GENOME-WIDE IDENTIFICATION OF SPLICE SITES BY LARIAT SEQUENCING

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

Nicholas Stepankiw

August 2016

GENOME-WIDE IDENTIFICATION OF SPLICE SITES BY LARIAT SEQUENCING

Nicholas Stepankiw, Ph.D.

Cornell University 2016

The coding regions of most eukaryotic genes are interrupted by regions of non-coding RNA, known as introns, which must be removed from precursor mRNAs (pre-mRNAs) for proper protein expression. Intron removal is catalyzed by the spliceosome, a large ribonucleoprotein that must identify intron boundaries with single nucleotide precision in order to faithfully generate translatable mRNAs. In higher eukaryotes it is now widely appreciated that proteome diversity is greatly expanded through alternative splicing, wherein the sites of spliceosomal activation are altered, often yielding multiple protein isoforms from a single pre-mRNA. Despite its importance for human biology, the mechanistic basis by which the spliceosome identifies and activates appropriate splice sites under a given set of conditions is poorly understood, owing at least in part to our limited abilities to identify the genome-wide complement of locations at which the spliceosome acts.

To better learn about splice site selection, I have used intron lariat sequencing, a technique developed in our lab that allows for high sensitivity detection of splicing events, to generate a comprehensive profile of splicing in the fission yeast, *Schizosaccharomyces pombe*. *S. pombe* provides an attractive model organism for studying splice site selection, both for its tractability and also for the similarity of it's splice site with humans splice sites. My work reveals an unprecedented level of alternative splicing for this model organism, including alterna-

tive splice site selection for over half of all annotated introns, hundreds of novel exon-skipping events, and thousands of novel locations of splicing activation. Moreover, the frequency of these events is far higher than previous estimates, with alternative splice sites on average activated at ~3% the rate of canonical sites. Although a subset of alternative sites are conserved in related species, implying functional potential, the majority are not detectably conserved, suggesting that these events reflect aberrant splice site activation. Interestingly, the rate of aberrant splicing is inversely related to expression level, with lowly expressed genes more prone to erroneous splicing, a finding for which the functional significance remains unknown. Together, my data suggest that the spliceosome possesses far lower fidelity than previously appreciated, highlighting the potential contributions of alternative splicing in generating novel gene structures. Having demonstrated the capacity of lariat sequencing to uncover novel splicing events across the genome, my ongoing work has focused on extending lariat sequencing to humans in order to more fully understand the mechanisms by which splicing alters the human transcriptome.

BIOGRAPHICAL SKETCH

Nick Stepankiw is a Texas born scientist who hails from Houston, Texas. In his youth, he was educated in schools following Montessori teaching philosophies. He holds a Bachelors of Arts in Computational and Applied Mathematics from Rice University. Following his undergraduate education, he worked at Baylor College of Medicine as a research technician in a group lead by Dr. David Bates and studied *Escherichia coli* chromosome segregation. Afterwards, Nick moved to Ithaca, NY and pursued his doctoral training in Biochemistry, Molecular, and Cellular Biology at Cornell University. His doctoral research was completed under the guidance of his co-mentors Jeffrey A. Pleiss, who specializes in yeast splicing, and under Andrew W. Grimson, who specializes in post-transcriptional RNA regulation. Nick investigated mRNA splicing from the viewpoint of RNA lariats and provided an incremental increase in our understanding of the extent of the fidelity of the spliceosome in selecting its splicing targets.

I dedicate this work to my parents for their love and support.

ACKNOWLEDGEMENTS

I grateful to and thank my parents, Elizabeth and Michael, for their love though all of my life. Their support has and will always be immensely important source of strength for the pursuit of my dreams. I must also thank my three siblings, Frank, Wendy and, Marika, for their continual friendship and encouragement in all matters of my life.

I thank my advisors, Dr. Andrew Grimson and Dr. Jeffrey Pleiss, for their support and their guidance in my training as a scientist. I also would like to thank my committee members Dr. Andrew Clark and Dr. Adam Siepel for their advice during my graduate research career.

To my friends I offer thanks for their support in my life both within and beyond science. To my colleagues I offer thanks for creating a strong and nurturing research environment during my graduate training. To the Stewart Little Cooperative, with which I was a member of for five years, I am grateful for the unique community living experience it provided during my graduate studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 The removal of introns is essential for gene expression

Living organisms depend on the utilization of genetic information found in genes. Genes are encoded as a sequence of DNA nucleotides and this sequence is used as a template for the creation of an RNA transcript corresponding to the DNA sequence. Some of these RNA transcripts are then used as templates for the creation of a protein for that gene. For many RNA transcripts, such as pre-messenger RNAs (pre-mRNAs), it is essential that the RNA sequence is altered to create a mature messenger RNA (mRNA), which is the functional RNA product of many genes. The process of splicing removes linear segments of the RNA, called introns, from a pre-mRNA transcript and rejoins the remaining portion of the transcript. Failure to remove these introns often results in mRNAs lead to the creation of aberrant protein products.

Alternative sequences may be recognized as introns which is a mechanism that regulates the function of many genes. This regulation of what sequence to conditional splice as an intron can depend on the cellular context and may result in a single gene producing multiple functional products[1]. This conditional splicing produces introns that are known as alternative introns and results in alternative mRNAs and this process is referred to as alternative splicing. Conditional or alternative splicing results in alternative splicing products that differ in their nucleotide sequences. Alternatively spliced mRNAs may result in production of alternative proteins and depending on the nucleotide and coding sequence can result in a single gene producing proteins with significantly

different and sometimes contrasting functions[2]. The flexibility to alternatively splice introns requires that the splicing machinery to precisely select the correct boundaries of introns for a given circumstance at one time, but also to alternatively splice the correct functional mRNA products in other circumstances. The flexibility of alternative splicing results in the splicing machinery recognizing a large range of sequences as introns, but the splicing machinery must also be able to avoid splicing of sequences that are similar to introns but are not actual introns. Understanding the balance of flexibility with precision in splicing is critical for understanding the regulatory actions of splicing. To improve our ability to understand how the spliceosome acts requires the identification of which sequences are spliced as introns. Understanding how introns are recognized is essential for understanding the functional activities of genes.

1.2 Introns are removed by splicing

RNA splicing acts on RNA transcripts to remove segments of nucleotides from the transcript, called introns, and to join the sequences flanking the intron together, which are known as exons. The simplest (in terms of number of factors involved) splicing of introns is found in group I and II introns[3], which both catalyze their own removal and rejoin the flanking exons together, each with their own different mechanism. Another type of intron, Group III introns, are not self-splicing. These introns accomplish their splicing with a chemistry and mechanism that is conceptually similar to group II introns, but with the aid of a large number of protein and RNA co-factors. Group III introns are prevalent in eukaryotes and a typical gene in many eukaryotes contain at least one intron. The splicing mechanism of group II introns is indicative of the chemistry of the

splicing of group III introns.

Group II self-splicing introns accomplish their self-splicing through the base pairing interactions between of several distant and conserved nucleotide regions in the intron. These base pairing interactions bring together several specific sites of the intron and promote two catalytic reactions during splicing. The first catalytic reaction is followed by rearrangements of the RNA tertiary structure and leads to a second catalytic reaction. These reactions remove the intron from the RNA and join the flanking exons into one contiguous RNA. Metal cofactors participate in the catalytic reactions, which are two sequential transesterification reactions. The location of these transesterifications define the boundaries of introns and exons (Figure 1.1. Additionally, these transesterification reactions define three specific sequences in an intron. These sequences are designated as the 5′ splice site (5′SS), branch point, and 3′ splice site (3′SS). The 5′SS is the 5′ most sequence of the intron the 3′SS is the 3′ most sequence of the intron. The branch point is a feature interior to the intron and is chemically involved with the 5′SS during the first transesterification reaction. The 3′SS is released when the upstream and downstream exons react during the second transesterification. Together, these sites participate in the catalysis of splicing and the removal of the self-splicing intron from the mRNA.

The reactions involved in of intron splicing results in both the removal of the intron and the creation of an unique RNA species called the intron lariat. First, the removal of introns is achieved by the two metal-ion facilitated transesterification reactions. During the first transesterification reaction, the phosphodiester bond at the 5′ position of the 5′SS participates in a nucleophilic attack on the phosphodiester bond at the 2′ position of the branch point. This attack

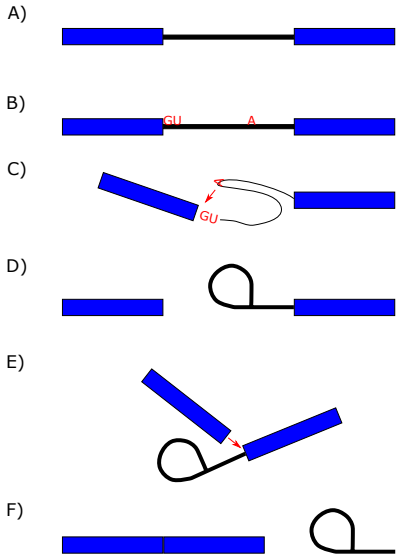

Figure 1.1: Steps in pre-mRNA splicing.

A) pre-mRNA diagram of exon-intron-exon structure. B) pre-mRNA diagram with GU dinucleotide of the 5′SS and the adenosine of the branch point indicated in red. C) The branch point adenosine performs a nucleophilic attack on the GU of the 5′SS. D) The result of the nucleophilic attack creates the splicing intermediate RNA lariat. E) The upstream exon's 3′ terminus performs a second nucleophilic attack on the phosphodiester bond between the 3′SS and the downstream exon. F) The result of the second nucleophilic attack is the release of the lariat and the joining of the two exons.

results in the freeing the 3′ position of the upstream exon and the formation of a covalent phosphodiester bond between the 5′SS and the branch point. The bond connecting the 5′SS and the 2′ position of intron interior branch point results in a circular structure with a RNA tail connected at the 3′ position of the branch point nucleotide. This circular RNA with a tail has a branched structure at the branch point and is known as an RNA lariat. At this stage the tail still contains the downstream exon. The next transesterification reaction occurs when the 3′-OH of the freed upstream exon participates with the RNA lariat in a second nucleophilic attack on the phosphodiester bond between the downstream exon

and the 3′ end of the 3′SS. This attack frees the 3′ termini of the lariat from the downstream exon. It also results in the formation of a covalent bond between the 3′ position of the upstream exon with the 5′ position of the downstream exon, ligating those two exons together in the spliced RNA transcript. The end product of these reactions separates the group II intron, as an RNA lariat, from the mRNA.

This splicing activity requires the specific positioning of nucleotides in the intron with the presence of metal cations in order to initiate and complete the reactions. Group II introns, acting as ribozymes, accomplish this task using cis secondary and tertiary structures. Many Eukaryotic introns are different group II introns and are designated as group III introns. These intron lack the RNA structures necessary for self-splicing and, instead, group III intron splicing is accomplished with the aid of a large complex of proteins and RNA referred to as the spliceosome.

1.3 The yeast spliceosome splices introns

The vast majority of eukaryotic introns are spliced by the spliceosome and are known as spliceosomal introns. The spliceosome is a large complex of RNAs and proteins that participate in the process of splicing introns. Similar to self-splicing group II introns, the spliceosomal introns are spliced by two metal-ion dependent transesterification reactions and produce both a spliced mRNA and a lariat intron. In order for the spliceosome to act on an intron, the spliceosome must recognize the splice sites involved in splicing the intron. Identification of the splice sites establishes the sequences needed for the transesterifications.

After identification of the splice sites, the splicesome also performs the rearrangements of the introns and exons that are involved in splicing the intron. Ultimately the spliceosome performs the same task as group II introns, both removing the intron and joining the adjacent exons. Much of our understanding of the spliceosome has come from work studying splicing in the yeast *S. cerevisiae*, owing to the tractability of yeast for studying the mechanisms of splicing.

The yeast *S. cerevisiae* spliceosome performs the splicing activities analogous to group II introns. These introns are relatively simple, with each intron sharing nearly identical 5'SS sequences, nearly identical branch point sequences, and nearly identical 3'SS sequences across the ~300 annotated introns. The nearly identical nature of the intron splice site sequences aids in the activity of the *S. cerevisiae* spliceosome in both recognizing introns and conducting the splicing reaction. In *S. cerevisiae*, the 5'SS and branch point are recognized by near-perfect base pairing interactions with the small nuclear ribonucleoproteins (snRNPs) components of the spliceosome, namely the U1 and U2 snRNPs (Figure 1.2). The identification of the branch point also generally identifies the 3'SS, often selecting the first downstream AG dinucleotide as the 3'SS, though this rule is not absolute[4]. After the splice sites engage in base pairing with U1 and U2, there are additional snRNPs that associate with the intron, namely U5/U4/U6, aids the yeast spliceosome in rearranging the structure of the intron to position the 5'SS and the branch point for the first transesterification reaction and results in the formation of the RNA lariat and the freeing of the upstream exon. Additional rearrangements occur to position the upstream exon and the 3'SS for the second transesterification reaction to both free the intron lariat and join the exons together. These structural and catalytic activities of the spliceosome require the aid of additional factors that associate at multiple stages dur-

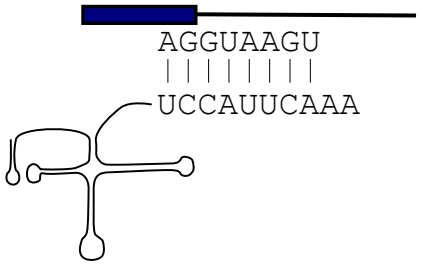

Figure 1.2: Diagram of U1 snRNA basepairing with 5′SS of an intron. Box indicates the exonic sequence and the line indicates the intronic sequence.

ing the splicing process making the spliceosome the largest ribonucleoprotein complex in the cell.

The splicing of spliceosomal introns is accomplished by over a hundred RNAs and proteins that interact with the spliceosome during different stages of the splicing process. The spliceosome is at its heart a ribozyme, with the U6 snRNA catalyzing both transesterifications [5]. The involvement of a large number of factors in the splicing process offers opportunities to regulate whether introns are spliced or retained in the final mRNA. How spliceosomal introns evolved as the main mechanism for the majority of intron splicing is still under debate, but one attractive theory is that spliceosomal introns arose from group II introns and that the spliceosome evolved as a more efficient mechanism for the splicing of these group II introns[6], but with extra regulatory options. While *S. cerevisiae* has few introns, the last eukaryotic common ancestor is thought to have had a high density of intron in its mRNAs, similar to modern eukaryotes like humans[6]. These introns likely had 5′SS and branch points that were degenerate, lacking near-perfect base pairing with the U1 and U2 snRNPs. Degeneracy in the sequence composition of splice sites necessitates the aid of ad-

ditional factors to correctly and efficiently identify which sequences are to be recognized as splice sites. The benefit of the involvement of such factors and their added complexity is the provision of additional opportunities for the regulation of intron splicing.

1.4 The human spliceosome acts on degenerate splice sites

Unlike introns in *S. cerevisiae*, the splice sites in human introns are degenerate. *S. cerevisiae* introns are often easily identified by near perfect base pairing between the splices sites and U1 and U2. To compensate for the degenerate splice sites in human introns, the human spliceosome utilizes additional trans-factors not present in *S. cerevisiae*, such as SR proteins and certain hnRNPs which may enhance or inhibit the usage of splice sites near where these trans-factors bind the RNA transcript[7, 8]. The human spliceosome display a high specificity in selecting which introns to splice.

SR proteins initially associate with the RNA polymerase and are transferred to the pre-mRNA after binding sites are transcribed[9]. The binding of SR proteins to the pre-mRNA often aids in the association of U1 and U2 to an intron, identifying the boundaries of the intron to splice, and enhancing the splicing of that intron in a given circumstance[7, 10]. The evolution of splice sites show evidence of coevolution with SR proteins, supporting their critical involvement in splicing[11]. Sometimes SR proteins are antagonistic and inhibit the splicing of an intron[12]. This enhancing or suppressing activity depends on where SR proteins associate relative to the splice sites[13]. Another class of proteins, hnRNPs, were initially thought to inhibit intron splicing but recent evidence

shows that they can also enhance the splicing of some introns[14]. The activity of SR proteins and hnRNPs as enhancers or inhibitors of splicing is complex, with some of this complexity being explained by the location of their binding sites relative to an intron[13]. Splicing enhancers and inhibitors are essential for the correct splicing of many human introns and give rise to a large number of transcripts[10, 15]. The involvement of SR and hnRNPs is essential for the correct activity of the human spliceosome.

The specific introns identified by the spliceosome may lead to an alternative transcripts(Figure 1.3). Modulating the location of splice sites can give rise to alternative introns and is often referred to as alternative splicing. Alternative mRNAs produced by alternate splicing may sometimes produce alternative proteins that can alter the function of the protein produced by a gene. In this manner, alternative splicing expands the proteome[16] and permits additional mechanisms for the regulation of gene expression[17]. The involvement of trans-acting factors is a known major factor in dictating whether an alternative intron is spliced.

1.5 Alternative splicing expands the proteome

Many human introns are alternatively spliced, where alternative splicing be the retention of the intron, or more commonly the intron is fundamentally altered through the use of a different 5′SS or 3′SS. The usage of alternative splice sites can lead to the splicing of a different, alternative intron that results in the removal as an alternative lariat intron. The language of alternative is used to indicate that there is another intron that generally needs to be removed by splicing

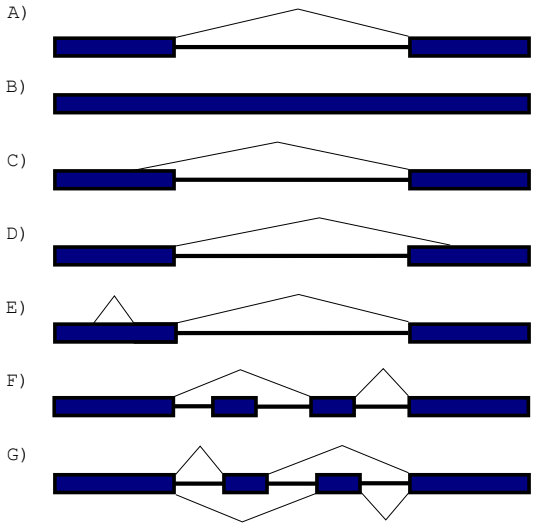

Figure 1.3: Diagram of alternative splicing in relation to intron-exon structure

A) Canonical intron splicing. B) Intron retention. C) Alternative 5′SS. D) Alternative 3′SS. E) Exitron. F) Exon skipping. G) mutually exclusive alternative exon skipping.

to produce a functional gene product. The removal of the alternative intron leads to an alternative mRNA isoform with nucleotide sequence that may code for different proteins, sometimes dramatically altering the function and activity of the gene.

There are many types of alternative introns. Alternative 5′SS is the pairing of an alternate 5′SS with a canonical branch point and alternative 3′SS is pairing of a canonical 5′SS with an alternate 3′SS with the choice of 3′SS frequently being the result of usage of an alternate branch point. Some alternative splicing is the result of introns in coding exons that are not normally spliced, referred to as exitrons[18]. Another type of alternative splicing is exon-skipping where the 5′SS of an upstream intron pairs with the branch point and 3′SS of a downstream intron and results in the skipping of all intermediate exons and introns. Because

much of the work to understand splicing has been done with the functional impact of coding potential of mRNAs, further types of exon-skipping alternative splicing have been described, namely mutually exclusive exon skipping and alternative first or last exons that change the UTR sequence of a genes mRNA[19]. Different organisms display different alternative splicing preferences, with *S. cerevisiae* using intron retention to control mRNA abundances and humans displaying almost all forms of alternative splicing, but however with a strong preference for gene regulation using exon skipping which can control both the function of a protein or even whether a functional protein is produced[19, 17]. Alternative splicing is a major regulator of the human transcriptome and the mechanisms of alternative splicing are of interest for understanding human gene expression, both in healthy tissues and diseased tissues. Alternative splicing in human is frequent in humans, with the majority of introns having minor isoforms of at least 10% in at least one tissue[20]. It is still under debate how much of alternative splicing is functional at the protein level and it may be that the benefit of having an abundance of alternative splicing provides a mechanism for the gain of regulated functional alternative splicing[19, 20].

Degeneracy of splice site sequences and alternative splicing is a property thought to be true of the greatest common ancestor and is present for many organisms today[21]. Splicing in the context of splice site sequence degeneracy requires the aid of additional trans-factors that are essential for the spliceosome to identify many of the introns spliced in humans, whereas the *S. cerevisiae* spliceosome alone accomplishes the task of splice site identification. For the catalytic mechanism of splicing, humans and *S. cerevisiae* share the same mechanisms for splicing. Like in group II introns and *S. cerevisiae* introns, human introns are spliced by the two transesterification reactions and splicing results in

both a spliced lariat intron and the spliced mRNA. The assembly of the human spliceosome is frequently dependent on the trans-factors that aid in selection of the initial splice sites. Other factors also regulate assembly of the spliceosome and splicing. These factors include RNA polymerase transcription speed, local secondary structure, chromatin, histones, gene promoters, and the abundances of various trans-factors[22, 23]. Together, a multitude of factors lead to complex patterns of alternative splicing that are hallmarks of human gene regulation by mRNA regulation. However, not all regulated alternative splicing events are functional and some may be detrimental. Cells have the ability to down-regulate alternative transcripts with aberrant behaviors.

1.6 mRNA surveillance removes some splicing mistakes

For coding genes, introns contain noncoding sequence that is removed from the final transcript by splicing. For canonical introns, failure to remove an intron by splicing may lead to a detrimental outcomes for the transcript due to the included intronic sequence disrupting the function of the gene product. Correctly spliced mRNAs will maintain their coding frame in sets of 3 nucleotides. Many introns are not a multiple of 3 nucleotides in length and a failure to splice these introns may lead to retention of that sequence in the mRNA and this retention may lead to a shift in the frame of the coding sequence that follows the intron, possibly resulting in the generation of a protein with no function and a highly disordered region. Often, this frame-shift leads to premature stop codons and during translation leads to the rapid decay of the transcript due to nonsense-mediated decay[24]. Additionally, this effect can even take place even when the retained intron is a multiple of 3 as intron sequences that are retained often

contain premature stop codons.

1.7 Discovering introns and annotating gene structure

Different organisms have different numbers of introns. The yeast *S. cerevisiae* has only a handful at ~300 with no annotated examples of alternative splicing producing functional alternative proteins. In contrast, the human genome has around 20,000-25,000 genes with an estimated 200,000 or more introns. Another yeast, the *Schizosaccharomyces pombe*, occupies a middle ground between *S. cerevisiae* and humans. *S. pombe* has about as many introns as genes. Other model organisms have differing intron densities. Additionally, these different organisms can have differing intron lengths, with *S. pombe* tending towards short introns around 80 nucleotides in average length, *S. cerevisiae* having their average length around several hundred nucleotides and humans having an average length of several thousand nucleotides. The prevalence of introns and their sometime complex regulation makes intron identification as important factor in understanding gene structure and gene function. The discovery and annotation of introns has been aided by several technologies that identify and quantitate splicing.

One approach for identifying introns is to examine the sequence of spliced RNA transcripts and comparing this sequence to the genes genome sequence as introns appear as gaps in the genomic alignment of the RNA transcript sequence[25]. However, because genome-wide RNA transcript sequencing samples RNA molecules based on their abundance, identifying all introns in a gene can prove challenging, especially if the intron being identified is an al-

ternative minor isoform[20]. Identifying introns is made more difficult as some introns may only be spliced in only a subset of possible conditions, such as stage in the cell-cycle, different stress conditions, stage of development, or tissue-specific intron splicing. Capturing the totality of splicing events requires both sequencing in high depth and also sequencing all possible cell states. While comparing RNA sequence to genomes has identify many introns, the use of comparative genomics has identified many introns based on properties of their genomic sequence.

Comparative genomics analysis of related genomes is a powerful tool that has identified many introns due to specific properties of their genomic sequence when that sequence is compared with the genomic sequence of related genomes[26, 27]. In general, there is a strong conservation of nucleotides in exons and a reduced conservation for the sequences of introns. Exon-skipping demonstrates greater levels of conservation of the interior of the intron compared with just the splice sites of the introns[28]. Canonical introns often show conservation of the nucleotide sequences at and around the splice sites and the rest of the intron sequence shows a significant lack of conservation. This strong conservation at the splice sites and lack of conservation in the rest of an intron is a signal that aids in predicting the presence of an intron. This approach is aided by computational and statistical analysis of other parameters that help define the likelihood of an intron existing at a particular sequence[29]. While computational approaches have greatly expanded our understanding of what sequences are identified as an intron, though current models do not account for the large number of infrequently occurring alternative splicing.

1.8 Intron splicing is typically quantified using mRNA based technologies

After an intron have been discovered and incorporated into a gene model, the next question regarding splicing is how frequently is the intron spliced for that gene. There are many ways to quantify splicing. Several of these methods require knowing the sequence of the spliced intron. Expressed sequence tags, genome tiling microarrays, exon-exon junction microarrays and reverse transcription PCR and quantitative PCR (RT-PCR and qPCR) are examples of such methods[30]. These techniques vary between looking at a specific intron and globally profiling splicing. A more recently developed technology, short read sequencing of reverse transcription products identifies introns by sequencing spliced mRNAs[31]. This technology is commonly referred to as RNAseq and is typically employed to quantitate mRNA sequence abundances. For quantitating mRNAs, RNAseq is typically first uses oligonucleotide dT selection to recover RNA that is poly-adenylated, then constructing a cDNA library from that RNA, and finally sequencing the cDNA using the Illumina short read sequencing platform technology[32]. Selecting for RNAs with a polyA tail is necessary to enrich mRNAs over rRNA, tRNA, and other noncoding RNAs making up the majority of RNA in a cell. Messenger RNA makes up around 1% of the RNA in a sample, depending on the organism, cell-type, and condition. Splicing is assayed by counting sequencing reads containing exon-exon junctions. Novel introns from previously unidentified exon-exon junctions can be discovered by mapping mRNAs to the genome. Because of the ease and utility of using RNAseq to measure splicing it is the leading technology for genome-wide intron quantitation.

Early uses of RNAseq generated sequencing libraries with several million short, approximately 30 nucleotide reads. More recently, libraries are sequenced to higher depth for the same cost, now at around a few hundred million reads with single read lengths extending above 100 nucleotides per read. Typically, though, experiments are done at a depth of tens of millions of reads for practical estimation of gene expression. Currently, 20-40 million short sequencing reads which align to the body of the gene transcript are typically used to estimate gene expression robustly for majority of gene expression. Deeper sequencing will provide better quantitation, but for the purpose of measuring the levels of expressed genes this sequencing depth is a good balance between cost and the benefit of sequencing. Learning about splicing events, however, can require more depth as only a small fraction of the reads align to a specific splicing junction, when compared to the total number of reads that align to the gene[20]. This limits quantitative utility of measuring splicing by RNAseq is the case of lowly expressed introns or alternative introns.

1.9 Splicing maybe be quantified using lariat cDNA

While RNAseq and other mRNA technologies have been popular for intron quantitation, a recently developed technology sequences cDNA from the spliced RNA lariat to quantify splicing events. The RNA lariat is unique in the cell due to having a circular and branched nature. The branch in the circle leads to a unique cDNA product of RNA lariats where the cDNA starts synthesis downstream of the 5'SS, synthesizes cDNA upstream until the reverse transcriptase encounters the 5'to 2'phosphodiester bond linking the 5'SS and branch point. At a low frequency it crosses this branch in the RNA, and continues syn-

thesizing cDNA upstream of the branch point 1.4. This lariat branch cDNA is unique as it gives both the branch point of the intron and identifies the boundaries of the intron with the pairing the branch point and 5'SS one a single read that can be mapped back to the genome across large distances[33, 34, 35, 36]. The 3'SS boundary is likely to be the first AG dinucleotide downstream of the BP, a property referred to as the AG dinucleotide exclusion zone (AGEZ)[37], though the extent that this rule applies to minor isoforms representing mistakes is unknown. One challenge in using lariat branch cDNA is that reverse transcriptase rarely makes this cDNA products due to the branched RNA not being a normal substrate. Additionally, there are frequent mutations associated with the branch point in these sequencing reads which can further complicate their mapping to the genome. An additional challenge to these approaches is that some introns are quite large and much of sequencing depth in organism with many large introns leads to a large fraction of the sequencing library consisting of lariat cDNA that does not correspond to the lariat branch. The low rate of synthesis limits the depth of sequencing possible using lariat branch cDNA, though valuable and novel information is discoverable even at low depth. This novel information includes the *in vivo* position of an introns branch point and that lariat abundances are likely not tied to the regulation of the abundance of their associated mRNAs.

Lariat derived cDNA has been used to address a number of biological questions. One such question is what are the position of an introns branch point? Another more involved question is what are the differences between splicing patterns assayed by lariats when compared to current RNAseq intron splicing analyses? Addressing this question is made difficult by limitations in performing high-depth lariat branch sequencing, primarily because of the reduced ability of

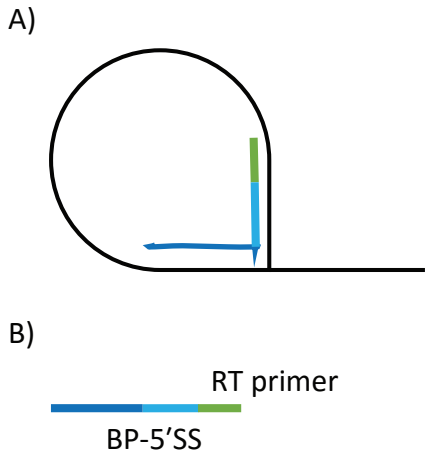

Figure 1.4: cDNA of lariat branch

A) A reverse transcription primer (green) on a lariat RNA initiates a cDNA product that contains sequence including the 5'SS (light blue), crosses the 2'-5'phosphodiester and contains sequence of the branch point (dark blue). B) The resulting cDNA.

reverse transcriptase to create lariat branch cDNA, and also because sequencing large intron lariats produces many lariat body reads that titer away sequencing potential from lariat branch cDNA. The limitation of sequencing depth currently makes it difficult to quantitate splicing events that require deep coverage. Lariat branch sequencing has been performed in several model organisms, such as the yeasts *S. pombe* and *S. cerevisiae*, *Drosophila melanogaster*, and also in humans. This leaves open questions regarding the lariat branches of other model organisms, such as *Caenorhabditis elegans*. Other questions also remain, such as are there examples of spliceosomal mediated decay of mRNAs where the second step of splicing is stalled and results in the pre-mRNA decaying[38]. The question of how accurately do lariats estimate actual splicing frequency remains unanswered, though it is known that some lariats accumulate disproportionately to others due to primary and secondary sequence factors[39]. Given both

that lariat decay is governed by an alternate pathway and the possibility that the stability of exceedingly long lariats may have a length dependency, it is hard to judge the meaning of the expression levels of lariat sequencing, though comparisons with transcriptional rates may give insight into the interpretation of the meaning of their expression levels. Regardless of the accuracy of expression levels, questions that involve comparative analyses between two samples will likely yield biological insights once the sequencing depth problem is solved. A whole class of questions can be investigated that involve investigating the quantitative regulation imparted by splicing factors that are mutated, knocked-down, or deleted. Looking at the effect of specific splicing factors on specific introns would be interesting as current mRNA based sequencing approaches for estimating intron splicing frequency are convoluted by the regulatory effects of all splicing isoforms.

1.10 Approaches that identify and quantify introns using lariats

The ability of reverse transcriptases to read across 2′ modifications and lariat branch points have been known for decades[40] One of the first attempts to globally characterize spliced intron branch points was done using reverse transcriptase PCR on dozens of human introns using primers upstream and downstream of the 5′SS-branch point junction[41][42]. These PCR products were then cloned into plasmids and were used as templates for Sanger sequencing. It was the then largest study of human branch points at the time and characterized the consensus of the human branch point to be the 5 nucleotide sequence yUnAy.

One observation noted by the study is that reverse transcriptase frequently produces mutations across the branch point, replacing what should be an adenosine with a uracil. At the DNA level, this results in an A to T transversion mutation that frequently characterizes lariat branch reads. This fact has been used by later studies as validation that the unique lariat branch cDNA is in fact the result of reverse transcriptase reading across this junction.

Several works have looked at lariats in high-throughput sequencing libraries. The first analysis reported found lariat branch crossing reads in polydT oligonucleotide selected human RNAseq libraries and rRNA depleted RNAseq without polydT oligonucleotide selection[33]. Presumably the lariats found in the polydT oligonucleotide selected RNA are the result of either splicing intermediates where the second step of splicing has yet to occur or else they are the result of spliced intron lariats that have significant stretches of adenosines and are therefore selectable by polydT oligonucleotides. Regardless of the source of these lariat branch reads , this meta-analysis found 2,118 lariat branch reads in 1.2 billions RNAseq reads. Approximately one in a million lariat branches represents an extremely small coverage of lariat branches in that publication, but as the first study to globally characterize lariat cDNA and it provided valuable insight into branch point biology. This data describes 864 unique lariats for 760 introns which leaves hundreds of thousands of lariat introns undescribed. These *in vivo* discovered branch points were used to examine properties of several RNA binding proteins that associate with the branch point, showing, for example, that hnRNP C binds both upstream and downstream of the selected branch point. Additionally, introns that utilize multiple branch points displayed lower levels of branch point conservation when compared to introns that only have one discovered branch point.

In addition to finding RNA lariats in normal RNAseq libraries, the enriched sequencing of RNA lariats may accomplished by several techniques designed to take advantage of the unique structure that is a lariat RNA. One such approach sequenced lariats in *S. pombe* and enriched for the lariats by both the differential electrophoretic separation of circular lariats from linear RNA through the use of 2D denaturing PAGE. This approach also further enriched for lariats by using a strain with reduced decay of spliced RNA lariats by deletion of the lariat debranching enzyme (Dbr1)[34]. The majority of data for this approach is for lariat body mapping reads, with a small fraction of the reads mapping to lariat branch cDNA. This small fraction was significantly more than the previous genome-wide work[33]. The lariats from this 2D gel approach were used to investigate whether there are novel regulated splicing events in the fission yeast *S. pombe* that have been missed in all the then current RNAseq data sets. Conserved and regulated exon skipping was discovered, demonstrating that this yeast displays mammalian-like exon skipping that is deeply conserved across many clades including humans. In one such example, an exon of alp41 demonstrates a conserved reading frame length across nearly all organisms that include that exonic, a fact supporting the conserved function of that exon. Another interesting exon-skipping involves the gene Srrm1 with a conserved frame-preserving skipped-exon that contains a PWI RNA binding domain, resulting in an alternate RNA binding protein with different RNA binding activity. The skipping of this exon is greatly increased by temperature shock, suggesting a possible regulated role for this exon skipping event. There is significant value in the enrichment for RNA lariats and a natural follow-up is to apply such enrichment strategies to human samples.

There are several general strategies for enriching for lariats in human RNA

samples. The fundamental problems with enriching for human lariats is their abundance is expected corresponds to a very small fraction of the transcriptome.. One useful approach for overcoming these limitations is to degrade non-circular RNA using the exonucleases RNAse R, a 3′-5′ exonuclease or the enzyme Xrn1, a 5′-3′ exonuclease. A second approach is to remove rRNA, which typically makes up over 85% of the RNA in a sample. Removal of rRNA is accomplished by hybridization pull-down or hybridization based cleavage by RNAse H. The removal of rRNA may be used in combination with the former approach. Other options for enrichment are to knockdown the lariat decay pathway[39, 43], though reducing lariat decay could perturb normal regulation of non-lariat RNAs making this approach of limited expected utility for addressing questions that depend on maintaining proper global gene regulation. Additionally, there is another enzyme with some lariat debranching activity, though the extent of its activity in humans is currently unknown[44]. Another approach to enrich for lariats is to perform oligonucleotide hybridization directly to the oligonucleotide lariat and to pull out that RNA from a sample[45]. These tools together have been used to answer a handful of questions regarding human RNA splicing and lariat biology.

It was found that spliced lariats can have regulatory potential while performing a genome-wide loss-of-function TDP-43 toxicity suppressor screen in *S. cerevisiae*. TDP-43 is a gene where some mutations lead to neuronal diseases and the question the researchers were intending to answer is what other proteins impact the toxicity phenotype. This screen found that reduced activity of DBR1, either deleted in yeast or additionally knocked-down in humans leads to suppression of the toxicity due to TDP-43 sequestering to some spliced intron lariats. Further work along these lines show that some circular intron lariats

(ciRNAs) that presumably escape debranching due to secondary structural elements will accumulate and display localization to sites of ankrd52 transcription where it positively regulates levels of the ankrd52 mRNA, indicating the potential of ciRNAs to be involved in unique regulatory roles[39].

Another study looking at *S. pombe* took the approach of not using 2D gels but instead doing rRNA depletion on yeast strains missing the lariat debranching enzyme[36]. This study yielded ~0.2% percent of its reads mapping to RNA lariat branches and identified many branch points for about half of the annotated introns. This work sought to answer what fraction of a Δdbr1 sequencing library is RNA lariats and to share a tool for rapidly mapping annotated lariats. One notable biological finding of this study was that the *in vivo* branch points in *S. pombe* have similar consensus to the predicted branch points. This study also discovered 94 lariats for exon skipping introns, further confirming the presence of exon skipping in yeast[34]. Additionally, this study sought to answer whether mis-regulation of intron splicing occurs due to the deletion of DBR1 by mapping the pre-mRNA to mRNA ratio using exon and 5'SS crossing reads verse the exon-exon junction reads. They found a significant increase in intron retention.

Another recent study used a different approach to pull down RNA lariats using oligonucleotide hybridization to sequences just downstream of known 5'SS sequence and sequences upstream of predicted branch points in intron lariats[45]. This work achieved an approximately 100-fold increase in lariat enrichment over background levels and with 133 million intron aligning reads they recovered 532,405 lariat branches, describing 59,359 high-confidence branch points for almost 25% of known introns. This work sought to answer the

question of whether there are populations of unique branch point sequence and surrounding features. The conservation and sequence surrounding the branch points was queried and five pentamer branch point submotifs were found to be significantly over-represented when compared to other pentamers that were both conserved and have similarly strong U2 binding strengths. These over-represented branch branch points are associated with different polypyrimidine tract sequence preferences and comparison with more distantly related metazoans reveals the evolutionary path for branch point sequences. These and other findings regarding the frequency of disease associated SNPs appearing in or around the branch point sequence offers mechanistic insights into branch point biology.

And yet another technique uses a debranching null *S. cera* strain or a WT strain with spliceosome co-immunoprecipitation to isolate lariats and follows isolation with debranching of the lariat[46]. This work sought to compare how well lariat branches could be characterized when there isn't deletion of the lariat debranching enzyme. This work chose to isolate lariats and then debranch which is notable because it avoids the severe penalty of creating cDNA across the lariat branch. The downside to the implementation of this approach is it decouples the 5'SS and branch point pairing information. The choice to characterize only the termini of the intron lariat gives the advantage that the entire branch point sequence is present in the debranched lariat. This contrasts with the mutations that lead to inexact inference of branch point sequence due to the property that reverse transcriptase crossing the lariat branch frequently results in mutations and deletions around the branch point. The choice to immunoprecipitate introns using antibodies to prp16 demonstrated that alternate methods can be used to enrich for lariats. The downside to this approach is it requires

working with large volumes of cells, but it may provide an avenue for isolating human lariat branch points and 5'SS with great depth.

1.11 Fidelity of introns splicing: sensitivity and specificity

The spliceosome acts with great sensitivity and specificity on a large number of mRNAs. One task of the spliceosome is to select the correct splice sites for intron splicing. This selection must find the correct site among the many sites that have sequences that are similar to the degenerate sequence of splice sites. This is done in part with the aid of snRNAs U1 and U2. In organisms with degenerate splice site sequences, additional factors are essential for the correct splicing of many introns, such as SR proteins and hnRNPs. Additional factors can also contribute to the identification of splice sites, for example the rate of transcription by RNA polymerase, the secondary structure of the RNA, or the levels of available trans-factors[47, 48]. The extent of the spliceosomes specificity in identifying introns may be a balancing act between the potential deleterious effects of selecting the wrong splice sites with the cost of maintaining hyper-accuracy in splice sites selection. Errors in splice site selection may represent a failure of the spliceosome to act specifically. The extent of errors in splice site selection have been estimated several ways, accounting for different types of errors. These error estimates range from one error in ten to one error in tens of thousands of splicing events. One caveat to these estimates is that they have been done using mRNA quantification of splicing events which is biased by processes that regulate mRNA decay. Regardless, these studies provide a first approximation of error rate in splicing and begin to answer several questions regarding the global rates of splicing errors. Questions range from broad, such as what is

the frequency of splicing errors, to more narrow questions such as what is the error in correctly performing exon skipping of transcripts without annotated minor isoforms, how frequently incorrect erroneous alternate 5'SS and 3'SS are selected, and what fraction of mRNAs produced are expected to have at least one error. The extent of the spliceosomes fidelity can depend on the severity of the impact of of low fidelity. This leads to additional questions mostly regarding what additional parameters impact fidelity and what are the properties of introns with extremes differences in their fidelity levels.

Evaluating the spliceosomes sensitivity and specificity can be challenging due to the uncertainty in identifying splicing events that results from false positive activation of splice sites as opposed to those that serve a biological role, either in producing a functional protein or functionally regulating the mRNAs expression level. Defining erroneous events requires categorizing events as functional or not. This can be difficult in organisms like humans, which display a high degree of alternative splicing complexity[20]. One possible guideline for evaluating whether a 5'SS or branch point has functionality is whether the utilized alternate splice site display a high degree of conservation. For exon skipping events, one possible guideline is whether the exon skipping mRNA isoform is annotated. Alternate splicing events partitioned into functional and non-functional categories for producing an estimate of the global splicing error rate, though in this case these might not represent errors, but could represent regulation of transcript abundance. Evaluating splicing errors using mRNA based splicing quantitation is a first approximation of the splicing error rate since errors are known to lead to degradation of splice mRNAs. Analyzing splicing sensitivity and specificity using spliced lariats may address these caveats as lariats display short half-lives and presumably the majority of lari-

ats not regulated in a manner similar to mRNAs. However, since mRNA-based splicing estimates are easy to perform, initial explorations of splicing specificity have been made using EST, quantitative RT-PCR, and RNAseq.

1.12 Splicing error estimates range from rare to frequent

To answer what is the spliceosomes intrinsic error-rate, one group looked at exon skipping events where a human cell type has a constitutive isoform, but there are also many upstream exons that could participate in exon skipping but does not based on the lack of an annotated isoform skipping those exons[49]. This work performed quantitative RT-PCR to estimate the intrinsic splicing error rates to be in the range of $2x10^{-3}$ to $6x10^{-6}$. Such a high error rate places many these splicing errors in the same order of magnitude as would be explained by transcription errors from RNA polymerase interrupting the intended constitutive splice site. This error rate was found to be modulated by levels of a master assembler of spliceosomal components, namely the gene Survival of Motor Neuron. Together these observations were considered support for the idea that all multi-intronic genes undergo exon-skipping at some level, possibly due to transcriptional errors modulating the strengths of splice site motifs.

Another work sought to answer what fraction of human mRNAs contain at least one error due to an erroneous splicing event and explored three models of the splicing error rate to address whether the splicing error rate is reasonably approximated as a constant across all introns[50]. The analysis was done using expressed sequence tags and microarrays to quantitate splicing events in a human cell line. This work operated under the assumption that only the ma-

jor isoform of a gene is the only correct mRNA, which is likely to not be true for many transcripts. Despite this limitation, the number of splicing errors in a transcript were measured it was found that an mRNAs erroneous splicing frequency is inversely linked to the number of introns in an mRNA and the level of expression of that mRNA. The decreased error-rate in splicing for mRNAs with a large number of splicing events makes sense as too high of a rate of erroneous mRNAs would be detrimental for the cell. This work estimated the mRNA error rate at around 1-10% a surprisingly high number, though it is likely that this is an overestimate as some genes will have multiple functional mRNA isoforms that in this study are being counted as part of the error-rate. Even so, the expectation is that the majority of alternate isoforms are errors and this work demonstrates that the degree of specificity of intron splicing is linked to both the total number of introns and the mRNA expression level.

A more recent study sought to answer what fraction of alternate splicing is non-functional[51]. To do this they investigated alternate splicing events that are not exon skipping events. These alternate events are mostly not conserved, suggesting they are not functional and are thus due to the spliceosomes background activation rate of near-cognate splice sites. Examining alternate introns from human RNAseq data and considering alternate sites showing weak or no conservation, this study found that the overall alternate rate is ~0.7%. The mean fraction of alternate nonconserved splicing events correlates with the length of introns, with larger introns having mean alternate rates above 2%. Splicing enhancers are enriched around these alternate splice sites, suggesting their possible involvement in activating the alternate splice sites.

Another recent study sought to answer questions regarding the prevalence

of alternative exon skipping in *S. pombe* and the possibility that exon skipping is widely regulated in some conditions. Previous work demonstrated the presence of condition responsive exon skipping *S. pombe*[34], a model organism that was previously regarded as not having complex alternative splicing[52, 53]. This study used a meta analysis of RNAseq data to examine the prevalence of exon-skipping events in *S. pombe* and found exon skipping events in *S. pombe* are widespread and pervasive, but alternative exon skipping of alternative events of specific transcripts are generally not highly abundant. Additionally, the conditions they analyzed did not demonstrate widespread condition specific regulation of introns. In general the alternate rate of exon skipping events was calculated to be around 0.2%. In two conditions analyzed, deletion of dhp1 and late meiotic differentiation, the alternate rate rose to 1.7%. These two conditions have reduced levels of the nuclear exosome, indicating a possible role of the nuclear exosome in suppressing apparent levels of alternative splicing and suggesting that the actual rate of alternative splicing is above the level present in normal conditions. Additionally, this work found the existence of additional alternative 5'SS and 3'SS that appear to also be regulated by these events. Whether the increase in these alternative events are artifacts of the conditions surveyed and how the nuclear exosome is able to identify these aberrant events that are down-regulated remains an open question.

Together these studies address some of the basic questions regarding the background rate of activation of alternate near-cognate splice sites. Parameters such as expression level, intron length, and the presence of splicing enhancers have been found to correlate with increased activation of presumably nonfunctional alternate splicing events. RNA regulatory mechanisms such as nonsense mediated decay may mask some of the alternative splicing that occur and the

nuclear exosome may further mask alternative splicing events in yeast. Understanding the transcriptome from the perspective of spliced RNA lariats may uncover additional examples of regulation due to stalled or stabilized RNA lariats. Additionally, some mRNAs may be vanishingly rare due to rapid decay of the spliced mRNA while the spliced lariat may decay at the normal rate. This leads to questions of what transcribed mRNAs are we missing due to their regulated rapid decay. One questions that further remains to be answered is what the background rate of nonfunctional splicing is when quantitated by spliced RNA lariats. This is of interest because discrepancies between mRNA splicing quantitation and lariat quantitation could further implicate the involvement of additional regulatory processes that mask the true error rate. In order to address this question requires sequencing lariats with considerable depth and independent of intron annotation, a feat currently possible with a specific lariat-stabilizing strain of the yeast *S. pombe*. To do this using human lariats requires the development of a high-depth lariat sequencing technology, a feat that has yet to be completed and may be difficult because of the rarity of spliced RNA lariats.

CHAPTER 2

LARIAT BRANCH SEQUENCING IN *S. POMBE*

This work is available online with NAR[35]

2.1 Abstract

Alternative splicing is an important and ancient feature of eukaryotic gene structure, the existence of which has likely facilitated eukaryotic proteome expansions. Here, we have used intron lariat sequencing to generate a comprehensive profile of splicing events in *Schizosaccharomyces pombe*, amongst the simplest organisms that possess mammalian-like splice site degeneracy. We reveal an unprecedented level of alternative splicing, including alternative splice site selection for over half of all annotated introns, hundreds of novel exon-skipping events, and thousands of novel introns. Moreover, the frequency of these events is far higher than previous estimates, with alternative splice sites on average activated at ~3% the rate of canonical sites. Although a subset of alternative sites are conserved in related species, implying functional potential, the majority are not detectably conserved. Interestingly, the rate of aberrant splicing is inversely related to expression level, with lowly expressed genes more prone to erroneous splicing. Although we validate many events with RNAseq, the proportion of alternative splicing discovered with lariat sequencing is far greater, a difference we attribute to preferential decay of aberrantly spliced transcripts. Together, these data suggest the spliceosome possesses far lower fidelity than previously appreciated, highlighting the potential contributions of alternative splicing in generating novel gene structures.

2.2 Introduction

The process of pre-messenger RNA (pre-mRNA) splicing, mediated by the spliceosome and accessory proteins, removes non-coding introns, which are found throughout most eukaryotic transcripts[54]. Interactions between the spliceosome and sequences within introns are central to the mechanism of splicing. The essential sequences are the 5′ and 3′ splice sites (5′ and 3′SS), found at the termini of the intron, and an internal sequence known as the branch-point (BP)[55]. Along with spliced exons, excised introns are released from the spliceosome as a lariat, in which the 5′SS is covalently attached via a 2′-5′ phosphodiester linkage to an essential adenosine within the BP sequence[56]. Thus, the progress and patterns of splicing can be monitored by methods that detect either the mature spliced mRNA or the intron lariats.

In higher eukaryotes, alternative splicing provides a powerful opportunity for both regulation of gene expression and generation of proteome diversity[57]. Mechanistically, alternative splicing derives from control of splice site selection[58]. The splice site sequences in mammalian introns are highly degenerate, often requiring the activity of auxiliary proteins to enhance activation of sub-optimal sequences[59]. Because of the low information content at many mammalian splice sites, nearby cryptic sites can be activated, resulting in the production of alternative splice products, many of which are subjected to degradation by RNA quality control pathways[60]. Several estimates of the intrinsic rate at which the spliceosome aberrantly generates alternative splicing products have been previously published. A quantitative RT-PCR based study of mammalian splicing suggested an exceptionally low rate of aberrant exon skipping, on the order of one error for every $10^{-3} - 10^{-5}$ splic-

ing events[49]. By contrast, computational modeling of exon skipping and alternative splice site usage within mammalian EST data predicted higher error rates that were variable, dependent upon expression level and intron number of the host transcript[50]. And finally, analyses of large RNA-seq datasets have identified alternative events at a rate of one in $10^{-2} - 10^{-3}$ splicing events[61]. Importantly, it is unclear whether any of these approaches are appropriate for defining the actual spliceosomal error rate given that they predominantly measure mature mRNA levels, which derive from both rates of splicing error and preferential decay of mispliced products.

Here we examine global splice site selection in the fission yeast, *Schizosaccaromyces pombe*. Unlike budding yeast, splice site sequences in *S. pombe* are highly degenerate and comparable to those found in mammalian introns[62, 63]. Furthermore, nearly half of all *S. pombe* genes contain an intron, and nearly half of those contain multiple introns. We previously described a lariat sequencing approach that allows for high depth analysis of splicing events[34]. Here, we have improved upon this approach to generate the most comprehensive dataset of experimentally identified splicing events in an organism with degenerate splice site sequences. We observe a large number of alternatively spliced isoforms, a subset of which correspond to conserved alternative splice sites. Remarkably, we also observe a rate of alternate splice site selection that is greatly higher than previous estimates. Together, these data provide compelling evidence suggesting that spliceosomal infidelity is widespread, and provides a powerful mechanism by which alternative gene structures can evolve[64].

2.3 MATERIALS AND METHODS

2.3.1 Strains used

yNZS005 was used as the WT strain and was from ATCC (Linder 972, #38366). yNZS006 contained a deletion of the dbr1 locus from yNS005 and was generated as previously described[34]. yNS008 contained a deletion of the upf1 locus, and was taken from the *S. pombe* Haploid Deletion Collection from Bioneer, and yNS007 was the matched wild type strain from this library.

2.3.2 Yeast cultures

Unless otherwise noted, yeast cultures were grown according to standard protocols[65]. Cells were harvested by filtration thru Millipore HAWP0025 filters when OD600 measurements were between 0.8 and 1.0. For heat shock samples, after reaching the noted OD600 the cultures were shifted to 37°C for 15 minutes and then harvested. For the diauxic shift samples, the cultures were allowed to grow until the OD600 reached 7.6.

2.3.3 Two-dimensional (2D) gel electrophoresis

Denaturing acrylamide gels were polymerized with 7.5 M Urea with varying acrylamide concentrations. Two sets of gels were run: one designed to isolate shorter lariats and the other to isolate longer lariats. For shorter lariats, the acrylamide percentages were 7.5% and 15% for the first and second dimensions,

respectively. For the longer lariats, the percentages were 4% and 8%. For both gel types, 40 – 60 μg of Δdbr1 RNA was loaded onto the first gel. The first gel was run at 200 V for 2 h and the second gel at 230 V for 2 h followed by 290 V for 1 h. The entire lane containing the RNA was cut from the gel, rotated, and recast at the top of the second gel. The second gels were run until the xylene cyanol dye migrated 5.5 or 9 cm for the short and long lariats, respectively. Gels were stained, visualized (Dark Reader or Typhoon) and lariat arcs were excised. RNA was extracted according to standard protocols.

2.3.4 Lariat sequencing library construction

The RNAs from 2D gels were used to construct non-stranded multiplexed libraries using a custom protocol that placed appropriate sequences for multiplexed Illumina-sequencing. The first strand and second strand synthesis was done using Invitrogen second strand synthesis kit and 500 ng of dN_9 primer. Because of the high levels of dN_9 present in our first strand reaction, we omitted the Escherichia coli DNA ligase from the second strand reaction. After second strand synthesis, the products were run on a native gel and the library was recovered between ~20 and 300 nucleotides. The sized material was eluted from the gel by adding 4x volume of 0.3 M NaOAc pH 5.3. The DNA was ethanol precipitated by adding 2.5 volumes of 95% ethanol, incubating at -20°C, spinning at 14,000 x g for 20 min, washing twice with 70% ethanol for 10 min and resuspending in 10 μl H_2O. For all subsequent steps, enzymatic reactions were purified using phenol:chloroform extraction followed by ethanol precipitation. DNA end repair was performed using NEB Next End Repair Module. Next, dA tailing was performed using the NEB Next dA-tailing Module. Adapter liga-

tion was performed using T4 DNA Ligase (Rapid) from Enzymatics and 1 μl of Illumina barcoded adapter. The resulting ligation product was sized on a denaturing urea gel between 130 and 300 nucleotides. This was precipitated and sequenced on an Illumina Hiseq 2000 with single-end 100 nucleotide reads.

2.3.5 Lariat sequencing genome alignment

Illumina sequence reads were trimmed of the 3' adapter using Trimmomatic[66] with parameters :3:30:10 and MINLEN:18. Trimmed reads were aligned to the *S. pombe* genome (pombase.org, Schizosaccharomyces_pombe.ASM294v2.21.dna.genome.fa) using Bowtie2[67]. Parameters of alignment were score-min L,-0.4,-0.4 -very-sensitive. End-to-end alignments were done for the initial genome alignment. Paired end alignments on split reads were done using -ff -I 20 -X 3000 -no-mixed -no-discordant.

2.3.6 Splice site scoring

Log-odds scores for splice sites were computed from a Position Weight Matrix (PWM)[59]. The 5'SS was scored either using the dinucleotides within the first nine nucleotides of annotated introns or using the dinucleotides within three nucleotides upstream the 5'SS through the first nine nucleotides of annotated introns for the foreground signal and the dinucleotides of annotated intron sequences for the background signal. The former was used during splitting reads only, while the later was used for all analysis after split read identification. The putative BP was scored using an 8 position dinucleotide model using branch

points identified from the aligned_introns.txt file from pombase.org for the foreground signal and the dinucleotide composition of introns for background signal. The BP and 9-nucleotide 5'SS scores have a similar maximal value in this scheme.

2.3.7 Branch spanning split read alignment

For identification of branch spanning reads, each read that failed to align to the genome in end-to-end fashion was split into paired-end reads at every GT dinucleotide, including those that appeared in the reverse complement of the strand. For alignment, we only considered the subset of these reads for which the splitting process produced two fragments of at least 10 nucleotides, and excluded all others. These reads were then assessed for alignment using Bowtie2 in paired-end mode with the fragment containing the GT as mate 1 and the other as mate 2, and using the previously noted parameters. Since a single read split in this fashion can yield several possible alignments, possible GT split alignments were collapsed into a best available paired-end alignment that minimizes the number of mismatches in the alignment. As an additional criterion to judge alignment quality, log-odds scores were calculated for the putative 5'SS (GT end of the alignment) and BP (the other end of the alignment): the combination of these scores was required to be greater than 0; this strategy was used to both break ties and reduce the number of artifactual alignments.

Reverse transcriptase frequently introduces deletions and mutations when creating the cDNA product that crosses the 5'SS to the BP[33]. To determine the likely branch point nucleotide, BP scores were computed between two nu-

cleotides upstream of the non-GT end of the paired alignment and three nucleotides downstream of the read end and the position of the BP was determined by the maximal BP score in that range. Due to the to the high rate of mismatches from reverse transcriptase reading across the branch point of the lariat, up to two mismatches within +/- 1 nucleotide of the putative branch-point adenosine were disregarded for total mismatch calculations. The position where the BP minimized the number of mismatches was considered the branch point. If the branch point score caused the total log-odds scores to fall below 0, then the read was considered to fail alignment.

At a low frequency, the heuristics of Bowtie2 are such that it can fail to find an alignment when an acceptable alignment is possible. At a low rate this results in incorrect alignments where a nearly correct version of the GT iteration with some mismatches has an alignment but the correct alignment failed to align. Found alignments were assessed for better nearby alignments by appropriately shifting sequence from one end of the alignment to the other side. This shifted sequence was checked to see if it decreased the number of total mismatches in the alignment. The BP was then reevaluated in this alignment, as described above. Additionally, since Bowtie2 may have failed to correctly align a read to the genome in end-to-end fashion, found split read alignments were assayed for this by anchoring one end of the mate-pair to the genome and checking if recreating the original read leads to the non-anchored mate aligning using the Smith-Waterman algorithm. If this original-format alignment was plausible, then the paired-end read alignment was removed from further analysis.

To estimate an upper bound for false alignments generated by our split-read approach, the alignment strategy described above was applied to the reads that

aligned to the genome with one difference: the initial reads were split at GA dinucleotides instead of GT. Of the ~78 million genome-aligning reads assessed this way, only 15,727 reads could be split and aligned to the genome, reflecting a total of 1267 intron branch intervals, yielding a false alignment rate of only ~0.02%. Importantly, this rate may well be an over estimate because of the propensity of some true split reads to incorrectly align to the genome with mismatches.

2.3.8 Branch read identification

Strandedness of a split read was determined by selecting the highest scoring 5′SS and BP scores for each direction of the read alignment. The branch point was identified by looking within +/- 2 nucleotides of the end of the read alignment for the best scoring BP. This allows for small deletions and insertions (a common property of reverse transcription across branchpoints). Split reads that make use of the same 5′SS and BP were aggregated together and considered to come from the same lariat. The associated 3′SS was determined to be the first AG dinucleotide found at least 5 nucleotides downstream of the BP.

2.3.9 Aggregation of introns

Overlapping split reads were aggregated by strand and by overlapping genomic coordinates. Annotated 5′SS and 3′SS were determined by the POMBASE Schizosaccharomyces_pombe.ASM294v2.21.gff3 file. Alternate sites were defined as any site that did not correspond to either the 5′SS or 3′SS of an an-

notated intron. Split reads that overlap annotated introns with both splice sites corresponding to a non-annotated location were not further considered. Alternate 5′SS within 5 nt of an annotated 5′SS are prone to mismapping and as an aggregate were ignored as a parsimonious approach to identifying unannotated splicing events. Novel introns were defined as split reads that do not overlap a known intron on the same strand. Exon skipping events were defined as split reads that overlap two introns in the same transcript. Annotated, alternate, and novel introns were required to have at least one read with a minimum of 30 nucleotides of total aligned sequence.

2.3.10 Alternate intron identification

Branches overlapping a single intron were aggregated. For each branch the first AG dinucleotide at least five nucleotides downstream of the intron was called as the 3′SS. For the RNAseq alternate introns, utilization was measured as alternate counts divided by the sum of alternate and annotated counts. When indicated, likelihood of alternate introns were computed with a binomial distribution created from the number of total reads alternate and annotated reads and an error rate being tested in the range 10^{-1} to 10^{-6}.

2.3.11 RNAseq

RNAseq[1] was done using TruSeq RNA Kit v2 or NEBNext Ultra Directional Kit. RNAseq alignments were done using Tophat suite version 2.0.9 with Schizosac-

[1]RNAseq library construction was either completed by Elizabeth A. Fogarty or the Cornell University RNA sequencing core

charomyces_pombe.ASM294v2.21.gff3 and with novel intron discovery. For novel introns, only GT-AG novel exon-exon junctions were further processed. Novel introns found to overlap a single annotated intron were aggregated. Reads for novel intron RNAseq junctions were required to have mapq scores of at least 30. Expression quartiles were computed for transcripts using RNAseq for the indicated RNAseq library by using the exon-exon junction counts from Tophat.

2.3.12 Comparative analyses

MAF format MultiZ alignments[52] from 01/04/2012 were acquired from the Broad Institute website and used to retrieve aligned sequences for computing log-odds scores. Log-odds scores for 5'SSs and BPs were computed, when possible, for each of the Schizosaccharomyces species considered. To calculate a background rate of sequence conservation, putative upstream 5'SS and downstream BP sequences were identified from the coding sequences of intronless genes in *S. pombe*. From these sequences, a large number of putative splice sites were initially chosen, the total number representing a 10-fold increase over the number of identified alternative sites. For each of these sites, the PWM score was determined for the *S. pombe* sequence, after which time a subset of these sites was selected that had a similar score distribution to the identified alternative sites. For this subset of sites, the PWM was then determined for the orthologous sequence in each of the Schizosaccharomyces species, and this score was assessed for conservation. A similar approach was used to determine background rates for the downstream 5'SS and upstream BPs, but using sequences found in *S. pombe* 3' UTRs that are shorter than 150nt as the source for the back-

ground distribution.

2.3.13 qPCR of lariat introns

Intron qPCR measurements[2] were made for two different introns from each of four different multi-intronic genes using RNA isolated from a Δdbr1 strain. Standard dilution curves using genomic DNA were generated for each primer pair, allowing for comparison of the relative levels of each RNA.

2.3.14 Weblogos

Web logos were generated using a command line version 3.4[68].

2.3.15 Accession codes

All sequencing data have been submitted at NCBI's Gene Expression Omnibus (GEO) repository with accession number GSE68345.

2.4 RESULTS

Global *S. pombe* splicing profiles revealed by intron lariat-sequencing To capture the global splicing profile of *S. pombe*, we used two-dimension gels (Figure 2.1A) to isolate intron lariats from a Δdbr1 strain grown under several growth

[2]qPCR experiments were completed by Madhura Raghavan

conditions (see Materials and Methods). Building upon our previous work[34], experimental conditions were optimized to recover both long and short introns. Purified lariat RNAs were converted, without debranching, into cDNA and sequenced on an Illumina HiSeq 2000, generating over 231 million sequencing reads. Alignment of these reads to the *S. pombe* genome revealed that 60% of genome-matching reads mapped to annotated introns (Figure 2.1B) and Supplementary Table S1[3]) with only a minority of reads mapping to exons, confirming the high level of lariat enrichment afforded by this approach.

As we and others have previously noted[34, 33, 36, 45], sequencing of lariat introns generates two distinct types of reads: one derived from the body of the intron, and the other derived from reverse transcription across the lariat branch (Figure 2.1C). Branch-spanning reads contain both the 5'SS and BP sequence within the lariat, and are thus diagnostic of a splicing event, somewhat analogous to exon-exon spanning reads within RNAseq data. While these reads are information rich, their identification is non-trivial because of both their inverted nature and the poor efficiency and reduced fidelity of reverse transcription across this junction[33]. To systematically identify these reads in our dataset, we developed an alignment pipeline whereby all reads that failed to directly map to the genome were divided at each GU dinucleotide, corresponding to possible 5'SSs, and the pairs of divided reads were re-assessed for alignment to the genome using a split-read mapping strategy (Figure 2.1D). A total of 3.7 million such reads were identified by this approach with a low false discovery rate (see Materials and Methods, and Supplementary Table S2[4]), making this by far the largest dataset of experimentally identified branch sites to date.

[3]see[35] for data
[4]see[35] for data

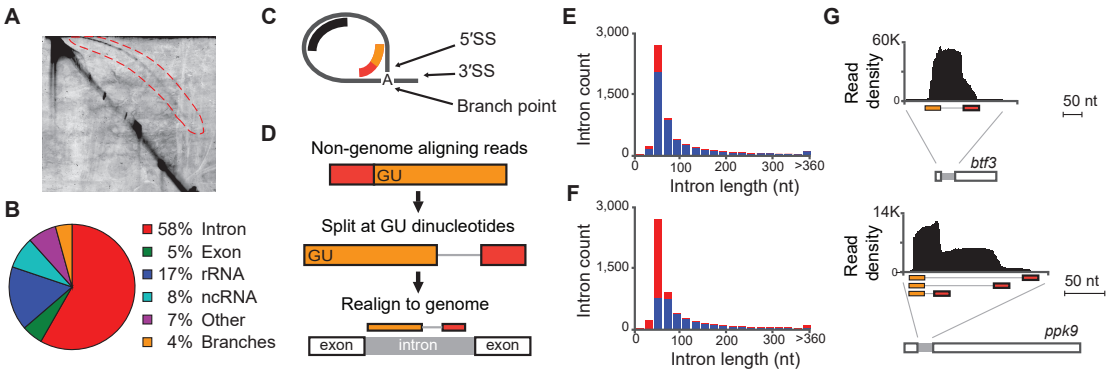

Figure 2.1: Intron lariat sequencing defines splicing patterns.

(A) Image of two-dimensional gel electrophoresis of RNA isolated from Δdbr1 *S. pombe*. Intron lariats (red-bounded region) were isolated and used as source material for sequencing. (B) Pie-chart summarizing allocation of lariat sequencing reads to indicated genomic regions. (C) Illustration of intron lariat and splice sites, depicting intron-mapping reads (in black) and branch-spanning reads (in red-orange). (D) Schematic of alignment strategy for candidate branch-spanning reads, together with illustration of aligned branch-spanning read. (E) Histogram of annotated introns counts (y-axis) separated by length (x-axis, 20 nucleotide bins), indicating introns recovered with intron-mapping reads (blue) and those not recovered (red). (F) Histogram of annotated introns counts (y-axis) separated by length (x-axis, 20 nucleotide bins), indicating introns precisely recovered with branch-spanning reads (blue) and those not recovered (red). (G) Intron-mapping reads (y-axis indicates read density) aligning (x-axis indicates alignment position) to indicated *S. pombe* pre-mRNAs, together with branch-spanning reads (orange-red) aligned with split-read mapping strategy. The btf3 peak density truncated at +/- 50 nt of intron boundaries. The ppk9 peak tapers just upstream of intron boundaries.

Importantly, the data generated from both the body-mapping and branch-spanning reads successfully identified the majority of known *S. pombe* introns. Over 85% of annotated introns had reads mapped to the body of the intron (Figure 2.1E and Supplementary Table S3[5]), while ~55% had branch-spanning reads that recover both the annotated 5'SS and predicted BP (Figure 2.1F and Supple-

[5]see[35] for data

mentary Table S3[6]). As we and others have previously seen, short intron lariats were particularly difficult to recover in these experiments[34, 36]. Nevertheless, because branch-reads contain the coupled information of both the 5'SS and BP sequence used to form the lariat, alternative splicing events that would be difficult to reliably predict from body-mapping reads can be definitively assigned by branch-reads. For example, whereas the peak of body-mapping reads for the btf3 transcript (Figure 2.1G, top) suggested a discreet 5'SS and BP, the spectrum for the ppk9 transcript (Figure 2.1G, bottom) suggested possible alternative splice sites. The use of branch-spanning reads readily resolved these different patterns by identifying a single 5'SS/BP combination for btf3, but three distinct combinations for ppk9. Because of the precision with which branch-reads define splicing events, we relied exclusively upon them for further analysis.

2.4.1 Lariat sequencing identifies widespread examples of alternative splicing

To characterize identified splice sites, a position weight matrix (PWM) scoring metric[59] was implemented (see Materials and Methods), based upon the ~5,000 splice sites annotated on PomBase[69]. Previous computational analyses had predicted the likely BP for >90% of annotated *S. pombe* introns; these sites were used as the basis of our PWM scoring for BPs. For the small number of introns for which multiple BPs were predicted, we considered the BP closest to the annotated 3'SS to be the primary BP, and used it for our analysis. As expected, our analysis of the set of annotated introns showed a wide range of scores for both the 5'SS and BP sequences, reflecting the degeneracy of splice

[6] see[35] for data

site sequences in *S. pombe* (Figure 2.2A)[62]. Interestingly, while the scores of the 5'SSs recovered by lariat sequencing closely match the distribution of all annotated scores, the recovered BPs included many more low scoring sequences. Remarkably, nearly 900 of the annotated introns recovered here revealed activation of multiple BPs, each of which was predicted to use the annotated 3'SS. Importantly, while this estimate represents the lower limit of the frequency of alternate BP activation because of the size limitations described previously, it is nevertheless significantly higher than a previous study indicated[36]. When considering the scores of the primary BPs we identified, defined as the closest BP upstream of the annotated 3'SS, there was little difference between those annotated introns for which only a single BP was identified and those with multiple branches (Figure 2.2B), suggesting that alternative BP selection is not driven simply by the strength of the primary BP. Not surprisingly, however, quality scores of alternative BPs tended to be weaker than those of primary branch sites (Figure 2.2B). Nevertheless, many annotated introns contain alternative BP sequences with scores comparable to those of the presumed primary branch sites; representative examples are shown for the spf47 intron 1 (Figure 2.2B) for which the identified alternative and presumed primary branch points both have scores near the middle of their respective distributions.

Whereas the alternative branch points noted above are not predicted to change intron-exon boundaries, an additional 2,923 alternative splicing events (associated with 1851 annotated introns) were identified that utilize one annotated site and one alternate site and are predicted to change the coding character of the resulting mRNA. Remarkably, these alternative splicing events implicate at least half of all annotated introns as subject to alternative splicing. Included among these were 1031 events (corresponding to 858 annotated introns) where

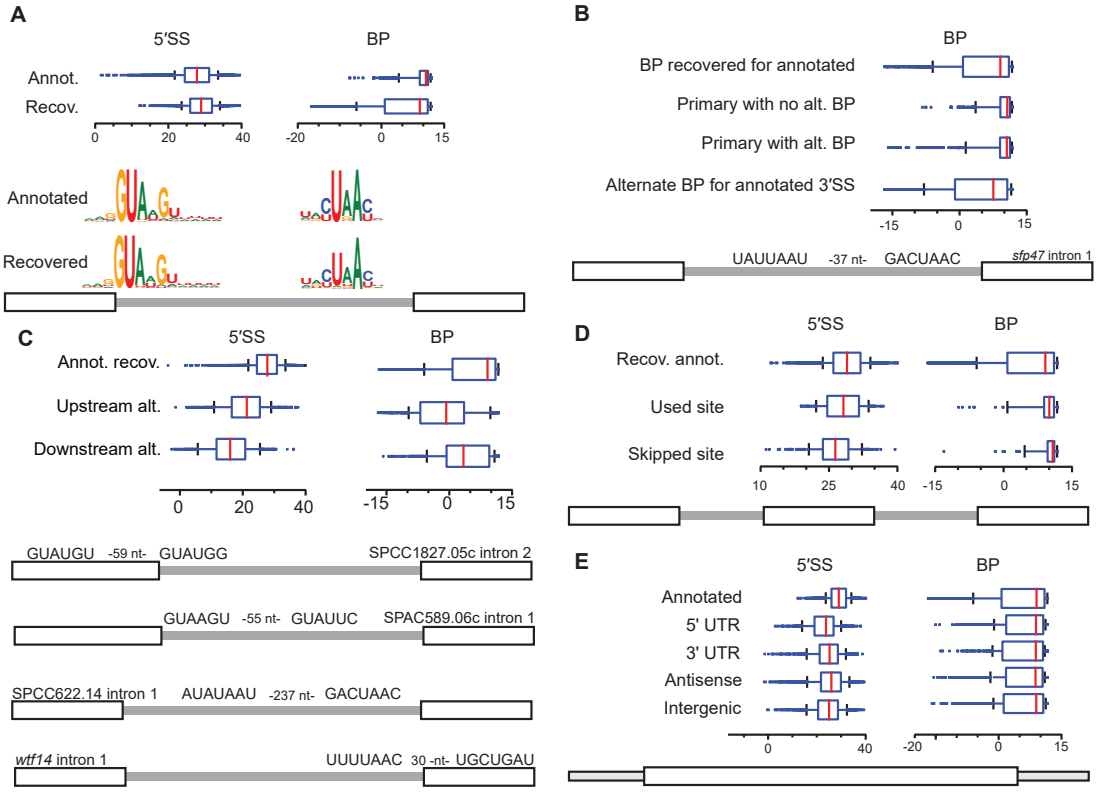

Figure 2.2: Global analysis of alternative and novel splice sites in *S. pombe*.

(A-E) Distribution of splice-site strengths as boxplots (x-axis indicates PWM scores), together with examples of alternative site sequences. (A) Splice site scores (5′SS and BP) corresponding to annotated introns (annot.), and sites corresponding to annotated introns recovered with branch-spanning reads (recov.) (B) BP scores for indicated categories of alternative splicing events associated with alternative BPs that are paired with annotated 3′SSs. (C) Splice site scores (5′SS and BP) corresponding to alternative splice site scores partitioned into upstream and downstream alternative intron boundary sites, compared to annotated and recovered sites (annot. recov.). (D) Splice site scores (5′SS and BP) corresponding to splice site scores associated with exon-skipping events partitioned into sites participating in exon-skipping (used site) and those skipped (skipped site), compared to annotated and recovered sites (annot. recov.). (E) Splice site scores (5′SS and BP) corresponding to sites found in novel introns in indicated genomic regions, compared to annotated and recovered sites (annot. recov.).

an alternative 5′SS is spliced to the canonical BP, and 1892 events (corresponding to 1276 annotated introns) where the canonical 5′SS is spliced to an alternative BP/3′SS; akin to mammalian alternative splicing. Interestingly, for both 5′SSs and BPs, alternative sites upstream of the canonical site were identified at nearly twice the frequency as they were downstream (Supplementary Table S4[7]), consistent with a first come, first served model of splice site selection[70].

Many of the alternative splice sites described above had sequence scores similar to those of annotated introns, however, we observed a clear relationship between the quality of the splice sites and their position relative to the annotated sites. For 5′SSs, alternative sites identified upstream of the annotated site had a distribution of scores that, while weaker, substantially overlapped those of the annotated sites (Figure 2.2C). In contrast, the distribution of scores corresponding to alternative downstream 5′SSs were significantly weaker than those of both the annotated and upstream 5′SS (Figure 2.2C). The pattern for alternate BPs was inverted: activated downstream sites had scores more similar to canonical BPs while those found upstream tended to be lower in strength (Figure 2.2C). Representative examples of these types of alternative splicing are shown (Figure 2.2C) where both the alternative and canonical events have scores near the middle of their respective distributions.

Having identified alternative 5′SSs upstream of nearly 15% of annotated introns, we wondered whether the failure to identify alternative sites for the remaining 85% of introns reflected the absence of an effective alternative splice site or the failure to utilize such sites. To address this question, we examined a 150 nucleotide window upstream of every annotated intron and identified the highest scoring potential 5′SS. Many annotated introns are flanked by upstream

[7] see[35] for data

sequences that contain high scoring candidate-alternative 5′SSs but for which we detected no alternative splicing (Supplementary Figure 2.3). Perhaps deeper sequencing might reveal usage of these potential sites; alternatively, additional sequence elements or secondary structures may be functioning to preclude alternative splicing at these locations[71, 72].

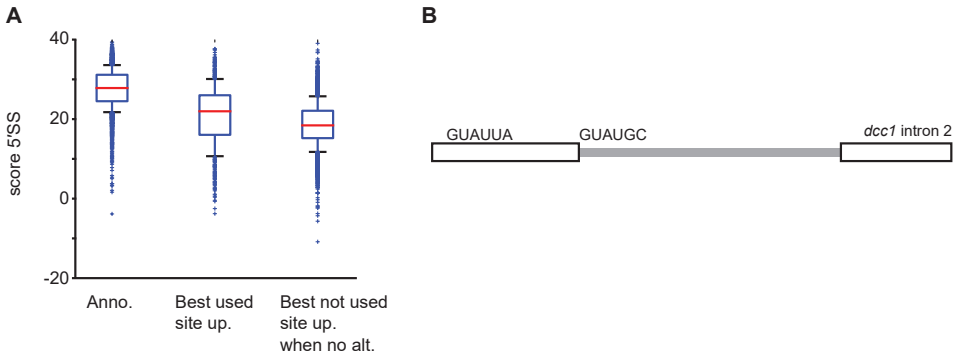

Figure 2.3: Unused upstream GT sites have similar splicing potential as sites that are used.

(a) The scores of annotated splice sites vs the best scoring 5′SS upstream within 150 nt of the annotated 5′SS for introns without a found upstream alternate 5′SS. (b) Gene model example showing an alternate upstream 5′SS and the annotated 5′SS.

Previous studies identified exon skipping in *S. pombe*; although relatively few such examples were discovered in earlier experiments[34, 36], more recent work identified over 100 high-confidence events[53]. Here, we found hundreds of additional exon skipping events, expanding the repertoire of confirmed exon skipping events in *S. pombe* (Supplementary Table S5[8]). Analyses of the splice site sequences associated with the events identified here (Figure 2.2D) revealed that the 5′SS of the upstream intron and the BP sequence of the downstream

[8]see[35] for data

intron had sequences that were nearly indistinguishable from the composite scores of annotated introns. Remarkably, however, the skipped BPs of the upstream introns were not characterized by low information sequences, but rather appeared to have slightly stronger scores in aggregate than annotated introns, inconsistent with expectations of intron-definition based models of exon skipping. By contrast, the skipped 5'SS of the downstream introns had significantly weaker splice site scores than annotated introns. This increased propensity of exon skipping events to be associated with weak downstream 5'SSs, together with the absence of weak BP sequences, implies that spliceosome assembly via exon definition may be a more prominent aspect of splicing in *S. pombe* than previously appreciated[73].

In addition to alternative splicing associated with annotated introns, our data also revealed an unprecedented level of splicing across the transcriptome at sites with no characterized introns. A total of 8113 splicing events were identified associated with 7412 novel introns (Supplementary Table S6[9]). These introns were located within the transcripts of protein-coding genes (including 857 within annotated 5' UTRs, 971 within annotated 3' UTRs, and 1567 within the coding regions of these transcripts), within anti-sense RNAs (2699), within non-coding RNAs (554), and within intergenic regions of the genome (1343). Remarkably, while the overall distribution of 5'SSs scores for these novel introns was noticeably lower than those associated with annotated introns, the majority of these novel events had 5'SSs sequences with strong PWM scores (Figure 2.2E). Similarly, the distribution of BP scores for novel introns was virtually indistinguishable from those found in annotated introns (Figure 2.2E). Importantly, although many novel events are recovered with low read counts, the splice-site

[9] see[35] for data

score distributions are similar across high and low read counts (Supplementary Figure 2.4).

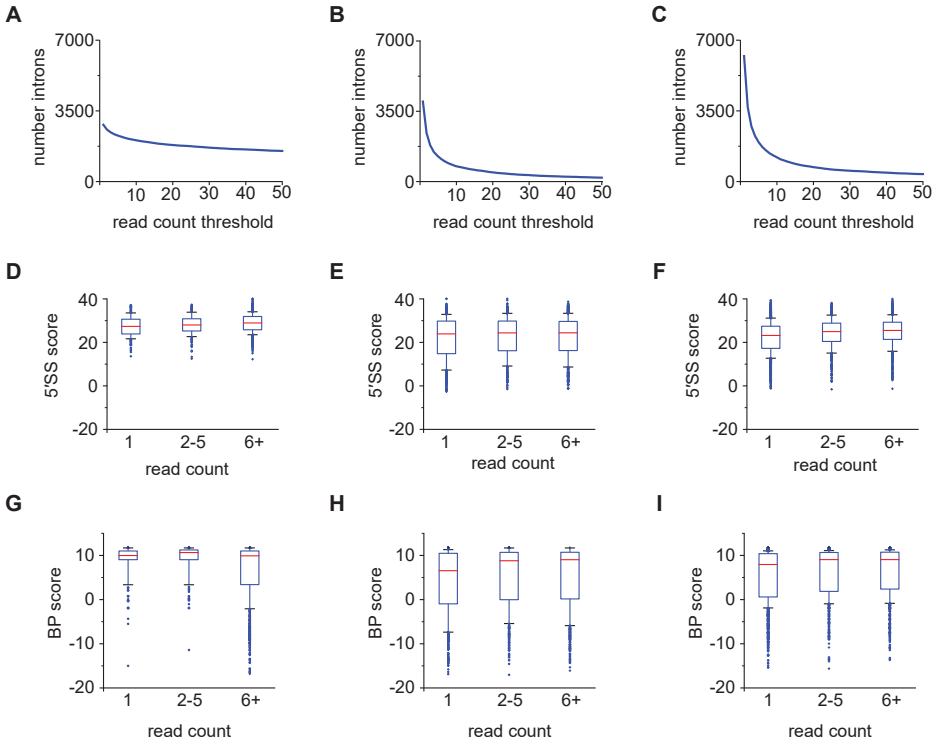

Figure 2.4: Relationship between read depth and introns recovered showing similarities across low and high read count events.

Number of introns recovered at a read threshold for annotated, alternate, and novel introns, respectively (a-c). Boxplots of 5'SS splice site scores (d-f) and BP scores (d-f) for introns with the given read count ranges for annotated, alternate, and novel introns, respectively. Lower read count events are likely a composite of poorer expressed introns and introns not in our datas size sensitivity range. The marked difference between BP scores for annotated introns is the result of the method picking up low frequency events with higher read counts in the size sensitivity range.

2.4.2 RNAseq validates widespread alternative splicing

Having identified widespread examples of alternative splicing via lariat sequencing, we turned to RNA sequencing as an orthogonal approach for validation. Datasets of poly(A)+ RNA were generated for both wild-type and Δdbr1 strains (Supplementary Table S7[10]). Importantly, although loss of Dbr1 perturbed transcript levels of a small subset of genes, the overall transcriptomes of wild-type and Δdbr1 strains as determined by RNAseq were extremely similar (Pearson correlation coefficient of 0.993, $P < 2.2 \times 10^{-16}$; Supplementary Figure 2.5). We used TopHat2 to identify transcripts harboring alternative 5'SSs[74], and then examined the extent to which these alternative 5'SSs overlapped with those identified by lariat sequencing. Approximately 40% of alternative 5'SSs identified by lariat sequencing were detectable by RNAseq, and 25% of alternative 5'SSs identified by RNAseq were detected by lariat sequencing (Figure 2.7A and Supplementary Table S4[11]), demonstrating that the two approaches provide complementary but not identical descriptions of the transcriptome. Importantly, the average quality of alternative 5'SSs, as judged using PWM scoring, was nearly identical when comparing alternative sites defined uniquely by lariat sequencing or RNAseq, whereas those sites identified by both methods generally corresponded to slightly stronger sites (Figure 2.7B). Because lariat sequencing directly identifies BPs but not 3'SS, and RNAseq identifies the inverse, to enable comparison of these datasets the first AG dinucleotide downstream of the BP identified by lariat sequencing was assumed to be the 3'SS. Using this approach, similar overlaps were also observed in the alternative BPs/3'SSs identified by lariat sequencing and RNAseq (Supplementary Figure 2.6).

[10] see[35] for data
[11] see[35] for data

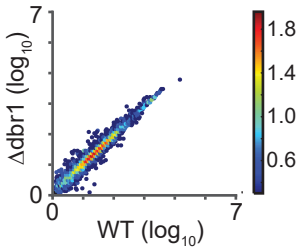

Figure 2.5: Scatter plot showing similarity between RNAseq transcript expression between Δdbr1 and WT *S. pombe*.

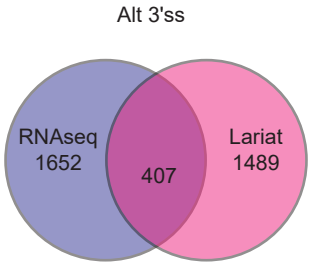

Figure 2.6: Venn diagram of alternate 3'SS found by RNAseq and lariat sequencing.

Multiple sources, both biological and technical, almost certainly contributed to the imperfect overlap between RNAseq and lariat sequencing-based definitions of the transcriptome. A major biological difference derives from the species sequenced: lariats versus mature transcript. This difference is likely to be particularly important for alternative products whose structures result in accelerated decay, leaving them poorly detected by RNAseq. Alternatively, because of the length bias of lariat sequencing, RNAseq is better positioned to capture very short or very long introns, as confirmed in Figure 2.7C. While investigating this length bias, however, we found an additional, and somewhat surprising, result: the levels of lariats derived from different introns of a com-

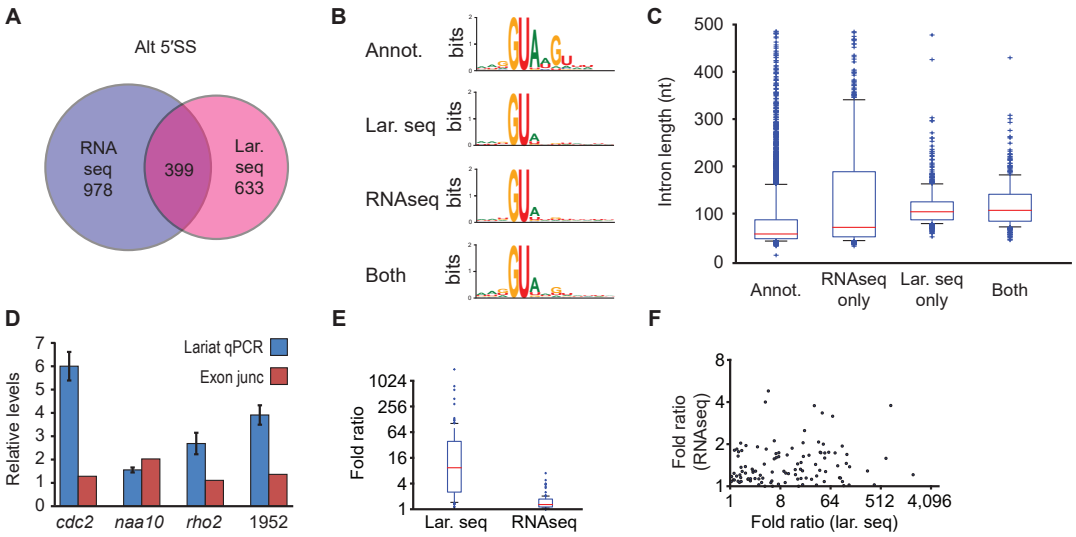

Figure 2.7: Cross validation and comparisons of alternative splicing detected using RNAseq and lariat sequencing.

(A) Venn diagram illustrating alternative 5'SSs identified by RNAseq, lariat sequencing, or both. (B) Web-logo comparisons of annotated 5'SSs compared to those identified by RNAseq, lariat sequencing, or both. (C) Intron length comparisons of annotated introns compared to those identified by RNAseq, lariat sequencing, or both. (D) qRT-PCR measurements of relative lariat levels compared for pairs of lariats from two-intron genes (blue), compared to RNAseq determinations of exon-exon junction reads for the corresponding splice junctions (SPAC1952.04c labeled as 1952). (E) Boxplots indicating distributions of fold ratios (y-axis) of branch-spanning read counts for pairs of introns from multi-intron genes, compared to ratios of exon-exon spanning read counts for splice junctions from multi-intronic genes with comparably sized intron lengths. (F) Scatter-plot of values shown in (e), relating RNAseq-derived ratios (y-axis) to lariat sequencing-derived ratios (x-axis) for genes whose introns are of comparable size.

mon pre-mRNA were detected to markedly different degrees in our data, and this difference was maintained even when looking at intron pairs that were of the same general size in genes with multiple introns. Importantly, we confirmed this result using qRT-PCR on four transcripts (Figure 2.7D) by comparing two introns within each transcript. For each mRNA tested, RNAseq indi-

cated that signal derived from the different exon-exon boundaries were narrowly distributed, as expected. In contrast, qRT-PCR measurements indicated that lariat introns derived from the same pre-mRNA were present at greatly different levels (Figure 2.7D). This result was confirmed genome-wide, using exon-exon spanning reads found in RNAseq and branch-spanning reads from lariat sequencing (Figure 2.7E and F). The biological basis for this result is unclear, but likely indicates variable decay rates for lariats in the absence of Dbr1. Importantly, while this result complicates quantitative comparisons for individual species detected using branch-spanning reads, genome-wide comparisons are less likely to be compromised.

In addition to evidence of extensive alternative splicing in *S. pombe*, our lariat sequencing data also revealed many thousands of novel introns (Figure 2.2E). As before, we wished to validate and compare novel introns found with lariat sequencing to those present in RNAseq. Similar to our previous findings, approximately 15% of novel introns found by lariat sequencing are also found within RNAseq, whereas ~20% of novel introns detected by RNAseq are also found within lariat sequencing (Figure 2.8A and Supplementary Table S8[12]). Importantly, the strengths of 5'SSs and BPs detected by both approaches were highly similar (Figure 2.8B and 2.8C, respectively), and the lengths of the novel introns recovered by the two approaches are consistent with the previously discussed length biases (Figure 2.8D). Taken together, these results imply the existence of many thousands of additional introns in *S. pombe* not found by either result.

During final preparation of our work, an analysis of publicly available RNAseq datasets from a variety of *S. pombe* growth conditions and mutants

[12]see[35] for data

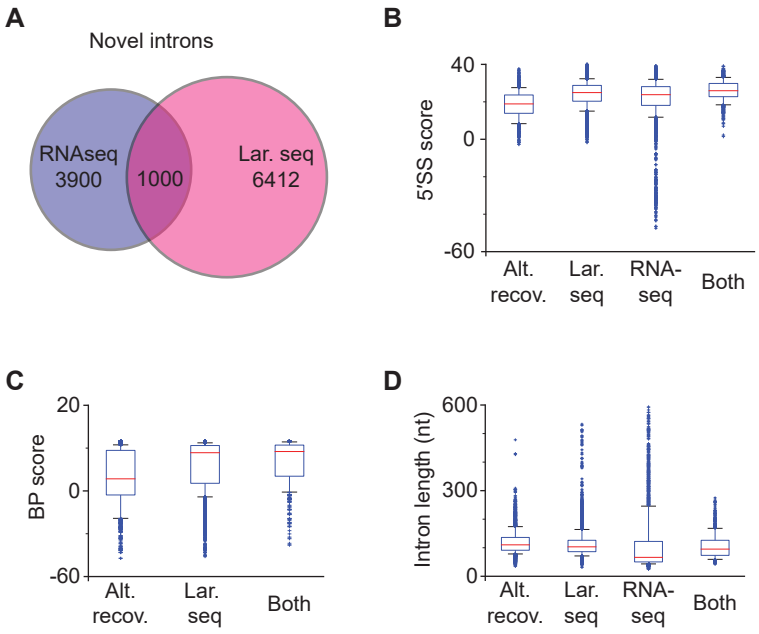

Figure 2.8: Cross validation and comparisons of novel introns detected using RNAseq and lariat sequencing.

(A) Venn diagram illustrating novel introns identified by RNAseq, lariat sequencing, or both. (B) Boxplot distributions of 5′SS strength (y-axis) for 5′SSs corresponding to alternative sites recovered using lariat sequencing (alt. recov.) for annotated introns, or for novel introns identified using: lariat sequencing (lar. seq); RNAseq; or both. (C) Boxplot distributions of BP sequence strength (y-axis) for categories indicated in (B). (D) Boxplot distributions of intron length (y-axis) for categories indicated in (B).

were examined for unannotated splicing events[53]. As with the analysis of our own RNAseq data, we sought to compare the novel splicing events identified by lariat sequencing with those identified in this new study. Remarkably, even though nearly 4 billion reads of RNAseq data were analyzed, comparison of these published data with our lariat sequencing generated a similar result: significant overlap of the novel splicing events was identified between the datasets, with each approach further identifying unique subsets of events. For example, of the 2923 alternative 5′SSs and 3′SSs associated with known introns identi-

fied that were identified by lariat sequencing, only ~15% were identified in the published RNAseq data (Supplementary Table S5[13]). By contrast, 2472 events were uniquely identified by lariat sequencing and 1207 events were uniquely identified in the published RNAseq data. Importantly, as before, the majority of the events that went undetected by lariat sequencing were expected to generate lariats of sizes not readily recovered in our experiment. Similar patterns were observed when comparing exon skipping events, and novel introns (Supplementary Tables S4 and S8[14]). Together, these results reinforce the idea that there are many additional locations within the *S. pombe* genome that are acted upon by the spliceosome but have not yet been identified by either method.

2.4.3 Estimating the frequency of alternative splicing

Having identified thousands of locations of alternative splicing, we next sought to characterize the frequency of these events. As a simple gauge of the extent of alternative splicing, we determined the total number of alternative splice site reads for every annotated intron relative to the sum of all annotated and alternate reads (from Supplementary Table S4[15]). Remarkably, this yielded an alternative rate of 2.8% (Figure 2.9A), far higher than comparable values estimated from exon-exon spanning reads in most RNAseq-based studies (Figure 2.9A and references[52, 53]), and moderately higher than values calculated in the background of nuclear decay mutants[53]. Importantly, when the alternative splicing percentages were recalculated considering only those alternative splicing events for which the canonical lariat was within the optimal size range

[13] see[35] for data
[14] see[35] for data
[15] see[35] for data

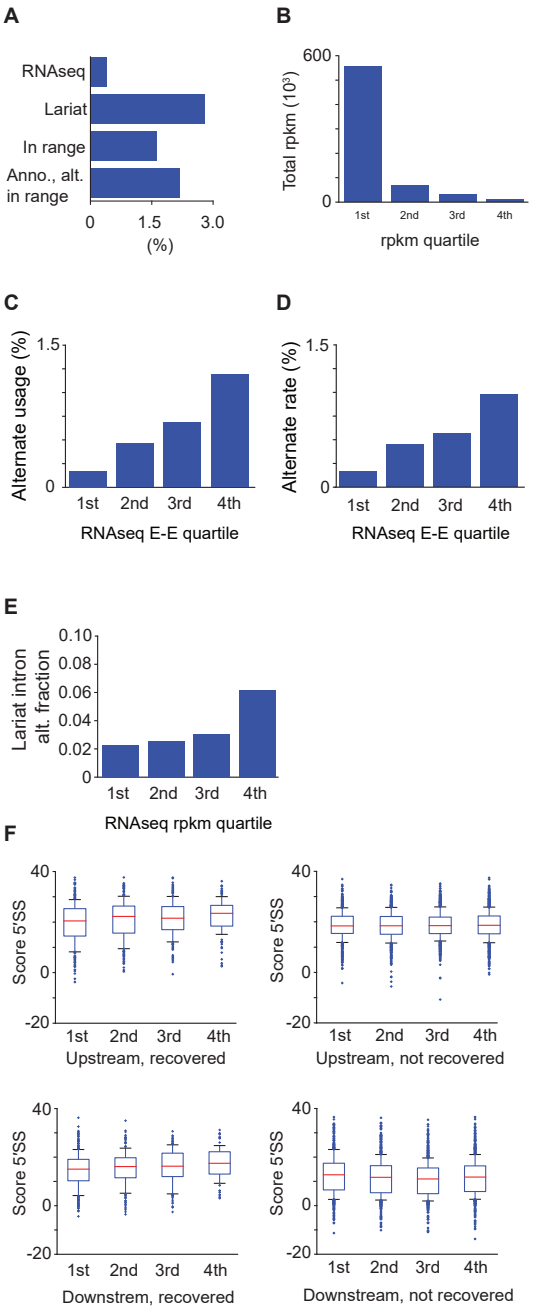

Figure 2.9: Extent of alternative splicing in *S. pombe*.

(A) Percentage of reads corresponding to alternative splice products detected by RNAseq or lariat sequencing (lar. seq), and in lariat sequencing restricting the analysis to introns for which the annotated intron has a size optimal for detection by lariat sequencing (lar. seq opt. size; 70-150 nucleotides). (B) total RNAseq RPKM values (y axis) for intron containing transcripts separated into expression quartiles (x axis). (C) Percentage of reads corresponding to alternative splice products detected by RNAseq (y axis), shown by expression quartiles (x axis). (D) Inferred rate of alternative splicing, modeled using RNAseq data, shown by expression quartiles (x axis). (E) Percentage of reads corresponding to alternative splice products detected by lariat sequencing, shown by expression quartiles (x axis). (F) Boxplot distributions of 5'SS strength (y-axis) for upstream and downstream alternative 5'SSs for each expression quartile (x-axis; recovered), and for best-scoring upstream and downstream candidate sites whose usage was not observed (not recovered).

for lariat sequencing, and separately only those for which both the alternative and canonical lariats were within the optimal range, estimated alternative rates were determined to be 1.6% and 2.2%, respectively, still well in excess of estimates derived from our RNAseq data (Figure 2.9A).

Although the overall rate of alternative splicing detected in our RNAseq data was significantly lower than what was measured by lariat sequencing, we noted in our data that a broad range of error rates were measured among the different transcripts. In particular, alternative splicing rates were relatively low for highly expressed genes, but more pronounced for those with low expression. Therefore, to account for any relationship between gene expression and fidelity of splicing, intron containing genes were separated into expression quartiles as determined by host-transcript RPKM values (Figure 2.9B). The proportion of exon-exon spanning reads in the RNAseq data that corresponded to alternative products was then separately recalculated for each of the four expression quartiles (Figure 2.9C). In addition, a maximum-likelihood based approach was used to estimate an alternative splicing rate within each quartile of genes (Figure 2.9D). Importantly, this strategy excluded from our analysis all genes for which we detected high proportions of alternative splice products, as determined by a likelihood scoring approach, reasoning that such events are more likely to correspond to bona fide alternative splicing rather than errors in splicing. Together, both of these approaches showed that highly expressed genes were spliced with extremely high fidelity, whereas the fidelity of splicing decreases as expression quartiles decrease, a result that is robust to different likelihood threshold calculations (Supplementary Figure 2.10). Importantly, the association of decreased fidelity of splicing with more lowly expressed genes is also observed when calculated using the percentage of alternative branch-spanning reads detected in lariat sequencing (Figure 2.9E), although the extent of alternative splicing detected using lariat sequencing far exceeded that detected with RNAseq, a result consistent with lariat sequencing possessing enhanced sensitivity to detect alternative isoforms subject to rapid decay. It is worth noting that no striking dif-

ferences are apparent between the quartiles when comparing the scores of upstream or downstream alternative 5'SSs (Figure 2.9F). Moreover, the top-scoring potential alternative 5'SSs within introns for which we observed no alternative splicing were comparable to the alternative sites used (Figure 2.9F).

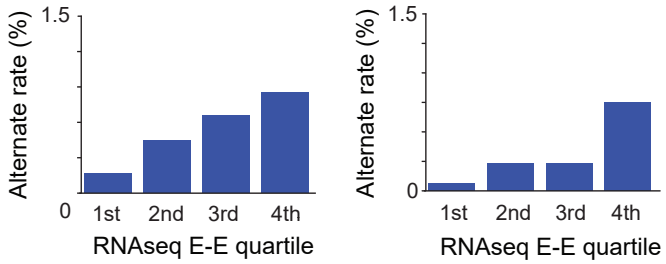

Figure 2.10: Alternate exon-exon junction rate for RNAseq expression quartiles.

Alternate exon-exon junction rate for RNAseq expression quartiles calculated by log-likelihood (left) and log likelihood with unlikely events excluded using an alternate criteria (right)

Given the large number of novel introns identified by both lariat sequencing and RNAseq, we determined their percent spliced index (PSI)[75], based upon our RNAseq data, in an effort to characterize whether these events reflected: bona fide introns whose annotations are incomplete, regulated splicing events with low PSI under standard growth conditions, or low-frequency events likely to represent splicing noise. For each novel intron, the RNAseq data were examined to identify reads spanning exon-exon or exon-intron boundaries, reflecting the spliced and unspliced isoforms, respectively. Interestingly, 93 of these introns showed a PSI of over 80%, consistent with the behavior of bona fide, canonical introns. These introns were distributed between coding and non-coding, sense and anti-sense transcripts, and argue for modifications of their

genome annotations (Supplementary Table S8[16]). Similarly, an additional 523 introns showed a PSI between 20% and 80%. While these PSI values were lower than expected for a canonical intron, they suggest the possibility that these are conditionally regulated splicing events. For the vast majority of the novel introns identified, PSI was below 20%. Although it is difficult to discern the functional significance of any given isoform simply on the basis of its PSI, we chose to refer to these low frequency events as aberrant.

2.4.4 The majority of alternative splicing events in *S. pombe* are not conserved in closely related species

To gain additional perspective on the potential functional relevance of alternative splicing in *S. pombe*, we investigated the extent to which alternative splice sites are significantly evolutionarily conserved. The PWM scores calculated for splice sites in *S. pombe* were compared with scores for the orthologous positions in three related Schizosaccharomyces species: *S. octosporus*, *S. cryophilus* and *S. japonicus* (Figure 2.12A)[52]. As expected, annotated 5'SSs and BPs in *S. pombe* overwhelmingly maintain their splice site identity in related species (Figure 2.12B and 2.12C, respectively). In contrast, a comparison of the alternative sites identified by lariat sequencing in *S. pombe* showed that a large fraction (68-89%, depending on the species compared) of 5'SSs in the related species have no potential to function as splice sites (Figure 2.12D and Supplementary Figure 2.11). There are, however, many sites whose sequences in related species closely match consensus 5'SSs used as alternative splice sites in *S. pombe* (Figure 2.12E and 2.12F, for upstream and downstream alternative 5'SSs, respectively).

[16] see[35] for data

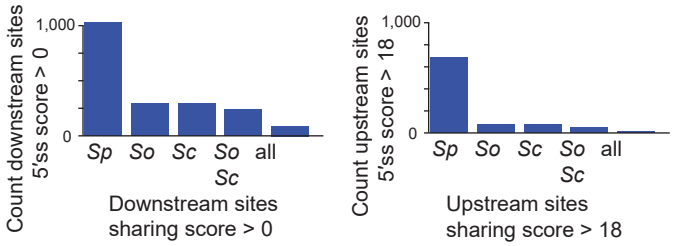

Figure 2.11: Alternate 5'SS sites in *S. pombe* and the number of sites showing similar conservation

Alternate 5'SS sites in *S. pombe* and the number of sites also in *S. octosporus*, *S. cryophilus*, and *S. japonicas* as indicated for downstream events (left) and upstream events (right)

To determine the background level of conservation that exists independent of possible functions as 5' splice sites, a similar analysis was performed on theoretical 5'SSs we found within coding sequences from genes for which no evidence of splicing exists, and separately, from noncoding sequence within 3' untranslated regions (UTR). We reasoned that the level of conservation we detected from such theoretical 5'SSs would be a suitable background estimate for the extent of conservation we detected for real sites, and thus enable us to estimate the number of sequences selectively maintained to function as splice sites. This approach suggested that a small minority of alternate 5'SSs, both upstream and downstream, were selectively maintained above our background estimate (Figure 2.12G and H; see Supplementary Table S9[17] for quantification), with up to ~10% of alternate sites potentially the result of conservation, presumably corresponding to biologically meaningful occurrences of alternative splicing. In contrast, potential conservation of orthologous alternate BPs more strongly resembled the background distribution (Figure 2.12 I, J and Supplementary Figure

[17] see[35] for data

2.13). The apparent lack of conservation might be complicated by the dilution of conservation signal due to the propensity for a given 3'SS to utilize one of several possible BPs. Regardless, the lack of strong orthologous splicing signals suggested that most alternative BP usage results from aberrant splicing.

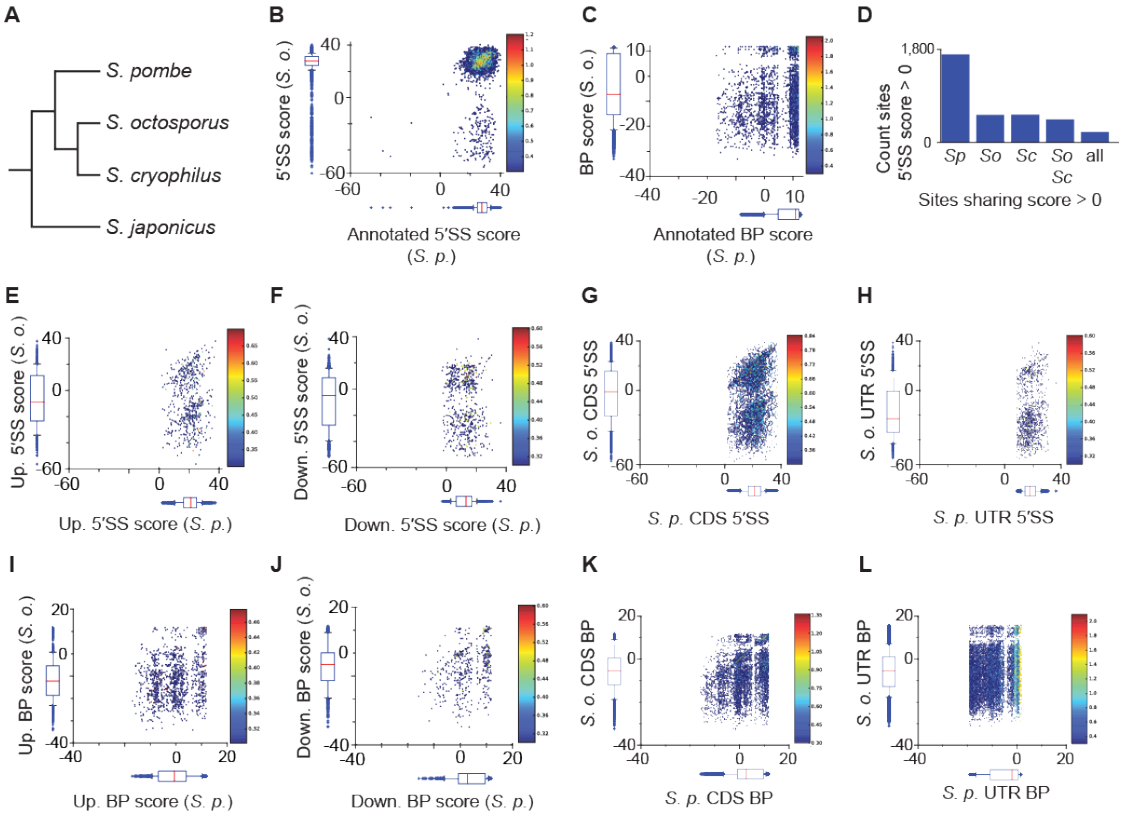

Figure 2.12: Comparative analyses of *S. pombe* splice sites.

(A) Cladogram illustrating *Schizosaccharomyces* species included in subsequent analyses. (B) Heat map and boxplots indicating relationship and distributions of annotated 5′SSs in *S. pombe* (x-axis) and *S. octosporus* (y-axis). (C) Recovered annotated intron BP sequences, plotted as in (B). (D) Counts of alternative 5′SSs in *S. pombe* exceeding a score of zero, together with counts for sites conserved in indicated species. (E-H) Comparison of 5′SS between *S. pombe* and *S. octosporus* for: alternate upstream 5′SSs (E), alternate downstream 5′SS (F), control upstream 5′SSs (G) and control downstream 5′SSs (H); plotted as in (B). (I-L) Comparison of BPs between *S. pombe* and *S. octosporus* for: alternate upstream BPs (I), alternate downstream BPs (J), control upstream BPs (K) and control downstream BPs (L); plotted as in (B).

		up. 5'ss	down. 5ss	down. bp	up bp
S. c.	background est.	42.03%	9.48%	31.16%	9.43%
	foreground	51.78%	18.42%	26.19%	8.52%
	excess foreground	9.75%	8.94%	-4.98%	-0.91%
S. o.	background est.	42.03%	9.48%	31.16%	9.43%
	foreground	54.37%	16.92%	28.43%	9.73%
	excess foreground	12.34%	7.44%	-2.73%	0.30%
S. j.	background est.	42.03%	9.48%	31.16%	9.43%
	foreground	38.51%	9.02%	16.21%	5.23%
	excess foreground	-3.52%	-0.46%	-14.95%	-4.20%

Figure 2.13: Estimated fraction of splice sites displaying conservation potential.

Table of fraction of sites in *S. pombe* recovered with orthologous splice site score above 0 in *S. octosporus*, *S. cryophilus*, and *S. japonicas* relative to a background estimate for either upstream or downstream sites for either 5'SS or BP.

2.4.5 Aberrant splicing in *S. pombe*

Our discovery of widespread alternative splicing in *S. pombe*, very little of which exhibited evidence of conservation, suggested to us that the majority of the observed alternative events represented aberrant splice site usage. At a high frequency, the alternative splicing events identified by lariat sequencing generated transcripts with expected reductions in overall stability. A total of 1661 of these alternate splicing events generated a frameshift of the resulting mRNA; an additional 376 events maintained coding frame but introduced premature stop codons in the mRNA. By contrast, only 559 of the alternative splicing events neither changed the reading frame nor introduced premature stop codons.

To better assess the stability of the aberrant splicing events we identified, additional RNAseq data were generated from a strain deficient for upf1, an essential component of the nonsense-mediated mRNA decay (NMD) pathway that selectively degrades erroneous transcripts (Supplementary Table S7[18]). To determine whether alternative isoforms were stabilized in this strain, we carefully examined a subset of transcripts that satisfied three criteria: the same alternative event was identified in both datasets, the expression level of the host transcript varied by less than 25% between datasets, and the total number of alternative reads in the Δupf1 dataset exceeded a threshold of 10 counts. As expected, when considering only those events that satisfied these criteria, the average alternate usage rate increased by over 50%, consistent with destabilization of these isoforms in wild type cells.

We considered it likely that the frequency at which an aberrant site was activated would be related to its strength as an alternative splice site. Somewhat

[18] see[35] for data

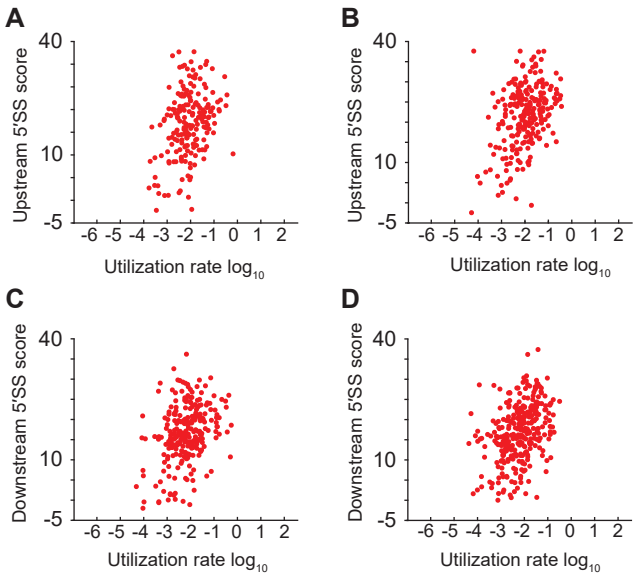

Figure 2.14: Influence of splice site strength on frequency of alternative splicing.

(A-D) Scatter-plots of proportion of alternative splicing events found in RNAseq data (x-axis) plotted against predicted strength of alternative 5'SS (y-axis). Alternative events analyzed separately for upstream (A, B) and downstream alterative 5'SSs (C, D), using RNAseq from wild-type (A, C) and NMD-deficient strains (B, D).

surprisingly, however, only a weak correlation was observed between upstream alternative 5'SS scores and their usage in the wild type RNAseq dataset (Pearson correlation coefficient of 0.12, p = 0.04; Figure 2.14A). A weak but more significant correlation was also seen for downstream alternative 5'SSs (ρ = 0.16; P = 0.01). By contrast, in the Δupf1 dataset the rate of utilization of both upstream and downstream alternative 5'SSs were more significantly, albeit still weakly, associated with the strength of alternative sites (ρ = 0.22, P = 5 x 10^{-4}, and ρ = 0.23, P = 5 x 10^{-5}, respectively). We also explored whether utilization rate of alternative sites, in RNAseq data from either wild-type or upf1-deficient cells, might correlate with conservation of splice sites; such analyses (Supplementary

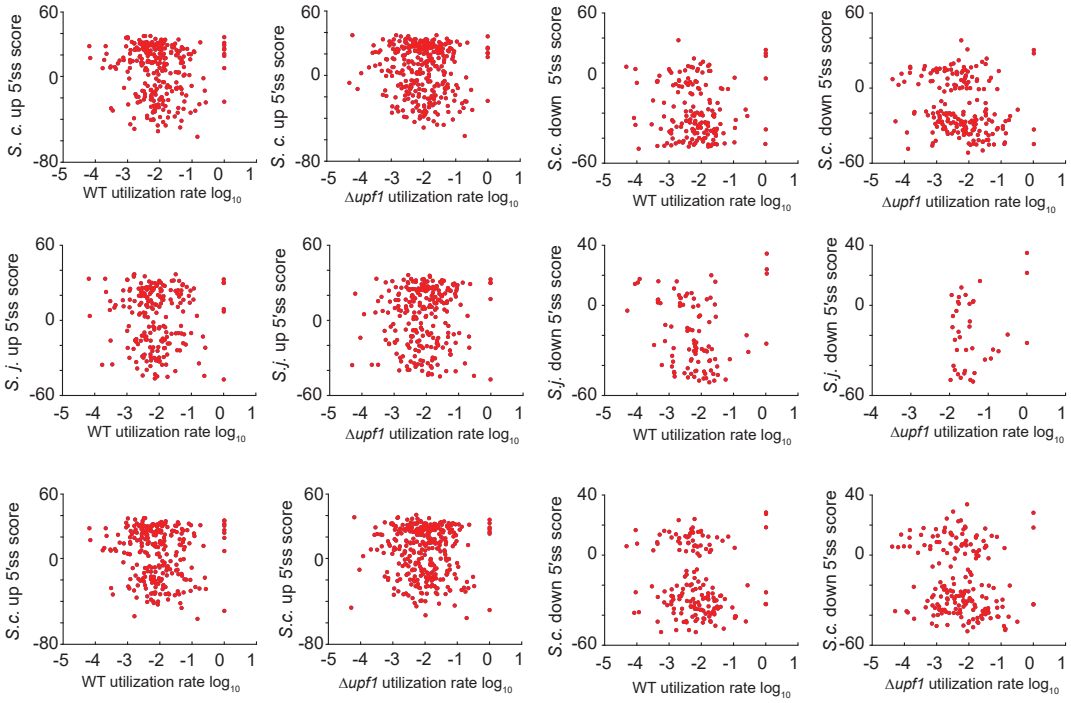

Figure 2.15: Orthologous 5′SS scores vs the rate of alternate splice site utilization in *S. pombe*.

Scatter plot of scores of alternate 5′SS in orthologous species (*S. octosporus*, *S. cryophilus*, and *S. japonicas*) vs the alternate splice site utilization rate ($\log_{10}$) in *S. pombe*. Utilization from either upstream (up) or downstream (down) sites for either WT or Δupf1 *S. pombe*.

Figure 2.15) identified no such correlations. Taken together, these results indicate extensive utilization of a wide range of splice sites in *S. pombe*, with many of the resulting aberrant splice products substrates for cellular decay pathways.

2.4.6 DISCUSSION

Pre-mRNA splicing is a critical component of eukaryotic gene expression. By temporally regulating the activation of different splice sites within a transcript, the process of alternative splicing provides a powerful opportunity for organisms to expand their proteomic diversity[57]. The importance of appropriate splice site selection is highlighted by the number of human diseases that are associated with mutations at or near these sites[76, 77]. Attempts to understand the rules that govern splice site selection have largely been driven by analysis of mRNAs present in cells, and inferences about the splice sites that were used to generate those mRNAs[78]; however, a major shortcoming of this approach is its failure to detect those splicing events that lead to destabilization of the spliced product.

Here, we have used lariat sequencing to enable an understanding of splicing in *S. pombe* not possible with RNAseq. As a tool for elucidating the diversity of substrates acted upon by the spliceosome, the major advantage of lariat sequencing derives from directly sequencing introns, rather than sequencing mature transcripts that are subject to RNA quality control pathways. Because lariat sequencing directly identifies both the 5'SS and the BP used during the splicing reaction, the data generated here represent the largest collection of experimentally identified splice sites in an organism with degenerate splice sites, allowing an opportunity to visualize the sequence constraints of the spliceosome in a way not previously possible.

Our data reveal a remarkable level of alternative and aberrant splice site activation across the *S. pombe* genome, including nearly 3000 examples of alternative splice site activation surrounding annotated introns, hundreds of novel

examples of exon skipping, and thousands of examples of novel introns (Figure 2.2). Importantly, because of the length limitations of lariat sequencing, these events underrepresent the total number of alternative and aberrant splicing events which must exist in the *S. pombe* genome. Together, our data suggest that rates of alternative splice site activation in *S. pombe* are around 3%, a value significantly higher than previous estimates[50, 61, 52]. The difference between our lariat derived error estimates and published RNAseq estimates is particularly noteworthy given the expectation that alternative splicing in the less complex genome of *S. pombe* is predicted to be lower than that seen in mammals[79].

The instances of alternative 5′ and 3′SS activation identified here are remarkable not only because of the frequency with which they occur, but also because of the relative strengths of the sites that are being activated. Indeed, given the strength of many of these alternative site sequences, the more surprising finding may be that they aren't activated at higher frequency. This observation, along with the failure to identify alternative splicing at many high scoring cryptic sites, underscores the significance of context in understanding splice site strength. The identification here of both activated and silent cryptic splice sites offers a powerful opportunity to better understand the constraints that drive splice site activation. Future experiments examining the subsets of activated and silent sites should provide insights into the mechanisms by which cryptic splice sites can be activated or repressed.

Our analyses of exon skipping events also provides a surprising insight into the mechanism of spliceosome assembly in *S. pombe*. In higher eukaryotes, where introns can be exceptionally large and exons tend to be shorter, strong evidence exists in support of an exon-definition model for spliceosome assem-

bly, wherein recognition of the 5'SSs by the U1 snRNP can facilitate recognition of an upstream BP/3'SSs[80]. By contrast, in lower eukaryotes like *S. pombe*, where introns are much shorter, assembly is thought to occur across introns, agnostic to the content of surrounding introns[73, 81]. In this context, our finding that exon skipping events tended to be enriched for poor scoring 5'SSs in the downstream introns was completely unexpected (Figure 2.2D). Interestingly, the simple model that the subset of introns identified here represent the few splicing events in *S. pombe* that utilize exon-definition for spliceosome assembly seems unlikely since the skipped BP sequences have scores which are largely indistinguishable from the global population. Alternatively, these results suggest the possibility that cross-exon interactions facilitate spliceosome assembly for many or all introns, including those presumed to assemble primarily by intron-definition, and that exons with weak downstream 5'SSs are more likely to exhibit exon skipping because of the decreased ability to utilize these cross-exon interactions.

Our results demonstrate that estimates of alternative splicing using RNAseq alone are likely to significantly underestimate the prevalence of alternative splicing. Splicing errors that generate aberrant transcripts will be significantly underestimated by RNAseq because they are likely to be subjected to RNA degradation pathways, including nonsense-mediated mRNA decay (NMD) and spliceosome-mediated decay (SMD)[38, 82]. Recent work in budding yeast, where splice sites conform to a strong consensus sequence, also revealed an unappreciated level of alternative splice site selection, much of which is masked by the NMD pathway[83]. Similarly, recent work in *S. pombe* identified widespread alternative splicing at rates approaching those detected here in the background of nuclear decay mutants[53]. Moreover, although we have associated the lari-

ats identified here with splicing events that have completed both chemical steps, a fascinating example of biologically-relevant, first-step only splicing has been demonstrated for the TER1 transcript in *S. pombe*[84]. As such, we cannot preclude the possibility that some of the lariats identified here are the products of reactions that have only undergone the first chemical step of the splicing pathway. Additional experiments will be necessary to fully understand the mechanisms by which these alternative splicing events are generated and subsequently linked with cellular decay pathways.

A surprising consequence of sequencing intron lariats was our discovery that lariats derived from different introns of multi-intronic genes have highly discrepant abundances, both as measured by lariat sequencing and confirmed with qRT-PCR and RNAseq. In organisms from yeast to humans, and including *S. pombe*, the nuclear processing of many non-coding RNAs is accomplished through endonucleolytic cleavage by RNase III homologs (Pac1 in *S. pombe*)[85]. Recent work in budding yeast demonstrates that Rnt1, the homolog of Pac1, cleaves more targets than previously expected[82, 86]; it remains unknown whether Pac1 or an as yet unidentified endonuclease contributes to the degradation of lariat introns.

While our data make it clear that alternative splicing is widespread in *S. pombe*, from the perspective of *S. pombe* biology, it is less clear that these events are functionally significant. Sites that have been selectively maintained over evolutionary time likely correspond to biologically meaningful alternative splicing events, whereas sites that have diverged at a neutral evolutionary rate are more likely to correspond to errors in splicing[87, 88]. Our comparative analyses of alternative splice site sequences indicate that the preponderance of al-

ternative splicing in *S. pombe* has not been maintained, even in closely related species. In the absence of evolutionary conservation, we conclude that the majority of the alternative splice sites we detected in *S. pombe* correspond either to rapidly evolving functional splicing events in *S. pombe*, likely true for only a very small subset of sites, or splicing errors that have arisen as a consequence of neutral genome evolution in the *S. pombe* lineage. Nevertheless, although the majority of the alternative events detected here likely have no physiological function in *S. pombe*, the widespread aberrant splicing identified here almost certainly plays an important role in genome evolution. Presumably, permissive alternative splicing, typically resulting in aberrant transcripts that are selectively degraded, provides the raw material from which advantageous events are selected during evolution. Our study suggests that the error rate intrinsic to splicing, acting upon cryptic splice sites, greatly exceeds previous estimates, perhaps facilitating more rapid acquisition of conserved alternative splicing events.

Although lariat sequencing is not readily amenable to sequencing introns from higher eukaryotes, chiefly because electrophoretic separation of large lariats is impractical, it will be important to develop variations on this approach that are suitable for intron sequencing from any species. Given the conservation of the splicing apparatus, and the overall similarity of splice site sequences between *S. pombe* and humans, we predict that a similar, or higher level of aberrant splice site activation will occur in humans as well. Knowledge of the locations and identities of splice sites activated in the human genome, together with the information derived here, will help in better understanding the mechanistic bases of splice site selection.

CHAPTER 3

DEVELOPMENT OF AN APPROACH FOR HUMAN LARIAT BRANCH SEQUENCING

3.1 Introduction

Much of what is know of the intron-exon structure of human genes and other higher eukaryotes comes from studies either identifying the locations of exon-exon junctions found in mRNA cDNA sequencing, and by examining the nucleotide conservation of related species to infer gene structure, and from predictive models that incorporate information from conservation, genome sequence, splice site signals, and other cis-information. These studies have greatly expanded our understanding of the locations of introns, but deeper sequencing suggests the existence of many more introns which are currently unidentified in intron annotations[20, 89]. The most prevalent and dominate mRNA isoforms of many gene models are robustly characterized at current levels of deep sequencing, however deeper sequencing experiments demonstrate there are many minor isoforms of alternative introns present in mRNAs. These minor isoforms demonstrate that the spliceosome acts in far more places than current annotations indicate, albeit at levels often far lower than primary isoform introns[20, 89].Some of these alternative minor isoforms may represent introns that become a dominate isoform in some yet unassayed circumstance. Some of the other minor isoforms may be spliced more prevalently, but are less represented due to processes that result in destabilization of spliced mRNAs [NMD]. The identification of these mRNA destabilization inducing introns and the quantitation of their splicing frequency can be challenging when using con-

ventional mRNA measurements. The sequencing of the cDNA of spliced RNA lariats is an alternative approach for understanding the location of introns. Additionally, introns identified using the exon-exon junctions in mRNAs do not identify the location of the branch point utilized during an intron's splicing. Computational approaches have sought to predict the location of branch points, however *in vivo* work has not yet verified the location of all the predicted branch points. Recently, several works have performed global sequencing of RNA lariats and have expanded the number of branch points identified *in vivo* to tens of thousands of introns[45], a quantity that is still far fewer than the several hundred thousand major isoform introns expected to be in the human genome. One limiting feature of these works is they often only consider the branch points of annotated introns, preventing the identification branch points in novel introns. While much has been learned about the splice sites found in mRNA, the lingering question still remains to be answered as to what are all of the sites utilized as introns in the human genome. More fully answering this question may further improve answers for other remaining questions regarding intron splicing.

Information about the locations and absolute quantification of human intron splicing currently remains incompletely determined and this presents a challenge for answering other questions regarding intron biology. One basic question is what are all the splice sites locations that the spliceosome identifies in human genes? Another globally unanswered question is what branch point used during splicing for each intron? Yet another question is are there introns that are frequently spliced, but are poorly quantified by mRNA steady state levels? Additionally, which splicing events are the result of errors? What do splicing errors indicate about how the spliceosome is selecting its targets? Expression levels of trans-factors are known to impact splice site selection[48]

and this leads to the question of how do different cellular contexts impact the selection of alternative and erroneous introns?

The quantification of splicing by measuring RNA lariats may provide insight in addressing the above questions. Identifying and measuring intron lariats provided an alternative quantification of splicing (see chapter 2) by identifying splicing events not quantified in typical mRNA deep sequencing experiments. A *de novo* genome-wide study of human splicing from the perspective of spliced lariats remains to be completed. Such a study may aid in identify introns that are currently unannotated, either due to the differential stability of the associated mRNA or RNA transcript, or else due to the splicing event that stalls at the splicing-intermediate step in intron splicing.

The technique employed to isolate and sequence lariats in *S. pombe* using 2D denaturing PAGE[35, 34] is not appropriate for the study of lariats in humans for several reasons. First, the 2D acrylamide gel system that enriches for lariats works well for *S. pombe* because the lengths of the introns in *S. pombe* are relatively short, with a median around 75 nucleotides in current annotations. This contrasts with human introns which are larger, with a median size over 1000 nucleotides. The resolving power of large RNAs in the high acrylamide concentrations of 2D acrylamide gels makes it difficult to separate large lariat RNAs from non-lariat RNA. Furthermore, the recovery of RNA from the gel for slowly migrating RNA often results in low RNA recovery from the gel. Another challenge in this isolation and sequencing approach is that intact lariat RNAs are recovered. Lariat cDNA that crosses the location of the lariat branch uniquely identifies the boundaries of a spliced intron is the most informative cDNA that can be sequenced. As intron length increases, short read sequencing reads from intact

lariats will more frequently correspond to the interior and therefore lesss informative portion of the intron. In this case, a smaller fraction of the sequencing reads will identify introns boundaries. The significant size difference between the median intron lengths of *S. pombe* and humans makes the 2D gel technology less useful for identifying introns. The short intron lengths in *S. pombe* results in a higher proportion of the sequencing reads containing sequence of the intron boundary. The longer human introns result in more of the sequencing depth recovering sequencing reads for the intron interior. To solve this issues another approach needs to be developed to isolate the lariat introns. One characteristic of this approach is the need for it to primarily sequences cDNA crossing the branch point as that information is the essential information required for identification of the spliced intron.

3.2 Results and discussion

3.2.1 A genome-wide approach for lariat branch sequencing

The unique structure of the RNA lariat differs from all other known RNAs in a cell and this unique substrate has properties that may be leveraged for the purification of lariats from non-lariat RNA. One purification strategy involves utilizing the circular properties of lariats[34, 46, 36, 39, 35] Alternatively, the branched portion of the lariat could be exploited for purification as it provides an unique RNA substrate with two 3′ termini and one 5′ terminus.

The large size of most human introns makes it unwieldy to use the circular nature of lariats to separate lariat from non-lariat RNA. The limit resolving

power of PAGE presents challenges that are likely not addressable due to human introns spanning six orders of magnitude in length. Enzymatic enrichment through degradation of non-circular RNA is possible for RNA lariats by using exonucleases that cannot degrade circular RNAs, such as the enzyme RNAse R[90, 46, 45]. This technique alone only enriches lariats and by a small fraction of a percent as some non-lariat RNAs are resistant to degradation by RNAse R. Additionally, larger lariats are more likely to randomly hydrolyze as part of the process, leading to their removal by the same exonucleases that degrade the non-circular RNAs. None of these mentioned approaches take specific advantage of the unique branched structure of lariats, which contains the 5'SS and branch point portion of the lariat.

To take advantage of the lariat branch, on option is to gently fragment the circular RNA lariat. Random RNA hydrolysis fragmentation causes the circular loop to fragment into many linear RNA fragments and one branched, or y-shaped, RNA [figure 3.5] which contains the lariat termini. This lariat branch has two 3′ termini and one 5′ terminus. One approach for lariat branch isolation is to attache different selectable labels to each of the two 3′ termini. Having two different labels provides an option for the tandem selection of RNAs with both of two different unique affinity labels on the 3′ termini. Non-branched RNA will have a single label and the non-branched RNA will fail to be selected during the tandem affinity label selection. Because this approach is annotation independent, it can allow for the *de novo* identification of splicing events using RNA lariat.

The general approach for tandem affinity selection for lariat branch isolation and sequencing protocol enriches for lariats as follows. First, RNA is frag-

mented at roughly one phosphodiester cleavage in 100 phosphodiesters. Second, one of two different labels are attached to each of the two 3 termini so that each terminus has a different label. Third, two different sequential tandem affinity purifications are performed. Fourth, reverse transcription is performed to generate cDNA. Fifth, this cDNA is then amplified using PCR to produce sufficient material for concentration quantification and subsequent sequencing. This general framework is amenable at different stages for additional strategies of RNA lariat purification. Additionally, there are also several choices available for what affinity labels are used for selection.

3.2.2 Lariat branch cDNA identification

Lariat branch cDNA identification was performed similarly as in chapter 2. Briefly, sequencing reads are first aligned to the genome. Those reads that do not align are candidates for lariat branch sequencing reads, which align to the genome at both the 5'SS and branch point of an intron. These reads are discovered through splitting the reads are the GT position, switching the order of the reads and then aligning them to the genome in paired-end mode. The alignments are generated using bowtie2[35] . The *S. pombe* samples were associated with known introns while the human samples were not.

3.2.3 Enrich for lariat RNA by optional rRNA depletion

The majority of RNA in a cell by mass is rRNA and this RNA often represents the vast majority of the RNA in the lariat branch sequencing experiments. One

straightforward approach is to remove rRNA from a sample by rRNA depletion. This depletion results from one of two typical strategies. Both strategies involve annealing tiled synthetic DNA oligonucleotide sequences that basepair with the rRNA. After annealing, one of two approaches is taken: either the sample is incubated with RNAse H to degrade the RNA in the rRNA-DNA base pairing hybrid, or else the rRNA is removed via hybridization pulldown of the rRNA from the sample. These approaches often remove about 95% of the rRNA in a sample. The advantage of this approach is the relative ease of these steps. The removal of rRNA is significant, though not sufficient alone, in reducing non-lariat cDNA from a sample. Even if rRNA was entirely removed at this stage, other noncoding RNAs make up a large proportion of the transcriptome relative to lariat RNA.

The removal of rRNA can be considered at multiple points in the protocol. One option is to deplete before starting the fragmentation step. The other option is to perform rRNA depletion sometime after RNA fragmentation. Several vendors provide rRNA depletion kits, such as the human/mouse/rat Ribozero rRNA depletion kit from Epicenter. Removal of rRNA by these kits works for *S. pombe* RNA due to the high conservation of eukaryotic rRNAs, though kits specialized to yeast also exist.

3.2.4 RNA fragmentation

RNA fragmentation is essential for making the two 3′ termini available for downstream labeling. This task can be accomplished by metal ion facilitated hydrolysis. This leads to the phosphate at the hydrolysis site resolving as a

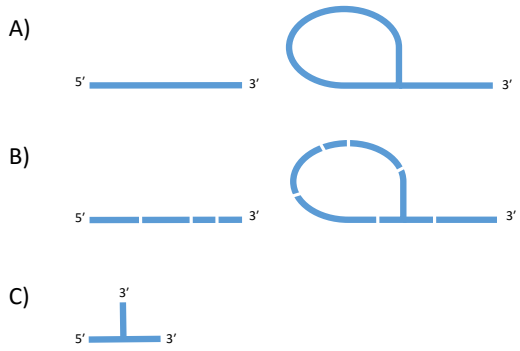

Figure 3.1: RNA fragmentation of lariats produces unique branched RNA

A) Example RNA and lariat RNA. B) Random RNA hydrolysis leads to the RNA in a sample fragmenting to different sizes. C) The lariat branch fragment is a unique fragment with two 3′termini and one terminus.

cyclic 2′-3′ phosphate. It is unclear how long the 2′3′ cyclic phosphate is stable before it resolves into either a 2′ or 3′ phosphate. Each of the possible phosphate groups, namely 2′, 3′, and 2′-3′cyclic, interfere with downstream RNA labeling as the requisite chemistry for all downstream labeling requires the RNA 3′ termini to have a 2′-OH and a 3′-OH. The unwanted phosphates may be removed using Calf Alkaline Phosphatase (CIP) or T4 Polynucleotide Kinase (PNK). CIP resolves single phosphates at the 2′ or 3′ positions of 3′ termini. T4 PNK (unmodified) resolves 2′-3′ cyclic phosphates, at a lower rate than it resolves 3′ phosphates. In the presence of ATP, T4 PNK places a mono phosphate at the 5′ termini, another required terminal chemistry. Sequential reaction of CIP and T4 PNK establish the appropriate terminal chemistries for RNA ligation, the next enzymatic step.

After RNA fragmentation, an optional step to size the RNA is available. In this case, the RNA is sized on a 7.5M urea denaturing acrylamide gel at 6%. The goal of RNA sizing is several fold. First, the cDNA that is eventually se-

quenced will be in a smaller range than the products of the RNA fragmentation reaction. Second, unwanted RNA may lead to co-purification of background material (discussed further in the affinity purification section). The denaturant and the low acrylamide gel should allow the fragmented lariat branches to run close to their nucleotide mass in which case the lariat RNAs may be sized between 30 nucleotides and somewhere above 150 nucleotides. A high percent of acrylamide may lead to the lariat branches migrating above what their nucleotide mass would predict, although no direct evidence has been presented to corroborate this concern or, if it is a concern, to know what concentration of acrylamide deferentially affects the migration of branch RNA.

3.2.5 RNA 3′ affinity label choice

This approach for lariat sequencing depends on the utilization of two different 3′ affinity labels. There are several labels can be employed for the affinity selection steps. Generally, these labels may be affixed to the RNA by ligating a 3′ labeled DNA adapter to the 3′ terminus of the RNA. Alternatively, an affinity label may be chemically attached to the 3′ terminus by periodate oxidation followed by conjugation with a labeled hydrazide molecule.

The simplest DNA adapter based label involves the reverse transcription oligonucleotide primer annealing to a ligated 3′ adapter. This label derives its specificity from the reverse transcription primer that should anneal specifically to the 3′ adapter and not in other RNAs.

An additional label purification option employs oligonucleotide hybridization and pulldown using the 3′ DNA adapter as the target of the oligonu-

cleotide. A more complicated labeling option involves incorporating an epitope tag into the DNA adapter and then selecting for it using antibody conjugation and pulldown purification. Alternatively, in place of an epitope tag, biotin can be incorporated into the DNA adapter and the biotin may be selected for with streptavidin purification.

An additional option for 3′ labeling of RNA is to chemically attach a hydrazide-biotin label to a 3′ terminus. Biotin labels are selected for by streptavidin purification and Because the biotin-streptavidin interaction is among the strongest affinity interactions, the use of biotin as a 3′ label is highly desirable. The chemical biotinylation reaction is a two step procedure. The first step chemically oxidizes the cis-diols at the 3′ terminal ribose of an RNA. Oxidation is achieved using sodium meta periodate ($NaIO_4$) and leads to the formation of a dialdehyde at the 3′ terminal ribose. The next step is the biotinylation reaction and that is achieved by reacting the oxidized ribose, which now has two aldehydes, with a biotin-hydrazide molecule. This reaction produces a hydrazone bond that conjugates an aldehyde to the hydrazide of the biotin-hydrazide (figure 3.2).

Affinity labels are attached to 3′ termini randomly until all the termini are labeled. Regardless of which affinity labels are chosen, it is important that each of the two affinity labels are each attached to ~50% of the 3′ termini. There are two general ways to accomplish this task. One option is for both tandem selections to utilize 3′ adapter ligated oligos and to equally mix the two adapters in the ligation reaction. Achieving an equal frequency in each adapter ligating may be difficult due to primary and secondary effects on the ligation reaction[91]. This is attractive due to being able to perform both labeling reaction simulta-

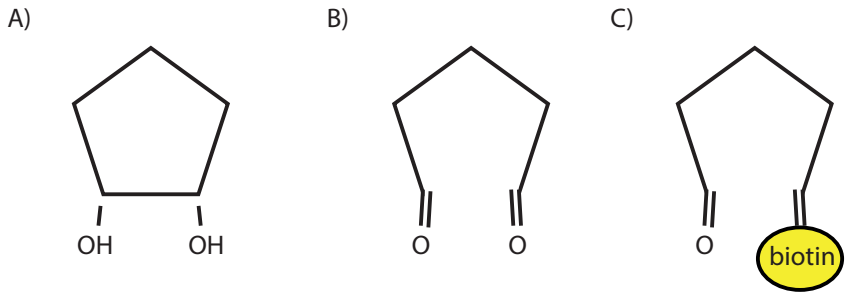

Figure 3.2: Periodate oxidation and biotinylation of RNA 3'terminus

A) Ribose of RNA 3'terminal nucleotide. B) Periodate peridoate oxidation of ribose generates two aldehydes. C) Biotin-hydrazide coupled to aldehyde

neously, thus simplifying and shortening the protocol. Alternatively, the two labels may be attached sequentially. In this case, either the 3′ adapter ligated oligonucleotide or the chemically attached biotin tag is applied to ~50% of completion for the 3′ termini. Next the other label is applied, this time to completion. This approach is more desirable if it is expected that spliced lariats have very short tails, as is the case for Δdbr1 yeast strains where RNA lariats are stabilized and are likely not protected from cellular 3′ to 5′ exonucleases. In this case, it is likely that the RNA lariats subjected exonucleases and have short tails. It is uncertain if lariats with short tails impede 3′ adapter ligation. The chemical labeling approach should work better in this case since it should work independently of the lariat tail length. If this is the case, then former approach adding both labels in a single ligation reaction may be more sensitive for capturing splicing intermediates or recently spliced lariats when there hasn't been enough time for exonucleases in the cell to shorten the lariat tail. It is uncertain if human lariats that accumulate from DBR1 knockdown would have lariat tails

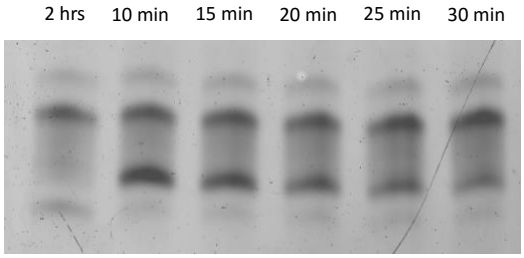

Figure 3.3: Optimizing periodate oxidation for 50% biotinylation

An 18 nucleotide RNA was subjected to periodate oxidation at pH 8.5 for the given times and then was subjected to biotinylation with biotin-hydrazide. This results in two dominant species, either the unoxidized RNA and the oxidized and biotinylated RNA.

lengths that are too short for the ligation approach. Optimizing a lariat branch sequencing protocol using the lariat enriched *S. pombe* Δdbr1 RNA, may require the use the chemical biotinylation approach for labeling the RNA.

Achieving 50% labeling in a RNA ligation requires estimating the number of picomoles of RNA in a sample and then limiting the 3′ adapter in the ligation reaction so that only ~50% of termini are labeled during the reaction. An alternative option is to alter the chemical biotinylation reaction to achieve ~50% periodate oxidation of 3′ temini and then performing the biotin-hydrazide coupling completion. One strategy for acheiving this limited periodate oxidation is to adjust the pH of the reaction conditions and altering the time of the reaction 3.3. In this case, 20 minutes at pH ~8.5 is sufficient for achieving ~50% biotinylation. This deviates from the standard protocol for periodate oxidation which performs biotinylation at pH 5.3 for 2 hours to completely oxidize all available 3′ termini.

Extra consideration is needed when working with periodate. Periodate is used to oxidize carbohydrates like ribose and due to its high sensitivity and specificity for reacting with the 3′ terminal ribose in RNA samples. However, periodate also reacts with glyocogen, a common nucleic acid carrier, and periodate also reacts with the different amino acids to varying degrees[92]. The lariat branch sequencing protocol utilizes several enzymatic steps. While complete removal of some enzymes in not necessary for many downstream steps, it is essential that enzymes are removed prior to periodate oxidation. Phenol-chloroform extractions are efficient for the removal of proteins from nucleic acids and phenol-chloroform extracted RNA adequately prepares the RNA for periodate oxidation.

A single selection of a 3′ affinity label may not be sufficient for complete removal of background RNAs. Multiple iterations of selection for the 3′ affinity label may help improve background reductions. Alternating between selecting for the two different labels (excluding the use of reverse transcription as a label until the final step) may improve the reduction of background more rapidly than multiple selections using the same label. The rationale for this improved reduction is expected to come from removing background RNA that co-purifies with for one label. Hybridization of a RNA with one label with an RNA with the other label essentially creates a hybridization pull down of such RNAs. Repeatably alternating between the labels should help mitigate co-purifying background material assuming that the hybridization is disrupted and the two RNAs are not able to hybridize with each other again.

3.2.6 5′ adapter ligation

After affinity purification, it is necessary to ensure that a 5′ RNA adapter is ligated to the 5′ position of the RNA. This may be done earlier, however, it is essential that the adapter is in place prior to reverse transcription.

3.2.7 Reverse transcription

The 2′-5′ phosphodiester linkage of the lariat branch RNA is an unique structure and research usage of reverse transcritpion enzymes is not currently optimized for the regularly creation of cDNA using lariat branch RNA as a template. It is known that the reverse transcription enzyme has difficulty recreating cDNA from this branched RNA product[43]. The goal of sequencing lariat branches is made more difficult by the reduced ability of reverse transcriptase to create cDNA from lariat branch RNA template. Effectively, the cDNA conversion penalty requires an increased purification of RNA lariat branches from non-branched RNA. It may be possible to mitigate the cDNA conversion penalty by optimizing the reverse transcription reaction to promote cDNA conversion of lariat branch RNA. Methyl groups located on the 2′ position of an RNA ribose are known to impede reverse transcription and the use of manganese in the reverse transcriptase reaction buffer is known increase cDNA conversion of 2′ methyl containing RNAs. Thus, the use of manganese in the reverse transcription reaction may be a suitable approach for increasing the enrichment of lariat branch cDNA generation.

MMLV reverse transcriptase is a commonly used reverse transcriptase. Invitrogen Superscript II is an MMLV reverse transcriptase with a mutated RNAse H

domain that has reduced RNAse activity. It is known that 2′ modified RNA are poor substrates for cDNA synthesis. Manganese is known to increase reverse transcription across 2′ positions of some modified riboses. The phosphodiester linking the 5′SS to the branch point is a 2′ modification and the efficiency with which lariat branch cDNA is created by reverse transcriptase is modulated by the presence of manganese in the reaction buffer (figure 3.4). For the conditions tested, the choice of manganese concentration correlates with the conversion frequency of lariat RNA into cDNA. A notable observation in this experiment is that cDNA corresponding to mRNAs decrease when Mn^{++} alone is used during reverse transcription as compared to the control that only uses the standard conditions of Mg^{++} only. A possible explanation for this behavior is that reverse transcription ends prematurely due to some effect of having Mn^{++} in the reaction conditions.

Given that the lariat branch levels increase in Mn^{++}, there may be value in not using Mg^{++} in the reaction buffer. However, it is possible that the loss of some exonic signal is may indicate reduced processivity of the reverse transcription enzyme. This may be relevant for optimizing the lariat branch sequencing method, though how important it is remains to be tested. Reverse transcription using a normal cation concentration but with one half of the cations being Mg^{++} and the other half being Mn^{++} leads to an appreciable increase in lariat branch cDNA levels. It also maintains the level of exonic cDNA when compared to the Mg^{++} condition.

The manganese dependent increase in lariat cDNA is significant, approximately 30-fold. One extra consideration to remember when working with Mn^{++} is that the presence of Mn^{++} interferes with the downstream PCR. Mn^{++} is some-

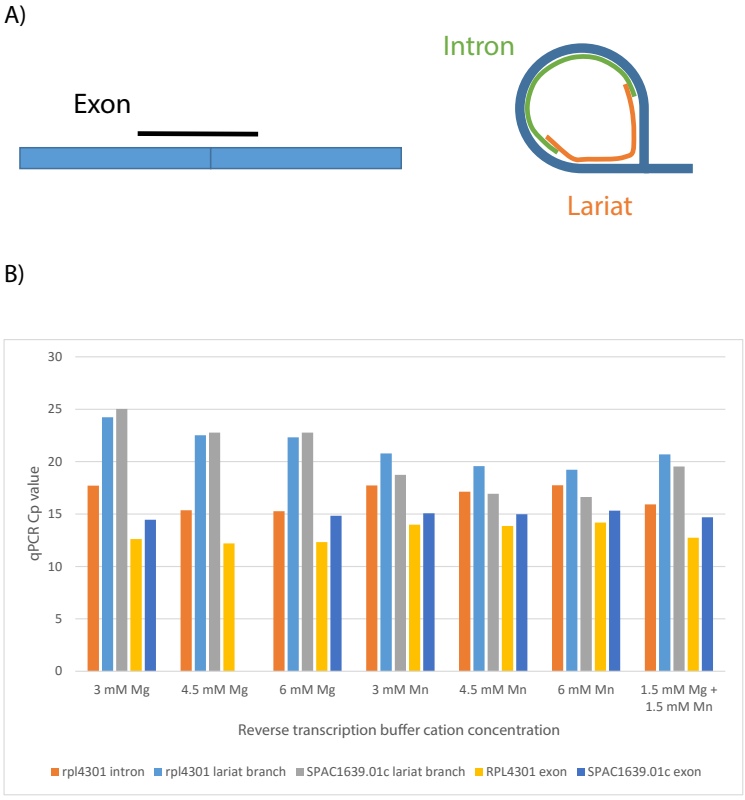

Figure 3.4: qPCR measurements of different reverse transcription conditions

qPCR measurements of cDNA produced using different types and concentrations of metal ions. A) example diagram of the dominant qPCR targets for exon, intron, lariat branch for a lariat debranching deficient *S. pombe* strain. B) qPCR measurements for an intron, its corresponding lariat branch, a second lariat branch, the exon former introns associated exon, the latter introns associated exon. Metal ions were either magnesium (Mg), Manganese (Mn), or Mg + Mn with concentrations indicated.

times utilized at low concentration during PCR amplification as a method to randomly introduce mutations in the PCR amplicon product. Since this behavior is unwanted for lariat branch sequencing, desalting either through precipitation or the use of desalting columns is needed to remove the Mn^{++} and prepares the cDNA sample for high-fidelity PCR.

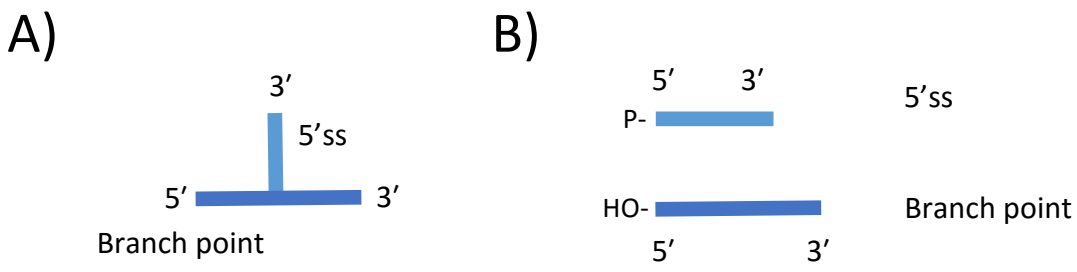

Figure 3.5: Terminal 5′ chemistry after debranching lariat branch.

A) Lariat branch fragment with the 5′ terminus and two 3′ termini with the branch point fragment (dark blue) and 5′SS fragment (light blue) indicated. B) Debranched lariat branch produces two fragments. The 5′SS fragment (light blue) has the correct 5′monophosphate terminal chemistry for RNA ligation and the branch point fragment (dark blue) has the same terminal chemistry prior to debranching, here indicated as a 5′ OH.

PCR amplification using PCR is done to generate a sufficient concentration of material for downstream Illumina high-throughput sequencing. The reverse transcribed cDNA contains two synthetic sequences corresponding to the 5′ adapter and 3′ adapter oligonucleotides. PCR amplification is performed using a DNA primer oligonucleotides corresponding to the reverse complement of the 3′ adapter oligonucleotide and the forward sequence of the 5′ adapter oligonucleotide. PCR amplication should be specific to only those molecules that received both adapters during the ligation steps.

It is preferable that sequencing reads contain the branch point to 5′SS transition. cDNA derived from lariat branch reads where the branch point is not in the sequenced portion leads prevents the identification of the 5′SS used by the lariat. The length of the sequencing reads is an important factor that impacts the frequency with which associated sequencing read contains the lariat branch

point in the read. Sizing of the final amplified sequencing library so that the molecules in the library have lengths below the sequenced length mitigates this challenge.

3.2.8 Library construction challenges

The first lariat sequencing experiment described in this work did not remove protein after the RNA ligation step. As 10 units of NEB RNA ligase I contains about 42.5 pmol of protein (personal communication with NEB), it is plausible that the biotinylation step was compromised and only a small fraction of the RNA 3′ termini was labeled with biotin. It is unclear if this concern is warranted without additional experiments to test the rate of periodate oxidation relative to amino acid oxidation. Given the ease in removing proteins from an RNA sample, the second iteration of lariat branch sequencing described here included a phenol chloroform extraction step to avoid this concern.

	genome aligning	unalign
Standard	2606187	1779414
Mg^{++}	5780347	1510421
rRNA depleted	1002162	491927
WT	5009757	2115096

Table 3.1: First lariat branch sequencing genome alignment

Total number of reads aligning and not aligning to the genome for first branch sequencing experiment .

3.2.9 The first lariat sequencing experiment

An initial lariat branch isolation and sequencing experiment was performed to test several potential optimizations. In this experiment, the RNA was first fragmented and then sized on a 5% urea acrylamide gel to recover RNA fragments with lengths in the range of 30-150 nucleotides. This size range was used to reduce the amount of RNA that was too long for downstream sequencing and to reduce the possibility that the presence of these large RNA fragments might increase the levels of background RNA. The choice of performing 3′ adapter ligation before biotinylation was used for this experiment. The optional RNAse R treatment was performed on these libraries to reduce the amount of linear RNA fragments available for hybridization pulldown of background RNAs. This experiment included several single-step optimizations for comparing the effectiveness of rRNA depletion with no rRNA depletion, to compare the usage of manganese in the reverse transcription reaction against the standard reverse transcription condition, and to compare the lariat levels between a Δdbr1 against a wild-type strain. The above samples utilized RNA extracted from *S. pombe*.

The removal of rRNA was accomplished using the Zymo RiboZero rRNA depletion kit. The experimental outcome of rRNA depletion showed a significant decrease in rRNA levels, effecting around ~7-fold increase in lariat branches (after normalizing sequencing depth using reads that align to the genome) when compared to the amount of lariat present in the standard library condition (table 3.1). One possible factor negatively impacting the removal of rRNA levels is that the distribution of rRNA fragment species recovered from the gel sizing step may not be proportional to the amount of the various species

of rRNA that the rRNA depletion kit is designed to remove. Even if this is the case, the approximately 7-fold enrichment for lariat branch RNA is significant and the utilization of rRNA depletion provides one avenue for lariat branch enrichment.

The effect of using Mn^{++} as the metal cation during reverse transcription was tested by lariat sequencing (table 3.1). For this experiment 4.5 mM Mn^{++} was used in place of 3 mM Mg^{++}. Lariat branch sequencing showed that 4.5 mM Mn^{++} and with no Mg^{++} provides a global increase in lariat branch cDNA levels when compared against the appropriate control. This effect is significant as it affords a 28.8-fold increase in lariat branch levels.

3.2.10 Evaluating the effectiveness of human lariat branch sequencing

It is possible estimate improvements in the branch sequencing using *S. pombe* alone, however to gauge how well branch sequencing recovers human introns requires sequencing human samples. Human rRNA depleted HEK293 RNA was sequenced using branch sequencing. These data were aligned to the human genome rev19 first to remove normal genome aligning reads. Then the reads that failed to align to the genome were aligned as paired split reads, similar to chapter 2, using a gap size of 50,000 nucleotides (data not shown). However, unlike in chapter 2, the reads were not checked for whether they could be normal sequencing reads that failed to align to the genome, a critical analysis step that removes many false split read alignments. The result of this sequencing data found an upper bound of 2.9% of the reads were split reads. A cursory

examination of the split reads showed that they rarely aligned to intron boundaries and it was decided that whatever the true lariat branch sequencing rate was, it is not high enough to constitute a success. A quick examination of the length of the read pairs found the majority of split read pairs were short, with 8.4% having one pair of the two pair split read having a length above 17 nucleotides. Short read lengths are more likely to be the result of genome aligning reads that failed the initial alignment. This suggests that the 2.9% figure should be revised down to an upper bound of 0.24%. A cursory examination of the sequencing reads from this set of sequencing reads was performed and the upper bound estimate was found have few lariat branch sequencing reads. One challenge in using this approach for the mapping of lariat branch sequencing reads is some human introns, especially exon skipping introns, will require the use of very large gap sizes when using a short read aligner. This both increases the possibility for false positive alignments and dramatically increases the time required to perform the alignments.

3.2.11 The second lariat branch sequencing experiment

Experimental setup

The second lariat branch sequencing experiment sought to explore the biotin first approach to constructing a sequencing library. It also sought to test whether mis-priming with the reverse transcription oligonucleotide is a significant source of the background RNA being sequenced. This test was accomplished by using a modified RA3 oligonucleotide that has 24 dA bases located on the 3′ end of the oligonucleotide to serve as the site for the reverse transcrip-

tion primer when using a dT(18) oligonucleotide for priming.

Results of a biotin first lariat branch sequencing experiment

The second lariat sequencing experiment tested two parameters using human or Δdbr1 *S. pombe* RNA. The first parameter was a binary check whether moving the biotinylation of the RNA prior to the 3′ adapter ligation would work. The second parameter was reverse transcription using the 3′ adapter sequence oligonucleotide or else using a poly dT oligonucleotide. This poly dT oligonucleotide is expected to cut down on sequencing reads resulting from mispriming of reverse transcription.

Evaluating the *S. pombe* lariat sequencing libraries

The second lariat branch sequencing dataset used Δdbr1 RNA that was not rRNA depleted. For the normal reverse transcription primer, 5,314,779 reads mapped to the genome and 106,191 mapped as split reads to known introns and their alternates. The rate of this alignment is ~2%. This serves as a lower bound for intron recovery, as intron skipping events and novel introns are not part of this count. In aggregate, this dataset is best compared to the standard *S. pombe* version from the first sequencing experiment which recovered ~0.19%. This us an 11-fold increase in lariat cDNA. Had rRNA depletion been included, we would expect approximately 6-20 fold more lariat branch cDNA reads. The lower estimate of 6-fold enrichment by rRNA depletion is based on the previous lariat branch sequencing experiment. The upper bound of 20-fold increase in lariat branches is based on vendor rRNA depletion estimates. Due to the im-

pact that gel sizing may have had on the first branch sequencing experiment, it would may be reasonable to expect ~20-fold depletion of rRNA if the rRNA depletion step was performed prior to RNA fragmentation. Taken together, these numbers indicate that lariat branches in a Δdbr1 *S. pombe* strain may be be possible to somewhere in 12-40% in total, representing two orders of magnitude improvement over the 0.2% reported in work published using a RNAse R and rRNA depletion of *S. pombe* Δdbr1 RNA[36].

Evidence of reverse transcription mis-primed cDNA

While linear RNAs should be removed during the lariat branch sequencing protocol, some amount of non-branched, linear RNA makes it through as background present during the reverse transcription reaction. Some fraction of the linear RNA containing the biotin label (and consequentially lacking the RA3 oligonucleotide adapter) may be a substrate for weak hybridization with the reverse transcription oligonucleotide primer. Depending on the strategy employed for removal of the linear RNA, there can be copious quantities of this background substrate for reverse transcription priming. When the oligonucleotide sequence of the reverse transcription primer is also the sequence of one of the PCR oligonucleotides, this leads to the introduction of the oligonucleotide primer sequence at that location in the reverse transcribed cDNA. When this primer also corresponds to the exact sequence of the 3′ adapter, it can lead to mis-primed background RNAs becoming substrates for downstream sequencing.

When these mis-primed RNAs become cDNA they can lower the level of lariat branch enrichment. The previously mentioned strategy to get around

this problem is to use a downstream portion of the PCR sequence for reverse transcription priming. In this case the 3′ portion of the PCR oligonucleotide sequence is not identical to the 3′ portion of the reverse transcription priming oligonucleotide. This change in oligonucleotide primer sequences makes it less likely that PCR will occur on mis-primed cDNA from background RNAs. Evidence of this mis-priming exists when comparing two sequencing libraries created using a 3′ adapter with 24 nucleotide polydA at the 3′ terminus. This library can be primed using both the Illumina small RNA cloning primer and also a 18mer oligonucleotide dT_{18} primer. An examination of aligned sequencing reads from both of these libraries shows examples of read peaks present in the Illumina reverse transcription oligonucleotide sequencing library that are missing from the poly dT_{18} oligonucleotide primed library, though at this time it is not possible to . It is important to note that the library strategy employed in this test did not attempt to remove biotinylated linear non-branched RNAs. It is not possible to determine how much of a problem mis-priming would be if lariat branch isolation is improved. Regardless, using the alternate reverse priming oligonucleotide solves this potential problem in future lariat branch sequencing efforts.

Two human samples were included in the second lariat branch sequencing experiment. The first utilized the normal reverse transcription primer. The second used the oligonucleotide dT_{18} primer. Similar to the first experiment, these reads were aligned to the genome and the reads that failed to align were utilized as candidates for split read alignment. Like in the first experiment, the human split reads were only evaluated for their ability to align to the genome in a paired-end split fashion. These alignments were not computationally associated with known introns and were not evaluated for whether they could

have failed the first genome alignment prior to being split due to heuristics in bowtie2. The normal RT primer generated ~0.4% split reads and the oligonucleotide dT_{18} primer generated a sequencing library with ~0.2% of the library mapping as split reads. Upon manual inspection, the vast majority of these alignments were unlikely to be true, intron aligning branches. For this reason, the data were not further analyzed for split reads.

Concluding remarks

The two orders of magnitude increase in lariat cDNA levels between the first and the second *S. pombe* lariat branch sequencing experiments demonstrate that lariat branch sequencing has the potential to replace 2D gels. While human lariat branch sequencing is not yet mature, there is room for additional steps to improve on lariat branch cDNA recovery. Additional rRNA could be removed using rRNA depletion. The problem of background RNA making it to the end of library construction could be mitigated by increasing rounds of selection for biotin and by introducing a second selection step, such as oligonucleotide dT_{25} magnetic beads capture of RNA with the 3′ poly dA DNA adapter. Incorporating an oligonucleotide dT_{25} capture between the two biotin selections should reduce the material that is co-hybridizing with the biotinylated background RNA. Reduction of co-hybridizing material should result in less background RNA that contains a 3′ adapter. While lariat branch sequencing in humans is not yet mature, there are opportunities remaining to further enrich for lariat cDNAs.

CHAPTER 4

CONCLUSIONS AND FUTURE DIRECTIONS

The sequencing of lariat branches has the potential to address several outstanding questions regarding the biology of RNA splicing and gene regulation. Currently, these technologies do not produce sequencing libraries with enough enrichment of RNA lariat branches for the purpose of a thorough examination of RNA splicing. A major challenge in the isolation and subsequent sequencing of RNA lariats is the effect that their relatively short half-lives have on their abundance relative to other RNA species in the cell. The lariat branch sequencing approach described in chapter 3 rivals the 2D gel approach in chapter 2 for the level of enrichment of lariat branches using the Δdbr1 *S. pombe* RNA. The lack of a size bias against long introns demonstrates that this approach outperforms the 2D gel approach describes in chapter 2. While this approach achieves a similar level of enrichment as 2D gels, the enrichment it achieves is not enough to for a deep and *de novo* characterization of spliced lariats in humans. Using this technology for human lariat branch isolation requires additional enrichment, approximately the level of enrichment that would be yielded by the *in vivo* accumulation of lariats resulting from the removal of the lariat debranching enzyme from the human cells. This removal of the lariat debranching enzyme is easily achieved in yeast, but it remains to be seen if it is possible to remove the lariat debranching enzyme in humans.

Looking forward, the goal of creating a highly enriching method for the sequencing of spliced lariats in humans needs to overcome several challenges. Several unanswered questions regarding *in vivo* lariats have implications for the feasibility of the purification of spliced lariats. One central question is how

many spliced lariats are in a given sample? Along this thought is the question of how much enrichment is necessary? Currently, there are no global estimates for the abundance or half-lives of spliced lariats. Without this information, it may not be possible to accurately predict how much enrichment will be necessary for a lariat branch sequencing technology to be considered a success. One possible strategy to improve on the current level of enrichment would be the sequential utilization of several approaches.

Alternate approaches exist for lariat enrichment, though the current state of these approaches only partially allow for the genome-wide discovery of lariats. Exonucleases, like RNAse R, can enrich for lariats by degrading linear RNA, but the quantitative recovery of large lariat RNAs might not be possible using this technique. Alternatively, the pull-down of lariats using the hybridization of oligonucleotide probes targeting the termini of lariat RNA could dramatically increase the purity of the starting RNA sample, though this strategy requires targeting annotated introns and thus will now allow for *de novo* discovery of lariat RNAs. Other options for enrichment may be possible. One possible approach would be the isolation of fragmented lariat branch RNAs with the use of gel mobility shift techniques (similar to 2D gels). This approach would work if branched RNA shows a level of gel mobility retardation that is similar to circular RNA and thus migrate in a size range resolvable from non-branched RNA.

APPENDIX A

APPENDIX

A.1 Protocols and details relevant to lariat branch sequencing

Estimating the molar concentration of fragmented RNA

Either a negative bionomial or an exponential approximation is useful for estimating the number of fragments resulting from the random RNA fragmentation reaction[93].

RNA isolation from human cells

RNA extractions are performed according to the TRIzol protocol for isolating RNA from human cell lines. Precipitate with linear acrlyamide in place of glycogen (all downstream precipitations must use linear acrylamide until after the preiodate oxidation step).

RNA fragmentation using $ZnCl_2$

RNA fragmentation is performed in 10 mM $ZnCl_2$ pH 7.6 for 15 minutes at 65°C. The reaction volume is 10 μl of H_2O per 10 μg of RNA. Reaction are quenched by moving to ice and adding 1/10th volume of 0.5 M EDTA pH 7.6.

The fragmentation reaction is precipitate prior to the next step. Precipitations are performed by raising the sample volume to 4-fold the fragmentation reaction volume, by adding 1/10th volume 3 M NaOAc pH 5.4, adding 6 μl of

linear acrylamide carrier (a nucleic acid carrier lacking cis-diols), and 2.5 volumes of of 95% ethanol. The precipitation is chilled in liquid nitrogen or incubated at -20°C for several hours. The precipitation reaction is spun down at 14,000 x g or greater for 20 minutes at 4°C. The resulting pellet is washed using 70% ethanol and spun down at 14,000 x g for 5 minutes at 4°C. The wash is repeated. The pellet is air dried and resuspended in an appropriate volume for subsequent gel sizing or CIP treatment.

Sizing fragmented RNA using 7.5 M urea acrylamide gels

The samples have their H_2O volume adjusted to final volumes of 25 μl sample per well mixed with 25 μl of 1X RNA denature gel loading buffer (95% formamide, 20 mM EDTA pH 8.5) per well. A ssRNA ladder that resolves the intended size range for gel excision is also prepared as above. Samples and ladder are heated at 95°C for 2 minutes and are then place on ice.

The 7.5 M urea acrylamide 1 x TBE gels are cast at a concentration of 5%. The gels are pre-run in 1X TBE loading buffer for 20 minutes at 40 mA. Wells are then blown out using TBE buffer and 50 μl of sample is loaded per sample well and 50 μl of ladder per ladder well.

It is ideal to size a short size range to make subsequent gel elutions easier. The gel is run until the running dyes (bromophenol blue and xylene cyanol) resolve to about 1.5- 2 cm apart.

The gels are stained in SYBR gold or another appropriate RNA stain. RNA sizing and gel elution is performed using a squeeze and freeze protocol.

Optional removal of rRNA using the Ribo-zero rRNA depletion kit

Starting with up to 4.5 μg of sized RNA, use the Zymo ribo-zero rRNA depletion kit to deplete rRNA per manufacturer directions. Precipitate the reaction and suspend in 45 μl H_2O. Note that a user should be cautious of how much fragmented RNA is loaded into the rRNA depletion kit because the kit is not designed for the molarities of rRNA present in sized and fragmented rRNA.

Removal of cyclic phosphates using CIP

The is no standardized reaction for CIP (Calf Aklaline phosphatase) enzyme treatment. In this work, 1 μl of CIP enzyme is used per 60 pmol of RNA in 1x the manufacturer's CIP reaction buffer. The reaction is incubated at 37°C for 60 minutes. The CIP reaction needs to be be cleaned by a phenol:chloroform extraction followed by ethanol precipitation.

5′ phosphorylation using Polynucleotide Kinase

T4 PNK (Polynucleodtide Kinase) is an enzyme used to ensure the presence of the correct terminal phosphorylation status of the RNA termini. T4 PNK reactions were set up in the manufacturer's 2x reaction buffer, 1 μl of enzyme per 60 pmol of RNA. The reaction was incubated at 37°C for 30 minutes. If the next step involves periodate oxidation, then the reaction needs to cleaned by phenol:chloroform extraction and ethanol precipitation, otherwise the reaction can be cleaned by ethanol precipitation.

Optional ordering of RA3 ligation or periodate oxidation with biotinylation

There is an option to perform either of the 3′ adapter ligation or the periodate oxidation first and then perform the other reaction second. There are several rationale's for which is performed first. In the end, the most important consideration is that the first reaction is performed to 50% of completion and the second reaction is performed to completion.

In the event that the 3′ adapter ligation is performed first, then the options to achieve 50% adapter ligation is accomplished by limiting the ratio of adapter to RNA termini. When periodate oxidation is performed first, then the approach for 50% biotin labeling is accomplished by changing the pH of the reaction buffer used during the periodate oxidation step. Both of these approaches are difficult to optimize. Performing RNA ligation to 50% completion depends on the population of RNA termini sequences and their biases in ligating with the adapter[91]. Periodate oxidation being done to 50% of completion has its difficulty in that it is pH dependent and small changes in pH can drastically impact the percent of completion achieved. To address this concern, it is prudent to retest the oxidation conditions when using new reagents or reagents prone to changing in pH with exposure to air. This is not a concern when periodate oxidation is done to completion using pH 5.4.

3′ adapter ligation

An oligonucleotide based on the Illumina RA3 oligonucleotide is used for 3′ adapter ligation. IDTdna may be used for oligonucleotide synthesis using the following description: /5rApp/TGG AAT TCT CGG GTG CCA AGG /3SpC3/

for the standard RA3 oligonucleotide.

When the ligation of the 3′ adapter is done to completion the RNA ligation reaction is set up using 1X RNA ligase buffer, with 1 μl of T4 RNA ligase I per 12.5 μl of reaction volume, and with 100 pmol of RA3 oligonucleotide per 50 pmol of fragmented RNA in the reaction. When the reaction is not done to completion, then the concentration of adapter should match the concentration of the RNA. The sample is briefly mixed and incubated at room temperature for 2 hours. It is cleaned up by a ethanol precipitation unless periodate oxidation follows. If periodate oxidation follows, then a phenol:chloroform extraction must to be performed prior to the ethanol precipitation. The phenol:chloroform extraction removes the enzyme which is likely provide a significant source of reaction substrates for the periodate oxidation reaction and which may be labeled with a high concentration of biotin labels.

Optional Rnase R treatment to remove RNA without a 3′ adapter

RNAse R may be used to remove RNAs that do not have a three prime adapter. This step is only relevant if the 3′ adapter is ligated prior to periodate oxidation and biotinylation. In this case, the lariat branch fragments that have a 3′ adapter on the 5′ SS branch will be expected to mimic the resistance to RNAse R degradation demonstrated by lariat RNAs that y-shaped RNAs lack[94].

RNAse R was conducted using 20 units of RNAse R in a 50 μl reaction and 1x RNAse R reaction buffer. The reaction is cleaned by ethanol precipitation. In the event that the next step is periodate oxidation, then a phenol:chloroform extraction is needed prior to the ethanol precipitation.

Periodate oxidation

Periodate oxidation is performed either to completion using sodium acetate pH 5.4 or by using another pH and buffer. Potassium phosphate dibasic is one option, though this is not a buffer.

Sodium acetate preparation:

Set up a 50 μl reaction with up to 146 pmol of RNA termini in 100 mM NaOAc pH 5.4 and 100 μM $NaIO_4$. Incubate at room temperature for 2 hours.

potassium phosphate dibasic preparation:

Set up a 50 μl reaction with up to 146 pmol of RNA termini in 100 mM K2HPO_4 and 100 μM $NaIO_4$. Incubate on ice for 20 minutes.

After either reaction, the reaction is quenched by adding 4 μl of 40% glucose and it is cleaned by ethanol precipitation. Because the potassium phosphate dibasic precipitates, a buffer exchange column (GE g-25 spin column) should be used when performing a potassium phosphate reaction.

Hydrazide biotinylation

A hydrazide biotin is chemically conjugated to the dialdehydes generated by periodate oxidation.

The reaction is performed in 1.25 mM biotin-PEG4-hydrazide in 50 mM NaOAc PH 5.3 and incubate at 4°C overnight.

This reaction may be cleaned using ethanol precipitation. A buffer exchange

column is an optional additional step that may aid in the further removal of free biotin.

RA5 adapter ligation

The final reaction is set up in 15 μl of 1x RNA ligase I buffer, 1 mM ATP, 400 pmol of RA5 oligonucleotide (5′ adapter should be heated for 2 minutes at 65°C and immdiately stored on ice), and 1 μl of RNA ligase I enzyme. Incubate at 28°C for 1 hr. The reaction is cleaned by ethanol precipitation.

IDTdna code for RA5 oligonucleotide: rGrUrUrCrArGrArGrUrUrCrUrArCrArGrUrCrCrGrArCrGrArUrC

General protocol for streptavidin binding and selection

Buffers:

Binding Buffer

Tris-Cl pH 7.4	10 mM
EDTA	1 mM
NaCl	300 mM
Triton X-100	0.1%

Low Salt Wash Buffer

1 M Tris-Cl	5 mM
EDTA	1 mM
Triton X-100	0.1%

High Salt Wash Buffer

1 M Tris-Cl pH 7.4	50 mM
EDTA	1 mM
Triton X-100	0.5%
NaCl	2 M

Formamide + EDTA Elution Buffer

Formamide	95%
EDTA pH 8.4	20 mM

Protocol:

The RNA is selected for using streptavidin coated magnetic beads. This selection is repeated twice.

0. Prewash beads using 55 μl of the invitrogen dynabeads C1 streptavidin beads. Wash twice with 100 mM NaCl. Incubate with a NaOH solution for 2 minutes. Wash once with 100 mM NaCl. Resuspend with 500 μl of Streptavidin binding buffer.

1. Add an equal volume (up to 100 μl) + 10 μl of formamide loading buffer (about 95% formamide + 25 mM EDTA)

2. Heat at 95°C for 2 minutes and move to ice. (This step leads to the loss of some of the biotinylation, likely due to the degradation of the hydrazone linkage and it would be appropriate to rethink what temperature to use)

3. Add sample to beads that have been pre-washed and are in 700-800 μl of biotin binding buffer.

4. Vortex briefly and rotate samples at room temperature for 15 minutes.

5. Collect beads on stand for 1 minute and remove supernatant.

6. Wash in 900 μl of biotin binding buffer. Collect beads on stand for 1 minute and remove supernatant.

7. Wash in 900 μl of high salt buffer. Collect beads on stand for 1 minute and remove supernatant.

8. Wash in 900 μl of low salt buffer. Collect beads on stand for 1 minute and remove supernatant.

9. Wash in 900 μl of low salt buffer. Collect beads on stand for 1 minute and remove supernatant.

10. Add 200 μl of formamide loading buffer + 6 μl of linear acrylamide and heat the samples for 65C for 2 minutes.

11. Collect the beads using a magnetic stand for 10 minutes. Gently remove the formamide using a pipette. Place the pippette tip near the magnet while slowly pippetting the formamide solution into a fresh microfuge tube for precipitation.

12. Add 70 μl of 5 M NaCl + 500 μl H_2O + 5 μl of linear acrylamide + 700 μl of cold isopropanol. mix well and incubate at -20°C overnight. Precipitate as normal with two 70% ethanol washes and resuspend in 25 μl H_2O.

Reverse transcription using manganese chloride

Set up RT as follows:

Add 10.6 μl of RNA to tube

Add 1 μl of 10X RT buffer (that does not contain magnesium)

Add 1 μl of 5 μM barcode RT primer OR oligonucleotide dt_{18} primer

Incubate at 65°C for 5 minutes and cool on bench.

For the next step, add the following to the tubes:

1 μl of 10 mM dNTPs

2 μl 100 mM DTT

1 μl 10 X RT buffer (that does not contain magnesium)

2.4 μl of 50 mM $MnCl_2$

1 μl of SSII enzyme (Invitrogen)

Gently mixed the reaction and incubate at 42°C for 1 hour.

Phusion PCR

The reverse transcribed cDNA is used for PCR amplification according to the Illumina Small RNA cloning Protocol using Phusion polymerase.

Library sizing using native PAGE

The amplified library is sized using 10% native polyacrylamide gels. Sized libraries are eluted using 0.3 M NaOAc pH 5.4 and with overnight rotation at room temperature. Eluted libraries are then precipitated using an ethanol precipitation.

Library submission

The precipitated libraries are then quantified for their mass using a Qubit and are diluted to 2 nM in preparation for Illumina single-end sequencing.

BIBLIOGRAPHY

[1] Gil Ast. How did alternative splicing evolve? 5(10):773–782.

[2] Adeeb M. Al-Zoubi, Elena V. Efimova, Shashi Kaithamana, Osvaldo Martinez, Mohammed El-Azami El-Idrissi, Rukiye E. Dogan, and Bellur S. Prabhakar. Contrasting effects of ig20 and its splice isoforms, madd and denn-sv, on tumor necrosis factor -induced apoptosis and activation of caspase-8 and -3. *Journal of Biological Chemistry*, 276(50):47202–47211, 2001.

[3] R Saldanha, G Mohr, M Belfort, and A M Lambowitz. Group i and group ii introns. *The FASEB Journal*, 7(1):15–24, 1993.

[4] Martin Akerman and Yael Mandel-Gutfreund. Alternative splicing regulation at tandem 3 splice sites. *Nucleic Acids Research*, 34(1):23–31, 2006.

[5] Sebastian M. Fica, Nicole Tuttle, Thaddeus Novak, Nan-Sheng Li, Jun Lu, Prakash Koodathingal, Qing Dai, Jonathan P. Staley, and Joseph A. Piccirilli. RNA catalyses nuclear pre-mRNA splicing. 503(7475):229–234.

[6] Manuel Irimia and Scott William Roy. Origin of spliceosomal introns and alternative splicing. *Cold Spring Harbor Perspectives in Biology*, 6(6), 2014.

[7] AM Zahler, WS Lane, JA Stolk, and M B Roth. Sr proteins: a conserved family of pre-mrna splicing factors. *Genes Dev*, 1992.

[8] M Takimoto, T Tomonaga, M Matunis, M Avigan, H Krutzsch, G Dreyfuss, and D Levens. Specific binding of heterogeneous ribonucleoprotein particle protein k to the human c-myc promoter, in vitro. *Journal of Biological Chemistry*, 268(24):18249–18258, 1993.

[9] Rita Das, Jiong Yu, Zuo Zhang, Melanie P. Gygi, Adrian R. Krainer, Steven P. Gygi, and Robin Reed. {SR} proteins function in coupling {RNAP} {II} transcription to pre-mRNA splicing. 26(6):867 – 881.

[10] Jennifer C. Long and Javier F. Caceres. The SR protein family of splicing factors: master regulators of gene expression. 417(1):15–27.

[11] Mireya Plass, Eneritz Agirre, Diana Reyes, Francisco Camara, and Eduardo Eyras. Co-evolution of the branch site and SR proteins in eukaryotes. 24(12):590–594.

[12] Arvydas Kanopka, Oliver Muhlemann, and Goran Akusjarvi. Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. 381(6582):535–538.

[13] Chaolin Zhang, Maria A. Frias, Aldo Mele, Matteo Ruggiu, Taesun Eom, Christina B. Marney, Huidong Wang, Donny D. Licatalosi, John J. Fak, and Robert B. Darnell. Integrative modeling defines the nova splicing-regulatory network and its combinatorial controls. 329(5990):439–443.

[14] R Martinez-Contreras, P Cloutier, L Shkreta, JF Fisette, T Revil, and B Chabot. hnrnp proteins and splicing control. *Adv Exp Med Biol*, 2007.

[15] Todd Bradley, Malcolm E. Cook, and Marco Blanchette. Sr proteins control a complex network of rna-processing events. *RNA*, 21(1):75–92, 2015.

[16] Brenton R. Graveley. Alternative splicing: increasing diversity in the proteomic world. 17(2):100–107.

[17] Benjamin P. Lewis, Richard E. Green, and Steven E. Brenner. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mrna decay in humans. *Proceedings of the National Academy of Sciences*, 100(1):189–192, 2003.

[18] Yamile Marquez, Markus Hpfler, Zahra Ayatollahi, Andrea Barta, and Maria Kalyna. Unmasking alternative splicing inside protein-coding exons defines exitrons and their role in proteome plasticity. *Genome Research*, 2015.

[19] Hadas Keren, Galit Lev-Maor, and Gil Ast. Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics*, 2010.

[20] Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 2008.

[21] M Irimia, J L Rukov, D Penny, and S W Roy. Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing. *BMC Evol Biol*, 2007.

[22] Paula Cramer, C. Gustavo Pesce, Francisco E. Baralle, and Alberto R. Ko-

rnblihtt. Functional association between promoter structure and transcript alternative splicing. *PNAS*, 1997.

[23] Xiang-Yang Zhong, Jian-Hua Ding, Joseph A. Adams, Gourisankar Ghosh, and Xiang-Dong Fu. Regulation of sr protein phosphorylation and alternative splicing by modulating kinetic interactions of srpk1 with molecular chaperones. *Genes and Development*, 2009.

[24] Kristian E Baker and Roy Parker. Nonsense-mediated mrna decay: terminating erroneous gene expression. *Current Opinion in Cell Biology*, 2004.

[25] Phillip A. Sharp. The discovery of split genes and RNA splicing. 30(6):279–281.

[26] Mario Stanke and Stephan Waack. Gene prediction with a hidden markov model and a new intron submodel. *Bioinformatics*, 2003.

[27] Serafim Batzoglou, Lior Pachter, Jill P. Mesirov, Bonnie Berger, and Eric S. Lander. Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Research*, 10(7):950–958, 2000.

[28] Charles W Sugnet, Karpagam Srinivasan, Tyson A Clark, Georgeann O'Brien, Melissa S Cline, Hui Wang, Alan Williams, David Kulp, John E Blume, David Haussler, and Manuel Ares, Jr. Unusual intron conservation near tissue-regulated exons found by splicing microarrays. 2(1):1–14.

[29] Jonathan E. Allen and Steven L. Salzberg. A phylogenetic generalized hidden markov model for predicting alternatively spliced exons. 1(1):1–13.

[30] Michael J. Moore and Pamela A. Silver. Global analysis of mrna splicing. *RNA*, 14(2):197–203, 2008.

[31] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-seq: a revolutionary tool for transcriptomics. 10(1):57–63.

[32] Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. 17(6):333–351.

[33] A.J. Taggart, A.M. DeSimone, J.S. Shih, M.E. Filloux, and W.G Fairbrother. Large-scale mapping of branchpoints in human pre-mrna transcripts in vivo. *Nat. Struct. Mol. Biol.*, 2012.

[34] A.R. Awan, A. Manfredo, and J.A. Pleiss. Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. *Proc. Natl. Acad. Sci. U.S.A.*, 2013.

[35] Nicholas Stepankiw, Madhura Raghavan, Elizabeth A. Fogarty, Andrew Grimson, and Jeffrey A. Pleiss. Widespread alternative and aberrant splicing revealed by lariat sequencing. *Nucleic Acids Research*, 2015.

[36] Danny A Bitton, Charalampos Rallis, Daniel C Jeffares, Graeme C Smith, Yuan YC Chen, Sandra Codlin, Samuel Marguerat, and Jurg Bahler. Lasso, a strategy for genome-wide mapping of intronic lariats and branch-points using rna-seq. *Genome Research*, 2014.

[37] Clare Gooding, Francis Clark, Matthew C Wollerton, Sushma-Nagaraja Grellscheid, Harriet Groom, and Christopher WJ Smith. A class of human exons with predicted distant branch points revealed by analysis of ag dinucleotide exclusion zones. *Genome Biol*, 2006.

[38] Adam Volanakis, Monica Passoni, Ralph D. Hector, Sneha Shah, Cornelia Kilchert, Sander Granneman, and Lidia Vasiljeva. Spliceosome-mediated decay (smd) regulates expression of nonintronic genes in budding yeast. *Genes & Development*, 27(18):2025–2038, 2013.

[39] Maria Armakola, Matthew J. Higgins, Matthew D. Figley, Sami J. Barmada, Emily A. Scarborough, Zamia Diaz, Xiaodong Fang, James Shorter, Nevan J. Krogan, Steven Finkbeiner, Jr. Robert V. Farese, and Aaron D. Gitler. Inhibition of rna lariat debranching enzyme suppresses tdp-43 toxicity in als disease models. *Nat Genet*, 2012.

[40] Jrg Vogel, Wolfgang R. Hess, and Thomas Brner. Precise branch point mapping and quantification of splicing intermediates. *NAR*, 1997.

[41] David P. Lorsch, John R. Bartel and Jack W. Szostak. Reverse transcriptase reads through a 2'5'linkage and a 2-thiphosphate in a template. *NAR*, 1995.

[42] Kaiping Gao, Akio Masuda, Tohru Matsuura, and Kinji Ohno. Human branch point consensus sequence is yunay. *NAR*, 2008.

[43] Jamie F. Conklin, Aaron Goldman, and A. Javier Lopez. Stabilization and analysis of intron lariats in vivo. *Methods*, 37(4):368 – 375, 2005. Post-transcriptional Regulation of Gene Expression.

[44] Stephen M. Garrey, Adam Katolik, Mantas Prekeris, Xueni Li, Kerri York, Sarah Bernards, Stanley Fields, Rui Zhao, Masad J. Damha, and Jay R. Hesselberth. A homolog of lariat-debranching enzyme modulates turnover of branched rna. *RNA*, 20(8):1337–1348, 2014.

[45] Tim R. Mercer, Michael B. Clark, Stacey B. Andersen, Marion E. Brunck, Wilfried Haerty, Joanna Crawford, Ryan J. Taft, Lars K. Nielsen, Marcel E. Dinger, and John S. Mattick. Genome-wide discovery of human splicing branchpoints. *Genome Research*, 2015.

[46] Daoming Qin, Lei Huang, Alissa Wlodaver, Jorge Andrade, and Jonathan P. Staley. Sequencing of lariat termini in s. cerevisiae reveals 5 splice sites, branch points, and novel splicing events. *RNA*, 22(2):237–253, 2016.

[47] Manuel de la Mata, Claudio R Alonso, Sebastin Kadener, Juan P Fededa, Matas Blaustein, Federico Pelisch, Paula Cramer, David Bentley, and Alberto R Kornblihtt. A slow {RNA} polymerase {II} affects alternative splicing in vivo. *Molecular Cell*, 12(2):525 – 532, 2003.

[48] JF Caceres, S Stamm, DM Helfman, and AR Krainer. Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors. *Science*, 265(5179):1706–1709, 1994.

[49] K.L. Fox-Walsh and K.J Hertel. Splice-site pairing is an intrinsically high fidelity process. *Proc. Natl. Acad. Sci. U.S.A.*, 2009.

[50] E. Melamud and J Moult. Stochastic noise in splicing machinery. *Nucleic Acids Res*, 2009.

[51] Joseph K. Pickrell, Athma A. Pai, Yoav Gilad, and Jonathan K. Pritchard. Noisy splicing drives mrna isoform diversity in human cells. *PNAS*, 2013.

[52] N. Rhind, Z. Chen, M. Yassour, D.A. Thompson, B.J. Haas, N. Habib, I. Wapinski, S. Roy, M.F. Lin, D.I. Heiman, and et al. Comparative functional genomics of the fission yeasts. *Science*, 2011.

[53] Danny A Bitton, Sophie R Atkinson, Charalampos Rallis, Graeme C Smith, David A Ellis, Yuan Chen, Michal Malecki, Sandra Codlin, Cristina Cotobal, Jean-Francois Lemay, Francois Bachand, Samuel Marguerat, Juan Mata, and Jurg Bahler. Widespread exon-skipping triggers degradation by nuclear rna surveillance in fission yeast. *Genome Research*, 2015.

[54] Yeon Lee and Donald C. Rio. Mechanisms and regulation of alternative pre-mrna splicing. *Annual Review of Biochemistry*, 84(1):291–323, 2015. PMID: 25784052.

[55] M. Moore, C. Query, and P Sharp. The rna world. vol. 1. *NY: Cold Spring Harbor Laboratory Press*, 1993.

[56] Horst Domdey, Barbara Apostol, Ren-Jang Lin, Andrew Newman, Edward Brody, and John Abelson. Lariat structures are in vivo intermediates in yeast pre-mRNA splicing. 39(3):611–621.

[57] Timothy W. Nilsen and Brenton R. Graveley. Expansion of the eukaryotic proteome by alternative splicing. 463(7280):457–463.

[58] Mo Chen and James L. Manley. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. 10(11):741–754.

[59] Lee P. Lim and Christopher B. Burge. A computational analysis of sequence features involved in recognition of short introns. *Proceedings of the National Academy of Sciences*, 98(20):11193–11198, 2001.

[60] Fursham M. Hamid and Eugene V. Makeyev. Emerging functions of alternative splicing coupled with nonsense-mediated decay. 42(4):1168–1173.

[61] J.K. Pickrell, A.A. Pai, Y. Gilad, and J.K Pritchard. Noisy splicing drives mrna isoform diversity in human cells. *PLoS Genet.*, 2010.

[62] V. Wood, R. Gwilliam, M.-A. Rajandream, M. Lyne, R. Lyne, A. Stewart, J. Sgouros, N. Peat, J. Hayles, S. Baker, and et al. The genome sequence of schizosaccharomyces pombe. *Nature*, 2001.

[63] A.N. Kuhn and N.F. Kufer. Pre-mrna splicing in schizosaccharomyces pombe. *Curr. Genet.*, 2002.

[64] W. Gilbert. Why genes in pieces? *Nature*, 1978.

[65] S.L. Forsburg and N. Rhind. Basic methods for fission yeast. *Yeast Chichester Engl.*, 2006.

[66] A.M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinforma. Oxf. Engl.*, 2014.

[67] B. Langmead and S.L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 2012.

[68] Gavin E. Crooks, Gary Hon, John-Marc Chandonia, and Steven E. Brenner. Weblogo: A sequence logo generator. *Genome Research*, 14(6):1188–1190, 2004.

[69] Valerie Wood, Midori A. Harris, Mark D. McDowall, Kim Rutherford, Brendan W. Vaughan, Daniel M. Staines, Martin Aslett, Antonia Lock, Jrg Bhler, Paul J. Kersey, and Stephen G. Oliver. Pombase: a comprehensive online resource for fission yeast. *Nucleic Acids Research*, 40(D1):D695–D699, 2012.

[70] M. Aebi and C. Weissman. Precision and orderliness in splicing. *Trends in Genetics*, 3:102 – 107.

[71] Christopher J. Webb, Charles M. Romfo, Willem J. van Heeckeren, and Jo Ann Wise. Exonic splicing enhancers in fission yeast: functional conservation demonstrates an early evolutionary origin. 19(2):242–254, 2005.

[72] Jorge Prez-Valle and Josep Vilardell. Intronic features that determine the selection of the 3 splice site. *Wiley Interdisciplinary Reviews: RNA*, 3(5):707–717, 2012.

[73] Charles M. Romfo, Consuelo J. Alvarez, Willem J. van Heeckeren, Christopher J. Webb, and Jo Ann Wise. Evidence for splice site pairing via intron definition in schizosaccharomyces pombe. *Molecular and Cellular Biology*, 20(21):7955–7970, 2000.

[74] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L. Salzberg. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):1–13, 2013.

[75] Julian P. Venables, Roscoe Klinck, Anne Bramard, Lyna Inkel, Geneviève Dufresne-Martin, ChuShin Koh, Julien Gervais-Bird, Elvy Lapointe, Ulrike Froehlich, Mathieu Durand, Daniel Gendron, Jean-Philippe Brosseau, Philippe Thibault, Jean-Francois Lucier, Karine Tremblay, Panagiotis Prinos, Raymund J. Wellinger, Benoit Chabot, Claudine Rancourt, and Sherif Abou Elela. Identification of alternative splicing markers for breast cancer. *Cancer Research*, 68(22):9525–9531, 2008.

[76] Thomas A. Cooper, Lili Wan, and Gideon Dreyfuss. {RNA} and disease. *Cell*, 136(4):777 – 793, 2009.

[77] Hui Y. Xiong, Babak Alipanahi, Leo J. Lee, Hannes Bretschneider, Daniele Merico, Ryan K. C. Yuen, Yimin Hua, Serge Gueroussov, Hamed S. Najafabadi, Timothy R. Hughes, Quaid Morris, Yoseph Barash, Adrian R. Krainer, Nebojsa Jojic, Stephen W. Scherer, Benjamin J. Blencowe, and Brendan J. Frey. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347(6218), 2015.

[78] Zefeng Wang and Christopher B. Burge. Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA*, 14(5):802–813, 2008.

[79] Eddo Kim, Alon Magen, and Gil Ast. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research*, 35(1):125–131, 2007.

[80] Laura De Conti, Marco Baralle, and Emanuele Buratti. Exon and intron definition in pre-mrna splicing. *Wiley Interdisciplinary Reviews: RNA*, 4(1):49–60, 2013.

[81] Wei Shao, Hyun-Soo Kim, Yang Cao, Yong-Zhen Xu, and Charles C. Query. A u1-u2 snrnp interaction network during intron definition. *Molecular and Cellular Biology*, 32(2):470–478, 2012.

[82] Kevin Roy and Guillaume Chanfreau. Stress-induced nuclear rna degradation pathways regulate yeast bromodomain factor 2 to promote cell survival. *PLoS Genet*, 10(9):1–15, 09 2014.

[83] Tadashi Kawashima, Stephen Douglass, Jason Gabunilas, Matteo Pellegrini, and Guillaume F. Chanfreau. Widespread use of non-productive alternative splice sites in ¡italic¿saccharomyces cerevisiae¡/italic¿. *PLoS Genet*, 10(4):1–15, 04 2014.

[84] Jessica A. Box, Jeremy T. Bunch, Wen Tang, and Peter Baumann. Spliceosomal cleavage generates the 3[prime] end of telomerase RNA. 456(7224):910–914.

[85] Guillaume Chanfreau. Conservation of rnase iii processing pathways and specificity in hemiascomycetes. *Eukaryotic Cell*, 2(5):901–909, 2003.

[86] Jules Gagnon, Mathieu Lavoie, Mathieu Catala, Francis Malenfant, and

Sherif Abou Elela. Transcriptome wide annotation of eukaryotic rnase iii reactivity and degradation signals. *PLoS Genet*, 11(2):1–29, 02 2015.

[87] Jason Merkin, Caitlin Russell, Ping Chen, and Christopher B. Burge. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, 338(6114):1593–1599, 2012.

[88] Nuno L. Barbosa-Morais, Manuel Irimia, Qun Pan, Hui Y. Xiong, Serge Gueroussov, Leo J. Lee, Valentina Slobodeniuc, Claudia Kutter, Stephen Watt, Recep Çolak, TaeHyung Kim, Christine M. Misquitta-Ali, Michael D. Wilson, Philip M. Kim, Duncan T. Odom, Brendan J. Frey, and Benjamin J. Blencowe. The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114):1587–1593, 2012.

[89] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. 40(12):1413–1415.

[90] Hitoshi Suzuki, Yuhong Zuo, Jinhua Wang, Michael Q. Zhang, Arun Malhotra, and Akila Mayeda. Characterization of rnase r-digested cellular rna source that consists of lariat and circular rnas from pre-mrna splicing. *NAR*, 2006.

[91] Fanglei Zhuang, Ryan T. Fuchs, Zhiyi Sun, Yu Zheng, and G. Brett Robb. Structural bias in t4 rna ligase-mediated 3-adapter ligation. *Nucleic Acids Research*, 2012.

[92] J. R. Clamp and L Hough. The periodate oxidation of amino acids with reference to studies on glycoproteins. *Biochem J*, 1965.

[93] Gregory L Moore and Costas D Maranas. Modeling dna mutation and recombination for directed evolution experiments. *Journal of Theoretical Biology*, 205(3):483 – 503, 2000.

[94] Hitoshi Suzuki, Yuhong Zuo, Jinhua Wang, Michael Q. Zhang, Arun Malhotra, and Akila Mayeda. Characterization of rnase r-digested cellular rna source that consists of lariat and circular rnas from pre-mrna splicing. *Nucleic Acids Research*, 34(8):e63, 2006.