GENOMIC APPROACHES TO CHARACTERIZE PRE-mRNA SPLICING REGULATION IN SCHIZOSACCHAROMYCES POMBE

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Amy Larson
May 2017
The protein-coding regions of most eukaryotic genes are interrupted by non-coding introns, which must be excised from the pre-messenger RNA (pre-mRNA) prior to translation. Pre-mRNA splicing is performed by the spliceosome, a macromolecular machine that assembles anew on every intron, catalyzing both the removal of the intron and ligation of exons. The central role of pre-mRNA splicing in regulating gene expression is seen in the large number of human diseases which are associated with mutations in this pathway. Nevertheless, the mechanisms by which the complex splicing pathway are regulated remain poorly understood. For example, while it has been established that the spliceosome is co-transcriptionally recruited to introns via interactions with the C-terminal domain of RNA polymerase II, the mechanism by which this enables spliceosome assembly is unknown.

To better understand the mechanisms of splicing regulation, our work harnesses the power of the unicellular fission yeast, *Schizosaccharomyces pombe*, which shares many features of mammalian splicing systems. The splice site sequences found within *S. pombe* introns are marked by a degeneracy similar to that seen in human introns. Additionally, nearly 50% of *S. pombe* genes contain an intron, and almost half of those contain multiple introns, a prerequisite for exon-skipping. Indeed, our lab recently
demonstrated the first examples in *S. pombe* of environmentally-regulated exon skipping of deeply evolutionarily conserved exons.

Here I describe our work developing a sequencing-based screen that uses a custom-designed barcoding scheme to simultaneously measure changes in *in vivo* levels of pre-mRNAs in the background of thousands of arrayed mutants. This approach was originally used to assay a library of non-essential genes in *S. pombe*, enabling identification of scores of strains that displayed defects in splicing of two endogenous pre-mRNAs. This work identified splicing phenotypes associated with factors with described roles in other RNA-processing pathways, such as heterochromatic silencing and 3’end processing, highlighting the interconnected nature of RNA processing. More recently, this approach has been used to interrogate a collection of conditional mutants generated by random mutagenesis to identify canonical splicing mutants. Identification of the causative mutations by whole genome sequencing has led to the identification of novel mutant alleles in the known splicing factors Sap61, Prp22, Cdc28, and Prp1. Our ongoing characterizations of these alleles coupled with recently published structures of several spliceosome complexes suggests that they will provide fascinating insights into the mechanisms by which the spliceosome can function to control regulation. Given the high level of homology between the splicing apparatus in *S. pombe* and humans, we expect that the results of these studies may provide important insights not only into *S. pombe* biology, but into mechanisms of mammalian regulation and disease.
BIOGRAPHICAL SKETCH

Amy Larson was born in Scranton, PA. After moving around several times throughout the early years of her life, she and her family finally settled down near Jim Thorpe, PA where she attended high school. It was during her sophomore biology course that she first realized she was interested in science. She enrolled in the forensic chemistry program at West Chester University of Pennsylvania. While enrolled at WCUPA, she learned two very important things about herself. The first was that she loves to help people and teaching is a passion of hers. At the end of her first year chemistry courses, her general chemistry professor recommended she become a Supplemental Instructor through the university’s Learning and Resource Center. Amy spent 3.5 years helping college students in general and organic chemistry.

The second important thing that Amy learned was after a summer internship working in an equine toxicology lab. It was there that she wanted to do more than spend every day performing the same techniques every day. Rather, she wanted to explore and push the boundaries of what is known. After taking a molecular genetics course that subsequent fall semester, she knew she wanted to be a molecular biologist and to learn more about the fascinating world of nucleic acids.

After graduating from WCUPA with a Bachelors in Forensic Chemistry and a Bachelors in Molecular and Cell biology, Amy enrolled in a PhD program in Biochemistry, Molecular and Cell Biology and Cornell University in August 2009. There she joined the lab of Jeff Pleiss and spent the next seven years exploring the
world of splicing in fission yeast. Additionally, she spent a lot of time teaching an undergraduate genetics course and improving her teaching skills through workshops at the Center for Teaching Excellence, and even served as a fellow there for two years. Amy hopes to someday combine the two things that she loves, teaching and research, working at a primarily undergraduate institution mentoring the next generation of scientists.
To Grandma. I miss you every day.
ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Jeff Pleiss, for accepting me into his lab and mentoring me over the first few years of my scientific career. Jeff pushed me beyond my comfort zone, and refused to let me say phrases like “I’m just not that good at math.” He believed in me as an experimentalist long before I ever did, and for that I am forever grateful. I would also like to thank my committee members, Eric Alani and Ailong Ke, for being interested in my work, and giving me important advice at many points during my graduate career.

I would also like to thank several members of the Pleiss lab with whom I have had the pleasure of working. First I need to thank Ben Fair, the co-first author on the work presented in chapters two and three of this work. Without his help, this dissertation would have been much more difficult. I would also like to thank Zach and Nick who collectively and individually spent several hours with me teaching me how to analyze sequencing data. I would like to thank the rest of the Pleiss lab members, Hansen, Madhura and Mike, for their constant support and help these past few years. I would also like to thank former lab members Mohan, Ali, and Laura who made the first few years in the lab incredibly enjoyable.

I would also like to thank my crazy family for their constant support over the past seven years and for understanding why I couldn’t make it to many family functions over the past few years.
To all of my wonderful and crazy friends that have become family over the past seven years who made these the best seven years of my life. From crazy wine tours, to helping me train for my first (and many other) half marathons, to performing in the Vagina Monologues with me, game nights, the many beer tastings, the long conversations about science, life, and love, thank you for all the wonderful memories. I love you all. To my teammates, thank you. Playing for the Cornell Club Field Hockey team gave me such joy reconnecting with a sport I grew up playing. Thank you to everyone I have played ice hockey with for all the fun on the ice, whether in the town adult league, women’s hockey tournaments, or just skating on the pond. These memories and the people I met were invaluable to my time here.

Finally, to the two most important people I have met in Ithaca, I couldn’t have done this without you. To Carolyn Jablonowski, who first became my roommate, then best friend, and finally my sister. Thank you for sticking by me through all of life’s ups and downs and believing in me when I didn’t. Graduate school was an amazing time in my life, and I owe so many of my amazing memories to you. Thank you just isn’t enough. Last but not least, Phil Gordon. To think you thought I hated you when we first met in our Intro to Hockey class five years ago. We eventually became friends, sharing a love of hockey together. Now, you are my absolute best friend, my rock, and my biggest cheerleader. Thank you for challenging me in every possible way and always pushing me to be the best version of myself. You helped me get through the last year in graduate school in so many ways. Thank you for what you do for me every single day.
TABLE OF CONTENTS

GENOMIC APPROACHES TO CHARACTERIZE PRE-mRNA SPICING REGULATION IN SCHIZOSACCHAROMYCES POMBE ............................................... 2

BIOGRAPHICAL SKETCH ................................................................................... iii

ACKNOWLEDGEMENTS ..................................................................................... vi

TABLE OF CONTENTS ....................................................................................... viii

CHAPTER 1 ............................................................................................................. 1
  1.1 Introduction to Splicing ................................................................................ 1
    1.1.1 Spliceosome Assembly ........................................................................ 4
    1.1.2 Splicing and Disease ........................................................................... 8
  1.2 The Power of Fission Yeast for Studying Splicing ....................................... 9
  1.3 Splicing Connections with Other mRNA Processing Events ..................... 15
    1.3.1 Spliceosome Assembly is Mechanistically Coupled with Transcription 16
    1.3.2 Promoter Architecture Influence on Splicing ....................................... 20
    1.3.3 Chromatin effects on Splicing .............................................................. 21
    1.3.4 Impacts of 3’ End Processing on Terminal Exon Definition ................. 22
  1.4 Genomic Approaches for Identifying Splicing Factors .............................. 24
  1.5 Structural Insights into Splicing Mechanism ............................................. 26
  1.6 Contribution ............................................................................................... 30

CHAPTER 2 ........................................................................................................... 33
  2.1 Abstract ....................................................................................................... 33
  2.2 Introduction .................................................................................................. 34
  2.3 Materials and Methods ................................................................................ 38
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.1</td>
<td>Strains and Cell Growth</td>
<td>38</td>
</tr>
<tr>
<td>2.3.2</td>
<td>cDNA Synthesis and Library Preparation</td>
<td>39</td>
</tr>
<tr>
<td>2.3.3</td>
<td>Data Processing</td>
<td>44</td>
</tr>
<tr>
<td>2.3.4</td>
<td>Splicing Sensitive Microarrays</td>
<td>45</td>
</tr>
<tr>
<td>2.4</td>
<td>Results and Discussion</td>
<td>46</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Known Splicing Factors Identified Here Display Global Splicing Defects</td>
<td>62</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Predicted Splicing Factors Also Display Global Splicing Defects</td>
<td>65</td>
</tr>
<tr>
<td>2.4.3</td>
<td>Genes involved in heterochromatin formation show a range of genome-wide splicing defects</td>
<td>66</td>
</tr>
<tr>
<td>2.4.4</td>
<td>3' end processing factors impact the splicing of both terminal and non-terminal introns</td>
<td>75</td>
</tr>
<tr>
<td>2.5</td>
<td>Conclusion</td>
<td>81</td>
</tr>
<tr>
<td>2.6</td>
<td>Author Contributions</td>
<td>82</td>
</tr>
<tr>
<td>3.1</td>
<td>Abstract</td>
<td>83</td>
</tr>
<tr>
<td>3.2</td>
<td>Introduction</td>
<td>84</td>
</tr>
<tr>
<td>3.3</td>
<td>Materials and Methods</td>
<td>87</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Strains, Cell Growth, Screen Libraries, and Data Processing</td>
<td>87</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Backcrossing and Random Spore Generation</td>
<td>89</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Whole Genome Sequencing to Identify the Causative Mutation</td>
<td>89</td>
</tr>
<tr>
<td>3.3.3.1</td>
<td>Spore Phenotyping</td>
<td>89</td>
</tr>
<tr>
<td>3.3.3.2</td>
<td>Genomic DNA and Sequencing Library Preparation</td>
<td>90</td>
</tr>
<tr>
<td>3.3.3.3</td>
<td>Data Processing</td>
<td>91</td>
</tr>
</tbody>
</table>
3.3.4 RNA-sequencing ................................................................. 93
3.3.5 Complementation Testing .................................................. 94
3.4 Results ..................................................................................... 94
3.4.1 Sequencing-Based Screen Identifies ts Strains with Defects in Splicing 94
3.4.2 Determining the relationship between splicing and temperature phenotypes ................................................................. 98
3.4.3 Bulked Segregant Analysis Followed by Whole Genome Sequencing Identifies the Putative Causative Mutation .............................. 104
3.4.4 Characterization of sap61-E7K ............................................. 108
  3.4.4.1 Sap61+ rescues the ts phenotype in sap61-E7K ................. 108
  3.4.4.2 The sap61-E7K strain displays a genome-wide splicing defect .... 110
3.4.5 Pooling by Temperature Sensitive Phenotype Shows Missense Mutations in Splicing Factors ...................................................... 112
3.5 Discussion .............................................................................. 119
3.6 Author Contributions .............................................................. 132

CHAPTER 4 ...................................................................................... 133
4.1 Spliceosome Assembly is Largely Co-Transcriptional ............... 133
4.2 Preliminary Data ..................................................................... 137
  4.2.1 Sequencing-Based Screen Shows Mutants Have Different Effects on Introns within a Transcript .................................................. 137
  4.2.2 Whole Genome Sequencing to Identify the Mutation Causing the Downstream Splicing Defect Reveals a SNP in prp10. .............. 149
4.3 Discussion and Future Directions ............................................. 151
4.4 Author Contributions .............................................................. 155
1.1 Introduction to Splicing

Most eukaryotic genes are interrupted by non-coding sequences called introns that must be removed to create a translatable message. The process of removing these introns and subsequent ligation of the flanking coding regions, called exons, is termed pre-mRNA splicing. This essential process in eukaryotic gene expression is catalyzed by the spliceosome, a highly dynamic, multi-megadalton ribonucleoprotein complex that recognizes conserved sequences at the intron-exon boundary to remove introns with nucleotide precision. The spliceosome is comprised of five small nuclear RNA-protein complexes (snRNPs), each consisting of a single RNA molecule and core protein factors, as well as upwards of 200 auxiliary proteins that aid in recognition of splicing signals, regulation of the splicing process, and integration of splicing with other mRNA biogenesis events in the nucleus\(^1\). Additionally, alternative splicing, which increases the complexity of the proteome, is pervasive in higher eukaryotes, and recent RNA-seq studies indicate that more than 95% of multi-intronic human genes undergo alternative splicing in a regulated, context-dependent manner underscoring the importance of splicing in regulating gene expression\(^1,2\). Understanding the molecular basis of pre-mRNA splicing and all of the players involved in the splicing pathway is of utmost importance to not only understanding eukaryotic gene
expression, but also for understanding the underlying mutational cause of diseases and cancer as aberrant pre-mRNA splicing is the basis for many human diseases\textsuperscript{3}.

The splicing reaction consists of two transesterification reactions that serve to both remove the intron and ligate the two exons. To carry out these two reactions, the spliceosome must first recognize the intron-exon boundaries. The U1 snRNP identifies the 5′ splice site (SS) sequence located at the 5′ boundary of the intron and the U1 snRNA base-pairs with this six nucleotide sequence. At the other end, the 3′SS delineates the 3′ end of the intron, while the branchpoint sequence (BP) sits about 18-40 nucleotides upstream. The U2 snRNA binds to the BP, causing a conserved adenosine residue to bulge out of the BP, and its 2′OH carries out a nucleophilic attack on the 5′SS, which cleaves the 5′ exon from the exon. While still tethered to the spliceosome, the upstream exon now has a free 3′OH group that can now attack the 3′SS, resulting in precise ligation of the two exons while also releasing the intron from the mRNA\textsuperscript{4}. These two reactions are illustrated in Figure 1.1.
Figure 1.1 Chemistry of splicing reaction. The branchpoint adenosine bulges out, and its 2’ OH groups makes a nucleophilic attack on the phosphate group of the first nucleotide of the intron to form a covalent 2’-5’ bond, releasing the upstream exon from the intron. The now free 3’OH group on the upstream exon makes a nucleophilic attack on the phosphate group of the first nucleotide in the downstream exon to ligate the two exons together forming a 3’-5’ bond, and releasing the intron from the transcript.
Remarkably, this chemistry of pre-mRNA splicing by the spliceosome resembles that seen in the group II introns, a self-splicing class of introns present in all domains of life. Group II introns undergo two transesterification reactions, with the first involving an attack of a 2’ OH from an adenosine on the first nucleotide of the intron, followed by the free 3’ OH of exon one attacking the intron-exon junction to release the ligated exon product and the intron as a lariat. The fact that the chemistry of splicing of Group II introns and spliceosomal introns are so similar strongly suggests that spliceosomal introns have evolved from Group II.

1.1.1 Spliceosome Assembly

The spliceosome must faithfully assemble anew on each intron, and it must do so by recognizing a limited number of nucleotides (the 5’SS, BP, and 3’SS) from the immense sequence space of the transcriptome. Many trans-acting factors associate with the pre-mRNA and components of the spliceosome prior to spliceosome assembly on the intron, and with the spliceosome throughout the splicing cycle.

In addition to the splice sites within an intron, there are other cis-acting elements that are classified as exonic splicing enhancers (ESEs) or exonic splicing silencers (ESSs), and intronic splicing enhancers (ISEs) or intronic splicing silencers (ISSs), to which regulatory proteins bind and stimulate or repress spliceosome assembly on nearby splice sites.

Spliceosome assembly is a dynamic and ordered process of the snRNPs and non-snRNP splicing factors interacting with each other and the pre-mRNA substrate.
Much of what is known about spliceosome assembly is derived from *in vitro* experiments characterizing discrete spliceosomal complexes from native gels. The U1 and U2 snRNPs are recruited to their respective splice sites, forming what is termed the A complex. After this, the U4/U6.U5 tri-snRNP is recruited to the intron, forming the pre-catalytic spliceosome, known as the pre-catalytic B complex. Several RNA-RNA and RNA-protein interactions must be interchanged and destabilize the U1 and U4 snRNPs associations with the spliceosome. Either during or before these conformational changes take place, an essential but non-snRNP complex, the NineTeen Complex (NTC, named after the splicing factor Prp19p) is recruited to the spliceosome and stays associated during the two chemical steps of splicing, helping to regulate spliceosome conformational changes and increase fidelity. Once the U1 and U4 snRNPs are displaced from the spliceosome and the NTC associates, an activated spliceosomal complex, termed B^{act}, is formed which consists of the U2, U5, and U6 snRNPs. The DEAH-box RNA helicase Prp2p then activates the spliceosome to form the catalytically activated B* complex which then catalyzes the first step of splicing.

At the end of the first step, the C complex is formed, which undergoes structural rearrangements catalyzed by the DEAH-box RNA helicase Prp16p to position the 5’ exon and the 3’ splice site at the catalytic center. This new splicing complex is the C* complex, and is poised to catalyze the second step of splicing. Immediately following the ligation of the two exons and the concomitant excision of the intron from the transcript, the P spliceosomal complex is formed. After undergoing additional
conformational changes catalyzed by the DEAH-box RNA helicase Prp22p, the ligated exons are released from the spliceosome, and the intron lariat remains associated with the spliceosome as part of the intron-lariat spliceosome, ILS, complex. Additional structural rearrangements occur to release the individual snRNPs from the intron-lariat, freeing them to participate in additional rounds of spliceosome assembly and catalysis. This process is illustrated in Figure 1.2.
Figure 1.2 Cross-intron assembly and disassembly of the spliceosome\textsuperscript{10}. 
An incredible number of proteins associate with and dissociate from the spliceosome during each transition of the splicing cycle. Mass spectrometry studies have been pivotal in determining the proteins associated with the spliceosome at each stage\textsuperscript{10,12}. Studies of purified human spliceosomal complexes indicate that over 170 different proteins associate with the metazoan spliceosome during the splicing cycle, while the number of proteins associating with the yeast spliceosome is closer to 100\textsuperscript{15}. The transition between each step of the splicing cycle is accompanied by extensive changes in protein composition, most of which are conserved from yeast to humans. The greatest change occurs during the transition from the precatalytic B complex to the activated B\textsuperscript{act} complex, where as many as 35 leave and 12 new proteins are recruited\textsuperscript{4}. The transition between each spliceosome complex is tightly regulated by the activity of essential DExD/H-box RNA helicases. These proteins increase the fidelity of the splicing process, sending aberrant and non-productive splicing intermediates down discard pathways\textsuperscript{16,17}.

1.1.2 Splicing and Disease

Splicing has gained recognition as an important component of gene regulation as more and more examples of splicing defects have become associated with the etiology of many human diseases. More than 200 human diseases are caused by point mutations that destroy or weaken splice sites, or activate cryptic splice sites\textsuperscript{18}. Xiong \textit{et al} identified more than 10,000 intronic single nucleotide polymorphisms (SNPs), as well as \textasciitilde70,000 missense and synonymous SNPs in exonic regions that affect
enhancer and silencer elements that potentially link splicing defects to human
disease\textsuperscript{19}. In addition to \textit{cis}-acting mutations, many mutations that affect \textit{trans}-acting
factors lead to splicing defects in a number of different ways. In the case of DM1 and
DM2 myotonic dystrophies, the protein muscleblind-like (MBNL) is recruited to
mRNAs that have expansions of CUG and CCUG repeats. This recruitment
sequesters MBNL, thus inhibiting its ability to bind to its normal mRNA targets, and
ultimately leads to a change in alternative splicing of hundreds of transcripts\textsuperscript{20–22}.
Many cancers have changes in the nuclear concentration of many splicing regulators,
which ultimately leads to changes that affect cell cycle control and apoptosis. More
importantly, mutations in core spliceosome components are becoming a recurrent
theme in many disease states, with studies relating to cancer and neuropathologies
being the most prevalent\textsuperscript{3}. How these mutations mechanistically lead to these disease
states, however, remains largely unknown.

1.2 The Power of Fission Yeast for Studying Splicing

Much of what is known about the core spliceosome and the mechanism of splicing
has come from studying the budding yeast, \textit{Saccharomyces cerevisiae}. The first
genetic screen that identified mutants defective in RNA production was published in
1967 by Leland Hartwell, where his lab performed a screen in \textit{S. cerevisiae} on
temperature sensitive (\textit{ts}) mutants and found 10 complementation groups\textsuperscript{23}. About
twenty years later, it was elucidated by John Abelson’s group that many of these
mutants were defective for pre-mRNA splicing and were renamed as pre-mRNA
processing (prp) mutants\textsuperscript{24}. Additionally, Christine Guthrie’s group screened through 340 cold sensitive (cs) mutants for splicing mutants, and identified 12 mutants in nine different complementation groups\textsuperscript{25}. Early genetic screens using libraries of randomly mutagenized ts and cs strains were beneficial in pioneering the identification of many of the core factors essential for pre-mRNA splicing. In addition, many suppressor screens of these cs and ts mutants have been instrumental in our understanding of how many of these factors function in the splicing pathway, and have led to initial understandings of how DexD/H-box proteins function in maintaining the fidelity of splicing, for example. Additionally, in vitro splicing reactions using purified biochemical extracts from budding yeast have proved instrumental in elucidating the mechanism of splicing catalysis\textsuperscript{26}.

Despite the great advances the splicing field has made utilizing budding yeast as a model system, many questions about complex, regulated splicing pertaining to higher eukaryotes remained unanswered largely because \textit{S. cerevisiae} lacks the complexity that is seen in metazoan splicing. The fission yeast, \textit{Schizosaccharomyces pombe}, on the other hand, more closely resembles human splicing in many ways, but resembles budding yeast in its genetic tractability. One of the most notable similarities between fission yeast and humans is its complex intron landscape. Nearly half of the genes in fission yeast contain an intron, compared to only 5\% of budding yeast genes, and greater than 95\% of human genes\textsuperscript{8,9}. Even more interestingly, almost half of the intron containing genes in fission yeast contain two or more introns, a prerequisite for exon
skipping events\textsuperscript{27}. In addition, the splice sites in fission yeast have a high degree of degeneracy, resembling the degeneracy seen in humans and in contrast to the tight consensus sequences observed in budding yeast\textsuperscript{28,29} (Figure 1.3). The difference between the two yeasts is more noticeable at the 3′ end of the intron. In fission yeast, the branchpoint sequence has more degeneracy, with a CURAY (where R is a purine and Y is a pyrimidine) motif that is also seen in humans, compared to the strict consensus UACUAAC in budding yeast. Also, in about 75% of fission yeast introns there is a polypyrimidine tract between the BP and 3′SS, another feature shared with human introns. Based on these observations, it is thought that the intron architecture in \textit{S. pombe} reflects the ancestral intron\textsuperscript{29}.
Figure 1.3 Comparison of splice sites between humans and yeast. WebLogos displayed here depict the consensus sequence for each splice site in the different species. The height of the letter is proportional to the conservation at that position\textsuperscript{82}.
The presence of more than one intron in a gene is a prerequisite for alternative splicing events. Also, degenerate splice sites allow for the possibility of alternative splicing events in addition to regulation of gene expression. Concordantly, *S. pombe* is the first unicellular eukaryote where environmentally regulated skipping of evolutionarily conserved exons has been observed, with many of these gene structures conserved in humans\(^{30}\). In addition to regulated alternative events, two recent studies have published that the fission yeast spliceosome actually activates alternative splice sites in 2-3% of splicing events, resulting in unstable mRNA isoforms\(^{31,32}\). While human pre-mRNAs undergo alternative splicing to a much higher degree, these results revealed that there are many more splice sites that the fission yeast spliceosome can activate than previously appreciated. This aberrant splice site activation displayed by the fission yeast spliceosome provides an exciting possibility for how more complex gene architectures, like those in humans, may evolve.

In addition to the similarities in intron architecture between fission yeast and humans, they also share many conserved proteins that are absent in the budding yeast splicing machinery, and these proteins often play important roles in splicing regulation\(^{15,33}\). A major class of proteins absent in the budding yeast but present in fission yeast is the SR protein class. SR proteins, so named because they all contain a protein domain of repeating serine and arginine residues, are correlated with complex splicing in metazoans. Humans have nine different SR proteins, while *S. pombe* only has two *bona-fide* SR proteins and one SR-like protein. SR proteins are involved in
regulating alternative splicing events, as well as recognizing and directing the core spliceosome to degenerate 3’ splice sites in constitutive splicing\textsuperscript{34}. The mechanism by which SR proteins perform these functions remains largely unknown in mammals. The genetic tractability of the fission yeast provides an exciting opportunity to study these proteins and elucidate their mechanism of action. Additionally, protein kinases that target spliceosomal proteins are present in fission yeast but absent in budding yeast. The \textit{S. pombe} kinase Prp4p was the first identified kinase with a role in pre-mRNA splicing, and has an important role in regulating the formation of the active spliceosome through phosphorylation of the Prp1p, a protein in the tri-snRNP complex\textsuperscript{35}. More interestingly, 38\% of the known factors involved in splicing in fission yeast are more similar in sequence to humans than they are to their budding yeast counterparts\textsuperscript{33}. Many of the proteins in this category function in identifying the 3’ SS, which aligns with the fact that the 3’ architecture of introns in both fission yeast and humans is degenerate. For example, the large subunits of U2AF are 45\% similar to humans, while the smaller subunits are as much as 75\% similar. The fission yeast Slu7p, a protein involved in recognizing the correct AG dinucleotide at the 3’ end of introns, is 51\% similar to humans.

Taken together, the intron architecture in \textit{S. pombe} likely reflects the ancestral intron, while the protein composition of the splicing machinery in fission yeast resembles the archetype of a spliceosome\textsuperscript{33}. Thus, studying the mechanism of splicing in fission yeast is more likely to provide greater insight into mammalian splicing.
While many of the proteins involved in splicing are known, the roles and mechanism of action for each protein largely remain unknown. The fission yeast provides a unique opportunity to study pre-mRNA splicing due to its genetic tractability, like the budding yeast, but with the advantage that it shares many similarities to metazoan splicing.

1.3 Splicing Connections with Other mRNA Processing Events

Expression of a protein-coding gene involves many steps: transcription, capping, pre-mRNA splicing, 3’ end processing, mRNA export, and translation. These processes are all performed by different multi-protein complexes, but there has been growing evidence over the past decade or so that many of the processes that occur in the nucleus are functionally coupled. This coupling is not mandatory because each system can be reconstructed individually in vitro, but it is thought that coupling occurs to increase the efficiency and fidelity of each step to create a highly regulated gene expression pathway. While a detailed mechanism of how splicing is functionally coupled to many of these nuclear processes remains to be determined, there is a plethora of evidence indicating that pre-mRNA splicing is spatially and temporally connected. Therefore, to understand how splicing is regulated, particularly in complex systems like in metazoans, it is imperative to understand how other processes impact the splicing pathway, as well as how splicing impacts those processes.
1.3.1 Spliceosome Assembly is Mechanistically Coupled with Transcription

Spliceosome assembly not only occurs simultaneously with transcription, but several lines of evidence indicate that they are mechanistically coupled. For example, splicing is highly inefficient when transcripts are synthesized by a polymerase other than Pol II \(^37\). The highly conserved CTD tail of Pol II is largely responsible for efficient splicing when transcripts are transcribed by Pol II. The CTD consists of heptad repeats of the consensus YSPTSPS and acts as a landing pad to recruit mRNA processing factors to the transcription elongation complex, and carry them during transcription\(^38\)–\(^40\). All the serine residues, as well as the tyrosine and threonine residues, can be reversibly phosphorylated, and phosphorylation of the different residues, particularly the serines, has been linked with different stages of transcription and pre-mRNA processing\(^39\)–\(^41\). When Pol II is first recruited to the promoter, the CTD is not phosphorylated, but becomes phosphorylated at the Ser5 and Ser7 residues prior to transcription initiation\(^38\). Phosphorylation states of the different residues change throughout the transcription cycle, and the phosphorylated CTD residues interact with a unique set of proteins that are involved in transcription and RNA processing. A recent study from the Churchman group demonstrated that Ser5 phosphorylation contributes to co-transcriptional spliceosome recruitment by showing an enrichment for U1 components bound to Ser5-P\(^42\). However, other recent studies have cast doubt on a direct connection between the CTD and splicing factors. The fact that U1 snRNPs is not detected on intronless genes in the budding yeast despite being transcribed by Pol II suggests that the CTD might not be solely responsible for
recruiting splicing factors to nascent transcripts\textsuperscript{43}. Perhaps the most convincing
evidence came from Inada \textit{et al.} who demonstrated that CTD mutants that are unable
to be phosphorylated show no defect in pre-mRNA splicing, highlighting the fact that
how CTD marks influence co-transcriptional RNA processing events is still not well
understood\textsuperscript{44}. 
Figure 1.4 Schematic of co-transcriptional assembly of the spliceosome on a nascent transcript occurring through interactions with CTD3.
Chromatin immunoprecipitation (ChIP) studies in budding yeast provided some of the first mechanistic insights into co-transcriptional spliceosome assembly. The fact that U1, U2, and U5 snRNPs can be pulled down in ChIP assays indicated that the spliceosome was assembling on the nascent transcript while it was still tethered to chromatin. Additionally, splicing factors were not seen to be recruited to intronless genes. Therefore, spliceosomal assembly appears to be functionally connected with transcription; however, the chemical steps of splicing can occur post-transcriptionally. While the majority of introns appear to be spliced co-transcriptionally, there is a substantial fraction that complete splicing post-transcriptionally.

The elongation rate of Pol II is another opportunity for transcription to affect splicing efficiency and decisions. In early studies using a mutant Pol II that slowed down the rate of transcription lead to an increase in the inclusion of alternative exons. Speeding up the rate of elongation showed increased skipping of those exons, indicating that kinetic coupling would affect splicing decisions by providing a “window of opportunity.” A general conclusion by several nascent RNA-seq studies is that introns that are alternatively spliced are spliced more slowly than constitutive introns, but they are still largely spliced co-transcriptionally. Other recent genome-wide studies have indicated that splicing and transcription rates are independent of each other, and that alternative exon inclusion in slow and fast Pol II mutants does not always occur in ways predicted by the “window of opportunity.”
model”, and that there might be an optimal elongation rate\textsuperscript{53,54}. Several genome-wide studies have indicated that transcriptional pausing coincides with splicing. NET-seq studies, which sequence the 3’ ends of nascent transcripts, revealed that many 3’ ends were close to splice sites, indicating the splicing signals might also be potent Pol II pausing signals\textsuperscript{55–57}. While the polymerase is paused, the phosphorylation pattern on the CTD tail shows an enrichment for S5 phosphorylation, which switches to an increase in S2 phosphorylation as Pol II is released from the pause. Whether pausing is a cause or a consequence of splicing is a question that remains to be answered.

### 1.3.2 Promoter Architecture Influence on Splicing

Other parts of the transcription pathway can also influence splicing. Several studies have indicated that promoter architecture can affect splicing outcomes\textsuperscript{58}. Different promoter-bound activators can recruit different splicing factors to the transcription machinery, or by affecting the elongation rate of Pol II\textsuperscript{59}. PGC-1 is a promoter activator that changes the splicing outcomes of genes involved in adaptive thermogenesis when it is bound to their promoters\textsuperscript{60}. Since PGC-1 has an RS domain, it is hypothesized that it is responsible for recruiting certain splicing factors to the promoter. A component of the mediator complex, MED23, has been shown to affect alternative splicing decisions by directly interacting with several splicing and polyadenylation factors\textsuperscript{61}. In order to fully understand how promoters can direct splicing outcomes, it will be imperative to ascertain all of the factors that bind to the
promoter, bridge these activators with Pol II, and the components of Pol II that can transmit the information to the pre-mRNA and change the splicing decisions.

1.3.3 Chromatin effects on Splicing

Chromatin can also indirectly affect decisions about splice site usage and exon inclusion. Chromatin can influence splicing by substitution of one of its canonical histone subunits for a non-allelic variant, through methylation patterns on the DNA, or through various posttranslational modifications, known as histone marks, on histone tails. It has been well known that these marks are not distributed evenly throughout the genome, and that there is a distinct pattern denoting promoters from gene bodies. More recently it has become apparent that there is a differential distribution of these marks with respect to exon-intron boundaries. There have been several global studies to characterize the distribution of histone marks within a gene\textsuperscript{62–64}. While some studies produce conflicting results, multiple studies have shown an increase of H3K36me3 in exons\textsuperscript{63,65–67}. Exons also have a higher GC content than the rest the genome, which in turn affects nucleosome occupancy\textsuperscript{68}. Additionally, nucleosomes that flank the 5' and 3' splice sites are enriched for the histone variant H2A.Bbd, indicating that nucleosomes within exons may have a different histone variant compositions than introns\textsuperscript{69}.

Chromatin can also function in recruiting splicing factors either through directly interacting or indirectly via adaptor proteins. Mass spectrometry studies reveal that the histone variant H2A.Bbd binds with many splicing factors, and that deletion of this
variant results in a global defect of alternative and constitutive splicing\textsuperscript{69}. The protein CHD1, a component of the U2 snRNP, actively binds to H3K34me3, the histone mark that denotes actively transcribing genes\textsuperscript{70}. This association between CHD1 and H3K34me3 increases splicing efficiency. These are just a few of the many examples in the literature. Together, they point to a mechanism by which the histone architecture of actively transcribing genes can have an impact on splicing by recruiting the spliceosome to actively transcribing genes and change splicing decisions.

1.3.4 Impacts of 3’ End Processing on Terminal Exon Definition

In higher eukaryotes, it has been well established that splicing of the terminal intron requires cross-talk with the 3’ processing machinery. U2AF binding to the 3’ end of the terminal intron recruits Cleavage Factor I to the poly(A) site (PAS) and directly interacts with the CTD of polyadenylate polymerase. This interaction stimulates both splicing of the terminal intron as well as cleavage and polyadenylation. Thus, recognition of the terminal 3’ SS enhances downstream poly(A) sites\textsuperscript{71,72}. Splicing factors SF3B and SRm160 both interact with the CPSF (cleavage/polyadenylation specificity factor) complex to influence 3’ end processing\textsuperscript{73,74}. Additionally, U1 snRNP has dual roles in splicing and preventing premature cleavage and polyadenylation. Screens for mutants that cause defects in splicing have implicated several 3’ end factors, such as Cft2 and Yth1, being functionally important for efficient splicing in budding yeast\textsuperscript{75}. 

22
Figure 1.5 Coupling connections between splicing and 3′ end formation\textsuperscript{76}. Splicing of terminal introns is coupled to 3 end processing through interactions between SR proteins and factors in the cleavage machinery, as well as interactions between U2AF and the cleavage machinery.
1.4 Genomic Approaches for Identifying Splicing Factors

It has become increasingly clear that individual transcripts may be subjected to their own unique set of splicing factors. However, there has been much value gained by performing large-scale screens in order to identify factors that are involved in the splicing pathway. Proteomic studies are instrumental in providing information not only about the composition of protein complexes, but their function and regulation as well. Over the past decade, quantitative mass spectrometry (MS) has become invaluable for identifying proteins that are associated with the spliceosome at the various stages of the splicing cycle \(^{77-79}\). The MS studies of the spliceosome were limited to \textit{in vitro} derived spliceosomes, and to only those proteins that were visible in 2D gels \(^{77}\). This strategy identified 17 known proteins from genetic studies, and 23 new splicing factors. The goal in the early stages of proteomic analyses of the spliceosome was to identify every protein associated with the core spliceosome regardless of stage. In the early 2000s, it was hard to purify specific spliceosomal subcomplexes, so amalgamations of spliceosome intermediates were used. As MS technologies improved and became more sensitive, the number of proteins detected that interacted with the spliceosome reached almost 300 different proteins \(^{80}\). The challenge, then, became to distinguish which proteins truly associated with the spliceosome, versus those that were simply contaminants, such as general RNA-binding proteins. Another challenge in these proteomic studies has been to try to
distinguish between proteins that are specific to splicing of certain transcripts, versus those that are general splicing factors.

In the past decade, a major goal of splicing proteomic studies has been to identify the proteins in each spliceosome subcomplex to better understand the mechanism of the splicing reaction. In 2014, a study from the Urlaub group used three different methods to quantify the protein composition of pre-catalytic B and catalytic C complexes in humans: metabolic labeling by SILAC, chemical labeling by iTRAQ, and label-free spectral count⁸¹. Using these three methods they identified almost 200 proteins in both the B and C complexes. However, only a few of these proteins corresponded to core components of these complexes. Researchers observed that most proteins join or leave the spliceosome during the transition between spliceosome subcomplexes, which shed light on how transitions in the spliceosome are regulated.

Protein-protein interactions, whether stable or transient, are important for not only the function of the spliceosome, but its regulation as well. Yeast two-hybrid screens have been invaluable in identifying new splicing factors early on, and identifying interactome networks over the past few years⁸². One of the first yeast two-hybrid studies performed to look for splicing interactors used the DEAH-splicing helicase Prp2 from S. cerevisiae as bait and screened through a yeast genomic library. This screen identified Spp2 repeatedly as an interactor of Prp2⁸³. Screening through a library of mutants of Spp2 lead to determining that Prp2 binds to the G-patch of Spp2, providing evidence for how this helicase associates with the spliceosome.
Genome-wide screens have also been used to identify factors that are involved in other cellular processes and interact with the spliceosome. Two of these approaches, Synthetic Genetic Array (SGA)\textsuperscript{84} Analysis and Epistatic MiniArray Profiling (E-MAP)\textsuperscript{85} have been important in identifying functional relationships between proteins in different cellular process, thus functionally linking these processes together. These two techniques take advantage of the ability to generate thousands of double-mutants in yeast. Those double mutants that result in synthetic lethality or synthetic sickness indicate that the interaction of those proteins is essential or important, respectively. Vo et al performed a yeast two-hybrid screen in \textit{S. pombe} to map the interactome in fission yeast and compare it to that of budding yeast and humans\textsuperscript{82}. Interestingly, protein-protein interactions are more conserved between fission yeast and humans than they are between fission and budding yeast. They also demonstrated that there is a physical interaction network between SR and SR-like proteins, which supports the thinking in the field that RS domains primarily interact with other RS domains\textsuperscript{86–89}.

1.5 Structural Insights into Splicing Mechanism

Understanding the detailed mechanism of the spliceosome and splicing reaction is aided by knowledge of the structure and composition of the spliceosome at each stage of the splicing cycle. Solving a high-resolution structure of the spliceosome seemed elusive for many years due to its large size, heterogeneous composition, and dynamic nature. X-ray crystallography has focused on individual splicing proteins, such as Prp8\textsuperscript{90} and Brr2\textsuperscript{91}, and subcomplexes, such as U1\textsuperscript{92,93} and U2 snRNPs\textsuperscript{94–96} for
example. These structures provided insights into the splicing mechanism and confirmed some biochemical observations, but still lacked the bigger picture look that a structure of the entire spliceosome would provide. More recently, electron microscopy (EM) studies of the intact spliceosome at several stages of the splicing cycle have been performed, but the greatest resolution achieved using this method was only 20-29Å\(^97–99\). Thus, only general descriptions of the overall structure of the spliceosome were possible, and a detailed molecular view of the spliceosome was still needed. Thanks to recent major developments in cryo-electron microscopy (cryo-EM), there has been an explosion of structures of splicing complexes within the past two years. Cryo-EM structures of the \textit{S. pombe} ILS complex, the \textit{S. cerevisiae} B\(^\text{act}\), C, and C* complexes, and even the human C* complex have been solved\(^100–106\).

Additionally, cryo-EM structures of both the \textit{S. cerevisiae} and human tri-snRNP complexes have been solved\(^107,108\). Together, these structures have provided enormous insight and details of the catalytic center, and the conformational changes that occur between each complex in the splicing cycle.

The first molecular insight into the architecture of the catalytic core of the spliceosome came from the \textit{S. pombe} ILS cryo-EM structure containing 37 distinct proteins. From this structure, it is clear that the proteins near the catalytic center had evolved in both placement and morphology to allow for and promote the dynamic process that is splicing. Additionally, this structure revealed that many spliceosomal proteins share an extended architecture, and that it is this feature that allows for such
conformational flexibility in addition to increasing the surface area of these proteins to increase the number of possible interactions with other splicing proteins\textsuperscript{100}.

Comparison of the B\textsuperscript{act}, C, C*, and ILS cryo-EM structures demonstrate that the spliceosomal complexes share a stable structural core throughout the splicing reaction, which consists of the U5 snRNP, U6 snRNA, 5′-end of the U2 snRNA, and about a dozen proteins of the NTC and NTR\textsuperscript{105}. Additionally, comparing these structures reveals the conformation changes that take place, from the rotation of the snRNA duplexes to proteins associating and dissociating with the spliceosome, between stages in the splicing cycle. Comparison of the B\textsuperscript{act} and C complex revealed details of the first step of the splicing reaction. In the B\textsuperscript{act} complex, the U2/U6 triplex and Prp8 catalytic core is already established. The 5′ splice site is already poised for the catalytic first step, but is blocked from the catalytic core by a protein density that is no longer present in the C complex\textsuperscript{102,104}. The RNA elements present in the catalytic core after the first splicing step are stabilized by 15 different proteins in the C complex\textsuperscript{103}. Comparison of the C and C* complexes revealed conformational changes that occur after the first step of splicing to activate the spliceosome for the second splicing reaction. One major conformational change that occurs is the lariat junction is translated 15-20Å away from the catalytic core, and the BP/U2 duplex rotates about 90° to also move away from the catalytic core, while step II factors such as Prp17 and Prp18 are recruited to the catalytic core, as well as shifts in the RNase H-like domain of Prp8 to move into its new position in the C* complex catalytic core\textsuperscript{106}. A common
feature among these structures is the placement of the highly conserved DExD/H RNA helicases Prp2, Prp16, and Prp22 on the periphery of their respective spliceosomal complexes. These helicases bind transiently and drive the conformational changes to convert the spliceosomal complexes to the next spliceosome stage. By binding to the periphery, they are at a considerable distance from their presumed targets at the catalytic core, and thus likely exert their functions at a distance.

While these manuscripts describing the structures of the different spliceosomal complexes have been instrumental in deciphering the catalytic core and provide tremendous insight to the molecular underpinnings of the dynamic conformational changes in the spliceosome to facilitate the splicing reaction, they are just scratching the surface for how many of the spliceosomal proteins function in the spliceosome. Preliminary analyses of these structures were provided by the authors. More detailed comparisons between each structure, as well as detailed analyses focusing on one or a subset of proteins will provide an even greater understanding of the dynamics of the spliceosome. Additionally, many of these structures contain only 28-37 proteins, while there have been upwards of 100 different proteins reported to associate with the spliceosome\textsuperscript{52,77,78}. How these other proteins interact with the spliceosome and exert their function remains to be determined. However, the abundance of structural data available in these cryo-EM structures of intact and distinct spliceosomal complexes as well as structural solutions of spliceosomal subcomplexes increases the utility of genetic and biochemical data for informing mechanistic underpinnings of the splicing
reaction. Additionally, comparisons of the yeast and human C* spliceosomal complexes reveals that conserved spliceosomal protein domains perform similar functions, such as stabilizing snRNA duplexes in the spliceosome, even if the domains are organized differently in human proteins\textsuperscript{105}. This observation underscores the utility of yeast as a system for studying splicing mechanism as it relates to human biology.

1.6 Contribution

Over the past two decades, advances in high-throughput methodologies, as well as computational power and approaches has increased our ability to ask questions about the cis elements that contribute to splicing regulation as well as identification of the protein factors that interact with the spliceosome to contribute to its regulation, and much work has been done to understand how splicing decisions are made. Since most human transcripts consist of several introns and are frequently subjected to alternative splicing decisions, I was interested in understanding how these complex splicing decisions are made. Primarily, I was interested in understanding how the spliceosome is recruited and assembled onto downstream introns in transcripts that contain more than one intron. While there is substantial evidence for co-transcriptional recruitment of components of the spliceosome, the molecular mechanism for how several components can be recruited co-transcriptionally for transcripts that contain more than one intron remain unclear.
The goal of my work was to identify mutants in *S. pombe* that are defective in downstream splicing of multi-intronic transcripts to begin to unravel mechanistic details of spliceosome recruitment. The interconnectedness of the splicing pathway with so many other nuclear processes underscores the need for an unbiased genome-wide approach to identify the full complement of factors that impact splicing of downstream introns. Using a liquid handling robot to automate cell growth, RNA isolation, and cDNA synthesis, and the power of next generation sequencing, I developed a sequencing-based method to screen through thousands of mutant yeast strains. One particularly strong advantage of this approach is that pre-mRNAs can be probed in their native environments to determine which mutants impact their splicing *in vivo*. In chapter 2, I describe the design and implementation of this method to screen through a non-essential gene deletion collection in *S. pombe* to identify mutants that impact canonical splicing. This work revealed novel connections between the splicing pathway and heterochromatin silencing, and with 3′ end processing factors. Chapter 3 depicts the work I did using this method to screen through a novel collection of randomly generated unknown conditional mutants in *S. pombe* to first identify canonical splicing mutant alleles. Bulk segregant analysis followed by whole genome sequencing was performed to identify the mutations causing the splicing defects. This work identified novel alleles in core splicing factors, such as Sap61p, a component of SF3a complex in the U2 snRNP, and subsequent structural analysis of these alleles revealed insights into spliceosome assembly, and demonstrated our ability to identify functionally relevant mutations. Finally, in chapter 4 I describe the preliminary data I
have generated screening through a conditional mutant library collection to identify mutants that are defective in downstream splicing. This sequencing method identified a component of the SF3b complex, Prp10p, which has a defect in downstream splicing but has no effect on a canonical splicing substrate. Additional biochemical studies are needed to understand how this mutant could show a defect in downstream splicing. Additional screening through the conditional library will hopefully reveal other mutants defective in downstream splicing and will provide a clearer picture of how the spliceosome is recruited to downstream introns.
CHAPTER 2

INTERCONNECTIONS BETWEEN RNA PROCESSING PATHWAYS REVEALED BY A SEQUENCING BASED GENETIC SCREEN FOR PRE-mRNA SPLICING MUTANTS IN FISSION YEAST

2.1 Abstract

Pre-mRNA splicing is an essential component of eukaryotic gene expression and is highly conserved from unicellular yeasts to humans. Here we present the development and implementation of a sequencing based reverse genetic screen designed to identify non-essential genes that impact pre-mRNA splicing in the fission yeast Schizosaccharomyces pombe, an organism that shares many of the complex features of splicing in higher eukaryotes. Using a custom-designed barcoding scheme, we simultaneously queried ~3000 mutant strains for their impact on the splicing efficiency of two endogenous pre-mRNAs. A total of 61 non-essential genes were identified whose deletions resulted in defects in pre-mRNA splicing; enriched among these were factors encoding known or predicted components of the spliceosome. Included among the candidates identified here are genes with well-characterized roles in other RNA-processing pathways, including heterochromatic silencing and 3’ end processing. Splicing-sensitive microarrays confirm broad splicing defects for many of these factors, revealing novel functional connections between these pathways.
2.2 Introduction

The protein-coding regions of most eukaryotic genes are interrupted by non-coding introns which must be precisely removed from the pre-mRNA in order to generate a translatable message. This essential process is carried out by the spliceosome, a dynamic macromolecular machine that recognizes specific sequence elements within the pre-mRNA, such as the short consensus sequences that mark intron boundaries, and catalyzes intron removal\(^\text{10}\). At its core, the spliceosome is composed of five snRNA-protein complexes (snRNPs), each comprised of a single RNA and multiple core protein factors. In addition, the human spliceosome associates with upwards of 200 auxiliary splicing proteins that aid in proper recognition of splice sites and catalysis\(^\text{4}\).

Over the last decade, it has become increasingly clear that splicing is integrated with other steps of pre-mRNA synthesis and maturation. Studies from yeast to humans suggest that the majority of splicing occurs co-transcriptionally while the RNA is still tethered to the polymerase\(^\text{46,52,111–113}\). Accordingly, multiple lines of evidence support the idea that transcriptional dynamics influence the splicing process. Mutations that alter polymerase elongation rate yield different splicing patterns\(^\text{49,50}\), suggesting a kinetic coupling between transcription rate and the ability of splicing factors to recognize and act upon splice sites. Genome-wide studies have also demonstrated that transcriptional pausing coincides with the splicing process\(^\text{111,114}\). In addition to a kinetic coupling of elongation and splicing, biochemical studies have shown that the
C-terminal domain (CTD) of RNA polymerase II can directly interact with splicing components\textsuperscript{115,116} and post-translational modifications of the CTD can differentially impact recruitment of splicing components\textsuperscript{117–119}. The chromatin environment encountered by the transcribing polymerase can also influence splicing, and genome-wide nucleosome positioning data from fission yeast to humans reveal an enrichment of nucleosome density in exons over introns\textsuperscript{66,67,120–123}. The mechanistic link between chromatin state and splicing could be explained in part by the relation between chromatin state and polymerase kinetics\textsuperscript{124,125}, but might also reflect direct interactions between chromatin marks and splicing factors. For example, the H3K4me3 mark interacts with the U2 snRNP through interactions with the adapter protein CHD1\textsuperscript{126}. In addition to chromatin-based interactions, it is clear that the cleavage and polyadenylation machinery\textsuperscript{127} at the 3′ end of transcripts can interact with splicing components to influence splicing. In higher eukaryotes, identification of splice sites in terminal introns requires interactions between splicing components and the cleavage and polyadenylation machinery. Recently, the cleavage and polyadenylation factor CPSF1 was found to regulate alternative splicing in human T-cells\textsuperscript{128}. These interconnections between splicing and other nuclear processes underscore the need for unbiased genome-wide approaches to identify the full complement of factors that functionally impact spliceosomal activity.

The fission yeast, \textit{Schizosaccharomyces pombe}, provides a powerful genetic system in which to examine the splicing pathway. Like the budding yeast, \textit{Saccharomyces}}
*cerevisiae*, fission yeast is genetically tractable, allowing for easy manipulation of its genome. The *S. cerevisiae* genome, however, has shed most of its introns, with only ~300 introns remaining\(^2^9\). In contrast, over 5000 introns have been identified in the *S. pombe* genome, and over 1000 genes are interrupted by multiple introns\(^2^7,1^2^9\).

Furthermore, whereas splice site sequences in budding yeast introns tend to conform to a strict consensus sequence, *S. pombe* splice sites are characterized by a far higher level of degeneracy, more closely resembling the degeneracy\(^2^9\) seen in human splice sites. Perhaps accordingly, sequence homology identifies many auxiliary components of the human spliceosome, such as SR proteins, that are present in the *S. pombe* genome but have been lost in the *S. cerevisiae* lineage\(^1^5,1^3^0\). These properties suggest that regulation of pre-mRNA splicing in *S. pombe* may be more similar to that seen in humans than in *S. cerevisiae*\(^1^3^1\). Indeed, some transcripts in *S. pombe* are subject to mammalian-like, environmentally regulated exon skipping\(^3^0\), and others respond to insertion of mammalian splicing enhancer elements\(^1^3^0\). Moreover, similar to observations in mammalian cells, widespread activation of cryptic splice sites has been demonstrated in *S. pombe*, highlighting the flexibility in the *S. pombe* spliceosome for selecting splice sites\(^3^1,3^2\). Although these features highlight the potential of *S. pombe* to serve as a model for understanding the complex splicing seen in higher eukaryotes, the precise factors responsible for regulating these splicing events remain largely unknown.
Components of the *S. pombe* spliceosome have been identified using a variety of approaches. Genetic screening of randomly mutagenized strains identified numerous core splicing factors\textsuperscript{35,132–134} and biochemical purifications followed by mass spectrometry have greatly added to the list of components\textsuperscript{135,136}. Although these strategies successfully identified core components of the spliceosome, they have been less effective at identifying factors that functionally connect splicing with other nuclear processes. More recently, a high-throughput genetic interaction mapping strategy examining non-essential *S. pombe* genes identified strong genetic interactions between U2 snRNP components of the spliceosome and chromatin remodeling enzymes such as the SWI/SNF complex\textsuperscript{85}, suggesting that a mechanistic coupling between chromatin modification and splicing also exists in *S. pombe*. In addition, recent systematic genome-wide yeast-two-hybrid interaction mapping strategies have correctly identified a handful of *S. pombe* genes as factors in the splicing pathway based on physical interactions with known spliceosome components\textsuperscript{82,137}. These high-throughput genetic and physical interaction strategies can yield a wealth of information and strongly hint at a gene’s involvement in the splicing pathway, but they do not provide a direct functional test of a factor’s impact on splicing.

We have previously described a reverse-genetic screening methodology in *S. cerevisiae* that couples high-throughput sample processing with quantitative RT-PCR to enable direct measurements of the splicing efficiency of endogenous pre-mRNA transcripts in the background of thousands of mutant strains\textsuperscript{75}. In addition to
identifying the majority of known splicing mutants, this work successfully identified splicing defects associated with components of the SWI/SNF complex as well as with components of the cleavage and polyadenylation machinery, confirming both the sensitivity of this approach and the evolutionarily conserved nature of these functional interactions. Here we present the results of a study designed to identify non-essential genes in the *S. pombe* genome whose deletion impacts the splicing efficiency of endogenous transcripts. We have developed and implemented a sequencing-based approach for monitoring splicing efficiency in the background of thousands of *S. pombe* strains, and describe the functional significance of those mutants identified.

### 2.3 Materials and Methods

#### 2.3.1 Strains and Cell Growth

All strains examined here were from the haploid deletion library from Bioneer\textsuperscript{138}, representing 3020 individual gene deletions. All strains were grown in supplemented rich growth medium (YES) at 32°C, according to standard procedures\textsuperscript{139}. Strains were recovered from glycerol stocks on solid media supplemented with 200 µg/mL G418. A manual pinning tool (V&P Scientific, cat.#: VP384FP6) was used to transfer cells from solid media into 384-well microtiter plates (Greiner BioOne, cat.#: 781271) for growth in liquid media. Liquid cultures were grown in an Infors HT Multitron plate shaker at 900 rpm with 80% constant humidity. Breathable adhesive tape (VWR, cat.#: 60941-086) was used to seal the plates and reduce evaporation. Because the growth rates of the strains being used varied substantially, an approach was developed to
enable the collection of a similar number of actively growing cells for every strain. Initial cultures (150 µL) of all strains were grown in microtiter plates for two days, allowing nearly all strains to reach saturation. The cell density for most strains is similar at saturation, allowing us to effectively ‘normalize’ cell numbers. Using a liquid handling robot (Biomek NX), 2 µL of saturated culture was used to inoculate 150 µL of fresh media in duplicate to create biological replicate cultures for each strain. After inoculation, cells were allowed to grow for eight hours, allowing most strains to reach A_{600} values near 0.5. Cells were harvested by centrifugation at 5000xg for 5 min, and pellets were flash frozen in liquid N₂ and stored at -80°C until further processing.

2.3.2 cDNA Synthesis and Library Preparation

RNA was isolated from cell pellets and cDNA was synthesized using random ninemers for primers, as previously described. The resulting cDNA was amplified by two sequential PCR reactions to generate products compatible with the Illumina HiSeq2000 Flow Cell as follows. For each cDNA sample, a 12 µL PCR reaction was prepared containing 1x Phusion HF buffer (New England Biolabs), 1x Phusion enzyme, 250 nM forward primer with custom plate-specific barcodes, 250 nM reverse primer, and 1% of the cDNA reaction. The plate specific barcodes sequences were designed as previously described. A complete list of the primers used in this study is available in Table 2.1. Cycling conditions for this first PCR reaction were as follows: 95°C for 3 min, then an empirically determined number of cycles of 98°C for
15 s, 62°C for 20 s, 72°C for 30 s. The number of amplification cycles required was determined in a separate QPCR reaction as the minimum number of cycles necessary to generate a detectable fluorescence signal; required cycle numbers varied from 18 to 21 for the different primers and plates. The products resulting from this first PCR contained plate-specific barcodes, but no well-specific barcodes (see Figure 2.1). For each target, the products from each plate of this first PCR reaction were pooled into a single 384-well microtiter plate, and 0.5 µL was used to seed a second PCR reaction (15 µL), during which well-specific Illumina-Nextera barcodes and Illumina Flow Cell binding sites were appended. This reaction contained 1x Phusion HF buffer (New England Biolabs), 1x Phusion enzyme, 200 nM forward Nextera index primer, and 200 nM Nextera reverse index primer. Cycling conditions were as follows: 95°C for 3 min, then five cycles of 98°C for 15 s, 68°C for 60 s. The PCR products were pooled, concentrated via ethanol precipitation, purified using glass fiber spin columns (Zymo Research), and separated on a 6% native acrylamide gel. Materials of the expected molecular weight ranges were excised from the gel and recovered by soaking crushed gel bits in 0.3 M sodium acetate followed by ethanol precipitation. The resulting DNA precipitate was dissolved in 25 µL water and sequenced on the Illumina HiSeq2000 with the assistance of the Cornell University Biotechnology Resource Center.
<table>
<thead>
<tr>
<th>Description</th>
<th>Primer sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 1</td>
<td>tcgctggcagctgagcagatgtgatgtgataagacacagACAAGCTAGGAAGGAATTGACCATAATTTGATC</td>
</tr>
<tr>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 2</td>
<td>tcgctggcagctgagcagatgtgatgtgataagacacagNAACATTGGGAAGGAATTGACCATAATTTGATC</td>
</tr>
<tr>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 3</td>
<td>tcgctggcagctgagcagatgtgatgtgataagacacagNNACATTGGGAAGGAATTGACCATAATTTGATC</td>
</tr>
<tr>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 4</td>
<td>tcgctggcagctgagcagatgtgatgtgataagacacagNNNACATTGGGAAGGAATTGACCATAATTTGATC</td>
</tr>
<tr>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 5</td>
<td>tcgctggcagctgagcagatgtgatgtgataagacacagNNNCATTGGGAAGGAATTGACCATAATTTGATC</td>
</tr>
<tr>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 6</td>
<td>tcgctggcagctgagcagatgtgatgtgataagacacagNNNNACATTGGGAAGGAATTGACCATAATTTGATC</td>
</tr>
<tr>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 7</td>
<td>tcgctggcagctgagcagatgtgatgtgataagacacagNNNNNACATTGGGAAGGAATTGACCATAATTTGATC</td>
</tr>
<tr>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 8</td>
<td>tcgctggcagctgagcagatgtgatgtgataagacacagAAGTACAAGGAAGGAATTGACCATAATTTGATC</td>
</tr>
<tr>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 9</td>
<td>tcgctggcagctgagcagatgtgatgtgataagacacagCTGTAGCCCGGAAGGAATTGACCATAATTTGATC</td>
</tr>
<tr>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 10</td>
<td>tcgctggcagctgagcagatgtgatgtgataagacacagNAGTACAAGGAAGGAATTGACCATAATTTGATC</td>
</tr>
<tr>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 11</td>
<td>tcgctggcagctgagcagatgtgatgtgataagacacagNCCATCAAGTGGAAGGAATTGACCATAATTTGATC</td>
</tr>
<tr>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 12</td>
<td>tcgctggcagctgagcagatgtgatgtgataagacacagNNNAGTGGTCAAGGAAGGAATTGACCATAATTTGATC</td>
</tr>
<tr>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 13</td>
<td>tcgctggcagctgagcagatgtgatgtgataagacacagAACACCAGGAAGGAATTGACCATAATTTGATC</td>
</tr>
<tr>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 14</td>
<td>tcgctggcagctgagcagatgtgatgtgataagacacagNAACCGAGAAGGAAGGAATTGACCATAATTTGATC</td>
</tr>
<tr>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 15</td>
<td>tcgctggcagctgagcagatgtgatgtgataagacacagNNAACGCTTABAGGAAGGAATTGACCATAATTTGATC</td>
</tr>
<tr>
<td>Sequence</td>
<td>Description</td>
</tr>
<tr>
<td>----------</td>
<td>-------------</td>
</tr>
<tr>
<td>tcgtcgccgcgtcagatgtgtgtaagagacagNNNAAAGACG GAGGAGGAATTGACCATATTAAAGTC</td>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 16</td>
</tr>
<tr>
<td>tcgtcgccgcgtcagatgtgtaagagacagAAAGGTACAG GAGGAGGAATTGACCATATTAAAGTC</td>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 17</td>
</tr>
<tr>
<td>tcgtcgccgcgtcagatgtgtaagagacagNACACAGAA GGAAGGAAATTGACCATATTAAAGTC</td>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 18</td>
</tr>
<tr>
<td>tcgtcgccgcgtcagatgtgtaagagacagNNACAGCAG AGGAGGAAATTGACCATATTAAAGTC</td>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 19</td>
</tr>
<tr>
<td>tcgtcgccgcgtcagatgtgtaagagacagNACCTCCCA GGAAGGAAATTGACCATATTAAAGTC</td>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 20</td>
</tr>
<tr>
<td>tcgtcgccgcgtcagatgtgtaagagacagACGCCTGAG GAGGAGGAATTGACCATATTAAAGTC</td>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 21</td>
</tr>
<tr>
<td>tcgtcgccgcgtcagatgtgtaagagacagNAGGTATCA CCAAGGTTTAAGGACTTCAAATCAC</td>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 22</td>
</tr>
<tr>
<td>tcgtcgccgcgtcagatgtgtaagagacagNAAACATCG CCAAGGTTTAAGGACTTCAAATCAC</td>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 23</td>
</tr>
<tr>
<td>tcgtcgccgcgtcagatgtgtaagagacagNNACATTGG CCAAGGTTTAAGGACTTCAAATCAC</td>
<td>fet5, Illumina forward, flanking intron, antisense, barcode 24</td>
</tr>
<tr>
<td>tcgtcgccgcgtcagatgtgtaagagacagNNNAGAGTC CCAAGGTTTAAGGACTTCAAATCAC</td>
<td>fet5, Illumina reverse, flanking intron, sense</td>
</tr>
<tr>
<td>tcgtcgccgcgtcagatgtgtaagagacagACCCAAAGT TTCACGAACCTAATTC</td>
<td>srrm1, Illumina forward, exon3, antisense, barcode 1</td>
</tr>
<tr>
<td>tcgtcgccgcgtcagatgtgtaagagacagAAAGCTAC CCAAGGTTTAAGGACTTCAAATCAC</td>
<td>srrm1, Illumina forward, exon3, antisense, barcode 2</td>
</tr>
<tr>
<td>tcgtcgccgcgtcagatgtgtaagagacagNAAACATCG CCAAGGTTTAAGGACTTCAAATCAC</td>
<td>srrm1, Illumina forward, exon3, antisense, barcode 3</td>
</tr>
<tr>
<td>tcgtcgccgcgtcagatgtgtaagagacagNNACATTGG CCAAGGTTTAAGGACTTCAAATCAC</td>
<td>srrm1, Illumina forward, exon3, antisense, barcode 4</td>
</tr>
<tr>
<td>tcgtcgccgcgtcagatgtgtaagagacagNNNAGAGTC CCAAGGTTTAAGGACTTCAAATCAC</td>
<td>srrm1, Illumina forward, exon3, antisense, barcode 5</td>
</tr>
<tr>
<td>tcgtcgccgcgtcagatgtgtaagagacagACCGTGATC CCAAGGTTTAAGGACTTCAAATCAC</td>
<td>srrm1, Illumina forward, exon3, antisense, barcode 6</td>
</tr>
<tr>
<td>tcgtcgccgcgtcagatgtgtaagagacagNNACAGATCG CCAAGGTTTAAGGACTTCAAATCAC</td>
<td>srrm1, Illumina forward, exon3, antisense, barcode 7</td>
</tr>
<tr>
<td>tcgtcgccgcgtcagatgtgtaagagacagNNATGCTT AACCAGGTGTTAAAGGACTTCAAATCAC</td>
<td>srrm1, Illumina forward, exon3, antisense, barcode 8</td>
</tr>
<tr>
<td>Primer</td>
<td>Sense/Antisense</td>
</tr>
<tr>
<td>--------</td>
<td>----------------</td>
</tr>
<tr>
<td>srrm1, Illumina forward, exon3, antisense, barcode 9</td>
<td>tcgtggcagegtcagatgtgtaagacagCTGTAGCCC CAAGGTTTAAGGACTTCAATATTCA</td>
</tr>
<tr>
<td>srrm1, Illumina forward, exon3, antisense, barcode 10</td>
<td>tcgtggcagegtcagatgtgtaagacagsNGTACAAG CCAAGGTTTAAGGACTTCAATATTCA</td>
</tr>
<tr>
<td>srrm1, Illumina forward, exon3, antisense, barcode 11</td>
<td>tcgtggcagegtcagatgtgtaagacagNNCATCAAG TCCAAGGTTTAAGGACTTCAATATTCA</td>
</tr>
<tr>
<td>srrm1, Illumina forward, exon3, antisense, barcode 12</td>
<td>tcgtggcagegtcagatgtgtaagacagNNNAGTGTC CACCAAGGTTTAAGGACTTCAATATTCA</td>
</tr>
<tr>
<td>srrm1, Illumina forward, exon3, antisense, barcode 13</td>
<td>tcgtggcagegtcagatgtgtaagacagAAACAACAC CAAGGTTTAAGGACTTCAATATTCA</td>
</tr>
<tr>
<td>srrm1, Illumina forward, exon3, antisense, barcode 14</td>
<td>tcgtggcagegtcagatgtgtaagacagNAACCGAGA CCAAGGTTTAAGGACTTCAATATTCA</td>
</tr>
<tr>
<td>srrm1, Illumina forward, exon3, antisense, barcode 15</td>
<td>tcgtggcagegtcagatgtgtaagacagNNAACGCTT ACCAAGGTTTAAGGACTTCAATATTCA</td>
</tr>
<tr>
<td>srrm1, Illumina forward, exon3, antisense, barcode 16</td>
<td>tcgtggcagegtcagatgtgtaagacagAAGGTACAC CAAGGTTTAAGGACTTCAATATTCA</td>
</tr>
<tr>
<td>srrm1, Illumina forward, exon3, antisense, barcode 17</td>
<td>tcgtggcagegtcagatgtgtaagacagNACACAGAA CCAAGGTTTAAGGACTTCAATATTCA</td>
</tr>
<tr>
<td>srrm1, Illumina forward, exon3, antisense, barcode 18</td>
<td>tcgtggcagegtcagatgtgtaagacagNNACAGCTC ACCAAGGTTTAAGGACTTCAATATTCA</td>
</tr>
<tr>
<td>srrm1, Illumina forward, exon3, antisense, barcode 19</td>
<td>tcgtggcagegtcagatgtgtaagacagNNNACCTCC ACCAAGGTTTAAGGACTTCAATATTCA</td>
</tr>
<tr>
<td>srrm1, Illumina forward, exon3, antisense, barcode 22</td>
<td>tcgtggcagegtcagatgtgtaagacagACGCTCGAC CAAGGTTTAAGGACTTCAATATTCA</td>
</tr>
<tr>
<td>srrm1, Illumina forward, exon3, antisense, barcode 23</td>
<td>tcgtggcagegtcagatgtgtaagacagNNAGTGTCACC ACGGTTTAAGGACTTCAATATTCA</td>
</tr>
<tr>
<td>srrm1, Illumina forward, exon3, antisense, barcode 24</td>
<td>tcgtggcagegtcagatgtgtaagacagNNNAGTGTC ACCAAGGTTTAAGGACTTCAATATTCA</td>
</tr>
<tr>
<td>srrm1, Illumina reverse, exon1, sense</td>
<td>gtcctgcggagtgcagatgtgtaagacagTTATTGCTG CTTTCTCATTTC</td>
</tr>
</tbody>
</table>

**Table 2.1 Primer sequences used.** The listed sense and antisense primers were used for the first PCR reaction in which the splicing event is amplified with a plate specific barcode. The second PCR reaction used primer sequences of the "Nextera XT index kit v2" available through Illumina.
2.3.3 Data Processing

Reads were demultiplexed using a combination of Nextera-specific indices and custom plate-specific barcodes within the insert read. The bwa-mem aligner was then used to align reads to a custom index containing both the spliced and unspliced isoforms of the two targets. A splice-index (SI) was calculated for \( \text{fet5\_intron1} \) and \( \text{pwi1\_intron2} \) in each sample by comparing the number of reads mapped to the unspliced isoform versus the spliced isoform as follows:

\[
SI = \frac{\text{unspliced read count}}{\text{spliced read count}}
\]

**Equation 2.1**

To determine SI relative to wild type while accounting for plate to plate variation, we assumed that the median SI within each 384-well plate was representative of wild-type. Therefore, the relative SI was calculated as:

\[
SI_{\text{relative}} = \frac{SI}{SI_{\text{PlateMedian}}}
\]

**Equation 2.2**

After determining the SI for each biological replicate, we filtered our dataset to include only those samples for which the standard deviation between the \( \log_2(SI_{\text{relative}}) \) was less than 1, and for which the combined read count was greater than 1000. A total of 3007 and 3005 strains (99.6% and 99.5%) passed these quality scores for the \( \text{fet5\_intron1} \) and \( \text{pwi1\_intron2} \) datasets, respectively. In order to determine the subset of strains which exhibit a \( \log_2(SI_{\text{relative}}) \) that was statistically significantly different
from wild-type, we considered how the precision of our $SI_{\text{rel}}$ measurements varied as a function of read count. In concept, this approach has similarities to algorithms commonly used for RNA-seq analysis which empirically estimate noise within a dataset to identify significant changes in gene-expression \(^{48}\). The log\(_2\)-transformed $SI_{\text{relative}}$ values were plotted as a function of log-transformed read count for each sample (see Figure 2.4). The dataset was then divided into 20 equal sized bins based on read count. Using the mean and standard deviation within the bins as data points, spline interpolation was used to estimate the log\(_2\)-transformed mean ($\mu_{\text{interpolated}}$) and standard deviation ($\sigma_{\text{interpolated}}$) of $SI_{\text{rel}}$ measurements at any given read count under the null hypothesis. For each strain, a $p$-value was then estimated by defining a $Z$-score as follows:

$$Z = \frac{\log_2(SI_{\text{relative}}) - \mu_{\text{interpolated}}}{\sigma_{\text{interpolated}}}$$

**Equation 2.3**

Strains were called as significant if the Benjamini-Hochberg corrected $p$-value was below 0.05. The 95% confidence intervals in Figure 3 represent $\log_2(SI_{\text{relative}}) \pm 2\sigma_{\text{interpolated}}$ given the read depth of that strain.

### 2.3.4 Splicing Sensitive Microarrays

All microarrays were performed as two-color arrays comparing mutant and wild type strains, each grown under identical conditions. Briefly, strains were grown to saturation at 30°C, then back-diluted in 25 mL cultures and allowed to grow at 30°C
until they reached an optical density of $A_{600} \approx 0.5$. Total cellular RNA samples were isolated, converted into cDNA, fluorescently labeled, and hybridized to the array as previously described$^{143}$. Biological replicate microarrays were performed for most mutant strains, with average expression measurements between biological replicates being presented in the figures. Both raw and processed microarray data are available through GEO using accession number GSE79153.

2.4 Results and Discussion

Here we sought to identify the full complement of non-essential genes that impact pre-mRNA splicing efficiency in *S. pombe*, an organism whose splicing properties closely resemble those seen in higher eukaryotes$^{131}$. To quantitatively measure the impact of mutations on the splicing pathway, we designed an assay that would allow for high-sensitivity detection of both spliced and unspliced isoforms in thousands of unique samples (see Figure 2.1).
Figure 2.1 Schematic of workflow for quantitatively measuring splicing in the fission yeast deletion collection. After cell growth, RNA isolation, and cDNA synthesis with random primers, consecutive PCR reactions are performed using primers that flank an intron to amplify both spliced and unspliced RNA while appending sample specific barcodes and Illumina compatible ends. Estimates of splicing efficiency in each strain are determined by counting the number of spliced and unspliced sequencing reads derived from each sample.
Briefly, cDNA from a given sample was used as template for a PCR reaction using primers that flank a splicing event, enabling amplification of both spliced and unspliced isoforms. By appending appropriately barcoded sequences, the resulting material was subjected to deep sequencing to count the number of molecules corresponding to both the spliced and unspliced isoforms for each sample. To demonstrate that this approach could provide a quantitative representation of the underlying species, we measured isoform ratios for samples that contained known ratios of different spliced isoforms. Across a large range of relative isoform abundances, this sequencing-based approach gave results that were both highly accurate and precise (Figure 2.2).
Figure 2.2 RT-PCR amplicon sequencing accurately measures splicing of in RNA mixtures containing different relative isoform abundances. Total cellular RNA was collected from two strains, each carrying a different splice isoform of the pwi1 gene ORF in place of the native pwi1 gene. Isoforms A and B differed by a length of 90 nucleotides. RNA from each strain was mixed in different proportions, and the relative abundance of each isoform was measured in each mixture by RT-PCR amplicon sequencing. Error bars
represent the standard deviation between three technical replicate RT-PCR reactions.

After determining that this sequencing approach was sensitive and quantitative, we turned to examining each of the ~3000 deletion strains available within the *S. pombe* haploid deletion collection\textsuperscript{138} to identify novel factors whose disruption impacts splicing. Primers were designed that would allow for the determination of the splicing efficiency of two introns: the single intron in the *fet5* transcript, a predicted GTPase involved in RNA polymerase localization, and the second intron in the *pwi1* transcript, a splicing co-activator. The *fet5* intron resembles a typical intron in *S. pombe*, in that the *fet5* transcript contains just a single intron with canonical 5’ splice site (GUAAGU), and branch point (UGCUAUU) sequences, and whose length (45 nt) is close to the median intron length (56 nt). The second intron in *pwi1* is also of typical length (59 nt) for an *S. pombe* intron, and has a typical branch point sequence (CAUUUAU) but has an atypical 5’ splice site sequence (GUACAA) which significantly deviates from the canonical sequence. Importantly, because these two introns are short, the PCR amplification efficiency of both the spliced and unspliced isoforms should be similar, reducing the likelihood of artifacts derived from amplification bias. In total, ~12,000 samples were generated, corresponding to each of these targets within each of these strains with biological replicates. As a convenient measure of splicing efficiency, we define the splice-index (SI) as the ratio of unspliced to spliced reads and looked to identify mutants that caused significant changes to the SI. Importantly, because this assay measures the steady state abundances of specific
RNA species, a high SI could indicate a defect in pre-mRNA splicing, or alternatively, a change in the relative stabilities of spliced or unspliced RNA. For both the \textit{fet5\_intron1} and \textit{pwi1\_intron2}, the unspliced pre-mRNA was present at about 2% of the spliced mRNA in the background of most strains, with respective median SI values of 0.018 and 0.025 (Figures 2.3A and B), consistent with the expectation that splicing occurs efficiently and the vast majority of these transcripts are present as the spliced isoform. Moreover, for the vast majority of strains, the measured SI was relatively close to the median value, with interquartile ranges across all samples of 0.004 and 0.011 for the \textit{fet5\_intron1} and \textit{pwi1\_intron2} targets, respectively (Figure 2.3A and B), consistent with the expectation that most genes examined here do not impact the splicing pathway. Across all strains, the biological replicates were correlated with $R^2$ values of 0.37 and 0.23 for \textit{fet5\_intron1} and \textit{pwi1\_intron2}, respectively (Figure 2.3C and D).
Figure 2.3 Splice index measurements for *fet5-intron1* and *pwi1-intron2* are reproducible. Histograms and box-whisker plots of the measured splice index for *fet5-intron1* (A) and *pwi1-intron2* (B) reveal that we measure a relatively consistent splice index in all samples. The splice index relative to the median was calculated for each biological replicate for each deletion strain and replicate scatter plots (C, D) show that the relative splice index measurements are largely reproducible among the 3000+ strains in the strain collection.
As with any RNA-sequencing experiment, the statistical power to identify changes in expression increases with greater read depth. In order to identify the subset of strains that exhibited a significant change in splicing, we developed a statistical test that assessed the observed change in SI as a function of read depth (see Materials and Methods, Figure 2.4).
Figure 2.4. Relative splice index measurements in deletion strains. The relative splice index for \textit{fet5\_intron1} (A) or \textit{pwi\_intron2} (B) is plotted as a function of read count for each of the ~3000 strains examined. Strains which significantly differed from wild-type after multiple hypothesis correction are colored red and labeled. A total of 61 strains were identified as having a significantly different splice index measurement for either \textit{fet5\_intron1} or \textit{pwi\_intron2}.
Using this approach, statistically significant changes in SI were identified for 57 and 18 deletion strains for the *fet5_intron1* and *pwi1_intron2*, respectively (see Figure 2.5). Importantly, of the 18 strains that affected splicing *pwi1_intron2*, 14 were also found to affect splicing of *fet5_intron1*. This significant degree of overlap ($p<3.81\times10^{-22}$, Fisher’s exact test) suggests that the splicing defect observed in many of the strains is not specific to a single gene.
Figure 2.5. Gene deletions which result in significant splice index changes in either fet5_intron1 or pwi1_intron2. The measured relative splice index is shown for fet5_intron1 and pwi1_intron2 with 95% confidence intervals for the 61 gene deletion strains which were significantly different than wild-type for at least one of splicing events examined. Notable Gene Ontology (GO) categories are indicated.
To better understand the functional significance of the genes identified through this screen, we asked whether there was enrichment for factors involved in similar pathways by analyzing their Gene Ontology (GO)\(^{144,145}\). Appropriately, the most highly enriched biological process identified was ‘mRNA splicing, via spliceosome’ \((p<6.63\times10^{-3}, \text{Table S3 in } \text{Larson et al.}^{140})\), confirming the ability of the method to positively identify known splicing factors. Consistent with our previous results in \(S. \text{cerevisiae}\), not all deletions of known splicing factors resulted in a measurable change in splicing efficiency of either of the tested introns. Although these might represent false negative discoveries, on the basis of our experience in \(S. \text{cerevisiae}\) we expect the more likely explanation is that these factors are not strictly required for efficient splicing of these specific introns under the conditions tested. Interestingly, significant overrepresentation of components of the SWR1 nucleosome remodeling complex was also uncovered \((p<6.75\times10^{-3})\), consistent with previous reports describing the role of SWR1 components in splicing\(^{146}\). Other GO categories that are well represented in the list of significant genes include ‘transcription from polymerase II promoter’, ‘mRNA cleavage and polyadenylation specificity factor complex’, and ‘chromatin remodeling’ (Figure 2.5 and Table S3 in \(\text{Larson et al.}^{140}\)).

Several of the genes identified here belong to seemingly unrelated GO categories. It seems important to reiterate that the approach implemented here doesn’t measure splicing defects \textit{per se}, but rather changes in the relative steady state levels of spliced
and unspliced isoforms. As such, while some of these candidates may represent false positive discoveries, it seems likely that many are true positives which impact splicing isoform abundances either through non-splicing related pathways, or by modulating the activity of *bona fide* splicing factors. For example, deletions of either *ski2* or *trs130* resulted in some of the most significant increases in SI for either of the tested splicing events. Ski2 is an RNA helicase and member of the SKI complex, a highly conserved complex necessary for 3’ to 5’ degradation of transcripts subject to the nonsense-mediated decay (NMD) pathway. The unspliced isoforms of *fet5* and *pwi1* contain premature stop codons and would be predicted targets of the NMD pathway, providing a plausible explanation for their accumulation in the Δ*ski2* strain. The *trs130* gene, by contrast, is involved in vesicle transport from the endoplasmic reticulum; the mechanism by which it might relate to altered splice isoform abundances is less clear. While no physical interactions have been documented between Trs130 and splicing-related proteins, epistatic genetic interactions between *trs130* and essential *bona-fide* splicing factors have been documented.

To better understand the evolutionary nature of the genes that we identified, we examined each of them to determine whether homologs could be identified in either *S. cerevisiae* or humans. Of the 61 candidates we identified, 17 have clear human homologs but appear to lack an *S. cerevisiae* homolog (Table 2.2), underscoring the potential that *S. pombe* provides as a model system for understanding the complex splicing seen in mammalian systems. Four of these genes, *cay1*, *cwf18*, *cwf19*, and
*pwi1* have previously been annotated as splicing factors, while two others, *SPAC20H4.06c* and *SPBC713.05*, have been implicated in the splicing pathway on the basis of homology to human counterparts. Here, we provide experimental evidence that these protein products functionally impact pre-mRNA splicing.
<table>
<thead>
<tr>
<th>Candidate</th>
<th>S. cerevisiae homolog(s)</th>
<th>H. saplens homolog</th>
</tr>
</thead>
<tbody>
<tr>
<td>ash2</td>
<td>BRE2</td>
<td>ASH2L</td>
</tr>
<tr>
<td>cay1</td>
<td>--</td>
<td>CACTIN</td>
</tr>
<tr>
<td>csx2</td>
<td>--</td>
<td>ACAP1</td>
</tr>
<tr>
<td>ctr1</td>
<td>--</td>
<td>CCDC174</td>
</tr>
<tr>
<td>cwf18</td>
<td>--</td>
<td>CCDC12</td>
</tr>
<tr>
<td>cwf19</td>
<td>--</td>
<td>CWF19L2/L1</td>
</tr>
<tr>
<td>est1</td>
<td>EST1</td>
<td>SMG6</td>
</tr>
<tr>
<td>hba2</td>
<td>PDR5/10/12/15/18</td>
<td>--</td>
</tr>
<tr>
<td>hns2</td>
<td>--</td>
<td>EIF2AK2/3</td>
</tr>
<tr>
<td>iex2</td>
<td>IES2</td>
<td>INO80B</td>
</tr>
<tr>
<td>imp1</td>
<td>SRP1</td>
<td>KPNA1/5/6</td>
</tr>
<tr>
<td>lem2</td>
<td>LEMD2</td>
<td>SRC1, HEH2</td>
</tr>
<tr>
<td>lsg1</td>
<td>CTK3</td>
<td>--</td>
</tr>
<tr>
<td>med20</td>
<td>SRB2</td>
<td>MED20</td>
</tr>
<tr>
<td>meu17</td>
<td>SGA1</td>
<td>--</td>
</tr>
<tr>
<td>mug154</td>
<td>NUR1</td>
<td>--</td>
</tr>
<tr>
<td>mug42</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>mug78</td>
<td>ATG13</td>
<td>ATG13</td>
</tr>
<tr>
<td>mug80</td>
<td>CLG1</td>
<td>--</td>
</tr>
<tr>
<td>not3</td>
<td>NOT3/5</td>
<td>CNOT3</td>
</tr>
<tr>
<td>nup170</td>
<td>NUP157, NUP170</td>
<td>NUP155</td>
</tr>
<tr>
<td>nup37</td>
<td>--</td>
<td>NUP37</td>
</tr>
<tr>
<td>nxt3</td>
<td>BRE5</td>
<td>G3BP1</td>
</tr>
<tr>
<td>oma4</td>
<td>PMT4</td>
<td>POMT1</td>
</tr>
<tr>
<td>pds5</td>
<td>PDS5</td>
<td>PDS5B</td>
</tr>
<tr>
<td>pht1</td>
<td>HTZ1</td>
<td>H2AFZ, H2AFV</td>
</tr>
<tr>
<td>png2</td>
<td>YNG1/2, PHO23</td>
<td>ING2/3/4/5</td>
</tr>
<tr>
<td>pof4</td>
<td>EKA1</td>
<td>TCEB3, TCEB3B</td>
</tr>
<tr>
<td>ppk29</td>
<td>PRK1, ARK1, AKL1</td>
<td>BMP2K, AAK1</td>
</tr>
<tr>
<td>ppp1</td>
<td>--</td>
<td>PPP1R10</td>
</tr>
<tr>
<td>prx</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>pwi1</td>
<td>--</td>
<td>SRRM1</td>
</tr>
<tr>
<td>red1</td>
<td>--</td>
<td>ZFC3H1</td>
</tr>
<tr>
<td>rhp61</td>
<td>RAD51</td>
<td>RAD51</td>
</tr>
<tr>
<td>mh201</td>
<td>RNH201</td>
<td>RNASEH2A</td>
</tr>
<tr>
<td>rpb9</td>
<td>RPB9</td>
<td>POLR21</td>
</tr>
<tr>
<td>rpl2102</td>
<td>RPL21A,B</td>
<td>RPL21</td>
</tr>
<tr>
<td>rpl4301</td>
<td>RPL43A,B</td>
<td>RPL37A</td>
</tr>
<tr>
<td>Screen Hit</td>
<td><em>S. cerevisiae</em> homolog(s)</td>
<td><em>H. sapiens</em> homolog</td>
</tr>
<tr>
<td>------------</td>
<td>----------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>rtf2</td>
<td>--</td>
<td>RTFDC1</td>
</tr>
<tr>
<td>rtt109</td>
<td>--</td>
<td>RTT109</td>
</tr>
<tr>
<td>sdu1</td>
<td>--</td>
<td>DES2</td>
</tr>
<tr>
<td>smd3</td>
<td>SMD3</td>
<td>SNRPD3</td>
</tr>
<tr>
<td>SPAC18G6.13</td>
<td>--</td>
<td>GPATCH1</td>
</tr>
<tr>
<td>SPAC20H4.06c</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>SPAC22H10.11c</td>
<td>IFH1, CRF1</td>
<td>SEL1L, SEL1L2, SEL1L3</td>
</tr>
<tr>
<td>SPBC1348.03</td>
<td>--</td>
<td>WRD83</td>
</tr>
<tr>
<td>SPBC28F2.08c</td>
<td>HRD3</td>
<td>WDR4</td>
</tr>
<tr>
<td>SPBC713.05</td>
<td>--</td>
<td>SKIV2L</td>
</tr>
<tr>
<td>SPCC18.13</td>
<td>TRM82</td>
<td>URI1</td>
</tr>
<tr>
<td>SPCC550.03c</td>
<td>SKI2</td>
<td>PARK7</td>
</tr>
<tr>
<td>SPCC736.07c</td>
<td>BUD27</td>
<td>DNAJC8</td>
</tr>
<tr>
<td>sdj1</td>
<td>--</td>
<td>SSU72</td>
</tr>
<tr>
<td>spf31</td>
<td>--</td>
<td>VPS72</td>
</tr>
<tr>
<td>ssu72</td>
<td>SSU72</td>
<td>VPS72</td>
</tr>
<tr>
<td>swc2</td>
<td>VPS72</td>
<td>TIPIN</td>
</tr>
<tr>
<td>swi3</td>
<td>CSm3</td>
<td></td>
</tr>
<tr>
<td>toc1</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>trs130</td>
<td>TRS130</td>
<td>TRAPPC10</td>
</tr>
<tr>
<td>ubr1</td>
<td>UBR1, UBR2</td>
<td>UBR1, UBR2</td>
</tr>
<tr>
<td>vps71</td>
<td>VPS71</td>
<td>ZNHT1</td>
</tr>
<tr>
<td>yaf9</td>
<td>YAF9</td>
<td>MLLT3, YEATS4</td>
</tr>
</tbody>
</table>

Table 2.2 Many candidates identified in our screen have no apparent *S. cerevisiae* homolog but have an apparent homolog in humans. Listed here are the 61 candidates identified in our screen with their corresponding homolog(s) in *S. cerevisiae* and in humans. The red asterisks indicate the 17 genes that have no homolog in *S. cerevisiae* but do in humans. Of those 17 genes, the green asterisks indicate those genes that are known splicing factors. The blue asterisks indicate those genes that were predicted to be involved in splicing based on homology and have now been confirmed in this study.
2.4.1 Known Splicing Factors Identified Here Display Global Splicing Defects

To better understand the impact of the genes identified here, splicing sensitive microarrays were used to determine the global changes in pre-mRNA splicing that result from their deletions. These microarrays contain probes that target an exonic region of every protein coding gene in the *S. pombe* genome, as well as probes targeting every intron and its corresponding exon-exon junction, allowing for measurements of changes in total expression, pre-mRNA, and mature mRNA levels, respectively (Figure 2.6A). As an initial test, we chose to examine strains harboring deletions in three known splicing factors: (1) *smd3*, a core component of the SM complex in the U1, U2, U4, and U5 snRNPs; (2) *aar2*, a component of the U5 snRNP; and (3) *pwi1*, a splicing co-activator. For both *smd3* and *aar2*, clear homologs exist in both *S. cerevisiae* and humans, and their specific roles in the splicing pathways have been well characterized\(^{39,149,150}\). Moreover, the *Δsmd3* strain showed a statistically significant increase in our screen for both the *fet5*_intron1 and the *pwi1*_intron2 pre-mRNAs, while the *Δaar2* strain showed increased pre-mRNA levels for both transcripts, albeit just below our cutoff for statistical significance. The third gene, *pwi1*, is the homolog of the human SRRM1 gene, a member of the SR-like family of proteins\(^{56}\). Unlike *smd3* and *aar2*, there is no homolog of *pwi1* in the *S. cerevisiae* genome. In our screen data, deletion of *pwi1* caused a statistically significant increase in the SI of the *fet5*_intron1. Deletion of any of these three genes resulted in global defects in pre-mRNA splicing, albeit each with unique properties (Figure 2.6B). For each of the mutants, increased levels of pre-mRNAs were detected for a majority of...
the events observed, and concomitant decreases were seen for many of the mature mRNA species, consistent with our expectations for a bona fide splicing mutant. Furthermore, similar to our screen data, the Δaar2 strain showed levels of pre-mRNA accumulation that were overall lower than in the Δsmd3 strain. Nevertheless, a similar number of splicing events was impaired by all three deletions.
Figure 2.6 Known and predicted splicing factors display global splicing defects. (A) Splicing sensitive microarrays contain probes for quantification of total (T), pre-mRNA (P), and mature (M) mRNA levels. (B) Deletion of known splicing factors \textit{smd3}, \textit{aar2}, and \textit{pwi1} each display global splicing defects. Each row represents the relative measurements for total, pre-mRNA, and mature mRNA for a particular splicing event. Numbers below each column represent the median value within the column. Rows from each sample are independently sorted by hierarchical clustering and displayed for only those events for which data were available for all three probe types. (C) Global splicing phenotypes of the \textit{ΔSPAC20H4.06c} and \textit{ΔSPCC162.01c} strains. The orange and purple bars highlight specific splicing events showing decreases or increases in splicing efficiency, respectively.
2.4.2 Predicted Splicing Factors Also Display Global Splicing Defects

Among the genes we identified in our screen whose deletions negatively impacted the splicing of either \textit{fet5\_intron1} or \textit{pwi1\_intron2} were several that are predicted based on homology studies to be involved in the splicing pathway. We chose to focus on two of these mutants, deletions of \textit{SPAC20H4.06c}, a RNA-binding protein, and \textit{SPCC162.01c}, a putative tri-snRNP component. Deletion of \textit{SPAC20H4.06c} resulted in a statistically significant increase in both \textit{fet5\_intron1} and \textit{pwi1\_intron2} pre-mRNA levels, whereas deletion of \textit{SPCC162.01c} also caused an increase in both pre-mRNAs, although just below our significance cutoff. As with \textit{pwi1}, there are no apparent \textit{S. cerevisiae} homologs for either \textit{SPAC20H4.06c} or \textit{SPCC162.01c}, but apparent human homologs do exist. For \textit{SPAC20H4.06c}, the human homolog is GPATCH1, a member of the G-patch containing family of proteins. G-patch containing proteins have been previously implicated in splicing, yet no direct evidence appears to exist that specifically couples GPATCH1 to splicing. By contrast, the human homolog of \textit{SPCC162.01c} is SNRNP27, a component of the U4/U6·U5 tri-snRNP complex and has a direct role in splicing\textsuperscript{151}. Interestingly, human SNRNP27 was previously shown to contain an N-terminal domain with strong homology to the SR domain of U170K; however, unlike U170K, SNRNP27 lacks an RNA-binding domain.

The global splicing defects of \textit{ΔSPAC20H4.06c} and \textit{ΔSPCC162.01c} revealed remarkably different phenotypes (Figure 2.6C). Deletion of \textit{SPAC20H4.06c} showed a canonical splicing defect with broad increases in pre-mRNA species and decreases in
mature mRNA species. The level to which pre-mRNAs accumulated is similar to that seen upon deletion of the canonical splicing factor *smd3* (Figure 2.5B). These data strongly suggest that the *SPAC20H4.06c* gene product is participating in the splicing pathway. By contrast, the global splicing profile resulting from deletion of *SPCC162.01c* looked quite different from the other splicing mutants examined here. Whereas a subset of splicing events appeared to be negatively affected by deletion of *SPCC162.01c*, as evidenced by the accumulation of pre-mRNA and loss of mature mRNA, a nearly equal number of splicing events seemed to be positively, albeit weakly, affected by its deletion, with pre-mRNA levels decreasing and mature mRNA levels increasing for these transcripts. These results are consistent with a model where SR proteins can function to either enhance or repress splice site activation at different introns. These data also suggest that a large number of *S. pombe* introns are spliced at suboptimal efficiency in wild type cells. Additional experiments will be necessary to understand the mechanistic basis by which this SNRNP27 homolog can impart these phenotypes.

### 2.4.3 Genes involved in heterochromatin formation show a range of genome-wide splicing defects

In examining the list of candidates identified in our screen, we were struck by the number of components with known roles involved in RNA silencing and heterochromatin formation. It has long been known that RNA plays a critical role in silencing in *S. pombe* via the RNA-induced initiation of transcriptional gene silencing
(RITS) complex\(^{152}\). While it has been suggested that splicing components facilitate RITS function\(^{153}\), it remains unclear whether these effects are direct or indirect\(^{154}\). Two groups recently described purifications of two related complexes involved in silencing: MTREC, which is involved in assembling heterochromatin\(^{155}\); and the Nuclear RNA Silencing complex, NURS. While these complexes each contain unique elements, they share in common both the essential RNA helicase Mtl1 and the non-essential zinc-finger protein Red1. In our work, deletion of red1 resulted in a statistically significant decrease in the splicing efficiency of both tested splicing events. In addition, affinity purification of Mtl1 as part of the MTREC complex co-purified Ctr1\(^{155}\), whereas affinity purification of Red1 as part of the NURS complex co-purified SPAC18G6.13\(^{156}\). In our screen, deletions of ctr1 and SPAC18G6.13 both resulted in statistically significant decreases in splicing efficiency of both target pre-mRNAs.

Ctr1 was previously implicated in splicing of TER1, the RNA component of telomerase\(^{155}\). Moreover, Ctr1 had been shown to physically interact with components of the Prp19 complex, including Cwf10, Cwf11, and Prp19. To determine whether Ctr1 had a more global effect on the splicing pathway we again turned to microarray analysis. Deletion of ctr1 resulted in a striking global splicing defect, strongly resembling that of a canonical splicing mutant (Figure 2.7A). A recent RNA-seq analysis of Actr1 and other MTREC mutants also revealed a global increase in unspliced transcript levels\(^{78}\). Interestingly, whereas our data reveal a broad decrease in
mature mRNA concomitant with the increase in unspliced isoform, the Zhou et al\textsuperscript{78}. study reported largely unchanged levels of spliced transcripts. Owing at least in part to this observation, Zhou and colleagues proposed that Ctr1/MTREC plays a role in targeting unspliced transcripts to the nuclear exosome, and that the pre-mRNA accumulation phenotype of the Δctr1 strain did not reflect a direct role for MTREC on splicing. The broad decreases in mature mRNA demonstrated by our microarray experiments are more consistent with a direct role for Ctr1 in the splicing pathway; additional experiments will be necessary to understand the discrepancy between these results, and the functional significance of Ctr1 in the pre-mRNA splicing pathway.

By contrast with Ctr1, far less is known about the relationship between SPAC18G6.13 and splicing. Whereas physical interactions have been described between \textit{SPAC18G6.13} and some splicing factors\textsuperscript{136}, the functional relevance of these interactions has not been previously described. Using microarray analysis, we showed that deletion of \textit{SPAC18G6.13} also resulted in a broad increase in unspliced messages (Figure 2.7B). Interestingly, whereas SPAC18G6.13 was co-purified with Red1 as part of the NURS complex, the same study also demonstrated that Mtl1 co-purifies with SPAC20H4.06c, homolog of the human GPATCH1 gene described in the section above. When the global splicing defects of the Δ\textit{ctr1}, Δ\textit{SPAC18G6.13}, and Δ\textit{SPAC20H4.06c} strains were analyzed together, the overlap in genome-wide splicing patterns was striking (Figure 2.7C). The physical interactions both between these proteins themselves, and with additional components of the spliceosome as observed
by others, as well as the splicing phenotypes we observed here in these mutants suggest that the functional relationship between splicing and heterochromatin formation may be more bi-directional than previously thought (Figure 2.7D).
Figure 2.7 Deletion of factors involved in heterochromatin formation strongly impact global splicing. Splicing sensitive microarrays for δctr1 (A) and δSPAC18G6.13 (B) reveal global splicing defects for each. Splicing events for each mutant were sorted independently using hierarchical clustering and displayed for only those events for which data were available for all three probe types. (C) A comparison of the splicing defects on common targets reveals a large overlap among all three of these deletion strains, with a subset of events highlighted by the orange bar. (D) Known physical interactions between several components of the silencing pathway and the splicing pathway. Red arrows indicate previously published one-way physical interactions. Green arrows indicate two-way interaction. Blue ovals represent splicing factors, while yellow ovals represent members of the NURS and/or MTREC complexes. Black outlines note the components whose deletions caused splicing defects in this study. Previously described physical interactions between known splicing factors and components of the NURS and MTREC complexes, together with our observations that deletion of these components result in large accumulations of unspliced transcripts and decreases in spliced transcripts, suggest that these components may have a more direct role in splicing regulation.
In addition to the NURS complex, heterochromatic silencing is accomplished in part through cooperation between the RNAi machinery and the heterochromatic factors Clr4 and the histone variant H2A.Z\textsuperscript{157–160}. While H2A.Z is generally thought of as a repressive mark, it is also associated with promoters and may play roles in recruiting RNAP II to genes\textsuperscript{161}. Interestingly, among our list of mutants that affected splicing of our targets were Δpht1, the fission yeast homolog of H2A.Z, as well as many components of the INO80/SWR1 complex, which is responsible for catalyzing H2A/H2A.Z exchange\textsuperscript{162}, including Δyaf9, Δies2, Δvps71, and Δswc2. Similarly, while the Set1C complex is responsible for catalyzing the addition of H3K4me marks, it is also necessary for proper silencing of subtelomeric regions in \textit{S. pombe}\textsuperscript{163}. Deletion of two components of the Set1 complex, \textit{ash}2 and \textit{swd}1, were identified in our screen as causing decreases in pre-mRNA splicing efficiency, although the Δ\textit{swd}1 effect was just below our significance cutoff (Table S1 in \textit{Larson} et al.\textsuperscript{140}).

To determine the effect that loss of these heterochromatic factors have on the splicing pathway, we again assessed the global splicing profiles of the Δpht1, Δ\textit{ash}2, and Δ\textit{swd}1 strains by microarray. On the basis of these experiments alone, it is difficult to say whether deletion of any of these factors is impacting pre-mRNA splicing (Figure 2.8). While small groups of transcripts can be seen to exhibit a canonical splicing defect, the large changes in total gene expression, both increases and decreases, that are associated with these mutations complicates their analysis.
Further studies will be necessary to better characterize the impact on pre-mRNA splicing of deletion of these genes.
Figure 2.8 3’ end processing factors affect the splicing of both terminal and not terminal introns equally. Splicing sensitive microarrays of 3’ end processing factors *ppn1* (A) and *ssu72* (B). Only transcripts that contained multiple introns were considered for this analysis. Introns were called ‘terminal’ when they were the last intron within a transcript, and all introns were grouped together as ‘not terminal’. Each heatmap is organized by hierarchical clustering. When comparing the ‘terminal’ to ‘not terminal’ heatmap for both Δ*ppn1* and Δ*ssu72*, splicing defects are observed regardless of intron position.
2.4.4 3’ end processing factors impact the splicing of both terminal and non-terminal introns

Here we identified two genes involved in the cleavage and polyadenylation pathway, *ssu72* and *ppn1*, whose deletions resulted in pre-mRNA splicing defects. The 3’ end processing and splicing pathways have been previously demonstrated to be functionally coupled together\textsuperscript{71,73}. Components of the U2 snRNP co-purify with cleavage and polyadenylation specificity factor, CPSF, demonstrating a physical interaction between the two pre-mRNA processing pathways\textsuperscript{73}. In addition, CPSF is necessary for efficient splicing activity, while binding of the U2 snRNP promotes efficient cleavage at the 3’ end. Importantly, Ppn1 and Ssu72 have been shown to physically interact with each other in *S. pombe* and to co-purify with the 3'end processing machinery\textsuperscript{164}. We determined the global splicing profiles of these two mutants using microarrays: deletion of both *ssu72* and *ppn1* resulted in pre-mRNA splicing defects for a large fraction of the events monitored (Figure 2.9A). The defects seen for these mutants was similar to those seen for deletion of the canonical splicing mutant *smd3*, both in terms of the number of transcripts affected and the levels of pre-mRNA accumulation.
Figure 2.9  Deletions of 3’ end processing factors result in global splicing defects.  Splicing sensitive microarrays for Δppn1 (A) and Δssu72 (B) strains each show broad splicing defects. Splicing events from each array were clustered independently using hierarchical clustering and displayed for only those events for which data were available for all three probe types. (C) The pre-mRNA levels of terminal and non-terminal introns within multi-intronic genes were compared for each mutant, revealing no obvious difference between their behaviors. (D) From our data, it is unclear how the CPF factors Ppn1 and Ssu72 are impacting splicing. Here we depict two possible explanations: deletion of these factors could either prevent proper phosphorylation of the CTD tail and thus disrupt the interaction between U2AF and the CTD tail, or their absence from the CPF complex could disrupt the physical interactions between Yth1 and U2AF. Orange ovals represent CPF factors that cause significant splicing defects in our screen. Yellow ovals indicate factors that caused increases in pre-mRNA levels but were not significant. Green circles represent factors that have been shown in S. cerevisiae to cause splicing defects upon deletion.
In higher eukaryotes, where exons are short and introns can be extraordinarily long, spliceosome assembly is hypothesized to occur by exon definition, wherein recognition of a downstream 5’ splice site can facilitate recognition of an upstream 3’ splice site by cross-exon interactions. For terminal introns, where no downstream 5’ splice site exists, it has been demonstrated that components of the cleavage and polyadenylation machinery can serve to facilitate recognition of the terminal 3’ splice site in a process termed terminal exon definition. Although it has been thought that the short introns in yeast would not require cross-exon interactions for efficient splicing, several studies have demonstrated that components of the cleavage and polyadenylation machinery do impact pre-mRNA splicing in yeast\textsuperscript{75,165}. Given the large number of multi-intronic genes in \textit{S. pombe}, we sought to determine whether the extent to which the pre-mRNA increases detected in these 3’ end mutants were dependent upon the locations of the intron. Each intron in the genome was classified as being either the last annotated intron (terminal) or not the last (non-terminal). The pre-mRNA increases we observed for terminal introns was not obviously different than the increases seen for non-terminal introns, neither in the \textit{Assu72} nor the \textit{Appn1} strains (Figures 2.9B, 2.9C, and 2.10).
Figure 2.10 3’ end processing factors affect the splicing of both terminal and not terminal introns equally. Splicing sensitive microarrays of 3’ end processing factors \( \text{ppn1} \) (A) and \( \text{ssu72} \) (B). Only transcripts that contained multiple introns were considered for this analysis. Introns were called ‘terminal’ when they were the last intron within a transcript, and all introns were grouped together as ‘not terminal’. Each heatmap is organized by hierarchical clustering. When comparing the ‘terminal’ to ‘not terminal’ heatmap for both \( \Delta \text{ppn1} \) and \( \Delta \text{ssu72} \), splicing defects are observed regardless of intron position.
The mechanistic bases by which Ssu72 and PPn1 influence pre-mRNA splicing remain unclear. Although they are physically parts of the CPSF complex, both Ssu72 and Ppn1 are phosphatases that target the CTD of RNA Pol II. Ssu72 preferentially targets the Ser5P modification, while Ppn1 acts upon both Ser2P and Ser5P via the PP1 Nuclear Targeting Subunit (PNUTS) complex. Phosphorylation of Ser5 is generally associated with promoter proximal pausing, and its dephosphorylation is important for escape into productive elongation. The Ser5 mark of the CTD has been shown to be important for efficient splicing in yeast and humans, perhaps by slowing or pausing the polymerase so as to allow more time for co-transcriptional splicing to occur. Given these roles for Ssu72 and Ppn1, the changes in splicing efficiency that accompany their deletions may not be a result of defects in cleavage and polyadenylation activity, per se, but rather changes in the CTD phosphorylation state. Alternatively, our understanding of the interactions between the cleavage and polyadenylation machinery and splicing may be incomplete, such that the interactions known to be important for terminal exon definition may in fact be important for general spliceosome assembly. In budding yeast, where introns are strongly biased towards the 5’ end of transcripts, mutations in the endonuclease Brr5/Ysh1, ortholog of human CPSF-73, yield a strong splicing defect, highlighting the capacity of bona fide 3’ end processing factors to influence splicing at distances far removed from locations of cleavage and polyadenylation. Moreover, affinity capture and mass-spectrometry analysis of the S. pombe cleavage and polyadenylation factor.
complex reveals physical interactions between PPN1 and both of the SR-protein orthologs in the *S. pombe* genome. More experiments will be necessary to understand the mechanistic bases by which 3’ end processing and splicing impact one another in *S. pombe*.

### 2.5 Conclusion

Here we described the development and implementation of a sequencing-based reverse genetic screen to identify the complement of non-essential genes in the fission yeast *S. pombe* that impact pre-mRNA splicing. Our ability to positively identify both known and predicted splicing factors demonstrates the ability of this approach to identify splicing mutants among a collection of thousands of diverse strains. Moreover, the identification here of scores of factors previously unknown to impact splicing highlights the potential of this approach for *de novo* discovery. As with all genetic screens, further characterization of the individual factors identified here will be necessary to understand the mechanistic bases by which each of them impacts the splicing pathway. Nevertheless, the broad interconnectivity of RNA-processing pathways revealed in this work is testimony to the utility of *S. pombe* as a genetic system for studying these processes. Moreover, the recently solved EM structure of the *S. pombe* spliceosome significantly enhances the ability of genetic data to inform about the mechanistic underpinnings of this process. Importantly, because many known components of the spliceosome are themselves essential, they have not been assayed in the screen described here. The creation of a temperature sensitive strain
collection and subsequent screening using methods similar to those described here will present the opportunity to explore those essential genes, and thus provide greater insight into the mechanisms of more complex splicing.

2.6 Author Contributions

B.J.F. developed the sequencing assay and data processing scheme. B.J.F and A.L. performed the RNA extraction, cDNA synthesis and library preparations. A.L. performed the microarray experiments and data processing and analysis. B.J.F., A.L., and J.A.P. prepared the written manuscript. B.J.F. made figures 2.1 to 2.5, and A.L. made figures 2.6 to 2.10.
CHAPTER 3

STRUCTURAL INSIGHTS INTO SPLICING REVEALED BY HIGH-THROUGHPUT SCREENING OF CONDITIONAL ALLELES IN FISSION YEAST

3.1 Abstract

Pre-mRNA splicing is an essential component of eukaryotic gene expression. The chemistry of splicing as well as the machinery that performs the reaction are both highly conserved from yeast to humans. Recent advances in cryo-electron microscopy has resulted in an explosion of structures of intact spliceosomal complexes, providing molecular insights into splicing mechanism as well as evolutionary conservation of the catalytic core of the spliceosome. Here we harness the power of fission yeast, a genetically tractable unicellular fungus that shares many important hallmarks of complex splicing present in higher eukaryotes, to identify mutants that are defective in pre-mRNA splicing. We present the implementation of a sequencing-based screen on a collection of conditional mutant strains generated by random mutagenesis. The identification of the causative alleles by whole genome sequencing revealed mutations in the known splicing factors Sap61, Prp22, Prp1, and Cdc28. Analysis of these mutants with the some of the published structures revealed that many of these alleles will likely provide mechanistic insights of spliceosome activation, function, and regulation. Given the high degree of homology between S. pombe and humans for
much of the splicing apparatus, coupled with important similarities in their intron landscape, these studies will likely provide invaluable insight into splicing regulation and diseases in humans.

3.2 Introduction

An essential step in eukaryotic gene expression is the removal of non-coding introns from pre-mRNA in order to form a translatable message. While the chemistry of intron removal proceeds through two simple transesterification reactions, intron excision is performed by a highly dynamic macromolecular machine called the spliceosome, which is composed of five small nuclear RNAs (snRNAs) with their core proteins (snRNPs), and in some instances upwards of 200 auxiliary proteins$^{4,10}$. Through a series of ordered interactions of the snRNPs and auxiliary proteins, the spliceosome assembles de novo for each intron that is removed, and requires the ordered interaction of the snRNPs to complete the splicing reaction$^{10}$. Understanding how the components of the spliceosome interact with each other at each stage of the splicing cycle to direct the splicing reaction is crucial for understanding the molecular mechanistic details of splicing.

X-ray crystallography has been a powerful tool for visualizing and elucidating mechanistic details of many macromolecular complexes in the cell. Because of the large size as well as large changes in composition and conformation of the spliceosome, until recently, detailed structural studies have largely been limited to individual spliceosomal proteins due to limitations in crystallization and resolution.
Recent advances in cryo-electron microscopy (cryo-EM), however, have led to single particle analysis of the spliceosome at several stages of the splicing cycle, revealing many details about the catalytic center of the spliceosome, as well as the conformational changes that several spliceosomal proteins and the snRNAs undergo\textsuperscript{101,103–107,170–172}. Comparison of the structures across the different stages of the splicing cycle revealed that the spliceosome maintains a stable structural core throughout the splicing reaction, consisting of the U5 snRNP, U6 snRNA, 5-end of the U2 snRNA, and a dozen proteins from the NTC and NTR\textsuperscript{101,103–106,170,171}. Comparisons between the activated spliceosomal complex, B\textsuperscript{act}, and the catalytically active spliceosomal complex, C, revealed details of the first step of the splicing reaction, while comparing the C complex with the C* complex, which is remodeled and poised for exon ligation, revealed details about the second step of splicing. While these structures have provided tremendous molecular insight into the dynamic conformational changes that the spliceosome must undergo throughout the splicing cycle, these structures are largely composed of only those proteins that are at the core of the spliceosome. Upwards of 100 different proteins can associate with the spliceosome throughout the splicing cycle, and how those proteins interact with the spliceosome remains to be determined. The abundance of structural data present in these structures, however, increases the utility of genetic and biochemical data for informing molecular details about the splicing reaction.
The fission yeast *Schizosaccharomyces pombe* is a powerful system for studying splicing mechanisms, particularly as they relate to human biology. While retaining the genetic tractability of the budding yeast, *S. pombe* has an intron landscape that resembles that of higher eukaryotes. Moreover, orthologs of several known human splicing factors are present in *S. pombe* but absent in budding yeast\(^{15,33,173}\). *S. pombe* has two orthologs of SR proteins, which is a class of auxiliary splicing factors that are involved in alternative splicing decisions in higher eukaryotes\(^{15}\). Additionally, there is a protein kinase, Prp4, in *S. pombe* that plays an important role in spliceosome activation through phosphorylation of Prp1, a component of the tri-snRNP\(^{33,35}\). Homologs of Prp4 exist in humans, but have not been retained in the budding yeast.

Recently, we published the results of an unbiased screen of a non-essential gene deletion collection in *S. pombe* to identify mutants that impacted the splicing pathway. Of the mutants that were identified in this screen, about a quarter of them had clear homologs in humans but not in *S. cerevisiae*\(^{140}\). These observations together underscore the utility of using *S. pombe* to study the protein components of the splicing pathway to further understand how each component functions as it relates to human splicing biology.

Most of the known splicing factors are essential genes, and thus conditional alleles of these proteins must be generated to perform to assess splicing-related phenotypes pf these mutant proteins. The use of conditional alleles has been a fundamental approach for obtaining a detailed understanding of the function of essential gene, as well as
providing insights into their molecular details. The generation of conditional alleles in splicing factors in *S. pombe* would allow for the investigation of how these mutants impact splicing *in vivo* in an organism that has many similarities with human splicing. Here, we describe the characterization of a library of randomly generated conditional mutants in *S. pombe* using the sequencing-based screening method we previously described\textsuperscript{140}, to identify mutants that display splicing defect phenotypes. Whole genome sequencing analysis of these mutants revealed novel alleles in known splicing factors. Structural analysis of these mutant alleles in the context of the spliceosome or spliceosomal complexes in the recently published structures revealed insights into how these mutants could impact spliceosome function.

### 3.3 Materials and Methods

#### 3.3.1 Strains, Cell Growth, Screen Libraries, and Data Processing

The work described here takes advantage of a library of temperature-sensitive strains, developed in concert with John Armstrong and Neil Bone (University of Sussex), and partly described in the FYSSION project. The library consists of ~4000 individually arrayed strains, each of which is capable of robust growth at 27°C, but is unable to support growth at 37°C. These strains were isolated using previously described approaches. Briefly, a culture of WT *S. pombe* was subjected to nitrosoguanidine mutagenesis, after which time, individual clones were tested for their ability to support growth at both 27°C and 37°C. Those strains which were unable to support growth at the higher temperature were saved as part of the library.
Strains were grown in liquid culture in biological replicate as previously described\textsuperscript{139,140}, with a few changes. Initial growth of the strains was performed at 25°C, allowing the cells to reach saturation. After reaching saturation, the cells were back-diluted into fresh media and allowed to grow for an additional 8 hours, a sufficient time to enable initiation of exponential growth. In order to elicit the molecular defects that accompany the temperature sensitivities of these strains, the temperature of the cell culture was raised to 37°C by addition of an equal volume of pre-warmed (50°C) media. The cells were incubated with shaking for 15 minutes in an Infors HT Multitron plate shaker at 900 rpm at this elevated temperature, after which the cells were harvested, flash frozen in liquid nitrogen, and stored at -80°C until further processing. RNA isolation from these cell pellets and cDNA synthesis was performed as previously described\textsuperscript{75}.

The RNA from these strains was analyzed using a sequencing-based approach similar to that previously described\textsuperscript{140} to identify strains in which pre-mRNA of the \textit{fet5} transcript accumulated. Two sequential rounds of PCR reactions were performed on the cDNA samples described above according to previous methods. The PCR primers flank the intron of \textit{fet5}, allowing for a determination of the levels of both the mature and pre-mature \textit{fet5} transcripts in the background of each of the strains. The resulting DNA libraries were sequenced on the Illumina NextSeq500 with the assistance of the Cornell University Biotechnology Resource Center. The sequence alignments and data processing were performed as previously described\textsuperscript{140}. 

88
3.3.2 Backcrossing and Random Spore Generation

Candidate strains were backcrossed to a WT strain, KGY461, and then subjected to random spore analysis. The backcrossing was performed by patching a small amount of the $h^{-}$ ts strain on Malt Extract with Supplements (MES) agar. The $h^{+}$ KGY461 strain was patched on top of the ts strain, and they were incubated at 25°C for two days. A small patch of the colony was placed in 1 mL of MilliQ H2O containing 2 uL Glusulase enzyme (PerkinElmer, cat#: NEE154001EA), which contains about 20 U beta-glucoronide sulfatase and 180 U glucoronidase. Glusulase treatment, which degrades the ascus wall but not the spore wall, was carried out overnight at 25°C with rotation. Cells were harvested by centrifugation at 13,000 x g for 3 minutes and resuspended in 100 uL of MilliQ water. The spores were plated at a variety of densities onto YES agar and incubated at 25°C for four days, allowing for isolation of 100’s of individual colonies, each representing a single spore generated from mating of the candidate strain with the WT strain. Typically, 96 individual offspring from each mating were selected into individual wells of a 96-well plate for further analysis and processing.

3.3.3 Whole Genome Sequencing to Identify the Causative Mutation

3.3.3.1 Spore Phenotyping

The temperature sensitivity of individual offspring of mating reactions was determined by assaying their capacities to support growth at both 25°C and 37°C. Cells were grown at 25°C in liquid culture for two days, allowing nearly all the strains
to reach saturation. Then, using a manual pinning tool, a small volume of culture was transferred onto two plates of YES agar. All cells were pinned in quadruplicate onto each plate, and one plate was incubated at 25°C, while the other plate was incubated at 37°C. After an appropriate growth period, the plates were scored for growth in each of the original wells. Growth was characterized as temperature-sensitive if none of the four replicates showed grow at 37°C while their four matched replicates did grow at 25°C. By contrast, growth was characterized as WT if all four replicates grew at both 25°C and 37°C. Strains that showed inconsistent growth phenotypes at either temperature were excluded from downstream analysis.

3.3.3.2 Genomic DNA and Sequencing Library Preparation

Genomic DNA was generated for whole genome sequencing from cells which had grown to saturation at 25°C in 1.2 mL of YES media. Individual offspring were pooled together based on their temperature sensitive phenotype, and 1 mL of culture was used for each offspring. An equal number of offspring were collected in the ts, and WT pools (totaling between 15 and 20 offspring for each pool). For each pool, cells were pelleted at 5000 x g for 5 min. Genomic DNA was harvested using a PrepEase Genomic DNA Isolation Kit (Affymetrix), according to manufacturer’s recommendations. The genomic DNA was prepared for sequencing using a TruSeq DNA PCR-Free Library Prep kit (Illumina), except that shearing of the DNA was accomplished using a Covaris S2 machine at intensity level 3. The library was
sequenced on the Illumina NextSeq500 with the assistance of the Cornell University Biotechnology Resource Center.

3.3.3.3 Data Processing

Analysis of the sequencing reads followed the GATK Best Practices from the Broad Institute\textsuperscript{175–177}. Sequencing reads were first aligned to the \textit{S. pombe} genome (version 2.26) using the Bwa mem aligner\textsuperscript{142}. Once aligned, local areas targeted for realignment were created using the RealignerTargetCreator tool, allowing for a reduction in the number of false discovery events resulting from insertions or deletions (indels). Reads were then realigned to the genome using the IndelRealigner tool. Finally, SNPs were called using the HaplotypeCaller tool. All of the default settings were used for all of these tools from GATK, and the process was performed separately for the phenotypic and aphenotypic pools. The data processing pipeline is outlined in Figure 3.1.
Figure 3.1 Whole genome sequencing read processing scheme. Briefly, the reads are aligned to the *S. pombe* genome using the bwa mem aligner. Once aligned, to reduce the number of local misalignments due to the presence of indels in the sequencing reads, local realignment is performed using the RealignerTargetCreator and IndelRealigner tools. Finally, the HaplotypeCaller tool is used to call SNPs.
3.3.4 RNA-sequencing

Strains were initially grown at 25°C until they reached OD$_{600}$ ~0.5, at which point they were shifted to 37°C in a shaking water bath for 15 minutes. Cells were collected by centrifugation at 5000 x g for 5 minutes in 10 mL aliquots. Cell pellets were flash frozen in liquid nitrogen. RNA isolation was performed as previously described$^{143}$. Sequencing libraries were made using TruSeq RNA Sample Prep Kit version 2 with 2 µg of total RNA as input. The libraries were sequenced on the Illumina NextSeq500 with the assistance of the Cornell University Biotechnology Resource Center. To assess the impact of these mutations on splicing, the splice index (SI) was calculated for every intron in the genome. The SI was defined as the level of pre-mRNA divided by the level of pre-mRNA plus mature mRNA. The level of pre-mRNA was determined by the total number of reads that map either directly to an intron, or to an intron-exon boundary. This number was divided by the length of the intron plus 75 (the length of each sequencing read), to account for total ‘alignable’ length. Similarly, the level of the mature mRNA was determined by the total number of reads that map directly to an exon-exon boundary, again divided by the read length (75). The fold change of SI is calculated by dividing the SI in the ts mutant divided by the SI in the WT sample and log$_2$ transformed.

$$SI = \frac{\text{unsplice read count}}{\text{unspliced read count} + \text{spliced read count}}$$

Equation 3.1
3.3.5 Complementation Testing

The *S. pombe* expression vector pSP2 containing a full-length copy of *sap61*+ including ~300 bp up- and downstream of the ORF was transformed into the *sap61*-E7K strain. In addition, *sap61*-E7K and the Bioneer ED666 consortium WT strains were transformed with empty vector. All strains were grown to OD₆₀₀ ~0.8 and eight four-fold dilutions were made, plated onto YES agar and assayed for growth at 25°C and 37°C.

3.4 Results

3.4.1 Sequencing-Based Screen Identifies ts Strains with Defects in Splicing

To identify conditional mutants that cause splicing defects, we designed PCR primers that flanked the only intron of *fet5* and the third intron of *SPBC21B10.09* in order to measure the endogenous levels of both the spliced and unspliced transcripts for each substrate. These PCR primers contained custom designed barcodes that uniquely identify all ~2200 mutant strains and enable them to be simultaneously assayed. The change in the relative abundance of the unspliced to spliced form, called the splice index, SI (see methods), allowed us to identify mutant strain that were defective for splicing. Splicing analyses were performed for each strain with biological replicates. Here we present the data from one 384-well plate of strains as a pilot of our approach. Whereas most strains on this plate showed little change in the SI relative to WT, 12 strains showed significant SI increases for *fet5* _intron1_ and 11 strains for *SPBC21B10.09* _intron3_, Figure 3.2. Between these two lists of strains
displaying splicing defects, six strains are common to both lists, bringing the total number of strains with splicing defects to 17, Figure 3.3.
Figure 3.2 Relative splice index measurements for each ts strain in plate 1 for fet5_intron1 and SPBC21B10.09_intron3. The relative splice index for fet5_intron1 (A) and SPBC21B10.09_intron3 (B) are plotted as functions of the sequencing depth generated for each ts strain. Those strains that show significant deviations in their SI values are colored in red. A total of 12 strains showed significant defects in splicing of fet5, and 11 for SPBC21B10.09_intron3.
Figure 3.3 Overlap in the number of strains with significant relative splice index measurements for fet5_int1 and SPBC21B10.09_intron3. Of the 12 strains that showed significant splicing defects for fet5_int1 and the 11 for SPBC21B10.09, six strains were common to both, for a total of 17 mutant ts strains displaying splicing defects.
3.4.2 Determining the relationship between splicing and temperature phenotypes.

Because our library of strains was generated by random mutagenesis across the genome, the mutant phenotypes we observed could have been the result of a single mutant locus, or through the interactions of two or more mutant loci. To determine if more than one mutant locus was responsible for causing the $ts$ phenotype, random spores were generated and phenotyped for temperature sensitivity to assess the segregation of the $ts$ phenotype. Phenotypes caused by a single genetic locus would be expected to generate an equal number of $ts$ and WT offspring. As an initial test, we focused on a single candidate strain, 103_N17, which showed a strong splicing defect in the original screen. This $ts$ strain was backcrossed to a WT strain, the resulting diploid strain was sporulated, and the haploid offspring were assessed for fitness of both temperature and splicing. Whereas all the aphenotypic (or background) mutations that were created by the original chemical mutagenesis freely segregate into both phenotypic and aphenotypic offspring, the mutation(s) responsible for the phenotype would specifically accumulate in offspring with the mutant phenotype (Figure 3.4).
Figure 3.4 Back-crossing and bulked segregant scheme. The haploid ts parent is backcrossed to a haploid WT parent to form a diploid. The blue lines in the WT parent are background variants in the WT genome. The green lines in the ts parent indicate either background mutations present in the strain that was mutagenized or are other mutations that were caused by the mutagenesis but are aphenotypic. The red line and asterisk indicate the SNP that causes the mutant phenotype. After random sporulation, this SNP will segregate into all of the spores that display the mutant phenotype, while the other mutations will independently assort into all of the spores. To identify the locus causing the mutant phenotype, those spores with the mutant phenotype will be pooled together and their pooled genomic DNA samples will be sequenced. The same will be done with the aphenotypic spores. The SNP that is enriched in the phenotypic pool will be the causative SNP.
After random spores were generated, 48 individual progeny were picked and assayed for growth at the permissive and non-permissive temperatures. Out of the 48 progeny tested, 16 were $ts$, giving a ratio of $2 \text{ WT} : 1 \ ts$ (Figure 3.5). Given this result alone, it was unclear whether the apparent deviation from a $2 \text{ WT} : 2 \ ts$ segregation ratio indicated that more than one locus was responsible for the $ts$ phenotype, or whether this simply represented random fluctuations in the populations.
Figure 3.5 Temperature sensitive phenotype of the spores. The spores from the mating between the ts parent and the WT parent were pinned onto rich media and grown at the permissive temperature and non-permissive temperature and were phenotyped for temperature sensitivity. Out of 48 spores, 16 were ts, resulting in a 2 WT : 1 ts ratio.
In order to assess the linkage between the \( ts \) phenotypes and the splicing defect, a small-scale version of the sequencing-based screen used to identify the \( ts \) candidates was performed. Using the same primer design, we measured the splicing efficiency of \( fet5 \) in each of these offspring in biological replicate. If the splicing defect was linked to the \( ts \) phenotype, then the offspring that were \( ts \) would have higher SI values compared to the WT offspring. Although the correlation wasn’t perfect, the observed SI measurements suggested a strong linkage between the splicing defect and the temperature sensitivity we detected in these offspring (Figure 3.6B). By plotting the SI value for each individual offspring, Figure 3.5B, it was clear that there are a handful of offspring that were phenotyped as WT which do in fact display splicing defects, and the converse was true for some of the \( ts \) offspring. This result could be seen because of several reasons. One is that the splicing and \( ts \) phenotypes were caused by two separate mutations that were not linked. Another is that the offspring were incorrectly identified in their temperature sensitivity, and those which showed a splicing defect were in fact temperature sensitive. Regardless, because our goal was to identify mutations that cause the splicing defects, we focused on those offspring with a two-fold or greater SI value, labeled with red asterisk in Figure 3.6B.
Figure 3.6 Splicing defect phenotype largely segregates with ts phenotype. 
(A) The median SI for *fet5* is greater than four-fold higher than the median SI for the WT spores. (B) Some of the spores that are phenotyped as WT do display splicing defects. Those spores that have a splicing defect of greater than two-fold, indicated with a red asterisk, were pooled together as the phenotypic pool for bulked segregant analysis.
3.4.3  Bulked Segregant Analysis Followed by Whole Genome Sequencing Identifies the Putative Causative Mutation

To determine the mutation responsible for the splicing defect, bulked segregant analysis (BSA) was performed. Here, progeny were pooled based on shared phenotypes and genetic analyses were performed on each pool separately. Because the original strain mutagenesis presumably resulted in many background mutations in addition to the phenotype causing mutation, BSA enables identification of the single nucleotide polymorphisms (SNPs) that are enriched in the pool with the phenotype of interest. Therefore, progeny were pooled based on those that displayed splicing defects and those that did not. The genomic DNA from each pool was sequenced in order to identify the SNP(s) enriched in the splicing defective progeny and depleted in the WT progeny. The sequencing reads were analyzed according to Figure 3.1. In the splicing defective pool, 471 SNPs were identified, and 502 SNPs were present in the WT pool.

To narrow in on the potential causative SNP, the list of SNPs was initially filtered such that only those SNPs that were in or near (+/- 500bp) open reading frames (ORFs) were present. From these subsets, SNPs that were common to both pools were filtered out, leaving only those SNPs that were unique to either the splicing defective or WT pool, Figure 3.7.
Figure 3.7 Number of SNPs in the Phenotypic and Aphenotypic Pools. Listed here is the total number of SNPs that is identified in each pool, the number of SNPs that are in or +/- 500 bp of an ORF, and the number of SNPs that are unique to either pool.
Because the list of unique SNPs is so small, we manually examined the data for each SNP individually. For simplicity, we began by considering only those mutations that resided within the mRNA of a protein-coding gene, and we looked at the functions of those gene products to identify any obvious splicing factor harboring a mutation. Within the 39 SNPs that were unique to the phenotypic pool, 20 of those SNPs were within coding regions. Indeed, Sap61, a U2 snRNP-associated protein, was the only splicing factor in the list of SNPs enriched in the splicing defective progeny. More importantly, the SNP in sap61 reflects a G-to-A transition mutation at position II:858506 that is completely penetrant in the splicing defective progeny and is mostly absent in the WT progeny, Figure 3.8. This distribution of reads between the splicing defective pool and the WT pool matched expectations for enrichment of the causative allele segregating with the splicing phenotype. Additionally, the nucleotide change results in a substitution of the glutamate residue at position 7 with a lysine residue, which has different chemical properties and is likely to cause structural changes to the protein. This allele will be referred to as sap61-E7K.
Figure 3.8 Bases called at the SNP location in *sap61*. Here is the distribution of the bases called at the SNP locus in *sao61* in both the temperature sensitive pool and the WT growth pooled sample. The G-to-A mutation is 100% penetrant in the phenotypic pool and depleted in the aphenotypic pool. This mutation results in a E7K missense mutation.
3.4.4 Characterization of sap61-E7K

3.4.4.1 Sap61$^+$ rescues the ts phenotype in sap61-E7K

To determine if the sap61 locus was responsible for the mutant phenotypes, complementation testing was first performed to determine if expression of WT Sap61 in the sap61-E7K strain could rescue the ts phenotype. The sap61-E7K strain with the empty vector is completely dead at 37°C. Expression of the sap61$^+$ in this strain rescues the growth defect, and growth at 37°C resembles that of the WT strain with an empty vector, Figure 3.9.
Figure 3.9 Expression of sap61\textsuperscript{+} in sap61-E7K restores WT growth at the restrictive temperature. The sap61-E7K strain is inviable at the restrictive temperature. Expressing the WT Sap61 in this mutant strain rescues the growth defect observed at 37°C.
3.4.4.2 The sap61-E7K strain displays a genome-wide splicing defect

To further characterize the splicing defect phenotype of the sap61-E7K strain, RNA-sequencing was performed to assess the splicing efficiencies of all of the introns in S. pombe. RNA-seq was performed on biological replicates of sap61-E7K and a matched WT strain. The splice index (SI) was calculated for all four samples as in Equation 3.1, and allowed for an assessment of global splicing, Figure 3.10A. Globally, the SI shifted to a greater value in the ts samples, indicating an increase in the amount of pre-mRNAs present in this mutant background and thus, a splicing defect. Additionally, when comparing the fold change of the SI value in sap61-E7K over the SI value in the WT sample, the median value shows a greater than four-fold increase, Figure 3.10B. These data indicate that the defect in splicing observed in sap61-E7K is not specific to the fet5 transcript, but rather, this mutant strain displays a large global splicing defect.
Figure 3.10 RNA-seq data show a global splicing defect in one of the candidate strains. (A) The SI for each intron is displayed in a histogram. The WT biological replicates are in blue, sap61-7EK is in red. The median SI value is shifted far to the right in the mutant strains, indicating a large splicing defect in these strains. These data are also displayed in boxplots on the right, (B). (C) Fold-change of the SI value for one of the sap61-E7K replicates over one of the WT replicates in displayed on the bottom. The top histogram is one WT replicate over the other WT replicate and indicates the spread of the noise in these measurements. In the bottom histogram, the greater than two-fold shift to a higher median SI value for sap61-E7K also demonstrates a large splicing defect, with nearly every intron being affected in this mutant strain.
3.4.5 Pooling by Temperature Sensitive Phenotype Shows Missense Mutations in Splicing Factors

Given the relatively small number of SNPs identified in our pilot experiment, and because of the relative ease with which temperature sensitivity can be assessed relative to splicing fitness, we chose to proceed with bulked segregant analysis for subsequent strains on the basis of temperature sensitivity alone. For this end, BSA was performed on three additional candidates: 103_A10, 103_M03, 103H15. When the three candidate strains’ ts and WT pools were sequenced, similar numbers of SNPs were seen in each pool as was observed in sap61-E7K, highlighted in Table 3.2. A comparison of all datasets generated revealed that 405 SNPs were common to three or more of the datasets, suggesting that these variants were initially present in one of the two parental strains. Given that these mutations were present in all strains, independent of phenotype, it suggested that they were unlikely to be related to our phenotype of interest, and as such were excluded from further consideration.
<table>
<thead>
<tr>
<th>Candidate</th>
<th>Total # SNPs</th>
<th>Total # Unique SNPs</th>
<th>SNPs unique to ts Pool</th>
<th>Putative Mutant Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>103_A10</td>
<td>466</td>
<td>85</td>
<td>20</td>
<td>prp22</td>
</tr>
<tr>
<td>103_M03</td>
<td>478</td>
<td>83</td>
<td>23</td>
<td>cdc28</td>
</tr>
<tr>
<td>103_H15</td>
<td>496</td>
<td>111</td>
<td>52</td>
<td>prp1</td>
</tr>
<tr>
<td>sap61_E7K</td>
<td>447</td>
<td>81</td>
<td>39</td>
<td>sap61</td>
</tr>
</tbody>
</table>

Table 3.2 Number of SNPs called for each candidate ts strain sequenced. Each ts strain was back-crossed to a WT strain and spores were pooled by their ts phenotype and sequenced. The numbers listed in this table are the SNPs that are present in either the phenotypic, aphenotypic, or both, unless otherwise noted.
For candidates 103_A10 and 103_M03, identifying a putative causative SNP was feasible by looking at the SNPs that were present only in the temperature sensitive pool dataset. Since these pools were made based on ts phenotype, this suggested that the ts phenotype and splicing defect were caused by the same SNP. For 103_A10, a SNP was identified in the gene encoding the splicing factor Prp22. This SNP was a C-to-T transition and resulted in a G917R missense mutation. This mutation was completely penetrant in the ts phenotype pool and depleted in the WT growth pooled sample, Figure 3.11. The sequencing of the pools from candidate 103_M03 also revealed a SNP enriched in the ts pool in the gene encoding the splicing factor Cdc28. This SNP was a T-to-C transition and resulted in a V735R missense mutation. Again, the mutant allele was completely penetrant in the ts sample, and very few reads (only 5 out of 57) were present in the WT growth sample, Figure 3.12.
Figure 3.11 Bases called at the SNP location in prp22. Here is the distribution of the bases called at the SNP locus in prp22 in both the temperature sensitive pool and the WT growth pooled sample. The C-to-T mutation is 100% penetrant in the phenotypic pool and depleted in the aphenotypic pool. This mutation results in a G917R missense mutation.
Figure 3.12 Bases called at the SNP location in *cdc28*. Here is the distribution of the bases called at the SNP locus in *cdc28* in both the temperature sensitive pool and the WT growth pooled sample. The T-to-C mutation is 100% penetrant in the phenotypic pool and depleted in the aphenotypic pool.
For the 103_H15 strain, unlike those described above, there was no obvious SNP within the enriched pool of mutations that resided within a known splicing factor. An expanded search of the ~100 SNPs specific to this mutant strain revealed a mutation in \(prp1\), a component of the U4/U6.U5 tri-snRNP important in tri-snRNP assembly.

There was a C-to-T transition mutation that was in 64% of the reads in the phenotypic sample, and only 23% of the reads in the aphenotypic sample, Figure 3.13. This SNP causes a change in amino acid 653 from serine to proline. It remains unclear whether the offspring were phenotyped incorrectly for temperature sensitivity, if the splicing defect and \(ts\) phenotype are un-linked, or whether a complex set of mutations is responsible for both the temperature and splicing defects in this strain.
Figure 3.13 Bases called at the SNP location in *prp1*. Here is the distribution of the bases called at the SNP locus in *prp1* in both the temperature sensitive pool and the WT growth pooled sample. The T-to-C SNP is present in 63% of the reads at this position in the temperature sensitive pool, and 23% of the reads in the WT growth pool.
3.5 Discussion

Pre-mRNA splicing is an essential step in eukaryotic gene expression, and complex splicing pathways are pervasive in higher eukaryotes. In order to decipher the mechanisms of complex splicing, it is important to not just identify all of the factors that impact the splicing pathway, but to identify and characterize the defects of specific mutant alleles, leading to even greater mechanistic insight. Here we have described the implementation of a sequencing-based screen on a novel conditional mutant library in \textit{S. pombe}, a unicellular eukaryote with a splicing pathway that resembles that of higher eukaryotes in many ways, to identify mutant strains that display canonical splicing defects. Bulked segregant analysis of \textit{ts} versus non-\textit{ts} progeny identified the putative causative SNP for several of the candidate strains, revealing novel conditional alleles in core spliceosome components. Identification of point mutations in core splicing factors coupled with the surge in structures of the spliceosome or subcomplexes of the spliceosome in the past year allows for a more informed hypothesis of how these mutants can affect the splicing pathway. Since 2015, there have been at least 30 splicing related structures published. With this abundance of new structural information for many different subcomplexes of the spliceosome and even a structure of the activated spliceosome, identification of novel splicing conditional alleles will reveal functionally important protein-protein and protein-RNA interactions.
Sap61 is an essential protein and is a component of the SF3a complex, a trimeric protein complex that is necessary for anchoring U2 snRNA to the branchpoint and subsequent formation of the A complex\textsuperscript{178,179}. Both budding yeast and humans have orthologs of Sap61, PRP9 and SF3A3\textsuperscript{180}, respectively. The amino acid sequence of Sap61 is highly conserved in both \textit{S. cerevisiae} and in humans. Our data revealed a SNP that results in an amino acid change at position seven, replacing a glutamate residue with a lysine residue. This glutamate residue, as well as many other residues near the N-terminus, is conserved in both budding yeast and humans, Figure 3.14. Expression of the WT Sap61 in the background of the \textit{sap61-E7K} rescues the \textit{ts} phenotype in this mutant, revealing that \textit{sap61-E7K} is a recessive allele. Expression of the empty vector in this mutant was completely inviable at the restrictive temperature, indicating that the \textit{ts} phenotype is in fact due to the \textit{sap61-E7K} and that suppression of that phenotype was not due to the presence of the vector. However, replacing the WT allele with the \textit{sap61-E7K} allele is the ultimate test for determining causality and remains to be done.
Figure 3.14 BLASTp alignments of the N-terminus of Sap61 with those of ScPRP9 and HsSF3A3. Here, the alignments of Sap61 with the budding yeast and human homologs are shown, respectively. The red box highlights the glutamate residue that is mutated in sap61-E7K. This residue is conserved near the N-terminus of all three homologs.
The structure of Prp9, the *S. cerevisiae* ortholog of Sap61, complexed with the other two *S. cerevisiae* components of the SF3a complex, Prp11 and Prp21, has been solved by X-ray crystallography\(^96\). In the protein structure, this glutamate in the N-terminal alpha-helix forms a salt bridge with the lysine 97 residue in Prp21, Figure 3.15. It has been previously demonstrated that the N-terminus of Prp9 is necessary for interaction with Prp21, and that the N-terminal alpha helix makes extensive contact with Prp21\(^{181,182}\). Therefore, it is plausible that a the E7K mutation in Sap61 disrupts a crucial thermodynamic interaction important for the formation of the SF3a complex, which in turn could affect the formation of SF3a and in turn, its ability to interact upstream of the branchpoint sequence, resulting in a global splicing defect. This is consistent with the broad defect in splicing observed in RNA-sequencing of the *sap61-E7K*. 
Figure 3.15 Structure of SF3a complex in *S. cerevisiae*. Here, the E5 residue in *S. cerevisiae* Prp9p, the magenta colored protein structure, is shown forming a salt bridge interaction with K97 in Prp21, the blue protein structure. The E5 residue is the conserved glutamate that is mutated in *S. pombe* sap61-E7K.
Analysis of another ts candidate defective in splicing revealed a SNP in Prp22, a conserved ATP-dependent DEAH-box RNA helicase. Prp22 is important for proofreading the exon-ligation reaction, and mediates the ATP-dependent mRNA release from the spliceosome by unwinding RNA duplexes. The first SNP causes a silent mutation while the second results in a missense mutation by replacing a glycine residue with an arginine residue at position 917. This glycine residue is conserved in both S. cerevisiae and humans. Because arginine is so drastically different in its chemical properties from glycine, the G917R mutation is likely to be the cause of the splicing defect phenotype. Recently, a structure of the S. cerevisiae spliceosome poised for the exon ligation was solved, and position and structure of Prp22 in this complex was solved. Prp22 sits at the periphery of this spliceosomal complex, and exerts its proofreading effects on the exon ligation reaction by binding to the 3’ exon. The glycine residue that is mutated in S. pombe is conserved in S. cerevisiae and is at position 889. This glycine residue is in a loop connecting two alpha helices at the protein surface, Figure 3.16. This part of the protein is distant from its active site. It is plausible that the substitution of an arginine at this residue in a loop would significantly alter the folding of this protein, and thus reducing its essential activity. Further biochemical characterization needs to be done to confirm that this mutation is responsible for the splicing defect and how it affects the proofreading function of Prp22.
Figure 3.16 Structure of Prp22 in the C* spliceosomal complex in *S. cerevisiae*\(^{171}\). Here, Prp22 is shown in blue on the periphery of the C* spliceosomal complex in *S. cerevisiae*. The G889 residue is orthologous to the G917 residue that is mutated to an arginine residue in *S. pombe* and is shown highlighted in green.
A conditional allele was revealed in another ATP-dependent RNA helicase, Cdc28, which is required for formation of the catalytically active spliceosome before the first trans-esterification reaction\textsuperscript{184}. Sequencing revealed a SNP likely induced by the original mutagenesis. This SNP is likely the cause of the splicing defect due to the substitution of a proline for an arginine residue at position 735. This arginine residue is conserved in both budding yeast and humans and is in the conserved C-terminal domain that is found in other helicases. The activated spliceosome, B\textsuperscript{act}, from budding yeast was recently solved by cryo-electron microscopy. Prp2, the budding yeast homolog of Cdc28, sits on the outside of B\textsuperscript{act} and contacts the spliceosome through the HEAT domains of Hsh155p, a component of U2 snRNP, Figure 3.17. The arginine at position 543 is the conserved arginine mutated in \textit{S. pombe cdc28-R735P}. A proline substitution could significantly change the local folding around this residue, which could affect how this protein interacts with spliceosome and impair its ability to form the catalytically active spliceosome. Further biochemical tests will reveal more mechanistic details about how this residue and domain contribute to the spliceosome remodeling function of Cdc28.
Figure 3.17 Structure of Prp2p in the B$^{act}$ spliceosomal complex in *S. cerevisiae*. Here, Prp2, the *S. cerevisiae* ortholog of Cdc28 in *S. pombe*, is shown in blue on the periphery of the B$^{act}$ spliceosomal complex in. The R543 residue is orthologous to the R735 residue that is mutated to a proline residue in *S. pombe* and is shown highlighted in orange.
Analysis of the ts strain that revealed a mutation in prp1 required analyzing the list of SNPs in the ts pooled sample and the WT sample because there was no single SNP that was completely penetrant in the phenotypic pool. While sequencing revealed a SNP in prp1 that resulted in a S652P missense mutation, this SNP did not segregate with the progeny that showed a ts phenotype. Identification of this SNP required analysis of all the SNPs unique to this mutant candidate, which was feasible because there are only ~100 SNPs per conditional mutant strain that were induced by the mutagenesis strategy. Even though this SNP in prp1 was not fully penetrant in the phenotypic sample, it is the most enriched SNP, with 67% of the sequencing reads containing the mutant allele. It is possible that this mutation results in the splicing defect but is not responsible for the ts phenotype observed. Additional tests will need to be performed to determine that this mutation is responsible for the splicing defect, but not the ts phenotype.

Prp1 is an important protein in formation of the U4/U6.U5 tri-snRNP. The C-terminus of Prp1 makes many contacts with the U4/U6, while the N-terminus interacts with components of U5 to bridge these two subcomplexes together and facilitate assembly of the tri-snRNP. A cryo-EM structure of the budding yeast tri-snRNP was recently published. In this structure, Prp6, the budding yeast ortholog of Prp1, is shown making important interactions with many components of the tri-snRNP. The asparagine at position 641 in Prp6 is the S. cerevisiae counterpart of the residue that is mutated in Prp1 in S. pombe, Figure 3.18. This residue is not involved in making
contacts with any other protein in the tri-snRNP. It sits in an α-helix at the surface of the complex. Interestingly, Nguyen et al. mapped the budding yeast tri-snRNP onto the human EM envelope of human complex B$^{107}$. Here, they show that the U2 snRNP binds near Prp6. The mutated residue identified in our study sits in the region of Prp6 that appears to make contacts with SF3b155, a protein that is important for anchoring U2 snRNP to the pre-mRNA. The splicing defect observed when this residue is mutated to a proline residue, which is likely to have a large effect on local protein-folding, might reveal an important contact between the U2 snRNP and the tri-snRNP. It is possible that this contact facilitates activation of the spliceosome.
Figure 3.18 Structure of Prp6p in the *S. cerevisiae* tri-snRNP. Here, Prp6, the *S. cerevisiae* ortholog of Prp1 in *S. pombe*, is shown in green in the tri-snRNP complex. The N641 residue is orthologous to the S652 residue that is mutated to a proline residue in *S. pombe* and is shown highlighted in purple.
An important strength in our approach is the ability to robustly identify conditional mutant strains that are defective in splicing pre-mRNAs in their natural contexts in high throughput, and subsequently identify the causative allele for each mutant candidate. Additionally, because the conditional mutant strains in this collection do not appear to be heavily mutagenized, increasing the throughput of causative SNP identification is feasible. Identification of the SNPs in each mutant candidate could be done by whole genome sequencing of just the parent strain. In this study, a list of SNPs that is common to all four mutant strains that were sequenced was generated, and this list can be used to filter out the natural variants present in the mutant candidate strain, and leaves ~100 SNPs that are the result of the mutagenesis. While this would not provide any information about the linkage between the ts and splicing phenotypes, the dataset for the prp1 mutant strain indicates that this may not always be necessary to identify the splicing defective SNP. This procedure removes the backcrossing and spore analysis portions of the experimental pipeline. Additionally, the whole genome library preparations could now be done on 96 candidate strains simultaneously using a plate format for genomic DNA isolation and library preparation. Together, this would increase the throughput for identifying the SNPs responsible for splicing defects, and would thus enable identification of scores of conditional alleles in splicing factors, providing invaluable mechanistic insights in the splicing cycle.
Taken together, the results presented in this study of the interrogation of a collection of conditional alleles in *S. pombe* provides evidence for a robust method for identifying novel alleles in known splicing factors. The identification of novel *ts* alleles in splicing factors coupled with the surge in structures of splicing relevant complexes being solved provides an exciting opportunity to study and learn about the mechanistic basis of many complex splicing decisions. Given the level of homology between splicing machinery in *S. pombe* and humans, as well as many important similarities in their intron landscapes, the mutations identified in these studies are likely to provide important mechanistic insights into splicing regulation in humans, where many diseases are the result of misregulation of splicing.

### 3.6 Author Contributions

B.J.F. developed the sequencing assay and data processing scheme. B.J.F and A.L. performed the RNA extraction, cDNA synthesis and library preparations. A.L. processed and analyzed the sequencing data. A.L. performed the backcrossing, *ts* and splicing phenotyping experiments, and BSA and whole genome sequencing and analysis. A.L. performed cloning and B.J.F and J.A.P. did the complementation experiment. A.L. and J.A.P. prepared the written analysis of this work. A.L. made the figures.
CHAPTER 4

SPlicing of multi-intronic transcripts: Preliminary data and future directions

4.1 Spliceosome Assembly is Largely Co-Transcriptional

Removal of introns from transcripts is an essential component of eukaryotic gene expression and is catalyzed by the spliceosome, a dynamic macromolecular complex. Understanding how the spliceosome is recruited to and assembled upon transcripts is critical for understanding how splicing can be regulated. Decades of research has established that the canonical pathway for spliceosome assembly is a stepwise process, with the U1 snRNP recognizing and binding to the 5’ splice site, then the U2 snRNP binding to 3’ end of intron, and finally the joining of the U4/U6:U5 tri-snRNP with subsequent conformational changes to form the catalytically active spliceosome for intron removal\textsuperscript{10,186}. Additionally, much work has been done over the past two decades has revealed the vast interconnectedness of splicing with other pathways in the nucleus\textsuperscript{75,140,187}. Less than two decades ago, it was believed that splicing largely occurred after transcription termination. Now, it is widely accepted that spliceosome recruitment and assembly is initiated co-transcriptionally and the substrate that the spliceosome encounters is a nascent RNA molecule emerging from the exit channel of Pol II\textsuperscript{55,56,188}.

Several lines of evidence indicate that splicing is functionally coupled with transcription. Electron microscopy images showed looped RNA attached to
chromatin\textsuperscript{189}. RNA in situ hybridization with probes that detect the exon-exon junction also detected spliced mRNAs at their gene loci\textsuperscript{190}. Since these early studies, it has been well established that spliceosome components are deposited onto nascent transcripts co-transcriptionally by Pol II\textsuperscript{191–193}. Genome-wide studies have identified that co-transcriptional recruitment of splicing complexes is the norm, rather than the exception\textsuperscript{45,46,113}. Other studies demonstrated that the co-transcriptional nature of spliceosome assembly improved splicing efficiency, even in co-transcriptionally coupled in vitro reaction systems\textsuperscript{43,194}. The integration of splicing with transcription allows for faster splice site recognition, and thus assembly of splicing components onto the nascent transcript to increase the efficiency of splicing.

First insights into a functional coupling between these two pathways was during studies that demonstrated that introns were inefficiently spliced when transcribed by either RNA polymerase I or III and not Pol II\textsuperscript{36,195}. It was later established that it was the CTD tail of Pol II that was responsible for coupling transcription and splicing because truncations of the CTD tail resulted in defects in RNA splicing\textsuperscript{196}. The CTD contains a highly conserved heptad repeat, YS\textsubscript{2}PTS\textsubscript{3}PS\textsubscript{7}, where budding yeast and fission yeast have 26 and 29 repeats, respectively, while humans have 56 repeats\textsuperscript{39,40}. Each residue in the CTD can be modified, and phosphorylation of several of these residues has been linked to different stages of transcription\textsuperscript{42,197}. The different CTD marks serve as a landing pad for many processing factors to recruit them to the elongating polymerase. Several physical interactions between the CTD and splicing
factors have been established, including binding of the U1 component Prp40\[Ref 117\], and U2AF65\[Ref 169\], a U2 component. However, how phosphorylation of these residues in the CTD affects co-transcriptional recruitment of splicing factors remains unclear. Recent studies have cast doubt on a direct connection between the CTD in splicing factors. Failure to detect U1 snRNP on intronless genes in budding yeast despite being transcribed by PolII suggests that the CTD might not be solely responsible for recruiting splicing factors to nascent transcripts\[198\]. Even more compelling evidence came recently from Inada et al, who demonstrated that neither Ser2 or Ser7 were required for efficient splicing in fission yeast, and that phosphomutants of these residues did not result in a genome-wide splicing defect\[44\]. Therefore, how the CTD impacts splicing still remains unclear.

Recruitment of splicing factors to nascent transcripts by chromatin marks and the Mediator complex has also been observed. The H3K36me3 histone mark is made co-transcriptionally\[83\]. In addition to the presence of this mark within alternative exons correlating with their inclusion, H3K36me3 recruits the polypyrimidine tract-binding protein to chromatin via interactions chromatin readers\[199,200\]. However, how this enables assembly on the nascent RNA remains unknown. Is there a mechanism that transfers these splicing factors from the chromatin onto the nascent RNA? The Mediator complex, a transcriptional coactivator that binds promoters and enhancers, has also been shown to impact alternative splicing decisions via its Med23 subunit, which physically interacts with the splicing proteins SF3b and hnRNPL\[61\].
In higher eukaryotes, most transcripts contain more than one intron, and how spliceosomal components are recruited to downstream introns efficiently is a question that remains largely unclear. Each splice site in a multi-intronic transcript is recognized independently\textsuperscript{193,201}. Several studies have looked at the order of intron removal in a multi-intronic transcript and have demonstrated that the introns do not have to be removed in a 5’ to 3’ order\textsuperscript{202,203}. However, these studies looked only at the completion of the splicing reactions and did not examine spliceosomal recruitment at each intron. Recently, another study demonstrated that the exon-junction complex is deposited soon after the splicing reaction is completed to ensure efficient splicing of downstream introns, but again did not look at recruitment of splicing components\textsuperscript{204}.

With the abundance of proteins involved in RNA processing and transcription elongation that have been implicated in interacting with the CTD tail, it is hard to imagine that all the splicing components necessary for recognizing every splice site within a multi-intronic transcript travel along with the elongating polymerase. Perhaps there is a recharging mechanism for replacing the splicing components back on the elongating Pol II complex after being deposited onto the nascent transcript. Alternatively, the chromatin environment demarcating exons with the gene might be responsible for placing splicing components in the near vicinity of recently transcribed splice sites. Because most human transcripts contain several introns, understanding how splicing components are recruited and assembled on each intron is an important question that needs to be answered to fully understand splicing regulation as it relates to human biology.
Here I describe preliminary work that we have done to identify proteins that impact splicing efficiency of downstream introns in a multi-intronic transcript. Using a sequencing-based screen we developed\textsuperscript{140}, we measured the splicing efficiency of each intron within a multi-intronic transcript, \textit{SPBC21B10.09}, in the background of a collection of thousands of temperature-sensitive strains of \textit{S. pombe}. We hypothesize that those mutants that affect recruitment of splicing factors to downstream introns would demonstrate decreased splicing efficiency of the downstream introns but have no effect on the splicing efficiency of the first intron. Our preliminary analyses identified a mutant of \textit{prp10}, a component of the U2 snRNP, that displays this splicing defect in downstream introns. I discuss a potential model for how this factor could affect downstream intron splicing, and then provide some future directions.

4.2 Preliminary Data

4.2.1 Sequencing-Based Screen Shows Mutants Have Different Effects on Introns within a Transcript

To identify mutants that are defective in recruiting the spliceosomal machinery to downstream introns, we chose the \textit{SPBC21B10.09} as our model transcript to test. This transcript contains three introns, all of which have splice site sequences that do not deviate from consensus significantly, and all the intron sizes fall within the normal intron size range of \textit{S. pombe} introns, Figure 4.1. We have designed PCR primers that monitor both the pre-mRNA and the mRNA species for each intron, and performed the sequencing based screen we designed as previously described to monitor the splicing
index, SI, of each intron in \textit{SPBC21B10.09}\textsuperscript{140}. As a pilot screen, we only screened through a subset of the randomly generated temperature sensitive collection, and growth conditions were performed as in chapter two. Mutants that cause a defect in splicing show increased SI values due to an increase in precursor RNA levels. When we analyzed the measured splice index level as a measure of read count depth for each strain, 9 conditional mutants showed significant defects in splicing of intron 1 (Figure 4.2), 12 mutants that affected intron 2 (Figure 4.3), and 11 mutants showed defects in splicing of the third intron (Figure 4.4).
Figure 4.1 Splice site sequences and size of all the introns in *SPBC21B10.09*. The splice site sequences for each intron in *SPBC21B10.09* conform to the consensus sequences of introns in *S. pombe*, as well as fall within the range of most intron lengths.
Figure 4.2 Splice index measurements for each ts strain in plate 1 for \textit{SPBC21B10.09_intron1}. The splice index for each mutant strain in one plate of \textit{ts} mutants was plotted as a function of the sequencing depth generated for each \textit{ts} strain measurement of \textit{SPBC21B10.09_intron1}. Those strains that showed significant increases in splice index are colored in red. A total of 9 \textit{ts} mutant strains showed defects for splicing the first intron.
Figure 4.3 Splice index measurements for each ts strain in plate 1 for SPBC21B10.09_intron2. The splice index for each mutant strain in one plate of ts mutants was plotted as a function of the sequencing depth generated for each ts strain measurement of SPBC21B10.09_intron2. Those strains that showed significant increases in splice index are colored in red. A total of 12 ts mutant strains showed defects for splicing the first intron.
Figure 4.4 Splice index measurements for each ts strain in plate 1 for SPBC21B10.09_intron3. The splice index for each mutant strain in one plate of ts mutants was plotted as a function of the sequencing depth generated for each ts strain measurement of SPBC21B10.09_intron1. Those strains that showed significant increases in splice index are colored in red. A total of 11 ts mutant strains showed defects for splicing the first intron.
Factors that affect the recruitment of spliceosome factors to downstream introns in multi-intronic transcripts are likely to show defects in the splicing of introns 2 and 3 of \( SPBC21B10.09 \), but would have no effect of on the splicing of the first intron in this transcript. Therefore, we compared the splicing indices for each intron in each \( ts \) mutant (Figure 4.5). For many strains, similar phenotypes for each intron are observed. However, there are several instances in which mutants affect only one or two of the introns, while have no effect on the other intron, indicating that not all mutations affect each intron within a transcript similarly. We were most interested in identifying those strains that showed defects in splicing the two downstream introns and had no effect on the first intron, and thus limited our analysis to those strains that had significant increases in splice index for introns 2 and 3 (Figures 4.3 and 4.4), and did not have a significant increase in the splice index for the first intron, (Figure 4.2). In comparing the datasets between all three introns, there were five \( ts \) strains that showed statistically significant defects in the splicing of all three introns, indicating that they were more likely to be general splicing mutants. More interestingly, there were five mutant strains that had significant defects in splicing of introns 2 and 3, but not intron 1. Further analysis was limited to this latter subset of mutants.
Figure 4.5 Heat map comparing the splice indices of all three introns of *SPBC21B10.09*. The 384 strains on the plater were filtered for those strains that had splice index values for all three introns, and ordered in descending order by the splice index values for intron 3 and intron 2 and ascending order for intron 1. Of the 364 strains analyzed, 5 *ts* mutant strains showed significant defects in splicing on introns 2 and 3, and not intron 1.
It is possible that these mutant strains, which displayed defects in splicing of the two downstream introns but no apparent defect in splicing of the first intron, represent ‘false positive’ discoveries which incorrectly showed a normal splicing phenotype for the first intron, for example because of an error in the PCR reaction of the precursor species for the first intron. To test whether these five mutant strains do in fact have splicing defects in downstream introns but not the first intron, qPCR was performed. Primers were designed to measure the pre-mRNA species for each intron, as well as actin transcript levels as a control. A WT strain was also measured, and each precursor level value was normalized to WT and log transformed. Of the five strains analyzed, only two strains, 103_E16 and 103_D24, recapitulated the phenotype that splicing of the first intron was unaffected but splicing the two downstream introns was defective, Figure 4.6. Additionally, to determine whether the effect seen in these two strains was specific to multi-intronic transcripts, analysis of the splicing efficiency for a canonical intron in a single intron containing transcript was performed on data previously obtained for \textit{fet5\_intron1} (Chapter 3). For both mutant strains, they showed no increased in SI values for \textit{fet5\_intron1}, suggesting that their splicing defective phenotype was for multi-intronic transcripts.
Figure 4.6 qPCR tests recapitulated the splicing defect in only downstream introns in *SPBC21B10.09* for two of the *ts* mutant strains. qPCR measurements of all three precursor levels in *SPBCB10.09* demonstrate the splicing defect in the downstream introns but not the first in only two of the candidate strains, 103_E16 and 103_D24.
Not all splicing mutants affect all transcripts in the same manner. Thus, to test if the observed splicing defects in mutant strains 103_E16 and 103_D24 had a more general effect and were not specific to splicing the three introns in *SPBC21B10.09*, qPCR primers were designed against each intron in a second multi-intronic transcript, *mug64*, which contains four introns. Again, this transcript was chosen because the introns had canonical features for *S. pombe*. The primers for the first and second introns behaved well, but the other two primers did not. Nonetheless, measurements of the precursor levels for the first and second introns in both mutants demonstrated that the first intron was unaffected in these mutants, while there was a defect in splicing *mug64_int2*, Figure 4.7, indicating a similar pattern seen in *SPBC21B10.09*. The observation of an unaffected first intron splicing event with a defect in splicing of a downstream intron suggested that the downstream splicing phenotype observed in the *ts* mutant strains 103_D24 and 103_E16 might reflect a broader phenotype for multi-intronic transcripts.
Figure 4.7 qPCR on a second multi-intronic transcript indicate that the downstream splicing defect not specific to *SPBC21B10.09*. qPCR measurements of precursor levels of second multi-intronic transcript, *mug64*, displayed a downstream splicing defective phenotype for both 103_E16 and 103_D24. Splicing of the first intron of *mug64* was unaffected in these mutants, but the splicing of the second intron was impaired, indicating a broader multi-intronic splicing phenotype.
4.2.2 Whole Genome Sequencing to Identify the Mutation Causing the Downstream Splicing Defect Reveals a SNP in *prp10*.

To determine the mutation responsible for the downstream splicing defect, bulked segregant analysis (BSA) followed by whole genome sequencing (WGS) was performed as described in Section 3.4.5. As a pilot test, BSA and WGS were done on only candidate 103_E16 to determine if we could identify the SNP causing the multintronic splicing defect. The *ts* progeny were pooled together and sequenced, and compared to the WT growth progeny that were also pooled together for sequencing. After filtering out the SNPs that were common to other mutant *ts* strains that were previously sequenced (Chapter 3), there were ~100 SNPs that were specific to *ts* strain 103_E16. No SNP showed segregation in only the *ts* pool and absent in the WT pool. However, analysis of the location of all the SNPs revealed a mutation in the U2 snRNP component, *prp10* (Figure 4.8). While this mutation does not appear to segregate with the *ts* phenotype, it is still possible that this SNP is the cause of the downstream splicing defect. This SNP results in the missense mutation E407K, which is in the HEAT repeat domains in Prp10205. Alternatively, there is the possibility that multiple mutant loci cause the splicing defect. Further genetic analyses and complementation tests will need to be performed to identify the SNP responsible for this phenotype.
Figure 4.8 Bases called at the SNP location in *prp10*. Here is the distribution of the bases called at the SNP locus in *prp10* in both the temperature sensitive pool and the WT growth pooled sample. The T-to-C mutation is 45% penetrant in the *ts* pool and 41% penetrant in the WT growth pool. This mutation results in a E407K missense mutation.
4.3 Discussion and Future Directions

In higher eukaryotes, most transcripts are interrupted by multiple introns, and spliceosome assembly occurs anew on each intron that it encounters within a transcript, underscoring the need to understand how the spliceosome is recruited to each intron within a multi-intronic transcript. Here I have described preliminary work to identify mutants that are defective in splicing of downstream introns within a multi-intronic transcript to begin to tease apart the mechanism by which the spliceosome assembles on these introns. A pilot screen through a conditional mutant library in *S. pombe*, an organism where almost half of its intron-containing genes have at least two introns, identified two *ts* strains that showed defects in splicing efficiency of the two downstream introns in the transcript *SPBC21B10.09*. Bulk segregant analysis and whole genome sequencing of the *ts* pool and the WT growth pools identified the putative causative mutation, a glutamate-to-lysine missense mutation at position 407 in *prp10*, a component of the U2 snRNP. This mutation was not enriched in the *ts* pool and absent in the WT pool, indicating that it may either not be the mutation causing the *ts* phenotype, or is not sufficient on its own to cause the *ts* phenotype. Additionally, whether the *ts* phenotype and splicing defect are genetically linked in this mutant strain has not been investigated. Complementation and other molecular tests will need to be performed in order to verify whether the *prp10-E407K* mutation is indeed causing the splicing defect observed. Introduction of the *prp10-E407K* allele into a wild-type background is the gold standard for demonstrating that this mutation causes the splicing defect observed and will need to be performed to determine that
the SNP identified is indeed causative for the downstream splicing defect. The rest of the discussion is based on the assumption that \textit{prp10-E407K} is the causative mutation for the downstream splicing defect.

Prp10 is an essential component of the U2 snRNP, as part of the SF3b complex, and it plays an important role in formation of the spliceosome commitment complex but directing the U2 snRNA to the branchpoint\textsuperscript{206}. Prp10 is a conserved in both budding yeast and humans. SF3b155, the human ortholog of Prp10, is the most frequently mutated spliceosome component in many different cancers\textsuperscript{207,208}. Recently, the structure of the SF3b complex in humans was solved by X-ray crystallography\textsuperscript{205}. This structure revealed the spatial organization of the cancer related mutations, and showed that the majority are clustered together in the HEAT domain repeats. Point mutations in this part of the protein lead to use of alternative branchpoint sequences, and subsequently cryptic 3’ splice sites, indicating how these cancer-associated mutations might promote the development of cancer. Interestingly, the glutamate residue that is mutated in our \textit{prp10-E407K} strain is conserved in human SF3b155 at position 545. This residue sits just upstream of the clusters of the previously identified cancer-associated mutations and is also in a HEAT domain, indicating that the \textit{prp10-E407K} strain might provide a resource to study how these mutations mechanistically affect splicing decisions and thus promote disease states in humans.

Additionally, SF3b155 has been demonstrated to associate with chromatin, associating with chromatin marks positioned over exons in metazoans. It has been
suggested that this association between SF3155 and chromatin facilitates the
spliceosome in identifying exons in the vast sequence space of metazoan transcripts.\textsuperscript{209}
The association of SF3b155 with chromatin was shown to be functionally important
for identification of splice sites, indicating that SF3b155 plays an important role in
conveying information encoded in the chromatin state to influence splicing decisions
and the recognition of exons. Disruption of SFb155 binding to nucleosomes resulted
in different alternative splicing patterns. These data are consistent with the
identification of a mutant Prp10 displaying defects in splicing of downstream introns.
If the \textit{prp10-E407K} mutant reduces the ability of Prp10 to associate with chromatin, it
could affect the ability of the spliceosome to be efficiently recruited and assembled
onto downstream introns, while the first intron may not be affected because of
spliceosomal recruitment by the CTD tail. Alternatively, if the associate of Prp10 with
the exonic region is important for identifying the exon or making cross exon
interaction, as is the case in spliceosomes that are recruited in an exon-definition
model, then the \textit{prp10-E407K} mutation would affect only downstream introns and not
the first.

Additional biochemical and genetic tests are necessary to elucidate the mechanistic
details underpinning how the SF3b complex functions in splice site decision and
spliceosome assembly, and the \textit{prp10-E407K} strain in \textit{S. pombe} provides a unique
opportunity to study these questions. One obvious next step would be to determine
the global splicing phenotype in this mutant strain. \textit{RNAseq} studies could be
performed to identify how splicing events are less efficient in this mutant strain. Comparisons of all the introns within each multi-intronic transcript would also provide useful insight into the splicing defect observed in this mutant. One could also compare the relative strength of the branchpoints in the introns that are spliced less efficiently and compare them to the strength of the branchpoints in the introns that are spliced more efficiently to determine if prp10-E407K primarily effects those introns with weaker branchpoint sequences, or if the defect is more general. Chromatin immunoprecipitation studies in both the WT and mutant strains would be interesting to perform to detect if there is a defect in association of Prp10-E407K with nucleosomes and if that is what causes the splicing defect. The genetic tractability of S. pombe provides an exciting opportunity to study mutations in splicing factors that have relevance in human disease, and it will be exciting to see what details about how SF3b affects splicing decisions can be learned from this mutant Prp10p strain.

Finally, components of the SF3b complex are not likely to be the only spliceosomal components necessary for efficient spliceosome assembly on downstream introns. This pilot screen performed on a subset of conditional mutants has provided us with a very interesting and exciting mutant candidate in prp10-E407K, but there are surely many more to be discovered in the remainder of the conditional mutants that have not been explored. Completion of the screen through the rest of the mutant library, followed by subsequent identification of the causative mutations will likely provide a plethora of interesting mutant candidates as well as the opportunity to study mutants
that will likely be relevant in human biology. Additionally, the sequencing-based screening method that we have developed can be used in conjunction with saturating mutagenesis of a single splicing factor of interest to elucidate the full complement of important residues within a single protein to fully understand its role within the splicing pathway. The conditional mutant library coupled with the wealth of structural information that has recently been solved, seem certain to provide exciting information about splicing mechanisms in the immediate future.

4.4 Author Contributions

B.J.F and A.L. performed the RNA extraction, cDNA synthesis and library preparations. A.L. processed and analyzed the sequencing data. A.L. performed the backcrossing, ts phenotyping experiments, and BSA and whole genome sequencing and analysis. A.L. and J.A.P. prepared the written analysis of this work. A.L. made the figures.
REFERENCES

30. Awan, A. R., Manfredo, A. & a Pleiss, J. Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily


44. Inada, M. et al. Phospho-site mutants of the RNA Polymerase II C-terminal


60. Monsalve, M. *et al.* Direct coupling of transcription and mRNA processing


71. Millevoi, S. *et al.* An interaction between U2AF 65 and CF I(m) links the splicing and 3’ end processing machineries. *EMBO J.* 25, 4854–4864 (2006).


161


111. Core, L. J. & Lis, J. T. Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science*. **319**, 1791–1792


122. Wilhelm, B. et al. Differential patterns of intronic and exonic DNA regions with respect to RNA polymerase II occupancy, nucleosome density and H3K36me3 marking in fission yeast. *Genome Biol* **12**, R82


160. Anver, S. *et al.* Yeast X-chromosome-associated protein 5 (Xap5) functions with H2A.Z to suppress aberrant transcripts. *EMBO reports* 15, 894–902


