AID*-9myc	MATa PRP3-AID*-9myc-KanMX URA3::TIR1-9myc his3-Δ200 leu2-3,112 lys2-801 trp1-1	DF5
721	PRP9-AID*-9myc	MATa PRP9-AID*-9myc-KanMX URA3::TIR1-9myc his3-Δ200 leu2-3,112 lys2-801 trp1-1	DF5

6. References

Ahn, S.H., Kim, M., and Buratowski, S. (2004). Phosphorylation of Serine 2 within the RNA Polymerase II C-Terminal Domain Couples Transcription and 3′ End Processing. Mol. Cell *13*, 67–76.

Aibara, S., Gordon, J.M.B., Riesterer, A.S., McLaughlin, S.H., and Stewart, M. (2017). Structural basis for the dimerization of Nab2 generated by RNA binding provides insight into its contribution to both poly(A) tail length determination and transcript compaction in Saccharomyces cerevisiae. Nucleic Acids Res. *45*, 1529–1538.

Alexander, R.D., Barrass, J.D., Dichtl, B., Kos, M., Obtulowicz, T., Robert, M.C., Koper, M., Karkusiewicz, I., Mariconti, L., Tollervey, D., et al. (2010). RiboSys, a high-resolution, quantitative approach to measure the in vivo kinetics of pre-mRNA splicing and 3′-end processing in Saccharomyces cerevisiae. RNA *16*, 2570–2580.

Alpert, T., Reimer, K., Straube, K., and Neugebauer, K. Long read sequencing of nascent RNA from budding and fission yeasts. Methods Molcular Biol. (accepted)

Ameur, A., Zaghlool, A., Halvardson, J., Wetterbom, A., Gyllensten, U., Cavelier, L., and Feuk, L. (2011). Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. Nat. Struct. Mol. Biol. *18*, 1435–1440.

Baejen, C., Torkler, P., Gressel, S., Essig, K., Söding, J., and Cramer, P. (2014). Transcriptome Maps of mRNP Biogenesis Factors Define Pre-mRNA Recognition. Mol. Cell *55*, 745–757.

Baejen, C., Andreani, J., Torkler, P., Battaglia, S., Schwalb, B., Lidschreiber, M., Maier, K.C., Boltendahl, A., Rus, P., Esslinger, S., et al. (2017). Genome-wide Analysis of RNA Polymerase II Termination at Protein-Coding Genes. Mol. Cell *66*, 38-49.e6.

Barillà, D., Lee, B.A., and Proudfoot, N.J. (2001). Cleavage/polyadenylation factor IA associates with the carboxyl-terminal domain of RNA polymerase II in Saccharomyces cerevisae. Proc. Natl. Acad. Sci. U. S. A. *98*, 445–450.

Barrass, J.D., Reid, J.E.A., Huang, Y., Hector, R.D., Sanguinetti, G., Beggs, J.D., and Granneman, S. (2015). Transcriptome-wide RNA processing kinetics revealed using extremely short 4tU labeling. Genome Biol. *16*.

Baserga, S.J., and Steitz, J. a (1993). The Diverse World of Small Ribonucleoproteins. RNA World 359–381.

Batisse, J., Batisse, C., Budd, A., Böttcher, B., and Hurt, E. (2009). Purification of nuclear poly(A)-binding protein Nab2 reveals association with the yeast transcriptome and a messenger ribonucleoprotein core structure. J. Biol. Chem. *284*, 34911–34917.

Beckmann, J.S., and Trifonov, E.N. (1991). Splice junctions follow a 205-base ladder. Proc. Natl. Acad. Sci. U. S. A. *88*, 2380–2383.

Berget, S.M. (1995). Exon recognition in vertebrate splicing. J. Biol. Chem. *270*, 2411–2414.

Beyer, A.L., and Osheim, Y.N. (1988). Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. Genes Dev. *2*, 754–765.

Bird, G., Zorio, D.A.R., and Bentley, D.L. (2004). RNA Polymerase II Carboxy-Terminal Domain Phosphorylation Is Required for Cotranscriptional Pre-mRNA Splicing and 3'-End

Formation. Mol. Cell. Biol. *24*, 8963–8969.

Braberg, H., Jin, H., Moehle, E.A., Chan, Y.A., Wang, S., Shales, M., Benschop, J.J., Morris, J.H., Qiu, C., Hu, F., et al. (2013). From structure to systems: High-resolution, quantitative genetic analysis of RNA polymerase II. Cell *154*, 775–788.

Braunschweig, U., Gueroussov, S., Plocik, A.M., Graveley, B.R., and Blencowe, B.J. (2013). Dynamic integration of splicing within gene regulatory pathways. Cell *152*, 1252–1269.

Brinster, R.L., Allen, J.M., Behringer, R.R., Gelinas, R.E., and Palmiter, R.D. (1988). Introns increase transcriptional efficiency in transgenic mice. Proc. Natl. Acad. Sci. U. S. A. *85*, 836–840.

Buratti, E., and Baralle, F.E. (2004). Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process. Mol. Cell. Biol. *24*, 10505–10514.

Carrillo Oesterreich, F., Preibisch, S., and Neugebauer, K.M. (2010). Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. Mol. Cell *40*, 571–581.

Carrillo Oesterreich, F., Herzel, L., Straube, K., Hujer, K., Howard, J., and Neugebauer, K.M. (2016). Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. Cell *165*, 372–381.

Chan, S., Choi, E.A., and Shi, Y. (2011). Pre-mRNA 3′-end processing complex assembly and function. Wiley Interdiscip. Rev. RNA *2*, 321–335.

Chathoth, K.T., Barrass, J.D., Webb, S., and Beggs, J.D. (2014). A Splicing-Dependent Transcriptional Checkpoint Associated with Prespliceosome Formation. Mol. Cell *53*, 779–790.

Churchman, L.S., and Weissman, J.S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature *469*, 368–373.

Connelly, S., and Manley, J.L. (1988). A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. Genes Dev. *2*, 440–452.

Cooke, C., Hans, H., and Alwine, J.C. (1999). Utilization of Splicing Elements and Polyadenylation Signal Elements in the Coupling of Polyadenylation and Last-Intron Removal. Mol. Cell. Biol. *19*, 4971–4979.

Corden, J.L., Cadena, D.L., Ahearn, J.M., and Dahmus, M.E. (1985). A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II. Proc. Natl. Acad. Sci. U. S. A. *82*, 7934–7938.

Coulon, A., Ferguson, M.L., De Turris, V., Palangat, M., Chow, C.C., and Larson, D.R. (2014). Kinetic competition during the transcription cycle results in stochastic RNA processing. Elife.

Cramer, P., Pesce, C.G., Baralle, F.E., and Kornblihtt, A.R. (1997). Functional association between promoter structure and transcript alternative splicing. Proc. Natl. Acad. Sci. U. S. A. *94*, 11456–11460.

Custódio, N., and Carmo-Fonseca, M. (2016). Co-transcriptional splicing and the CTD code. Crit. Rev. Biochem. Mol. Biol. *51*, 395–411.

Davidson, L., and West, S. (2013). Splicing-coupled 3′ end formation requires a terminal splice acceptor site, but not intron excision. Nucleic Acids Res. *41*, 7101–7114.

Deutsch, M., and Long, M. (1999). Intron-exon structures of eukaryotic model organisms. Nucleic

Acids Res. *27*, 3219–3228.

Dominski, Z. (2010). The hunt for the 3′ endonuclease. Wiley Interdiscip. Rev. RNA *1*, 325–340.

Drexler, H.L., Choquet, K., and Churchman, L.S. (2019). Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. Mol. Cell *77*, 985–998.

Dye, M.J., and Proudfoot, N.J. (1999). Terminal exon definition occurs cotranscriptionally and promotes termination of RNA polymerase II. Mol. Cell *3*, 371–378.

Enge, M., Arda, H.E., Mignardi, M., Beausang, J., Bottino, R., Kim, S.K., and Quake, S.R. (2017). Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. Cell *171*, 321-330.e14.

Eperon, L.P., Graham, I.R., Griffiths, A.D., and Eperon, I.C. (1988). Effects of RNA secondary structure on alternative splicing of Pre-mRNA: Is folding limited to a region behind the transcribing RNA polymerase? Cell *54*, 393–401.

Eser, P., Wachutka, L., Maier, K.C., Demel, C., Boroni, M., Iyer, S., Cramer, P., and Gagneur, J. (2016). Determinants of RNA metabolism in the Schizosaccharomyces pombe genome. Mol. Syst. Biol. *12*, 857.

Fong, N., and Bentley, D.L. (2001). Capping, splicing, and 3′ processing are independently stimulated by RNA polymerase II: Different functions for different segments of the CTD. Genes Dev. *15*, 1783–1795.

Fong, N., Kim, H., Zhou, Y., Ji, X., Qiu, J., Saldi, T., Diener, K., Jones, K., Fu, X.D., and Bentley, D.L. (2014). Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. Genes Dev. *28*, 2663–2676.

Garalde, D.R., Snell, E.A., Jachimowicz, D., Sipos, B., Lloyd, J.H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. Nat. Methods *15*, 201–206.

Gonatopoulos-Pournatzis, T., and Cowling, V.H. (2014). Cap-binding complex (CBC). Biochem. J. *457*, 231–242.

Görnemann, J., Kotovic, K.M., Hujer, K., and Neugebauer, K.M. (2005). Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex. Mol. Cell *19*, 53–63.

Green, D.M., Marfatia, K.A., Crafton, E.B., Zhang, X., Cheng, X., and Corbett, A.H. (2002). Nab2p is required for poly(A) RNA export in Saccharomyces cerevisiae and is regulated by arginine methylation via Hmt1p. J. Biol. Chem. *277*, 7752–7760.

Grosso, A.R., Leite, A.P., Carvalho, S., Matos, M.R., Martins, F.B., Vítor, A.C., Desterro, J.M.P., Carmo-Fonseca, M., and de Almeida, S.F. (2015). Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. Elife.

Gunderson, S.I., Polycarpou-Schwarz, M., and Mattaj, I.W. (1998). U1 snRNP inhibits pre-mRNA polyadenylation through a direct interaction between U1 70K and poly(A) polymerase. Mol. Cell *1*, 255–264.

Harlen, K.M., Trotta, K.L., Smith, E.E., Mosaheb, M.M., Fuchs, S.M., and Churchman, L.S. (2016). Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue. Cell Rep. *15*, 2147–2158.

Haruki, H., Nishikawa, J., and Laemmli, U.K. (2008). The Anchor-Away Technique: Rapid, Conditional Establishment of Yeast Mutant Phenotypes. Mol. Cell *31*, 925–932.

Hawkin, J.D. (1988). A survey on intron and exon lengths. Nucleic Acids Res. *16*, 9893–9908.

Hérissant, L., Moehle, E.A., Bertaccini, D., Van Dorsselaer, A., Schaeffer-Reiss, C., Guthrie, C., and Dargemont, C. (2014). H2B ubiquitylation modulates spliceosome assembly and function in budding yeast. Biol. Cell *106*, 126–138.

Herzel, L., Ottoz, D.S.M., Alpert, T., and Neugebauer, K.M. (2017). Splicing and transcription touch base: Co-transcriptional spliceosome assembly and function. Nat. Rev. Mol. Cell Biol. *18*, 637–650.

Herzel, L., Straube, K., and Neugebauer, K.M. (2018). Long-read sequencing of nascent RNA reveals coupling among RNA processing events. Genome Res. *28*, 1008–1019.

Huff, J.T., Zilberman, D., and Roy, S.W. (2016). Mechanism for DNA transposons to generate introns on genomic scales. Nature *538*, 533–536.

Huranová, M., Ivani, I., Benda, A., Poser, I., Brody, Y., Hof, M., Shav-Tal, Y., Neugebauer, K.M., and Staněk, D. (2010). The differential interaction of snRNPs with pre-mRNA reveals splicing kinetics in living cells. J. Cell Biol. *191*, 75–86.

Izaurralde, E., Lewis, J., McGuigan, C., Jankowska, M., Darzynkiewicz, E., and Mattaj, I.W. (1994). A nuclear cap binding protein complex involved in pre-mRNA splicing. Cell *78*, 657–668.

Izaurralde, E., Lewis, J., Gamberi, C., Jarmolowski, A., McGuigan, C., and Mattaj, I.W. (1995). A cap-binding protein complex mediating U snRNA export. Nature *376*, 709–712.

Jonkers, I., and Lis, J.T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. Nat. Rev. Mol. Cell Biol. *16*, 167–177.

Jonkers, I., Kwak, H., and Lis, J.T. (2014). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. Elife.

Kaida, D. (2016). The reciprocal regulation between splicing and 3′-end processing. Wiley Interdiscip. Rev. RNA *7*, 499–511.

Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L., and Dreyfuss, G. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature *468*, 664–668.

Kfir, N., Lev-Maor, G., Glaich, O., Alajem, A., Datta, A., Sze, S.K., Meshorer, E., and Ast, G. (2015). SF3B1 Association with Chromatin Determines Splicing Outcomes. Cell Rep. *11*, 618–629.

Khodor, Y.L., Rodriguez, J., Abruzzi, K.C., Tang, C.H.A., Marr, M.T., and Rosbash, M. (2011). Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. Genes Dev. *25*, 2502–2512.

Khodor, Y.L., Menet, J.S., Tolan, M., and Rosbash, M. (2012). Cotranscriptional splicing efficiency differs dramatically between Drosophila and mouse. RNA *18*, 2174–2186.

Kielkopf, C.L., Rodionova, N.A., Green, M.R., and Burley, S.K. (2001). A novel peptide recognition mode revealed by the X-ray structure of a core U2AF35/U2AF65 heterodimer. Cell *106*, 595–605.

Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907–915.

Kim, H., Erickson, B., Luo, W., Seward, D., Graber, J.H., Pollock, D.D., Megee, P.C., and Bentley, D.L. (2010). Gene-specific RNA polymerase II phosphorylation and the CTD code. Nat. Struct. Mol. Biol. *17*, 1279–1286.

Kim, M., Ahn, S.H., Krogan, N.J., Greenblatt, J.F., and Buratowski, S. (2004). Transitions in RNA polymerase II elongation complexes at the 3′ ends of genes. EMBO J. *23*, 354–364.

Konarska, M.M., Padgett, R.A., and Sharp, P.A. (1984). Recognition of cap structure in splicing in vitro of mRNA precursors. Cell *38*, 731–736.

Kress, T.L., Krogan, N.J., and Guthrie, C. (2008). A Single SR-like Protein, Npl3, Promotes Pre-mRNA Splicing in Budding Yeast. Mol. Cell *32*, 727–734.

Kwak, H., and Lis, J.T. (2013). Control of Transcriptional Elongation. Annu. Rev. Genet. *47*, 483–508.

Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science (80-.). *339*, 950–953.

Kyburz, A., Friedlein, A., Langen, H., and Keller, W. (2006). Direct Interactions between Subunits of CPSF and the U2 snRNP Contribute to the Coupling of Pre-mRNA 3′ End Processing and Splicing. Mol. Cell *23*, 195–205.

Lacadie, S.A., and Rosbash, M. (2005). Cotranscriptional spliceosome assembly dynamics and the role of U1 snRNA:5′ss base pairing in yeast. Mol. Cell *19*, 65–75.

Lacadie, S.A., Tardiff, D.F., Kadener, S., and Rosbash, M. (2006). In vivo commitment to yeast cotranscriptional splicing is sensitive to transcription elongation mutants. Genes Dev. *20*, 2055–2066.

Lai, D., Proctor, J.R., and Meyer, I.M. (2013). On the importance of cotranscriptional RNA structure formation. RNA *19*, 1461–1473.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

Li, Y., Chen, Z.-Y., Wang, W., Baker, C.C., and Krug, R.M. (2001). The 3′-end-processing factor CPSF is required for the splicing of single-intron pre-mRNAS in vivo. RNA *7*, 920–931.

Licatalosi, D.D., Geiger, G., Minet, M., Schroeder, S., Cilli, K., McNeil, J.B., and Bentley, D.L. (2002). Functional interaction of yeast pre-mRNA 3' end processing factors with RNA Polymerase II. Mol. Cell *9*, 1101–1111.

Listerman, I., Sapra, A.K., and Neugebauer, K.M. (2006). Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. Nat. Struct. Mol. Biol. *13*, 815–822.

Liu, S.R., Hu, C.G., and Zhang, J.Z. (2016). Regulatory effects of cotranscriptional RNA structure formation and transitions. Wiley Interdiscip. Rev. RNA *7*, 562–574.

Mandel, C.R., Bai, Y., and Tong, L. (2008). Protein factors in pre-mRNA 3′-end processing. Cell. Mol. Life Sci. *65*, 1099–1122.

Manning, K.S., and Cooper, T.A. (2017). The roles of RNA processing in translating genotype to phenotype. Nat. Rev. Mol. Cell Biol. *18*, 102–114.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal.

Martin, R.M., Rino, J., Carvalho, C., Kirchhausen, T., and Carmo-Fonseca, M. (2013). Live-Cell Visualization of Pre-mRNA Splicing with Single-Molecule Sensitivity. Cell Rep. *4*, 1144–1155.

Martinez-Rucobo, F.W., Kohler, R., van de Waterbeemd, M., Heck, A.J.R., Hemann, M., Herzog, F., Stark, H., and Cramer, P. (2015). Molecular Basis of Transcription-Coupled Pre-mRNA Capping. Mol. Cell *58*, 1079–1089.

Martins, S.B., Rino, J., Carvalho, T., Carvalho, C., Yoshida, M., Klose, J.M., De Almeida, S.F., and Carmo-Fonseca, M. (2010). Spliceosome assembly is coupled to RNA polymerase II dynamics at the 3′ end of human genes. Nat. Struct. Mol. Biol. *18*, 1115–1123.

Mason, P.B., and Struhl, K. (2005). Distinction and relationship between elongation rate and processivity of RNA polymerase II in vivo. Mol. Cell *17*, 831–840.

Mayer, A., Di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J.A., and Churchman, L.S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell *161*, 541–554.

Mercer, T.R., Clark, M.B., Andersen, S.B., Brunck, M.E., Haerty, W., Crawford, J., Taft, R.J., Nielsen, L.K., Dinger, M.E., and Mattick, J.S. (2015). Genome-wide discovery of human splicing branchpoints. Genome Res. *25*, 290–303.

Meyer, M., Plass, M., Pérez-Valle, J., Eyras, E., and Vilardell, J. (2011). Deciphering 3'ss Selection in the Yeast Genome Reveals an RNA Thermosensor that Mediates Alternative Splicing. Mol. Cell *43*, 1033–1039.

Millevoi, S., Geraghty, F., Idowu, B., Tam, J.L.Y., Antoniou, M., and Vagner, S. (2002). A novel function for the U2AF 65 splicing factor in promoting pre-mRNA 3′-end processing. EMBO Rep. *3*, 869–874.

Milligan, L., Huynh-Thu, V.A., Delan-Forino, C., Tuck, A., Petfalski, E., Lombraña, R., Sanguinetti, G., Kudla, G., and Tollervey, D. (2016). Strand-specific, high-resolution mapping of modified RNA polymerase II. Mol. Syst. Biol. *12*, 874.

Moldón, A., Malapeira, J., Gabrielli, N., Gogol, M., Gómez-Escoda, B., Ivanova, T., Seidel, C., and Ayté, J. (2008). Promoter-driven splicing regulation in fission yeast. Nature *455*, 997–1000.

Morawska, M., and Ulrich, H.D. (2013). An expanded tool kit for the auxin-inducible degron system in budding yeast. Yeast *30*, 341–351.

Morris, K.J., and Corbett, A.H. (2018). The polyadenosine RNA-binding protein ZC3H14 interacts with the THO complex and coordinately regulates the processing of neuronal transcripts. Nucleic Acids Res. *46*, 6561–6575.

Müller-McNicoll, M., and Neugebauer, K.M. (2013). How cells get the message: Dynamic

assembly and function of mRNA-protein complexes. Nat. Rev. Genet. *14*, 275–287.

Muniz, L., Deb, M.K., Aguirrebengoa, M., Lazorthes, S., Trouche, D., and Nicolas, E. (2017). Control of Gene Expression in Senescence through Transcriptional Read-Through of Convergent Protein-Coding Genes. Cell Rep. *21*, 2433–2446.

Naftelberg, S., Schor, I.E., Ast, G., and Kornblihtt, A.R. (2015). Regulation of Alternative Splicing Through Coupling with Transcription and Chromatin Structure. Annu. Rev. Biochem. *84*, 165–198.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science (80-.). *320*, 1344–1349.

Nedelcheva-Veleva, M.N., Sarov, M., Yanakiev, I., Mihailovska, E., Ivanov, M.P., Panova, G.C., and Stoynov, S.S. (2013). The thermodynamic patterns of eukaryotic genes suggest a mechanism for intron-exon recognition. Nat. Commun. *4*, 2101.

Neves, L.T., Douglass, S., Spreafico, R., Venkataramanan, S., Kress, T.L., and Johnson, T.L. (2017). The histone variant H2A.Z promotes efficient cotranscriptional splicing in S. cerevisiae. Genes Dev. *31*, 702–717.

Nissen, K.E. (2017). Genetic studies of co-transcriptional pre-mRNA splicing regulation. UCSF.

Nissen, K.E., Homer, C.M., Ryan, C.J., Shales, M., Krogan, N.J., Patrick, K.L., and Guthrie, C. (2017). The histone variant H2A.Z promotes splicing of weak introns. Genes Dev. *31*, 688–701.

Niwa, M., and Berget, S.M. (1991). Mutation of the AAUAAA polyadenylation signal depresses in vitro splicing of proximal but not distal introns. Genes Dev. *5*, 2086–2095.

Nojima, T., Gomes, T., Grosso, A.R.F., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca, M., and Proudfoot, N.J. (2015). Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. Cell *161*, 526–540.

Nojima, T., Gomes, T., Carmo-Fonseca, M., and Proudfoot, N.J. (2016). Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. Nat. Protoc.

Nojima, T., Rebelo, K., Gomes, T., Grosso, A.R., Proudfoot, N.J., and Carmo-Fonseca, M. (2018). RNA Polymerase II Phosphorylated on CTD Serine 5 Interacts with the Spliceosome during Co-transcriptional Splicing. Mol. Cell *72*, 369–379.

Nott, A., Meislin, S.H., and Moore, M.J. (2003). A quantitative analysis of intron effects on mammalian gene expression. RNA *9*, 607–617.

O'Duibhir, E., Lijnzaad, P., Benschop, J.J., Lenstra, T.L., Leenen, D., Groot Koerkamp, M.J., Margaritis, T., Brok, M.O., Kemmeren, P., and Holstege, F.C. (2014). Cell cycle population effects in perturbation studies. Mol. Syst. Biol. *10*, 732.

Osheim, Y.N., O.L. Miller, J., and Beyer, A.L. (1985). RNP particles at splice junction sequences on Drosophila chorion transcripts. Cell *43*, 143–151.

Pabis, M., Neufeld, N., Steiner, M.C., Bojic, T., Shav-Tal, Y., and Neugebauer, K.M. (2013). The nuclear cap-binding complex interacts with the U4/U6·U5 tri-snRNP and promotes spliceosome assembly in mammalian cells. RNA *19*, 1054–1063.

Pai, A.A., Henriques, T., McCue, K., Burkholder, A., Adelman, K., and Burge, C.B. (2017). The kinetics of pre-mRNA splicing in the Drosophila genome and the influence of gene

architecture. Elife *6*, 1–26.

Pak, C.H., Garshasbi, M., Kahrizi, K., Gross, C., Apponi, L.H., Noto, J.J., Kelly, S.M., Leung, S.W., Tzschach, A., Behjati, F., et al. (2011). Mutation of the conserved polyadenosine RNA binding protein, ZC3H14/dNab2, impairs neural function in Drosophila and humans. Proc. Natl. Acad. Sci. U. S. A. *108*, 12390–12395.

Plaschka, C., Newman, A.J., and Nagai, K. (2019). Structural basis of nuclear pre-mRNA splicing: Lessons from Yeast. Cold Spring Harb. Perspect. Biol.

Proudfoot, N.J. (2016). Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. Science (80-.). *352*.

Rabani, M., Levin, J.Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., et al. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. Nat. Biotechnol. *29*, 436–442.

Rabani, M., Raychowdhury, R., Jovanovic, M., Rooney, M., Stumpo, D.J., Pauli, A., Hacohen, N., Schier, A.F., Blackshear, P.J., Friedman, N., et al. (2014). High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. Cell *159*, 1698–1710.

Reimer, K., Mimoso, C., Adelman, K., and Neugebauer, K. Rapid and Efficient Co-Transcriptional Splicing Enhances Mammalian Gene Expression. Submitted.

Rigo, F., and Martinson, H.G. (2008). Functional Coupling of Last-Intron Splicing and 3'-End Processing to Transcription In Vitro: the Poly(A) Signal Couples to Splicing before Committing to Cleavage. Mol. Cell. Biol. *28*, 849–862.

Rutkowski, A.J., Erhard, F., L'Hernault, A., Bonfert, T., Schilhabel, M., Crump, C., Rosenstiel, P., Efstathiou, S., Zimmer, R., Friedel, C.C., et al. (2015). Widespread disruption of host transcription termination in HSV-1 infection. Nat. Commun. *20*, 7126.

Ryan, K., Calvo, O., and Manley, J.L. (2004). Evidence that polyadenylation factor CPSF-73 is the mRNA 3′ processing endonuclease. RNA *10*, 565–573.

Sadowski, M., Dichtl, B., Hübner, W., and Keller, W. (2003). Independent functions of yeast Pcf11p in pre-mRNA 3′ end processing and in transcription termination. EMBO J. *22*, 2167–2177.

Sakharkar, M.K., Perumal, B.S., Sakharkar, K.R., and Kangueane, P. (2005). An analysis on gene architecture in human and mouse genomes. In Silico Biol. *5*, 347–365.

Saldi, T., Cortazar, M.A., Sheridan, R.M., and Bentley, D.L. (2016). Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. J. Mol. Biol. *428*, 2623–2635.

Schmid, M., Olszewski, P., Pelechano, V., Gupta, I., Steinmetz, L.M., and Jensen, T.H. (2015). The Nuclear PolyA-Binding Protein Nab2p Is Essential for mRNA Production. Cell Rep. *12*, 128–139.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–864.

Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure. Nat. Struct. Mol. Biol. *16*, 990–995.

Selenko, P., Gregorovic, G., Sprangers, R., Stier, G., Rhani, Z., Krämer, A., and Sattler, M. (2003). Structural basis for the molecular recognition between human splicing factors U2AF65 and

SF1/mBBP. Mol. Cell *11*, 965–976.

Singh, J., and Padgett, R.A. (2009). Rates of in situ transcription and splicing in large human genes. Nat. Struct. Mol. Biol. *16*, 1128–1133.

Singh, G., Pratt, G., Yeo, G.W., and Moore, M.J. (2015). The Clothes Make the mRNA: Past and Present Trends in mRNP Fashion. Annu. Rev. Biochem. *84*, 325–354.

Soucek, S., Zeng, Y., Bellur, D.L., Bergkessel, M., Morris, K.J., Deng, Q., Duong, D., Seyfried, N.T., Guthrie, C., Staley, J.P., et al. (2016). Evolutionarily Conserved Polyadenosine RNA Binding Protein Nab2 Cooperates with Splicing Machinery To Regulate the Fate of Pre-mRNA. Mol. Cell. Biol. *36*, 2697–2714.

Talerico, M., and Berget, S.M. (1994). Intron definition in splicing of small Drosophila introns. Mol. Cell. Biol. *14*, 3434–3445.

Tardiff, D.F., and Rosbash, M. (2006). Arrested yeast splicing complexes indicate stepwise snRNP recruitment during in vivo spliceosome assembly. RNA *12*, 968–979.

Tardiff, D.F., Lacadie, S.A., and Rosbash, M. (2006). A Genome-Wide Analysis Indicates that Yeast Pre-mRNA Splicing Is Predominantly Posttranscriptional. Mol. Cell *24*, 917–929.

Teng, X., Dayhoff-Brannigan, M., Cheng, W.C., Gilbert, C.E., Sing, C.N., Diny, N.L., Wheelan, S.J., Dunham, M.J., Boeke, J.D., Pineda, F.J., et al. (2013). Genome-wide consequences of deleting any single gene. Mol. Cell *52*, 485–494.

Tibshirani, R. (1994). Regression Shrinkage and Selection Via the Lasso. J. R. Stat. Soc. Ser. B *58*, 267–288.

Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcárcel, J., and Guigó, R. (2009). Nucleosome positioning as a determinant of exon recognition. Nat. Struct. Mol. Biol. *16*, 996–1001.

Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res. *22*, 1616–1625.

Tilgner, H., Jahanbani, F., Gupta, I., Collier, P., Wei, E., Rasmussen, M., and Snyder, M. (2018). Microfluidic isoform sequencing shows widespread splicing coordination in the human transcriptome. Genome Res. *28*, 231–242.

Tuck, A.C., and Tollervey, D. (2013). A transcriptome-wide atlas of RNP composition reveals diverse classes of mRNAs and lncRNAs. Cell *154*, 996–1009.

Újvári, A., and Luse, D.S. (2004). Newly initiated RNA encounters a factor involved in splicing immediately upon emerging from within RNA polymerase II. J. Biol. Chem. *279*, 49773–49779.

Vagner, S., Vagner, C., and Mattaj, I.W. (2000). The carboxyl terminus of vertebrate poly(A) polymerase interacts with U2AF 65 to couple 3'-end processing and splicing. Genes Dev. *14*, 403–413.

Veloso, A., Kirkconnell, K.S., Magnuson, B., Biewen, B., Paulsen, M.T., Wilson, T.E., and Ljungman, M. (2014). Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. Genome Res. *24*, 896–905.

Venkatesh, S., and Workman, J.L. (2015). Histone exchange, chromatin structure and the

regulation of transcription. Nat. Rev. Mol. Cell Biol. *16*, 178–189.

Vilborg, A., Passarelli, M.C., Yario, T.A., Tycowski, K.T., and Steitz, J.A. (2015). Widespread Inducible Transcription Downstream of Human Genes. Mol. Cell *59*, 449–461.

Viphakone, N., Voisinet-Hakil, F., and Minvielle-Sebastia, L. (2008). Molecular dissection of mRNA poly(A) tail length control in yeast. Nucleic Acids Res. *36*, 2418–2433.

Visa, N., Izaurralde, E., Ferreira, J., Daneholt, B., and Mattaj, I.W. (1996). A nuclear cap-binding complex binds Balbiani ring pre-mRNA cotranscriptionally and accompanies the ribonucleoprotein particle during nuclear export. J. Cell Biol. *133*, 5–14.

Wahl, M.C., Will, C.L., and Lührmann, R. (2009). The Spliceosome: Design Principles of a Dynamic RNP Machine. Cell *136*, 701–718.

Warf, M.B., and Berglund, J.A. (2010). Role of RNA structure in regulating pre-mRNA splicing. Trends Biochem. Sci. *35*, 169–178.

Weber, C.M., Ramachandran, S., and Henikoff, S. (2014). Nucleosomes are context-specific, H2A.Z-Modulated barriers to RNA polymerase. Mol. Cell *53*, 819–830.

Wetterberg, I., Zhao, J., Masich, S., Wieslander, L., and Skoglund, U. (2001). In situ transcription and splicing in the Balbiani ring 3 gene. EMBO J. *20*, 2564–2574.

Wick, R.R., Judd, L.M., and Holt, K.E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. *20*, 129.

Will, C.L., and Lührmann, R. (2011). Spliceosome structure and function. Cold Spring Harb. Perspect. Biol. *3*.

Windhager, L., Bonfert, T., Burger, K., Ruzsics, Z., Krebs, S., Kaufmann, S., Malterer, G., L'Hernault, A., Schilhabel, M., Schreiber, S., et al. (2012). Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. Genome Res. *22*, 2031–2042.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., et al. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science (80-.). *285*, 901–906.

Yu, Y., and Reed, R. (2015). FUS functions in coupling transcription to splicing by mediating an interaction between RNAP II and U1 snRNP. Proc. Natl. Acad. Sci. U. S. A. *112*, 8608–8613.

Zaborowska, J., Egloff, S., and Murphy, S. (2016). The pol II CTD: New twists in the tail. Nat. Struct. Mol. Biol. *23*, 771–777.

Zeisel, A., Köstler, W.J., Molotski, N., Tsai, J.M., Krauthgamer, R., Jacob-Hirsch, J., Rechavi, G., Soen, Y., Jung, S., Yarden, Y., et al. (2011). Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. Mol. Syst. Biol. *7*, 529.

Zhao, C., Liu, F., and Pyle, A.M. (2018). An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. RNA *24*.