High Resolution Global Analyses of the Molecular Mechanisms of pre-mRNA Splicing Regulation

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

Zachary W. Dwyer

May 2021

High Resolution Global Analyses of the Molecular Mechanisms of pre-mRNA Splicing Regulation

Zachary W. Dwyer, Ph.D.

Cornell University 2021

Pre-mRNA splicing is an essential component of eukaryotic gene expression, and complex patterns of alternative splicing in higher eukaryotes significantly enhance proteome diversity. Mutations in the splicing pathway have been increasingly identified as drivers of human disease, yet the mechanistic consequences of these mutations remain poorly understood, inhibiting our understanding of these diseases. The fission yeast *Schizosaccharomyces pombe* offers an outstanding model for understanding complex splicing patterns seen in humans: its genome is rich with introns with highly degenerate splice site sequences, closely resembling those seen in higher eukaryotes, while its facile genetics enable interrogations of the highly conserved splicing machinery. A broad focus of my work has been understanding the mechanisms by which *bona fide* examples of regulated alternative splicing of specific *S. pombe* transcripts occurs.

While Next Generation Sequencing technologies have had a transformative impact on studies of pre-mRNA splicing, currently a major obstacle in the field which remains poorly appreciated to this day, regards the quantitative limitations associated with these approaches, particularly in detecting and monitoring rare RNA species. To alleviate this problem, we developed a targeted RNA sequencing method termed Multiplexed Primer Extension sequencing (MPE-seq) which enriches for splicing informative reads, allowing for improved quantitative assessments of splicing isoforms,

including the intermediates generated during the splicing reaction. I have leveraged this approach to interrogate different aspects of splice site recognition, with a particular focus on alternative splicing. Here I describe the results of a high-throughput forward genetic screen of thousands of temperature-sensitive *S. pombe* strains designed to identify those with defects in canonical and/or alternative pre-mRNA splicing. We identified scores of alleles, some causing global defects of canonical splicing, while others lead to specific changes in alternative splicing of select transcripts. Among others, whole genome sequencing of candidate strains revealed a pair of mutations in *prp10*, the *S. pombe* ortholog to the human protein SF3B1, one of the most commonly mutated genes in myelodysplastic syndromes. Remarkably, while these two variants lie close in three-dimensional space to one another, using MPE-seq I demonstrate the markedly different impacts of these mutations on pre-mRNA splicing patterns genome-wide. These studies provide important insights into how the spliceosome activates its cognate targets, and how disease-related mutations may mis-regulate this process.

BIOGRAPHICAL SKETCH

Zach was born in Lansing, NY and graduated from Lansing High School where he developed a passion for both Biology and Computer Science. As an undergraduate, he continued to pursue both fields while majoring in Molecular and Cellular Biology and minored in Computer Science at Johns Hopkins University in Baltimore, Maryland. He got his first taste for research one summer break while studying how *Salmonella* respond to the environmental cues of the gut to induce virulence in the lab of Craig Altier at Cornell University. Throughout college, he continued to pursue research in the labs of Young-Sam Lee (which focused on identifying metabolic signaling pathways that contribute to human disease) and Sarah Wheelan (which focused on developing techniques for analysis and biological interpretation of sequencing data). After starting graduate school at Cornell University, he joined the lab of Jeff Pleiss where he learned to combine molecular biology, genetics, and computational approaches to dissect the regulatory codes that underpin pre-mRNA splicing. Upon receiving his Ph.D., Zach looks forward to continuing to focus on pre-mRNA splicing as he starts a career as a Bioinformatic Scientist at Skyhawk Therapeutics an RNA therapeutics company in Boston, Massachusetts.

ACKNOWLEGEMENTS

First and foremost, I would like to thank my advisor Dr. Jeffrey Pleiss for his unending support and guidance as I developed in my scientific career. He provided an environment that was not only academically rewarding but was also socially enjoyable. He allowed for independent scientific development, but his door was always open when I needed help.

Second, I would like to thank the rest of the Pleiss lab whose ideas and recommendations can be found throughout my work: Sara Downs, Ben Fair, Mike Gildea, Bec Grace, Maki Inada, Amy Larson, Kelly Murray, Madhura Raghavan, Rachel Sandman, Nick Stepankiw and Hansen Xu.

Finally, I would like to thank my friends and family for making all of this possible. Notably: Greg Booth, Ian Rose, Megan Rothstein, Florencia Schlamp, Roman Spektor, Jessica West, and Dan Zinshteyn

TABLE OF CONTENTS

Chapter I: Multiplexed primer extension sequencing - A targeted RNA-seq method that enables high-precision quantitation of mRNA splicing isoforms and rare pre-mRNA splicing intermediates

Alternative citation:

Gildea MA*, Dwyer ZW*, Pleiss JA. (2020). Multiplexed primer extension sequencing: A targeted RNA-seq method that enables high-precision quantitation of mRNA splicing isoforms and rare pre-mRNA splicing intermediates. *Methods*. 176,34-45. https://doi.org/10.1016/j.ymeth.2019.05.013
* Denotes equal contribution

Abstract

The study of pre-mRNA splicing has been greatly aided by the advent of RNA sequencing (RNA-seq), which enables the genome-wide detection of discrete splice isoforms. Quantification of these splice isoforms requires analysis of splicing informative sequencing reads, those that unambiguously map to a single splice isoform, including exon-intron spanning alignments corresponding to retained introns, as well as exon-exon junction spanning alignments corresponding to either canonically- or alternatively-spliced isoforms. Because most RNA-seq experiments are designed to produce sequencing alignments that uniformly cover the entirety of transcripts, only a comparatively small number of splicing informative alignments are generated for any given splice site, leading to a decreased ability to detect and/or robustly quantify many splice isoforms. To address this problem, we have recently described a method termed Multiplexed Primer Extension sequencing, or MPE-seq, which uses pools of reverse transcription primers to target sequencing to user selected loci. By targeting reverse transcription to pre-mRNA splice junctions, this approach enables a dramatic enrichment in the fraction of splicing informative alignments generated per splicing

event, yielding an increase in both the precision with which splicing efficiency can be measured, and in the detection of splice isoforms including rare splicing intermediates. Here we provide a brief review of the shortcomings associated with RNA-seq that drove our development of MPE-seq, as well as a detailed protocol for implementation of MPE-seq.

Introduction

The past decade has witnessed a growing appreciation for the role that pre-mRNA splicing plays in regulating eukaryotic gene expression. Contradicting the original concept that 'one gene equals one protein', it is now clear that alternative splicing plays a critical role in expanding the proteomes of eukaryotic organisms, enabling production of wide varieties of protein isoforms from single genetic loci (Nilsen & Graveley, 2010). Indeed, ever-increasing numbers of alternatively spliced variants are being identified, often associated with changing developmental or cellular states (E. T. Wang et al., 2008), (Pan et al., 2008). Consistent with its widespread role in gene regulation, the last several years have similarly seen an explosion of connections between human disease and mutations in the splicing pathway (David & Manley, 2010; Faustino, 2003; Padgett, 2012; Tazi et al., 2009; Zhang & Manley, 2013). Included among these are mutations in the splice site sequences and regulatory cis-elements of individual transcripts that influence their expression, as well as mutations in components of the core machinery that catalyzes pre-mRNA splicing, which potentially impact the splicing of many transcripts (Scotti & Swanson, 2016). Nevertheless, despite its significance, our field currently lacks answers to many critically important questions about pre-mRNA splicing,

ranging from aspects of its basic mechanisms to understanding the mechanistic and physiological implications of disease-related mutations.

The advent of short-read sequencing technologies has dramatically impacted our understanding of the role of pre-mRNA splicing in eukaryotic biology (Z. Wang et al., 2009). By identifying the subset of alignments that traverse a splicing boundary, RNA sequencing (RNA-seq) has been used to identify countless novel splice variants, expanding our understanding of the expressed proteome and the role that splicing plays in generating this diversity (Nilsen & Graveley, 2010). Yet while this approach has been unquestionably successful in identification of novel transcripts and splice isoforms, most standard experiments lack sufficient power to quantitatively assess changes in many splicing events. Because of the uniform alignment coverage across transcripts that traditional RNA-seq generates, only a small number of splicing informative alignments – those which cross splice junctions or align within an intron – are generated for any given splicing event. This low sampling results in poor precision in quantification of many splice isoforms, confounding studies aimed at identifying the subsets of splicing events that are impacted by a particular treatment. We have recently overcome this deficiency by developing an approach we termed Multiplexed Primer Extension sequencing (MPE-seq) that targets reads to user selected splice junctions, greatly increasing alignment coverage at selected loci (Xu et al., 2019). Here we present a brief review of the challenges associated with quantifying splice isoform abundance from traditional RNA-seq approaches as well as provide a detailed protocol for MPE-seq.

Results

Splicing Status is Often Poorly Sampled by RNA-seq Experiments

Over the past decade, RNA-seq has become the tool of choice for genome-wide analyses of the transcriptome and has led to the identification of scores of previously unidentified transcripts and splice isoforms in many organisms (Xu et al., 2019). The identification of sequencing alignments that unambiguously support different splice isoforms or pre-mRNA intermediates can be used to monitor changes in splicing status under changing conditions (Fig.1A). Yet most standard RNA-seq experiments, having historically been optimized to monitor transcript levels, undersample many if not most splicing events. Importantly, alignments that unambiguously distinguish between mature and premature mRNAs, or between canonically and alternatively spliced isoforms, are present at only a small fraction of the total alignments per transcript (Fig. 1B). As such, the coefficient of variation associated with counting these splicing informative alignments can be very high, particularly for lowly expressed transcripts (Fig. 1C).

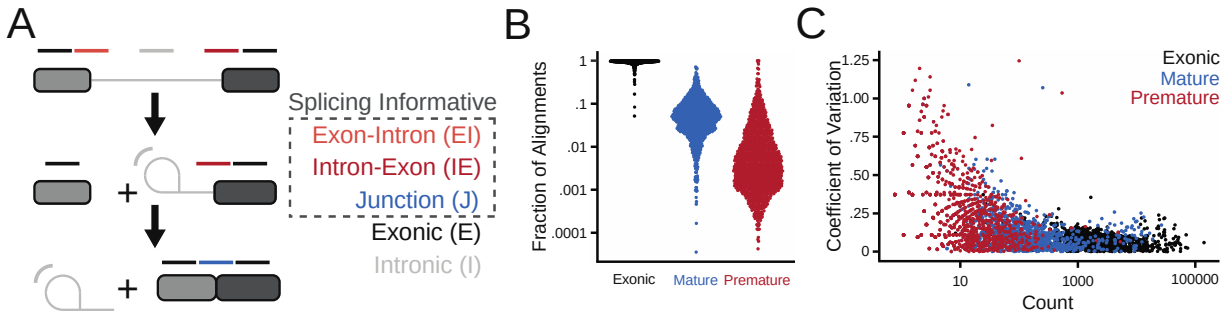


Figure 1: Quantitation of splicing isoforms from RNA-seq data (**A**) Only a subset of alignments are splicing informative: those which can be unambiguously mapped to premature (including completely unspliced and lariat intermediate) or mature (fully spliced) molecules. Black arrows denote the conversion between these forms in the 1st and 2nd chemical steps of the pre-mRNA splicing reaction. (**B**) The fraction of alignments from a standard RNA-seq experiment that are completely exonic (E), mature (J), and premature (EI + IE) is shown for all single intron containing genes in *S. pombe*. (**C**) The coefficient of variation of replicate measurements of each event in (B) is plotted as a function of the (geometric) mean count depth for the isoform. RNA-seq data are from Xu, et al. (2019).

The high, but non-uniform variance associated with most splicing informative alignments complicates statistical analyses of their abundance (Fig. 1C). Data sets such as these are characterized by high false discovery rates: false positive discoveries

derived from high (and unbalanced) variance between samples; and false negatives derived from events that are too poorly sampled to properly reveal the underlying difference between samples. Whereas high confidence events that are changed between two samples can often be identified from these data, these events are much more likely to be drawn from highly expressed transcripts because of the lower variance associated with their measurements, introducing selection biases which can dramatically influence the results of the study (Oshlack & Wakefield, 2009). Moreover, many investigators seek to compare 'affected' versus 'unaffected' pools of events to better understand the properties of the affected events (e.g. the presence of cis-regulatory sequences, the gene ontology of the parent transcript, etc.). But such an analysis requires high confidence knowledge of both categories of events, and far too often in this field the absence of evidence of a change in isoform levels from an RNA-seq experiment is incorrectly inferred to mean that there is evidence of absence of a change. While statistical approaches have been published that work to reduce the challenges associated with this problem in isoform detection (Katz et al., 2010; Li & Dewey, 2011; Patro et al., 2017; Pertea et al., 2016), none of these approaches can fully solve the problems associated with low read-count events. For these reasons, we set out to develop an approach that would increase the sampling of these otherwise poorly sampled regions of the transcriptome.

MPE-seq Method Overview

We recently described MPE-seq as a targeted RNA sequencing method based on primer extension that focuses sequencing reads at up to thousands of user selected loci (Xu et al., 2019). By targeting reads to splice junctions (Fig. 2A), MPE-seq greatly

increases the number of splicing informative alignments generated at a given site for a given experiment, allowing for the detection of rare splicing isoforms including pre-mRNA intermediates (Fig. 2B), and increasing the precision with which splicing efficiency can be measured (Fig. 2C). While we have recently applied this method to the budding yeast *S. cerevisiae* and the fission yeast S. pombe, yielding increases in the accuracy and precision of splice isoform measurements genome-wide, we anticipate that it will be readily translatable to other organisms, and here include suggestions for steps that might benefit from modification.

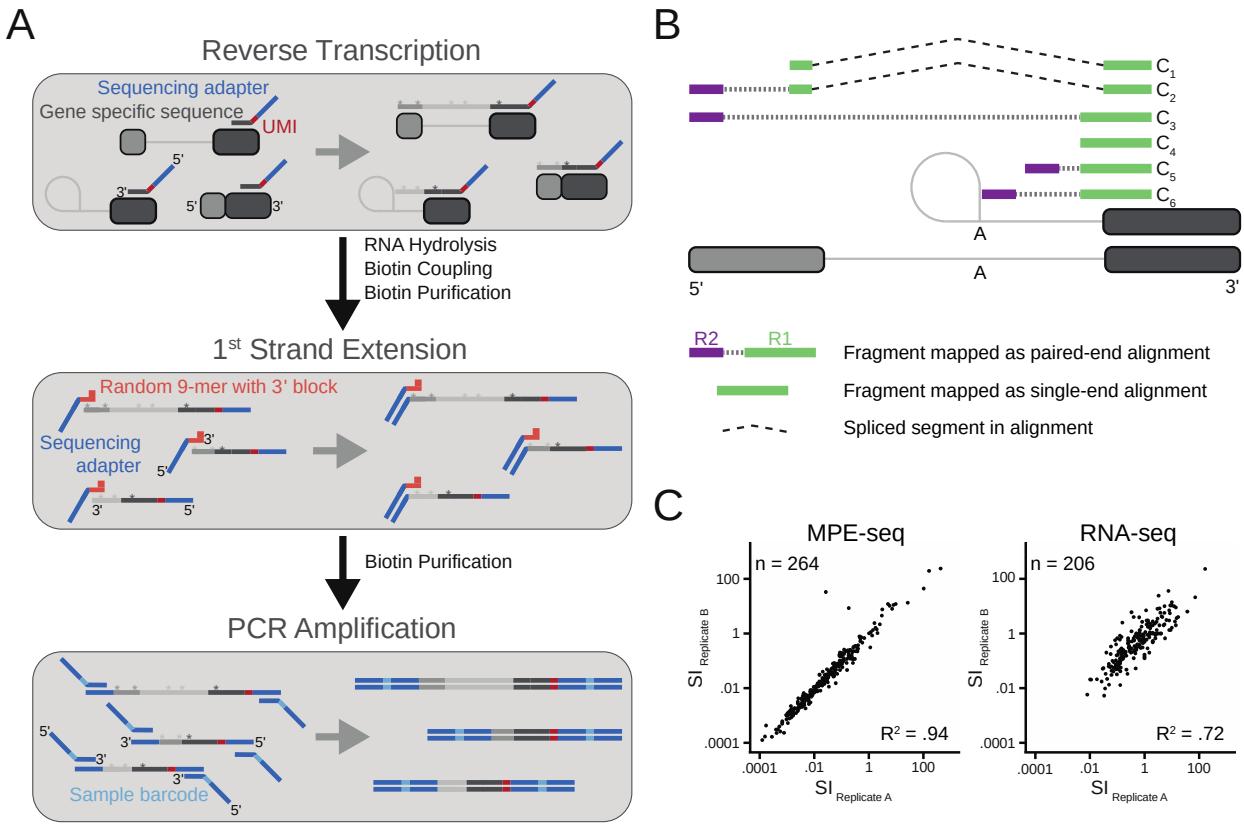


Figure 2: MPE-seq targets reads to splicing informative loci, increasing the precision of measurements of splicing efficiency compared to standard RNA-seq. (**A**) Steps in MPE-seq library preparation. Asterisk indicates the presence of aminoallyl-deoxyuridine, incorporated during reverse transcription. (**B**) Categories of MPE-seq alignments for quantifying the abundance of molecules at each chemical step of splicing for a single splicing event. (**C**) The correlation between measurements of splice index (SI, calculated as total unspliced reads divided by total spliced reads) for each splicing event is presented for replicate MPE-seq and RNA-seq *S. cerevisiae* libraries. Libraries were down sampled to five million reads each. The number of splicing events for which at least one premature and mature alignment exists in both replicates is represented by 'n'. R^2 values were calculated using linear regression.

Three key features distinguish MPE-seq from other approaches and enable the enrichment of sequencing alignments at splice junctions. First, reads are targeted to regions of interest by designing unique reverse transcription (RT) primers to each desired location (Fig. 2A). To effectively monitor splicing efficiency, we designed primers to regions just downstream of each targeted intron such that the cDNA resulting from reverse transcription would contain information about the corresponding transcript's splicing status: those cDNA molecules that cross the upstream intron-exon junction are derived from unspliced mRNA, while those that cross an upstream exon-exon junction are derived from either a canonically or alternatively spliced mRNA. In addition to the gene specific sequence, each reverse transcription primer includes a portion of a next-generation sequencing adapter and a unique molecular adapter (UMI), the latter allowing for compression of PCR duplicates (Fig. 2A) (Kivioja et al., 2012). Second, a modified nucleotide, aminoallyl-dUTP, is included in the reverse transcription reaction, which allows for purification of extended cDNAs and removal of excess unextended primer molecules (Fig. 2A): NHS-Biotin is coupled to the aminoallyl-deoxyuridine followed by streptavidin bead purification. Finally, addition of a portion of the 2nd sequencing adapter is accomplished via a strand extension step analogous to template switching, wherein the adapter is appended to the 3′ end of the original cDNA molecule (Fig. 2A). Combining this approach with paired end sequencing allows for querying of both the 5′ and 3′ ends of cDNAs, enabling identification of pre-mRNA splicing intermediates.

MPE-seq Assay Design Considerations

<u>RNA input</u>

Whereas traditional RNA-seq approaches require an initial step to enrich for mRNAs, either by positively selecting for poly(A)+ RNA or by negatively selecting against rRNAs, one advantage of MPE-seq is its ability to be used with unfractionated RNA. We have successfully generated high quality MPE-seq libraries using either total, unfractionated RNA as input, or poly(A)+ RNA (Xu et al., 2019). Poly(A)+ RNA was used to reduce the frequency of off-target priming on rRNA, a variable that will depend upon the quality of the primers and the reaction conditions that are used during reverse transcription (see additional discussion below). It should be noted that selection of poly(A)+ RNA will remove nascent RNAs that have yet to receive a poly(A) tail, potentially introducing a bias in the detection and/or quantitation of certain isoforms.

While MPE-seq can be successfully applied using a wide variety of starting quantities of RNA, we have typically used 10 µg of total cellular RNA as our initial input. We have generally started with higher total quantities (50 µg) when using poly(A)+ selection to account for inefficiency in recovery of target RNAs in this process. The appropriate level of starting RNA may also vary based on the number of targets chosen and their relative expression within the organism/cell type of interest.

Fragmentation of RNA input is also encouraged to prevent previously described sequencing artefacts that generate a bias against long DNA molecules. In our initial work in *S. cerevisiae*, we did not fragment the RNA due to the strong positional bias of introns: because the majority of introns in this organism are located near to the transcription start site, nearly all of the cDNAs generated in our experiments were expected to be short. By contrast, for most other organisms where introns are more

equally distributed throughout gene bodies, fragmentation of the RNA prior to reverse transcription should be performed.

Target Selection

MPE-seq provides the unique opportunity to select the specific subset of transcripts for study. In selecting our target locations, we considered several design parameters. First, with increasing number of targets, each target will make up a smaller portion of the overall library. As such, adequate sampling of some targets may require greater sequencing depth. Secondly, the relative expression levels of different targets can have a large impact on the quality of the data generated. If highly and lowly expressed targets are selected in the same experiment, the highly expressed targets are likely to dominate the resulting library, potentially leaving lowly expressed transcripts undersampled. In this instance, it is possible to group targets based on similar relative expression levels, and then generate separate MPE-seq libraries for different expression quanta to ensure adequate sampling of all targets. The budding and fission yeast genomes contain roughly 300 and 5000 annotated introns, respectively, and we found that targeting all junctions in both species yielded adequate sampling of at least 90% of targeted unspliced and spliced isoform. By contrast, because the human genome contains ~200,000 annotated introns, a complete global analysis of each intron via MPE-seq would require read depth approaching standard RNA-seq to yield quantitative data on most introns. In this case binning targets based on expression level or features of interest (e.g. function) would allow for enrichment of splicing information of desired targets.

RT Primer Design & Reverse Transcription

It is important to emphasize that the design of reverse transcription primers is one of the most crucial steps in this method. Poorly designed primers can easily cross-hybridize with undesired RNAs, leading to a decreased fraction of on-target alignments. In designing primers that target a particular region, three primary design constraints should be considered. First, primers should be designed within short windows immediately downstream of the exon boundary of interest such that short-read sequencing is sufficient to cross the exon-exon or intron-exon boundary of interest. In our work, we have designed primers between 24 and 26 nucleotides in length that are complementary to regions within a 50 nucleotide window downstream of an intron-exon boundary, enabling analysis with 75 nucleotide reads using an Illumina NextSeq sequencer. The locations and lengths of these primer sequences can of course be changed to better address different questions. Second, the melting temperatures of the target regions of the primers should be as homogenous as possible, and as high as possible given the properties of the reverse transcriptase being used. In our experience, the use of thermostable enzymes such as Superscript IV in conjunction with the highest reaction temperature that retained efficient on-target annealing yielded the highest number of on-target alignments with the lowest abundance of off-target alignments in the sequencing data. Third, robust bioinformatic approaches should be employed to ensure that primers are as specific to the target as possible. Many tools exist for batch primer design that allow the user to select multiple regions in which to design primers with parameters for melting temperature and specificity. Particular attention should be paid to rRNA sequences as these are large sources for potential non-specific primer hybridization. Oligowiz, a program originally designed for microarray primer design, was

used to design the primers for our initial MPE-seq study, using the basic constraints described above (Wernersson et al., 2007).

We add a note here specifically regarding primer design for groups of paralogous genes. As with all RNA-seq methods, differentiating between paralogous genes with MPE-seq can be difficult or impossible as unambiguous assignment requires alignments that reveal sequence differences between the paralogs. When designing MPE-seq reverse transcription primers to paralogous genes it is more important for differences between the paralogs to be contained within the extension region rather than the primer annealing region to maximize confidence in alignment allocation. Small differences in the primer annealing region could be sensitive to mis-priming during reverse transcription, allowing for extension from the incorrect paralog and obfuscating the assignment of the resulting alignment. By contrast, a single primer that hybridizes to an identical region of paralogs, but whose extension reveals even single nucleotide differences between the paralogs will enable higher quality differentiation between paralogous transcripts. Nevertheless, for paralogous genes that lack such differences but rather have identical sequences surrounding the junctions of interest, the inability to unambiguously determine the parental locus has led us to exclude these from our studies, either by removing targeting primers during experimental design, or by excluding unambiguous assignments during data analysis.

As noted above, reverse transcription is performed using gene specific primers that contain a UMI and a portion of the Nextera i5 sequencing adapter appended onto their 5′ ends (Fig. 2A). The UMI is a stretch of random nucleotides that is later used to identify reads resulting from PCR duplicates (Kivioja et al., 2012). The number of

random nucleotides included on these primers can be increased or decreased based on the user's needs. In selecting this number, it is important to consider the number of expected molecules versus the number of possible unique molecules potentially quantified by the UMI.

Primers for use in MPE-seq can either be synthesized individually and then pooled or synthesized in a pooled setting using a variety of commercial sources. In our original study, primers targeting *S. cerevisiae* splice junctions (309 in total) were synthesized both individually (by IDT) and in pooled format (by LC Sciences), while primers targeting S. pombe splice junctions (~4000 in total) were synthesized only in pooled format. Because the scale of pooled synthesis is small, larger quantities of the primers need to be generated prior to their use. This is made possible by 2 additional sequence features appended onto the 3′ ends of the synthesized primers: (1) a common sequence that enables PCR amplification and (2) a restriction site (SapI) that enables generation of an appropriate 3′ end on the final products (Xu et al., 2019). The complex pool of oligos is first amplified via PCR with a forward primer containing a chemically blocked 5′ end and a reverse primer which is biotinylated at the 5′ end. The double stranded amplicon generated in this reaction is then subjected to digestion with SapI, which generates the mature 3′ end of the desired primers. To generate single stranded DNA, a digestion using Lamda exonuclease is performed. The presence of the 5′ block on the forward primer during amplification precludes degradation of the desired forward strand. Finally, a streptavidin purification is performed to specifically remove the undesirable cleaved DNA which contain a 5′ biotin from the reverse primer used during amplification. A protocol for amplification of bulk-synthesized primers is detailed below.

See Table 12 for example sequences of an individually synthesized and a bulk-synthesized reverse transcription primer.

In designing the reverse transcription reaction, an important consideration is the concentration of each primer to include. For our original *S. cerevisiae* libraries, a total of 1 µg of pooled primer was used. This pool contained 309 individual primers at equimolar concentration, yielding roughly 160 fmol of each primer in the 1st strand synthesis reaction. For S. pombe, 200 ng of the bulk-synthesized primer pool was included in each 1st strand synthesis reaction. This pool contained 3918 primers mixed at equimolar concentration, yielding roughly 2.5 fmol of each primer in the 1st strand synthesis reaction. It is important to note that while the primer concentrations in both of these instances is significantly higher than the concentrations of their target RNAs, the primers are nevertheless present at concentrations that are likely subsaturating for complete binding of the RNAs. Moreover, the level of saturation achieved for different RNA/primer combinations is likely to vary as a function of the thermodynamic properties of these different complexes. These differences should be meaningless when investigators are comparing the levels of two different isoforms, each of which is targeted by a common primer. That is to say, the efficiency of hybridization of a single primer should be identical between its matched spliced and unspliced (or alternatively spliced) targets. As such, splicing efficiency measurements should be indifferent to the absolute efficiency of any primer. By contrast, if investigators wish to directly compare the counts derived from two different primers, additional experiments would be necessary to establish the absolute efficiency of each primer pair.

The reaction conditions during reverse transcription are crucial for minimizing the abundance of off-target alignments and for maximizing cDNA yield. While the RNA template, buffer, and primer pool are mixed at room temperature, the primers are annealed to the RNA template by raising the temperature to 70 °C for 1 min followed by a 65 °C incubation for 5 min. The reaction is then cooled to the optimal reaction temperature, ideally between 50 and 55 °C, depending upon the enzyme being used and the design of the primers. To reduce non-specific annealing, it is essential that the temperature of this mixture is never allowed to go lower than the temperature at which reverse transcription takes place. In parallel, the reverse transcription enzyme along with the other reaction components are pre-heated to the reaction temperature, after which they are added to the primer-annealed RNA mixture. Here we re-emphasize the importance of maintaining all the reaction components at the reverse transcription temperature prior to mixing so that primers aren't afforded the opportunity to mis-hybridize even for short times at lower temperatures. The choice of enzyme for use in reverse transcription is important for determining the optimal reaction temperature. Many commercial enzymes have different optimal synthesis temperatures and buffer components which will affect the melting temperature of the reverse transcription primers. This step may need to be optimized by the user by generating libraries after using a gradient of temperatures during reverse transcription. The temperature that yields the best library yield with lowest off-target and primer dimer fraction should be selected. For our studies we have used both Superscript III and IV reverse transcriptases. Reverse transcription was performed at 55 °C and 50 °C for *S. cerevisiae* and *S. pombe* libraries, respectively.

Sequencing Considerations

Depending on the targets selected, the sequencing parameters desired for MPE-seq may change (e.g. read depth, single vs paired end, read length, etc.). The read depth required for MPE-seq experiments is much lower than for standard RNA-seq. For example, *S. cerevisiae* libraries made by targeting every intron in the genome require only ~5 million reads to adequately sample the unspliced and spliced isoforms of most transcripts, whereas we estimate nearly 500 million reads of RNA-seq data would be necessary to yield splicing information of similar quality (Xu et al., 2019). Nevertheless, as the number of MPE-seq targets increases so may the read depth required to adequately sample the desired isoforms. The isoforms of interest will also dictate the use of single versus paired-end reads. If the user is considering 'simple' measurements of unspliced versus spliced (or canonically versus alternatively spliced), single end reads that align across the splice junction will be sufficient. When considering read length, reads should be long enough to cross junctions with enough mapped bases on each side to ensure confidence in junction mapping. In our experience, reads that align with >3 nucleotides on each side of a junction can be confidently mapped. However, this may be organism and target dependent. By contrast, paired-end reads enable differentiation between pre-first step pre-mRNAs and lariat intermediate pre-mRNAs (Fig. 2C). The beginning position at which the paired read aligns is the position at which reverse transcription stopped and is informative as to which step of splicing has occurred for a pre-mRNA (Xu et al., 2019).

Optimization and Troubleshooting

When optimization and troubleshooting are required, the efficiency and progress of most steps of the MPE-seq library prep process can be assayed. In this section, we suggest several methods for assaying library prep progress, and we note the steps during the protocol that are amenable to these methods. In our development and optimization of MPE-seq, quantitative PCR (qPCR) was the most useful method to assay each step of the protocol. At virtually every point in the protocol qPCR can be performed to assess the abundance of specific cDNA molecules. There are two primary qPCR methods that we utilized. The first is SYBR based qPCR, which has the advantage of being widely available and straight forward to use. One drawback of SYBR based qPCR is the inability to directly quantify the absolute abundance of DNA molecules in a sample. For this reason, we also used digital droplet PCR (dPCR), which allows for absolute quantification of target molecules in a sample with superior accuracy and precision compared to SYBR based qPCR (Strain et al., 2013). Both methods require primers to be designed against specific targets. A typical MPE-seq experiment will contain many targets that may behave differently at each step of the protocol and for that reason we suggest designing PCR primer sets against multiple targets. For example, it may be useful to design PCR primers against targets across a range of expression levels. The position of the amplicons in relation to the reverse transcription priming sites should also be considered. For example, if using RNA that has been fragmented to a mean length of 200 nt, a quantitative PCR amplicon greater than 200 nt from the reverse transcription priming site, would underestimate the abundance of that target cDNA.

One of the most crucial steps in the MPE-seq protocol is the 1st strand synthesis reaction. We recommend that this be the starting point in any optimization/troubleshooting. As previously noted, an appropriate reaction temperature for reverse transcription is crucial for maximizing cDNA yield as well as minimizing the abundance of off-target priming events. cDNA synthesis efficiency can be assayed immediately after the 1st strand synthesis reaction. From this point, the efficiency of cDNA synthesis can be determined, and a starting number of cDNA molecules can be discerned using qPCR. Quantitative PCR can then be performed at any point downstream and compared to the initial cDNA synthesis values to assess the efficiency of each step in the library prep process. Off target priming events are best analyzed through sequencing test libraries and counting off target alignments. For a typical MPE-seq library from *S. cerevisiae* RNA, we expect to see ~10–20% of reads mapping to off target sites.

Before performing the final library PCR amplification, SYBR based qPCR is performed to determine the appropriate number of amplification cycles. This step is important for minimizing PCR duplicates but also gives an indication of the quality of the library prep. The threshold cycle (Ct) is proportional to the abundance of sequenceable molecules present in the library. Any erroneous step in the protocol could lead to a high Ct value and would be indicative of a poor library. For example, poor cDNA synthesis efficiency would lead to a high Ct value. In our experience, *S. cerevisiae* libraries with a Ct of less than 18 will result in quality datasets, although this value may differ depending on the qPCR machine/software and assay design. Libraries that require more amplification than this are likely to contain a large proportion of primer dimers made

from unextended reverse transcription primers which were not removed prior to the 1st strand extension step of library prep. Another similar indicator of overall library quality is the number of unique molecules sequenced, which is discerned from the number of unique UMIs for each target. A low number of UMIs indicates a low number of sequenceable molecules and/or PCR over amplification.

Methods

Required Equipment

Table 1. Equipment required for MPE-seq

Equipment	Description	Cat#
Thermo-cycler		
Zymo-5 columns	Fiberglass columns for nucleic acid purification from Zymo Research	D4013
Magnetic Stand	Preferably 2 stands: one that can handle both 1.5 mL and 2.0 mL Eppendorf tubes and one that can handle 0.2 mL PCR tubes.	
Rotator	For bead binding incubations	
RT-PCR Machine	For determining library number of amplification cycles	
Qubit	For library quantification and pooling	
Acrylamide gel apparatus	For library size selection	
Eppendorf tubes	2.0 mL, 1.5 mL, 0.5 mL	
PCR tubes	Strips or plates for thermocycler	
Centrifuge	For Eppendorf and PCR tubes	
Bioanalyzer or Fragment analyzer	Library size distribution and pooling	
Bioinformatics capable machine or server access	Access to appropriate software (e.g. Quality assessment, trimming, alignment, feature counting, etc.)	

Materials and Reagents

Table 2. Materials and reagents required for streptavidin bead purification

Reagent	Stock Concentration	Vendor	Cat #
Dyna Beads MyOne Streptavidin C1	10 mg/mL	ThermoFisher	65,001
2x Bind and Wash Buffer	10 mM Tris-HCl pH 7.5, 1 mM EDTA, 2 M NaCl		
1x Bind and Wash Buffer	5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl		
1x SSC	150 mM NaCl, 15 mM Sodium Citrate		
Denaturing Solution	0.1 M NaOH		
TE Buffer	10 mM Tris pH 7.5, 1 mM EDTA		
Bead Elution Buffer	95% Formamide, 10 mM EDTA, pH 8.2		

Table 3. Materials and reagents required for nucleic acid column purification

Reagent	Stock Concentration
Binding Buffer	2 M Guanidinium-HCl, 75% Isopropanol
Washing Buffer	10 mM TRIS pH 8.0, 80% Ethanol

Table 4. Materials and reagents required for amplification and processing of array synthesized 1st strand synthesis primers

Reagent	Stock Concentration	Vendor	Cat #
PCR Primers (oHX093 & oHX094)*	100 µM & 10 µM	IDT	
Phusion polymerase	100x	NEB	M0530S
Phusion buffer	5x	NEB	M0530S
Array synthesized oligos		LC Sciences	
DMSO	100%		
dNTP mix	10 mM each (dATP, dGTP, dCTP, dTTP)		
Sap1 restriction enzyme		NEB	R0569S
Cutsmart buffer	10x	NEB	R0569S
Lambda exonuclease		NEB	M0262S
Lambda exonuclease buffer	10x	NEB	M0262S
Isopropanol	100%		
Ethanol	70%		
Sodium Acetate	3 M, pH 5.3		
Zymo DNA binding buffer		Zymo Research	D4003-1-L

* Primer sequences can be found in Table 12

Table 5. Materials and reagents required for RNA fragmentation

Reagent	Stock Concentration
10x Fragmentation Buffer	100 mM $ZnCl_2$, 100 mM Tris-HCl pH 7.0
Fragmentation stop buffer	0.5 M EGTA

Table 6. Materials and reagents required for 1st strand synthesis

Reagent	Stock Concentration	Vendor	Cat #
SSIV RT Buffer	5x	ThermoFisher	18,090,050
Super Script IV		ThermoFisher	18,090,050
1st strand primer pool	Each primer in *S. cerevisiae* pool at ~80 nM (309 primers). For *S. pombe* each primer at ~2.5 nM (3918 primers)	IDT	
DTT	0.1 M		
Individual dNTPs	100 mM	Sigma-Aldrich	
5-(3-Aminoallyl)-dUTP	50 mM	ThermoFisher (Ambion)	AM8439
Aminoallyl-dNTP mix	10 mM dATP, dGTP, dCTP, 6 mM dTTP, 4 mM Aminoallyl-dUTP	Mixed from above components	
DEPC Water			

Table 7. Materials and reagents required for RNA hydrolysis

Reagent	Stock Concentration
RNA Hydrolysis Solution	0.3 M NaOH, 0.03 M EDTA
Neutralization Solution	0.3 M HCl

Table 8. Materials and reagents required for biotin coupling

Reagent	Stock Concentration	Vendor	Cat #
Sodium Bicarbonate	1.0 M pH 9.0		
EZ-link NHS-Biotin	Diluted to 300 nmol/µL in DMSO	ThermoFisher	20,217

Table 9. Materials and reagents required for 1st strand extension

Reagent	Stock Concentration	Vendor	Cat #
NEB Buffer 2	10x	NEB	MO212L
Klenow exo-	5000 units/mL	NEB	MO212L
dNTPs	10 mM each (dATP, dGTP, dCTP, dTTP)		
1st Strand Extension Primer	100 µM	IDT	

Table 10. Materials and reagents required for library qPCR and amplification

Reagent	Stock Concentration	Vendor	Cat #
Phusion Buffer	5x	NEB	M0530S
Phusion Polymerase	100x	NEB	M0530S
dNTPs	10 mM each (dATP, dGTP, dCTP, dTTP)		
DMSO	100%		
Nextera i5 primer	10 µM	IDT	
Nextera i7 primer	10 µM	IDT	
SyBr Green	100x		

Table 11. Materials and reagents required for gel purification

Reagent	Stock Concentration	Vendor	Cat #
Acrylamide/Bis	29:1		
10X TBE	1 M Tris, 1 M Boric acid, 0.02 M EDTA		
1x TBE	100 mM Tris, 100 mM Boric acid, 2 mM EDTA		
Ammonium persulfate (APS)	10%		
TEMED			
SyBr Gold	10,000x		
100 bp Ladder		NEB	N3231S
Gel Extraction Solution	0.3 M Sodium Acetate		
Glycogen	20 mg/mL		
Isopropanol	100%		
AmpureXP beads		Beckman	A63880
Ethanol	70%		

Table 12. Oligonucleotide sequences

Oligo Name	Description	Sequence 5′ – 3′
YBL092W	Example array-based synthesized RT primer	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNGACAATCTTTGGGTGAGGTAAGGACCTCGAAGAGCATTACGGCTCCTCGCTGCAG
oHX093	Array-based oligo pool amplification primer forward	/5SpC3/TCGTCGGCAGCGTCAGATGTGTATAAGA
oHX094	Array-based oligo pool amplification primer reverse	/5Bioag/CTGCAGCGAGGAGCCGTAATGC
oJP788	1st strand extension oligo (dN9)	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNN/3c6/
YBL092W	Example individually synthesized RT primer	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNGACAATCTTTGGGTGAGGTAAGGA

General Protocols

Below are 2 general protocols that are repeated several times throughout the library preparation. They are listed here for reference.

Column Clean-up

1. Add 7 volumes of binding buffer to each sample and mix well
2. Transfer each sample to a zymo-5 column placed inside a collection tube
3. Spin samples at 14 K RPM for 1 min
4. Discard flow through and place zymo column back in collection tube
5. Wash columns by adding 700 µL washing buffer to each column
6. Spin for 1 min at 14 k RPM
7. Discard flow through
8. Repeat washing for a total of 2 washes
9. Dry column by spinning for an additional 1 min at 14 k RPM
10. Add appropriate elution buffer volume to each column and incubate for ∼1–2 min at room temperature

11. Place columns in labeled collection tubes and spin at 14 K RPM for 1 min

Bead Purification

1. Vortex beads vigorously to resuspend
2. Transfer 20 µL of beads per sample to be purified into a 1.5 mL Eppendorf tube
3. Place tube on magnetic stand and incubate for 1 min
4. While leaving tube on magnetic stand, aspirate buffer
5. Remove tube from magnetic stand
6. Wash beads by adding 500 µL of 1x bind and wash buffer to tube containing the beads and mix well with a pipette
7. Place tube on magnetic stand and incubate for 1 min
8. Aspirate buffer
9. Repeat wash for a total of 2 washes
10. Resuspend beads in 50 µL of 2x bind and wash buffer per sample to be purified
11. Mix beads well with a pipette
12. Aliquot 50 µL of beads into a new tube. 1 for each sample to be purified
13. Add each 50 µL sample to an aliquot of washed beads
14. Place on rotator and incubate at room temperature for 30 min
15. Place tubes on magnetic stand and incubate for 1 min
16. Aspirate buffer
17. Remove samples from magnetic stand
18. Add 500 µL of 1x bind and wash buffer and mix well with pipette
19. Place on magnetic stand and incubate for 1 min

20. Aspirate buffer
21. Repeat washing for a total of 2 washes
22. Wash each sample with 100 µL 1x SSC, mixing well by pipetting
23. Place on magnetic stand for 1 min
24. Aspirate buffer
25. Remove samples from magnetic stand
26. Add 100 µL of denaturing solution to each sample
27. Incubate each sample for 10 min at room temperature
28. Place samples on magnetic stand for 1 min and aspirate buffer
29. Wash by adding 100 µL of denaturation solution to each sample. Mix well by pipetting
30. Place samples on magnetic stand for 1 min and aspirate buffer
31. Add 100 µL of TE buffer and mix well with pipette
32. Place on magnetic stand and incubate for 1 min
33. Aspirate buffer
34. Repeat TE buffer wash 2 times for a total of 3 washes
35. Add 100 µL of bead elution buffer to each sample
36. Incubate samples at 90 °C for 2 min
37. Immediately place samples on magnetic stand and incubate for 1 min
38. Aspirate each sample and transfer to a new tube

Complex Oligo Array Amplification

1st PCR Amplification

Perform a SYBR based qPCR experiment to establish the number of PCR cycles to perform to prevent the reaction from plateauing:

1. Mix the qPCR mix detailed in Table 13

Table 13. qPCR amplification reaction mix

Reagent	Volume (µL)
Phusion buffer	10
dNTP mix	1
oHX093 (10 µM)	1.25
oHX094 (10 µM)	1.25
LC oligo mix	0.125% total mass
DMSO	0.5
SyBr Green	5
Phusion	0.5
H20	Up to 50

2. Aliquot 15 µL of qPCR reaction mix into 3 wells of a qPCR plate
3. Run the following cycle program:

Initial Denaturation

98 °C 30 s

Amplification (40x)

98 °C 10 s

60 °C 20 s

72 °C 30 s

Final extension

72 °C 3 min

4. Chose the cycle number closest (but prior) to the plateau of the amplification curve for all 3 samples. Use this cycle number for the 1st PCR amplification.
5. Mix the PCR mix detailed in Table 14

Table 14. 1st PCR amplification reaction mix.

Reagent	Volume (µL)
Phusion buffer	80
dNTP mix	8
oHX093 (10 µM)	10
oHX094 (10 µM)	10
LC oligo mix	1% mass*
DMSO	4
Phusion	4
H20	Up to 400

* We chose to use 1% of the total oligo mass to maximize the amount of RT primer generated from an individual synthesis. This amount can be optimized by the user.

6. Run the following cycle program:

   Initial Denaturation

   98 °C 30 s

   Amplification (Cycle number determined in step 6)

   98 °C 10 s

   60 °C 20 s

   72 °C 30 s

   Final extension

   72 °C 3 min

7. Repeat steps 1–6 using 0.5 µL of 1st PCR reaction as template to determine amplification cycle number for 2nd PCR

2nd PCR Amplification

1. Prepare a 2nd PCR mix using the product from the 1st PCR as the template. Mix the reaction according to Table 15

Table 15. Reaction mix for 2nd PCR amplification.

Reagent	Volume (µL)
Phusion buffer	8000
dNTP mix	800
oHX093 (100 µM)	100
oHX094 (100 µM)	100
Product from 1st PCR	400
DMSO	400
Phusion	400
H20	29,800

2. Aliquot 100 µL volumes of the reaction mix into four 96 well plates and run the following cycling program:

Initial Denaturation

98 °C 30 s

Amplification (Cycle number determined as in step 8 of above)

98 °C 10 s

60 °C 20 s

72 °C 30 s

Final extension

72 °C 3 min

Concentrate DNA

1. Pool the reactions and split equal volumes into 50 mL falcon tubes
2. Isopropanol precipitate each sample by adding a 1/10th volume of 3 M sodium acetate followed by an equal volume of isopropanol
3. Let sample precipitate at room temperature for ~30 min
4. Spin in a capable centrifuge at max speed for 1 h

5. Wash pellets twice with 70% ethanol centrifuging at max speed for 5 min in between washes
6. Allow pellet to dry
7. Dissolve each pellet in 700 µL H20
8. Pool samples together and isopropanol precipitate again
9. Dissolve the pellet in 300 µL H20
10. Perform column purification on sample using 1500 µL of Zymo DNA binding buffer Once binding buffer is added split sample equally into 4 Zymo-25 columns and follow column purification protocol from section above starting at step 2.
11. Elute from each column with 105 µL H20
12. Combine eluents

SapI Digestion

1. Add 50 µL 10X smart cut buffer
2. Add 30 µL SapI enzyme.
3. Incubate sample at 37 °C for 15 h
4. Isopropanol precipitate sample
5. Resuspend pellet in 125 µL H20

Lambda Exonuclease Digestion

1. Add 15 µL 10X lambda exo- buffer to sample
2. Add 10 µL lambda exo-
3. Digest at 37 °C for 2 h
4. Total volume is 150 µL.

5. Perform column purification of sample. Once binding buffer is added, split the sample equally into 2 Zymo-25 columns
6. Elute DNA from each column by adding 25 µL H20 and combine eluants

Streptavidin Bead Purification

1. Vortex beads vigorously to resuspend
2. Transfer 50 µL of beads per sample to be purified into a 1.5 mL Eppendorf tube
3. Place tube on magnetic stand and incubate for 1 min
4. While leaving tube on magnetic stand, aspirate buffer
5. Remove tube from magnetic stand
6. Wash beads by adding 500 µL of 1x bind and wash buffer to tube containing the beads and mix well with a pipette
7. Place tube on magnetic stand and incubate for 1 min
8. Aspirate buffer
9. Repeat wash for a total of 2 washes
10. Resuspend the beads in 50 µL 2x bind and wash buffer
11. Add the sample and mix well with a pipette
12. Place sample on a tube rotator and incubate sample for 15 min at room temperature
13. Heat sample to 65 °C and incubate for 2 min
14. Place tube on magnetic stand and allow beads to separate for 1 min
15. Aspirate supernatant and place in a new Eppendorf tube. This is the purified pool of RT primers
16. Isopropanol precipitate samples and dissolve pellet in 30 µL H20

17. Check DNA concentration via nano-drop or Qubit

    Note: For *S. pombe*, the purified array synthesized primers were diluted to roughly 100 ng/µL

1st Strand Synthesis

Note: The 1st strand synthesis protocol detailed here is for an RT reaction temperature of 55 °C. See section 3.2 for design considerations.

1. Make a primer/template mix for each sample as detailed in Table 16

Table 16. Reaction mix for 1st strand synthesis primer annealing.

Reagent	Volume (µL)
5X RT buffer	4
1st strand primer pool	2
Total RNA	10 µg
DEPC Water	Up to 20
total	20

2. Place samples in thermo-cycler and run the following cycle:

    70 °C 1 min

    65 °C 5 min

    55 °C Hold

3. Make the enzyme dNTP mix for each sample as detailed in Table 17

Table 17. Reaction mix for 1st strand synthesis.

Reagent	Volume (µL)
5X RT buffer	4
DTT	2
Aminoallyl-dNTP mix	2
SSIV	2
DEPC Water	Up to 20
Total	20

4. Heat dNTP Enzyme mix to 55 °C by placing it in the thermo-cycler containing the template/primer mix

5. Directly add 20 µL dNTP/enzyme mix to the template/primer mix, ensuring that both samples are kept at 55 °C
6. Allow the 1st strand synthesis reaction to proceed for 10 min at 55 °C
7. Incubate sample at 80 °C for 10 min to inactivate the enzyme

   Note: Sample volume is now 40 µL

RNA Hydrolysis

1. Add 20 µL of RNA hydrolysis solution to each sample
2. Incubate at 65 °C for 15 min
3. Add 20 µL of neutralization solution to each sample
4. Note: Sample volume is now 80 µL
5. Perform Column clean up as detailed in the protocol above
6. Elute each sample with 16 µL H20

Biotin Coupling

1. Add 2 µL of sodium bicarbonate to each sample
2. Add 2 µL of NHS-biotin to each sample
3. Incubate at 65 °C for 1 h in the dark
4. Note: Perform the next few steps quickly to prevent biotin precipitation
5. Briefly spin tubes in a centrifuge
6. Add 40 µL H20
7. Note: Sample volume is now 60 µL
8. Perform column cleanup
9. Elute in 50 µL H20

10. Perform streptavidin bead purification on each sample
11. Perform column purification and elute in 40 µL H20
12. You now have purified 1st strand cDNA

Note: The protocol can be paused here by storing the samples at −20 °C

1st Strand Extension

1. Prepare the following reaction mix described in Table 18 for each sample

Table 18. Reaction mix for 1st strand extension.

Reagent	Volume (µL)
10X NEB Buffer 2	5
10 mM dNTP Mix	1
cDNA from previous	40
1st Strand Extension Primer	1
Total	47

2. Incubate each sample at 65 °C for 2 min
3. Cool samples to room temperature by placing on bench top for ∼5 min
4. Add 3 µL of Klenow exo- enzyme to each sample
5. Incubate samples at room temperature for 5 min
6. Heat samples to 37 °C and incubate for 30 min
7. Heat samples to 75 °C and incubate for 20 min to inactivate enzyme
8. Perform bead purification on each sample
9. Perform column purification on each sample eluting with 33 µL H20

Note: This is a safe stopping point. Samples can be stored at −20 °C

PCR Amplification

qPCR

1. Perform triplicate qPCR reactions on each sample by mixing the qPCR reaction mix detailed in Table 19

Table 19. qPCR reaction mix.

Reagent	Volume (µL)
5x Phusion Buffer	10
dNTP Mix	1
1st Strand Extension Product	2
Nextera i5 Primer	2.5
Nextera i7 Primer	2.5
Phusion Polymerase	0.5
DMSO	0.5
SyBr Green	0.5
H20	30.5

2. Aliquot 15 µL of each qPCR reaction mix into 3 separate wells of a qPCR plate
3. Perform the following qPCR cycling:

    Denaturation:

        98 °C 30 s

    Amplification (40x):

        98 °C 10 s

        62 °C 20 s

        72 °C 30 s

    Final elongation:

        72 °C 3 min

4. After the qPCR calculate the Ct value for each sample and average the 3 values for each library
5. Calculate the number of amplification cycles to perform by averaging the 3 replicate Ct values for each library, adding 3 and rounding to the nearest whole number

Note: The number of PCR cycles can be reduced by 2–3 if bead purification is performed instead of gel purification as this results in more efficiency purification

PCR Amplification

1. For each sample, prepare the following PCR reaction mix detailed in Table 20

Table 20. Reaction mix for library PCR amplification.

Reagent	Volume (µL)
5x Phusion Buffer	10
dNTP Mix	1
1st Strand Extension Product	10
Nextera i5 Primer	2.5
Nextera i7 Primer	2.5
Phusion Polymerase	0.5
DMSO	0.5
H20	23

2. Run the following cycling conditions on each PCR reaction mix:

Denaturation

98 °C 30 s

Amplification (Cycle # determined in step 5 of qPCR):

98 °C 10 s

62 °C 20 s

72 °C 30 s

Final elongation:

72 °C 3 min

4 °C Hold

Size Selection, Clean-up, and Quality Check

1. Prepare a 6% native acrylamide gel

2. The entirety of each PCR reaction will be run in a single well so make sure the wells can comfortably hold 60 µL (50 µL sample + 10 µL loading dye)
3. Run a 100 bp ladder for sizing.
4. Once samples are loaded on the gel, run at 250 V for 70 min.
5. Stain in 1x SyBr Gold for 15 min
6. Image gel for documentation
7. For each sample, cut out the lane from 200 bp to 800 bp and place in a 2 mL tube
8. Add 900 µL of 0.3 M NaOAc to each sample and place tubes on rocker overnight.
9. Added 900 µL isopropanol to precipitate.
10. Add 1 µL glycogen to each sample
11. Allow sample to precipitate.
12. Spin at 14 k RPM for 30 min
13. Wash with 500 µL of 70% ethanol twice
14. Dry pellet and resuspend in 40 µL H20
15. Assay DNA concentration using Qubit
16. It is also helpful to run the samples on a Bioanalyzer or similar instrument to check the size distribution of the library
17. Pool libraries for sequencing

    Note: The gel running conditions will likely vary between equipment and so careful attention should paid at first. The goal is to get as much separation as possible. We run the gel until the 100 bp ladder band is ~1–2 cm from the bottom of the gel. After purification, library concentration is generally somewhere between 0.5 ng/µL to

5 ng/µL. Fig. 3A shows a Bioanalyzer trace of an S. cerevisiae library after size selection with a gel.

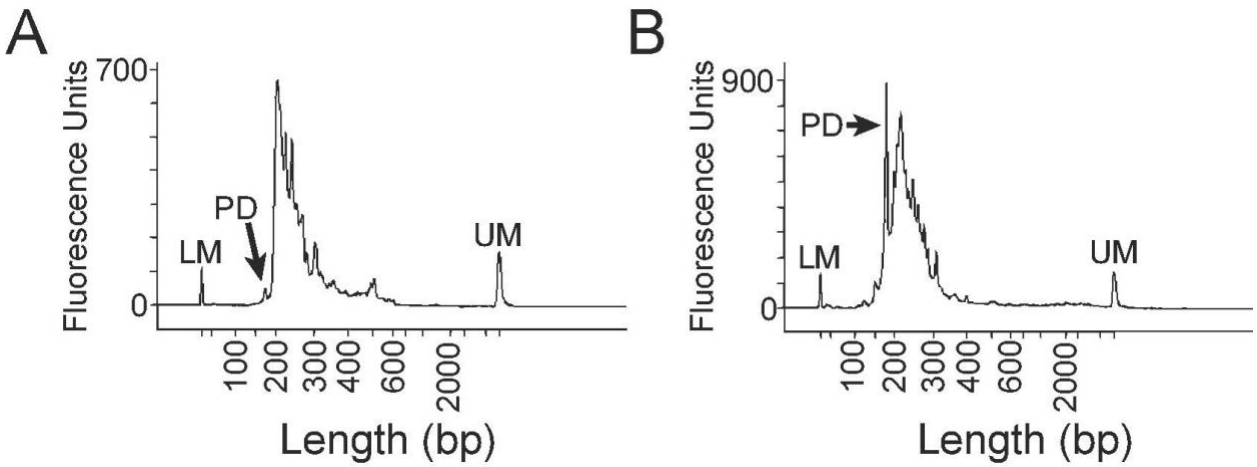


Figure 3: Bioanalyzer traces of replicate *S. cerevisiae* MPE-seq libraries after size selection by (**A**) acrylamide gel, or (**B**) AmpureXP beads. LM and UM denote the lower and upper markers, respectively. PD denotes the peak resulting from primer dimer molecules.

Alternative Size Selection using AmpureXP Beads

It may be preferable to purify libraries using AmpureXP beads or an equivalent. The purpose of gel sizing is 2-fold: (1) It removes excess PCR primers and, (2) It removes primer dimers that run at ~180 bp. Depending on the user's design (e.g. number of targets, input RNA, target abundance, etc.) primer dimers may make up a very small portion of the library and not need to be excluded. This can be determined by the user. One downside to gel sizing worth considering is it may introduce a bias against short cDNAs derived from features of interest such as lariat intermediates. Fig. 3B shows a Bioanalyzer trace of an S. cerevisiae library with size selection via AmpureXP beads.

Data analysis

In the following sections we describe a general pipeline for MPE-seq data analysis and suggest publicly available software packages for each step.

Read Processing and Quality Control

The first step in MPE-seq data processing is to de-multiplex reads into their respective samples. This is achieved by parsing reads into fastq files based on their respective barcodes. Depending on the sequencing platform and center, this processing may be provided before the data are returned to the user. Next, we suggest assessing the overall quality of the reads, which can be done with tools like FastQC. The output of this software provides overall quality scores, read lengths, adapter dimer content, and sequence duplication levels. The next step in processing is compressing reads resulting from PCR duplications into a single read. For this step, the entirety of each sequence, which contains the UMI and the rest of the read, is considered. If multiple reads have identical sequences, they are considered PCR duplicates and compressed into a single read for counting purposes (Kivioja et al., 2012). For this processing we use custom python scripts that search for reads with identical sequences within a fastq file and generate a new fastq file containing those unique reads (see Code Availability section). After this step, the reads are trimmed to remove any regions that correspond to the Illumina-based sequencing adapters. This step can be accomplished using publicly available software such as Trimmomatic (Bolger et al., 2014). In addition, the UMI sequence is also trimmed in this step to prevent mapping artefacts. It may be useful to move the trimmed UMI to the sequence name using a custom script so that it can be accessed in the SAM file after alignment. An alternative approach is to clip the UMI sequence during alignment.

Alignment

Properly trimmed reads can be aligned to a reference genome using a variety of splicing aware aligners such as HISAT (Kim et al., 2019) or STAR (Dobin et al., 2013). There are a few important parameters to consider when aligning these reads. First, the minimum alignment length on either side of a splice junction, also known as anchor length, should be specified. We suggest an anchor length of 3, as we have found this best reduces erroneous mappings. Second, we generally include in the output unmapped and multimapping reads so they can be used to quantify alignments resulting from off-target priming events during cDNA synthesis, which is useful for trouble shooting and optimization. Third, reads resulting from primers that were unextended by reverse transcriptase during 1st strand synthesis and were unsuccessfully removed during library prep purifications can generate erroneous alignment counts and should be removed from the analysis. These reads will contain the gene specific regions used for reverse transcription and will align once adapter sequences have been trimmed. Additionally, a small number of bases can be added during 1st strand extension due to an overhang of the dN9 region of the 1st strand extension primer hybridized to the reverse transcription primer. For this reason, in some cases, the aligned region of these reads may overlap with a splice junction and result in erroneous counts. These reads can either be filtered out during or post alignment by requiring an insert size of >30 bases. This can be accomplished with alignment parameters or by post alignment filtering with custom scripts.

Read Allocation and Isoform Quantitation

After reads have been aligned, the abundance of each target isoform can be quantified. Fig. 2C shows an example of how paired end sequencing alignments can be partitioned into 6 categories (labeled C1 to C6 on Fig. 2C) which support each isoform for a single splicing event (Fig. 2C). The abundance of Spliced (S), Unspliced (U), Pre-first step (P), and Lariat intermediate (L) isoforms can then be calculated as follows:

$$S = C_1 + C_2$$

$$U = C_3 + C_4 + C_5 + C_6$$

$$P = C_3 + \left((C_4 + C_5)\left(\frac{C_3}{C_6}\right)\right)$$

$$L = C_6 + \left((C_4 + C_5)\left(\frac{C_6}{C_3}\right)\right)$$

To count the number of S isoforms for a target, the CIGAR string in SAM/BAM alignment files can be utilized. Alignments which contain a gap due to splicing will contain an ‘N’ in the CIGAR string, from which the positions of the spliced intron can be discerned. The abundance of S isoforms can then be counted as the number of alignments that contain an ‘N’ in the CIGAR string corresponding to a gap resulting from the excised intron (C1, C2) (Fig. 2C). U isoforms will not contain an ‘N’ in the CIGAR string and will map within the intron (C3 – C6) (Fig. 2C). By utilizing paired end reads, MPE-seq affords the ability to differentiate unspliced isoforms further into those that have not undergone the 1st step of splicing from those that have undergone the 1st step of splicing but not the 2nd step of splicing (Xu et al., 2019). P and L isoforms can be quantified by considering the mapping location of the first base of the second read of paired-end reads (Fig. 2C). cDNA synthesis will terminate at the branch point adenosine of lariat intermediates due to the inability of reverse transcriptase to process through the

branched structure. If the first base of a second read maps within +/− 4 bases from the branch point adenosine, the read pair is unambiguously counted as a lariat intermediate (C6) (Fig. 2C). A small window around the branch point adenosine is suggested due to our observation that second read ends don't pile up precisely at annotated branch point adenosines (Xu et al., 2019). All read pairs which have a second read mapping location upstream of the branch point window are unambiguously counted as pre-first step (C3) (Fig. 2C). Read pairs with a second read that maps downstream of the branch point window (C5) or does not map but has a first read that maps within the intron (C4) are considered ambiguous and are assigned to lariat intermediate or pre-first step based on the proportion of unambiguous lariat intermediates to pre-first step read pairs. This relationship is shown in the equations above. Each MPE-seq read corresponds to a single reverse transcription priming event on a single transcript, precluding the necessity for length normalization. As a result, counts can be used directly to calculate splicing efficiency as the user sees fit. Data generated in this way can subsequently be compared between experimental conditions using appropriate statistical methods.

The simple quantification method detailed above assumes that the frequency of random reverse transcription termination events and RNA fragmentation that occur within the branch point window is minimal. It also assumes that reverse transcriptase reads through lariat branches at a very low frequency. It is important to note that this quantification also requires a priori knowledge of the position of the branch point adenosine and may not be possible in organisms with degenerate or non-annotated branch points.

As noted earlier, MPE-seq can be used to quantify the frequency of canonical as well as alternative splice isoforms, such as those generated by exon skipping and alternative 3′ and 5′ splice site usage, provided that these events are located within the sequencing window of a targeted region. To quantify these events, the CIGAR string of BAM/SAM files can be utilized. Alignments which contain junctions or gaps denoted by an ‘N’ in the CIGAR string at the coordinates of specific alternative events can be filtered and counted. Some aligners, such as STAR, generate a file containing the genomic positions and counts of all splice junctions identified, removing the need for custom scripts or additional software packages to count these events (Dobin et al., 2013). These counts can then be compared to quantify the frequency of specific alternative splice isoforms.

A variety of counting programs are available that can be utilized with MPE-seq data such as HTSeq (Anders et al., 2015), featureCounts (Liao et al., 2014), and BEDTools (Quinlan & Hall, 2010). However, these software packages are designed to count reads that overlap specific features (e.g. genes) and do not differentiate isoforms within a feature (e.g. spliced vs unspliced). For that reason, some combination of data filtering, modification of counting software, and/or use of custom scripts is required for properly counting and analyzing MPE-seq data.

Conclusions

The study of RNA splicing has been greatly enhanced by the adaptation of next-generation sequencing. RNA-seq has been pivotal in genome wide analysis of the transcriptome, leading to the identification of countless numbers of novel splicing isoforms. While RNA-seq has been crucial for the analysis of transcript abundance, the

unbiased nature of cDNA generation in most library generation methods often results in poor sampling of splicing informative loci. By targeting cDNA synthesis to splicing informative loci of choice, MPE-seq provides unique opportunities to measure splicing efficiency with high precision and to detect splice isoforms that are otherwise poorly sampled with standard RNA-seq. It is important to note that the targeted nature of MPE-seq renders it poor at identifying novel splicing events outside of targeted regions. Nevertheless, MPE-seq provides the unique ability to detect lariat intermediates, which allows for the analysis of the efficiencies of the 1st and 2nd steps of splicing, a level of resolution that is unavailable with standard RNA-seq. Moreover, the increase in read depth afforded by MPE-seq has the additional advantage of greatly reducing sequencing costs and allows for multiplexing of many samples on a single sequencing lane. Beyond the analysis of steady-state total RNA, MPE-seq can readily be applied to study splicing efficiency in a variety of biological contexts through use of fractionated RNA sources such as, poly(A)+ selected RNA, polysome associated RNA, nuclear/cytoplasmic RNA, and nascent RNA purified with metabolic labels.

Code Availability

Details of all data analysis associated with this paper can be found at https://github.com/zdwyer/MPE-seq-methods.

Data Availability

Raw sequencing data are available through Gene Expression Omnibus (GEO accession: GSE126583)

Acknowledgements

We thank B. Fair and H. Xu for their work on the development of MPE-seq. We thank members of the J.A.P., H. Kwak, and A. Grimson laboratories for helpful discussions on development of this method. We thank P. Schweitzer, J. Grenier, and the BRC Genomics Facility at Cornell for outstanding technical support with Illumina sequencing.

Funding

This work was funded by a Research Scholars Grant from the American Cancer Society (to J.A.P.) and NIH grant R01GM098634 (to J.A.P.)

Works Cited

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

David, C. J., & Manley, J. L. (2010). Alternative pre-mRNA splicing regulation in cancer: Pathways and programs unhinged. *Genes & Development*, *24*(21), 2343–2364. https://doi.org/10.1101/gad.1973010

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Faustino, N. A. (2003). Pre-mRNA splicing and human disease. *Genes & Development*, *17*(4), 419–437. https://doi.org/10.1101/gad.1048803

Katz, Y., Wang, E. T., Airoldi, E. M., & Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, *7*(12), 1009–1015. https://doi.org/10.1038/nmeth.1528

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., & Taipale, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, *9*(1), 72–74.

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(323).

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Nilsen, T. W., & Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature*, *463*(7280), 457–463. https://doi.org/10.1038/nature08909

Oshlack, A., & Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, *4*(1), 14. https://doi.org/10.1186/1745-6150-4-14

Padgett, R. A. (2012). New connections between splicing and human disease. *Trends in Genetics*, *28*(4), 147–154. https://doi.org/10.1016/j.tig.2012.01.001

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, *40*(12), 1413–1415. https://doi.org/10.1038/ng.259

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, T., & Salzberg, S. L. (2016). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842.

Scotti, M. M., & Swanson, M. S. (2016). RNA mis-splicing in disease. *Nature Reviews Genetics*, *17*(1), 19–32. https://doi.org/10.1038/nrg.2015.3

Strain, M. C., Lada, S. M., Luong, T., Rought, S. E., Gianella, S., Terry, V. H., Spina, C. A., Woelk, C. H., & Richman, D. D. (2013). Highly Precise Measurement of HIV DNA by Droplet Digital PCR. *PLoS ONE*, *8*(4), e55943. https://doi.org/10.1371/journal.pone.0055943

Tazi, J., Bakkour, N., & Stamm, S. (2009). Alternative splicing and disease. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, *1792*(1), 14–26. https://doi.org/10.1016/j.bbadis.2008.09.017

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., & Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–476. https://doi.org/10.1038/nature07509

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Wernersson, R., Juncker, A. S., & Nielsen, H. B. (2007). Probe selection for DNA microarrays using OligoWiz. *Nature Protocols*, *2*(11), 2677–2691.

Xu, H., Fair, B. J., Dwyer, Z. W., Gildea, M., & Pleiss, J. A. (2019). Detection of splice isoforms and rare intermediates using multiplexed primer extension sequencing. *Nature Methods*, *16*(1), 55–58. https://doi.org/10.1038/s41592-018-0258-x

Zhang, J., & Manley, J. L. (2013). Misregulation of Pre-mRNA Alternative Splicing in Cancer. *Cancer Discovery*, *3*(11), 1228–1237. https://doi.org/10.1158/2159-8290.CD-13-0253

Chapter II: The problem of selection bias in studies of pre-mRNA splicing

Alternative Citation

Dwyer ZW and Pleiss JA. The problem of selection bias in studies of pre-mRNA Splicing. *In Review*

Abstract

Here we demonstrate how selection bias in studies of pre-mRNA splicing can easily generate biologically flawed conclusions which nevertheless appear statistically robust. We argue that the widespread nature of this problem has important and deleterious consequences for the field.

Introduction

We write here to raise awareness of the problem of selection bias in analyses of next-generation sequencing studies designed to understand quantitative changes in pre-mRNA splicing. Selection bias, also sometimes referred to as sample bias or sample selection bias, generally refers to a distortion (bias) of statistical testing that results from the way that samples are collected (selection). While this problem has been well understood in clinical and social sciences for quite some time, (Antman et al., 1985) its significance in molecular biology has been less widely appreciated. Work from Oshlack's group (Oshlack & Wakefield, 2009; Young et al., 2010) provides perhaps the most compelling demonstration of the problem of selection bias in molecular biology, wherein they demonstrate its impact on RNA-seq experiments where analyses of differential gene expression are coupled to analyses of GO-term enrichment. Oshlack's group highlights a major pitfall in studies like this which is that not all transcripts are measured with the same statistical power in a typical RNA-seq experiment: because

longer and more highly expressed transcripts are sampled more frequently than are shorter and lower expressed transcripts, there is more statistical power to identify long and/or highly expressed transcripts as being differentially expressed. The result of this is a distortion in the statistics that measure enrichment: the set of transcripts identified as differentially expressed is dependent not only on their biological behavior but on unrelated properties that enhance their capacity for detection in the experiment. Importantly, because splicing-informative reads are so rare within standard RNA-seq datasets, the problem of selection bias is particularly problematic in studies involving pre-mRNA splicing.

Selection bias, in principle

To demonstrate both the problem of selection bias and the deleterious consequences of this bias in studies of pre-mRNA splicing, we present here the results from a simple experiment designed to examine changes in splicing in the background of a well characterized genetic variant of a canonical spliceosomal component, the RNA helicase Prp2. Work from many groups has established a role for Prp2 in rearranging the spliceosome prior to the first catalytic step, and as such a reasonable expectation is that loss of Prp2 function would result in defective splicing for all (or nearly all) expressed transcripts. Using a targeted sequencing approach termed Multiplexed Primer Extension Sequencing, or MPE-seq (Xu et al., 2019), which massively enriches for splicing informative reads, we generated rich datasets, equivalent to ~ an entire lane of NextSeq550 sequencing for each of triplicate samples from a budding yeast strain harboring the conditional prp2-1 allele and a matched wildtype strain. To demonstrate the effect of sequencing depth on experimental outcome, we then computationally

downsampled this large experiment to generate three smaller subsets of data, equivalent in an RNA-seq setting to what could be considered high, medium, and low sized experiments, or ~80, 40, and 20 million reads per replicate, respectively.

To analyze these datasets, for each intron-containing gene in the genome we calculated the fold change in abundance of reads corresponding to both premature and mature isoforms and then assessed these for differential expression using DESeq2 (Love et al., 2014). As seen in Figure 1A, of the 272 events profiled in the full dataset, 261 demonstrated statistically-significant differential splicing in the mutant relative to wildtype. For the majority of these (211), both the premature and mature versions of the transcript were detected as differentially expressed, whereas a smaller number of transcripts displayed differential expression of only one of the two isoforms, presumably reflecting different intrinsic properties of the rates of synthesis or degradation of these transcripts. Importantly, the absence of evidence for differential expression of either splicing isoform for the remaining 11 transcripts in this experiment cannot be interpreted as evidence of an absence of an impact of the prp2-1 variant on these transcripts. While such a biological conclusion might be true, it may simply be that the design of this experiment was ‘biologically flawed’, perhaps because these transcripts were not actively expressed under these conditions. Equally plausible, however, is the possibility that the experiment was ‘statistically flawed’ because even at this high sequencing depth it lacked sufficient statistical power to detect real changes in the splicing efficiency of these transcripts.

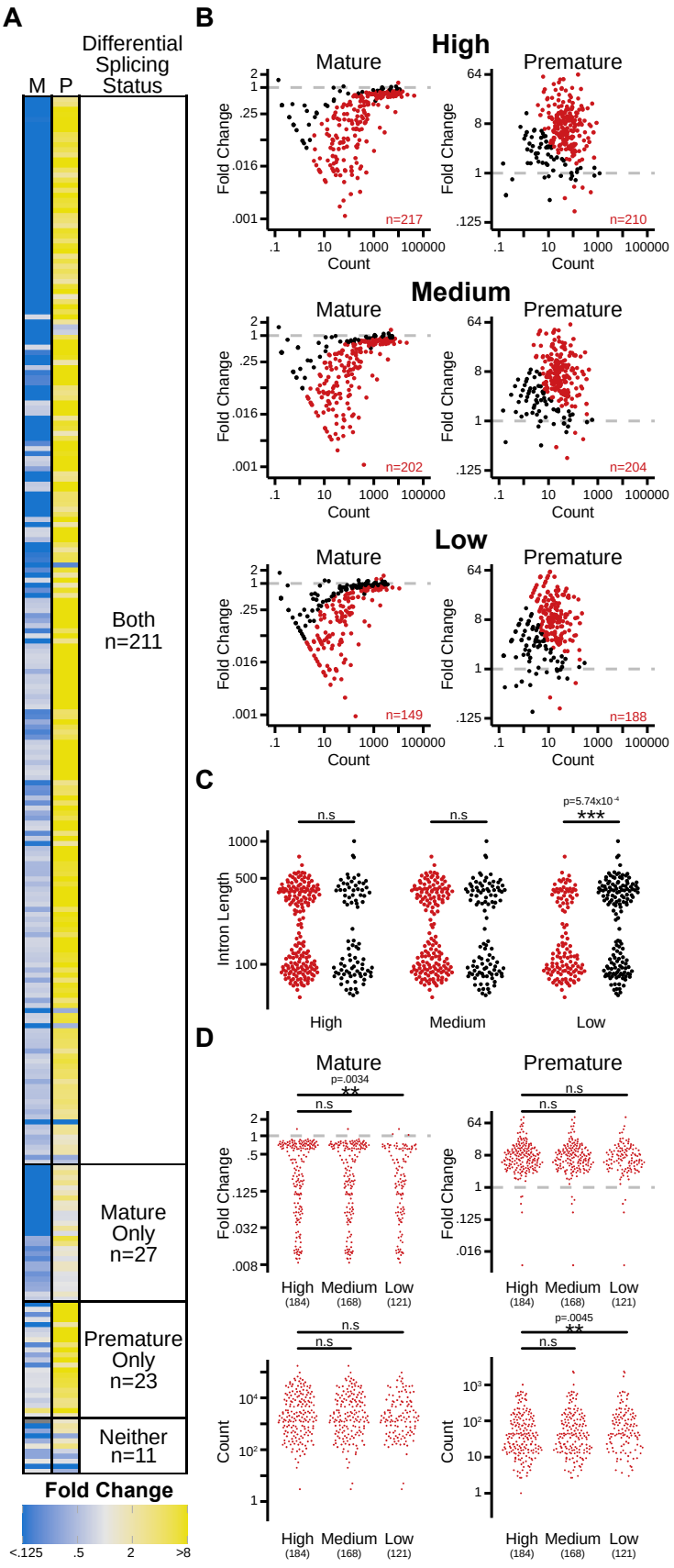

A
Differential Splicing Status
M P
Both n=211
Mature Only n=27
Premature Only n=23
Neither n=11
Fold Change
<.125 .5 2 >8
B
High
Medium
Low
Mature
Premature
Fold Change
Count
n=217
n=210
n=202
n=204
n=149
n=188
C
Intron Length
n.s
p=5.74x10-4
***
High
Medium
Low
D
Mature
Premature
p=.0034
**
p=.0045
**
n.s
High (184)
Medium (168)
Low (121)
Count

Figure 1: Selection bias introduces false correlations at insufficient read depth. Comparison of genome-wide splicing status in a *prp2-1* harboring strain relative to a matched wildtype strain after a ten-minute shift to the non-permissive temperature (37°C). **(A)** Heat map of fold change broken into categories of introns where both mature and premature, only mature, only premature, and neither mature nor premature supporting reads are significantly different between the Prp2 mutant and wild type as measured by DESeq2 at a multiple hypothesis corrected value of .05. **(B)** Fold change in number of mature or premature supporting reads as a function of expression between Prp2 mutant and wild type after downsampling to library sizes of 800,000 (High) MPE-seq reads to 400,000 (Medium) and 200,000 (Low) reads as measured by DESeq2. Red points are statistically significantly different at a multiple hypothesis corrected value of .05. **(C)** Length of introns that have (red) or do not have (black) significant difference in both premature and mature supporting reads as a function of read depth. One-sided Mann-Whitney test performed to determine significance. **(D)** Fold change (upper) and count total (lower) for mature and premature supporting reads of introns that are significantly different in both mature and premature supporting counts at each downsampling. Two-sided Mann-Whitney test performed to determine significance.

The consequences of decreased statistical power become readily apparent when considering our analyses of the downsampled datasets, as shown in Figure 1B, wherein ever decreasing numbers of events were detected as differentially expressed with statistical significance as the read depth decreased. Whereas analysis of the full dataset demonstrated with statistical significance the widespread impact of prp2-1 on the genome-wide splicing outcome, in standard sized experiments many of these splicing events lack the statistical power to be 'selected' within the class of transcripts considered impacted by prp2-1: the selection bias problem. Importantly, as Oshlak's group previously demonstrated for standard gene expression studies, loss of statistical power does not occur evenly across the complement of genome-wide events being monitored, but rather occurs as a function of intrinsic properties of those targets which may or may not be important in the context of the biological problem being examined. For example, Figure 1C shows a comparison of the lengths of the introns that were identified as impacted or not at each of the different experimental depths. Whereas no length difference was apparent between these classes at the High and Middle sizes, in the Low dataset a strong and statistically significant difference in the intron lengths was observed between the classes. While this result might suggest that short introns are more sensitized to loss of Prp2 function, the underlying data are more consistent with

this being a biologically meaningless artifact of the loss of statistical power. As with most approaches for statistical testing, DESeq2 considers two important properties of the data in determining significance: the effect-size, or difference in expression between the experimental and control samples; and the variance associated with the underlying measurements. As seen in Figure 1D, for the relatively highly sampled mature isoforms (left), the subset of introns identified as differentially expressed are not characterized by higher read counts, but rather by larger fold-changes: small fold-changes in expression of the mature mRNA only surpass the significance threshold in the highly sampled dataset where overall variance is decreased. By contrast, for the relatively rare premature isoforms, the subset identified as differentially expressed is biased towards those that are highly sampled: even large fold-changes in differential expression fail to be deemed statistically significant if read depth is low (where variance is naturally higher).

Selection bias, in practice

While the above data demonstrate how selection bias *can* impact splicing studies, it is reasonable to ask whether in practice it ever *does* influence the field. We argue that this problem is indeed widespread, and with apologies for negatively highlighting their work we consider here a single study from Corsini *et al.* (Corsini et al., 2018) which examined the role of the splicing factor HTATSF1, the ortholog of yeast Cus2. As a core component of the U2 snRNP, and building off of significant prior work demonstrating a role for Cus2 in stabilizing a core structure of the U2 snRNA (Rodgers et al., 2016; Yan et al., 1998), it was a reasonable expectation that loss of HTATSF1/Cus2 activity would lead to decreased splicing efficiency across the

complement of genome-wide substrates, akin to our expectations and observations for loss of Prp2 function as shown above. By contrast, based on a knock-down experiment in mouse embryonic stem cells, Corsini *et al.* suggested that HTATSF1 functioned as a "regulator of intron retention specifically in ribosomal proteins," a conclusion echoed in an accompanying perspective by Sharma and Blencowe. (Sharma & Blencowe, 2018) However, while Corsini *et al.* provide compelling statistical support for intron retention within 45 different transcripts, many of which are involved in ribosome biogenesis and assembly, in the context of selection bias it is important to understand whether these transcripts were indeed uniquely impacted by loss of HTATSF1, or whether they simply reflect the subset of transcripts for which there was sufficient statistical power in their experiment to detect a change in splicing efficiency. We therefore asked whether the underlying data from the Corsini study suggested a bias in the subset of selected events by examining two parameters which are expected to influence pre-mRNA detection: expression level of the host transcript; and distance between the end of the affected intron and the polyadenylation site for that transcript. Our motivation for examining this second feature is that most RNA-seq protocols, including the one employed by Corsini *et al.,* utilize a poly(A)$^+$ enrichment step: because splicing is coupled to transcription, the likelihood of a retained intron being detected within a poly(A)$^+$ pool of RNA is expected to be highest for those located closest to the polyadenylation site, as these would have the least amount of time for removal prior to polyadenylation. As seen in Figure 2, the events identified in Corsini *et al.* are indeed massively biased for each of these properties such that they are likely to have much greater statistical power for detecting differential expression than most of the other events in the genome. Importantly, while it

is true that Ribosomal Protein Genes (RPGs) are naturally skewed relative to the global population for both of these features, we note that nearly half of the retained introns identified in Corsini *et al.* are within non-RPGs, and that these show a nearly identical bias within the data. Taken together, we argue that there is insufficient data within this experiment to support authors' hypothesis that HTATSF1 "specifically controls splicing and intron retention in ribosomal proteins". Rather, building off of earlier work with Cus2, and absent any data suggesting otherwise, we expect that a sufficiently powered study would reveal a defect in the splicing of a broad collection of genome-wide targets.

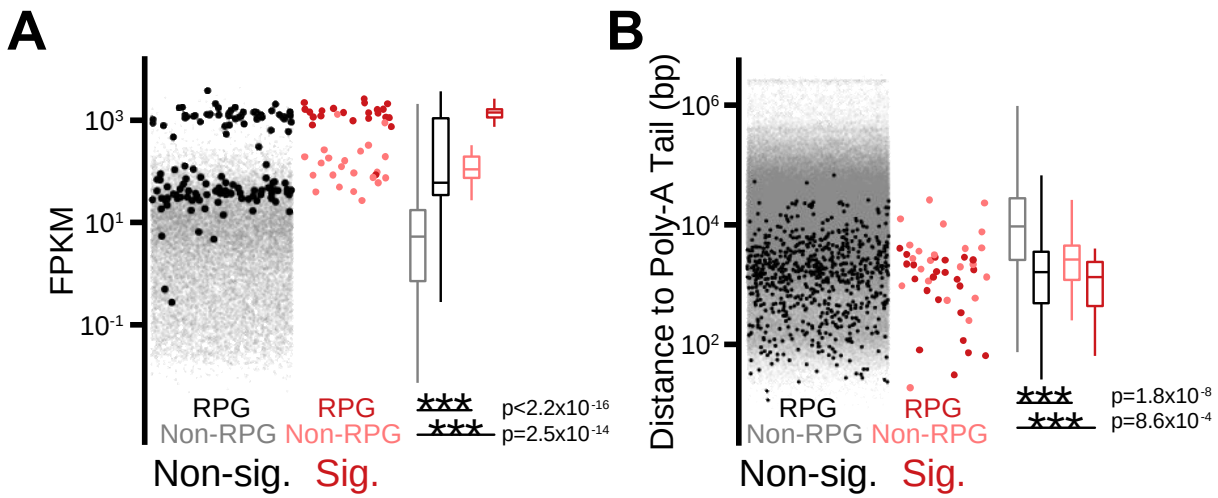


Figure 2: Selection bias in identified intron retention events. (**A**) Expression (measured in Fragments per Kilobase per Million Reads) of all introns versus those identified as retained broken out by ribosomal protein genes (RPG) and non-ribosomal protein genes (Non-RPG). (**B**) Distance from intron to poly-A Tail for all introns versus those identified as retained, broken out by RPG and Non-RPG. Two sided Mann-Whitney test performed to determine significance. For all boxplots: Median is represented by the center line, box limits represent the 25th and 75th quartiles, and whiskers are 1.5x the interquartile range.

Concluding remarks

Here we argue that selection bias not only can have deleterious impacts on RNA-seq based studies of pre-mRNA splicing, but in fact has had and will continue to have such impacts unless and until knowledge of this problem and its potential solutions becomes widely appreciated. Indeed, while this problem has been previously described

in the context of standard gene expression studies, a cursory examination of the literature nevertheless highlights the extent to which this problem continues to infect current work. Regretfully, whereas Oshlack provided an elegant mathematical approach for mitigating the impacts of selection bias in GO-term enrichment analyses, we offer no such solution here. While we hope that others might provide such a solution in the future, we note that the simplest solution to this problem for now is the use of approaches that either enrich for, or otherwise increase the number of, splicing-informative reads across the complement of genomic substrates, thereby reducing the differences in sampling across the dataset. Importantly, while all of the work examined here involved short-read, Illumina-based sequencing, we note that this is not a problem unique to this platform, but rather reflects a fundamental statistical challenge associated with analyzing datasets with small numbers of measured events. As such, we expect that this problem will be even greater in analyses of datasets from long-read sequencing platforms where the number of reads per experiment is typically much lower, and likewise in single-cell experiments where the number of reads per cell is dramatically reduced; indeed, evidence of such bias in single-cell experiments has been recently demonstrated (Buen Abad Najar et al., 2020). Finally, we note that as ‘users’ of these technologies, whether that be that as experimentalists generating and analyzing such data, or as ‘consumers’ evaluating the work of others, it will be essential to consider the possibility that an apparently statistically-significant conclusion may not reflect a meaningful biological property but rather a meaningless technical artifact.

Methods

Cell Growth

Wild-type cells and those harboring the prp2-1 allele were streaked from glycerol stocks onto solid rich media (YPD) and grown at the permissive temperature (25˚C) for three days. In triplicate, single colonies were inoculated into 5 mL of YPD and grown at 25˚C with shaking at 200 rpm overnight. Cultures were backdiluted into 20 mL of YPD to an OD600 of 0.05 and incubated at 25˚C with shaking at 200 rpm. Upon reaching an OD600 of approximately 0.75, cultures were transferred to a 37˚C shaking water bath (200 rpm) for 10 minutes. Cells were collected via vacuum filtration and pellets were immediately flash-frozen in liquid nitrogen and stored at -80˚C.

MPE-seq Library Preparation

RNA was purified and MPE-seq libraries were prepared as previously described (Gildea et al., 2019) with the exception that biotin-11-dUTP was used in place of aminoallyl-dUTP during reverse transcription such that no separate biotin coupling step was necessary. Instead, following hydroxide treatment an elution volume of 50 µL was used during the zymo column clean-up which went immediately into the first bead purification.

Downsampling

Using a custom script, each read was assigned a random number from 0 to 1 and sorted based on their random number. The top 800,000, 400,000, and 200,000 were included in the high, medium, and low datasets, respectfully. Random number generation used a seed value of 1 to allow future reproducibility.

Alignment and Quantification

The full and downsampled datasets were processed as follows: Reads were trimmed of sequencing adapters using fastp (Chen et al., 2018) with the following parameters:

```
--adapter_sequence CTGTCTCTTATACACATCT
--adapter_sequence_r2 CTGTCTCTTATACACATCT
```

Trimmed reads were aligned to the R64-2-1 genome release from SGD with hisat2(Kim et al., 2019) with the following parameters:

```
 --max-intronlen 2000 --no-unal
```

and reads with MAPQ scores below 5 were removed with samtools(Li et al., 2009). Unspliced and spliced counts were obtained with a custom script based on HTSeq-count (Anders et al., 2015). DESeq2 (Love et al., 2014) was used to assess differential splicing.

Corsini et al. Data Processing

FPKM values for host genes and identified retained introns were obtained from Corsini *et al. (*GEO: GSM2535498 and Table S2, respectfully*).* All mm9 UCSC introns were broken into groups based on whether they were identified as significant and whether they are ribosomal protein genes. Additionally, the distance from the end of each distinct (as determined by their chromosome, start, and stop positions) intron and the end of the host transcript was calculated. In the case that an intron existed within multiple transcripts, the shortest transcript was used.

Code Availability

Code for basic analysis steps is available at https://github.com/zdwyer/Problem-of-Selection-Bias.

Data Availability

All sequencing data are available through NCBI's Gene Expression Omnibus (GEO) at accession number GSE160046.

Competing Interests Statement

The authors have no competing interests to declare.

Acknowledgements

We thank members of the Pleiss lab for critical feedback on this work. We thank P. Schweitzer and the BRC Genomics Facility at Cornell for outstanding technical support with Illumina sequencing. This work was funded by NIH grant R01GM098634 (to J.A.P.).

Works Cited

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Antman, K., Amato, D., Wood, W., Carson, J., Suit, H., Proppe, K., Carey, R., Greenberger, J., Wilson, R., & Frei, E. (1985). Selection bias in clinical trials. *Journal of Clinical Oncology*, *3*(8), 1142–1147. https://doi.org/10.1200/JCO.1985.3.8.1142

Buen Abad Najar, C. F., Yosef, N., & Lareau, L. F. (2020). Coverage-dependent bias creates the appearance of binary splicing in single cells. *ELife*, *9*. https://doi.org/10.7554/eLife.54603

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Corsini, N. S., Peer, A. M., Moeseneder, P., Roiuk, M., Burkard, T. R., Theussl, H.-C., Moll, I., & Knoblich, J. A. (2018). Coordinated Control of mRNA and rRNA Processing Controls Embryonic Stem Cell Pluripotency and Differentiation. *Cell Stem Cell*, *22*(4), 543-558.e12. https://doi.org/10.1016/j.stem.2018.03.002

Gildea, M. A., Dwyer, Z. W., & Pleiss, J. A. (2019). Multiplexed primer extension sequencing: A targeted RNA-seq method that enables high-precision quantitation of mRNA splicing isoforms and rare pre-mRNA splicing intermediates. *Methods (San Diego, Calif.)*. https://doi.org/10.1016/j.ymeth.2019.05.013

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Oshlack, A., & Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, *4*, 14. https://doi.org/10.1186/1745-6150-4-14

Rodgers, M. L., Tretbar, U. S., Dehaven, A., Alwan, A. A., Luo, G., Mast, H. M., & Hoskins, A. A. (2016). Conformational dynamics of stem II of the U2 snRNA. *RNA (New York, N.Y.)*, *22*(2), 225–236. https://doi.org/10.1261/rna.052233.115

Sharma, E., & Blencowe, B. J. (2018). Orchestrating Ribosomal Subunit Coordination to Control Stem Cell Fate. *Cell Stem Cell*, *22*(4), 471–473. https://doi.org/10.1016/j.stem.2018.03.019

Xu, H., Fair, B. J., Dwyer, Z. W., Gildea, M., & Pleiss, J. A. (2019). Detection of splice isoforms and rare intermediates using multiplexed primer extension sequencing. *Nature Methods*, *16*(1), 55–58. https://doi.org/10.1038/s41592-018-0258-x

Yan, D., Perriman, R., Igel, H., Howe, K. J., Neville, M., & Ares, M. (1998). CUS2, a yeast homolog of human Tat-SF1, rescues function of misfolded U2 through an unusual RNA recognition motif. *Molecular and Cellular Biology*, *18*(9), 5000–5009. https://doi.org/10.1128/mcb.18.9.5000

Young, M. D., Wakefield, M. J., Smyth, G. K., & Oshlack, A. (2010). Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biology*, *11*(2), R14. https://doi.org/10.1186/gb-2010-11-2-r14

Chapter III: Characterization of cancer-related mutations in the pre-mRNA splicing pathway revealed by high-throughput screening for alternative splice variants in *S. pombe*

Alternative citation:

Dwyer ZW*, Fair BJ*, Larson A, Pleiss JA. Characterization of cancer-related mutations in the pre-mRNA splicing pathway revealed by high-throughput screening for alternative splice variants in *S. pombe.* (In Prep)
*Denotes equal contribution

Abstract

Pre-mRNA splicing, a highly conserved process from unicellular yeasts to humans, is an essential regulator of eukaryotic gene expression. Here we present a sequencing based forward genetic screen to identify alleles that impact exon skipping in the yeast species *Schizosaccharomyces pombe*, an organism which shares many aspects of the splicing pathway with higher eukaryotes. We queried about two thousand temperature sensitive *S. pombe* strains to identify those which harbor mutations that increase the exon skipping in the *pwi1* transcript. We identified 36 candidate strains, including a pair of strains with different alleles in the prp10 gene that had different splicing defect signatures. Further genome-wide characterization of the splicing defects led to the identification of other *S. pombe* transcripts that are sensitized to exon skipping in the prp10-E407K allele and share a common set of features. Profiling of these sensitive transcripts identified a common set of features: A strong branchpoint sequence in the intron upstream of the skipped exon and a weak 5' splice site in the intron downstream of the skipped exon.

Introduction

The coding sequence of most eukaryotic genes is not continuous, but rather segments of coding sequence are interrupted by non-coding sequences known as introns (Berget et al., 1977). Prior to translation, introns are removed from the mRNA transcript in a process known as splicing which is carried out by the spliceosome, a huge macro-molecular machine consisting of five small nuclear ribonucleic proteins (snRNPs) and scores of auxiliary factors, which must assemble anew upon each premature transcript in a stepwise manner dependent on three conserved sequence elements: the 5′ and 3′ splice sites which are at the beginning and end of the intron respectively, and the branch point which lies a few nucleotides upstream of the 3′ splice site (reviewed in Will & Lührmann, 2011; Matera & Wang, 2014). Spliceosome assembly begins with the binding of the U1 snRNP to the 5′ splice site via complementary base pairing between the 5′ splice site and the U1 snRNA. In a similar fashion, the U2 snRNP binds to the branch point sequence through complementary base pairing between the branch point sequence and the U2 snRNA. These two recognition events play a crucial role in the splicing reaction as they dictate the exact location of splicing activity and enable further assembly of the spliceosome. Once assembled, the spliceosome catalyzes two transesterification reactions, excising the intron and splicing together the two coding sequences.

The complexity of splice site selection is increased by alternative splicing events where variant 5′ or 3′ splice sites are activated upon in place of the constitutive sites. In the higher eukaryotes this mainly occurs through exon skipping events, where in a multi-intronic gene, the 5′ splice site of an upstream intron is used along with the 3′ splice site of a downstream intron, thereby excising the upstream intron, downstream

intron, and the intervening exon(s). Understanding how the spliceosome regulates alternative splicing is of utmost importance as RNA-seq datasets estimate that as many as 95% of human genes undergo alternative splicing events (Pan et al., 2008) and about 15% of human disease causing point mutations involve defects in splicing (Faustino & Cooper, 2003), furthermore, alterations to alternative splicing are a hallmark of many human cancers (Makishima et al., 2012).

Historically, genetic screens in the yeast species *Saccharomyces cerevisiae* have played an important role in the identification of genes in the splicing pathway (Hartwell et al., 1970; Vijayraghavan et al., 1989; Noble & Guthrie, 1996; Hossain & Johnson, 2014). Many of such screens rely on temperature-sensitive alleles of splicing genes which are functional at a permissive temperature (often 25˚C) but no longer viable at non-permissive temperatures (either higher or lower than the permissive temperature). Once identified, these conditional alleles are useful tools in understanding the biochemical processes within the splicing pathway (Lin et al., 1987; Libri et al., 2001). However, *S. cerevisiae* can have limitations to understanding the splicing pathway as their intronic landscape has undergone a severe reduction relative to higher eukaryotes, with many splicing proteins no longer existing in the *S. cerevisiae* genome. A distantly related fission yeast, *Schizosaccharomyces pombe*, serves as a strong model organism to better understand the splicing pathway as it shares the genetic tractability of *S. cerevisiae*, while it retains the bulk of the splicing pathway components seen in higher eukaryotes (Kuhn & Käufer, 2003; Fair & Pleiss, 2017). Previous work from our lab screens a haploid non-essential gene-deletion library in *S. pombe* (Larson et al., 2016) to identify genes that play a role in the splicing pathway. However, as most

known genes in the fission yeast splicing pathway are essential (D Kim et al., 2010), work highlighted in Appendix III extends this screen to a library of temperature sensitive *S. pombe* strains to identify essential proteins that contribute to the splicing pathway as well as provide a toolkit of interesting splicing alleles that can be used for detailed biochemical analysis of the pathway. Work in this chapter looks specifically at proteins that are involved in alternative splicing, which although is rare in *S. pombe*, has been demonstrated to exist in specific transcripts *(Awan et al., 2013; Stepankiw et al., 2015)*, one such event is an environmentally regulated skipping event in the *pwi1* transcript.

Results

Identifying temperature sensitive alleles which increase exon skipping

In order to identify the regulatory mechanisms by which exon skipping is permitted in S. pombe, we screened a library of about two thousand randomly mutagenized S. pombe strains for those that harbor mutations that lead to an increase in exon skipping of the *pwi1* transcript with a similar method as outlined in Larson et al. (2016) and highlighted in Appendix 3. Briefly, to quantitatively measure exon skipping, following a temperature shift for 15 minutes, cDNA from biological replicates of each individual strain was used as a template for a PCR reaction using primers sitting in exon 1 and exon 3 of the *pwi1* transcript enabling amplification of both constitutively spliced transcripts as well as those that had skipping of the second exon. In a second PCR reaction, individual barcodes were appended to each strain and libraries were subjected to deep sequencing. For each strain, the skip index, a measurement that increases with the prevalence of exon skipping events, was calculated as defined in the methods section. Because most mutations would not be expected to affect the splicing pathway,

we identified strains whose skip index was statistically different than the average adjusted for read depth (Figure 1A). In addition to the *pwi1* transcript, in order to identify mutations that were leading to exon skipping specifically, not just those that leave the splicing pathway in complete disarray, we also screened against *fkh1*, an "average" intron based on properties like splice site sequences, intron length, etc. For this target, primers sat in exon 3 and exon 4, monitoring the splicing of the third intron. In this screen, the splice index is calculated, a similar metric to the skip index, where an increase is indicative of a splicing defect (Figure 1B).

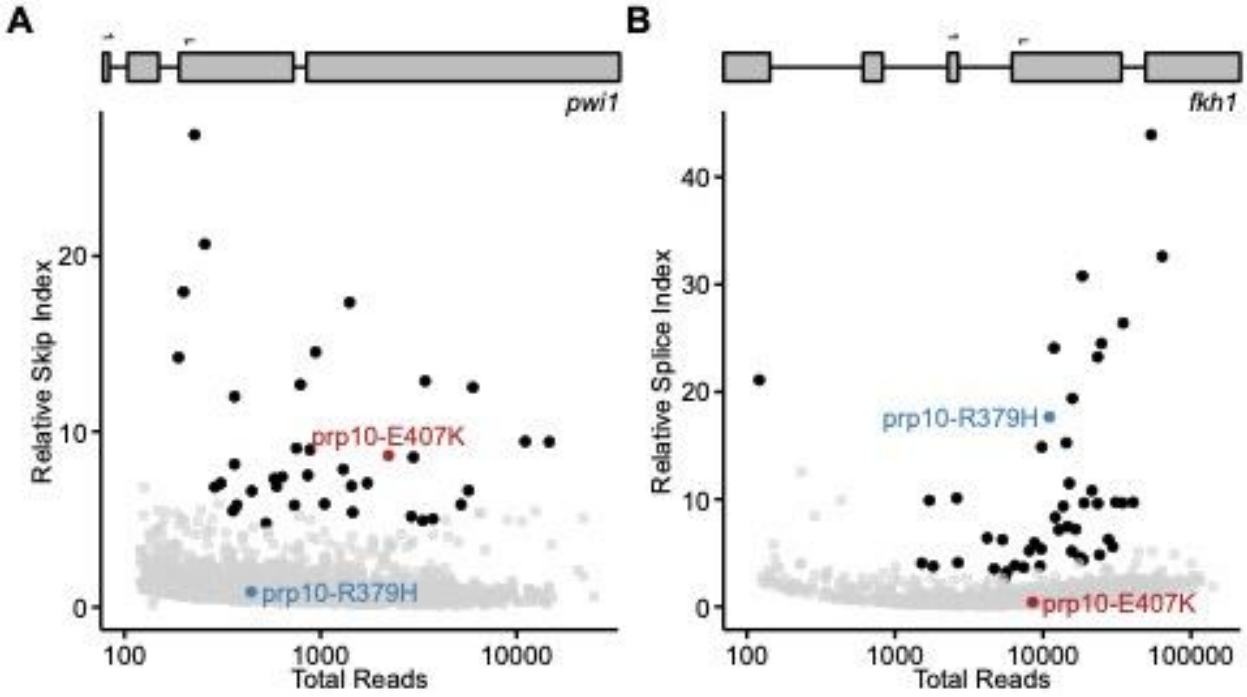


Figure 1: Screen for splice-altering mutations in a library of temperature sensitive mutant strains (A) The relative skip index for each strain as a function of total counts for that strain when screening against the *pwi1* transcript. **(B)** The relative splice index for each strain as a function of total counts for that strain when screening against the *fkh1* transcript. Arrows on gene bodies mark primer locations during the first PCR.

Overall, we identified 38 strains which had an increase in exon skipping, of which 36 did not have a defect in the splicing of *fkh1*. In conjunction with the screens highlighted in Appendix III, a handful of candidate strains were selected for whole

genome sequencing so the causative mutation could be inferred from identified SNPs and validated experimentally (see appendix III Table SI). Of note, we identified a pair of strains that each had a mutation in *prp10*, one which caused an increase in exon skipping without affecting the splicing of *fkh1*, prp10-E407K, while the other led to a splicing defect in *fkh1* without affecting exon skipping, prp10-R379H (Figure 1). This protein is of interest as it is the *S. pombe* ortholog to human SF3B1, mutations in which are highly associated with a variety of blood borne cancers such as Myelodysplastic syndromes (MDS) and Chronic Lymphocytic Leukemia (CLL) (Papaemmanuil et al., 2011; Wang et al., 2011; Yoshida et al., 2011; Quesada et al., 2012). Mutations that disrupt SF3B1's role in branch point recognition lead to alternative splicing (Cretu et al., 2016; Kesarwani et al., 2017)

High resolution characterization of genome-wide splicing defects

To extend the characterization of the two mutants to a genome-wide level, we prepared triplicate MPE-seq (Xu et al., 2019; Gildea et al., 2019) libraries for both candidate *prp10* mutant strains as well as a matched wildtype using reverse transcription primers targeting a representative subset of introns within the *S. pombe* genome. For each intron, a splice index was calculated as the ratio of unspliced to spliced read counts and then compared to the corresponding splice index from the wildtype strain. Here an increase in relative splice index is indicative of a splicing defect in the mutant strain. In the prp10-E407K strain we demonstrate that most splicing events do not show a statistically significant splicing defect relative to wildtype (Figure 2A) while in the prp10-R379H strain the majority of splicing events are disrupted relative

to wildtype (Figure 2B), consistent with the hypothesis that prp10-R379H causes a general splicing defect.

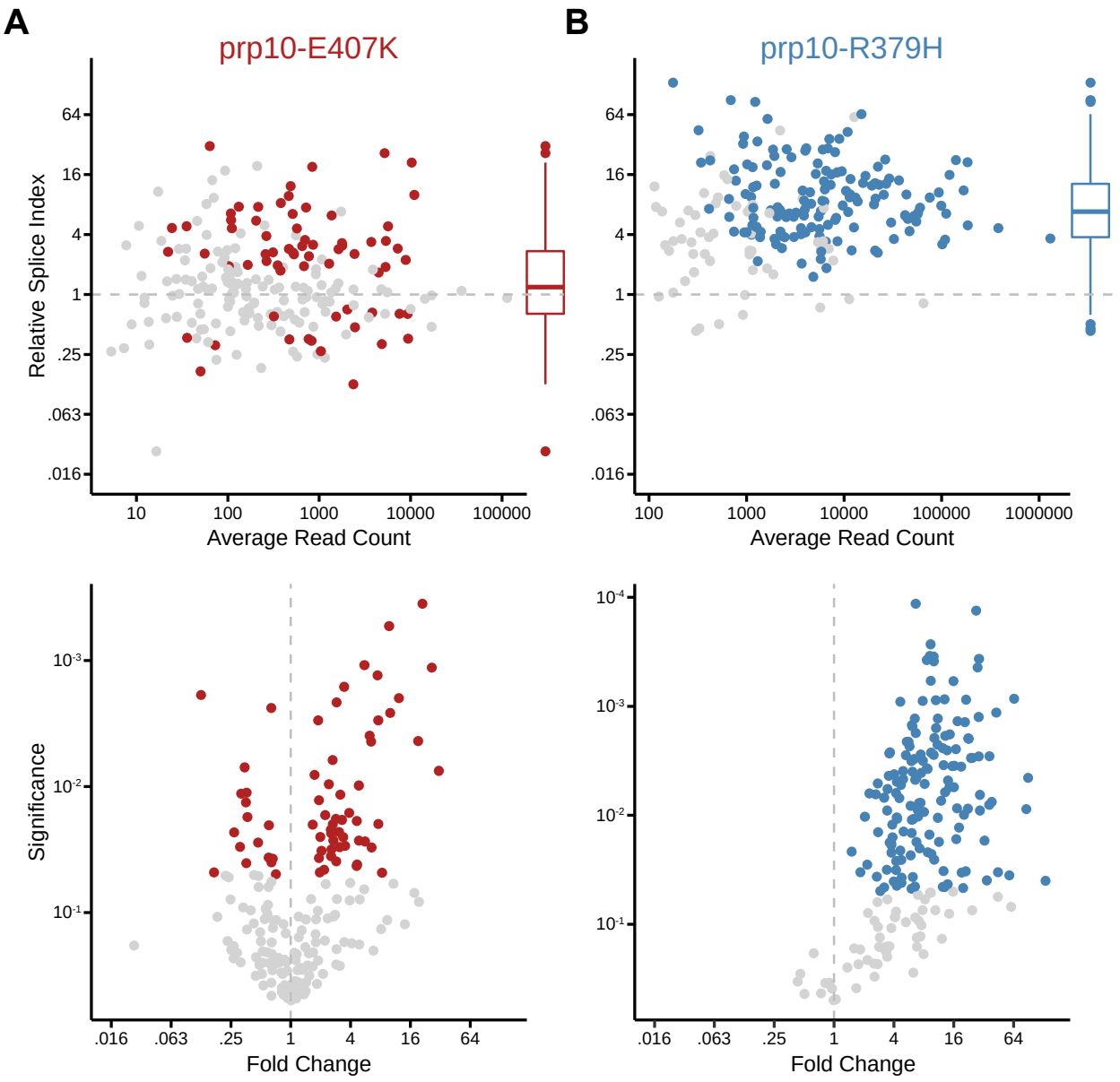


Figure 2: Quantitative measurement of genome-wide splicing relative to wildtype in mutant strains. Splicing index relative to wildtype as a function of read depth (upper) and corresponding significance (lower) of change for (**A**) prp10-E407K (**B**) prp10-R379H. Colored points notate a statistically significant difference from wildtype as measured by a student t-test on the log transformation of splice index evaluated at a threshold of .05.

In addition to looking at changes in splice index, we performed a more qualitative analysis by looking at which junctions were activated upon. Here, for each intron, counts

for every junction were tabulated and compared to junction activation in wildtype with DESeq2 (Love et al., 2014). This analysis is sensitive to both changes in transcript expression as well as changes in the use of specific junctions. In the prp10-E407K strain, in addition to a few modest changes in expression of constitutive events, there is also a statistically significant increase in four exon skipping events (Figure 3A). Conversely, in prp10-R379H, we do not see any changes in junction utilization beyond changes in expression of host genes (Figure 3B).

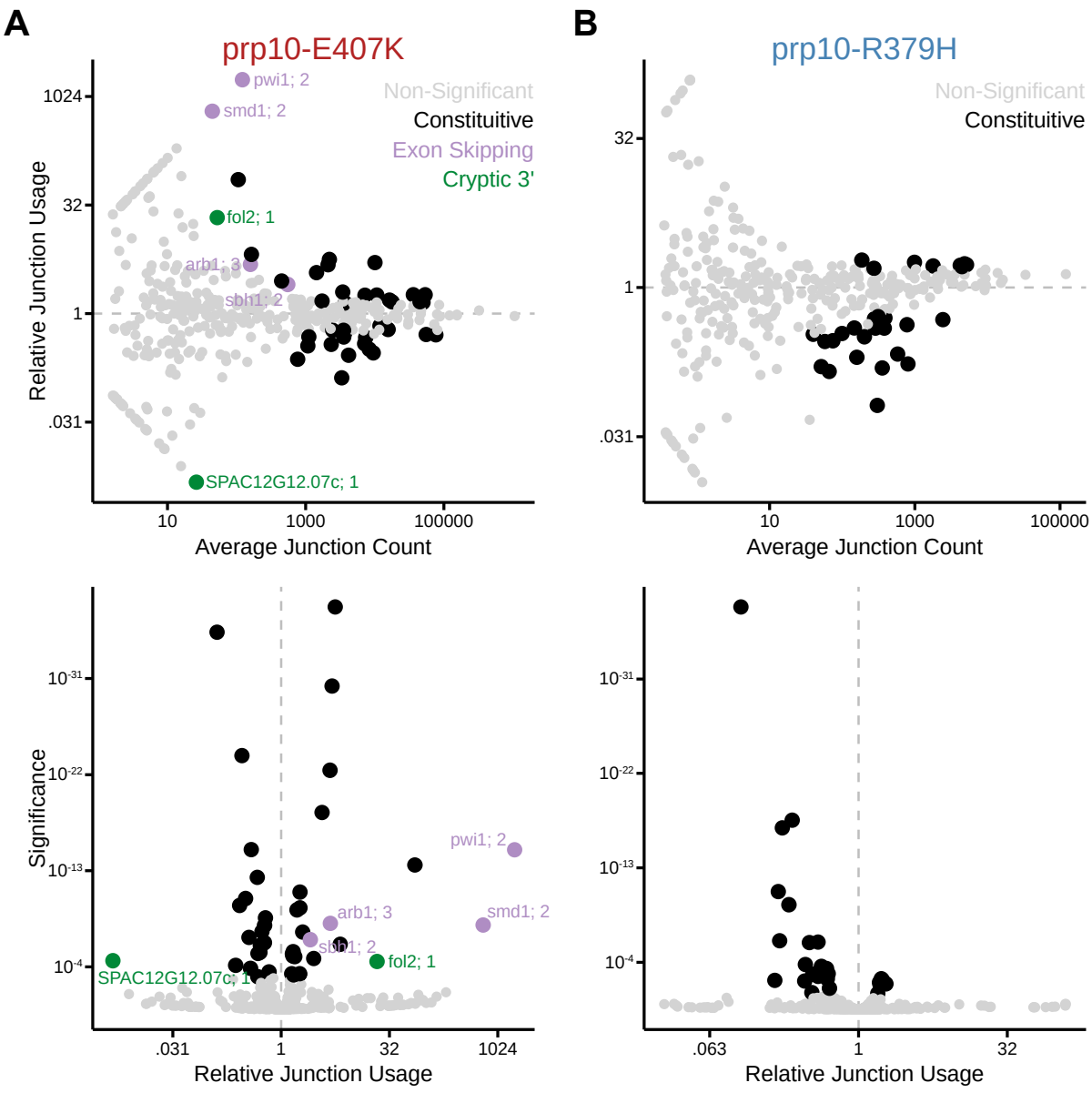


Figure 3: Qualitative measurement of genome-wide splicing relative to wildtype in mutant strains. Skipping index relative to wildtype as a function of read depth (upper) and corresponding significance (lower) of change for (**A**) prp10-E407K (**B**) prp10-R379H. Colored points notate a statistically significant difference from wildtype as measured by DESeq2 (Love et al., 2014) at a significance level of .05. For exon skipping events the number indicates the skipped exon. For cryptic 3′ splice sites the number indicates the affected intron.

Taken together, the relative splicing (Figure 2) and junction utilization changes (Figure 3) demonstrate that prp10-E407K leads to a transcript-specific splicing defect including an increase in exon skipping events while prp10-R379H leads to a general splicing defect in the splicing of all introns.

Determining shared properties of skipped exons

To better understand what it is about the prp10-E407K mutation that allows for an increase in exon skipping we looked at properties that were shared by the four specific exon skipping events (pwi1, arb1, sbh1, smd1) identified in the analysis of junction utilization. Of note, relative to the genome-wide distributions, three of the four sensitized skipping events (pwi1, arb1, sbh1) came from transcripts where there was a strong branch point in the intron upstream of the skipped exon (Figure 4A) and weak 5′ splice site in the intron downstream of the skipped exon (Figure 4B).

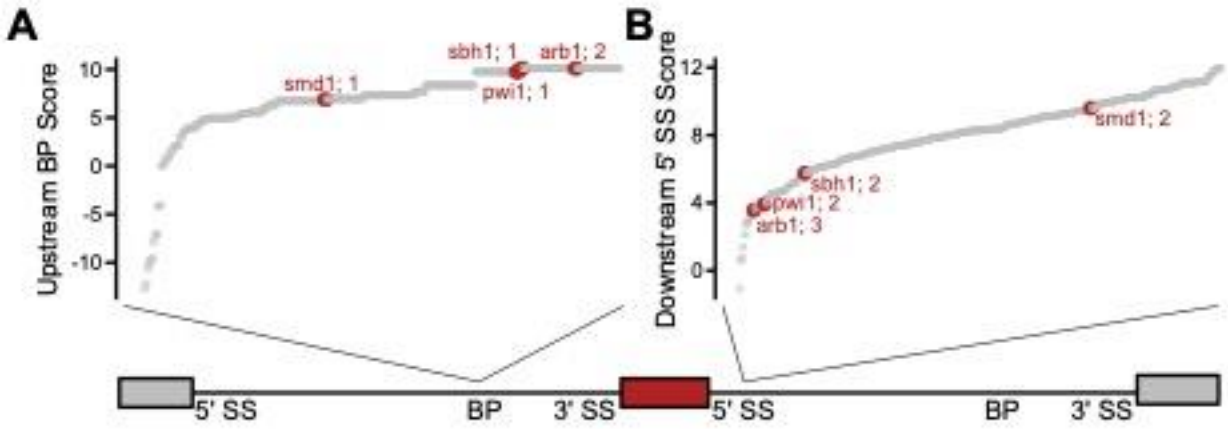


Figure 4: Properties of representative set of introns (**A**) Branch point scores for all introns targeted in MPE-seq experiment. (**B**) 5′ splice site scores for all introns targeted in MPE-seq experiment. Previously described exon skipping events are highlighted in red where the number represents the specific intron within the transcript. Splice site scores are from Barrass et al. (2014).

Discussion

Here we identified 36 temperature sensitive strains which led to an increase in exon skipping of the *pwi1* transcript including a mutation in prp10, the *S. pombe* ortholog to SF3B1, a protein which is known to lead to alternative splicing changes causing various human blood borne cancers. Further characterization of this mutation demonstrates that it leads to a transcript specific splicing defect and increased exon skipping in four splicing events. Extrapolating from the 5% of the splicing events in the

genome that we sampled, we expect there to be on the order of a hundred potential exon skipping events that arise from this mutation. We demonstrate that the identified transcripts that are sensitized to exon skipping, share a common property of having a strong branch point in the intron upstream of the skipped exon and a weak 5′ splice site in the intron downstream of the skipped exon. While the latter is consistent with the failure to recognize the downstream 5′ splice site (which would make it more likely for the upstream 5′ splice site to be activated upon with the downstream 3′ splice site) however, the former is inconsistent with exon skipping. Stronger upstream branch point sequences should be more likely to be activated upon by the spliceosome. Ongoing work aims to better understand the relationship between the prp10-E407K mutation and exon skipping by strengthening the 5' splice site of the intron downstream of the skipped exon. Ultimately a better understanding of prp10's modulation of splice site activation (or more importantly it's misactivation) serve as an important step in the discovery of therapeutic treatments to human cancers arising from the SF3B1 mutation.

Methods

Sequencing Preparation for Screening Library for Splicing Defects

Cell Growth

An arrayed library of 2,304 temperature sensitive strains originally created via nitrosoguanidine mutagenesis was obtained from J. Armstrong (Armstrong et al., 2007) and re-arrayed and stocked in six 384-well plates. Strains were grown and processed with liquid handling robotic protocols as previously described (Larson et al., 2016) with the following exceptions: Strains were pinned from initial glycerol stocks onto solid rich media (YES) and grown at the permissive temperature (25˚C) for three days. Cells were

then pinned into liquid YES and grown to saturation for three days to normalize cell density. Cultures were backdiluted to an OD_{600} of approximately 0.1 and grown for about 12 hours to an OD_{600} of approximately 0.8.

Temperature Shifting

Cells were shifted to the non-permissive temperature (37˚C) for 15 minutes by mixing 100 µL of culture into 100 µL of pre-warmed (45˚C) YES, followed by incubation (37˚C) with shaking for 15 minutes prior to cell collection via centrifugation as previously described in (Larson et al., 2016).

Reverse Transcription

RNA was isolated from cell pellets and reverse transcription was primed with random ninemers as described in (Albulescu et al., 2012).

RT-PCR and Sequencing

An initial PCR reaction with gene and plate-specific primers was performed as follows: 10 mM TRIS pH8.3, 50 mM KCl, 1.5 mM $MgCl_2$, .2 mM dNTPs, .25x SYBR Green I, 250 nM forward and reverse primer, 150 nM Hot-Start aptamer (Noma et al., 2006), and 1x Taq polymerase. The reactions were pooled (preserving well position) into a single 384-well plate and cleaned via glass-fiber column by the addition of two volumes of DNA Binding Buffer (5M Guanidinium HCl, 30% Isopropanol, 90 mM KOH, 150 mM acetic acid), mixing, addition to column, and centrifugation (2 minutes, 2000g). Two sequential washes (80% ethanol, 10 mM TRIS) and a dry spin proceeded elution in water with the original volume. Elutions were diluted five-fold in water and used to seed a second PCR reaction to append on sequencing adapters. This reaction was completed as above. Final libraries were pooled per target, concentrated via ethanol

precipitation, cleaned on a Zymo-25 column as described above, eluted in water, and sequenced on the Illumina platform.

Data Processing for Screening Library for Splicing Defects

Read Processing

The first 11 bases (contain plate barcode) and sequencing adapters were moved to the read information line of the sequencing fastq file using a custom script.

Trimming

Processed reads were trimmed using fastp (Chen et al., 2018) with the following parameters:

```
--adapter_sequence CTGTCTCTTATACACATCT
```

Alignment

Trimmed reads were aligned to a custom genome containing the full genomic sequences of the target transcripts obtained from the 2020-11-01 Pombase release using hisat2 (Daehwan Kim et al., 2019) with the following parameters:

```
--max-intronlen 2000 --no-unal --rna-strandness RF
```

And filtered with samtools (Li et al., 2009) for reads with a mapping score (MAPQ) score above 5.

Feature Counting

In the case of pwi1, junction utilization was determined using a custom script that tabulates the boundaries of each junction identified by the aligner for each strain independently (as determined by the plate, row, and column barcodes) and characterizes each junction as either being constitutively spliced (exon 2 joined to exon 3) or exon skipped (exon 1 joined to exon 3).

Calculation of Relative Skip Index

The skip index was independently calculated for each replicate of every strain:

$$\text{Skip Index} = \frac{\text{Exon 1 to Exon 3 Junction Count}}{\text{Exon 2 to Exon 3 Junction Count}}$$

Since most strains were not expected to harbor mutations that alter exon skipping, the plate median of each 384-well plate was assumed to be representative of wild-type. For each strain the relative skip index was calculated as:

$$\text{Relative Skip Index} = \frac{\text{Skip Index}}{\text{Skip Index}_{\text{Plate Median}}}$$

The calculated relative skip indexes were log-normally distributed, and their precision was a function of read depth. Therefore, to determine which strains exhibited a skipping defect: First, the relative skip index of the two replicates were averaged to create a single relative skip index for each strain. The dataset was then divided into ten equally sized bins based on read count and the mean ($\mu_{\text{interpolated}}$) and standard deviation ($\sigma_{\text{interpolated}}$) of the average relative skip index was calculated for each bin. Finally, a z-score was calculated for each strain comparing the average relative skipping index to the interpolated mean and standard deviation for that bin:

$$z = \frac{\log_2(\text{average relative skip index}) - \log_2(\mu_{\text{interpolated}})}{\log_2(\sigma_{\text{interpolated}})}$$

A one-sided z-test was performed, and strains were considered significant if they had a p-value below .01.

Calculation of Relative Splice Index

The splice index was independently calculated for each replicate of every strain:

$$\text{Splice Index} = \frac{\text{premature count}}{\text{mature count}}$$

The exact processing for the calculation of the relative skip index was repeated on these skip indexes.

MPE-seq Library Preparation for prp10 Mutant Strains

Cell Growth

Strains were streaked out from initial glycerol stocks onto solid rich media (YES) and grown at the permissive temperature (25˚C) for three days. A single colony was transferred into liquid YES and grown overnight at the permissive temperature. Cultures were backdiluted to an OD_{600} of approximately 0.05 and grown to an OD_{600} of approximately 0.80.

Temperature Shifting

Cultures were transferred to a shaking water bath at 37˚C for 15 minutes prior to collection with vacuum filtration. Pellets were flash frozen in liquid nitrogen and stored at -80˚C.

Library Preparation

Total RNA was extracted via Phenol:Chloroform extraction. To fragment the RNA, $ZnCl_2$ and TRIS-HCl pH 7.0 were each added to a final concentration of 10 mM and incubated for 10 minutes at 65˚C. Fragmentation was stopped by addition of EGTA to a final concentration of 50 mM and fragmented RNA was cleaned up via ethanol precipitation. MPE-seq libraries were prepared as outlined in Gildea et al. (2020) with the following modification: biotin-dUTP was used in place of aminoallyl-dUTP thus no biotin coupling was necessary. No size selection was performed.

Data Processing for MPE-seq of prp10 Mutant Strains

Trimming

Reads were trimmed using fastp (Chen et al., 2018) with the following parameters:

```
-w 1 --umi --umi_loc=read1 --umi_len=10
--adapter_sequence CTGTCTCTTATACACATCT
--adapter_sequence_r2 CTGTCTCTTATACACATCT
```

Alignment

Trimmed reads were aligned to the 2020-11-01 Pombase release of the *S. pombe* genome using hisat2 (Daehwan Kim et al., 2019) with the following parameters:

```
--max-intronlen 2000 --no-unal --rna-strandness RF
```

And filtered with samtools (Li et al., 2009) for reads with a mapping score (MAPQ) score above 5.

Feature Counting

Premature and mature reads were counted with a custom script based on HTSeq-count (Anders et al., 2015).

Junction Counting

Junctions were counted using a custom script which counts the number of times each specific junction start and stop were used and characterizes them as constitutive, exon skipping, cryptic 5′ splice site, cryptic 3′ splice site, or unannotated.

Calculation of Relative Skip Index

The splice index was defined as:

$$\text{splice index} = \frac{\text{premate counts}}{\text{mature counts}}$$

Splice indexes were log normally distributed. To assess significance, a two-sample students t-test was performed on the replicates and evaluated at a significance level of .05.

Calculation of Relative Junction Usage

Changes in junction usage were assessed by DESeq2 (Love et al., 2014) using raw counts and assessed as significant at an multiple hypothesis corrected level of .001.

Acknowledgements and Author Contributions

JA and NB donated the library of temperature sensitive strains. AL, BJF, ZWD, and JAP conceived and designed experiments. BJF and ZWD performed experiments. ZWD analyzed data.

Works Cited

Albulescu, L., Sabet, N., Gudipati, M., Stepankiw, N., Bergman, Z. J., Huffaker, T. C., & Pleiss, J. A. (2012). A Quantitative, High-Throughput Reverse Genetic Screen Reveals Novel Connections between Pre–mRNA Splicing and 5′ and 3′ End Transcript Determinants. *PLoS Genetics*, *8*(3), e1002530. https://doi.org/10.1371/journal.pgen.1002530

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Armstrong, J., Bone, N., Dodgson, J., & Beck, T. (2007). The role and aims of the FYSSION project. *Briefings in Functional Genomics and Proteomics*, *6*(1), 3–7. https://doi.org/10.1093/bfgp/elm004

Awan, A. R., Manfredo, A., & Pleiss, J. A. (2013). Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. *Proceedings of the National Academy of Sciences*, *110*(31), 12762–12767. https://doi.org/10.1073/pnas.1218353110

Berget, S. M., Moore, C., & Sharp, P. A. (1977). Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences*, *74*(8), 3171–3175. https://doi.org/10.1073/pnas.74.8.3171

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Cretu, C., Schmitzová, J., Ponce-Salvatierra, A., Dybkov, O., De Laurentiis, E. I., Sharma, K., Will, C. L., Urlaub, H., Lührmann, R., & Pena, V. (2016). Molecular Architecture of SF3b and Structural Consequences of Its Cancer-Related Mutations. *Molecular Cell*, *64*(2), 307–319. https://doi.org/10.1016/j.molcel.2016.08.036

Fair, B. J., & Pleiss, J. A. (2017). The power of fission: Yeast as a tool for understanding complex splicing. *Current Genetics*, *63*(3), 375–380. https://doi.org/10.1007/s00294-016-0647-6

Faustino, N. A., & Cooper, T. A. (2003). Pre-mRNA splicing and human disease. *Genes & Development*, *17*(4), 419–437. https://doi.org/10.1101/gad.1048803

Gildea, M. A., Dwyer, Z. W., & Pleiss, J. A. (2019). Multiplexed primer extension sequencing: A targeted RNA-seq method that enables high-precision quantitation of mRNA splicing isoforms and rare pre-mRNA splicing intermediates. *Methods*, *18*, 30383–30389. https://doi.org/10.1016/j.ymeth.2019.05.013

Hartwell, L. H., McLaughlin, C. S., & Warner, J. R. (1970). Identification of ten genes that control ribosome formation in yeast. *Molecular and General Genetics MGG*, *109*(1), 42–56. https://doi.org/10.1007/BF00334045

Hossain, M. A., & Johnson, T. L. (2014). Using Yeast Genetics to Study Splicing Mechanisms. In K. J. Hertel (Ed.), *Spliceosomal Pre-mRNA Splicing* (Vol. 1126, pp. 285–298). Humana Press. https://doi.org/10.1007/978-1-62703-980-2_21

Kesarwani, A. K., Ramirez, O., Gupta, A. K., Yang, X., Murthy, T., Minella, A. C., & Pillai, M. M. (2017). Cancer-associated SF3B1 mutants recognize otherwise

inaccessible cryptic 3′ splice sites within RNA secondary structures. *Oncogene*, *36*(8), 1123–1133. https://doi.org/10.1038/onc.2016.279

Kim, D, Hayles, J., Kim, D., Wood, V., Park, H.-O., Won, M., Yoo, H.-S., Duhig, T., Nam, M., Palmer, G., Han, S., Jeffery, L., Baek, S.-T., Lee, H., Shim, Y. S., Lee, M., Kim, L., Heo, K.-S., Noh, E. J., … Hoe, K.-L. (2010). Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. *Nature Biotechnology*, *28*(6), 617–623. https://doi.org/10.1038/nbt.1628

Kim, Daehwan, Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Kuhn, A. N., & Käufer, N. F. (2003). Pre-mRNA splicing in Schizosaccharomyces pombe: Regulatory role of a kinase conserved from fission yeast to mammals. *Current Genetics*, *42*(5), 241–251. https://doi.org/10.1007/s00294-002-0355-2

Larson, A., Fair, B. J., & Pleiss, J. A. (2016). Interconnections Between RNA-Processing Pathways Revealed by a Sequencing-Based Genetic Screen for Pre-mRNA Splicing Mutants in Fission Yeast. *Genes, Genomes, Genetics*, *6*(6), 1513–1523. https://doi.org/10.1534/g3.116.027508

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Libri, D., Graziani, N., Saguez, C., & Boulay, J. (2001). Multiple roles for the yeast SUB2/yUAP56 gene in splicing. *Genes & Development*, *15*(1), 36–41. https://doi.org/10.1101/gad.852101

Lin, R.-J., Lustig, A. J., & Abelson, J. (1987). Splicing of yeast nuclear pre-mRNA in viro requires a functional 40S spliceosome and several extrinsic factors. *Genes & Development*, *1*, 7–18.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Makishima, H., Visconte, V., Sakaguchi, H., Jankowska, A. M., Abu Kar, S., Jerez, A., Przychodzen, B., Bupathi, M., Guinta, K., Afable, M. G., Sekeres, M. A., Padgett, R. A., Tiu, R. V., & Maciejewski, J. P. (2012). Mutations in the spliceosome machinery, a novel and ubiquitous pathway in leukemogenesis. *Blood*, *119*(14), 3203–3210. https://doi.org/10.1182/blood-2011-12-399774

Matera, A. G., & Wang, Z. (2014). A day in the life of the spliceosome. *Nature Reviews Molecular Cell Biology*, *15*(2), 108–121. https://doi.org/10.1038/nrm3742

Noble, S. M., & Guthrie, C. (1996). Identification of Novel Genes Required for Yeast Pre-mRNA Splicing by Means of Cold-Sensitive Mutations. *Genetics*, *143*, 67–80.

Noma, T., Sode, K., & Ikebukuro, K. (2006). Characterization and application of aptamers for Taq DNA polymerase selected using an evolution-mimicking algorithm. *Biotechnology Letters*, *28*(23), 1939–1944. https://doi.org/10.1007/s10529-006-9178-4

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, *40*(12), 1413–1415. https://doi.org/10.1038/ng.259

Papaemmanuil, E., Cazzola, M., Boultwood, J., Malcovati, L., Vyas, P., Bowen, D., Pellagatti, A., Wainscoat, J. S., Hellstrom-Lindberg, E., Gambacorti-Passerini, C., Godfrey, A. L., Rapado, I., Cvejic, A., Rance, R., McGee, C., Ellis, P., Mudie, L. J., Stephens, P. J., McLaren, S., … Campbell, P. J. (2011). Somatic SF3B1 Mutation in Myelodysplasia with Ring Sideroblasts. *New England Journal of Medicine*, *365*(15), 1384–1395. https://doi.org/10.1056/NEJMoa1103283

Quesada, V., Conde, L., Villamor, N., Ordóñez, G. R., Jares, P., Bassaganyas, L., Ramsay, A. J., Beà, S., Pinyol, M., Martínez-Trillos, A., López-Guerra, M., Colomer, D., Navarro, A., Baumann, T., Aymerich, M., Rozman, M., Delgado, J., Giné, E., Hernández, J. M., … López-Otín, C. (2012). Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nature Genetics*, *44*(1), 47–52. https://doi.org/10.1038/ng.1032

Stepankiw, N., Raghavan, M., Fogarty, E. A., Grimson, A., & Pleiss, J. A. (2015). Widespread alternative and aberrant splicing revealed by lariat sequencing. *Nucleic Acids Research*, *43*(17), 8488–8501. https://doi.org/10.1093/nar/gkv763

Vijayraghavan, U., Company, M., & Abelson, J. (1989). Isolation and characterization of pre-mRNA splicing mutants of Saccharomyces cerevisiae. *Genes & Development*, *3*(8), 1206–1216. https://doi.org/10.1101/gad.3.8.1206

Wang, L., Lawrence, M. S., Wan, Y., Stojanov, P., Sougnez, C., Stevenson, K., Werner, L., Sivachenko, A., DeLuca, D. S., Zhang, L., Zhang, W., Vartanov, A. R., Fernandes, S. M., Goldstein, N. R., Folco, E. G., Cibulskis, K., Tesar, B., Sievers, Q. L., Shefler, E., … Wu, C. J. (2011). SF3B1 and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia. *New England Journal of Medicine*, *365*(26), 2497–2506. https://doi.org/10.1056/NEJMoa1109016

Will, C. L., & Lührmann, R. (2011). Spliceosome Structure and Function. *Cold Spring Harbor Protocols*, 1–23. https://doi.org/10.1101/cshperspect.a003707

Xu, H., Fair, B. J., Dwyer, Z. W., Gildea, M., & Pleiss, J. A. (2019). Detection of splice isoforms and rare intermediates using multiplexed primer extension sequencing. *Nature Methods*, *16*(1), 55–58. https://doi.org/10.1038/s41592-018-0258-x

Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-Otsubo, A., Kon, A., Nagasaki, M., Chalkidis, G., Suzuki, Y., Shiosaka, M., Kawahata, R., Yamaguchi, T., Otsu, M., Obara, N., Sakata-Yanagimoto, M., Ishiyama, K., … Ogawa, S. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*, *478*(7367), 64–69. https://doi.org/10.1038/nature10496

Chapter IV: Identification and characterization of mutations in library of thousands of temperature sensitive *S. pombe* strains

Alternative citation:

Dwyer ZW, Fair BJ, Larson A, Armstrong J, Bone N, Gallert B, Pleiss JA. Identification and characterization of mutations in library of thousands of temperature sensitive *S. pombe* strains. *In prep*

Abstract

Unicellular yeasts have long provided a power laboratory system for understanding eukaryotic genomes. Forward genetic screens in yeast have provided insights into virtually all pathways. Historically, the rate limiting step has been the identification of mutations that causally linked to a molecular phenotype. Rapid advancements in genome sequencing has enabled the capacity to solve this problem in a high-throughput way. Here we provide whole genome sequencing and mutational analysis for approximately 2000 temperature sensitive strains of *Schizosaccharomyces pombe.*

Introduction

Conditional mutations have provided as a way to determine the function of specific genes, especially those that are essential. A common form of conditional mutations is the creation of alleles that are functional at a permissive temperature but become inactive as the temperature shifts either upwards or downwards to a non-permissive temperature. This allows quick perturbation of a protein's function, in an otherwise normal environment.

The fission yeast, *Schizosaccharomyces pombe* are a good model organism for studying many aspects of eukaryotic gene regulation. While they maintain many of the pathways and proteins that exist in higher eukaryotes (Moreno et al., 1991; Kuhn & Käufer, 2003; Fair & Pleiss, 2017), they are easily genetically manipulatable and many strains can be simultaneous grown and manipulated with low special requirements.

A conditional library of temperature sensitive *Schizosaccharomyces pombe* was generated by N-methyl-N'-nitro-N-nitrosoguanidine (nitrosoguanidine) mutagenesis Beáta Gallert and Paul Nurse as outlined in the methods section. Here we describe the simultaneous production of whole genome sequencing libraries and identification of mutations in thousands of temperature sensitive *S. pombe* strains.

Results and Discussion

Parallel generation of Whole Genome Sequencing libraires

In order to prepare Whole Genome Sequencing libraries on the complete collection of temperature sensitive *S. pombe* strains we took advantage of previously developed liquid handling robot assisted methods (Albulescu et al., 2012) to collect individual cell pellets for each strain arrayed out in a collection of multi-well plates. Genomic DNA was collected using a modified version of the MasterPure Yeast DNA Purification Kit (Epicentre) and tagmentation libraries were prepared as described in the methods.

Initial sequencing to assess library quality

As an initial test of library quality and representation of the strains, we sequenced the full library at a level far below what is required to call SNPs (Figure 1).

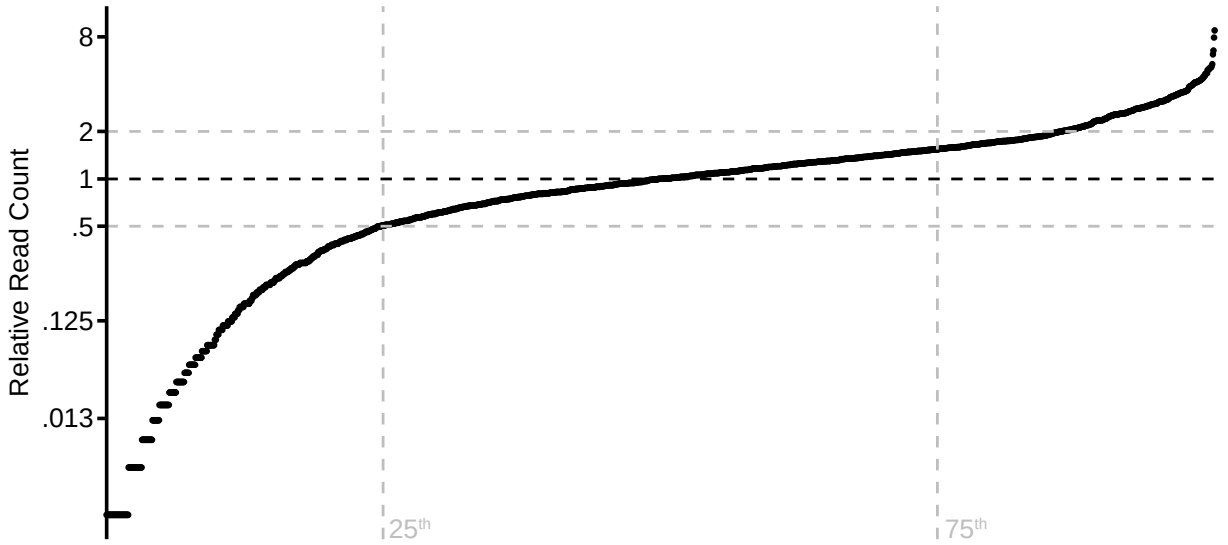


Figure 1: Relative read count per strain. Normalized to the median read count. Vertical lines represent the 25th and 75th percentile. Horizontal lines show two-fold change in either direction around the median.

In this initial test, we found that about 61% of strains were within a two-fold change from the median level, and that the vast majority of strains are within eight-fold of the median. Although this is a reasonable range given the difficulty in ensuring the same amount of genomic DNA went into each tagmentation reaction, in order to maximize sequencing cost efficiency, we will repool using the following scheme: All strains within a two-fold change in either direction of the median will be combined based on their relative amount. On either side of that window, the process will continue, with the creation of groups where all strains are within two-fold of the group's median (or a maximum of a four-fold change). This four-fold window was selected as it allows volumes that will be between 5 µL and 20 µL, which work well with the liquid handling robot. The repooled groups will then be combined equally based molarity as measured by digital droplet using a primer / probe set that amplifies all sequenceable material.

Mitochondria derived reads are overrepresented

In the course of this initial test, unexpectedly, we learned that roughly 78% of the reads came from the mitochondria. This significantly increases the depth that each strain must be sequenced at to get adequate coverage for SNP calling. In order to reduce the number of mitochondrial reads within the library, we will use Depletion of Abundant Sequences by Hybridization (DASH) (Gu et al., 2016), an *in vitro* approach which harnesses Cas9 to cleave sequencing libraries at user-defined locations. Using our initial sequencing as a model, we selected the 60 potential Cas9 guide RNA targets sites which will disrupt the maximal number of mitochondria derived molecules (Figure 2A). At full DASH efficiency we estimate that we can remove upwards of 96% of mitochondrial reads.

In order to evaluate the efficacy of DASH on our library, we DASHed the originally sequenced library at all 60 DASH targets and then performed qPCR with a pair of primers flanking one specific DASH guide RNA target (Figure 2B) as well as a pair of primers targeting a specific genomic locus as a loading control. Following DASH less than 7% of the mitochondrial reads that crossed the guide RNA region were remaining. Assuming all guide RNAs cleave with the same efficiency, we estimate that we can convert the libraries to about 80% chromosome derived reads.

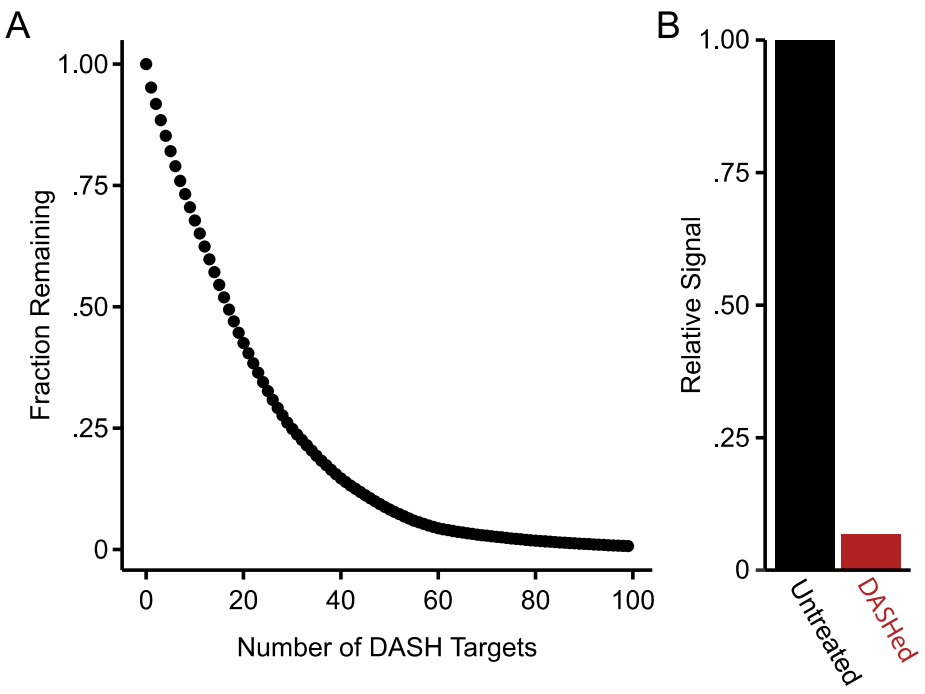


Figure 2: DASH reduces mitochondrial reads (**A**) Theoretical fraction of mitochondrial reads remaining depending on number of guide RNAs used in DASH (**B**) Post DASH signal at a specific mitochondrial loci normalized to untreated library.

Mutations are equally distributed throughout the genome

Although we don't have enough sequencing depth to call SNPs across the library, we have sequenced a handful of individual strains that were candidates from the screens outlined in Chapter III and Appendix III. By using these 26 strains as a proxy for the entire library, we can extrapolate some properties of the overall library. An important consideration is that these 26 strains were not selected randomly, rather they are a biased set towards strains that have splicing defects. Across the 26 strains, the median number of mutations per stain was 35 (Figure 2A). If all the mutations are pooled from the 26 proxy strains, we see a fairly random distribution of mutations across the chromosomes (Figure 2B). With 35 mutations evenly distributed mutations per stain and approximately 2000 strains in the library, we expect to have roughly upwards of 80,000 total mutations which, if evenly distributed would give us a mutation about every 150

bases. This would correspond to about 13 mutations per gene using the average gene length.

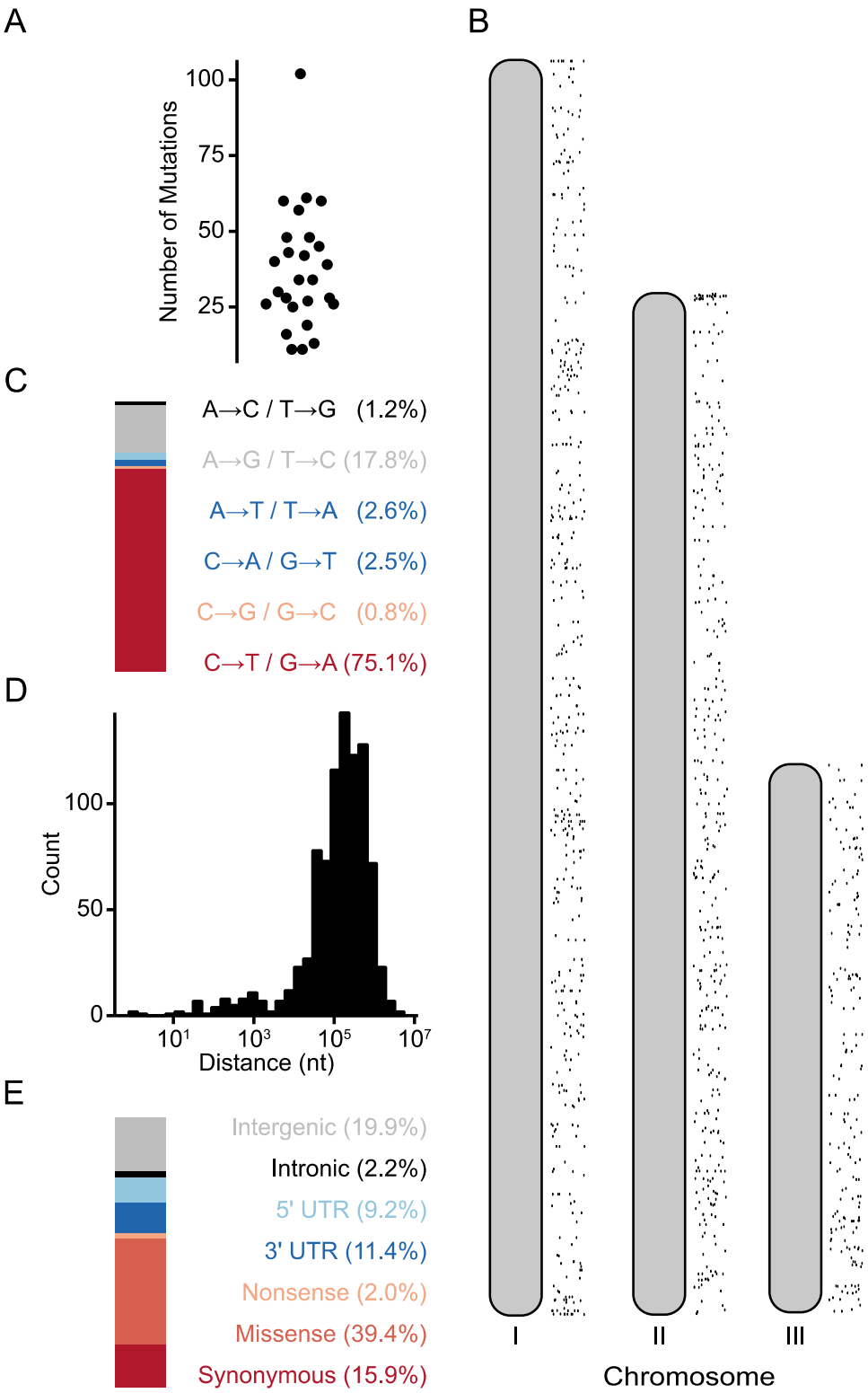


Figure 3: Properties of mutations found in subset of sequenced strains (**A**) Number of mutations identified in each strain. (**B**) Location of mutations across all sequenced strains. (**C**) Signature of mutations at DNA level (**D**) Histogram of distance (in nucleotides) between all consecutive mutations within each strain. (**E**) Consequences of mutations across all sequenced strains.

Nitrosoguanidine mutagenesis has a signature

When looking at the DNA mutations that occurred (Figure 3B), we see a large bias towards $\mathrm{C} \rightarrow \mathrm{G}$ or $\mathrm{G} \rightarrow \mathrm{A}$ mutations (which we cannot distinguish between since we do not know on which strand the original mutation occurred). Nitrosoguanidine imparts its mutagenic capabilities by affecting replication forks, so sometimes the same replication fork will receive multiple mutations, causing multiple mutations within a small window. We can detect this when looking at the distance between mutations within the same strain (Figure 3D), we see a bimodal distribution with a major peak around 100,000 with a smaller peak around 1,000.

Mutations lead to coding changes

Of the mutations identified in the 26 strains identified 39.4% led to coding changes (Figure 3E). These are the most likely cause of the temperature sensitive phenotype, but an additional 40% of mutations are found within UTRs or between genes that have regulatory potential.

Methods

Library Preparation

Library Creation

Temperature sensitive strains were mutated by Beáta Gallert using nitrosoguanidine as previously outlined (Gallert & Nurse, 1997; Moreno et al., 1991). Strains were arrayed out and sent to use by John Armstrong (Armstrong et al., 2007)

Cell Growth

Cells were pinned from glycerol stocks onto solid rich media (YES) and incubated for three days at the permissive temperature (25˚C). Colonies were repined

onto solid YES media to normalize colony size and incubated for three days at 25˚C. Following the second round of incubation, colonies were pinned into 150 µL YES in 384-well plates and grown with shaking at 900 rpm at 25˚C for three days to saturation.

Genomic DNA Extraction

Cells were collected via centrifugation and the supernatant was discarded. DNA was extracted with a liquid handling robot using the Epicentre Yeast DNA isolation kit with the following modifications to accommodate for scale: Cells were resuspended in 40 µL lysis solution and incubated at 65˚C for 30 minutes before being cooled on ice. 20 µL protein precipitation solution was added, and then the lysate was centrifuged for 25 minutes at 5000g. 45 µL of the supernatant was transferred into 50 µL isopropanol and mixed via repeat pipetting. DNA was pelleted with 25 minutes of centrifugation at 5000g and the supernatant was removed. Pellets were washed once with 100 µL 70% ethanol and then resuspended in 50 µL 0.1x TE (1 mM TRIS-HCl pH 8.0, 100 µM EDTA).

Tagmentation

5 µL of genomic DNA (about 50 ng on average) was added to 20 µL 1.25x Tagmentation Buffer (12.5 mM TRIS HCl pH 7.5, 6.25 mM $MgCl_2$, 12.5% DMF) and incubated for eight minutes at 55˚C. The reaction was stopped by the addition of 2.5 µL 1% SDS and stored at -20˚C.

Amplification

To reduce the inhibition of SDS on PCR, prior to amplification, 6.67 µL of tagmented genomic DNA was added to 33.3 µL mH_2O and 5 µL of this diluted mixture was used as template in a 15 µL PCR reaction with the following final concentrations: 1x Phusion Buffer (Thermo), dNTPs (200 µM each), forward and reverse primer (500 nM

each), and 1x Phusion (purified in-house). The reaction was vortexed briefly and spun down before being amplified with the following conditions:

1 cycle:	72˚C	3 mintues
	98˚C	30 seconds
15 cycles:	98˚C	10 seconds
	63˚C	30 seconds
	72˚C	3 minutes

Pooling

5 µL of each reaction was diluted into 95 µL 0.1x TE (1 mM TRIS-HCl pH 8.0, 100 µM EDTA). 20 µL of diluted reaction from all 2,304 strains was combined into a single tube and concentrated via ethanol precipitation with a final resuspension in 1 mL of .1x TE (1 mM TRIS-HCl pH 8.0, 100 µM EDTA)

Size Selection

Fragments of 500 nucleotides and larger were selected for with Ampure XP beads following manufacturer's protocol by the addition of 500 µL of beads to the 1 mL of pooled library with a final elution in 100 µL of .1x TE (1 mM TRIS-HCl pH 8.0, 100 µM EDTA).

Data Processing

Individual fastq files were received for each individual sample and processed as follows:

Trimming

Fastq files were trimmed of illumina adapters using fastp (Chen et al., 2018) with the following parameters:

```
--adapter_sequence CCGAGCCCACGAGAC
--adapter_sequence_r2 GACGCTGCCGACGA
```

Alignment

Trimmed reads were aligned to the 2020-11-01 Pombase release of the *S. pombe* genome using hisat2 (Kim et al., 2019) with the following parameters:

```
--maxins 3000 --no-spliced-alignment --rg-id <sample>
--rg ID:<sample> --rg LB:20181202 --rg PL:illumina
--rg SM:<sample> --rg PU:<sample>
```

And filtered with samtools (Li et al., 2009) for reads with a mapping score (MAPQ) above 5.

SNP Calling

SNPs were called using following the “best practices” outlines by GATK (DePristo et al., 2011) using the HaplotypeCaller program with the following parameters:

```
-ploidy 1
```

Works Cited

Albulescu, L., Sabet, N., Gudipati, M., Stepankiw, N., Bergman, Z. J., Huffaker, T. C., & Pleiss, J. A. (2012). A Quantitative, High-Throughput Reverse Genetic Screen Reveals Novel Connections between Pre–mRNA Splicing and 5′ and 3′ End Transcript Determinants. *PLoS Genetics*, *8*(3), e1002530. https://doi.org/10.1371/journal.pgen.1002530

Armstrong, J., Bone, N., Dodgson, J., & Beck, T. (2007). The role and aims of the FYSSION project. *Briefings in Functional Genomics and Proteomics*, *6*(1), 3–7. https://doi.org/10.1093/bfgp/elm004

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Fair, B. J., & Pleiss, J. A. (2017). The power of fission: Yeast as a tool for understanding complex splicing. *Current Genetics*, *63*(3), 375–380. https://doi.org/10.1007/s00294-016-0647-6

Gallert, B., & Nurse, P. (1997). An approach to identify functional homologues and suppressors of genes in fission yeast. *Current Genetics*, *32*(1), 27–31. https://doi.org/10.1007/s002940050244

Gu, W., Crawford, E. D., O'Donovan, B. D., Wilson, M. R., Chow, E. D., Retallack, H., & DeRisi, J. L. (2016). Depletion of Abundant Sequences by Hybridization (DASH): Using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biology*, *17*(1), 41. https://doi.org/10.1186/s13059-016-0904-5

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Kuhn, A. N., & Käufer, N. F. (2003). Pre-mRNA splicing in Schizosaccharomyces pombe: Regulatory role of a kinase conserved from fission yeast to mammals. *Current Genetics*, *42*(5), 241–251. https://doi.org/10.1007/s00294-002-0355-2

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Moreno, S., Klar, A., & Nurse, P. (1991). Molecular Genetic Analysis of Fission Yeast Schizosaccharomyces pombe. *Methods in Enzymology*, *174*, 795–823.

Appendix I: Detection of splice isoforms and rare intermediates using multiplexed primer extension sequencing

Alternative citation:

Xu, H., Fair, B. J., Dwyer, Z. W., Gildea, M., & Pleiss, J. A. (2019). Detection of splice isoforms and rare intermediates using multiplexed primer extension sequencing. *Nature Methods*, *16*(1), 55–58. https://doi.org/10.1038/s41592-018-0258-x

Abstract

Targeted RNA sequencing (RNA-seq) aims to focus coverage on areas of interest that are inadequately sampled in standard RNA-seq experiments. Here we present multiplexed primer extension sequencing (MPE-seq), an approach for targeted RNA-seq that uses complex pools of reverse-transcription primers to enable sequencing enrichment at user-selected locations across the genome. We targeted hundreds to thousands of pre-mRNA splice junctions and obtained high-precision detection of splice isoforms, including rare pre-mRNA splicing intermediates.

Main

Identification of the small subset of RNA-seq reads that span exon–exon junctions in transcripts has allowed the unambiguous detection of vast numbers of novel splice isoforms in scores of organisms (Barbosa-Morais et al., 2012; Merkin et al., 2012). Yet in spite of the power of this approach, the sequencing depth necessary to quantitatively detect many splicing events is substantially higher than what most experiments generate. Although this limitation of whole-transcriptome profiling has been addressed in part by methods that use antisense probes (Mercer et al., 2012, 2014) or

PCR enrichment (Blomquist et al., 2013) to target sequencing coverage to genomic regions of interest, a deeper understanding of the basic mechanisms by which splicing is regulated, and of the pathological consequences of its misregulation, will be facilitated by methods that enable the detection of splicing states within cells with higher resolution and greater precision. Toward this end, we have designed MPE-seq, a targeted sequencing method based on primer extension that improves splice-junction detection and allows for resolution of splicing intermediates. We demonstrate the ability to multiplex hundreds to thousands of primer extension assays and evaluate the products by deep sequencing (Fig. 1a). User-selected primers are extended to generate complementary DNA (cDNA) during a reverse-transcription reaction, thus enabling the user to target RNA regions of interest. The use of elevated temperatures during reverse transcription minimizes nonspecific primer annealing (Supplementary Fig. 1), and each primer is appended with a next-generation sequencing adaptor and a unique molecular identifier (Kivioja et al., 2012). A strand-extension step similar to template switching (Zhu et al., 2001) appends the second sequencing adaptor onto the 3' terminus of each cDNA molecule. Coupling this approach with paired-end sequencing allows for simultaneous querying of the 5' and 3' ends of the cDNAs from targeted regions (full details are presented in the Methods).

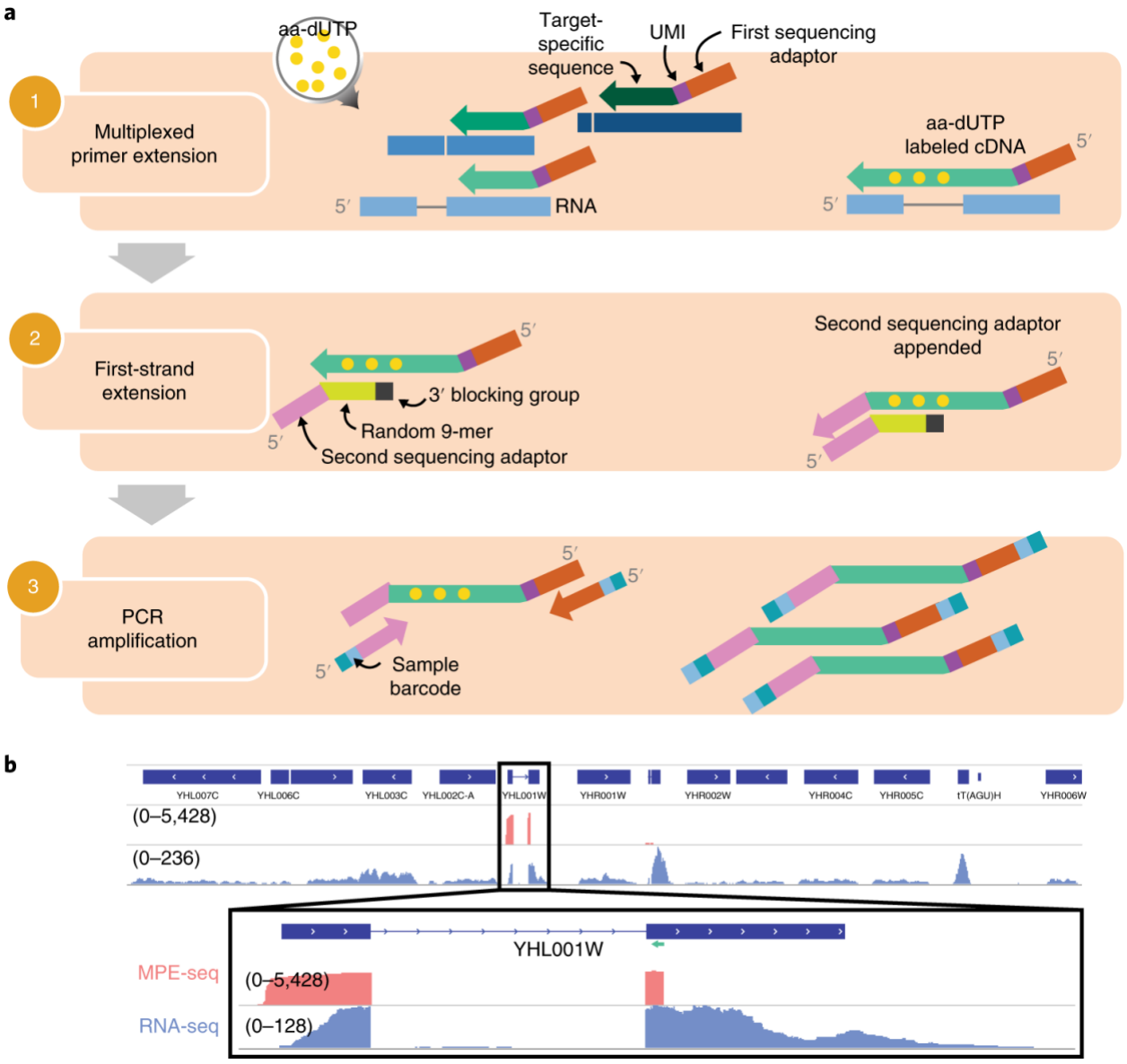


Fig. 1: MPE-seq uses complex pools of reverse-transcription primers to target sequencing to regions of interest. (a) Outline of the MPE-seq protocol. UMI, unique molecular identifier. **(b)** Genome browser screenshot of a targeted region in MPE-seq (pink) and conventional RNA-seq (purple). The location of a targeting primer is indicated by a green arrow.

As an initial demonstration of MPE-seq, we examined pre-mRNA splicing in the budding yeast *Saccharomyces cerevisae*. For each of the 309 annotated introns in the yeast genome, primers were systematically designed within a 50-nt window immediately downstream of the 3' splice site, ensuring that short extensions would cross splice junctions. Primers were pooled at equimolar concentration, and MPE-seq libraries were

generated with total cellular RNA from wild-type yeast and sequenced to a depth of only ~5 million reads. As a comparative reference, we generated conventional RNA-seq libraries using poly(A)-selected RNA and sequenced them to ~40 million reads. Whereas the conventional RNA-seq libraries yielded read coverage that comprised full gene bodies across the transcriptome, MPE-seq coverage was focused on the selected genes, precisely targeted to the regions upstream of the designed primers (Fig. 1b). Just over 75% of sequenced fragments from MPE-seq mapped to targeted regions (Supplementary Fig. 2, Supplementary Table 1 (not included)), resulting on average in >100-fold enrichment in sequencing depth at these regions compared with that obtained with RNA-seq (Fig. 2a, Supplementary Fig. 3). Although the fold enrichment varied on a target-by-target basis, it was similar across transcripts with a wide range of expression levels (Supplementary Fig. 3). From these data, we extrapolate that a standard RNA-seq experiment would require ~500 million sequencing reads to achieve a level of coverage over the targeted regions similar to what these 5 million MPE-seq reads provided. Given the increased read depth achieved over targeted regions with MPE-seq, we asked how well unspliced isoforms were sampled. Measurements of the fraction of unspliced messages from replicate libraries obtained with MPE-seq showed superior internal reproducibility compared with that in the larger, replicate RNA-seq libraries (Fig. 2b), probably reflecting the sampling noise associated with RNA-seq data with reduced sequencing depth over the targeted regions. Moreover, although MPE-seq is not amenable to de novo discovery of novel splicing events across the entire genome, it did allow for the identification of scores of rare, previously unannotated splicing events at the targeted regions (Supplementary Table 2 (not included)). Nevertheless, although

MPE-seq provided increased sensitivity and reproducibility of splicing measurements, estimates of the unspliced fraction determined from MPE-seq in a wild-type strain only modestly correlated with those determined by RNA-seq (Supplementary Fig. 4a,b). Notably, this correlation improved when we compared these techniques' measurement of changes in splicing between samples assayed by the same methodology (Supplementary Fig. 4c), presumably reflecting inherent technical biases (Zheng et al., 2011) present in one or both approaches that are internally well controlled.

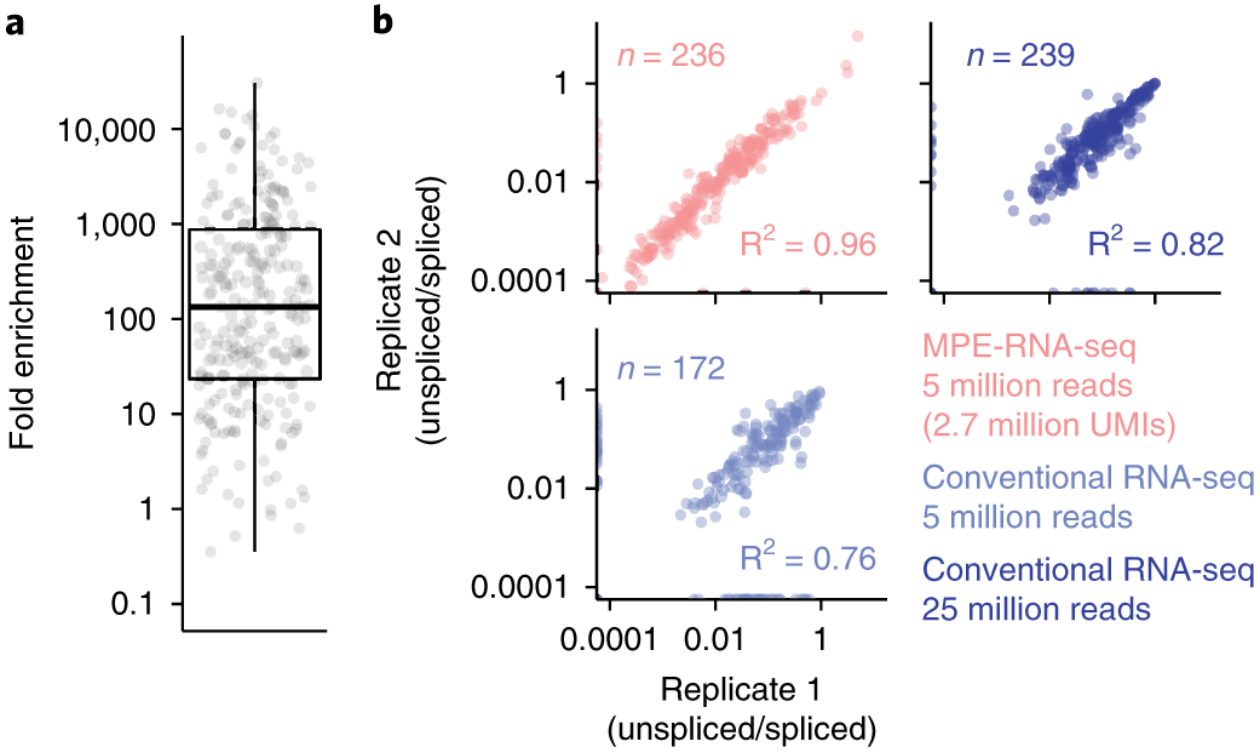


Fig. 2: MPE-seq enrichment allows high-precision measurements of splicing. (a) Each point represents the fold enrichment of a target region in MPE-seq compared with values for conventional RNA-seq. In the box plot, the center line represents the 50th percentile, and lower and upper hinges represent the 25th and 75th percentiles, respectively. Whiskers end at the 0th and 100th percentiles. *n* = 249 target regions that were detected with at least one read in both RNA-seq and MPE-seq libraries for comparison. **(b)** Scatter plots depicting intron-retention measurements in replicate libraries made from biologically independent samples in MPE-seq and conventional RNA-seq at matched or greater read depth. Pearson correlation coefficients (R^2) are indicated. *n* is the number of quantified intron-retention events, with at least one spliced read and one unspliced read required in both experiments.

We next sought to determine whether we could detect splicing intermediates with MPE-seq. Primer extension reactions, which can reveal the locations of reverse-

transcription stops, have historically been used to map a variety of biological features such as transcription start sites (Carey et al., 2013) and the locations of branch sites within the lariat intermediate (LI) species of the pre-mRNA splicing reaction (Coombes, 2005; Padgett et al., 1985) (Supplementary Fig. 5a). Our approach anticipated the possibility of mapping the 3' ends of the cDNA molecules, and indeed we found in our MPE-seq libraries that the 3' ends of many cDNAs accumulated at the transcription start sites, as determined by an orthologous method (Booth et al., 2016) (Supplementary Fig. 6), indicating that reverse transcription generally proceeded to the 5' terminus of the RNA. We also observed many cDNAs that terminated at or near the annotated branch-point motifs in introns, with decreased read coverage upstream of the motifs, consistent with the inability of reverse transcriptase to read past the branched adenosine in the LI (Fig. 3a,b). This drop in read coverage was not apparent in MPE-seq libraries generated from a strain that harbored a conditional mutation in Prp2, an RNA helicase required for catalysis of the first step of splicing (Kim & Lin, 1996), thus corroborating that these cDNAs originate from LIs. We note that these LI-derived cDNAs often contained at the 3' terminus a unique signature of mismatches incorporated by reverse transcriptase at the branched adenosine (Supplementary Fig. 7), which may serve as a tag for de novo identification of branch sites in organisms with less well-annotated branch sites (Chen et al., 2018) . The ability of MPE-seq to differentiate between unspliced isoforms allowed us to estimate that ~10% of unspliced pre-mRNAs are of the LI form genome-wide under steady-state conditions (Supplementary Fig. 5b, Supplementary Table 3 (not included), Methods), albeit with considerable variation between individual pre-mRNAs (Fig. 3b,c). Although we identified correlations between transcript- and intron-level

features and the abundances of these species (Supplementary Fig. 8), none of these correlations held when we considered the abundance of pre-first-step RNA relative to LIs, a metric that we expect would reflect variation between the relative catalytic rates of the first and second steps of splicing. A more complete understanding of the determinants of in vivo splicing efficiency will require kinetic measurements of the individual steps of splicing, rather than the steady-state levels measured here. The ability of MPE-seq to robustly distinguish these splice isoforms provides an opportunity to do just this.

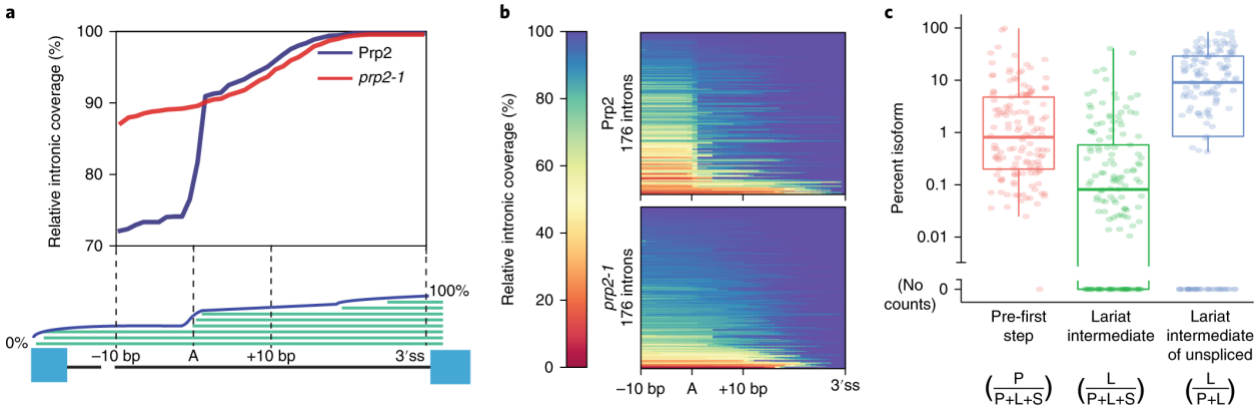


Fig. 3: MPE-seq allows genome-wide profiling of lariat intermediates. a, Meta-intron coverage plot surrounding predicted branch points in a wild-type (Prp2) and step-1 splicing mutant strain (*prp2-1*). The region between the +10 position downstream of the annotated branch point and the 3' splice site (3'ss) was rescaled for each intron. **b**, Heat maps showing the relative coverage at each intron for which lariat intermediate reads were detected. **c**, Estimates of the relative abundance of each isoform for each targeted intron for which reads were detected (P, pre-first-step RNA; L, lariat intermediate; S, spliced mRNA). In box plots, the center line represents the 50th percentile, and lower and upper hinges represent the 25th and 75th percentiles, respectively. Whiskers end at the 0th and 100th percentiles. $n = 141$ introns for which we attempted lariat quantification and found at least one spliced read.

In our initial experiments we used individually synthesized oligonucleotides as primers; we next sought to increase the utility of this approach by examining methods that would facilitate an increase in the number of targeted regions. We developed an approach that used pools of primers derived from array-based syntheses of thousands of oligonucleotides (Supplementary Fig. 9a,b). Using this approach, we re-created the 309 previously described *S. cerevisiae* primers, and generated an additional 3,918

primers that targeted splice junctions in the relatively intron-rich fission yeast *Schizosaccharomyces pombe*. Genome-wide splicing efficiencies determined from MPE-seq libraries generated with primers from pooled syntheses correlated highly with those in libraries derived from individually synthesized oligos (Supplementary Fig. 9), thus validating the utility of this approach. Moreover, MPE-seq libraries generated with primers derived from pooled synthesis also showed strong enrichment for the targeted regions, with levels on par with what we observed with individually synthesized oligonucleotide primers (Supplementary Figs. 2 and 9c). We observed a modest increase in off-target reads when we used primers from the pooled synthesis, consistent with the decreased sequence fidelity of array-based oligo synthesis (Wan et al., 2017) and the increased capacity of these aberrant oligos to prime reverse transcription at undesirable locations. Additionally, as the fraction of the transcriptome that is targeted becomes larger, the fold enrichment over RNA-seq is naturally expected to decrease. Accordingly, when we used the ~4,000 targeting primers in fission yeast, we achieved a median enrichment of sixfold at targeted regions (Supplementary Table 4 (not included)). Nevertheless, this enrichment enabled us to detect rare but natural alternative splicing events (Stepankiw et al., 2015) that are poorly sampled with standard RNA-seq library-preparation methods (Supplementary Fig. 9e). Although we see no de facto limitation to the number of unique primer sequences or species that could be used for MPE-seq, with increasing numbers of primers comes increasing potential for their cross-reactivity with undesirable RNA targets, which highlights the importance of specificity and fidelity in primer design and synthesis.

The improved sensitivity of MPE-seq is perhaps best exemplified by our ability to detect the LI products of the pre-mRNA splicing pathway. In contrast to studies using other recently described methods (Burke et al., 2018; Chen et al., 2018; Nojima et al., 2015) that have reported large-scale detection of upstream-exon splice intermediates and excised lariats, MPE-seq uniquely detects LIs, not excised lariats, from unfractionated cellular RNA. Moreover, these profiling methods that detect RNAs physically associated with the spliceosome require protein tagging and/or purification steps that necessitate large amounts of starting material, which limits their application. Conversely, MPE-seq can be implemented in virtually any system of interest with a need for only microgram quantities of RNA. Additionally, the ability of MPE-seq to query RNA from a wide variety of sources (e.g., cytoplasmic/nuclear fractionated RNA, polysome-fractionated RNA, poly(A)-selected RNA, metabolically labeled RNA) allows for analysis of the cellular location, translational or polyadenylation status, and turnover rates of splice isoforms and intermediates. Overall, we expect that the sensitivity, precision, and flexibility of this approach will lead to a higher-resolution understanding of the splicing pathway. Likewise, primer extension assays have been used to assay RNA secondary structure after in vitro (Lucks et al., 2011) or in vivo (Rouskin et al., 2014) chemical probing, and we expect that MPE-seq could be readily adapted to RNA-structure interrogation and other approaches where primer extension assays or targeted RNA sequencing is applicable.

Methods

Strain maintenance and growth conditions

Unless otherwise indicated, all S. cerevisiae experiments used the wild-type strain BY4741 (MATa, his2Δ1, leu2Δ0, met15Δ0, ura3Δ0). Single colonies were inoculated into liquid YPD media and grown overnight at 30 °C. Overnight cultures were then inoculated into fresh liquid YPD media, with cultures seeded at OD600 ~ 0.05. Cells were collected by vacuum filtration once cultures reached OD600 ~ 0.7 and were then immediately flash-frozen in liquid nitrogen. Cell pellets were stored at –80 °C. For the temperature-sensitive strain harboring the prp2-1 mutation (Hartwell et al., 1970), we grew cultures as described above, but at 25 °C. Once cultures reached OD600 ~ 0.7, an equal volume of fresh 50 °C YPD media was added to shift cells to the nonpermissive temperature of 37 °C. The cultures were then maintained at 37 °C for 15 min before cell collection as described above. All S. pombe experiments used the wild-type strain JP002 (h+, ade6-M210, leu1-32, ura4-D18). Single colonies were inoculated into liquid YES media and grown overnight at 30 °C. Overnight cultures were then inoculated into fresh liquid YES media, seeded at OD600 ~ 0.05. Cells were collected by vacuum filtration when they reached OD600 ~ 0.5 and then were immediately flash-frozen in liquid nitrogen. Cell pellets were stored at –80 °C.

Gene-specific reverse-transcription primer design

For each of the 309 annotated spliceosomal introns in the S. cerevisiae genome (annotations obtained from UCSC SacCer3) and for a subset of introns (3,918 in total) in the S. pombe genome (annotations obtained from Ensemble ASM294v2.37), a reverse-transcription primer was designed within the first 50 nt downstream of the intron. Targeting to this region ensured that short-read sequencing of the products generated from reverse transcription with these primers would cross the upstream

exon–exon or exon–intron boundaries, thereby allowing for determination of the splicing status. Primers were designed with OligoWiz, a program initially developed for microarray probe design that also enables one to select primer sequences optimized for target specificity relative to a designated genomic background (Wernersson & Nielsen, 2003). We used the stand-alone version of OligoWiz with default parameters for short (24–26 bp) oligo design to obtain optimal sequences within each 50-bp window. To the 5' end of each of these sequences we appended two additional sequence elements: a random 7-nt unique molecular identifier (UMI) that allowed for the detection and removal of amplification artifacts arising from library preparation (Kivioja et al., 2012), and the P5 region of the Illumina sequencing primer to allow sequencing of the reverse-transcription products. Each of the primers targeting S. cerevisiae junctions was individually synthesized by Integrated DNA Technologies (IDT); the full sequences are provided in Supplementary Table 5 (not included). Array-based oligonucleotide synthesis was done by LC Sciences using individual OligoMix syntheses for primers from each species (Supplementary Table 6 (not included)).

Complex oligo-mix amplification method

Array-based oligos are synthesized at vastly lower quantities than required for cDNA synthesis in MPE-seq. To generate a sufficiently large quantity of primer pool, we used PCR amplification along with several processing steps (Supplementary Fig. 9a). This was made possible by the addition of two key sequence elements appended on the 3' end of the individually synthesized oligo primers detailed above: from the 5' to 3' direction, (1) a SapI restriction site and (2) a PCR amplification sequence (Supplementary Table 5 (not included)). The oligos were amplified in a standard PCR

reaction with Phusion polymerase. This 400-µl PCR reaction contained 1% of the pooled oligonucleotides from LC Sciences as a template, a forward primer (oHX093) containing a C3 spacer at its 5' end, and a reverse-amplification primer (oHX094) containing a biotin label at its 5' end (Supplementary Table 7 (not included)). A total of 14 amplification cycles were performed, each consisting of the following conditions: denaturation at 95 °C for 10 s, annealing at 60 °C for 20 s, and extension at 72 °C for 30 s. Upon completion of this initial reaction, the entire reaction was used as a template to seed a larger (40-ml) PCR reaction. For efficient amplification, this large reaction was carried out in four 96-well plates with 100 µl in each well. Reaction conditions were identical to those described for the first reaction, but a total of 15 cycles was used for this second amplification. Reactions were purified and concentrated by isopropanol precipitation. To generate single-stranded primers for use in MPE-seq, we first digested the double-stranded amplicons with SapI (NEB R0569) in a 150-µl reaction containing 30 µl of enzyme. The reaction was incubated at 37 °C overnight, after which the reaction products were concentrated by ethanol precipitation. Next, the 5'-to-3' lambda exonuclease (NEB M0262) was used to preferentially degrade the two strands containing unmodified 5' ends. This reaction was carried out at 37 °C for 2 h according to the manufacturer's protocol. The products of this reaction were then purified on Zymo columns with a 7× volume of binding buffer (2 M guanidinium-HCl, 75% isopropanol). After this step, the remaining DNA consisted of the desired single-stranded reverse-transcription primer and an undesired single-stranded section containing the SapI site plus the amplification primer. Making use of the 5' biotin tag on the amplification primer, we removed these undesired oligos by affinity capture with streptavidin beads.

Specifically, we accomplished this by using 50 µl of Dynabeads MyOne Streptavidin C1 according to the manufacturer's protocol. The unbound supernatant fraction was retained, as it contained the desired products. The recovered material was precipitated and verified by 6% native PAGE stained with SyBr Gold (Supplementary Fig. 9b).

First-strand-extension template oligo design

The oligos were designed with three key features from the 5' to 3' end of the oligo. First, we used a portion of the Nextera P7 sequencing adaptor. Of the entirety of the P7 adaptor, the region 3' of the i7 barcode was used. This allowed for independent barcoding and amplification of discrete sequencing libraries. Second, a dN_9 or dN_{12} anchor on the 3' end of the oligo allowed it to randomly anneal to cDNA products. Third, a 3' carbon block modification (hexanediol; IDT) was added to preclude the ability of Klenow to extend this primer. As a result, the oligo could be used only as a template to append the Nextera sequencing adaptor onto the end of first-strand cDNAs. The full sequence of this primer can be found in Supplementary Table 7 (not included).

MPE-seq library prep

<u>cDNA synthesis</u>

For *S. cerevisiae* libraries, RNA was isolated after hot acid phenol extraction (Collart & Oliviero, 1993). Each library was generated with 10 µg of total RNA. From this RNA, we synthesized cDNA by mixing 1 µg of the gene-specific primer pool described above with each RNA sample in a 20-µl reaction containing 50 mM Tris-HCl (pH 8.5), 75 mM KCl. The primers were then annealed in a thermocycler with the following cycle: 70 °C for 1 min, 65 °C for 5 min, and a hold at 47 °C. An equivalent volume of MMLV reverse-transcriptase enzyme mix containing 1 mM dATP, 1 mM dGTP, 1 mM dCTP,

0.4 mM aminoallyl-dUTP, 0.6 mM dTTP, 50 mM Tris-HCl (pH 8.5), 150 mM KCl, 6 mM $MgCl_2$, 10 mM DTT was preheated to 47 °C and added to the primer-annealed RNA mix, resulting in a total reaction volume of 40 µl. It was essential to maintain the samples at 47 °C to reduce off-target cDNA synthesis. Reactions were incubated at 47 °C for 3 h and then subjected to heat inactivation at 85 °C for 5 min. The remaining RNA was hydrolyzed by the addition of a half-volume of 0.3 M NaOH, 0.03 M EDTA and incubation at 65 °C for 15 min. After neutralization with half the original volume of 0.3 M HCl, the cDNA was purified on a Zymo-5 column with a 7× volume of binding buffer (2 M guanidinium HCl, 75% isopropanol). Purified cDNA samples were dried to completion in a SpeedVac. For *S. pombe* libraries, RNA was isolated as described above. Polyadenylated RNA was then isolated from 60 µg of total RNA with the NEBNext Poly(A) mRNA Magnetic Isolation Module. RNA was then fragmented to an average size of 200 nt by incubation in 10 mM $ZnCl_2$, 10 mM Tris-HCl (pH 7.0) for 10 min at 65 °C. The reaction was then quenched by the addition of EGTA (pH 8.0) to a final concentration of 50 mM. The cDNA synthesis reactions were performed as above, with some modifications. For reasons described below, 4 µl of Superscript III (Thermo Fisher) reverse transcriptase was used along with the manufacturer-supplied 5× buffer. For primer annealing and extension, samples were held at 55 °C for 1 h and then subjected to heat inactivation at 85 °C for 5 min.

NHS-ester biotin coupling

Dried cDNA samples were resuspended in 18 µl of fresh 0.1 M sodium bicarbonate (pH 9.0) to which 2 µl of 0.1 mg/µl NHS-biotin (Thermo Fisher; 20217) was added. Reactions were incubated at 65 °C for 1 h, after which biotin-coupled cDNA was

purified from unreacted NHS-biotin on Zymo-5 columns with a 7× volume of binding buffer (2 M guanidinium HCl, 75% isopropanol).

Streptavidin–biotin purification

20 µl of Dynabeads MyOne streptavidin C1 (Thermo Fisher; 65602) per sample was prewashed twice in 500 µl of 1× bind and wash buffer (5 mM Tris-HCl (pH 7.5), 0.5 mM EDTA, and 1 M NaCl) per the manufacturer's protocol. Washed beads were resuspended in 50 µl of 2× bind and wash buffer per sample, and 50 µl was combined with each 50 µl of purified cDNA sample. Biotin–streptavidin binding was allowed to proceed for 30 min at room temperature with rotation. Bound material was washed twice with 500 µl of 1× bind and wash buffer, and then once with 100 µl of 1× SSC. To ensure purification of only single-stranded cDNAs, beads were then incubated with 0.1 M NaOH for two consecutive room-temperature washes for 10 min and 1 min, respectively. Finally, the bound material was washed three times with 100 µl of 1× TE. The cDNA was eluted from the beads by heating to 90 °C for 2 min in the presence of 100 µl of 95% formamide, 10 mM EDTA. The eluate was then purified on Zymo-5 columns as described above, and the cDNA was eluted from columns in 40 µl of water.

First-strand extension

We annealed primers to purified cDNA by combining 1 µl of first-strand extension oligo (100 µM of oJP788 for *S. cerevisiae* and oJP789 for *S. pombe*), 5 µl of 10× NEB buffer 2, 40 µl of purified cDNA sample, and 1 µl of 10 mM (each) dNTP mix. Samples were incubated at 65 °C for 5 min and then cooled to room temperature on the benchtop. To each sample we added 3 µl of Klenow exo- fragment (NEB M0212), and then we incubated the reactions for 5 min at room temperature and subsequently 37 °C

for 30 min. Samples were then purified with streptavidin beads according to the protocol described above. Samples were concentrated on Zymo-5 columns and eluted in 33 µl of water.

PCR amplification

We amplified the reaction products by using 10 µl of the purified material generated in the first-strand extension reaction as a template in a PCR reaction. Illumina Nextera i5 and i7 indexing primers were used in a standard 50-µl PCR reaction with Phusion polymerase (Thermo Fisher; F530S). Cycling conditions were as follows: denaturation at 95 °C for 10 s, annealing at 62 °C for 20 s, and extension at 72 °C for 30 s. Libraries typically required between 14 and 20 cycles of amplification, depending on the efficiency of library preparation. Each PCR reaction was then run on a 6% native polyacrylamide gel, and the DNA was resolved by staining with SyBr gold. Libraries were size-selected from 200 bp to 800 bp, and DNA was extracted from gel fragments via passive diffusion overnight in 0.3 M sodium acetate (pH 5.3). Libraries were then ethanol-precipitated and quantified.

cDNA synthesis temperature experiment

Because of the target-specific nature of MPE-seq cDNA synthesis, any reverse-transcription (RT) events at nontarget sites will reduce the fraction of on-target reads. Indeed, these off-target events contribute substantially to the nonspecific class reads in a typical MPE-seq experiment (Fig. 2a). One way to reduce off-target RT events is to increase the specificity of the RT primers. We assessed this by testing the effect of increased temperature during the RT reaction on off-target sequencing reads. MPE-seq libraries were generated via the above-described protocol, with one primary difference:

increased reaction temperatures required the use of a thermostable enzyme. For this reason, we used Superscript III (Thermo Fisher) along with the manufacturer-supplied buffer (reaction concentrations: 50 mM Tris-HCl (pH 8.3), 75 mM KCl, 3 mM $MgCl_2$). Primer annealing and reactions were carried out at 47 °C, 51 °C, and 55 °C in replicate.

MPE-seq data analysis

Sequencing and alignment

S. cerevisiae MPE-seq libraries were sequenced on the NextSeq platform by the BRC Genomics Facility at Cornell University with 60-bp (P5) + 15-bp (P7) paired-end chemistry. We removed PCR duplicates from the dataset by filtering out non-unique reads with respect to all base calls in both reads, including the 7-bp UMI. In other words, for each set of identical paired-end reads, a single read pair was retained for analysis. MPE-seq reads were aligned to the yeast genome (reference genome assembly R64-1-1 (Engel et al., 2014)) with the STAR aligner (Dobin et al., 2013) with the following alignment parameters: `{--alignEndsType EndToEnd --alignIntronMin 20 --alignIntronMax 1000 --alignMatesGapMax 400 --alignSplicedMateMapLmin 16 --alignSJDBoverhangMin 1 --outSAMmultNmax 1 --outFilterMismatchNmax 3 --clip3pAdapterSeq CTGTCTCTTATACACATCTCCGAGCCCACGAGAC --clip5pNbases 7 0}`. Alignment files were filtered to exclude read mappings deriving from inserts of less than 30 bases. We believe that these short fragments represent unextended RT primers that were retained in the sequencing libraries. These small fragments can sometimes erroneously map to splice junctions or target introns, even though we believe that they are not derived from cellular RNA. *S. pombe* MPE-seq libraries were sequenced on the MiSeq

platform by the BRC Genomics Facility at Cornell University using 100-bp (P5) + 50-bp (P7) paired-end chemistry. Reads were trimmed to 60 bp and 15 bp and processed as described above for Supplementary Fig.9c, whereas full-length reads were processed as described above for Supplementary Fig. 9e.

S. cerevisiae RNA-seq libraries were sequenced on an Illumina HiSeq 2500 by the BRC Genomics Facility at Cornell University using 100-bp single-end reads. *S. pombe* RNA-seq data (Rhind et al., 2011) were downloaded from the NCBI BioSample database (accession SRS167019), and read 2 of read pairs was discarded to make read lengths comparable to those in our other libraries. Reads were aligned with the STAR aligner with the following alignment parameters: `{--alignEndsType EndToEnd --alignIntronMin 20 --alignIntronMax 1000 --alignSJDBoverhangMin 1 --outSAMmultNmax 1 --outFilterMismatchNmax 3 --clip3pAdapterSeq CTGTCTCTTATACACATCTCCGAGCCCACGAGAC}`.

When applicable, replicate libraries were combined before alignment. However, for assessment of the technical reproducibility of MPE-seq, replicate libraries were subsampled to varying read depths, aligned separately, and compared with RNA-seq libraries also subsampled to varying read depths.

Estimating the fraction of on-target reads and MPE-seq enrichment

With bedtools (Quinlan & Hall, 2010), read 1 alignments extending into a targeted intron or crossing a targeted exon–exon junction were considered on-target. Read 1 alignments that mapped downstream of a targeted intron but did not extend into an intron or cross an exon–exon junction were considered unextended primers. Read 1

alignments that mapped to the genome at nontargeted loci were considered off-target. Unmapped reads were then realigned to the genome with the same parameters as above (except `with --clip5pNbases 31 0`), and subsequent read 1 alignments to nontargeted loci were also considered off-target. The remaining reads were considered unmapped. We calculated enrichment by dividing the number of read-count-normalized exon–exon junctions found in MPE-seq by the number of read-count-normalized exon–exon junctions in RNA-seq datasets for each targeted intron.

Estimating splice isoform abundances from MPE-seq data

For each intron, we determined the relative abundance of spliced and unspliced isoforms by counting spliced and unspliced reads. Spliced reads (*S*) were counted with the SJ.out.tab file created by the aligner. Unspliced reads were counted with bedtools (Quinlan & Hall, 2010), which we used to count the number of reads that covered any part of the intron, considering only the first read of paired-end reads. Unspliced read counts were further categorized as deriving from an LI (*L*) or pre-first-step RNA (*P*) on the basis of the mapping location of the second read of the paired-end reads, which we observed to often terminate near the transcription start site or, in the case of an LI-derived cDNA, near branch point A of the intron. On the basis of paired-end mapping locations, each fragment was categorized into one of six categories (Supplementary Fig. 5), and the counts within those six categories (C_1–C_6) were used to calculate *S*, *P*, and *L* as follows:

$$S = C_1 + C_2$$

$$P = C_3 \left(1 + \frac{C_5 + C_6}{C_3 + C_4}\right)$$

$$= C_4 \left(1 + \frac{C_5 + C_6}{C_3 + C_4}\right)$$

We determined the locations of branch points (Supplementary Table 8 (not included)) by consolidating the most used branch point from lariat sequencing data (Mayerle et al., 2017) and previously described branch locations based on sequence motif searches (Grate & Ares, 2002).

Heat maps and meta-gene plots

To generate meta-gene plots, which illustrate read coverage around features of interest, we used the deepTools ComputeMatrix command (Ramírez et al., 2016) in conjunction with a BigWig coverage file of the 3' terminating bases and a bedfile containing transcription start site positions as determined by PRO-cap (Booth et al., 2016) or a bedfile containing the annotated branch-point regions detailed above. Importantly, this bedfile was filtered to include only branch-point regions that would produce an LI within the size range captured by library size-selection of MPE-seq libraries (see column “AttemptedLariatQuantification?” in Supplementary Table 3).

RNA-seq experiments

Library prep

For each RNA-seq library, 1 µg of total RNA was input into the NEBNext Ultra Directional RNA library prep kit (Illumina). Libraries were prepared according to the manufacturer’s protocol.

Estimating splice isoform abundances from RNA-seq data

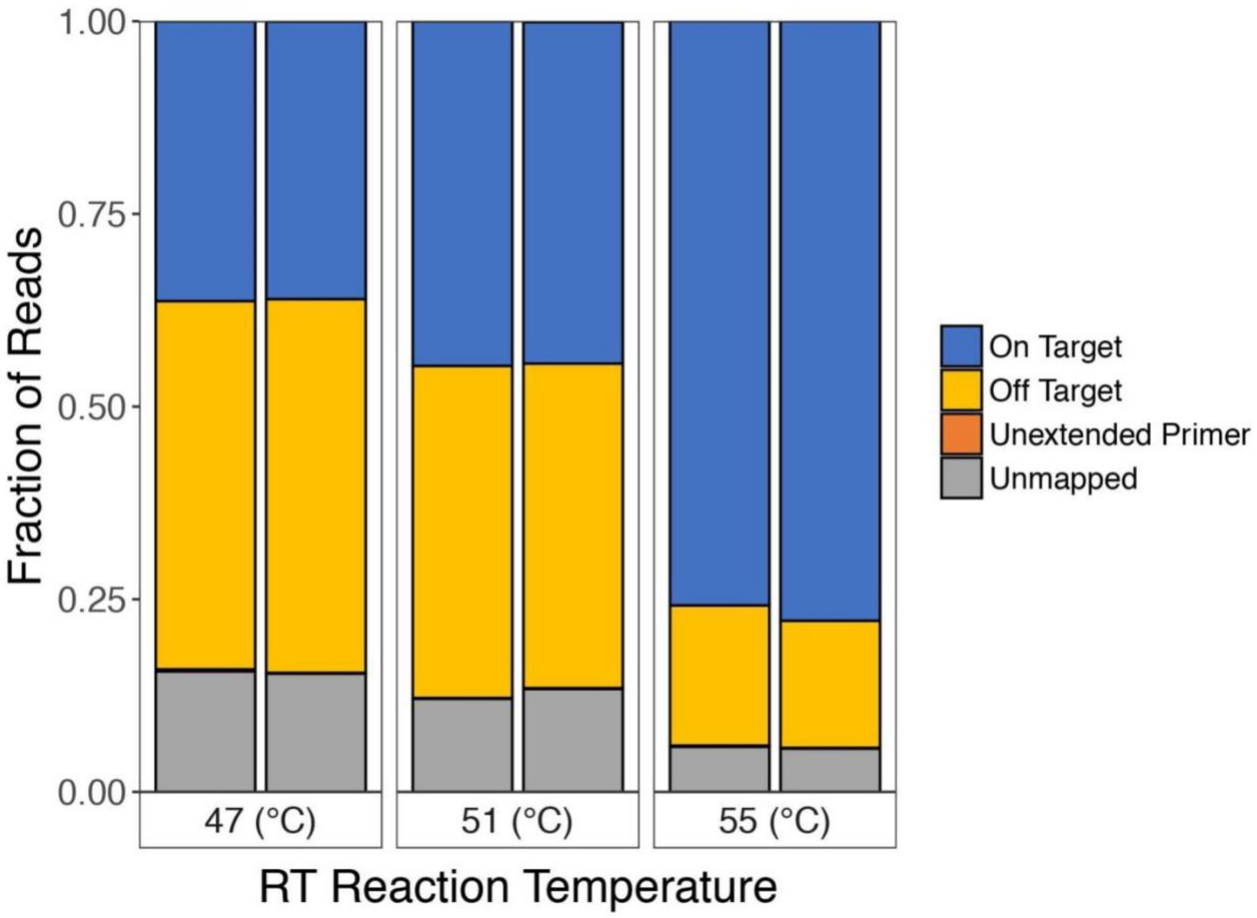
Similarly to MPE-seq data, spliced reads from target introns were counted with the SJ.out.tab file created by the aligner. Unspliced reads were counted with the bedtools software package (Quinlan & Hall, 2010), which counted the number of reads

that overlapped an intron. Spliced and unspliced read counts for each intron were then length-normalized for the feature's potential mapping space. The potential mapping space for a spliced read is equal to 2 × the read length minus the minimum splice-junction overhang length. The potential mapping space for an unspliced read is equal to 2 × the read length minus the minimum splice-junction overhang length plus the length of the intron. Read counts assigned to each feature were then divided by the length. The unspliced fraction was calculated for each intron as the quotient of length-normalized unspliced reads and spliced reads.

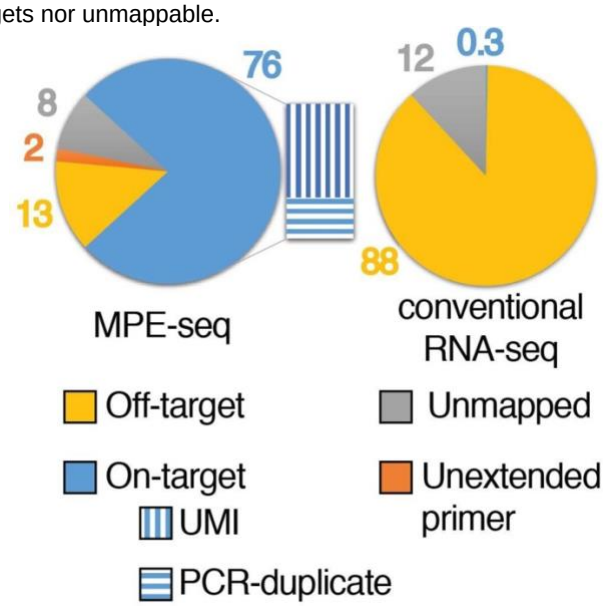
Gene expression normalization

Relative transcript expression was calculated from RNA-seq data via transcripts per million (TPM) normalization (Conesa et al., 2016), considering only exonic reads and exonic gene lengths. For *S. cerevisiae* MPE-seq data, we calculated a similar TPM metric by summing the reads per gene and dividing by the number of library mapped reads. Given that a single RNA corresponds to a single primer extension event, and because nearly all targeted transcripts have only a single targeting primer, normalization by gene length was not done in this calculation of TPM.

Supplementary Figures

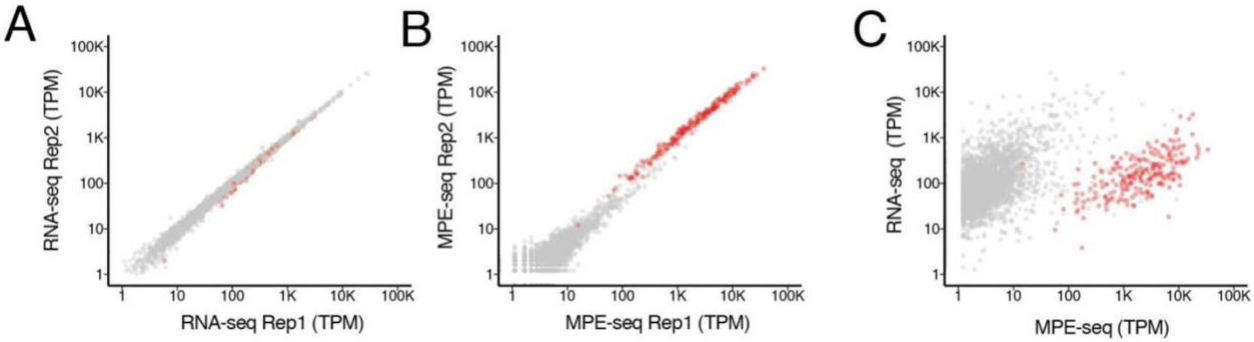


Supplementary Figure 1: Elevated temperatures in reverse-transcription reactions increase specificity. The fraction of on-target and off-target reads from replicate MPE-seq libraries generated from reverse-transcription reactions carried out at various temperatures. A small fraction of reads were categorized as “Unextended primer,” which corresponds to short primer extension products (0–5 bases extended past the primer), and thus were categorized as neither cDNAs derived from RNA targets nor unmappable.

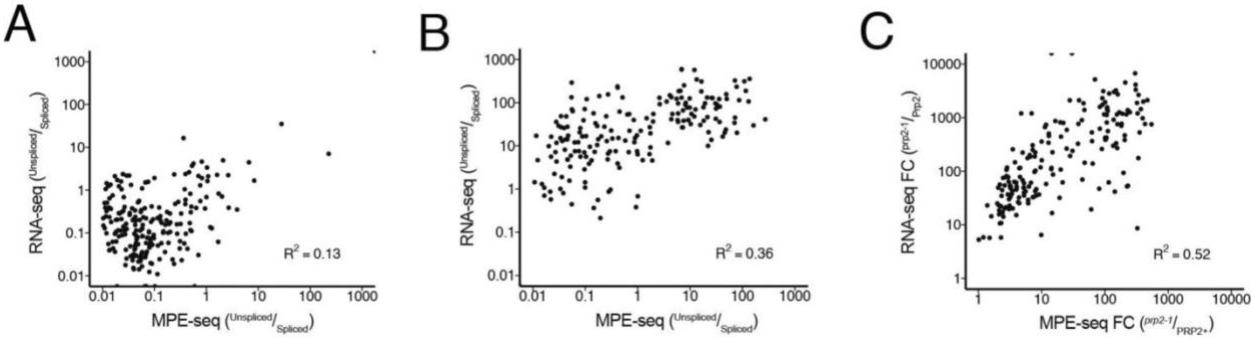


Supplementary Figure 2: Percentage of reads mapping to targeted regions. The percentage of reads mapped to target and off-target regions is depicted for MPE-seq and conventional RNA-seq. In MPE-seq a small fraction of

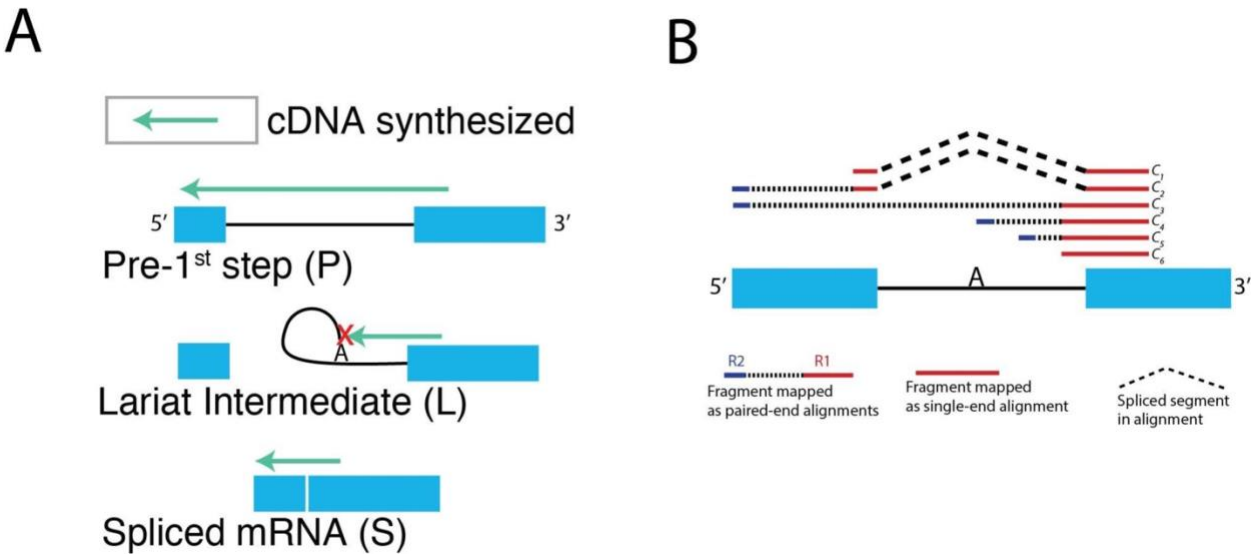
reads were categorized as “Unextended primer,” which corresponds to short primer extension products (0–5 bases extended past the primer), and thus were not categorized as cDNAs derived from RNA targets.



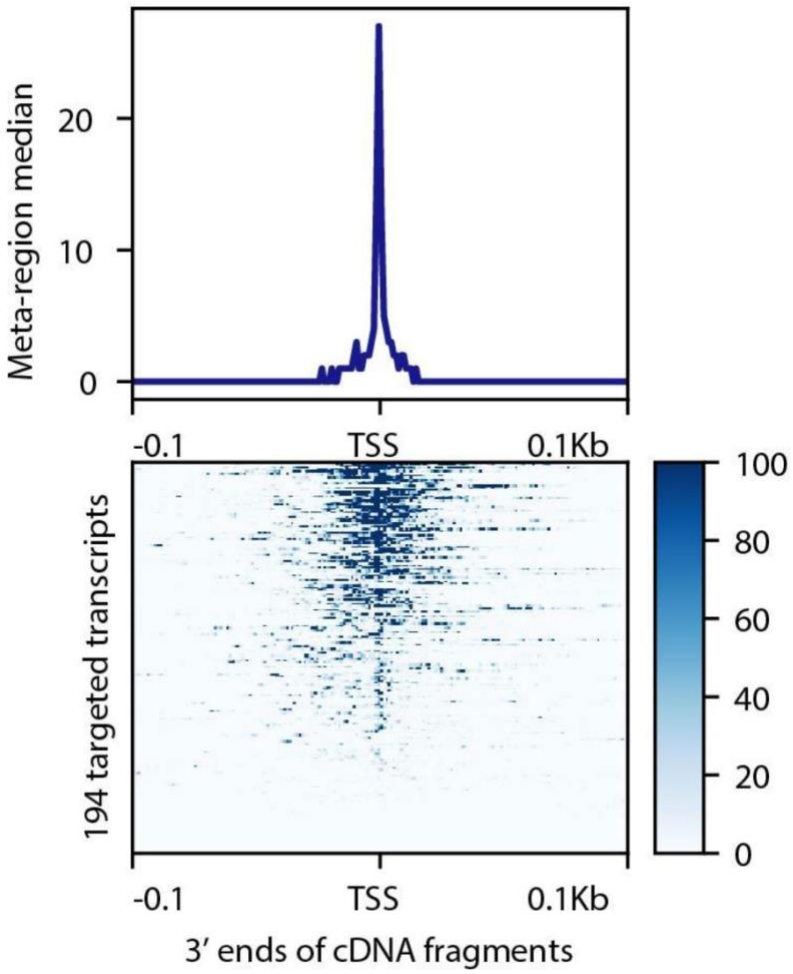
Supplementary Figure 3: Expression measurements as determined by MPE-seq and RNA-seq. (a) A scatter plot depicts gene expression measurements (RNA-seq) in replicate datasets. Genes containing splice events that were among those chosen for targeted sequencing are depicted in red. These targeted genes range in expression level by orders of magnitude. **(b)** A scatter plot depicts gene expression measurements in replicate MPE-seq datasets (red). Similar to conventional RNA-seq, expression measurements in MPE-seq are highly reproducible between replicates, even for the small proportion of mis-priming events that map to off-target locations (gray). **(c)** A scatter plot depicts gene expression measurements in RNA-seq and MPE-seq. The right shift of targeted genes reflects successful enrichment of targets by orders of magnitude. The observation that even highly expressed genes as measured by RNA-seq are proportionally highly expressed in MPE-seq suggests that primers are not limiting during reverse transcription.



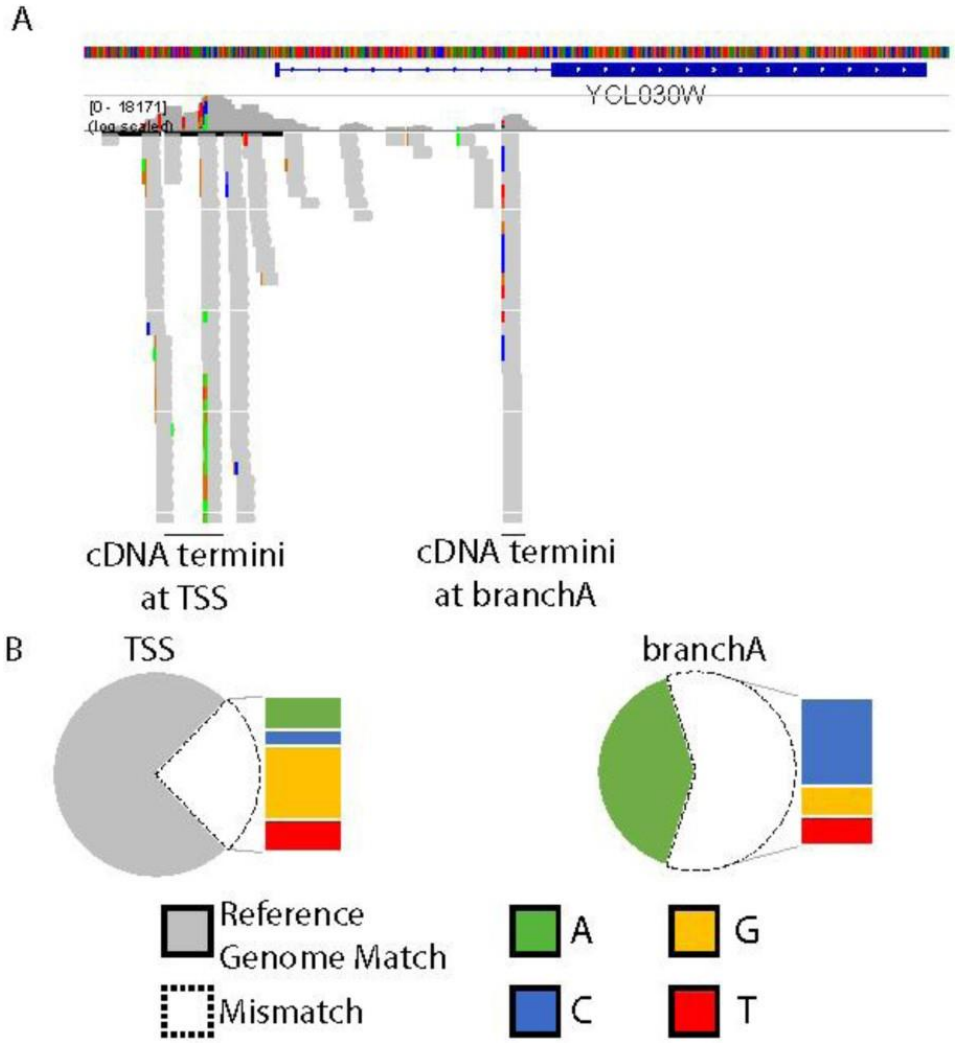
Supplementary Figure 4: Splicing measurements as determined by MPE-seq and RNA-seq. (a) A scatter plot depicts intron-retention measurements in MPE-seq and conventional RNA-seq using wild-type (Prp2) RNA. For calculation of R^2, n = 252 intron-retention events that were quantified, requiring at least one spliced read and one unspliced read in both experiments. **(b)** A scatter plot depicts intron-retention measurements in MPE-seq and conventional RNA-seq using RNA from a splicing mutant strain (*prp2-1*). For calculation of R^2, n = 193 intron-retention events that were quantified, requiring at least one spliced read and one unspliced read in both experiments**. (c)** A scatter plot depicts the fold-change (*prp2-1*/Prp2) in intron retention as measured by MPE-seq and conventional RNA-seq. For calculation of R^2, n = 203 intron-retention events that were quantified, requiring that both RNA-seq and MPE-seq be quantifiable in the wild-type (Prp2) dataset.



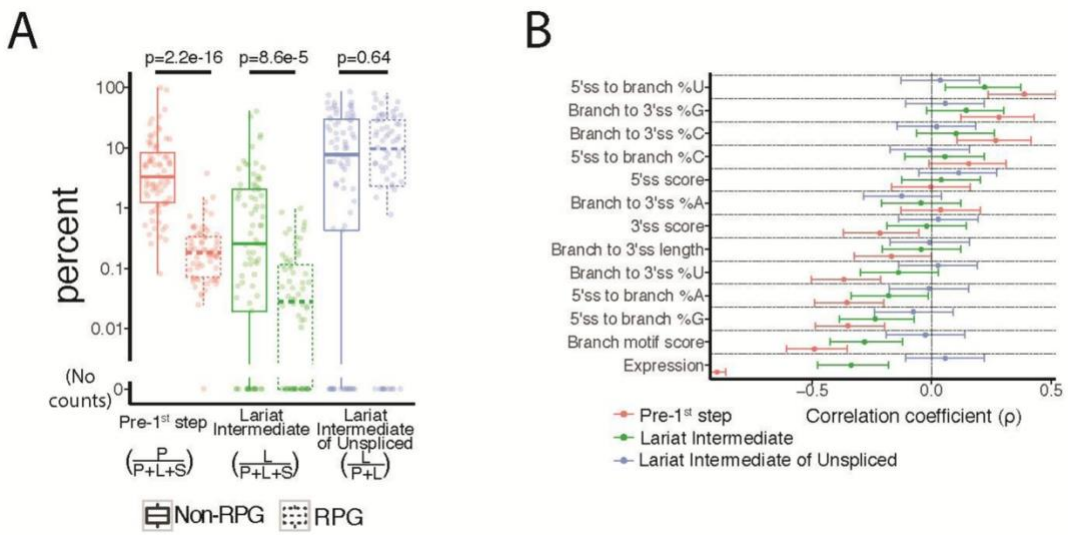
Supplementary Figure 5: Schematic for assigning reads to splice intermediate isoforms. (a) Schematic depicting cDNA products derived from pre-first-step (P), lariat intermediate (L), and spliced mRNA (S) isoforms. **(b)** To quantify the abundance of P, L, and S isoforms for each targeted splice event, we counted read fragments and categorized them into six classes based on paired-end alignments. Fragments containing a splice junction (C_1 and C_2) are indicative of S. Fragments that are unspliced and traverse the branch-point region (C_3) are classified as P. Fragments that are unspliced but terminate within a window of –3 to +5 bp from the previously determined branch point (C_4) are classified as L. Fragments that are unspliced and either terminate downstream of the branch point (C_5) or for which the terminus could not be mapped (C_6) are ambiguous between P and L. Therefore, for accounting purposes, the counts for these fragments were coerced into P and L classifications based on the ratio of P and L determined by unambiguous mappings (C_3 and C_4). See Methods for more details.



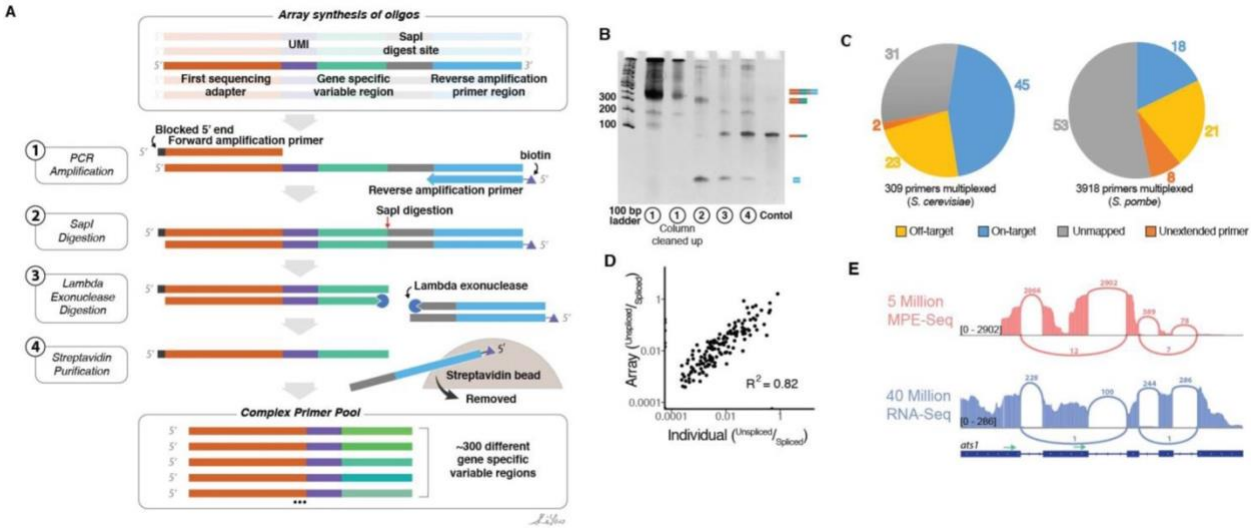
Supplementary Figure 6: Transcription start site profiling by MPE-seq. Metagene profile of 3' ends mapped by MPE-seq, centered on transcription start sites (TSSs) as determined by PRO-cap, an orthologous method for mapping TSSs. The high abundance of read ends that pile up at TSSs indicates that MPE-seq can be used to profile cDNA termini.



Supplementary Figure 7: Lariat-intermediate-derived cDNAs contain a unique signature of mismatches. (a) Genome browser screenshot of the 3' ends of reads from paired-end sequenced fragments illustrates the unique signature of non-templated base incorporation by reverse transcriptase at a branched adenosine versus the 5' RNA terminus. (b) Genome-wide quantification of the mismatch frequencies at 3' termini of cDNAs near the TSS (left) versus at the annotated branch point (right).



Supplementary Figure 8: Transcript features that correlate with the abundance of lariat intermediates. (a) The abundance of pre-first-step RNA and lariat intermediate RNA is significantly correlated with the classification of introns into those that are in ribosomal protein genes (RPG) and non-RPG. However, the abundance of lariat intermediate relative to pre-first-step RNA, a metric of the efficiency of the second step of splicing, does not correlate. Horizontal lines in box plots represent the 25th, 50th, and 75th percentiles. Whiskers end at the 0th and 100th percentiles. A two-sided Mann–Whitney U-test was used to test differences between the distributions of these metrics. For each metric (pre-first-step, lariat intermediate, and lariat intermediate of unspliced), n = 141 introns for which we attempted lariat quantification and found at least one spliced read, of which 64 were RPG and 77 were non-RPG. (b) Spearman correlations of various features to the abundance of pre-first-step RNA, lariat intermediate, or the abundance of lariat intermediate relative to pre-first-step RNA. None of these features significantly correlate with the abundance of lariat intermediate relative to pre-first-step RNA, a metric of the efficiency of the second step of splicing. Error bars indicate 95% confidence intervals as estimated by Fisher transformation of Spearman's correlation coefficient. For each metric (pre-first-step, lariat intermediate, and lariat intermediate of unspliced), n = 141 introns for which we attempted lariat quantification and found at least one spliced read.



Supplementary Figure 9: Array-based oligonucleotide synthesis can be used to generate primer pools for use in MPE-seq. (a) Obtaining adequate amounts of primer pools for MPE-seq from cost-effective array-based oligonucleotide synthesis can be achieved in four steps. (1) PCR amplification of the oligonucleotide synthesis pool using a 5' blocked sense primer and a biotinylated antisense primer. (2) Restriction digestion to cleave off the PCR primer handle. (3) Lambda exonuclease digestion of free 5' ends. (4) Streptavidin purification of biotinylated PCR handle. The unbound fraction is the desired primer pool product. **(b)** Steps during the amplification and purification of array-synthesized primer pools are monitored via native gel electrophoresis. The control lane represents a pool of

individually synthesized MPE-seq primers which did not require amplification and purification. Lanes refer to products of each individual step in the protocol. The unbound fraction is the desired primer pool product. Similar results have been consistently obtained in >3 independent experiments. **(c)** The percentage of reads mapped to target and off-target regions is depicted for MPE-seq using array-synthesized primers. **(d)** A scatter plot compares the fraction of unspliced mRNAs measured by MPE-seq libraries which used individually synthesized primer pools versus array-based synthesis of primer pools. For calculation of Pearson's correlation coefficient, n = 140 intron-retention events which were quantified, requiring at least one spliced read and one unspliced read in both experiments. Sashimi plot of a targeted region within the *ats1* gene locus demonstrates the capacity of MPE-seq to reveal complex alternative splicing patterns with higher sensitivity than RNA-seq, despite having lower total sequencing depth.

Works Cited

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., Kim, T., Misquitta-Ali, C. M., Wilson, M. D., Kim, P. M., Odom, D. T., Frey, B. J., & Blencowe, B. J. (2012). The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science*, *338*(6114), 1587–1593. https://doi.org/10.1126/science.1230612

Blomquist, T. M., Crawford, E. L., Lovett, J. L., Yeo, J., Stanoszek, L. M., Levin, A., Li, J., Lu, M., Shi, L., Muldrew, K., & Willey, J. C. (2013). Targeted RNA-Sequencing with Competitive Multiplex- PCR Amplicon Libraries. *PLOS ONE*, *8*(11).

Booth, G. T., Wang, I. X., Cheung, V. G., & Lis, J. T. (2016). Divergence of a conserved elongation factor and transcription regulation in budding and fission yeast. *Genome Research*, *26*(6), 799–811. https://doi.org/10.1101/gr.204578.116

Burke, J. E., Longhurst, A. D., Merkurjev, D., Sales-Lee, J., Rao, B., Moresco, J. J., Yates, J. R., Li, J. J., & Madhani, H. (2018). Spliceosome Profiling Visualizes Operations of a Dynamic RNP at Nucleotide Resolution. *Cell*, *173*, 1014–1030. https://doi.org/10.1016/j.cell.2018.03.020

Carey, M. F., Peterson, C. L., & Smale, S. T. (2013). The Primer Extension Assay. *Cold Spring Harbor Protocols*, *2013*(2), 164–173. https://doi.org/10.1101/pdb.prot071902

Chen, W., Moore, J., Ozadam, H., Shulha, H. P., Rhind, N., Weng, Z., & Moore, M. J. (2018). Transcriptome-wide Interrogation of the Functional Intronome by Spliceosome Profiling. *Cell*, *173*, 1031–1044. https://doi.org/10.1016/j.cell.2018.03.062

Collart, M. A., & Oliviero, S. (1993). Preparation of Yeast RNA. *Current Protocols in Molecular Biology*, *23*(1), 13.12.1-13.12.5. https://doi.org/10.1002/0471142727.mb1312s23

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 13. https://doi.org/10.1186/s13059-016-0881-8

Coombes, C. E. (2005). An evaluation of detection methods for large lariat RNAs. *RNA*, *11*(3), 323–331. https://doi.org/10.1261/rna.7124405

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Engel, S. R., Dietrich, F. S., Fisk, D. G., Binkley, G., Balakrishnan, R., Costanzo, M. C., Dwight, S. S., Hitz, B. C., Karra, K., Nash, R. S., Weng, S., Wong, E. D., Lloyd, P., Skrzypek, M. S., Miyasato, S. R., Simison, M., & Cherry, J. M. (2014). The Reference Genome Sequence of Saccharomyces cerevisiae: Then and Now. *G3*, *4*(3), 389–398. https://doi.org/10.1534/g3.113.008995

Grate, L., & Ares, M. (2002). *Searching Yeast Intron Data at Ares Lab Web Site*. 13.

Hartwell, L. H., McLaughlin, C. S., & Warner, J. R. (1970). Identification of ten genes that control ribosome formation in yeast. *Molecular and General Genetics MGG*, *109*(1), 42–56. https://doi.org/10.1007/BF00334045

Kim, S.-H., & Lin, R.-J. (1996). Spliceosome Activation by PRP2 ATPase prior to the First Transesterification Reaction of Pre-mRNA Splicing. *Molecular and Cellular Biology*, *16*(12), 6810–6819.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., & Taipale, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, *9*(1), 72–74.

Lucks, J. B., Mortimer, S. A., Trapnell, C., Luo, S., Aviran, S., Schroth, G. P., Pachter, L., Doudna, J. A., & Arkin, A. P. (2011). Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences*, *108*(27), 11063–11068. https://doi.org/10.1073/pnas.1106501108

Mayerle, M., Raghavan, M., Ledoux, S., Price, A., Stepankiw, N., Hadjivassiliou, H., Moehle, E. A., Mendoza, S. D., Pleiss, J. A., Guthrie, C., & Abelson, J. (2017). *Structural toggle in the RNaseH domain of Prp8 helps balance splicing fidelity and catalytic efficiency*. 6.

Mercer, T. R., Clark, M. B., Crawford, J., Brunck, M. E., Gerhardt, D. J., Taft, R. J., Nielsen, L. K., Dinger, M. E., & Mattick, J. S. (2014). Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nature Protocols*, *9*(5), 989–1009. https://doi.org/10.1038/nprot.2014.058

Mercer, T. R., Gerhardt, D. J., Dinger, M. E., Crawford, J., Trapnell, C., Jeddeloh, J. A., Mattick, J. S., & Rinn, J. L. (2012). Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nature Biotechnology*, *30*(1), 99–104. https://doi.org/10.1038/nbt.2024

Merkin, J., Russell, C., Chen, P., & Burge, C. B. (2012). Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science*, *338*(6114), 1593–1599. https://doi.org/10.1126/science.1228186

Nojima, T., Gomes, T., Grosso, A. R., Kimura, H., Dye, M. J., Dhir, S., Carmo-Fonseca, M., & Proudfoot, N. J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell*, *161*, 526–540.

Padgett, R. A., Konarska, M. M., Aebi, M., Hornig, H., Weissmann, C., & Sharp, P. A. (1985). Nonconsensus branch-site sequences in the in vitro splicing of transcripts of mutant rabbit beta-globin genes. *Proc. Natl. Acad. Sci. USA*, *82*, 8349–8353.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842.

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., & Manke, T. (2016). deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, *44*(W1), W160–W165. https://doi.org/10.1093/nar/gkw257

Rhind, N., Chen, Z., Yassour, M., Thompson, D. A., Haas, B. J., Habib, N., Wapinski, I., Roy, S., Lin, M. F., Heiman, D. I., Young, S. K., Furuya, K., Guo, Y., Pidoux, A., Chen, H. M., Robbertse, B., Goldberg, J. M., Aoki, K., Bayne, E. H., … Nusbaum, C. (2011). Comparative Functional Genomics of the Fission Yeasts. *Science*, *332*(6032), 930–936. https://doi.org/10.1126/science.1203357

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., & Weissman, J. S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, *505*(7485), 701–705. https://doi.org/10.1038/nature12894

Stepankiw, N., Raghavan, M., Fogarty, E. A., Grimson, A., & Pleiss, J. A. (2015). Widespread alternative and aberrant splicing revealed by lariat sequencing. *Nucleic Acids Research*, *43*(17), 8488–8501. https://doi.org/10.1093/nar/gkv763

Wan, W., Lu, M., Wang, D., Gao, X., & Hong, J. (2017). High-fidelity de novo synthesis of pathways using microchip-synthesized oligonucleotides and general molecular biology equipment. *Scientific Reports*, *7*(1), 6119. https://doi.org/10.1038/s41598-017-06428-0

Wernersson, R., & Nielsen, H. B. (2003). OligoWiz 2.0—Integrating sequence feature annotation into the design of microarray probes. *Genes & Development*, *17*, 419–437.

Zheng, W., Chung, L. M., & Zhao, H. (2011). Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics*, *12*(1), 290. https://doi.org/10.1186/1471-2105-12-290

Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R., & Siebert, P. D. (2001). Reverse Transcriptase Template Switching: A SMART™ Approach for Full-Length cDNA Library Construction. *BioTechniques*, *30*(4), 892–897. https://doi.org/10.2144/01304pf02

Appendix II: Transcript-specific determinants of pre-mRNA splicing revealed through *in vivo* kinetic analyses of the 1st and 2nd chemical steps

Alternative citation:

Gildea MA, Dwyer ZW, Pleiss JA (2021). Transcript-specific determinants of pre-mRNA splicing revealed through *in vivo* kinetic analyses of the 1st and 2nd chemical steps. *In prep.*

Summary

Understanding how the spliceosome processes its composite of pre-mRNA substrates through the two chemical steps required for mature mRNA production will be essential to deciphering splicing regulation, and its mis-regulation in human disease. To this end, here we have measured the *in vivo* rates of each step of pre-mRNA splicing across the genome-wide complement of splicing substrates by coupling metabolic RNA labeling, multiplexed primer extension sequencing (MPE-seq), and first order kinetic modeling. We demonstrate that there exists a wide variety of rates by which different introns are removed, that splice site sequences are primary determinants of 1st step rates, and that the 2nd step is generally faster than the 1st step. Additionally, we find that the ribosomal protein genes (RPGs) are spliced faster than non-RPGs (nRPGs) at each step, and that RPGs share distinct and evolutionarily conserved cis-features that differentiate them from nRPGs and may contribute to their faster splicing. Using a genetic variant which is defective for the 1st step of the splicing pathway, we observe the expected, genome-wide defect in the 1st step, but an unexpected, transcript-specific, change in the 2nd step wherein RPGs are significantly more slowed as compared to nRPGs. Importantly, these data uncover a coupling between transcription and 1st and

2nd step splicing rates that suggests co-transcriptional splicing is an important determinant of splicing rates.

Introduction

In the time since the discovery of pre-messenger RNA (pre-mRNA) splicing, our understanding of its importance in regulating gene expression has grown dramatically. This is perhaps best exemplified by the ever-increasing number of human diseases that are found to be associated with mutations in the splicing pathway (Montes et al., 2019, Scotti and Swanson, 2016). Critical to our understanding of these diseases is a comprehensive knowledge of the mechanisms by which the spliceosome processes its diverse set of substrates: for many such diseases, the fundamental question of whether the disease-state results from a defect in the splicing of the global complement of transcripts, or in the splicing of a specific but critical subset of transcripts, remains unresolved. Importantly, many of these details remain unclear due to the difficult problem the complexity of splicing regulation presents. Pre-mRNA splicing requires that the spliceosome, which catalyzes intron removal, accurately defines, assembles upon, and activates the appropriate splice site sequences in the background of a sea of non-cognate, cryptic sites (Lee and Rio, 2015, Wahl et al., 2009). Moreover, the spliceosome must catalyze the reaction in a timely manner, striking a balance between fidelity and speed to achieve maximal efficiency (Semlow and Staley, 2012). An understanding of the relative speeds or rates with which the complement of spliceosomal substrates are processed will allow for elucidation of substrate specific features that enable their regulation.

The spliceosome removes introns by catalyzing two stepwise transesterification reactions. Decades of research has established that cis sequence elements within introns and transcripts influence splicing outcomes. Notably, the 5' splice site (5'SS), 3' splice site (3'SS), and branch point (BP) sequences define introns by directly base pairing to partially complimentary sequences within the spliceosomal snRNAs, and as a result couple cis-elements within an intron with splice site selection (Wahl et al., 2009). However, the influence of these sequences, among other cis-elements, on the rate with which an intron is processed through each chemical step of splicing *in vivo* is poorly understood. Though many advancements have been made in our understanding of splicing and the spliceosome through detailed structural and functional studies, methodological limitations have made it difficult to deconvolute the influence of cis-features on the rates of the 1st and 2nd steps of splicing *in vivo* (Fica and Nagai, 2017, Shi, 2017, Lee and Rio, 2015, Wahl et al., 2009, Mayerle and Guthrie, 2017). For example, RNA-seq is commonly used to assess the impact of specific cis and/or trans-acting factors on splicing efficiency by comparing changes in reads representing unspliced and spliced transcripts. Yet, as commonly employed this approach is intrinsically limited by its reliance on steady-state measurements. While splicing rate and fidelity influence the steady-state abundances of unspliced and spliced species, so too do other processes such as degradation of the mature species, making it impossible to deconvolute changes in splicing rate from other degradation rates using steady-state measurements alone (Wachutka and Gagneur, 2017). By contrast, many investigators historically used cellular extracts to measure the kinetics of *in vitro* splicing of reporter constructs, avoiding the pitfalls of steady state analyses (Mayerle and Guthrie, 2017,

Hicks et al., 2005). But while such studies have provided important insights into the splicing process, it remains unknown how the results from these studies, where fully formed transcripts are presented to an unfractionated mixture of cellular components, relate to *in vivo* conditions across the global complement of substrates.

In addition to the role of *cis*-regulatory elements on splicing efficiency, it is well established that other pre-mRNA processing events are functionally coupled to splicing (Herzel et al., 2017, Naftelberg et al., 2015, Bentley, 2014). Current data overwhelmingly support a model wherein splicing is temporally, spatially, and functionally connected with transcription (Herzel et al., 2017, Naftelberg et al., 2015, Bentley, 2014). Indeed, *in vivo* studies have demonstrated the capacity of transcription to influence splicing outcomes (Naftelberg et al., 2015, Bentley, 2014). Nevertheless, estimates of when splicing completes in relation to transcription range widely. Whereas the seminal studies which established the co-transcriptional nature of splicing showed that assembly of the spliceosome occurs in a largely co-transcriptional fashion, these studies nevertheless concluded that the chemical steps generally complete post-transcriptionally (Lacadie et al., 2006, Tardiff et al., 2006, Moore et al., 2006). More recently however, studies using a variety of methods have yielded widely disparate estimates of when splicing completes with respect to the position of RNA polymerase. For example, whereas work from the Neugebauer group suggested that the chemistry of splicing may complete almost concurrently with production of the 3'SS (Oesterreich et al., 2016, Reimer et al., 2020), work from the Churchman group using a similar method suggested that the chemistry of splicing generally completes after kilobases of RNA have been transcribed (Drexler et al., 2020). A clearer picture of the functional coupling

of splicing and transcription as well as the influence of splicing on gene expression would be bolstered by robust measurements of splicing kinetics at individual step resolution.

Recently, the use of metabolic RNA labeling to estimate RNA processing rates *in vivo* has gained popularity and greatly expanded our knowledge of pre-mRNA processing kinetics in a variety of organisms (Duffy et al., 2019). This method enables the time-resolved isolation of nascent RNA and utilizes kinetic modeling of approach to equilibrium curves to estimate the rates of production and degradation of different RNA isoforms. This approach has been applied to measure splicing rates in a variety of organisms, generally using standard RNA-seq methods (Barrass et al., 2015, Wachutka et al., 2019, Pai et al., 2017, Eser et al., 2016, Rabani et al., 2014, Windhager et al., 2012). Despite these advances, the uniformity of RNA-seq read coverage across transcripts produces a low abundance of splicing informative reads per splicing event, resulting in poor quantification of many splicing events and thus their splicing rates. To improve upon these limitations, our lab has recently developed Multiplexed Primer Extension sequencing, or MPE-seq, a targeted RNA-seq method that provides a significant enrichment for splicing informative reads in RNA-seq data sets and enables differentiation of completely unspliced, lariat intermediate, and spliced isoforms genome-wide (Xu et al., 2019, Gildea et al., 2019). Here, we report on our measurements of the 1^{st} and 2^{nd} step splicing rates for the genome-wide compliment of substrates in the budding yeast *Saccharomyces cerevisiae* that were generated by combining rapid metabolic RNA labeling, MPE-seq, and modeling of approach to steady state kinetics.

Results

To better understand the relative efficiencies by which the spliceosome processes its composite of substrates, we designed a strategy to determine the genome-wide rates of the 1st and 2nd steps of splicing. As outlined in **Figure 1**, our approach incorporated three principal components, briefly described here and more fully described in the Methods section. First, we employed a rapid metabolic labelling approach using 4-thiouracil (4tu) to isolate nascent RNA at time intervals ranging from seconds to minutes. Second, we used multiplexed primer extension sequencing (MPE-seq) to quantify splicing intermediates and track their processing through time. Our lab has previously demonstrated two important properties of MPE-seq that are essential to this work: a dramatic increase in the sensitivity for detecting splicing-informative reads, enabling high precision measurements of their relative abundances; and the capacity to distinguish between unspliced, lariat intermediate, and spliced isoforms of a given transcript. Finally, these data were fit to first order kinetic models to estimate the half-lives of the 1st and 2nd steps of splicing genome-wide: the high resolution afforded by MPE-seq allowed us to robustly measure the change in abundance of these three splicing isoforms for each intron in the genome through time, enabling generation of high-quality kinetic models for these rates.

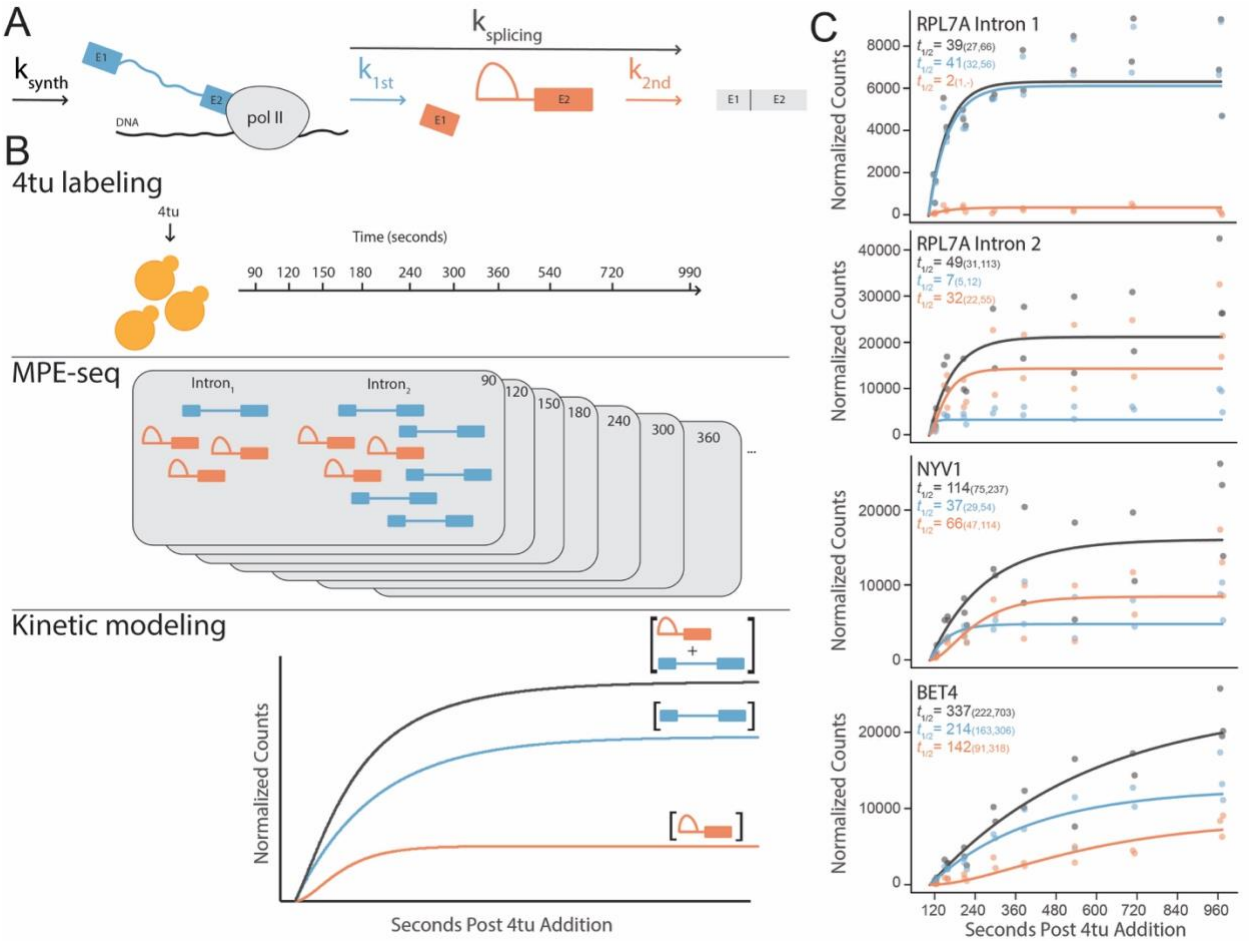


Figure 1: Determining the rates of 1st and 2nd Chemical Step of pre-mRNA Splicing: (**A**) Schematic of the rates measured in this study. (**B**) Experimental workflow. 4-thiouracil (4tu) is introduced to actively growing yeast cells followed by sample collection at specified time points. Nascent 4tu-labeled RNA is purified from each sample and MPE-seq libraries are prepared. The pre 1st step and lariat intermediate isoforms are quantified genome wide, and kinetic models are fit to the approach to equilibrium curves to estimate the total (grey), 1st (blue), and 2nd (orange) step rates of pre-mRNA splicing. (**C**) Normalized counts of total unspliced (grey), pre 1st step (blue), lariat intermediate (orange) versus time in seconds after addition of 4tu for RPL7A intron 1, RPL7A intron 2, NYV1, and BET4. Solid lines are model fits and corresponding half-lives are provided with 90% confidence intervals in parenthesis.

Measuring genome-wide splicing rates in vivo

Our method was first employed on exponentially growing wild type (BY4741) S. cerevisiae cells grown in synthetic medium at 22°C. Time points after 4tu addition were chosen at intervals as short as 30 seconds in order to maximize the number of splicing events for which sufficient data could be generated within the approach to steady state curves to enable high confidence model fits. We note that 22°C is a standard

temperature for yeast growth in the wild, and while many laboratory studies use 30°C to optimize growth rate, our initial experiments suggested that we could better model the kinetic data generated at 22°C (see also Methods and Discussion for further consideration). To measure the absolute abundance of each RNA species in each sample, a fixed amount of an exogenous pool of in vitro transcribed, 4-thiouridine (4su) containing RNAs was added to each sample as a spike-in control prior to RNA purification. As expected, the abundance of 4su-containing RNAs isolated from our total yeast RNA pools increased over time (Figure S1A), consistent with increasing incorporation of 4su in nascent RNAs. MPE-seq libraries were prepared from the 4su-containing RNA purified from each sample, and the libraries were sequenced using paired-end sequencing on an Illumina NextSeq platform. Roughly 10 million total reads were obtained for each of three replicate samples of each time point (Table S1), equivalent to ~1 billion reads of standard RNAseq per individual replicate sample (Gildea et al., 2019). Sequencing reads were aligned and those representing the pre 1st step, lariat intermediate, and spliced isoforms of each intron-containing gene were quantified. To account for variations in processing efficiency between time point and replicate samples, the counts for each species were normalized to the read counts corresponding to the spike-in control RNAs. Normalized reads for each species in the samples collected prior to addition of 4tu to the media were used to estimate and correct for the background signal for each RNA species in each sample. Consistent with the isolation of nascent RNA, the fraction of reads corresponding to unspliced RNA (pre 1st step + lariat intermediate) relative to spliced RNA per intron was highest in the early time points and decreased over time (Figure S1B).

To determine the rates of each chemical step in the splicing pathway, we used first-order kinetic models to fit the approach to equilibrium curves generated for each of the spliced isoforms. Simply stated, these models describe how the rate of increase in abundance of a specific RNA species over time becomes balanced by decay due to splicing and/or degradation (see methods section for a detailed description of the models and model fitting procedures). In order to determine the overall splicing rates, as well as the rates of the 1st and 2nd chemical steps for each splicing event, we established an initial set of definitions. First, the synthesis rate for each intron was defined as the number of complete introns synthesized per unit time. Second, the overall splicing rate was defined as the time from completion of synthesis to completion of the 2nd chemical step of splicing. Third, the rate of the 1st chemical step was defined as the time from completion of synthesis to completion of the 1st chemical step. And finally, the rate of the 2nd chemical step was defined as the time from completion of the 1st chemical step to completion of the 2nd chemical step. In addition to these guiding definitions, a constant time offset was included due to the delay in the time between the addition of 4-thiouracil (4tu) to the growth medium and the steady state availability of 4-thiouridine (4su) to the transcription machinery in the nucleus. Using this approach, the abundance of total unspliced counts were modeled to estimate the total rate of splicing, akin to other RNA-seq based kinetics studies (Eser et al., 2016). The abundance of pre 1st step and lariat intermediate counts were similarly modeled to estimate the rates of the synthesis, 1st and 2nd steps of splicing. Splicing rates are represented as half-lives throughout this manuscript. High confidence synthesis, total, 1st, and 2nd step rate estimates were obtained for the majority of introns in the genome (Figure 1B and Table

S2). Confidence intervals for parameter estimates were used to evaluate model fits (Table S2).

Genome-wide rates reveal a wide variation in splicing efficiency, and that the 2nd step is generally faster than the 1st step

We first assessed the global distributions of the half-lives for the total, 1st step, and 2nd step of pre-mRNA splicing. A median half-life of 135, 83, and 41 seconds was estimated for each of these steps, respectively (Figure 2A). Remarkably, these data revealed a >30-fold variation in individual splicing rates across the genome (Figure 2A). Additionally, a large variation in relative 1st and 2nd step rates were apparent between individual introns, including in some cases even between two introns housed within the same transcript. For example, both introns in the Rpl7A pre-mRNA were excised with similar overall splicing rates, however, that overall rate was achieved through contrasting 1st and 2nd step rates (Figure 1C), suggesting that the spliceosome and its substrates may have evolved similar total rates of pre-mRNA splicing through contrasting means. Nevertheless, when considering all spliceosomal substrates, no significant correlation was observed between the measured 1st and 2nd step rates (Figure S2).

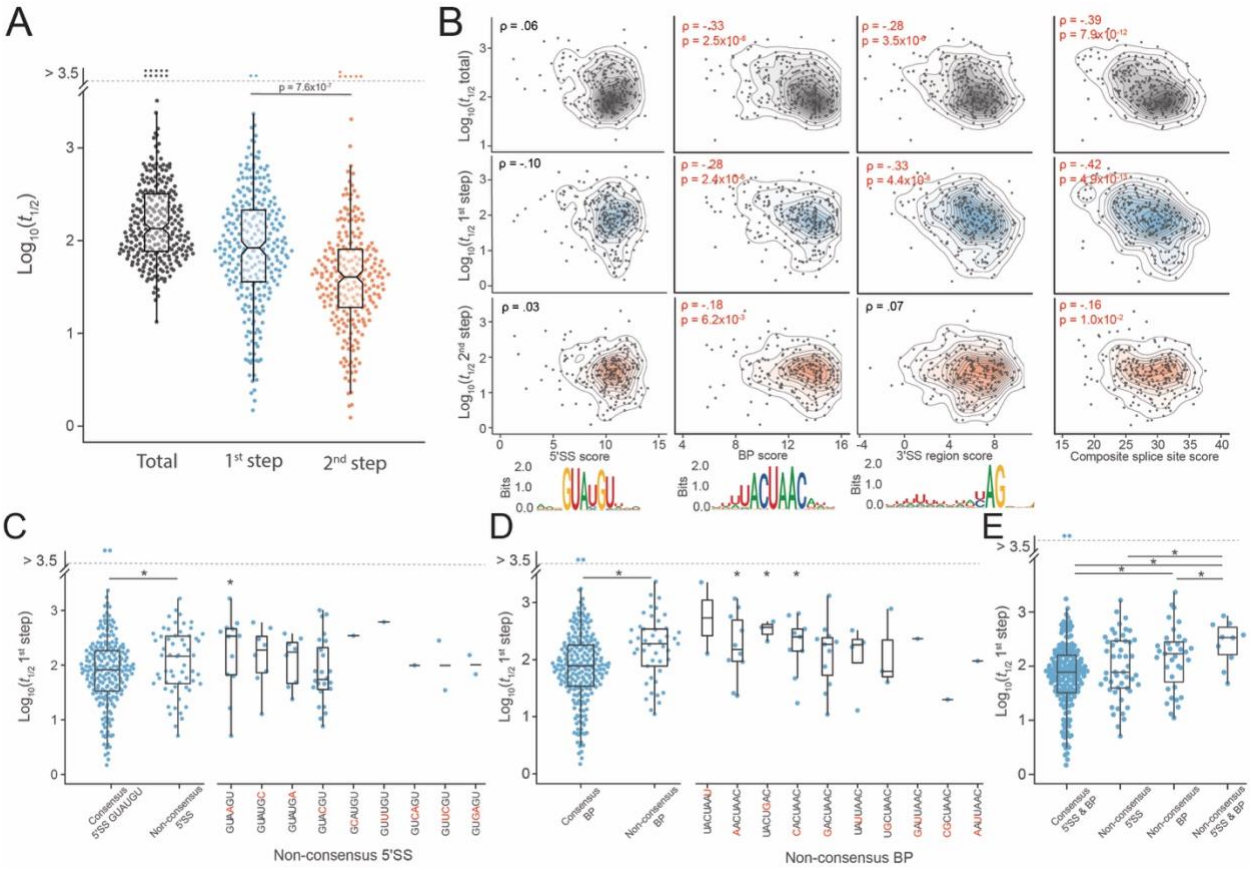


Figure 2: Splice Site Sequences are Major Determinants of the Rate of the 1st Step of pre-mRNA Splicing (**A**) Distributions of the half-lives for the total (n = 278), 1st step (n = 272), and 2nd step (n = 241) of pre-mRNA splicing. The 2nd step on average is significantly faster than the 1st step. (**B**) Comparison of 5′SS, BP, 3′SS, and composite splice site scores with total, 1st, and 2nd step rates. Web logos for 5′SS, BP, and 3′SS are included. Spearman correlation coefficients (ρ) and associated p-values (p) are included. Red text and highlighted axes indicate significant correlations ($p < 0.05$). (**C**) Comparison of 1st step splicing rates between introns with consensus (n = 213) and non-consensus (n = 59) 5′SS sequences. Non-consensus 5′SS sequences are further partitioned by sequence. (**D**) Comparison of 1st step splicing rates between introns with consensus (n = 227) and non-consensus (n = 45) BP sequences. Non-consensus BP sequences are further partitioned by sequence. (**E**) Comparison of 1st step splicing rates between introns with consensus 5′SS and BP sequences, introns with non-consensus 5′SS and consensus BP sequences, introns with consensus 5′SS and non-consensus BP sequences, and introns with non-consensus 5′SS and non-consensus BP sequences. Statistical significance in panels C-E was calculated with one-sided Mann-Whitney signed-rank test. Statistical significance for panel A was calculated with one-sided Wilcoxon signed-rank test. *p<0.05.

Cis-transcript features contribute to splicing kinetics

In order to understand what drives the observed variation in splicing rates, we investigated the influence of cis transcript features with overall and individual step splicing rates. We first investigated the influence of conserved cis-elements within introns that are recognized by components of the spliceosome and known to be crucial for efficient splicing, including the 5’ splice site (5’SS), branch point (BP), and 3’ splice

site (3'SS) sequences. To investigate the influence of these sequences we calculated position weight matrix (PWM) based scores for the 5'SS, BP, and 3'SS regions and compared them to the measured half-lives for the total, 1st, and 2nd steps **(Table S3)**. We observed a significant correlation between the scores of the BP and 3'SS region sequences and the half-lives of both the total and 1st steps of splicing **(Figure 2B)**. By contrast, only a mild correlation was observed between the BP scores and the half-lives of the 2nd step. Together, these data suggest that splice site strength is primarily a determinant of 1st step rates. The composite of all three splice site scores showed an increased correlation with the half-lives of the total and 1st steps, indicating that these sequences act to either additively or synergistically influence the 1st step splicing rate **(Figure 2B)**.

Surprisingly, a significant correlation was not observed between the 5'SS scores and the half-lives of the 1st step **(Figure 2B)**. However, we note that *S. cerevisiae* introns are defined by highly conserved splice site sequences, and over 75% of introns contain a GUAUGU sequence at their 5'SS. Indeed, we saw that those introns with the consensus 5'SS were spliced significantly faster than those with non-consensus 5'SS sequences **(Figure 2C)**, suggesting that the lack of correlation observed for the 5'SS scores likely reflects an absence of resolving power in the PWM scores. Interestingly, by further comparing individual non-consensus 5'SS and 1st step rates, we found that a substitution at the 4th position of the 5'SS from a U to an A, which results in perfect complementarity to the U1 snRNA, resulted in the longest 1st step half-lives of the non-consensus 5'SS for which there exists a sufficient sample size **(Figure 2C)**. By contrast, a substitution at the 4th position from a U to a C, which is the most common non-

consensus 5'SS and retains a mis-match with the U1 snRNA, showed no significant difference in 1st step half-life from the consensus 5'SS **(Figure 2C)**.

A similar analysis was performed for the BP consensus (UACUAAC) sequence, and as expected, non-consensus BP sequences resulted in a significantly longer 1st step half-life **(Figure 2D)**. We did not observe a significant difference in 1st step half-lives between individual non-consensus BP sequences **(Figure 2D)**, however small sample sizes may have obfuscated meaningful differences. Introns with both non-consensus 5'SS and BP sequences were spliced significantly slower than all other introns and the magnitude of the difference was greater than for introns with only one non-consensus sequence **(Figure 2E)**. The difference between introns with a non-consensus 5'SS and a consensus BP and those with both consensus 5'SS and BP was not statistically significant, likely because of differences in individual non-consensus 5'SS sequences **(Figures 2C and 2E)**.

Finally, we assessed the role of the 3'SS (YAG), as well as a short (six nucleotide) poly-U tract just upstream of the 3'SS. PWM scores were generated using bases -6 through -11 from the 3'SS for the poly-U tract, while the YAG plus 2 bases upstream and 3 bases downstream were used in determining the 3'SS score **(Figure S3A)**. The strength of both the poly-U tract and 3'SS showed a significant correlation with the half-lives of the 1st step of splicing but not the 2nd **(Figure S3B).**

Ribosomal protein genes are spliced faster at both steps

The set of intron-containing genes in S. cerevisiae can be partitioned into 2 functional classes: ribosomal protein genes (RPGs) and non-ribosomal protein genes (nRPGs). Of the roughly 280 annotated introns in S. cerevisiae, more than one third are

found within RPGs. RPGs are highly expressed under normal conditions, and in response to environmental changes or stress they have been shown to be coordinately regulated both at the level of transcription and splicing(Parenteau et al., 2011, Pleiss et al., 2007a, Parenteau et al., 2019, Gasch et al., 2000, Causton et al., 2001, Reja et al., 2015). Consistent with previous studies, we find that the total rate of splicing is fast for the RPGs relative to the nRPGs, with median half-lives of 84 and 218 seconds, respectively (Figure 3A) (Barrass et al., 2015). Additionally, a significantly higher synthesis rate was observed for RPGs compared to nRPGs (Figure 3B), consistent with their high level of expression. Whereas previous studies lacked the resolution to investigate the two individual steps of splicing, here we saw that splicing of RPGs was significantly faster at both the 1st and the 2nd steps of splicing (Figure 3A). Remarkably, the half-lives of the 1st and 2nd steps showed a significant negative correlation with one another within the RPG introns but not in the nRPGs introns, suggesting that RPGs may have evolved mechanisms to homogenize total splicing rates through optimization of different steps (Figure S4A). Consistent with their coordinated regulation, the variation in total splicing rates within the RPG class is considerably smaller when compared to the nRPGs.

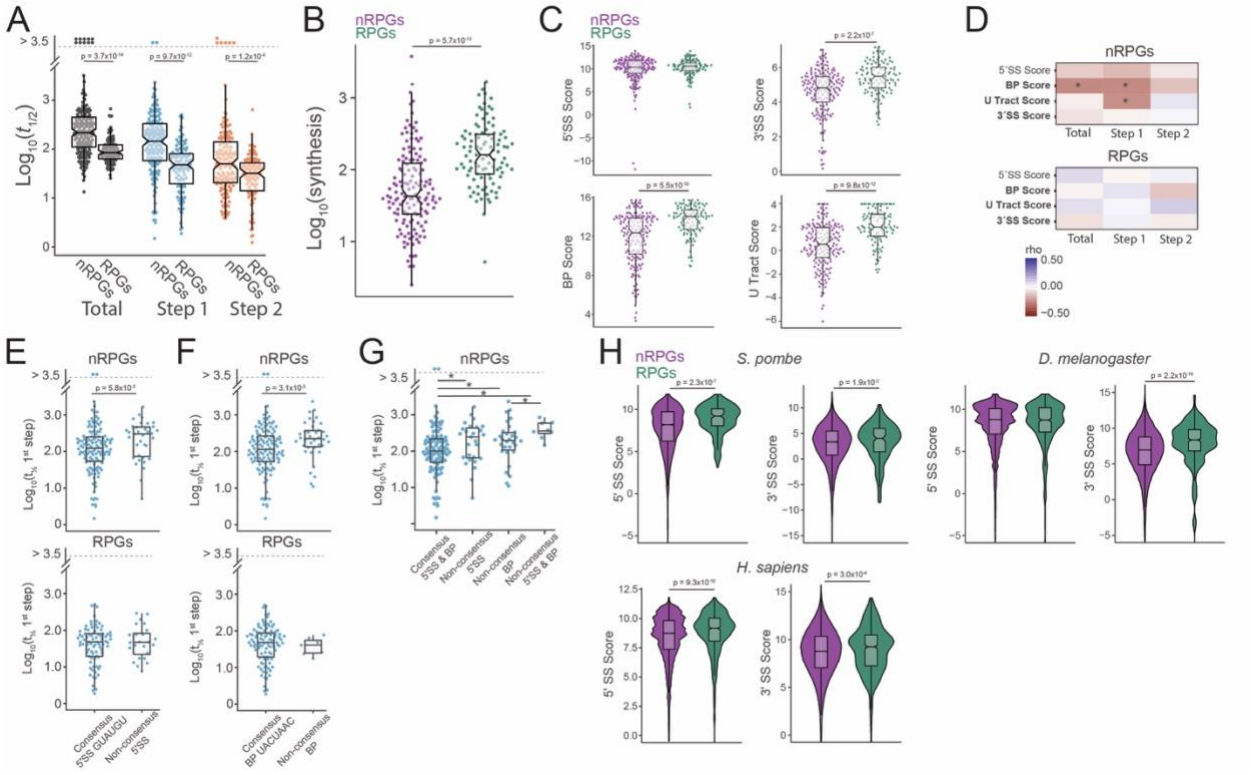


Figure 3: Both the Properties and the Splicing of RPG Introns is Distinct from nRPG Introns. (**A**) Distributions of the half-lives for the total (RPG n = 105, nRPG n = 173), 1st step (RPG n = 105, nRPG n = 167), and 2nd step (RPG n = 99, nRPG n = 142) of pre-mRNA splicing. (**B**) Synthesis rates for RPGs (n = 105) and nRPGs (n =173).(**C**) The PWM scores for 5′SS, BP, 3′SS, and U-tract regions compared between RPGs and nRPGs. (**D**) Spearman correlations between 5′SS, BP, 3′SS, and U-tract scores and half-lives for the total, 1st step, and 2nd step of pre-mRNA splicing for RPGs and nRPGs. (**E**) Comparison of the half-lives for the 1st step of splicing between introns with consensus (RPG n = 81, nRPG n = 137) and non-consensus (RPG n = 24, nRPG n = 40) 5′SS sequences. (**F**) Comparison of half-lives for the 1st step of splicing between introns with consensus (RPG n = 98, nRPG n = 135) and non-consensus (RPG n = 7, nRPG n = 42) BP sequences. (**G**) Among nRPG introns, a comparison of the half-lives of the 1st step of splicing between those with consensus 5′SS and BP sequences, those with non-consensus 5′SS but consensus BP sequences, those with consensus 5′SS but non-consensus BP sequences, and those with non-consensus 5′SS and BP sequences. (**H**) The PWM scores distributions for 5′SS and 3′SS among RPG and nRPG introns in *S. pombe* (RPG n = 134, nRPG n = 5160), *H. sapiens* (RPG n = 801, nRPG = 423110), and *D. melanogaster* (RPG n = 499, nRPG = 149895). Statistical significance in panels A-C, and D-H was calculated with one-sided Mann-Whitney signed-rank test. *p<0.05.

To better understand what drives the relative efficiency of the 1st and 2nd steps of RPG splicing and building off of previous observations that RPG introns have distinct features compared to nRPG introns (Spingola et al., 1999, Parker and Patterson, 1987), we asked how cis transcript features differed between RPGs and nRPGs. Notably, the RPGs as a class contained significantly stronger BP, 3'SS, and poly-U tract scores when compared to nRPGs, whereas no significant difference was observed in 5'SS scores **(Figures 3C and S4B)**. We then asked whether these features were correlated

with the observed splicing rates for the RPGs and nRPGs as independent groups. Interestingly, within the nRPGs, which as a class have lower scoring splice site sequences, we observed significant correlations between BP scores and the half-lives of the total and 1st step, and U-tract scores with the half-lives of the 1st step **(Figure 3D)**. By contrast, no correlations were apparent between the rates of RPG splicing and any of these features. Moreover, within the nRPG group, introns with non-consensus 5'SS were spliced significantly slower compared to those with consensus sequences, but somewhat surprisingly no such difference was seen within the RPGs **(Figure 3E)**; here we note, however, that the most common non-consensus 5'SS in RPGs is a variant harboring a C at the 4th position (15 of the 24 variants) which also showed little impact on 1st step half-lives in nRPGs. Similarly, introns with non-consensus BP sequences were also spliced significantly slower than those with consensus sequences within the nRPG class **(Figure 3F)**, whereas no such effect was seen in the RPGs, although here we note that the small population of RPGs bearing non-consensus BP sequences leaves the significance of this later result less clear. Nevertheless, within the nRPGs, an additive effect is observed in introns with both non-consensus 5'SS and BP as compared to those with only 1 non-consensus sequence **(Figure. 3G)**.

Together, the previous results suggested to us that RPGs in *S. cerevisiae* may have evolved strong BP, 3'SS, and poly-U tract sequences to facilitate efficient splicing, and we hypothesized that this may be a conserved feature of RPGs due to their critical role in all organisms. To investigate this possibility, we asked about the relative strengths of the 5'SS and 3'SS sequences in three other organisms with well-annotated genomes and well characterized intronomes: the fission yeast *Schizosaccharomyces*

pombe, *Drosophila melanogaster* and *Homo sapiens*. For *S. pombe*, we generated PWM scores for the 5'SS and 3'SS as described earlier (see also Methods), whereas for *D. melanogaster* and *H. sapiens* we used previously calculated values. Despite large differences in intron distribution and architecture between *S. cerevisiae* and each of these organisms, a similar trend was seen when comparing RPGs with nRPGs in each of these organisms. Like *S. cerevisiae*, the 3'SS scores for RPGs were significantly stronger in all three organisms as compared with the nRPG scores. **(Figure 3H and Table S4)**. Interestingly we also observed significantly stronger 5'SS scores in *H. sapiens* RPGs compared to nRPGs **(Figure 3H)**.

While these data support the notion that 5'SS and BP sequences are major determinants of 1st step splicing rates, these features alone cannot account for the large differences in 1st and 2nd step rates observed between the RPGs and nRPGs. We hypothesized that there may be cis features of RPGs that distinguish them from nRPGs and contribute to increased 1st and 2nd step splicing rates. Significant differences exist in both intron position and gene structure between RPGs and nRPGs **(Figure S4C)**. Specifically, RPGs have significantly shorter TSS to 5'SS distances and 3'SS to polyadenylation site (PAS) and shorter 3'UTR lengths when compared to nRPGs **(Figure S4C)**. However, RPGs have significantly longer introns compared to nRPGs **(Figure S4C)**, such that there is no significant difference in the unspliced transcript lengths between RPGs and nRPGs **(Figure S4C)**. Within introns, RPGs have significantly longer BP to 3'SS lengths and the nucleotide content in that region is significantly different with RPGs having a higher uridine and lower guanidine and cytidine content **(Figure S4C)**. In contrast, RPGs have a significantly lower fraction of

pyrimidines and higher fraction of purines in their full-length intron sequences **(Figure S4C)**.

We then examined these features relative to observed splicing half-lives within the context of the RPG and nRPG classes separately **(Figure 4A)**. Specific to the RPG class, we observed a significant negative correlation between the 3'SS to PAS lengths and the half-lives of the total splicing reaction **(Figures 4A and 4B)**, suggesting that splicing is inefficient when the 3'SS to PAS length is short. Specific to the nRPG class, we found significant negative correlations between the 1st step half-lives and: the length of the 5'UTR, the overall length of the intron, the distance from the 5'UTR to the 5'SS, and the distance between the BP and the 3'SS **(Figures 4A and 4C)**. Together these data suggest that longer distances from the TSS to the end of the intron promote more efficient 1st step splicing in nRPGs. Nucleotide content within the BP to 3'SS region correlated with splicing half-lives in both RPGs and nRPGs. Specifically, the fraction of adenosines in this region showed a significant positive correlation with the 1st step but a significant negative correlation with the 2nd step for both RPG and nRPG classes. The opposite effect was observed for uridine content **(Figures 4A and 4D)**. Surprisingly, within the RPG class, the synthesis rate showed a significant negative correlation with the half-life of the 1st step, but a significant positive correlation with the half-life of the 2nd step **(Figure 4A)**. Neither of these correlations was seen in the nRPG class **(Figure 4A)**.

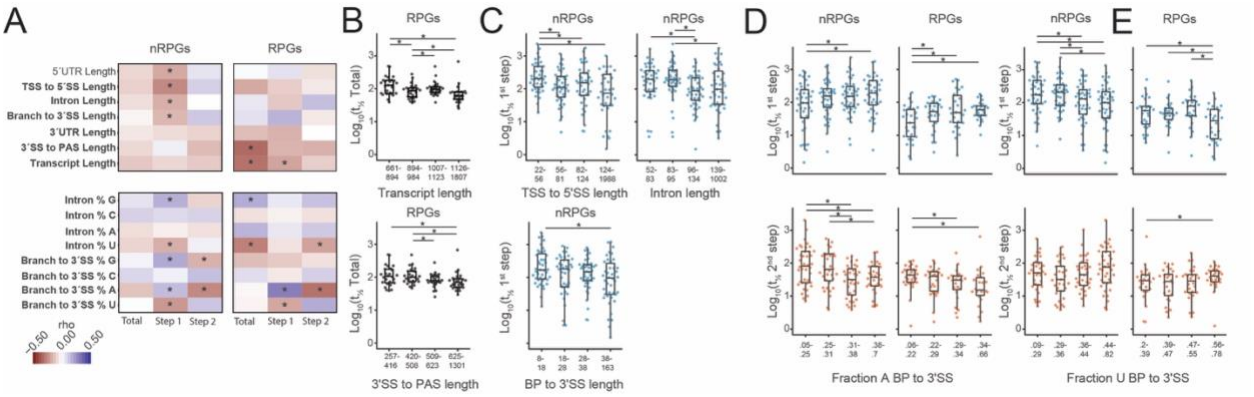


Figure 4: Distinct Transcript Features in RPGs and nRPGs are Correlated with Splicing Rates. (A) Spearman correlations between transcript features and the half-lives for the total, 1st step, and 2nd step of pre-mRNA splicing for RPGs and nRPGs. Bolded features show significant differences between RPGs and nRPGs shown in figure S4. (B) Quartiles of unspliced transcript length and 3′SS to PAS length versus the half-lives for total splicing within RPGs. (C) Quartiles of TSS to 5′SS, intron, and BP to 3′SS lengths versus half-lives of the 1st step of splicing within nRPGs. (D) Quartiles of fraction adenosine in the BP to 3′SS region versus half-lives for the 1st and 2nd steps of splicing for both RPGs and nRPGs. (E) Quartiles of fraction uracil in the BP to 3′SS region versus half-lives for the 1st and 2nd steps of splicing for both RPGs and nRPGs. Statistical significance was calculated with one-sided Mann-Whitney signed-rank test. * $p < 0.05$.

A genetic variant reveals expected impacts on the 1st step but unexpected impacts on the 2nd

To both confirm the biological significance of our measured rates and assess our ability to detect changes in splicing rates we re-measured global splicing rates in a strain harboring a mutation in the splicing factor Prp2, a DEAH-Box ATPase required for remodeling the spliceosome into a catalytically active state prior to the 1st chemical step. Cells harboring the prp2-1 allele are inviable at elevated temperatures and show a strong defect in genome-wide pre-mRNA splicing, consistent with its general role in spliceosomal activation. While these cells are viable at room temperature (22°C), their growth is impaired compared to WT, consistent with the strain exhibiting a modest molecular splicing defect even at the viable temperature (Pleiss et al., 2007b). To test the efficiency of 1st step splicing in this strain, we measured splicing rates when it was grown at the permissive temperature of 22°C, the same growth temperature that was used for the WT time course. Consistent with the role of Prp2 in the 1st step of splicing

we observed a significantly slower median half-life for the 1st step of splicing in the prp2-1 strain (286 sec) compared to WT (83 sec) (Figure 5A and Table S5). Surprisingly, we also observed a significantly slower median half-life for the 2nd step of splicing in the prp2-1 strain (233 sec) compared to WT (41 sec) (Figure 5A). Both trends were observed for the RPG and nRPG classes alike (Figure 5A), however RPGs were significantly more impacted at both steps when compared to nRPGs (Figure 5B).

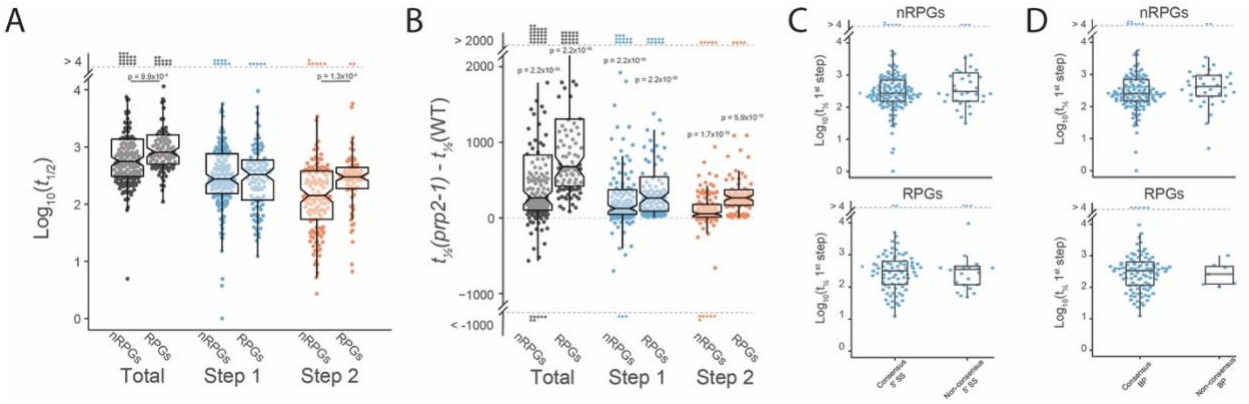


Figure 5: The *prp2-1* allele significantly slows the rate of splicing and differentially impacts RPG and nRPG processing. (**A**) Distributions of the half-lives for the total (RPG n = 105, nRPG n = 170), 1st step (RPG n = 104, nRPG n = 164), and 2nd step (RPG n = 79, nRPG n = 131) of pre-mRNA splicing in cells harboring the *prp2-1* allele. (**B**) Distributions of the differences in half-lives between WT and *prp2-1* cells for the total, 1st step, and 2nd step of splicing of RPG and nRPG introns. (**C**) Comparison of the half-lives for the 1st step of splicing between introns with consensus (RPG n = 80, nRPG n = 130) and non-consensus (RPG n = 24, nRPG n = 32) 5′SS sequences. (**D**) Comparison of the half-lives for the 1st step of splicing between introns with consensus (RPG n = 86, nRPG n = 126) and non-consensus (RPG n = 7, nRPG n = 38) BP sequences. Statistical significance was calculated with one-sided Mann-Whitney signed-rank test. * $p < 0.05$

Because Prp2 functions after intron recognition and spliceosome assembly but prior to the 1st catalytic step, we expected that those cis-features which were well correlated with the 1st step rate in WT cells would no longer be determinative of rate when the activity of *prp2-1* became limiting for the 1st step. Consistent with this hypothesis no difference in the half-lives of the 1st step was observed between introns with non-consensus splice site sequences (5'SS or BP) from those with respective consensus sequences in either the RPG or nRPG classes **(Figures 5C and 5D)**. Moreover, previously observed correlations between half-lives and BP, 3'SS, U-tract,

and 5'SS scores were no longer apparent in the *prp2-1* strain **(Figure 6A)**. We did, however, observe a significant negative correlation between 1st step half-lives and U-tract scores in nRPGs, albeit much milder than what was observed in WT cells **(Figures 3D and 6A).** Interestingly, we observed a positive correlation between U-tract scores and total splicing half-lives in RPGs **(Figure 6A)**. In addition, we saw the same effect when we assessed the correlation between U-tract score and the difference in half-lives between WT and *prp2-1* in RPGs, suggesting that RPG introns with stronger U-tract scores are more impacted by the *prp2-1* mutation **(Figure 6A)**.

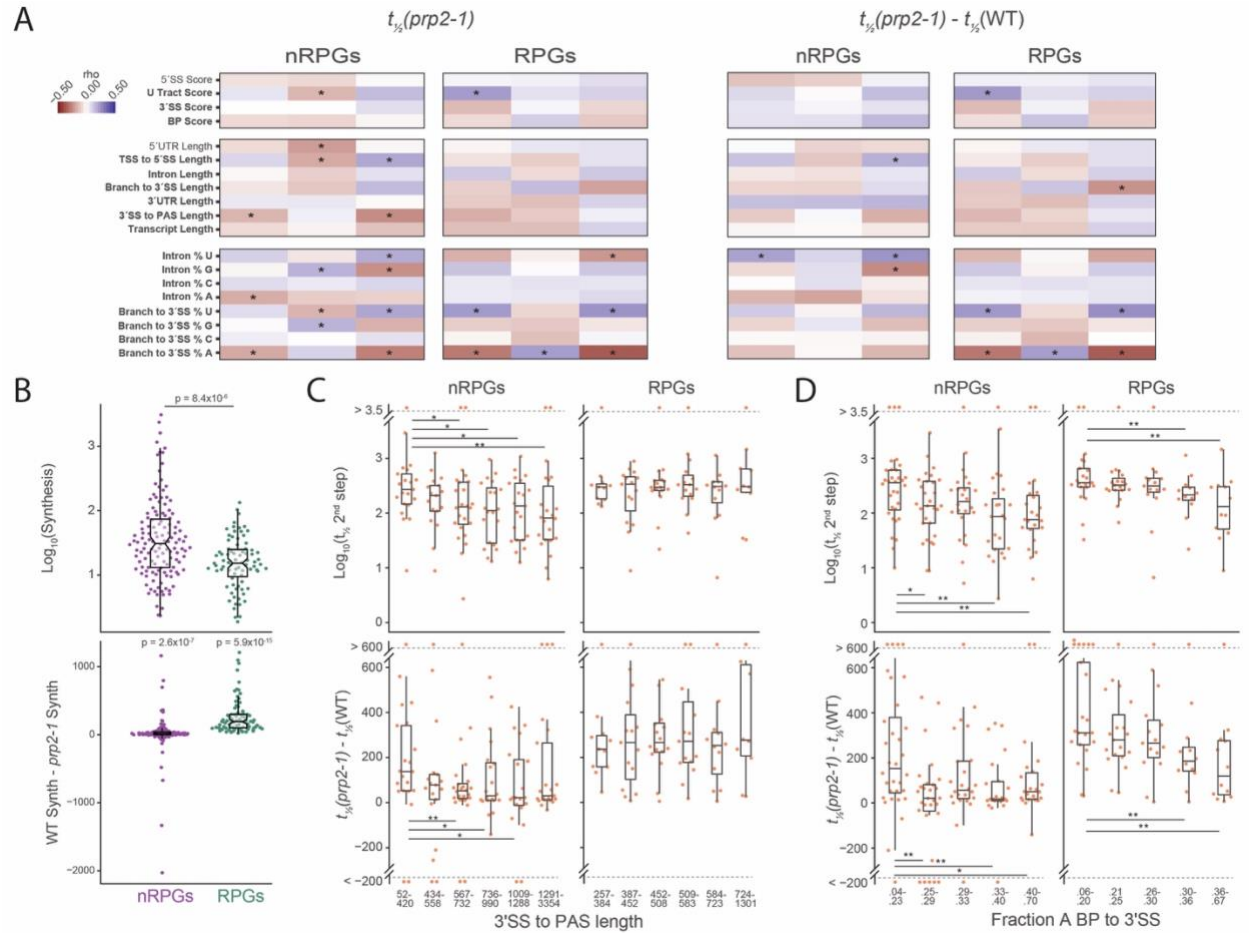


Figure 6: Transcript features correlate with magnitude of *prp2-1* impact on splicing rates. (**A**) Spearman correlations between transcript features and half-lives and half-live differences between WT and *prp2-1*. Data are partitioned into RPGs and nRPGs. Bolded features show significant differences between RPGs and nRPGs shown in figure S4. (**B**) nRPG and RPG synthesis rates in *prp2*-1. Synthesis rate difference between WT and *prp2-1.* (**C**) Sextiles of 3′SS to PAS length in RPGs and nRPGs versus 2nd step half-lives in *prp2-1* and 2nd step half-life

differences between WT and *prp2-1.* (**D**) Quintiles of fraction adenosine in the BP to 3′SS region in RPGs and nRPGs versus 2nd step half-lives in *prp2-1* and 2nd step half-life differences between WT and *prp2-1*.Statistical significance for panel B was calculated with one-sided Wilcoxon signed-rank test. Statistical significance for panels B, C, and D was calculated with one-sided Mann-Whitney signed-rank test. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Consistent with its slow growth phenotype, the synthesis rates of RPGs in the *prp2-1* strain were significantly slower when compared to their rates in WT cells **(Figure 6B)**. In fact, whereas RPGs showed faster synthesis rates than nRPGs in the WT strain, they showed significantly slower synthesis rates than nRPGs in the *prp2-1*-containing strain. By contrast, nRPG synthesis rates were largely unchanged from WT **(Figure 6B).** Importantly, a significant negative correlation was observed between the half-lives for the 1st step of splicing and synthesis rates of RPGs. Likewise, the differences in half-lives between WT and *prp2-1* cells for the 1st step of splicing of RPG introns showed a significant negative correlation with synthesis rates. These data indicate that RPGs are significantly repressed via transcription in the *prp2-1* strain, and those transcripts whose synthesis rates are more impacted in this strain are also more impacted at the 1st step of splicing. Interestingly we no longer observed a significant negative correlation between transcript lengths and 1st step rates in RPGs in *prp2-1* **(Figure 6A)**. However, we observed a significant negative correlation between 3'SS to PAS lengths and 2nd step half-lives in nRPGs, suggesting that introns within transcripts with longer 3'SS to PAS lengths undergo the 2nd step faster than those with short 3'SS to PAS lengths **(Figures 6A and 6C)**. No similar correlation was seen in RPGs; however we note that RPGs as a class have significantly shorter 3'SS to PAS lengths compared to nRPGs **(Figures 6A and 6C and S4C)**.

Similar to observations in the WT data, many of the same correlations were observed in the *prp2-1* strain between nucleotide content in the BP to 3'SS regions and 1st and 2nd step half-lives. Specifically, a significant negative correlation was observed

between the fraction of adenosines in this region and the 2nd step half-lives in both RPGs and nRPGs **(Figures 6A and 6D)**. Additionally, significant negative correlations were observed between the fraction of adenosines in the BP to 3'SS region and the difference in the 2nd step half-lives between the WT and *prp2-1* strains in both the RPG and RPG classes **(Figures 6A and 6D)**. To a lesser extent, the inverse was observed with the fraction of uridines in this region.

Discussion

The process of pre-mRNA splicing is pervasive throughout eukaryotes and provides an important control point for regulating gene expression. The capacity of the spliceosome to efficiently process full complement of substrates is poorly understood, yet knowledge of this will be critical for solving problems of both basic and clinical importance. Here we have combined rapid metabolic labeling and first order kinetic modeling with a targeted sequencing approach termed MPE seq to gain high resolution information about the efficiency by which this enzyme processes its full complement of substrates in the budding yeast *S. cerevisiae* **(Figures 1A and 1B)**. While the basic machinery that catalyzes intron removal is highly conserved across eukaryotes, in yeast the intron structures and splicing patterns have greatly simplified over time, facilitating a comprehensive analysis of the splicing targets (Spingola et al., 1999, Wahl et al., 2009). Whereas others have taken a similar approach to understanding this problem in a variety of organisms including yeast, the resolution provided by MPE-seq in the current work enables unmatched resolution of the temporal, global, and chemical aspects of this process (Barrass et al., 2015, Eser et al., 2016, Pai et al., 2017, Wachutka et al., 2019, Windhager et al., 2012, Rabani et al., 2014).

Splicing efficiency is highly variable across the genome-wide complement of substrates

An important and somewhat unexpected conclusion from the data presented here is that there exists a great variability in the efficiency with which the spliceosome processes its global complement of substrates, even within the relatively simplified system present in budding yeast. In considering this statement, it seems important to acknowledge certain limitations inherent to experiments such as those presented here, in particular the ability to accurately measure very fast and very slow events. Although the time points included in our experiments were selected because they optimized the fraction of splicing events that would be well sampled within our data, there remain some number of events for which the accuracy of our measurements is lower than desired. Nevertheless, the window through which we can robustly measure these rates constitutes a >30-fold difference between the fastest (with half-lives ~30s in a WT strain) and slowest (with half-lives ~15m) events (Figure 2A and Table S2). Indeed, our data reveal a continuum of rates for splicing ranging between and beyond these limits, highlighting the dramatic variability of speeds through which different introns transit this process (Figure 2A and Table S2). Importantly, these experiments demonstrate that this variability isn't the result of variability in just one of the chemical steps, rather significant variation is apparent for both chemical steps of splicing (Figure 2A). While a global view of our data shows that the 2nd step is roughly twice as fast as the first step, suggesting that the 1st step is generally rate limiting, there are many splicing events for which the 2nd step is slower than the 1st step, suggesting splicing has been optimized differently for different substrates (Figures 1C and 2A).

Cis-elements are important for, but not fully determinative of, splicing efficiency

Importantly, the kinetic measurements reported here are consistent with the long-held idea that splice site sequences play an important role in facilitating splicing efficiency. Indeed, the efficiencies of the 1st chemical step measured here are strongly correlated with the strengths of the splice site sequences, both individually and in composite **(Figure 2B-E)**. Here again, however, it is important to note that while the use of Position Weight Matrix scores enables a powerful approach for comparing the relative activities of these substrates, there are limitations and caveats to such an approach. First and foremost, it is important to emphasize that scores generated in this way are not direct measures of activity *per se*, but rather are measures of frequency; as such a rare but otherwise efficient sequence would appear as low scoring in this approach. Moreover, as noted earlier, the high similarity of sequences found within *S. cerevisiae* introns, particularly at the 5'SS, reduces the capacity of this approach to effectively differentiate between activities. Nevertheless, and these caveats notwithstanding, we note that splice site sequences alone are insufficient to fully the explain complement of 1st step rates determined here. This fact is perhaps best exemplified by the number of introns with fully canonical splice site sequences which nevertheless show disparate splicing efficiencies **(Tables S2 and S3)**. Similarly, the absence of strong correlations between splice site sequences and 2nd step efficiencies highlights the important roles that other substrate features must play in driving splicing rates **(Figure 2B)**.

Beyond the simple global analyses, our analyses of individual variants also provide insights into mechanisms of spliceosomal activation by revealing the differential impact of non-consensus splice site sequences on splicing efficiency **(Figures 2B-2E)**. This was particularly apparent for variants of the 5'SS and their impact on the 1st step of

splicing. Among introns with a non-consensus 5'SS, those with a U to A substitution at the 4th position had the slowest 1st step splicing half-lives, whereas those with a C substitution at the 4th position showed no difference from introns with consensus sequences (Figure 2C). The consensus 5'SS in budding yeast is perfectly complimentary to the U1 snRNA, except at the 4th position **(Figure 2B)**. By contrast, the canonical U at this 4th position is complementary to the U6 snRNA in the B_{act} complex (Zhang et al., 2019). Interestingly, while both the A and C mutations result in a mismatch with the U6 snRNA, an A at the 4th position of the 5'SS results in perfect complementarity to the U1 snRNA, whereas a C maintains the mismatch. The observation here that U4A variants are inefficiently spliced is consistent with previous studies that showed that extending the regions of complementarity between the U1 snRNA and the 5'SS beyond the canonical, 6-nucleotide region resulted in inefficient splicing (Staley and Guthrie, 1999). Indeed, the data presented here suggest that increased stabilization of the 5'SS:U1 snRNA duplex even within this 6-nucleotide region may be sufficient to impede the capacity of Prp28 to disrupt this region in the pre-B to B complex transition. It is more difficult to draw conclusions about the importance of the 5'SS:U6 snRNA duplex on the basis of these experiments. Whereas the observation here that U4C variants are spliced with similar efficiency to consensus substrates might mean that the 5'SS:U6 snRNA duplex is tolerant to mismatches, it is also possible that these mis-matches could become rate-limiting under different cellular conditions **(Figure 2C)**. By contrast with the 5'SS sequences, we did not observe a differential impact between the compliment of non-consensus BP sequences on 1st step splicing half-lives

(Figure 2D), although low sample numbers could be obfuscating meaningful differences.

The data presented here further reveal the presence of a small, conserved region of uridines just upstream of the 3'SS of some introns which is correlated with faster 1st step splicing **(Figures 3D and S3A and S3B)**. Budding yeast are generally considered to lack a canonical poly-pyrimidine tract (pY) in that no strong pY sequence element is apparent between the BP and 3'SS of most introns, and there is no homolog of U2AF, the protein responsible for pY binding in other eukaryotes (Kupfer et al., 2004, Spingola et al., 1999, Lopez and Séraphin, 1999, Parker and Patterson, 1987). However, *S. cerevisiae* contains a functional equivalent of $U2AF^{65}$, Mud2p, which together with the branchpoint binding protein (BBP) binds to the BP and downstream region in the initial steps of spliceosome assembly (Abovich et al., 1994). Our data suggest that this U-tract region may function similarly to the pY tract in other eukaryotes.

Beyond the U-tract element, our data also suggest that overall nucleotide content within the BP to 3'SS region may be important for splicing efficiency at both steps. Introns with high A content in this region tend to be processed slowly through the 1st step but fast through the 2nd step, whereas the inverse was observed with U content **(Figures 4A and 4D)**. To be sure, it is difficult to deconvolute the presence of a strong U-tract element from the more general property of A and U content within the region, leaving it unclear the mechanisms by which nucleotide content within this region might influence splicing rates. Nevertheless, it is important to note that throughout the splicing cycle this region is in close proximity to many spliceosomal proteins which might impart selective behavior on select transcripts based on these properties (Will and Lührmann,

2011, Fica and Nagai, 2017, Wahl et al., 2009). Similarly, it is well established that secondary structures can impact splicing efficiency, and these correlations may reflect increased or decreased capacity for formation of such structures (Warf and Berglund, 2010, Barrass et al., 2015).

Importantly, our work with a strain containing a conditional genetic variant of the essential splicing factor Prp2 further reinforces the role of cis-regulatory elements in splicing activity. Prp2 acts just prior to the 1st chemical step of splicing after recognition of the 5'SS and BP sequences and assembly of the spliceosome (Wahl et al., 2009). As expected, we observed a global slowing of 1st step splicing rates **(Figures 5A and 5B)**. However, consistent with the late action of Prp2 relative to splice site recognition, we no longer observed a correlation between splice site strength and 1st step rates, suggesting that the rate limiting step is Prp2 function rather than splice site recognition **(Figures 5C and 5D)**.

Evolution has tuned intronic and genic features to facilitate splicing of classes of transcripts

The relative contribution of splice site sequences in splicing efficiency is perhaps most notable when considering the processing of different classes of transcripts. It has long been known that the ribosomal protein genes (RPGs) constitute a unique category within the subset of intron-containing genes in budding yeast. Whereas only ~5% of all genes contain introns, nearly 70% of RPGs contain them. Indeed, because of the high transcriptional frequency of RPGs it has been noted that nearly 35% of all transcriptional flux requires processing by the spliceosome, in spite of this low population of introns within the genome (Warner, 1999). Moreover, it was long ago

noticed that certain properties of RPG introns were distinct from those in nRPGs, notably that intron lengths are markedly longer in the RPGs (Spingola et al., 1999, Parker and Patterson, 1987). Here we further note that the combination of the long intron lengths but relatively short coding and UTR lengths of RPG results in transcripts whose overall unspliced lengths are virtually indistinguishable from nRPGs, but importantly where the 3'SSs are both farther from the transcription start sites (TSSs) and closer to the cleavage and polyadenylation sites (PAS) (Figure S4). Moreover, we also show here that RPGs as a class are characterized by stronger scoring splice site sequences than nRPGs at the 5'SS, BP, U-tract, and 3'SS regions in budding yeast introns (Figure S4). Importantly, while far from a rigorous evolutionary analysis, an examination of three divergent organisms for which robust genome-wide information about intronic sequence elements is available suggests that strong splice site sequences may be an evolutionarily conserved property of RPGs.

Our data further show that RPG introns transit through both chemical steps of splicing faster than nRPG introns (Figure 3A). Surprisingly however, and somewhat paradoxically, we did not detect any significant correlation between the splicing rates for RPGs and any of the cis-features within their introns. By contrast, within the relatively weaker scoring nRPG introns, 1st step rates were strongly correlated with both 5'SS and BP scores (Figures 3D-3E). Additionally, the combination of nRPG introns with non-consensus 5'SS and BP sequences showed an additive impact on 1st step half-lives (Figure 3G). Moreover, U-tract score correlated strongly with 1st step half-lives in nRPGs, whereas no significant correlation was observed in RPGs.

How then to understand the presence of such strong splice site sequences within the RPG introns, and the high efficiency with which they are processed, but the apparent insensitivity of the spliceosome to the non-consensus variants within this class? As described below, we suggest that the data presented here are most consistent with a model wherein the overall architecture of the ribosomal protein genes has been tuned to enable efficient splicing, in particular by connecting their splicing to transcription, such that under optimized growth conditions the spliceosome is insensitive to suboptimal splice site sequences. Presumably the strong splice site features remain positive determinants of efficient RPG splicing under suboptimal growth.

Two main observations drive this hypothesis. First, the data presented here show that transcript length upstream of the 3'SS is a general effector of 1st step splicing rates, suggesting that RPGs may have selected for long introns to maximize 1st step splicing efficiency (Figures 4A and 4C). Mechanistically, increased length may provide additional time for association of the U1 snRNP and other early spliceosome components with the nascent transcript and transcription machinery, allowing for quicker progression through the splicing cycle once the full intron has been synthesized. Second, RPGs as a class are subject to higher transcription frequency, measured here as synthesis rates (Figure 3B). While it remains poorly understood how transcriptional activity and splicing rates are coupled, over the past decade increasing evidence has accumulated pointing to a role for liquid phase separation in gene expression, in particular during the early stages of transcription (Boehning et al., 2018, Lu et al., 2019, Harlen and Churchman, 2017, Hnisz et al., 2017). We propose that the high transcriptional activity of RPGs in WT cells is accompanied by liquid phase-separated

compartments containing higher concentrations of splicing factors leading to fast splicing rates relative to nRPGs. A comparison of splicing rates in WT cells versus those harboring the prp2-1 variant further supports this idea. Here, a strong correlation is observed between the change in transcriptional frequency and the change in splicing rate: as transcription rate slows, so does splicing. Decreased transcription is presumably accompanied by a reduction of these phase-separated compartments, leading to slower splicing.

Splicing is fast, but occurs over the length of the transcript, completing when polymerase is thousands of bases downstream

While the experiments presented here evaluate the kinetic properties of splicing intermediates, and as such do not directly evaluate the position of RNA polymerase with respect to this process, as noted in the previous section much can be inferred from these data about the relationship between splicing and transcription. Importantly, while the current inconsistencies within the literature regarding the relationship between polymerase location and splicing status might have been discounted as merely species-specific differences in the pathways of spliceosome assembly and chemistry, the data presented here for budding yeast are more consistent with the observations from higher eukaryotes wherein spliceosome assembly is presumed to occur in a co-transcriptional manner across the length of the downstream exon, with the chemistry of splicing completing at a position thousands of nucleotides downstream of the intron. Three different aspects of our data drive this conclusion.

First and foremost, the splicing half-lives measured here are on the order of minutes, with a median half-life of just over two minutes. While estimates of the

elongation rate of RNA polymerase vary by organism and growth condition, it is generally accepted that this occurs at a rate of roughly 1-4 kilobases per minute (Mason and Struhl, 2005, Oesterreich et al., 2011). To be sure, elongation rates are likely to be on the slower end of these estimates for yeast growing at 22°C, but even at the low end of this spectrum the polymerase is expected to be more than two kilobases downstream of the intron when the median splicing event completes. Moreover, while RNA polymerase does not transcribe at a uniform rate through a gene and transient pausing has been observed, the lifetime of transcriptional elongation pauses are measured in seconds, or fractions thereof, and as such are unlikely to significantly impact the location of the polymerase relative the times measured here (Oesterreich et al., 2011, Kwak et al., 2013).

Second, as noted above, the overall splicing rates presented here show a strong correlation with splice site strengths, cis-elements long established to influence splicing efficiency **(Figures 2B-E)**. The poor annotations of global splice site sequences, particularly BPs, in most higher eukaryotes has presumably precluded a careful assessment of the correlations between these features and polymerase location in earlier studies. However, we note that we detect no correlation between splice site strengths and the median splice distance in the yeast studies that report the completion of splicing in close proximity to the 3'SS (Oesterreich et al., 2016).

And finally, the combination of our measurements in both WT and *prp2-1* cells reveals an important correlation between the length of the transcript between the 3'SS and the PAS and the efficiency of splicing, wherein the longer the distance the more efficient the splicing. Such an observation would not be expected if splicing was

completed very rapidly in relation to transcription of the 3'SS. Rather, the simplest explanation for this observation is that the splicing process is facilitated by its connection with the polymerase, and the longer the region downstream of the intron but upstream of the cleavage site the longer-lived is the connection with the polymerase. Indeed, our observation that RPG transcripts with long 3'SS to PAS lengths are spliced significantly faster than short ones suggests that the more time in which the nascent RNA is engaged in transcription the more efficient the splicing becomes **(Figures 4A and 4B)**. Moreover, while no correlation was detected between 3'SS to PAS length and splicing of nRPG introns in WT cells, the observation in the *prp2-1* strain that RPG introns as a whole and the subset of nRPG introns with short 3'SS to PAS lengths were significantly slowed at the 2nd step compared to WT again points to this important relationship **(Figure 6C)**. Specifically, nRPG introns with 3'SS to PAS lengths of less than roughly 750nt were slowed in a length dependent manner, while those with 3'SS to PAS lengths greater than 750nt were not **(Figure 6C)**. Importantly, nearly 90% of RPGs have 3'SS to PAS lengths less than 750nt.

Data and Code Availability

Raw sequencing data is available at NCBI GEO at GSE159665. Python and r code used for processing and analyzing data in this study can be found at https://github.com/mgildea87/4tu_MPEseq

Methods

Strain growth and 4tu time course

BY4741 (WT) and *prp2-1* cells were streaked onto YPD plates from glycerol stocks stored at -80°C and grown for roughly 48 hours at 30°C. Five colonies were

inoculated into 50 mL of liquid YPD medium in 250mL flasks and grown in a shaking incubator at 30°C overnight. Triplicate cultures were started for each strain by back diluting the saturated overnight culture to an OD_{600} of 0.1 in 600mL of fresh complete synthetic medium containing 178 µM uracil in a 2.8L flask. Cultures were grown in a shaking incubator at 30°C. After one doubling, cultures were shifted to 22.5°C for an additional two doublings. 4-thiouracil was added to a final concentration of 500 µM to the log phase cells to start the time course. At each time point, 50mL of culture was extracted, vacuum filtered, and flash frozen in liquid nitrogen. Extracted cells were stored at -80°C. Time 0 samples were collected before addition of 4-thiouracil.

In vitro transcription of 4sU labeled spike-ins

PCR primer sets were designed to amplify five *S. pombe* genes **(Table S6).** The T7 promoter sequence was appended with PCR such that the resulting amplified DNA could be directly used in *in vitro* transcription. After an initial denaturation for 30 seconds at 98°C, 40 cycles of PCR were performed on a 200 µL reaction for each primer set, including 400 ng of *S. pombe* genomic DNA as template, Phusion polymerase (Thermo Fisher) with buffer, a final concentration of 200 nM each primer, and 200 µM dNTPs. Each cycle consisted of 10 seconds at 98°C, 20 seconds at 62-65°C, followed by 30 seconds at 72°C. This was followed by a final extension for 5 minutes at 72°C. All reactions were run on 0.8% agarose gels and bands were gel purified (Invitrogen Purelink) followed by ethanol precipitation. *In vitro* transcription was performed using the NEB HiScribe T7 high yield synthesis kit following the manufacturers protocol using 400 ng of each DNA template in 20 µL reactions. RNA was 4-thiouracil labeled by including 4-thiouridine-5'-triphosphate (4tUTP) at 1:3 with uridine-5'-triphosphate (UTP).

1:3 was empirically chosen because it afforded maximum purification efficiency of 4sU labeled spike-in transcripts compared to other tested ratios (data not shown). RNA was phenol chloroform purified following the manufacturers protocol. The length of 4sU labeled spike-in transcripts was assayed using the 2100 Bionalyzer (Agilent). The vast majority of transcripts were the expected full length (data not shown). 4sU labeled spike-ins were pooled at equal mass and stored at -80°C in 10 mM Tris-HCl pH 7.4, 1 mM EDTA.

RNA Extraction

RNA was extracted by adding 2 mL acid phenol (pH 5.3) to each cell pellet followed by 2 mL of AES buffer (50 mM sodium acetate pH 5.3, 10 mM EDTA, 1% SDS). Samples were incubated for seven minutes at 65°C with periodic mixing via vortexing. This was followed by a five minute incubation on ice. Samples were transferred to 15 mL PLG tubes and spun for five minutes at 3000g. 2 mL of phenol:chloroform:IAA (25:24:1) was added to the aqueous phase in each tube and samples were mixed and centrifuged for five minutes at 3000g. To remove excess phenol, one more extraction performed with 2 mL chloroform was performed. RNA was then isopropanol precipitated by addition of 200 µL 3 M Sodium acetate (pH 5.3) followed by 2.5 mL isopropanol. RNA was pelleted, washed twice with 70% ethanol, dried, and dissolved in 10 mM Tris-HCl (pH 7.4), 1 mM EDTA.

Biotin coupling to 4-thiouracil labeled RNA

For each sample, 400 µg of total RNA was mixed with 2.5 ng of 4-thiouracil labeled spike-in mix. Biotin coupling was performed following a previously published protocol(Dolken et al., 2008). 100 µL of 100 mM Tris-HCl (pH 7.4), 10 mM EDTA was

added to each sample followed by DEPC water up to 800 µL. This was followed by the addition of 200 µL of HPDP-biotin (1 mg/mL in dimethyl formamide). Each sample was mixed well and placed on a rotator at room temperature in the dark for three hours. Two 1 mL chloroform extractions in 2 mL PLG tubes were performed to remove excess unreacted HPDP-biotin. Biotin coupled RNA was isopropanol precipitated and RNA pellets were dissolved in 100 mL of 10mM Tris-HCl (pH 7.4), 1 mM EDTA.

Biotin purification

Biotin purification was performed according to a previously published protocol (Dolken et al., 2008). 100 µL of streptavidin beads (Dynabeads C1 Invitrogen) per sample were transferred to tubes and placed on a magnetic stand for one minute. The supernatant was aspirated, and the tubes were removed from the magnetic stand. The beads were washed by resuspending the pellet in 1 mL of bead wash buffer (100mM Tris-HCl pH 7.4, 10mM EDTA, 1M NaCl, 0.1% Tween) followed by mixing via pipetting. The samples were placed back on the magnetic stand for 1 min and the buffer was aspirated. Washing was repeated 2 more times for a total of 3 washes. Beads were resuspended in a 100 µL per sample volume of bead binding buffer (10 mM Tris-HCl pH 7.4, 1 mM EDTA, 2 M NaCl, 0.1% Tween) and mixed well by pipetting. 100 µL of resuspended beads were added to each biotin coupled RNA sample and placed on a tube rotator in the dark for 30-min at room temperature. Beads were washed 3 times with bead wash buffer pre-warmed to 65°C and three additional washes with room temperature bead wash buffer. To elute RNA, beads were resuspended in 100 µL of freshly prepared 100 mM DTT. Samples were mixed well and incubated at room temperature for 2-min. Eluted RNA was aspirated and transferred to a new tube. The

elution was repeated for a total of two elutions. Each RNA sample was purified by adding 1400 µL of binding buffer (2 M Guanidinium-HCl, 75% isopropanol), mixed well by vortexing, and transferred to Zymo-spin I columns. Samples were spun at 14k x g for one minute. This was followed by 2 washes with column was buffer (10 mM Tris-HCl pH 8.0, 80% ethanol). Columns were dried by spinning at 14K x g for an additional one minute. Purified RNA was eluted by adding 16 µL DEPC water. RNA was collected by spinning at 14k for one minute. RNA concentration was assayed using Qubit.

MPE-seq library preparation and sequencing

MPE-seq libraries were prepared as previously described with a few key differences (Gildea et al., 2019). 14 µL of each 4-tu purified RNA sample was included in the reverse transcription (RT) reaction. Five additional MPE-seq RT primers targeting the five spike-in RNAs were added to the S. cerevisiae pool at equimolar concentration (80 nM) (Table S6). Libraries were barcoded via PCR amplification and pooled. Paired-end sequencing was performed on the NextSeq 500 (Illumina) with a read 1 (P5) length of 61-bp and a read 2 (P7) length of 17-bp. Sequencing was performed by the BRC Genomics Facility at Cornell University.

MPE-seq data analysis

Read processing and alignment

Reads were demultiplexed by the sequencing center. First, overall read quality was assessed by running FastQC on each fastq file(Andrews, (n.d)). Illumina sequencing adapter sequences were trimmed from reads using version 0.20.0 of fastp (Chen et al., 2018) with the following parameters:

```
 --disable_trim_poly_g --adapter_sequence
CTGTCTCTTATACACATCT --adapter_sequence_r2
CTGTCTCTTATACACATCT
```

Reads were aligned to the yeast genome (reference genome assembly R64-1-124 (Engel et al., 2014)) with version 2.7.2b STAR (Dobin et al., 2013) with the following parameters:

```
--clip5pNbases 7 0 --peOverlapNbasesMin 5 --
peOverlapMMp 0.1 --outFilterMultimapNmax 1 --
alignIntronMin 10 --alignIntronMax 1100 --
outSAMattributes All --runThreadN 8
--outSAMunmapped Within KeepPairs --alignSJoverhangMin
3 --alignSplicedMateMapLmin 3 --alignMatesGapMax 3000
--alignEndsType EndToEnd
```

Only concordantly aligned reads were considered for further analysis except for calculation of the spike-in normalization factor wherein total on target reads were used (concordant + non-concordant).

Estimating splice isoform abundance

Reads derived from targeted spliced transcripts were identified and counted based on the presence of an 'N' in their CIGAR string and an extension of at least 3 bases across the splice junction into the upstream exon. Reads derived from unspliced transcripts were identified by a concordantly mapped read pair in which read 1 aligned to a primer target site and extended into an intron and was on the appropriate strand. Each intron was considered individually. For example, if a read pair aligned to 2 or more introns or splice junctions within a multi-intronic gene, only the targeted intron or splice junction was counted based on the position of read 1. Unspliced reads were further partitioned into lariat intermediate, pre-1st step, and ambiguous unspliced based on the position of read 2. Lariat intermediate reads were those whose 1st mapped base of read 2 aligned within an 8 base pair window around the annotated branch point adenosine.

This window used was 5 bases downstream and 3 bases upstream of the annotated branchpoint adenosine. Ambiguous reads were those whose 1st mapped base of read 2 aligned downstream of the lariat intermediate window and the 3'SS. Pre-1st step reads were than quantified by subtracting the sum of lariat intermediate and ambiguous unspliced from total unspliced reads. Finally, ambiguous reads were assigned to lariat intermediate or branched based on the ratio of unambiguous lariat intermediate to pre-1st step reads. To remove reads derived from primers not extended by RT, read pairs with insert sizes less than 33nt were discarded. Isoform quantification was performed using a custom python script that can be found in the Data and Code Availability section.

Read normalization

To monitor changes in absolute abundance of each RNA isoform over time read counts were normalized to *S. pombe* spike-in counts.

1. For each spike-in *k* in each replicate time course *j*, for each time point *t*. The fraction spike-in counts *A* of total counts:

$$A_{tjk} = \frac{S_{tjk}}{\sum_{i=1}^{n_{tj}} C_{tji} + \sum_{i=1}^{k_{tj}} S_{tji}}$$

   Where *k* = 1-5 for each spike-in, *j* = 1-3 for each time course replicate, *S* = spike-in count, *C* = target intron counts, and *n* = intron targets.

2. We then computed the ratio *B* of each *A* to the *A* of time point 0 t_0 in each replicate time course for each spike-in:

$$B_{jkt} = \frac{A_{jkt_0}}{A_{jkt}}$$

3. To remove outliers (spike-in counts that behaved far from the norm) we combined *B* values across replicates *j* and spike-ins *k* for each time point *t* (15 values/time point). Outliers were identified as values outside of the 1.5 x IQR (identified with the boxplot.stats() function in r) and removed. In the WT data 10 of 150 poorly behaved values were removed. 9 values were removed from the *prp2-1* data.

4. The scaling factor *F* was computed for each time point *t* by computing the geometric mean of *B* for the 5 spike-ins and 3 replicates:

$$F_t = \mu_{geometric}(B_t)$$

5. To normalize read counts *C* for each intron target *n* in each replicate *j* in each time point *t*, read counts were multiplied by the scaling factor *F*:

$$Normalized\ read\ counts = F_t \bullet C_{ntj}$$

Background correction

To estimate and correct for purification of unlabeled RNA during the 4-tu purification, samples were taken before the addition of 4-tu to the culture media. The mean normalized counts for total unspliced, lariat intermediate, and pre 1st step isoforms for each target were calculated across the time 0 samples and were subtracted from the corresponding counts in each other time point.

Model Fitting

The abundance of labeled RNA increases with time primarily due to transcription however, the cells division rate also increases labeled RNA abundance. However, the

replication rate of WT and *prp2-1* cells was roughly 100 times slower than the estimated splicing rates. This would be a constant correction across all splicing rates and would only minimally change the rate estimates. For this reason, we did not adjust for growth rate in our model.

Total splicing rate model

To model the total rate of splicing, a first order model was fit to background corrected normalized total unspliced counts to estimate the synthesis rate k_{synth} and total splicing rate $k_{splicing}$:

$$\frac{d[unspliced]}{dt} = k_{synth} - k_{splicing}[unspliced]$$

$$[unspliced]_t = \frac{k_{synth}}{k_{splicing}}\left(1 - e^{-k_{splicing}(t - t_{off})}\right)$$

The model was fit to the composite data across the three replicates using non-linear least squares with weights equal to $1/t^2$. This was done to place more weight on earlier time points that generally contain a large portion of the approach to equilibrium curve. The models were fit in r using the nlsLM() function in the minpack.lm package. Several samples behaved far from expected consistently across all targets and were removed from the total splicing rate and further models. 4 of 30 samples were removed from WT and 3 of 30 samples were removed from *prp2-1*.

Time offset

When 4-tu is added to a culture of cells, it must be transported into the cell and processed before it becomes available to the transcription machinery in the nucleus. This lag time (t_{off}) was estimated by fitting the above described model to pre-1st step counts and allowing it to estimate t_{off} in addition to the other parameters for every

intron. To estimate total splicing rates, the modeling procedure was repeated with total unspliced counts using the median t_{off} as a fixed parameter. t_{off} was 105s and 182s for WT and *prp2-1*, respectively.

Individual step coupled splicing rate model

The abundance of pre 1st step counts $[pre\ 1st\ step]$ over time is influenced by k_{synth} and decay via the k_{1st} . The abundance of pre 1st step counts over time was modeled using a simple 1st order model as follows:

$$\frac{d[pre\ 1st\ step]}{dt} = k_{synth} - k_{1st}[pre\ 1st\ step]$$

$$[pre\ 1st\ step]_t = \frac{k_{synth}}{k_{1st}}\left(1 - e^{-k_{1st}(t-t_{off})}\right)$$

The abundance of lariat counts over time are influenced by k_{synth}, k_{1st}, and k_{2nd}. The abundance of lariat intermediate counts over time was modeled using a 2 consecutive 1st order reactions model described below:

$$\frac{d[lariat]}{dt} = k_{synth} - k_{1st}[pre\ 1st\ step] - k_{2nd}[lariat]$$

$$[lariat]_t = \frac{k_{synth}}{k_{2nd}(k_{2nd} - k_{1st})}\left(k_{2nd}\left(1 - e^{-k_{1st}(t-t_{off})}\right) - k_{1st}\left(1 - e^{-k_{2nd}(t-t_{off})}\right)\right)$$

We assume that the same k_{synth}, k_{1st} influence pre 1st step and lariat intermediate counts for a given intron. As such, we estimated k_{synth}, k_{1st}, and k_{2nd} by simultaneously fitting both pre 1st step and lariat intermediate models with k_{synth} and k_{1st} as shared parameters. The models were fit to the composite data across all 3 replicates. As before, NLS was used with weights as described above. 90% confidence intervals for parameter estimates were calculated using the confint2() function in r and reported in the supplemental table **(Tables S2,5)**

As described, k_{synth} was estimated in both the total splicing rate and individual step coupled models. Both synthesis rate estimates correlated very well and k_{synth} from the total splicing rate model was used in synthesis rate analysis throughout this paper.

Splice site scores

For S. cerevisiae splice site scores (5'SS, BP, 3'SS region, 3'SS, and U-tract) position weight matrices were generated. Mononucleotide probabilities across all introns were used to generate scores and background nucleotide probabilities were calculated from the combination of all intron sequences (Table S3). Composite splice site scores were calculated by adding 5'SS, BP, and 3'SS region scores for each intron (Table S3). S. pombe 5'SS and 3'SS scores were calculated using the Burge lab's MaxEntScan tool's maximum entropy model(Yeo and Burge, 2004). H. sapiens and D. melanogaster 5'SS and 3'SS scores were obtained from Drexler H., et al(Drexler et al., 2020). Sequence logos were generated from mononucleotide position probability matrices using the ggseqlogo package in R (Figures 2 and S3A and S4B)(Wagih, 2017)

Intron and transcript annotations

S. cerevisiae Intron, open reading frame, and function (RPG or nRPG) annotations were extracted from the .gff feature file associated with the current genome release (Engel et al., 2014)(R64-2-1 downloaded from *Saccharomyces* Genome Database)(Engel et al., 2014). TSS and PAS annotations were extrapolated from published TIF-Seq median 5'UTR and 3'UTR lengths(Pelechano et al., 2013). S. pombe intron annotations and splice site sequences were extracted from the .gff3 file supplied by pombase and the current genome release(Lock et al., 2018, Wood et al., 2002)

Quantification and Statistical Analysis

Statistical analysis, quantification approaches, and RNA processing rate modeling associated with sequencing data are presented in the Method Details section. Sequencing data processing and transcript/splice isoform quantification were performed using the software described above along with custom python scripts that can be found in the

Supplemental Tables

Table S1. Key statistics of 4su labeled RNA purification and MPE-seq libraries sequenced in this study, related to STAR methods. Not Included.

Table S2. WT splicing half-life estimates related to Figures 2-6. Splicing half-life estimates from the total splicing rate and individual step coupled splicing rate models. Upper and lower 90% confidence intervals are included. Not Included.

Table S3. S. cerevisiae cis transcript features related to Figures 2-6. Not Included.

Table S4. S. pombe 5'SS and 3'SS scores related to Figure 3. Not Included.

Table S5. prp2-1 splicing half-life estimates related to Figures 5-6. Splicing half-life estimates from the total splicing rate and individual step coupled splicing rate models. Upper and lower 90% confidence intervals are included. Not Included.

Table S6 Oligonucleotide sequences used in this study, related to STAR methods. Includes oligonucleotide primer sequences for generation of IVT spike-ins and all RT primers used to prepare MPE-seq libraries. Not Included

Supplemental Figures

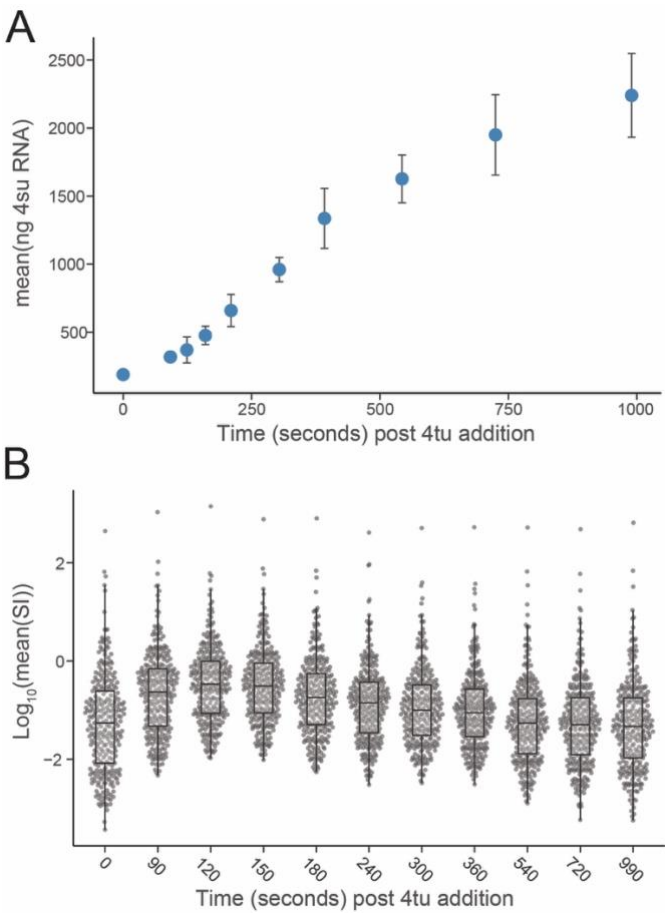


Figure S1: 4su labeling purifies nascent RNA. Related to Figure 1. (**A**) Mass of 4su labeled RNA purified versus time after addition of 4tu to the media. RNA mass is represented as the mean across 3 replicates for each time point. Error bars are 1 standard deviation. (**B**) Mean splice index (SI) versus time after addition of 4tu to the media. SI is defined as total unspliced reads divided by total spliced reads for each intron. SI is represented as the mean across 3 replicates for each time point.

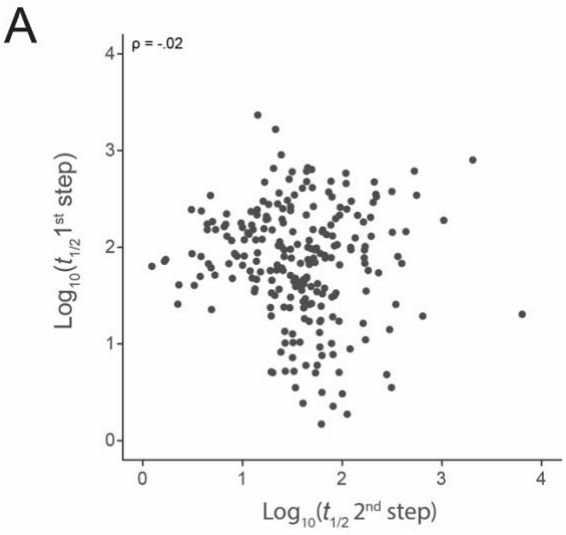


Figure S2. 1st and 2nd step rates are not correlated. Related to Figure 2. (A) Log10 transformed 1st step half-lives versus Log10 transformed 2nd step half-lives. Spearman correlation coefficient is included (ρ)

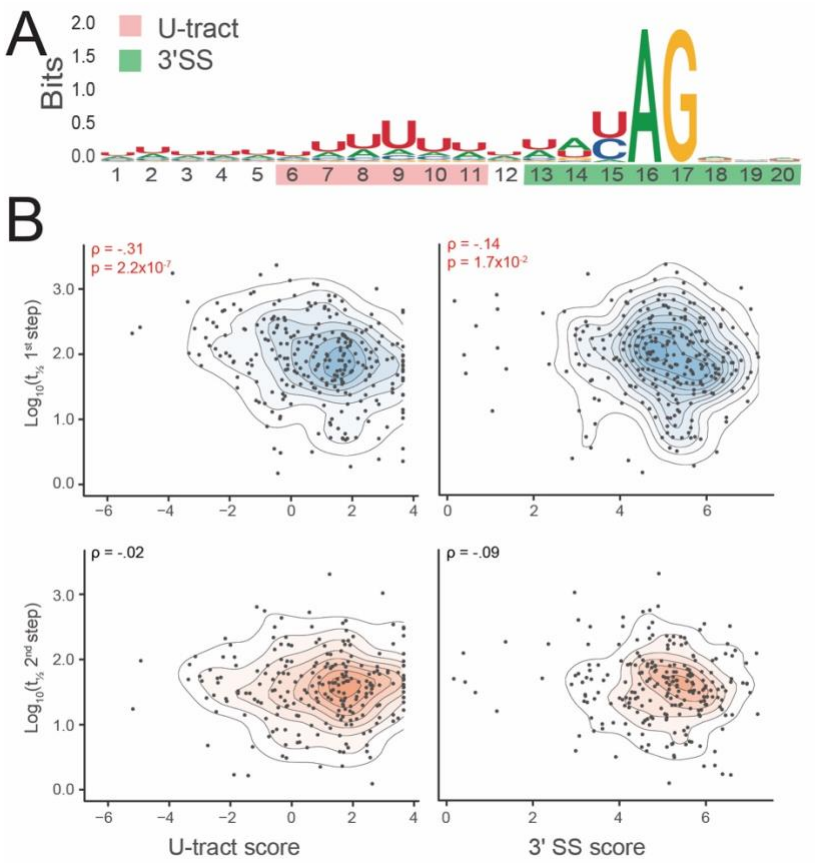


Figure S3. U-tract strength correlates with 1st but not 2nd step rates. Related to Figure 3. (**A**) Sequence logo generated from all analyzed introns for a region around the 3'SS. U-tract and 3'SS portions used for splice site score calculations are highlighted. (**B**) Comparison of U-tract and 3'SS scores with 1st, and 2nd step rates (see also Table S3). Spearman correlation coefficients (ρ) and associated p-values (p) are included. Contour lines were drawn from 2d kernel density estimation implemented by the stat_density_2d() function in the R package. Color fill gradient corresponds to density level. Red text and highlighted axes indicate significant correlations ($p < 0.05$).

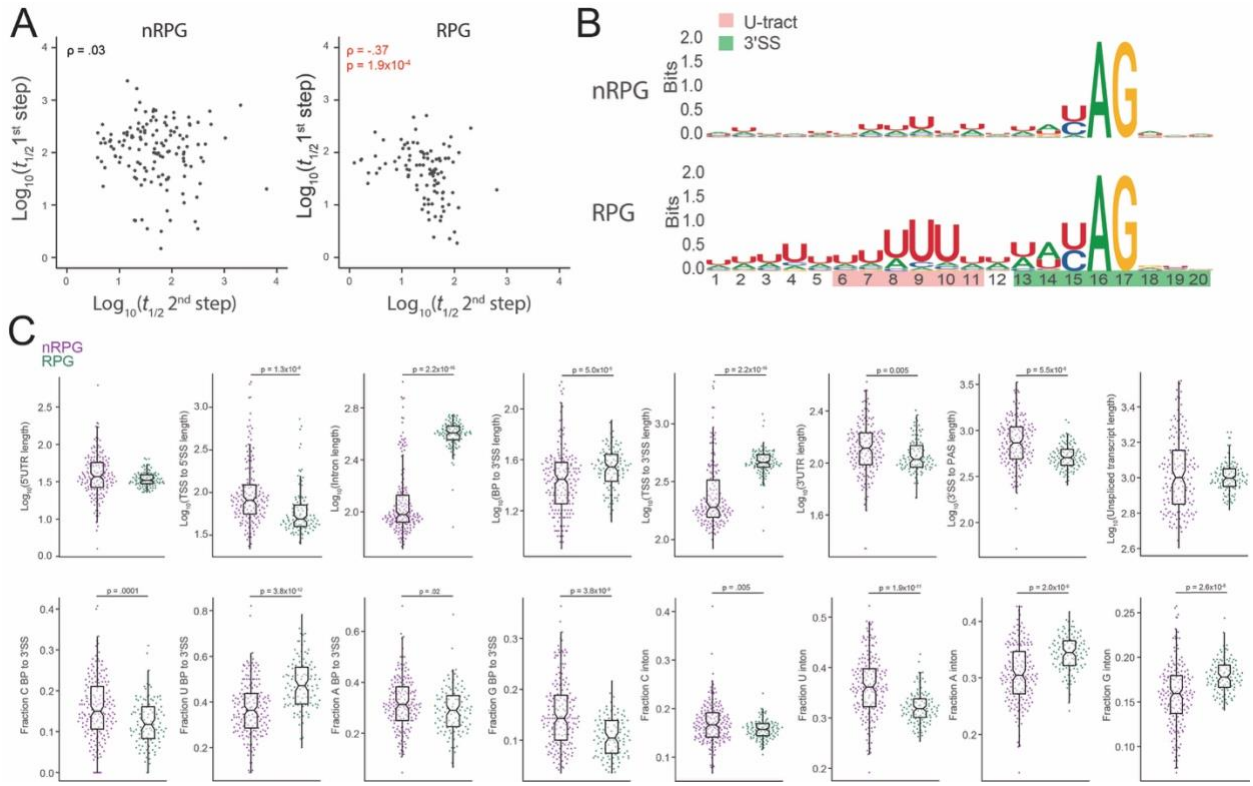


Figure S4. 1st and 2nd step rates are not correlated. Related to Figure 2. (**A**) Log_{10} transformed 1st step half-lives versus Log_{10} transformed 2nd step half-lives for RPGs and nRPGs. Spearman correlation coefficients (ρ) and associated p-values (p) are included. Red text and highlighted axes indicate significant correlations ($p < 0.05$). (**B**) Sequence logos generated from RPG or nRPG introns for a region around the 3'SS. U-tract and 3'SS portions used for splice site score calculations are highlighted. (**C**) Boxplots comparing transcript features between nRPG and RPGs. Statistical significance was calculated with one-sided Mann-Whitney signed-rank test and associated p-values (p) were included were $p < 0.05$.

Works Cited

ABOVICH, N., LIAO, X. C. & ROSBASH, M. 1994. The yeast MUD2 protein: an interaction with PRP11 defines a bridge between commitment complexes and U2 snRNP addition. *Genes Dev,* 8**,** 843-54.

ANDERS, S., PYL, P. T. & HUBER, W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England),* 31**,** 166-169.

ANDREWS, S. (n.d). FASTQC. *https://www.bioinformatics.babraham.ac.uk/projects/fastqc/*.

BARRASS, J. D., REID, J. E. A., HUANG, Y., HECTOR, R. D., SANGUINETTI, G., BEGGS, J. D. & GRANNEMAN, S. 2015. Transcriptome-wide RNA processing kinetics revealed using extremely short 4tU labeling. *Genome Biology,* 16**,** 282.

BENTLEY, D. L. 2014. Coupling mRNA processing with transcription in time and space. *Nature reviews. Genetics,* 15**,** 163-175.

BOEHNING, M., DUGAST-DARZACQ, C., RANKOVIC, M., HANSEN, A. S., YU, T., MARIE-NELLY, H., MCSWIGGEN, D. T., KOKIC, G., DAILEY, G. M., CRAMER, P., DARZACQ, X. & ZWECKSTETTER, M. 2018. RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nature Structural & Molecular Biology,* 25**,** 833-840.

CAUSTON, H. C., REN, B., KOH, S. S., HARBISON, C. T., KANIN, E., JENNINGS, E. G., LEE, T. I., TRUE, H. L., LANDER, E. S. & YOUNG, R. A. 2001. Remodeling of yeast genome expression in response to environmental changes. *Molecular biology of the cell,* 12**,** 323-337.

CHEN, S., ZHOU, Y., CHEN, Y. & GU, J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics,* 34**,** i884-i890.

DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. & GINGERAS, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* 29**,** 15-21.

DOLKEN, L., RUZSICS, Z., RADLE, B., FRIEDEL, C. C., ZIMMER, R., MAGES, J., HOFFMANN, R., DICKINSON, P., FORSTER, T., GHAZAL, P. & KOSZINOWSKI, U. H. 2008. High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *Rna,* 14**,** 1959-72.

DREXLER, H. L., CHOQUET, K. & CHURCHMAN, L. S. 2020. Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. *Molecular Cell,* 77**,** 985-998.e8.

DUFFY, E. E., SCHOFIELD, J. A. & SIMON, M. D. 2019. Gaining insight into transcriptome-wide RNA population dynamics through the chemistry of 4-thiouridine. *WIREs RNA,* 10**,** e1513.

ENGEL, S. R., DIETRICH, F. S., FISK, D. G., BINKLEY, G., BALAKRISHNAN, R., COSTANZO, M. C., DWIGHT, S. S., HITZ, B. C., KARRA, K., NASH, R. S., WENG, S., WONG, E. D., LLOYD, P., SKRZYPEK, M. S., MIYASATO, S. R., SIMISON, M. & CHERRY, J. M. 2014. The Reference Genome Sequence of Saccharomyces cerevisiae: Then and Now. *G3: Genes|Genomes|Genetics,* 4**,** 389-398.

ESER, P., WACHUTKA, L., MAIER, K. C., DEMEL, C., BORONI, M., IYER, S., CRAMER, P. & GAGNEUR, J. 2016. Determinants of RNA metabolism in the Schizosaccharomyces pombe genome. *Molecular systems biology,* 12**,** 857-857.

FICA, S. M. & NAGAI, K. 2017. Cryo-electron microscopy snapshots of the spliceosome: structural insights into a dynamic ribonucleoprotein machine. *Nature Structural & Molecular Biology,* 24**,** 791-799.

GASCH, A. P., SPELLMAN, P. T., KAO, C. M., CARMEL-HAREL, O., EISEN, M. B., STORZ, G., BOTSTEIN, D. & BROWN, P. O. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell,* 11**,** 4241-4257.

GILDEA, M. A., DWYER, Z. W. & PLEISS, J. A. 2019. Multiplexed primer extension sequencing: A targeted RNA-seq method that enables high-precision quantitation of mRNA splicing isoforms and rare pre-mRNA splicing intermediates. *Methods*.

HARLEN, K. M. & CHURCHMAN, L. S. 2017. The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nature Reviews Molecular Cell Biology,* 18**,** 263-273.

HERZEL, L., OTTOZ, D. S. M., ALPERT, T. & NEUGEBAUER, K. M. 2017. Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. *Nature Reviews Molecular Cell Biology,* 18**,** 637-650.

HICKS, M. J., LAM, B. J. & HERTEL, K. J. 2005. Analyzing mechanisms of alternative pre-mRNA splicing using in vitro splicing assays. *Methods,* 37**,** 306-313.

HNISZ, D., SHRINIVAS, K., YOUNG, R. A., CHAKRABORTY, A. K. & SHARP, P. A. 2017. A Phase Separation Model for Transcriptional Control. *Cell,* 169**,** 13-23.

KUPFER, D. M., DRABENSTOT, S. D., BUCHANAN, K. L., LAI, H., ZHU, H., DYER, D. W., ROE, B. A. & MURPHY, J. W. 2004. Introns and splicing elements of five diverse fungi. *Eukaryotic cell,* 3**,** 1088-1100.

KWAK, H., FUDA, N. J., CORE, L. J. & LIS, J. T. 2013. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science,* 339**,** 950-3.

LACADIE, S. A., TARDIFF, D. F., KADENER, S. & ROSBASH, M. 2006. In vivo commitment to yeast cotranscriptional splicing is sensitive to transcription elongation mutants. *Genes & development,* 20**,** 2055-2066.

LEE, Y. & RIO, D. C. 2015. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu Rev Biochem,* 84**,** 291-323.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & SUBGROUP, G. P. D. P. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-2079.

LOCK, A., RUTHERFORD, K., HARRIS, M. A., HAYLES, J., OLIVER, S. G., BÄHLER, J. & WOOD, V. 2018. PomBase 2018: user-driven reimplementation of the fission yeast database provides rapid and intuitive access to diverse, interconnected information. *Nucleic Acids Research,* 47**,** D821-D827.

LOPEZ, P. J. & SÉRAPHIN, B. 1999. Genomic-scale quantitative analysis of yeast pre-mRNA splicing: implications for splice-site recognition. *Rna,* 5**,** 1135-7.

LU, F., PORTZ, B. & GILMOUR, D. S. 2019. The C-Terminal Domain of RNA Polymerase II Is a Multivalent Targeting Sequence that Supports Drosophila Development with Only Consensus Heptads. *Molecular Cell,* 73**,** 1232-1242.e4.

MASON, P. B. & STRUHL, K. 2005. Distinction and Relationship between Elongation Rate and Processivity of RNA Polymerase II In Vivo. *Molecular Cell,* 17**,** 831-840.
MAYERLE, M. & GUTHRIE, C. 2017. Genetics and biochemistry remain essential in the structural era of the spliceosome. *Methods,* 125**,** 3-9.
MONTES, M., SANFORD, B. L., COMISKEY, D. F. & CHANDLER, D. S. 2019. RNA Splicing and Disease: Animal Models to Therapies. *Trends in Genetics,* 35**,** 68-87.
MOORE, M. J., SCHWARTZFARB, E. M., SILVER, P. A. & YU, M. C. 2006. Differential recruitment of the splicing machinery during transcription predicts genome-wide patterns of mRNA splicing. *Mol Cell,* 24**,** 903-15.
NAFTELBERG, S., SCHOR, I. E., AST, G. & KORNBLIHTT, A. R. 2015. Regulation of Alternative Splicing Through Coupling with Transcription and Chromatin Structure. *Annual Review of Biochemistry,* 84**,** 165-198.
OESTERREICH, F. C., BIEBERSTEIN, N. & NEUGEBAUER, K. M. 2011. Pause locally, splice globally. *Trends in Cell Biology,* 21**,** 328-335.
OESTERREICH, F. C., HERZEL, L., STRAUBE, K., HUJER, K., HOWARD, J. & NEUGEBAUER, K. M. 2016. Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell,* 165**,** 372-381.
PAI, A. A., HENRIQUES, T., MCCUE, K., BURKHOLDER, A., ADELMAN, K. & BURGE, C. B. 2017. The kinetics of pre-mRNA splicing in the Drosophila genome and the influence of gene architecture. *eLife,* 6**,** e32537.
PARENTEAU, J., DURAND, M., MORIN, G., GAGNON, J., LUCIER, J.-F., WELLINGER, RAYMUND J., CHABOT, B. & ABOU ELELA, S. 2011. Introns within Ribosomal Protein Genes Regulate the Production and Function of Yeast Ribosomes. *Cell,* 147**,** 320-331.
PARENTEAU, J., MAIGNON, L., BERTHOUMIEUX, M., CATALA, M., GAGNON, V. & ABOU ELELA, S. 2019. Introns are mediators of cell response to starvation. *Nature,* 565**,** 612-617.
PARKER, R. O. Y. & PATTERSON, B. 1987. 9 - Architecture of Fungal Introns: Implications for Spliceosome Assembly. *In:* INOUYE, M. & DUDOCK, B. S. (eds.) *Molecular Biology of RNA.* Academic Press.
PELECHANO, V., WEI, W. & STEINMETZ, L. M. 2013. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature,* 497**,** 127-131.
PLEISS, J. A., WHITWORTH, G. B., BERGKESSEL, M. & GUTHRIE, C. 2007a. Rapid, transcript-specific changes in splicing in response to environmental stress. *Molecular cell,* 27**,** 928-937.
PLEISS, J. A., WHITWORTH, G. B., BERGKESSEL, M. & GUTHRIE, C. 2007b. Transcript specificity in yeast pre-mRNA splicing revealed by mutations in core spliceosomal components. *PLoS Biol,* 5**,** e90.
RABANI, M., RAYCHOWDHURY, R., JOVANOVIC, M., ROONEY, M., STUMPO, D. J., PAULI, A., HACOHEN, N., SCHIER, A. F., BLACKSHEAR, P. J., FRIEDMAN, N., AMIT, I. & REGEV, A. 2014. High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell,* 159**,** 1698-1710.
REIMER, K. A., MIMOSO, C., ADELMAN, K. & NEUGEBAUER, K. M. 2020. Rapid and Efficient Co-Transcriptional Splicing Enhances Mammalian Gene Expression. *bioRxiv***,** 2020.02.11.944595.

REJA, R., VINAYACHANDRAN, V., GHOSH, S. & PUGH, B. F. 2015. Molecular mechanisms of ribosomal protein gene coregulation. *Genes & Development,* 29**,** 1942-1954.
SCOTTI, M. M. & SWANSON, M. S. 2016. RNA mis-splicing in disease. *Nature Reviews Genetics,* 17**,** 19-32.
SEMLOW, D. R. & STALEY, J. P. 2012. Staying on message: ensuring fidelity in pre-mRNA splicing. *Trends in biochemical sciences,* 37**,** 263-273.
SHI, Y. 2017. Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nature Reviews Molecular Cell Biology,* 18**,** 655-670.
SPINGOLA, M., GRATE, L., HAUSSLER, D. & ARES, M., JR. 1999. Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. *Rna,* 5**,** 221-34.
STALEY, J. P. & GUTHRIE, C. 1999. An RNA switch at the 5' splice site requires ATP and the DEAD box protein Prp28p. *Mol Cell,* 3**,** 55-64.
TARDIFF, D. F., LACADIE, S. A. & ROSBASH, M. 2006. A Genome-Wide Analysis Indicates that Yeast Pre-mRNA Splicing Is Predominantly Posttranscriptional. *Molecular Cell,* 24**,** 917-929.
VIJAYRAGHAVAN, U. & ABELSON, J. 1989. Isolation and characterization of pre-mRNA splicing mutants of Saccharomyces cerevisiae. *Genes & Development,* 3**,** 1206-1216.
WACHUTKA, L., CAIZZI, L., GAGNEUR, J. & CRAMER, P. 2019. Global donor and acceptor splicing site kinetics in human cells. *eLife,* 8**,** e45056.
WACHUTKA, L. & GAGNEUR, J. 2017. Measures of RNA metabolism rates: Toward a definition at the level of single bonds. *Transcription,* 8**,** 75-80.
WAGIH, O. 2017. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics,* 33**,** 3645-3647.
WAHL, M. C., WILL, C. L. & LÜHRMANN, R. 2009. The spliceosome: design principles of a dynamic RNP machine. *Cell,* 136**,** 701-718.
WARF, M. B. & BERGLUND, J. A. 2010. Role of RNA structure in regulating pre-mRNA splicing. *Trends in biochemical sciences,* 35**,** 169-178.
WARNER, J. R. 1999. The economics of ribosome biosynthesis in yeast. *Trends in Biochemical Sciences,* 24**,** 437-440.
WICKHAM, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
WILL, C. L. & LÜHRMANN, R. 2011. Spliceosome structure and function. *Cold Spring Harbor perspectives in biology,* 3**,** a003707.
WINDHAGER, L., BONFERT, T., BURGER, K., RUZSICS, Z., KREBS, S., KAUFMANN, S., MALTERER, G., L'HERNAULT, A., SCHILHABEL, M., SCHREIBER, S., ROSENSTIEL, P., ZIMMER, R., EICK, D., FRIEDEL, C. C. & DÖLKEN, L. 2012. Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome research,* 22**,** 2031-2042.
WOOD, V., GWILLIAM, R., RAJANDREAM, M. A., LYNE, M., LYNE, R., STEWART, A., SGOUROS, J., PEAT, N., HAYLES, J., BAKER, S., BASHAM, D., BOWMAN, S., BROOKS, K., BROWN, D., BROWN, S., CHILLINGWORTH, T., CHURCHER, C., COLLINS, M., CONNOR, R., CRONIN, A., DAVIS, P., FELTWELL, T.,

FRASER, A., GENTLES, S., GOBLE, A., HAMLIN, N., HARRIS, D., HIDALGO, J., HODGSON, G., HOLROYD, S., HORNSBY, T., HOWARTH, S., HUCKLE, E. J., HUNT, S., JAGELS, K., JAMES, K., JONES, L., JONES, M., LEATHER, S., MCDONALD, S., MCLEAN, J., MOONEY, P., MOULE, S., MUNGALL, K., MURPHY, L., NIBLETT, D., ODELL, C., OLIVER, K., O'NEIL, S., PEARSON, D., QUAIL, M. A., RABBINOWITSCH, E., RUTHERFORD, K., RUTTER, S., SAUNDERS, D., SEEGER, K., SHARP, S., SKELTON, J., SIMMONDS, M., SQUARES, R., SQUARES, S., STEVENS, K., TAYLOR, K., TAYLOR, R. G., TIVEY, A., WALSH, S., WARREN, T., WHITEHEAD, S., WOODWARD, J., VOLCKAERT, G., AERT, R., ROBBEN, J., GRYMONPREZ, B., WELTJENS, I., VANSTREELS, E., RIEGER, M., SCHÄFER, M., MÜLLER-AUER, S., GABEL, C., FUCHS, M., DÜSTERHÖFT, A., FRITZC, C., HOLZER, E., MOESTL, D., HILBERT, H., BORZYM, K., LANGER, I., BECK, A., LEHRACH, H., REINHARDT, R., POHL, T. M., EGER, P., ZIMMERMANN, W., WEDLER, H., WAMBUTT, R., PURNELLE, B., GOFFEAU, A., CADIEU, E., DRÉANO, S., GLOUX, S., et al. 2002. The genome sequence of Schizosaccharomyces pombe. *Nature,* 415**,** 871-80.

XU, H., FAIR, B. J., DWYER, Z. W., GILDEA, M. & PLEISS, J. A. 2019. Detection of splice isoforms and rare intermediates using multiplexed primer extension sequencing. *Nature Methods,* 16**,** 55-58.

YEO, G. & BURGE, C. B. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol,* 11**,** 377-94.

ZHANG, L., VIELLE, A., ESPINOSA, S. & ZHAO, R. 2019. RNAs in the spliceosome: Insight from cryoEM structures. *Wiley interdisciplinary reviews. RNA,* 10**,** e1523-e1523.

Appendix III: Identification and characterization of novel conditional splicing alleles in fission yeast

Alternative citation for this appendix:

Fair BJ*, Larson A*, Dwyer ZW, Armstrong J, Bone N, Pleiss JA. Isolation of scores of novel temperature-sensitive splicing pathway mutations in fission yeast (In preparation)
*Denotes equal contribution

Abstract

To identify conditional alleles which disrupt the splicing pathway, we screened a collection of ~2000 chemically mutagenized temperature-sensitive fission yeast isolates for intron accumulation in three introns, including a naturally occurring U2-dependent AT-AC intron. We identified 54 strains which accumulate unspliced message at the non-permissive temperature. A combination of mutation effect prediction and genetic mapping techniques revealed missense mutations in core splicing genes which function at various steps of spliceosome assembly and activation including: *prp10, prp28, sap61, ntr1, spp42, sap114, cwf22,* and *prp22*. Interrogation of structural data and genome-wide analysis of the splicing defects in these mutations revealed intron-specific defects which suggest the biochemical basis for these mutations' effects on splicing.

Introduction

Pre-mRNA splicing is catalyzed by the spliceosome, a dynamic multi-megadalton complex composed of five small nuclear ribonucleoprotein complexes (snRNPs) and hundreds of auxiliary proteins (reviewed in Will and Lührmann 2011; Matera and Wang 2014). Each snRNP (U1, U2, U4, U5 and U6) is composed of a small nuclear RNA

(snRNA) complexed with proteins. The snRNPs assemble anew on each intron in a stepwise manner dependent on conserved sequence elements of introns, namely the 5'splice site (5'ss), the 3' splice site (3'ss), and the branchpoint sequence (BPS) which typically lies 10-40 nucleotides upstream of the 3'ss (Taggart *et al.* 2012; Qin *et al.* 2016). Spliceosome assembly begins with binding of the U1 snRNP at the 5'ss, which nearly always begins with a GU dinucleotide and base-pairs with the 5' end of U1 snRNA. A rare class of introns, aptly named AT-AC introns, contain AU at the 5'ss and AC at the 3'ss. These introns are often spliced in a non-canonical pathway by an analogous set snRNPs (U11, U12, U4atac, U5, U6atac) which are collectively called the minor spliceosome (Turunen *et al.* 2013). In the canonical pathway, after U1 binding, the U2 snRNP assembles on the pre-mRNA in an ATP-dependent manner, resulting in a base-paired duplex between the BPS and the U2 snRNA wherein a conserved catalytic adenosine in the BPS is bulged. This duplex is stabilized by SF3A and SF3B, heteromeric protein subcomplexes that associate with the U2 snRNP. The SF3B complex shields the reactive bulged adenosine until catalytic activation (Rauhut *et al.* 2016). The pre-formed U4/U5/U6 tri-snRNP assembles on the intron, and a cascade of structural and base-pairing rearrangements are catalyzed by ATP-dependent RNA-helicases. U4 snRNP and U1 snRNP dissociate, and the U1:5'ss interaction is replaced by U6:5'ss base-pairing. Concomitantly, the NineTeen Complex (NTC), a complex of proteins associated with Prp19, joins the spliceosome. The RNA helicase Prp2 disassociates the SF3A/B complexes (Kim and Lin 1993; Liu and Cheng 2012), allowing for transesterification between the reactive bulged adenosine and the 5'ss, resulting in a lariat intermediate. The RNA helicases Prp16 and Prp22 facilitate a structural

rearrangement that leads to the second transesterification between the upstream exon and the 3'ss, resulting in exon-ligation and excision of the lariat (Schwer and Gross 1998; Tseng *et al.* 2011). Spliceosome disassembly is catalyzed by the helicase Prp43 and its cofactors, Ntr1 and Ntr2, resulting in separation of U2, U5, U6, NTC and the excised lariat (Tsai *et al.* 2005; Boon *et al.* 2006).

Many of the genes in the splicing pathway were initially identified by genetic screens for *S. cerevisiae* strains which harbor defects in pre-mRNA processing, thus the naming of the Prp genes (Hartwell *et al.* 1970; Roshbash *et al.* 1981; Vijayraghavan *et al.* 1989; Noble and Guthrie 1996; Hossain and Johnson 2014). Often these screens result in the isolation of conditional splicing alleles: cold-sensitive or temperature-sensitive (ts) alleles of splicing genes which are active at a permissive temperature (typically 25°C) but inactive and inviable for cell growth at the non-permissive temperature. These conditional alleles have allowed for temperature-selectable genetic screens for extragenic suppressor mutations, leading to identification of additional factors (Jamieson *et al.* 1991; Lybarger *et al.* 1999; Villa and Guthrie 2005). Furthermore, these conditional alleles serve as in-activateable splicing factors for *in vitro* splicing assays to delineate the biochemical requirements for each ordered step in the pathway (Lustig *et al.* 1986; Lin *et al.* 1987; Libri *et al.* 2001). Most recently, the cryo-EM structure of a late-stage yeast spliceosome was achieved by stalling the spliceosome immediately after to exon-ligation through utilization of a dominant negative ts-allele of Prp22 (Schwer 2008; Wilkinson *et al.* 2017).

The distantly related fission yeast *Schizosaccharomyces pombe* has also been used to identify conditional mutants via Northern blot screening for pre-mRNA

accumulation in libraries of ts-strains (Potashkin *et al.* 1989; Urushiyama *et al.* 1996). *S. pombe* is similarly genetically tractable as *S. cerevisiae*, yet retains many features of splicing that have been lost in *the S. cerevisiae* lineage (Kuhn and Käufer 2003; Fair and Pleiss 2017), including many spliceosome genes for which there exists a human homolog but not an *S. cerevisiae* homolog (Käufer and Potashkin 2000; Webb and Wise 2004). We previously screened a haploid non-essential gene-deletion library in *S. pombe* for genes which affect the splicing pathway (Larson *et al.* 2016). As most known splicing pathway genes in fission yeast are essential (Kim *et al.* 2010), we sought to gain mutational access to essential splicing pathway genes. Here we apply a similar quantitative screening methodology to a library of ~2000 ts-strains and report the identification of novel ts-alleles of known core splicing factors. After characterizing the genome-wide *in vivo* splicing phenotypes of some of these core factor mutations, we describe the relative dependencies of different subsets of introns for different splicing factors.

Results

Screening for splicing defects in canonical and non-canonical (AT-AC) introns identifies 54 ts mutant strains

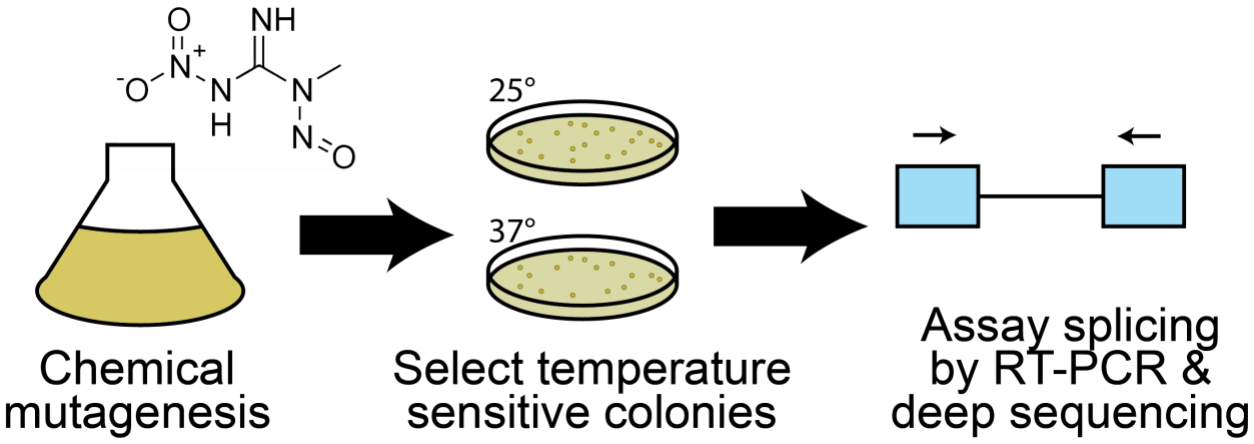


Figure 1: Workflow of a forward genetic screen for ts-alleles which display splicing defects. Yeast cells were randomly mutagenized with nitrosoguanidine and isolates were picked after replica plating at 27°C and 37°C to obtain an arrayed library of ~2000 ts-isolates. Strains were assayed for splicing defects by RT-PCR with primers that flank an intron, followed by deep sequencing the RT-PCR products to measure the relative amount of intron retention.

An arrayed library of ~2000 ts-isolates was obtained by chemical mutagenesis and replica plating, selecting for isolates which are viable at 27°C but not at 37°C (Armstrong *et al.* 2007). To identify the ts-strains harboring splicing defects, we performed RT-PCR and deep sequencing of the RT-PCR products to quantify the splicing status of naturally occurring introns in each strain after a 15-minute shift to the non-permissive temperature (Figure1, methods). To capture potential substrate-specific splicing defects, we chose three introns with varying features to quantify splicing in each strain. Firstly, we screened for strains with increased intron retention in a ribosomal protein gene, *rpl39*_intron1. This intron is of normal length for fission yeast (62bp) and contains splice sites that match the consensus splice site motifs for fission yeast. Importantly, this intron lies within the 3'UTR and thus, intron retention would not interrupt the open reading frame. We therefore expect accumulation of this intron retention isoform to specifically reflect defects in splicing, rather than defects in the nonsense mediated decay (NMD) pathway which strongly affects the abundance of most other intron-retained transcripts (Bitton *et al.* 2015). Consistent with this intron retention event being a weak target for surveillance by the NMD pathway, we observe a relatively high natural abundance, ~7%, of intron-retained transcript (Fig2A). We identified a significant increase in intron-retention in 42 strains, with intron retention rates as high as 40 percent-spliced-in (PSI). Given the relatively high sequencing depth targeted at a single locus in this screening approach, we were additionally able to detect ultra-low-frequency, unannotated splicing events that occur between the RT-PCR

primers. For example, we found an AG dinucleotide 17 bases downstream of the annotated 3'ss that gets utilized at a mere 0.001% of the annotated 3'ss, suggesting very strong constraints for the distance between the branchpoint and the 3'ss when multiple AG dinucleotides are present.

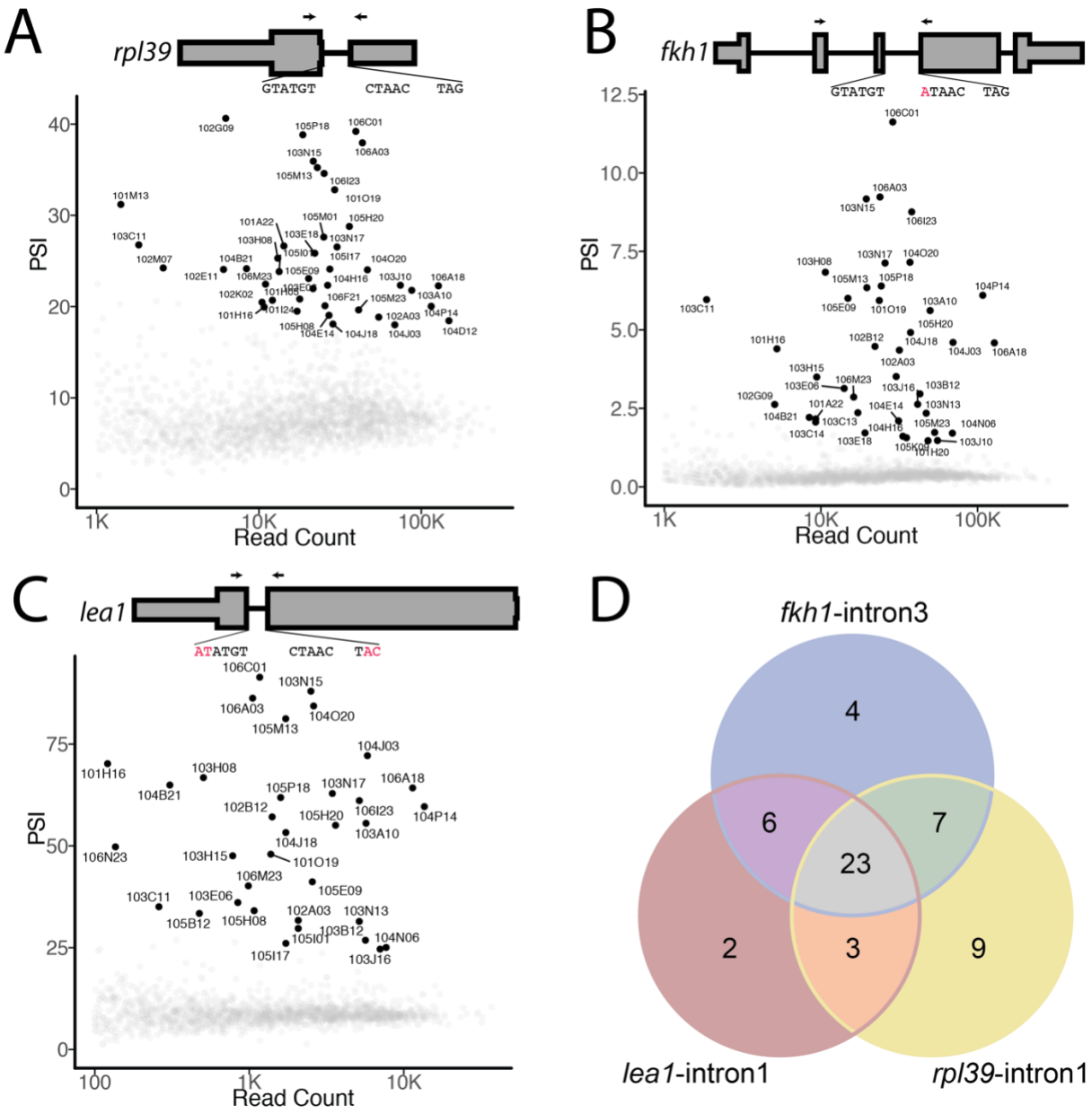
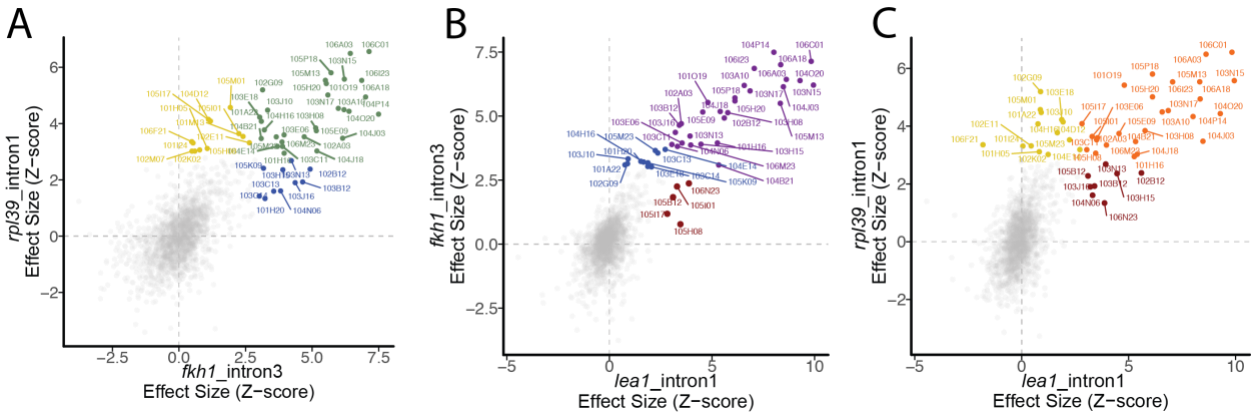


Figure2: Screening for intron retention defects in different intron substrates reveals different sets of strains. The intron retention percent spliced-in (PSI), a measure of the percent of transcripts that have the intron retained, is plotted for three different substrates for each of the ~2000 strains screened. Each strain is represented by a dot. Strains which have a significant increase in PSI, given read depth, are bolded and labelled according to an arbitrary strain identifier. The three intron substrates screened against are **(A)** *rpl39*-intron1, which has consensus motifs for the 5'ss, branchpoint motif and 3'ss; **(B)** *fkh1*-intron3, which has a non-consensus branchpoint motif; **(C)** and *lea1*-intron1 which is a major-spliceosome-dependent AT-AC intron. **(D)** The overlap among the 54 strains identified as exhibiting a significant splicing defect in the three substrates screened.

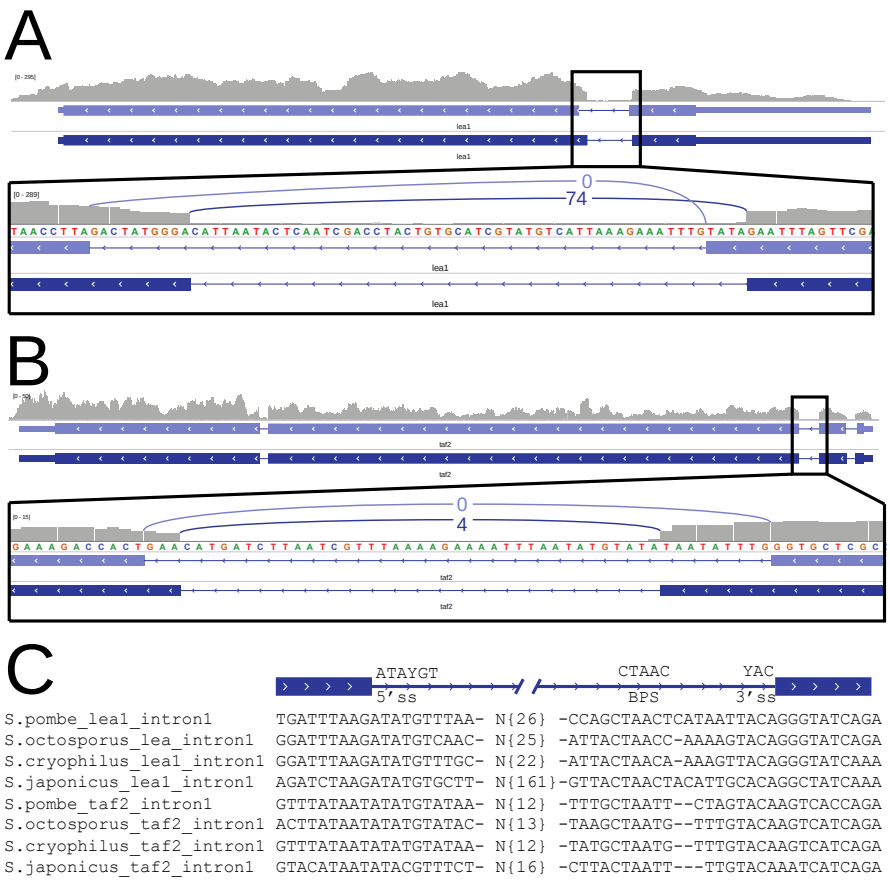
Next, we screened for strains which have increased intron retention in intron 3 of *fkh1*, a peptidylprolyl isomerase. This intron has a notably non-consensus BPS, due to the adenosine at position -3, a position reserved for pyrimidines in 99.5% of *S. pombe* introns (Fig1B). Nonetheless, this intron is efficiently spliced (PSI<1%) in most strains. Of the 40 strains identified with significant increases in intron-retained transcript, 30 were also identified in the screen against *rpl39*_intron1 (Fig2D). Changes in intron retention in rpl39_intron1 and *fkh1*_intron3 correlate well, suggesting most of the identified mutations affect splicing of many or all introns, albeit with slightly different effect sizes for different introns (FigS1).



FigureS1: The effect size of splicing defect as measured by thee different introns. The measured effect size for intron retention in each strain is plotted as a point in pairwise scatter plots for the three introns assayed. Each point is a strain. Strains with significant intron retention on either of the axes are bolded. Points are colored red for significant intron retention in *lea1*_intron1, yellow for *rpl39*_intron1, and blue for *fkh1*_intron3. The intersection of two significant sets of strains is the shown as green, purple, or orange.

Lastly, in view of the recent discovery of an AT-AC intron (with an unusual AU at the 5'ss and AC at the 3'ss) in the *lea1* gene (Chen *et al.* 2018), we screened for intron retention in this unusual splicing substrate. While considering alternative introns to

screen against, we discovered an additional AT-AC intron in the *taf2* gene (Supplemental FigS2).



FigureS2: lea1 and taf2 contain AT-AC introns in the fission yeast clade. (A) The lea1 intron is annotated (light blue gene structures, Ensembl release ASM294v2) as a canonical GT-AG intron. However, RNA-seq read alignments suggest this intron is actually spliced at nearby AT-AC splice sites (dark blue gene structure). RNA-seq coverage is shown above, with the number of spliced junction reads supporting each splice site pairing shown as arcs in the zoomed inset (B) taf2 intron2 is also annotated as a GT-AG intron, despite a number RNA-seq reads supporting an AT-AC splice site. (C) Alignments of inferred AT-AC splice amongst the Schizosachharomyces clade displays hallmarks of the AT-AC splice sites being utilized across the clade: conserved splice site motifs that conserve coding sequence, and typical CTAA[C/T] branchpoint (BPS) placed at a typical distance from the inferred AC 3'ss.

These introns, although unusual in their 5'ss and 3'ss sequence, must be substrates of the canonical U1/U2/U4/U5/U6 spliceosome, as fission yeast does not contain the U11/U12/U4atac/U5/U6atac minor spliceosome machinery known to catalyze splicing of many AT-AC introns in plants and humans (Turunen *et al.* 2013). Furthermore, the *lea1*_intron1 was found physically associated with the post-catalytic

U2-U5-U6 spliceosome as an excised lariat (Chen *et al.* 2018). In our screen, we identified only subtle differences in the mutant strains identified by screening against *lea1*_intron1; 32 of the 34 strains with significant increases in *lea1* intron retention (Fig2C, 2D) were also identified in the screens against more canonical intron substrates, consistent with this AT-AC intron being spliced by the canonical set of spliceosome factors. The two strains uniquely identified as significant for *lea1*_intron1 retention also had clear effects in *rpl39*_intron1 and *fkh1*_intron3 that just missed our significance thresholds for *lea1*_intron1 (FigS1). A more in-depth examination of differences in the relative effect sizes of different spliceosome factor mutations on this intron versus more canonical introns may reveal insights towards mechanistic distinctions in the splicing of GT-AG and AT-AC introns.

High resolution mapping of the causative mutations identifies novel alleles of core splicing factors

We identified a total of 54 strains carrying a significant increase in intron retention amongst at least one of the three introns examined. We initially picked three strains, each referred to hereafter by an arbitrary identifier, to map the causal mutation(s) for: 103A10, 103H15, and 103N15. Under the assumption that the causal mutation(s) for the ts-phenotype are also causing the observed splicing defects, we turned to bulk segregant analysis with whole-genome sequencing to map the the causal mutation(s) that segregate into ts and non-ts progeny. Each strain was outcrossed and F1 progeny from each cross were separated into bulk populations based on the progeny's ts-phenotype. In each cross, the F1 progeny appeared in roughly equal numbers of ts and non-ts, suggesting a single gene is responsible (data not shown). Whole genome

sequencing of each bulk comprehensively identified the mutations enriched in the ts-bulk and de-enriched in the non-ts-bulk. Genome-wide association analysis revealed *prp22*-G917R;V801M as a segregating mutations in strain 103A10 (Fig3A,B). The only other mutation with significant association signal is a non-coding mutation tightly linked to *prp22*. Given that the only significant coding mutations were in *prp22*, a DEAH-box RNA-helicase required for the second catalytic step of splicing (Mayas *et al.* 2006), we conclude that the splicing-defect and ts-phenotype are both caused by this single-gene mutation. We obtained a similarly simple genetic explanation for the phenotype of strain 103N15, where the significant association signal could be mapped to a single amino acid substitution in *sap61*, a component of the U2 snRNP, orthologous to S. cerevisiae Prp19 and human SF3A3. We did not find any significant association signal to explain the phenotype of strain 103H15, possibly owing to insufficient sequencing coverage to adequately genotype allele frequencies in bulk populations at the causal locus, despite ~30X genome coverage in each bulk (see methods).

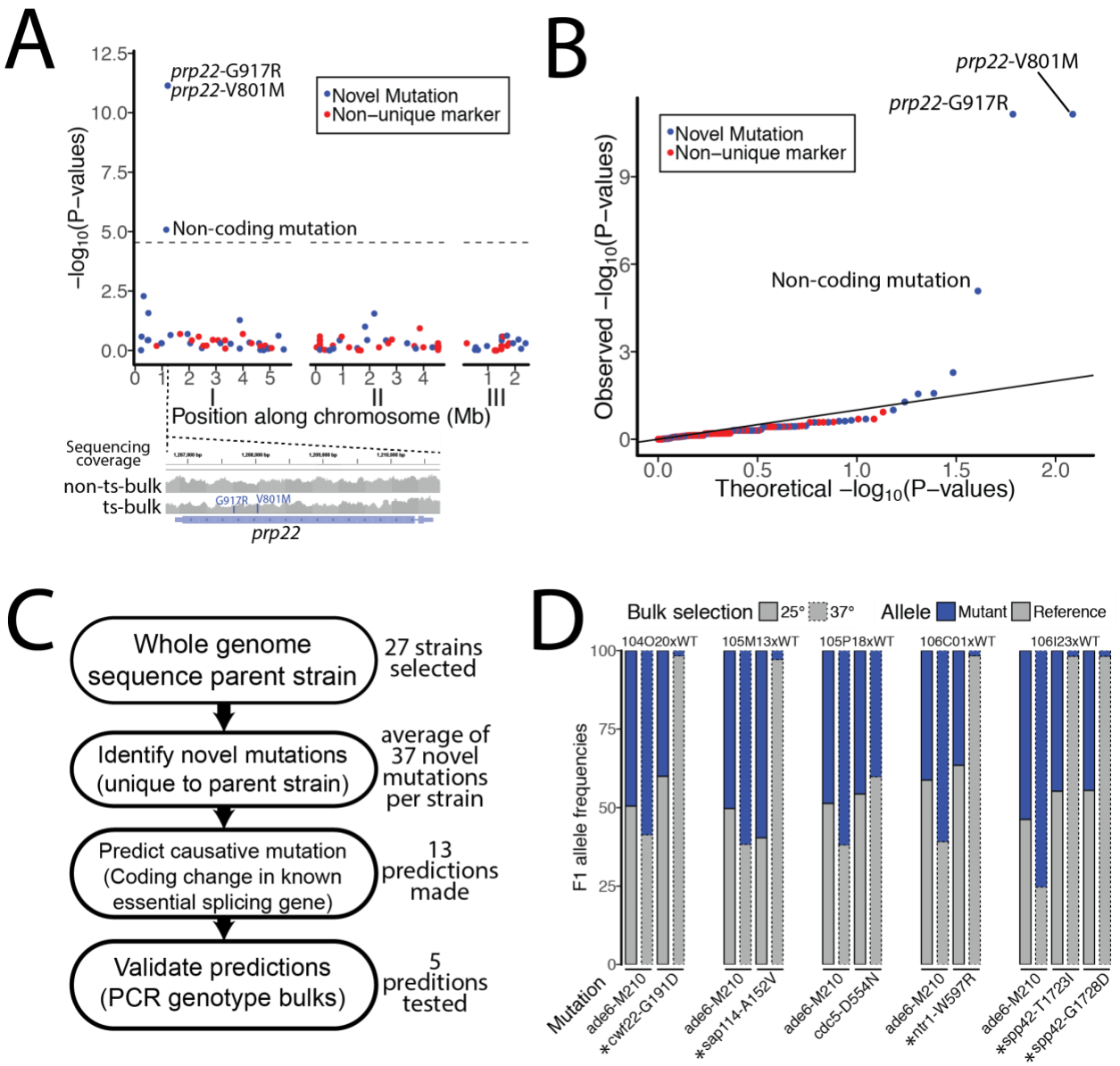
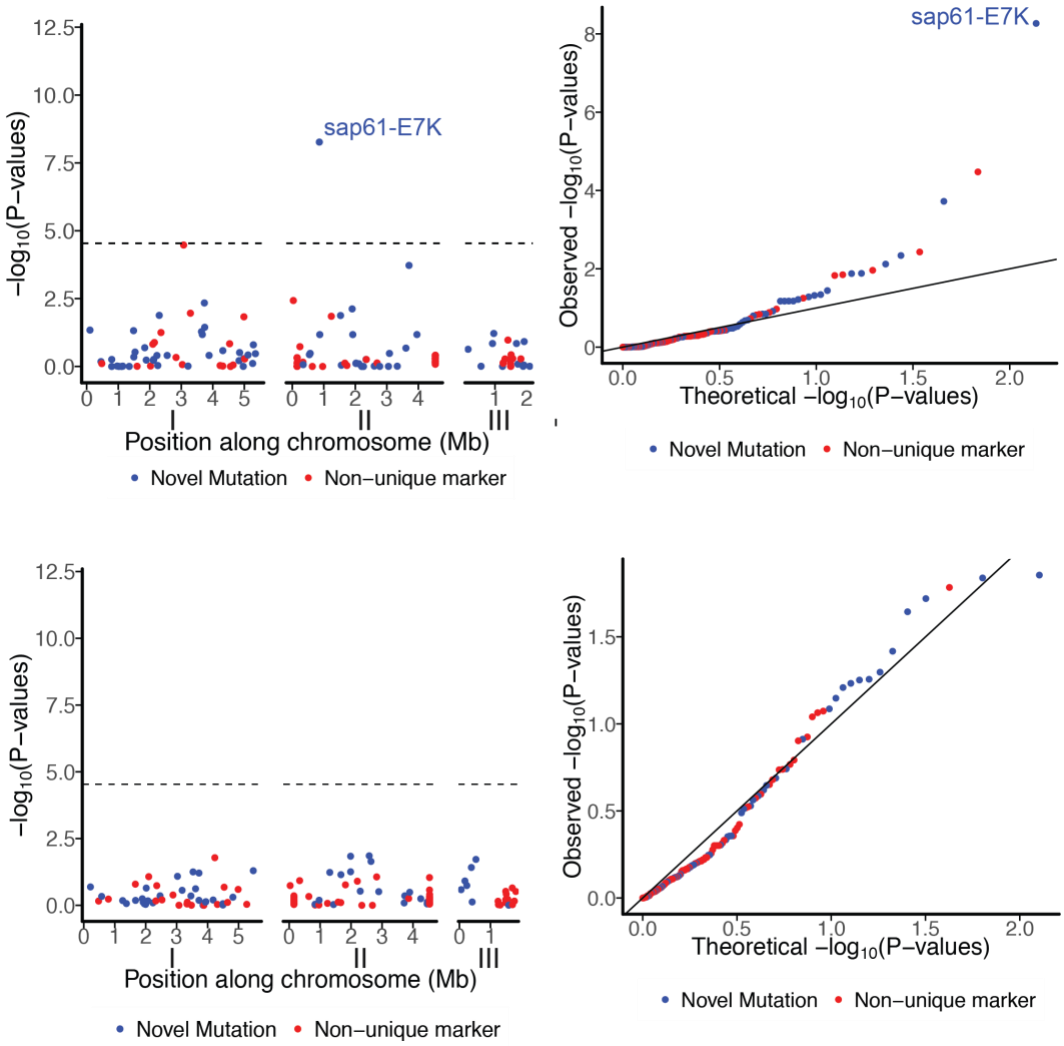


Figure 3: High resolution mapping strategies employed to identify the causative mutation of splicing-defective strains. (A) Manhattan plot of the association strength of mutations for being enriched in the of ts F1 progeny of a 103A10 x wildtype cross. Mutations which were specific to the progeny of this cross are novel mutations, while mutations that are shared across different crosses are classified as non-unique markers which were likely present in either parental strain prior to mutagenesis. Mutations that are significantly enriched in the ts-bulk and depleted from the non-ts bulk are above the Bonferonni-corrected significance threshold (dotted line). Raw read coverage with mismatches at the *prp22* locus is shown in the inset. **(B)** QQ-plot of theoretical and observed P-values of the association analysis in (A). **(C)** An alternative strategy that was used to map causative loci involves whole-genome-sequencing the strain of interest, and making predictions on which of the ~40 mutations could be causative, assuming simple the ts-phenotype and splicing defect are simple traits caused by the same mutation. **(D)** Validation of some of the predicted candidate mutations by bulk segregant analysis followed by targeted sequencing to genotype allele frequencies in each bulk. Allele frequencies in each bulk are indicated by stacked bar chart. As a control, an unrelated locus, ade6-M210, was genotyped in each bulk. Asterisks indicate mutations which are closely linked to the ts-phenotype.



FigureS3: High resolution mapping strategies employed to identify the causative mutation of splicing-defective strains. (A) Manhattan plot of associated mutations of the F1 progeny (separated into ts and non-ts bulks) of a 103N15 x wildtype cross. Mutations which were specific to this cross are novel mutations, while mutations that are shared across different crosses are inferred as non-unique markers which must have been present in either parental strain prior to mutagenesis. Mutations that are significantly enriched in the ts-bulk and depleted from the non-ts bulk are above the Bonferonni-corrected significance threshold (dotted line). **(B)** QQ-plot of theoretical and observed P-values of the association analysis in (A). **(C)** Manhattan plot of association analysis on 103H15. **(D)** QQ-plot of theoretical and observed P-values of the association analysis in (C).

To overcome the high sequencing coverage required for high-resolution mapping by whole genome sequencing large bulks, we developed an alternative strategy to mapping mutations. Given that whole genome sequencing of bulks detected only 50 to 70 novel mutations in each ts-strain (Fig3A,3B), of which only about half are

coding mutations, we surmised that the causative mutation could be predicted by whole genome sequencing the parental (F0) ts-strain and using gene annotations to make verifiable predictions for the causal mutations (Fig 3C). To demonstrate this approach, we performed whole genome sequencing on each of the strains identified as carrying a splicing defect the screen, of which, 26 of these strains had sufficient coverage (median 9X genome coverage) to comprehensively identify mutations genome-wide (TableS1). Given that causal mutations that were introduced during mutagenesis should be unique to the strain of interest, we excluded mutations common to many strains from further analysis, leaving a median of 36.5 novel mutations per strain. There was a median of 12 genes containing novel coding mutations per strain, with a median of 4.5 of these mutated genes being essential for viability. In 13 of these strains, one of the mutated essential genes was in a known splicing factor, which we predicted to be the causative mutation. We chose a subset of these strains for validation by confirming that the predicted causal mutation segregates with the ts progeny when crossed to wildtype. Of the 5 of strains selected for validation, 4 of the candidate mutations were tightly linked to the ts-phenotype, confirming our prediction (Fig3D). The strain for which the predicted mutation was not linked to the ts-phenotype (105P18) could represent an incorrect prediction for the splicing-defect mutation, or alternatively, a strain which contains a splicing-phenotype and a ts-phenotype which are unlinked. In total we identified, and confirmed via linkage, 6 novel alleles in core splicing factors: prp22-V801M;G917R, sap61-E7K; cwf22-G191D; sap114-A152V; ntr1-W597R and spp42-T1723I;G1728D. Additionally, we have strong predictions on the causative mutation for 6 strains, including one strain which harbors a predicted causative mutation in *prp10*, the ortholog

of human SF3B1 which is found commonly mutated in patients with chronic lymphocytic leukemia, myelodysplastic syndrome, and breast cancer (Network 2012; Quesada *et al.* 2012; Cazzola *et al.* 2013). The mutation identified in this study, *prp10*-R379H, maps to a region of the protein that contacts the U2 snRNA:BPS duplex formed prior to the first step of splicing (Fig4A).

TableS1: Causal mutations predicted and validated in strains subject to whole genome sequencing

Strain	Median base coverage in whole genome sequencing	Percent bases covered by at least 2 reads	Number of novel mutations identified	Genes with coding mutations	Essential genes with coding mutations	Essential, splicing genes with coding mutations	Causative mutation predicted without linkage analysis data	Mutations experimentally validated as linked to ts-phenotype
101A22	9	97.438	34	13	2	1	prp1-G705D	None tested
101I24	6	91.6416	26	9	0	0	No prediction	None tested
102A03	11	99.3911	43	6	1	0	No prediction	None tested
102B12	10	98.6822	61	23	5	2	prp10-R379H	None tested
102M07	15	99.3303	11	1	0	0	No prediction	None tested
103A10	4	97.4002	9	2	0	0	No prediction	prp22-V801M;G917R
103H08	4	95.355	48	18	5	1	cwf22-G468D	None tested
103H15	NA	NA	NA	NA	NA	NA	NA	Tested whole genome, none identified
103N15	8	97.4083	39	19	4	1	sap61-E7K	sap61-E7K
103B12	9	97.9299	57	17	5	0	No prediction	None tested
103C11	6	94.2691	60	27	6	0	No prediction	None tested
103C13	7	95.2521	16	7	2	0	No prediction	None tested
103J10	9	97.5016	27	7	5	1	brr2-G1527D	None tested
103J16	12	99.1053	11	4	1	0	No prediction	None tested
103N13	16	99.6166	48	21	7	1	prp1-G705D	None tested
104J03	9	98.3559	102	46	16	2	prp28-P383S	None tested
104J18	11	98.9073	30	8	3	0	No prediction	None tested
104O20	7	95.9661	26	10	7	1	cwf22-G191D	cwf22-G191D
104P14	11	98.7842	42	8	2	0	No prediction	None tested
104N06	6	92.5183	13	5	3	0	No prediction	None tested
105H20	9	97.6107	25	12	1	0	No prediction	None tested
105M13	8	97.2089	40	18	7	1	sap114-A152V	sap114-A152V
105P18	10	98.6747	43	16	6	1	cdc5-D554N	Tested 1 locus, none identified
106A03	11	99.1203	45	21	5	0	No prediction	None tested
106C01	12	99.2234	34	14	5	1	ntr1-W597R	ntr1-W597R
106I23	10	98.4374	28	7	4	1	spp42-T1723I;G1728D	spp42-T1723I;G1728D
106A18	19	99.7531	60	25	9	1	prp1-G705D	None tested

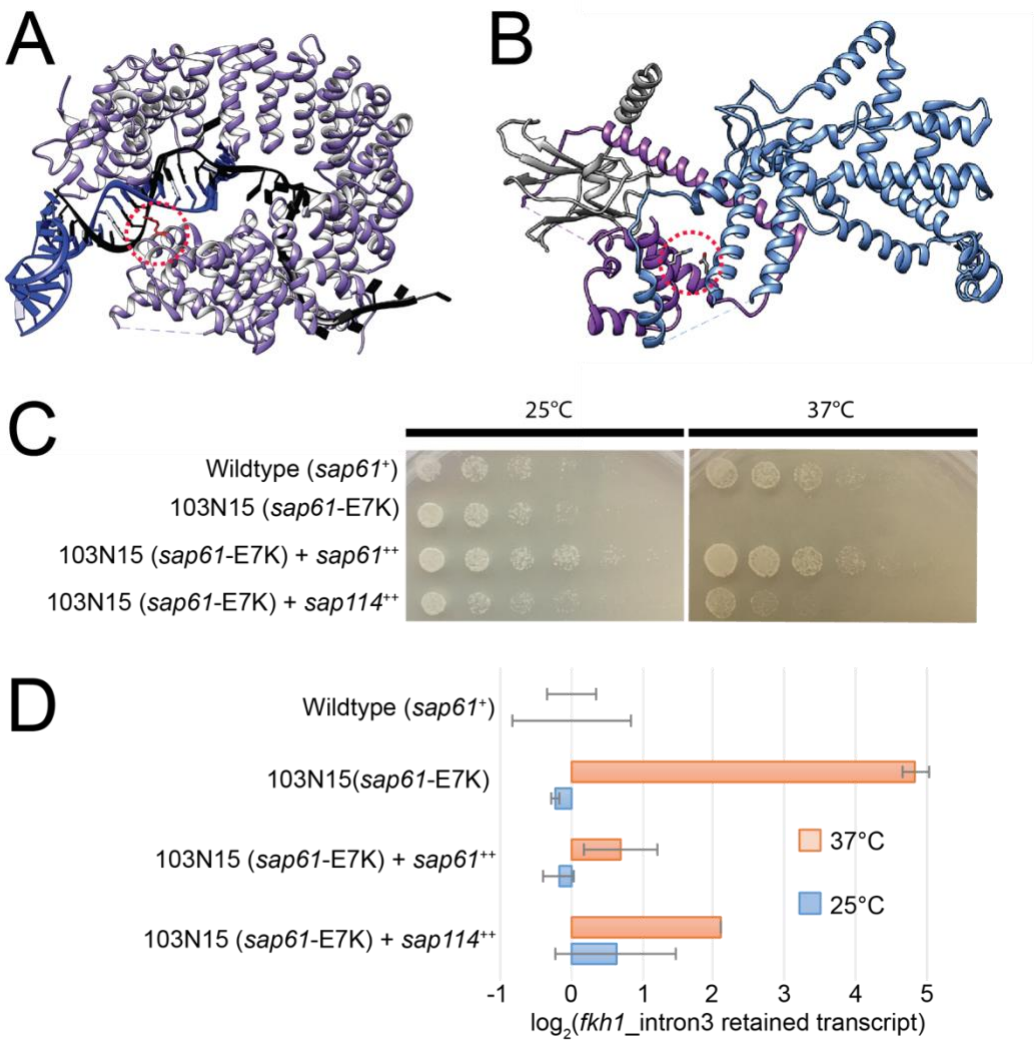


Figure 4: Structural insights to the nature of conditional mutations. (A) The published structure of S. cerevisiae HSH155 (orthologous to S. pombe *prp10*, human SF3B1) is shown covering the U2 snRNA:pre-mRNA duplex formed at the branch site in the B^{act} spliceosome complex. The residue orthologous to *prp10*-R379 is highlighted in red. U2 snRNA in blue. pre-mRNA in black. **(B)** The published structure of S. cerevisiae SF3A complex (Lin and Xu 2012), composed of Prp9 (blue), Prp21 (purple), and Prp11 (gray). The residue orthologous to S. pombe sap61-E7 (Prp9-E5) is shown making a salt bridge, highlighted in red, with Prp21 (*sap114*). **(C)** Yeast serial dilution growth assays of ts-strain 105N15 with various plasmids (indicated by + after the genotype) demonstrate that the ts-phenotype can be suppressed by exogenous over-expression (plasmid genotype marked as $^{++}$) of *sap61*, as well as *sap114* **(D)** qRT-PCR measurement of *fkh1*-intron3-retained transcript levels demonstrate that the splicing defect at the non-permissive temperature of strain 105M13 can be suppressed by expression of *sap61*, as well as sap114.

Mapping the location of the mutated residues of these ts-alleles onto recent atomic-level structures (Lin and Xu 2012; Yan *et al.* 2015) suggested the mechanistic basis for splicing defects. For example, when we mapped the *sap61*-E7K mutation we identified to the human crystal structure, we noted that this mutation disrupts a salt bridge between the *sap61* (SF3A3 in humans) and *sap114* (SF3A1 in

humans) genes of the hetero-trimeric SF3A sub-complex of the U2 snRNP (Fig4B). Armed with this structural knowledge, we hypothesized that over-expression of *sap114* could restore sufficient functional SF3A complex. Consistent with this hypothesis, exogenous overexpression of *sap114* in the sap61-E7K strain suppressed ts-phenotype (Fig4C) and partially suppressed the splicing defect as measured by qRT-PCR of fkh1_intron3 (Fig4D).

Genome-wide analysis of core splicing factor mutations reveal distinct in vivo splicing signatures

To better understand the splicing defects caused by the identified mutations, we used RNA-seq to measure genome-wide patterns in 6 of the mutant strains and two replicates of wildtype yeast after a 15-minute shift to the non-permissive temperature. All of the mutant strains increased intron retention isoforms genome-wide (Fig5A). The global levels of other forms of alternative splicing (Fig5A) constituted <1% of splicing events in all strains examined and no strains had a clear preference for any particular form of alternative splicing besides intron retention (alternative 5'ss, alternative 3'ss, etc.). The global levels of intron retention ranged from ~30% PSI in strain 106A18 (mutation in *prp1*-G705D) to ~75% PSI in 106C01 (mutation in *ntr1*-W597R). *prp1* (orthologous to *S. cerevisiae* Prp6) is a snRNP maturation factor which aids in bridging interactions between U4/U6 and U5 snRNPs to form the functional tri-snRNP. Given the relatively short (15 minute) inactivation time at the non-permissive temperature, it is perhaps unsurprising that 106C01 (*prp1*-G705D) has the weakest of the genome-wide intron-accumulation phenotype compared to the other strains which presumably directly

inhibit splicing reactions catalyzed by pre-formed tri-snRNPs. A longer inactivation period in this strain may be required to observe a stronger molecular phenotype.

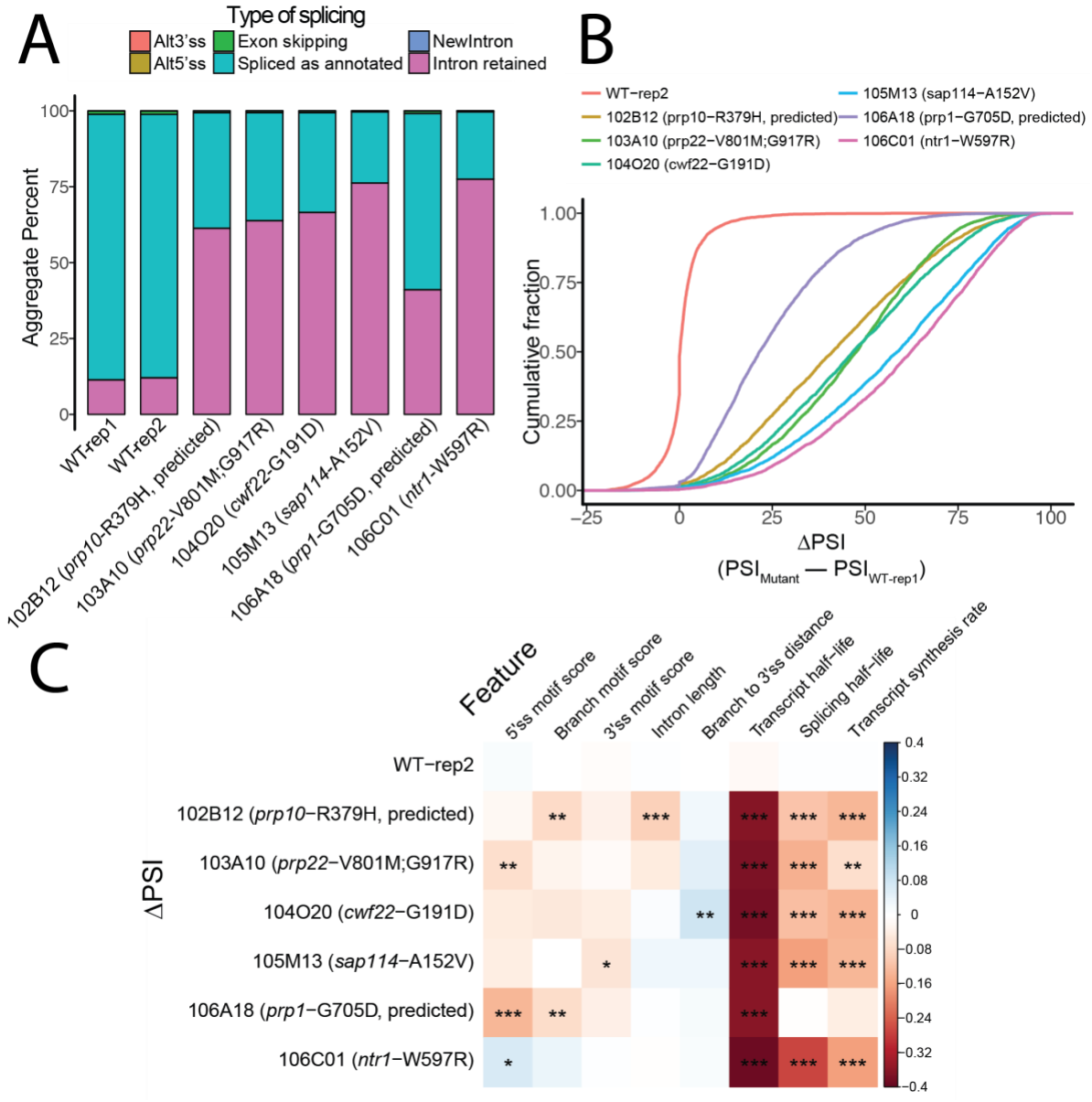
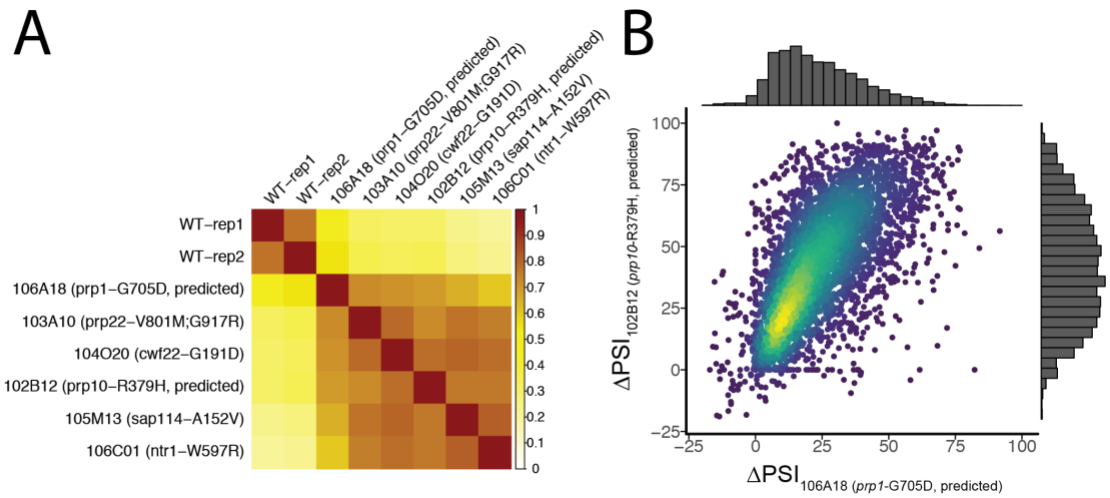


Figure 5: RNA-seq reveals distinct global splicing defects. (A) Stacked bar charts indicate the percentage of splice events of various alternative splicing categories as measured by RNA-seq after 15 minutes of cells shifted to the non-permissive temperature. Genotypes inferred as causative are labelled in parenthesis and marked as 'predicted' if tight linkage to the ts-phenotype was not demonstrated for that genotype **(B)** Empirical cumulative distribution plots of the change in intron retention ($\Delta PSI = PSI_{Mutant} - PSI_{WT\text{-}rep1}$) for annotated introns in each of the mutants as well as the other wildtype strain (WT-rep2) as a control. All strains generally increase intron retention, though the change in intron retention varies over a wide range. **(C)** Heatmap of pairwise spearman correlation coefficients between ΔPSI and features of the introns or their parent transcripts. Asterisks indicate levels of significance after Bonferroni-Holm correction: P<0.05*, 0.001**, 0.00001***.

Although global intron-retention levels increased in all 6 mutant strains, the changes in intron retention for each intron varied greatly from relatively unchanged (low ΔPSI) to greatly changed (high ΔPSI), with a median interquartile ΔPSI range of 35% amongst the 6 mutant strains (Fig5B). Interestingly, even for strains with different levels of global intron, such as 103B12 and 106A18 (FigS5), there was a high degree of correlation in the change in intron-retention as compared to wildtype (FigS5). In other words, generally, the same set of introns are most affected in each of the mutant strains, suggesting there exists common features that explain intron-to-intron variation of splicing defects in all strains. To identify these features, we tested pairwise correlations of ΔPSI versus various features of the intron, including splice site motifs, intron length, and published rates of transcription, mRNA degradation, and intron splicing rate (Eser *et al.* 2016). None of these features significantly correlated when comparing the ΔPSI of two replicates, confirming that the biological and technical noise between samples does not substantially bias this type of association analysis (Fig5C). The most significant correlating feature for all strains was mRNA degradation rate, wherein shortly lived transcripts are associated with the greatest changes in intron retention. This is consistent with the idea that the ability to measure a change in splicing after a short *in vivo* inactivation period is limited by the length of time for which the over-abundant spliced mRNA is degraded, as well as the rate at which new, initially unspliced, transcripts can be synthesized. Splicing time and transcription rate, also previously measured by a metabolic labeling time course in wildtype cells (Eser *et al.* 2016), also correlated strongly to changes ΔPSI. The *prp1*-G705D strain was unique in

that changes in splicing were not significantly correlated to splicing time or synthesis rate, consistent with this mutation affecting splicing in a time-delayed manner.



FigureS5: All ts-alleles examined display similar profile of global intron retention. **(A)** Hierarchal clustered spearman correlation matrix of intron retention (PSI) values genome-wide. All mutant strains cluster similarly **(B)** Scatterplot and marginal histograms of the changes in intron retention compared to wildtype (ΔPSI) for all introns in strain 106A18 and strain 102B12. Even though 106A18 experiences a lower degree of intron retention, the degree to which each intron experiences a change (ΔPSI) is highly correlated with 102B12.

Some other distinctive correlations that we identified also suggest specific mechanisms of splicing defects. For example, consistent with the observation that the 102B12 strain (causative mutation predicted in *prp10*) was identified in the screen for rpl39_intron1 retention (Fig1A) which carries a consensus branch motif, but not the screen for fkh1_intron3 retention (Fig1B) which has a non-consensus branch motif, the genome-wide ΔPSI measurements for this strain negatively correlate with branch motif strength (Fig5C). This perhaps is unsurprising, given the mutated residue in *prp10* contacts the U2-snRNA:BPS duplex at about the -3 to -4 position from the bulged adenosine (Fig3A). Changes in intron retention in strain 103A10 (*prp22*-V801M) negatively correlate with 5'ss strength. Studies in *S. cerevisiae* suggest that Prp22, can proofread and reject substrates with mutations in any of the consensus splice site motifs, leading to discard of lariat intermediates (Mayas *et al.* 2006; Semlow and Staley

2012). However these results (Fig5C) suggest that *prp22* proofreading is dependent more so on the relative strength of the 5'ss motif, rather than the branch or 3'ss motifs in *S. pombe*.

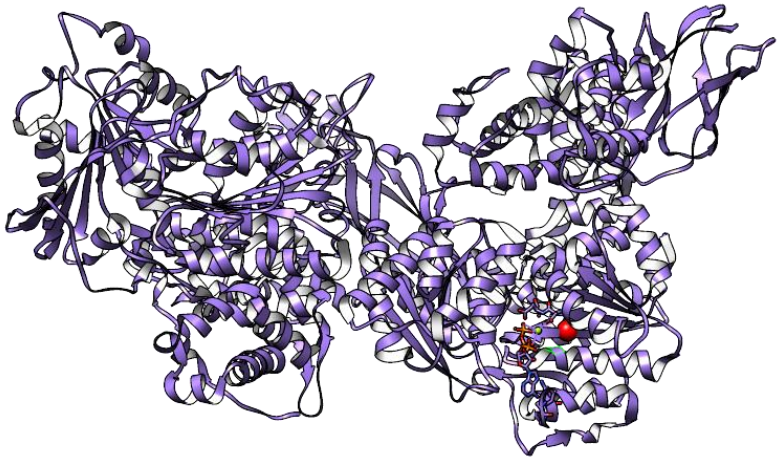
Intron retention in 104O20 (*cwf22*-G191D) is associated with longer distances between the branch and 3'ss. A relatively weaker pairwise associations was observed between 105M13 (*sap114*-A152V) intron retention and weaker 3'ss motifs. Interestingly, 106C01 (ntr1-W597) which exhibited the most intron retention globally (Fig5A,B), also exhibited a slight positive association with strength of splice site motifs. Additionally, this strain has a uniquely strong association with splicing time as measured in wildtype cells (Fig5C). Some of the intron features presented in this association analysis, namely 5'ss motif score and branch motif score, independently correlated with splicing time, or transcript synthesis and degradation rates (Eser *et al.* 2016). Therefore, the significant pairwise positive correlation between 5'ss strength and intron retention in this strain may not be a causal association, but rather a result of confounding or interacting variables. This would be most consistent with *ntr1*'s role as a spliceosome-disassembly factor acting on the excised-intron-lariat post-spliceosome complex (Tsai *et al.* 2005; Boon *et al.* 2006) wherein *ntr1* would have limited ability to discriminate splicing of substrates with varying splice site motifs. None of the strains examined here had obvious unique affects on the *lea1*_intron1 or *taf2*_intron1 (data not shown), the only two AT-AC introns *S. pombe*. However, given the effect sizes observed for other intron features, and the fact that there are only two instances of AT-AC introns, the power to detect a true association between splicing factor mutations and a tolerance for AT-AC splicing is severely limited.

Discussion

Here we screened an S. pombe library of ts-isolates for splicing defects in three introns, including an intron with canonical splice site motif, an intron with a weak BPS, and a U1/U2 dependent AT-AC intron. There was a high degree of correlation between all screens (Fig2D, FigS1), suggesting there were not any factors completely specific to splicing of the AT-AC intron. Together these screens identified 54 ts-strains with significant splicing defects, and from these we mapped with high-resolution six novel ts-alleles in splicing factors (tableS1), including mutations in *prp22*, *cwf22*, *sap61*, *sap114*, *ntr1*, and *spp42* (orthologous to budding yeast Prp22, Cwc22, Prp21, Spp382, and Prp8, respectively). Additionally, we made predictions on the causative alleles in an additional 6-ts isolates with splicing defects and demonstrated a 4/5 validation rate for our prediction method (Fig3C, Fig3D). The strains which did not harbor candidate mutations in known splicing genes may have mutations in periphery or unknown splicing-related genes. The application of more traditional mapping strategies, like gene complementation plasmid libraries, may lead to the identification of a more complete set of trans-factors involved in splicing.

Some of the alleles we mapped offer more mechanistic insights when placed in the context of published biochemical and structural data. For example, brr2 is a core helicase component of the spliceosome which contains an N-terminal and a C-terminal ATPase helicase domains. Biochemical studies have demonstrated that only the N-terminal helicase has ATPase activity. Interestingly, the brr2-G1527D allele we identified sits in the ATP binding pocket of the C-terminal helicase domain, suggesting ATP binding in the C-terminal domain is still critical for function. This is consistent with

biochemical studies which suggest that ATP binding in the C-terminal helicase may stimulate for the ATPase activity at the N-terminal domain which is required for splicing.



FigureS6: The *brr2* allele we identified likely affects ATP binding or release in the catalytically dead C-terminal helicase domain. The structure of *brr2* (Santos *et al.* 2012) with the residue orthologous to G1527, the position mutated in strain 103J10, shown as a red sphere. The N-terminal helicase domain is the left lobe of the structure, the C-terminal helicase domain, soaked with an ATP molecule shown in sticks, is the right lobe.

We also isolated an allele of *sap61*, a component of the SF3A complex, which disrupts the binding interface with complex member *sap114*. Overexpression of *sap114* suppressed the *sap61* mutant phenotype. We expect that identification of additional extragenic suppressors of these alleles will serve to clarify functional relationships between spliceosome components. The mutation identified in *spp42*, the largest spliceosome protein, maps to a functional hotspot in the highly conserved RNAseH domain of the protein. Alleles in this region, often with opposing phenotypes, have been well studied in *S. cerevisiae*. These studies suggest multiple conformations of this domain in mediating catalysis and proofreading of the 1st and 2nd steps of splicing and suggest dynamic conformational states for this region during different steps of splicing (Schellenberg *et al.* 2013; Galej *et al.* 2013, 2014; Mayerle *et al.* 2017). Our

identification of a similarly positioned allele in *S. pombe*, an organism with a larger variation in splicing substrate motifs, may allow for a more extensive *in vivo* analysis of how the RNAseH domain proofreads substrates at the 1st or 2nd step of splicing.

Genome-wide assessment of splicing in the mutant strains resolved substrate-specific defects caused by mutations in *prp10*, *prp22*, *cwf22*, *sap114*, *prp1*, and *ntr1*. Unlike the cancer-associated mutations which naturally occur most often in the SF3B1 alpha-helical HEAT domain repeats 6-8, the *prp10*-R379H allele in alpha-helical HEAT domain repeat 2 does not result any noticeable increase in alternative 3'ss selection (data not shown). Though, as with all of the mutations we examined, it remains to be seen if orthologous mutations would similarly manifest themselves in the human spliceosome which is more permissive to alternative splicing. Unsurprisingly, the *prp10*-R379H allele, which structurally maps to the -3 to -4 position of the BPS:U2snRNA duplex in activated spliceosome, causes intron retention preferentially in substrates with a weak BPS. Interestingly, this mutation also correlates with increased intron retention of long-introns. It is yet unclear if this association is explained by an *in vivo* confounding factor such, as turnover rate of affected transcripts, or by a mechanism wherein the intron length is sensed by *prp10*. We also observed an association wherein a mutation in *cwf22* (*S. cerevisiae* Cwc22), a protein loosely associated with NTC and essential for Prp2-mediated displacement of the SF3A/SF3B complexes (Yeh *et al.* 2011), causes intron retention preferentially in introns with a long distance between the BPS and 3'ss. Given the placement of *cwf22* in the activated spliceosome (Yan *et al.* 2016), a BPS-to-3'ss length-sensing mechanism for cwf22 is

plausible, though it may be mediated by additional proteins such as the non-essential N-terminal region of *prp10* (Habara *et al.* 1998).

Recently, the atomic structure of a spliceosome immediately after exon-ligation revealed the non-Watson-Crick base-pair interactions that must occur between the 5'ss G(+1) and the 3'ss G(-1) for exon ligation (Wilkinson *et al.* 2017) and suggesting a mechanism by which Prp22 may proofread and reject non-optimal substrates at this interaction site. This structure also suggested how 5'ss A(+1) and 3'ss C(-1) mutations could similarly satisfy these non-Watson-Crick interactions, reconciling the decades-old 'observation that the major spliceosome can splice AT-AC introns (Parker and Siliciano 1993; Chanfreau *et al.* 1994; Dietrich *et al.* 1997). Interestingly, we do not see unique accumulation of *lea1*_intron1 or *taf2*_intron1 retention in the *prp22*-V801 allele, consistent with *prp22* tolerating AT-AC introns within the sensitivity of our measurements. However, we observed preferential intron accumulation in introns with weak 5'ss, consistent with previous observations on the ability of *prp22* to proofread and discard substrates at the lariat intermediate stage (Mayas *et al.* 2006; Semlow and Staley 2012). If *prp22* does not reject AT-AC introns, it remains to be elucidated which steps of spliceosome assembly and/or catalysis and which factors are most responsible for rejection of potential AT-AC splice sites, as there likely exists many more potential AT-AC splice site pairs than actually get utilized by the cell. A simple explanation is that these introns fail to efficiently recruit U1 snRNP to the 5'ss. An estimation of the prevalence (or absence) of AT-AC intron lariat intermediates, relative to the prevalence of potential AT-AC introns, may clarify at which step AT-AC segments are rejected from the splicing pathway. More generally, an investigation to the degree of accumulation of

canonical (GT-AG) intron splice intermediates in optimal and non-optimal introns may be necessary to better understand the proofreading capabilities of the identified factors at different steps in splicing.

Historically, similar analyses of proofreading capabilities have often been performed by measuring the relative abundances of unspliced, intermediate, and spliced product of a labelled pre-mRNA substrate after *in vitro* reconstitution of the splicing reaction from yeast or human cell extracts (Hicks *et al.* 2005; Dunn and Rader 2014). However, fission yeast extracts largely contains late-stage spliceosomes which have proven thus far incompetent for full reconstitution of the splicing reaction *in vitro* (Huang *et al.* 2002; Dunn and Rader 2014). Use of conditional alleles to stall spliceosomes may be lead to insights as to why fission yeast extracts are incompetent for splicing. Recently published methods, including spliceosome profiling (Burke *et al.* 2018; Chen *et al.* 2018) and splice-isoform targeted sequencing (Xu *et al.* 2018), were utilized to simultaneously measure *in vivo* splice intermediates and splice products on a genome-wide scale in yeast. These methods may also yield higher-resolution insights to the proofreading capabilities of these mutants in various spliceosome stages.

Methods

Screening library for splicing defects

An arrayed library of approximately 2000 ts-isolates, originally created by the group of P. Nurse via nitrosoguanidine mutagenesis and replica plating, was obtained from J. Armstrong (Armstrong *et al.* 2007) and re-arrayed and stocked into 384-well format. We then used robotically assisted protocols to grow strains and perform RT-PCR and deep sequencing as previously described (Larson *et al.* 2016) with the

following exceptions to accommodate the growth characteristics of ts-strains: All strains were grown at the permissive temperature of 25°C for 3 days, rather than 2 days at 32°C, for initial pinning from glycerol stocks onto solid YES media. Cells were grown for 3 days at 25°C to saturation in liquid media to normalize cell density. After cultures reached saturation, cultures were back-diluted to OD_{600}~0.1 and grown for ~12 hours at 25°C (until OD_{600} ~0.8) before shifting cells to the non-permissive temperature of 37°C for 15 minutes. The duration of this temperature shift before cell collection was chosen as 15 minutes because previous work (Pleiss citation) has demonstrated that this is sufficient to observe genome-wide splicing defects in established yeast ts-alleles, while we reasoned that longer temperature shifts may elicit additional indirect effects on the transcriptome. The temperature shift was achieved by mixing 100uL of culture to 100uL of pre-warmed culture plates containing 100uL of 45°C liquid YES media. Cells were then incubated with shaking at 37°C for an additional 15 minutes prior to cell collection by centrifugation as previously described (Larson *et al.* 2016).

The RT-PCR primers and thermocycling conditions used for screening the splicing phenotype of *fkh1*-intron3, *rpl39*-intron1, and *lea1*-intron1 are listed in tableS2. We took careful care to prevent primer dimer formation during PCR reactions by employing empirically determined PCR conditions to be as follows: The initial PCR reaction containing gene-specific primers with plate-specific barcoded primers contained 10mM Tris (pH 8.3), 50mM KCl, 1.5mM $MgCl_2$ 0.2mM dNTPs, 0.25x SYBR Green I, 250nM Fwd and Rev primer, 150nM Hot-Start aptamer (Noma *et al.* 2006) (TableS1) and 1x Taq polymerase. The first PCRs for each gene-target were pooled into a single 384-well plate, preserving 384-well position, and cleaned via glass-fiber

columns (Whatman cat#7700-1101) by adding 2 volumes DNA binding buffer (5M Guanidinium HCl, 30% isopropanol, 90mM KOH, 150mM acetic acid) to the pooled samples followed by applying samples to the columns by centrifugation (2min, 2000xg) and 2 sequential wash steps with wash buffer (80% ethanol, 10mM Tris) and a dry spin. Samples were eluted in The original sample volume and diluted 5 fold prior to the next PCR which was used to append Nextera v2 indices to samples in a well-specific manner. This PCR was performed with 10mM Tris (pH 8.3), 50mM KCl, 1.5mM $MgCl_2$ 0.2mM dNTPs, 0.25x SYBR Green I, 250nM Fwd and Rev primer, and 1x Taq with cycling conditions described in tableS2. PCR reactions for each gene-target were pooled, concentrated via ethanol precipitation, cleaned via Zymo-25 glass fiber column, and resuspended in water. Libraries were sequenced to ~50M reads per gene target (~10,000 reads per individual sample per gene target) on a NextSeq lane with 75bp single-end reads.

Data processing for screen

Reads corresponding to individual samples were demultiplexed into separate fastq files via a custom script which takes into account Illumina dual index reads as well as the 8-12 base plate-barcode that is part of the insert read. Fastq files were aligned with the STAR aligner (Dobin *et al.* 2013) to a manually generated genome file containing only the targeted genes with their splice sites annotated. We had to manually re-annotate the splice sites in the *lea1* gene, as the Ensembl genome annotations (ASM294v2) erroneously list GT-AG splice sites for this gene, while we and others (Chen *et al.* 2018) empirically determined that splicing occurs at AT-AG splice sites. We used default STAR aligner parameters with the following exceptions:

```
--clip5pNbases 23 --alignEndsType EndToEnd
--clip3pAdapterSeq {ADAPTERSEQ} --genomeLoad LoadAndKeep
--limitBAMsortRAM 2000000000 --alignIntronMax 290
```

For each sample, spliced read counts were taken from the SJ.out.tab file created by the aligner, while read counts for unspliced read counts were determined using bedtools coverage (Quinlan and Hall 2010) to sum the number of reads which read into the intron. Spliced and unspliced read counts for biological replicates were combined, and strains which had a combined (unspliced + spliced) read count under 1000 (fkh1-intron3), 1000 (rpl39-intron1), or 100 (lea1-intron1) were discarded from further analysis. We determined the strains which have a significantly changed splice index (SI, equal to unspliced/spliced read count), using the statistical approach previously described to account for biases that change as a function of read depth and gene-target. Unlike our previous description of this statistical approach, in this work we performed inference using a right-sided Z-test using the interpolated Z-score as the test statistic, as we observed that the strains which were inferred as significantly increasing splicing efficiency (left-sided significant) are likely false-positives as they were generally unreproducible between the biological replicates. We attribute this to stochastic noise between biological replicates with low reads counts of the minor (unspliced) isoform which is unaccounted for in the statistical model. Multiple hypothesis correction was performed by converting P-values to Q-values (Storey and Tibshirani 2003), controlling for FDR at 0.05. For purposes of presenting the screen data in an easily interpretable manner (Figure2), we transformed the relative splice index (ratio of unspliced/spliced reads) into an estimate of percent spliced in (PSI) by the following function:

$$PSI = 100\left(\frac{SI_{relative}SI_{median}}{1 + SI_{relative}SI_{median}}\right)$$

where $SI_{relative}$ is a ratio of a particular sample's SI to an estimate of the true wildtype SI given read a particular depth ($\mu_{interpolated}$). SI_{median} is the median SI across all samples at all read depths. The $\mu_{interpolated}$ is used as a normalization factor to account for the variation in SI that is a function of read count, possibly originating from length-dependent and RNA-input-amount-dependent PCR biases. Calculation of $\mu_{interpolated}$ has been previously described previously (Larson *et al.* 2016). A count matrix of spliced and unspliced read conuts for each of the three introns assayed and test statistics are saved in tableS3.

Mapping mutations via 'bulk-segregant analysis first' approach

For 3 strains, we performed bulk segregant analysis to isolate the causative allele(s) to separate bulks of ts and non-ts phenotype, followed by genotyping the bulks by whole genome sequencing. For this approach, we first outcrossed the ts-strain to a WT strain (ED666, Bioneer) and isolated spores via glusalase treatment as described previously (Forsburg and Rhind 2006) with the exception that we found empirically that glusalase treatment required only 0.2% glusalase (Perkin Elmer cat#NEE154001EA) as the final concentration. We diluted and plated the spores to single colonies to isolate ~48 F1 isolates for each mating. Each isolate was then replica plated in quadruplicate at 25°C and 37°C to assay for ts-phenotype. Spores which were ts and non-ts were pooled into separate bulks of about *n*=20 and grown to saturation in a 2mL culture at 25°C. After cell collection and DNA isolation (Lucigen/Epicentre cat#MPY80200), we produced WGS libraries (TruSeq PCR free kit) to genotype the bulks. Bulks were sequenced to ~30X coverage using 75bp single-end reads on the Illumina NextSeq platform. Reads were aligned to the reference genome (ASM294v2) using bwa-mem aligner (Li and

Durbin 2009) with default settings. Variant calling was performed using CRISP (Bansal 2010) with an appropriate pool size (*n*) used as the --poolsize parameter for pooled genotyping. Read counts for the reference and non-reference allele were summed for each variant for each bulk to estimate allele frequency in each bulk population.

We then performed genome-wide association analysis. Given that each bulk population of n~20 individuals was only sequenced to ~30X coverage, we acknowledge that our estimates of allele frequency is limited by sequencing depth at many loci. In order to account for noise due to random sampling of both reads and spores in the bulk, we first estimated allele counts in each bulk as follows: If the read count at a variant is larger than the bulk size, the sampling error is likely driven mostly by the bulk size. Therefore, to estimate allele counts in the bulk, we coerced the allelic-ratio of read counts for each variant to the number of individuals in the bulk via simple rounding. For example, an allelic ratio in read counts of 18p:22q will correspond to an allele counts of 9p:11q if the bulk size is 20 individuals. If the read count at a variant is smaller than the bulk size, the random sampling error is likely driven by low read count so we simply used the allelic ratio of read counts as the number of allele counts in the population. Variants with a minor allele frequency less than 0.15 were discarded. Variants which were present in bulks from separate matings were categorized as markers which may be useful for mapping the causative loci but are unlikely to be the exact causative variant from a novel mutation. To look for statistically significant enrichments, we used the allele counts of each variant to create contingency tables for a one-sided Fisher exact test, looking for enrichment of the non-reference allele into the ts-bulk.

Mapping mutations via 'whole genome sequencing first' approach

Each strain of the 52 strains with significantly altered splicing was grown to saturation at 25°C in 100uL cultures and DNA was isolated as described above with proportionally scaled down volumes and whole genome sequencing libraries were generated via a tagmentation protocol similar to previously published protocols (Picelli *et al.* 2014) using ~5ng DNA as input: 5ng of DNA in Tagmentation buffer and Tn5, and indexed adapters were appended via PCR. Libraries were sequenced either as 65+10 paired end reads on a Illumina NextSeq platform or as 38+38 paired end reads on an Illumina NextSeq platform to ~10X coverage of the genome. PCR duplicates (non-unique reads with respect to every sequenced base) were removed from the dataset prior to alignment, and reads were aligned using HiSat2 (Kim *et al.* 2015). Samples with >90% genomic positions covered by >2 unique reads and >4X median coverage were retained for further analysis. Variant calling was performed using the GATK pipeline according to their 'best practices' protocols (Depristo *et al.* 2011) using the HaplotypeCaller tool with {–ploidy 1}. Variants that were shared by 4 or more strains were discarded as they likely do not represent novel mutations that are plausible candidates for the causative mutation. We choose 4 as a threshold to allow for mutations that could be present in 2 or 3 strains due to cross-contamination between neighboring wells of the strain library. We used the VariantEffectPredictor software (McLaren *et al.* 2016) to annotate the consequence of each mutation as coding or non-coding. We made predictions on the causative mutation (TableS1) for 13/27 strains if there existed a gene containing coding mutations that is a known splicing gene [belongs to the gene ontology category 'mRNA cis splicing, via spliceosome' (GO:0045292)] and is essential [associated with the PomBase gene annotation 'inviable cell population'

(FYPO:0002059)]. In the case of 102B12 and 104J03 there existed two mutations in essential splicing genes, but in both strains one of the mutations was a shared mutation in the *cwf2* gene. Reasoning that the causative gene is more likely to be a unique mutation, the causative mutation prediction listed in tableS1 for these two strains is the other splicing essential gene mutation.

Confirming predicted mutations via bulk segregant analysis

We outcrossed the ts-strain of interest to a wildtype strain (ED666, Bioneer) and isolated spores via glusalase treatment as described above. The glusalase-treated spores were divided into two equal volumes, each of which was used to inoculate a larger bulk culture. Plating assays suggest that the population size of these bulk cultures is >10000 individuals. The bulks were grown at either 25°C or 37°C in 2mL of YES media to either maintain the initial allele frequency in the bulk or to select against the causative mutation. Cells were collected after 3 days of growth or until saturated (OD_{600} ~ 12) and DNA was isolated as described above. We genotyped the candidate causal mutation in each bulk by PCR-targeted deep sequencing: PCR primers with Illumina overhangs that flank the locus of interest were designed and PCR was performed with the same reaction conditions as used for screening, with the exception that each PCR used thermocycler conditions listed in tableS2. As with the screen protocol, PCRs were diluted and used in a sequential PCR for indexing with identical reaction conditions and cleanup procedures. Reads were deep sequenced on a MiSeq platform with 150bp single end reads to a depth of ~20000 reads per sample. After alignment to the genome, allelic-ratio of read counts was determined and linkage

of the candidate mutation to the ts-phenotype was inferred if the allele is relatively depleted (0-10% allele frequency) from the 37°C bulk.

RNA-seq

25mL cultures of each strain were grown in a single replicate, with the exception of the wildtype strain (ED666, Bioneer) which was grown in two 25mL replicates. Cultures were grown at 25°C to OD_{600} ~ 0.8 and subsequently shifted to a shaking water bath incubator set at 37°C for 15 minutes before cell collection by vacuum filtration and snap frozen until later processing. RNA was isolated by hot-phenol extraction. 1ug of RNA, without DNAse treatment, was used as input into a poly-A selection protocol (NEB cat#E7490S) and a stranded RNA-seq protocol (NEB cat#E7760S), following the manufacturer's instructions. Libraries were sequenced on a NextSeq platform with 38+38 paired end reads to a depth of 20M reads per sample. RNA-seq data is deposited with accession code GSE161333.

Reads were aligned with the STAR aligner in two-pass mode with the following parameters: {--alignIntronMin 10 --alignIntronMax 1000 --sjdbFileChrStartEnd $junctions } where $junctions is a file from the aggregated first-pass spliced alignments to allow for sensitive detection of novel splice junctions. Counts of spliced reads at annotated and unannotated junctions were taken from the SJ.out.tab file generated by the aligner. Unspliced reads which infer an intron retention event were counted by using bedtools intersect to sum the counts of read alignments that overlapped >=3 bases of an annotated intron. Spliced reads were categorized as either annotated, alt3'ss (annotated 5'ss, unannotated 3'ss), alt5'ss (unannotated 5'ss, annotated 3'ss), exon-skipping (annotated 5'ss and annotated 3'ss but in novel pairing) or new intron

(unannotated 5’ss, unannotated 3’ss). Read counts of spliced and intron-retention events were normalized to the length the feature’s potential mapping space, which in the case of intron retention reads is 2(R-O)+L where R is read length, O is minimum overhang length (3) and L is length of the intron. In the case of spliced reads, the mapping space is 2(R-O). Length-normalized read counts were used for all further analysis.

PSI is calculated for each intron as the length-normalized count of intron retention reads divided by the sum of length-normalized intron retention and annotated spliced reads. ΔPSI was calculated for each sample (less WT-rep1) as the difference in PSI values between PSI_{Mutant} and $PSI_{WT\text{-}rep1}$. Motif scores were calculated for the 5’ss, branchsite, and 3’ss using a position weight matrix and log-odds scoring scheme (Lim and Burge 2001) using bases (-2 to +7) for 5’ss, (-4 to +2) for the branchsite, and (-6 to +2) for 3’ss. As branchsites for each intron are not annotated on PomBase, creating a position weight matrix first required identification branchsites. To accomplish this, we first created a position weight matrix using branch motif frequencies (PomBase) to search for the best match within a window (-40 to 0 from the 3’ss). These predicted branch locations are saved in Supplemental Table S4. From these branch locations, a new position weight matrix was created (-2 to +7). Estimates of transcription synthesis, splicing half-life, and transcript half-life were obtained from a published dataset (Eser *et al.* 2016).

Acknowledgements and Author Contributions

JA and NB donated the ts-library of strains. AL, BJF, ZD, and JP conceived and designed experiments. AL, BJF and ZD performed experiments and wrote manuscript.

Works Cited

Armstrong, J., N. Bone, J. Dodgson, and T. Beck, 2007 The role and aims of the FYSSION project. Briefings Funct. Genomics Proteomics 6: 3–7.

Bansal, V., 2010 A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics 26:.

Bitton, D. A., S. R. Atkinson, C. Rallis, G. C. Smith, D. A. Ellis *et al.*, 2015 Widespread exon-skipping triggers degradation by nuclear RNA surveillance in fission yeast. Genome Res. 884–896.

Boon, K.-L., T. Auchynnikava, G. Edwalds-Gilbert, J. D. Barrass, A. P. Droop *et al.*, 2006 Yeast Ntr1/Spp382 Mediates Prp43 Function in Postspliceosomes. Mol. Cell. Biol.

Burke, J., A. Longhurst, D. Merkurjev, J. Sales-Lee, B. Rao *et al.*, 2018 Spliceosome profiling visualizes the operations of a dynamic RNP in vivo at nucleotide resolution. Cell 173: 1014–1030.

Cazzola, M., M. Rossi, and L. Malcovati, 2013 Biologic and clinical significance of somatic mutations of SF3B1 in myeloid and lymphoid neoplasms. Blood.

Chanfreau, G., P. Legrain, B. Dujon, and A. Jacquier, 1994 Interaction between the first and last nucleotides of pre-mRNA introns is a determinant of 3′ splice site selection in S.cerevisiae. Nucleic Acids Res.

Chen, W., J. Moore, H. Ozadam, H. P. Shulha, N. Rhind *et al.*, 2018 Transcriptome-wide interrogation of the functional intronome by spliceosome profiling. Cell 173: 1031–1044.

Depristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, 2011 A

framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43: 491–501.

Dietrich, R. C., R. Incorvaia, and R. A. Padgett, 1997 Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. Mol. Cell.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski *et al.*, 2013 STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21.

Dunn, E. A., and S. D. Rader, 2014 Preparation of yeast whole cell splicing extract. Methods Mol. Biol.

Eser, P., L. Wachutka, K. C. Maier, C. Demel, M. Boroni *et al.*, 2016 Determinants of RNA metabolism in the Schizosaccharomyces pombe genome. Mol. Syst. Biol. 12: 857–857.

Fair, B. J., and J. A. Pleiss, 2017 The power of fission: yeast as a tool for understanding complex splicing. Curr. Genet.

Forsburg, S. L., and N. Rhind, 2006 Basic methods for fission yeast. Yeast 23: 173–183.

Galej, W. P., T. H. D. Nguyen, A. J. Newman, and K. Nagai, 2014 Structural studies of the spliceosome: Zooming into the heart of the machine. Curr. Opin. Struct. Biol.

Galej, W. P., C. Oubridge, A. J. Newman, and K. Nagai, 2013 Crystal structure of Prp8 reveals active site cavity of the spliceosome. Nature 493: 638–43.

Habara, Y., S. Urushiyama, T. Tani, and Y. Ohshima, 1998 The fission yeast prp10(+) gene involved in pre-mRNA splicing encodes a homologue of highly conserved splicing factor, SAP155. Nucleic Acids Res. 26: 5662–5669.

Hartwell, L. H., C. S. McLaughlin, and J. R. Warner, 1970 Identification of ten genes that

control ribosome formation in yeast. MGG Mol. Gen. Genet. 109: 42–56.

Hicks, M. J., B. J. Lam, and K. J. Hertel, 2005 Analyzing mechanisms of alternative pre-mRNA splicing using in vitro splicing assays. Methods.

Hossain, M. A., and T. L. Johnson, 2014 Using yeast genetics to study splicing mechanisms. Methods Mol. Biol.

Huang, T., J. Vilardell, and C. C. Query, 2002 Pre-spliceosome formation in S.pombe requires a stable complex of SF1-U2AF59-U2AF23. EMBO J. 21: 5516–5526.

Jamieson, D. J., B. Rahe, J. Pringle, and J. D. Beggs, 1991 A suppressor of a yeast splicing mutation (prp8-1) encodes a putative ATP-dependent RNA helicase. Nature.

Käufer, N. F., and J. Potashkin, 2000 Analysis of the splicing machinery in fission yeast: a comparison with budding yeast and mammals. Nucleic Acids Res. 28: 3003–3010.

Kim, D.-U., J. Hayles, D. Kim, V. Wood, H.-O. Park *et al.*, 2010 Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. Nat. Biotechnol. 28: 617–623.

Kim, D., B. Langmead, and S. L. Salzberg, 2015 HISAT: A fast spliced aligner with low memory requirements. Nat. Methods 12: 357–360.

Kim, S. H., and R. J. Lin, 1993 Pre-mRNA splicing within an assembled yeast spliceosome requires an RNA-dependent ATPase and ATP hydrolysis. Proc. Natl. Acad. Sci. U. S. A.

Kuhn, A. N., and N. F. Käufer, 2003 Pre-mRNA splicing in Schizosaccharomyces pombe: regulatory role of a kinase conserved from fission yeast to mammals. Curr.

Genet. 42: 241–251.

Larson, A., B. J. Fair, and J. A. Pleiss, 2016 Interconnections Between RNA-Processing Pathways Revealed by a Sequencing-Based Genetic Screen for Pre-mRNA Splicing Mutants in Fission Yeast. G3 (Bethesda). 6: 1513–23.

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–60.

Libri, D., N. Graziani, C. Saguez, and J. Boulay, 2001 Multiple roles for the yeast SUB2/yUAP56 gene in splicing. Genes Dev.

Lim, L. P., and C. B. Burge, 2001 A computational analysis of sequence features involved in recognition of short introns. Proc. Natl. Acad. Sci. U. S. A. 98: 11193–8.

Lin, R. J., A. J. Lustig, and J. Abelson, 1987 Splicing of yeast nuclear pre-mRNA in vitro requires a functional 40S spliceosome and several extrinsic factors. Genes Dev.

Lin, P. C., and R. M. Xu, 2012 Structure and assembly of the SF3a splicing factor complex of U2 snRNP. EMBO J.

Liu, H.-L., and S.-C. Cheng, 2012 The Interaction of Prp2 with a Defined Region of the Intron Is Required for the First Splicing Reaction. Mol. Cell. Biol.

Lustig, a J., R. J. Lin, and J. Abelson, 1986 The yeast RNA gene products are essential for mRNA splicing in vitro. Cell 47: 953–63.

Lybarger, S., K. Beickman, V. Brown, N. Dembla-Rajpal, K. Morey *et al.*, 1999 Elevated levels of a U4/U6.U5 snRNP-associated protein, Spp381p, rescue a mutant defective in spliceosome maturation. Mol. Cell. Biol.

Matera, A. G., and Z. Wang, 2014 A day in the life of the spliceosome. Nat. Rev. Mol. Cell Biol.

Mayas, R. M., H. Maita, and J. P. Staley, 2006 Exon ligation is proofread by the DExD/H-box ATPase Prp22p. Nat. Struct. Mol. Biol.

Mayerle, M., M. Raghavan, S. Ledoux, A. Price, N. Stepankiw *et al.*, 2017 Structural toggle in the RNaseH domain of Prp8 helps balance splicing fidelity and catalytic efficiency. Proc. Natl. Acad. Sci. 114: 4739–4744.

McLaren, W., L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie *et al.*, 2016 The Ensembl Variant Effect Predictor. Genome Biol. 17:.

Network, T. C. G. A., 2012 Comprehensive molecular portraits of human breast tumors. Nature.

Noble, S. M., and C. Guthrie, 1996 Identification of novel genes required for yeast pre-mRNA splicing by means of cold-sensitive mutations. Genetics 143: 67–80.

Noma, T., K. Sode, and K. Ikebukuro, 2006 Characterization and application of aptamers for Taq DNA polymerase selected using an evolution-mimicking algorithm. Biotechnol. Lett. 28: 1939–1944.

Parker, R., and P. G. Siliciano, 1993 Evidence for an essential non-Watson-Crick interaction between the first and last nucleotides of a nuclear pre-mRNA intron. Nature.

Picelli, S., A. K. Björklund, B. Reinius, S. Sagasser, G. Winberg *et al.*, 2014 Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. 24: 2033–2040.

Potashkin, J., R. Li, and D. Frendewey, 1989 Pre-mRNA splicing mutants of Schizosaccharomyces pombe. EMBO J. 8: 551–9.

Qin, D., L. Huang, A. Wlodaver, J. Andrade, and J. P. Staley, 2016 Sequencing of lariat

termini in S. cerevisiae reveals 5’ splice sites, branch points, and novel splicing events. RNA.

Quesada, V., A. J. Ramsay, and C. Lopez-Otin, 2012 Chronic lymphocytic leukemia with SF3B1 mutation. N. Engl. J. Med.

Quinlan, A. R., and I. M. Hall, 2010 BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

Rauhut, R., P. Fabrizio, O. Dybkov, K. Hartmuth, V. Pena *et al.*, 2016 Molecular architecture of the Saccharomyces cerevisiae activated spliceosome. Science (80-.).

Roshbash, M., P. K. Harris, J. L. Woolford, and J. L. Teem, 1981 The effect of temperature-sensitive RNA mutants on the transcription products from cloned ribosomal protein genes of yeast. Cell 24: 679–686.

Schellenberg, M. J., T. Wu, D. B. Ritchie, S. Fica, J. P. Staley *et al.*, 2013 A conformational switch in PRP8 mediates metal ion coordination that promotes pre-mRNA exon ligation. Nat. Struct. Mol. Biol.

Schwer, B., 2008 A Conformational Rearrangement in the Spliceosome Sets the Stage for Prp22-Dependent mRNA Release. Mol. Cell.

Schwer, B., and C. H. Gross, 1998 Prp22, a DExH-box RNA helicase, plays two distinct roles in yeast pre-mRNA splicing. EMBO J.

Semlow, D. R., and J. P. Staley, 2012 Staying on message: Ensuring fidelity in pre-mRNA splicing. Trends Biochem. Sci.

Storey, J. D., and R. Tibshirani, 2003 Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. 100: 9440–9445.

Taggart, A. J., A. M. DeSimone, J. S. Shih, M. E. Filloux, and W. G. Fairbrother, 2012 Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. Nat. Struct. Mol. Biol. 19: 719–721.

Tsai, R. T., R. H. Fu, F. L. Yeh, C. K. Tseng, Y. C. Lin *et al.*, 2005 Spliceosome disassembly catalyzed by Prp43 and its associated components Ntr1 and Ntr2. Genes Dev.

Tseng, C. K., H. L. Liu, and S. C. Cheng, 2011 DEAH-box ATPase Prp16 has dual roles in remodeling of the spliceosome in catalytic steps. RNA.

Turunen, J. J., E. H. Niemelä, B. Verma, and M. J. Frilander, 2013 The significant other: Splicing by the minor spliceosome. Wiley Interdiscip. Rev. RNA.

Urushiyama, S., T. Tani, and Y. Ohshima, 1996 Isolation of novel pre-mRNA splicing mutants of Schizosaccharomyces pombe. Mol. Gen. Genet. 253: 118–127.

Vijayraghavan, U., M. Company, and J. Abelson, 1989 Isolation and characterization of pre-mRNA splicing mutants of Saccharomyces cerevisiae. Genes Dev. 3: 1206–1216.

Villa, T., and C. Guthrie, 2005 The Isy1p component of the NineTeen Complex interacts with the ATPase Prp16p to regulate the fidelity of pre-mRNA splicing. Genes Dev.

Webb, C. J., and J. A. Wise, 2004 The Splicing Factor U2AF Small Subunit Is Functionally Conserved between Fission Yeast and Humans. Mol. Cell. Biol. 24: 4229–4240.

Wilkinson, M. E., S. M. Fica, W. P. Galej, C. M. Norman, A. J. Newman *et al.*, 2017 Postcatalytic spliceosome structure reveals mechanism of 3 ’-splice site selection. Science (80-.).

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

Xu, H., B. J. Fair, Z. Dwyer, M. Gildea, and J. A. Pleiss, 2018 Multiplexed Primer Extension Sequencing Enables High Precision Detection of Rare Splice Isoforms. BioRxiv.

Yan, C., J. Hang, R. Wan, M. Huang, C. C. L. Wong *et al.*, 2015 Structure of a yeast spliceosome at 3.6-angstrom resolution. Science 349: 1182–91.

Yan, C., R. Wan, R. Bai, G. Huang, and Y. Shi, 2016 Structure of a yeast activated spliceosome at 3.5 Å resolution. Science (80-.).

Yeh, T.-C., H.-L. Liu, C.-S. Chung, N.-Y. Wu, Y.-C. Liu *et al.*, 2011 Splicing Factor Cwc22 Is Required for the Function of Prp2 and for the Spliceosome To Escape from a Futile Pathway. Mol. Cell. Biol.