Sequencing of lariat termini in *S. cerevisiae* reveals 5' splice sites, branch points, and novel splicing events

DAOMING QIN,¹ LEI HUANG,² ALISSA WLODAVER,¹ JORGE ANDRADE,² and JONATHAN P. STALEY¹

¹Department of Molecular Genetics and Cell Biology, University of Chicago, Chicago, Illinois 60637, USA

²Center for Research Informatics, University of Chicago, Chicago, Illinois 60637, USA

ABSTRACT

Pre-mRNA splicing is a central step in the shaping of the eukaryotic transcriptome and in the regulation of gene expression. Yet, due to a focus on fully processed mRNA, common approaches for defining pre-mRNA splicing genome-wide are suboptimal—especially with respect to defining the branch point sequence, a key *cis*-element that initiates the chemistry of splicing. Here, we report a complementary intron-centered approach designed to more efficiently, simply, and directly define splicing events genome-wide. Specifically, we developed a method distinguished by deep sequencing of lariat intron termini (LIT-seq). In a test of LIT-seq using the budding yeast *Saccharomyces cerevisiae*, we not only successfully captured the majority of annotated, expressed splicing events but also uncovered 45 novel splicing events, establishing the sensitivity of LIT-seq. Moreover, our libraries were highly enriched with reads that reported on splice sites; by a simple and direct inspection of sequencing reads, we empirically defined both 5' splice sites and branch sites, as well as their consensus sequences, with nucleotide resolution. Additionally, our study revealed that the 3' termini of lariat introns are subject to nontemplated addition of adenosines, characteristic of signals sensed by 3' to 5' RNA turnover machinery. Collectively, this work defines a novel, genome-wide approach for analyzing splicing with unprecedented depth, specificity, and resolution.

Keywords: RNA-seq; lariat intron; pre-mRNA splicing; splice site; branch-point

INTRODUCTION

In eukaryotes, primary transcripts undergo substantial processing before export to the cytoplasm for translation. In the most dramatic processing step, pre-messenger RNAs (pre-mRNAs) are spliced to mature RNAs (mRNAs) through the removal of interrupting sequences (introns) and the concomitant ligation of flanking sequences (exons, Wahl et al. 2009). Pre-mRNA splicing is pervasive in most eukaryotic transcriptomes, such as in humans where genes are interrupted by eight introns, on average (Deutsch and Long 1999; Roy and Gilbert 2006). Additionally, splicing by alternative pathways impacts gene expression qualitatively and quantitatively, by reshaping coding potential or by introducing nonsense codons that induce nonsense mediated decay (NMD; Wang et al. 2008; Nilsen and Graveley 2010; Kalsotra and Cooper 2011). A subset of alternative splicing events are regulated and functionally involved in many key cellular processes including cell division, differentiation, and stress responses (Kalsotra and Cooper 2011). To understand fully the role of splicing in gene expression, it is necessary to monitor splicing genome-wide.

Pre-mRNA splicing proceeds through a two-step transesterification reaction that engages three intronic sites—the 5' splice site, branchpoint, and 3' splice site (Staley and Guthrie 1998). These sites are defined by consensus sequences that are recognized by the spliceosome, the cellular machinery that catalyzes the splicing reaction. In the first transesterification, the 2' OH of the branchpoint attacks the 5' splice site, generating a cleaved 5' exon and a lariat intermediate that is looped by an atypical 2'–5' phosphodiester bond. Then, the 3' OH of the 5' exon attacks the 3' splice site, ligating the exons as linear mRNA and releasing the intron as a branched lariat. Significantly, both the mRNA and lariat products report on the splice sites utilized, while only the lariat product reports on the site of branching.

Most of our current knowledge of splicing on a genome-wide scale comes from studies of the mRNA product, which is relatively abundant, easy to enrich, and biologically relevant for translation. In array-based investigations of mRNA, known splicing events are monitored by probes that specifically target the annotated junctions of mRNA

Corresponding author: jstaley@uchicago.edu

Article published online ahead of print. Article and publication date are at <http://www.rnajournal.org/cgi/doi/10.1261/rna.052829.115>.

© 2016 Qin et al. This article is distributed exclusively by the RNA Society for the first 12 months after the full-issue publication date (see <http://rnajournal.cshlp.org/site/misc/terms.xhtml>). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

(Clark et al. 2002; Pleiss et al. 2007a,b). In a widely employed application of RNA-seq to poly(A)-enriched mRNA, splicing events for a gene are either inferred by a gap in read coverage or observed through exon–exon junction reads that can be split and mapped to two discontinuous locations within the gene (e.g., Nagalakshmi et al. 2008; Wang et al. 2008; Kawashima et al. 2014). While sequencing of mRNA has facilitated the relative quantification of splicing events, enabled the genome-wide assembly of mRNA *de novo*, and allowed the refinement of existing genome annotations (Martin and Wang 2011), such approaches suffer from several drawbacks for studies that focus on splicing. First, mRNA sequencing underrepresents a significant population of splicing events that trigger decay of an mRNA, such as by NMD (McGlinchy and Smith 2008; Kawashima et al. 2014). Second, because most short mRNA reads do not span spliced junctions, mRNA sequencing reports on splicing inefficiently (e.g., ~0.3% in budding yeast and 5% in human; Wang et al. 2008; Nagalakshmi et al. 2008). Lastly, sequencing of mRNA fails to reveal branch sites. In a recent analysis of hundreds of mRNA-seq data sets, contaminating intronic RNA did reveal branch sites but only at a frequency of one read per million (Taggart et al. 2012). While some of these limitations of sequencing the mRNA products of splicing may be mitigated by longer reads or deeper sequencing, these improvements would come at a financial cost.

For an investigation focused on splicing, the lariat intron, the other product of splicing, offers a number of advantages over mRNA. First, the lariat intron will more effectively report on splicing events that result in NMD-sensitive mRNA isoforms (Awan et al. 2013). Second, the lariat intron reports on splicing efficiently, because every read from a lariat intron reflects a splicing event. Further, targeted sequencing of lariat termini offers a potential for every read to reflect a unique splicing event, thereby increasing the depth for detecting splicing. Third, excised introns can reveal all of the splicing signals used; the 5' terminus of the lariat reports on the 5' splice site, the 3' terminus of an intact lariat intron defines the 3' splice site, and the 3' terminus of a lariat intron, after 3'–5' exonuclease trimming *in vivo* or *in vitro*, reveals the branchpoint. Thus, the lariat intron product offers a faithful, efficient, and direct alternative to mRNA for studying splicing genome-wide. Importantly, capturing the full landscape of introns, or the intronome, can provide insight into the production of not only mRNAs but also intron-encoded ncRNAs, such as small nucleolar RNAs, microRNAs, and some intronic circRNAs (Zhang et al. 2013; Hesselberth 2014; Talhouarne and Gall 2014).

In studies of splicing from the perspective of excised introns, introns have been enriched by stabilization in a strain in which the gene *DBR1* was deleted (Spingola et al. 1999; Juneau et al. 2007; Zhang et al. 2007; Awan et al. 2013; Bitton et al. 2014; Stepankiw et al. 2015); *DBR1* encodes a debranchase that linearizes a lariat through hydrolysis of the unique 2'–5' phosphodiester bond—a prerequisite for

rapid, exonucleolytic degradation (Chapman and Boeke 1991). Early microarray-based studies inferred the location of introns as those genomic regions in which RNA levels increased in the *dbp1Δ* strain, relative to a wild-type strain (Juneau et al. 2007; Zhang et al. 2007). While these studies verified annotated introns and uncovered novel introns, these array-based methods were limited by low sensitivity, narrow dynamic range, and poor nucleotide resolution, which complicated a precise definition of splicing signals. More recently, in *Schizosaccharomyces pombe*, random priming-based RNA-seq was applied to lariat introns that were similarly enriched using a *dbp1Δ* strain, leading to the detection of splicing events genome-wide including hundreds of novel splicing events (Awan et al. 2013; Bitton et al. 2014; Stepankiw et al. 2015). This approach, along with an earlier large-scale analysis of human transcriptome data, also revealed both 5' splice sites and branch sites through branched junction reads that traversed the 2'–5' phosphodiester bond (Taggart et al. 2012; Awan et al. 2013; Bitton et al. 2014; Stepankiw et al. 2015). Recently, deep sequencing was focused further on the branched structure of lariats by 3'–5' exonuclease treatment or the capture of branched junction reads with probes that anneal immediately downstream from annotated 5' splice sites or upstream of 3' splice sites. Consequently, ~25% of human branch sites were mapped and analyzed (Mercer et al. 2015). However, these intron-centered, RNA-seq approaches to study splicing genome wide have not yet exploited the full depth of next-generation sequencing. First, targeting the entire intronic RNA by random priming resulted in reads that redundantly reported on splicing events and failed to focus on the branched junction reads that defined splicing signals. While probes can specifically capture such branched junction reads, this method demands prior knowledge of 5' and 3' splice sites. Second, targeting the entire intronic RNA biased toward large introns, because large introns were preferentially primed. Third, because reverse transcriptase generally traverses a branch site inefficiently, the fraction of reads that reported on branch points were few—only ~3% of total mapped reads, at best (Awan et al. 2013; Bitton et al. 2014; Mercer et al. 2015; Stepankiw et al. 2015). Finally, with one exception (Mercer et al. 2015), these intron-based approaches rely on strains that lack debranchase activity.

Here, using *Saccharomyces cerevisiae* as a model system, we circumvented these limitations by developing LIT-seq, a novel approach that centers on lariat termini and focuses the full capacity of RNA-seq on splicing events. Despite extensive studies of introns in *S. cerevisiae*, our approach revealed 45 novel splicing events while capturing the majority of annotated, expressed introns, establishing the sensitivity of the approach. Further, our approach empirically defined, in addition to 5' splice sites, 60% of intronic branch sites; indeed, up to 90% of total mapped reads reported on branch sites, establishing the utility of this method for broad and efficient branch-site mapping. In addition, our approach

revealed that the 3' termini of lariat introns were appended with nontemplated adenosines, a hallmark of RNA targeted by the exosome for decay (Jia et al. 2011; Schneider et al. 2012). Finally, by enriching for introns through coimmunoprecipitation of the spliceosome, we established that LIT-seq reveals the intronome even in a yeast strain expressing debranchase activity, thereby implying a broad and general utility for LIT-seq.

RESULTS

Strategy to capture and sequence lariat termini

To efficiently detect splicing events and to directly define splicing signals, we developed a strategy to enrich for and capture lariat termini (Fig. 1). To test our strategy, we utilized the model organism budding yeast, given its well-characterized genome and transcriptome. Since lariats turn over rapidly and represent low-abundance species in total RNA, we at first initiated the enrichment of lariats by a genetic approach utilized in earlier intronome studies, in which a *dbp1Δ* mutant strain was employed to stabilize lariat introns; subsequently, we validated an alternative, more general approach (see below). In the *dbp1Δ* mutant strain, the termini of a lariat reveal the 5' splice site and branch site, because the RNA 3' of the branch site is subject to trimming by endogenous 3'–5' exonucleases (Fig. 1A; Chapman and Boeke 1991; Zhang et al. 2007). To further enrich for lariat introns from whole-cell RNA of *dbp1Δ* cells, we discriminated against linear RNAs having a 5' phosphate that would incorporate into the sequencing library (see below). To do so, we first degraded linear RNAs bearing a 5' phosphate using XRN-1, a 5'–3' exonuclease, and then removed the 5' phosphate of remaining linear RNAs by calf-intestinal alkaline phosphatase (CIP). Note that, similar to lariat RNAs, other RNA species with a blocked 5' end (e.g., mRNA) will resist XRN-1 and CIP treatment. Thus, we additionally enriched for lariats by selecting against 5' blocked, linear RNAs by requiring ligation of an adapter to the 5' end of an RNA. To enable ligation of an adapter to the 5' end of introns, after ligation of an adapter to the 3' end of RNAs, we debranched lariats with rDbr1 (or not, in a control), which conveniently reveals a 5' phosphate required for adapter ligation. By RT-PCR, we then selectively amplified intronic species having adapters ligated to both termini. To capture the termini of introns, we employed a restriction enzyme MmeI to bind asymmetric sites engineered into each adapter and to direct cuts 20–21 nt into the intron, generating 5' and 3' terminal lariat tags. We then ligated adapters to the MmeI overhangs and prepared the library for sequencing (see Materials and Methods). Overall, four libraries were constructed, namely 5'(rDbr1+), 3'(rDbr1+), 5'(rDbr1–), and 3'(rDbr1–), representing the 5' and 3' terminal tag libraries treated or not with rDbr1p. We term this approach LIT-seq, for lariat intron termini sequencing.

FIGURE 1. LIT-seq: a strategy for sequencing the intronome. (A) Fate of an excised intron in wild-type *DBP1* and mutant *dbp1Δ* strains. In wild-type cells, the lariat RNA is debranched and degraded by exonucleases. In *dbp1Δ* cells, lariat RNAs accumulate in whole-cell RNA due to loss of debranching activity, although the RNA downstream from the branch point is exonucleolytically degraded. (B) LIT-seq workflow. Lariat RNA is initially enriched by harvesting RNA from *dbp1Δ* cells or by immunoprecipitating spliceosomes. The RNA is then treated with XRN-1 and CIP and, optionally, RNase R. The remaining RNA is ligated to a 3' adapter (green); debranched by rDbr1, which reveals a 5' phosphate, and then ligated to a 5' adapter (peridot). After RT-PCR, the products are digested with MmeI to generate two fragments corresponding to the 5' and 3' termini of an intronic RNA. These splicing signal tags are ligated to appropriate duplex adapters and amplified by PCR to finalize the libraries for deep sequencing. Barcodes (ruby) are introduced either during the first RT step, for the 3' library, or during duplex ligation, for the 5' library (see Materials and Methods for details).

LIT-seq identified the majority of annotated introns in budding yeast

In our test of LIT-seq, we obtained roughly 10 million reads for each of the four libraries (Supplemental Table S1). We first mapped the reads to the genome demanding a perfect match. Because we found that the 3' termini of the lariat introns were frequently tailed with several nontemplated adenosines (see below), we trimmed adenosines from the 3' termini of unmapped 3' tag reads one by one, collecting reads that subsequently mapped perfectly. We refer to these reads as “A-tailed” reads and to the initially mapped reads as “tailless” reads; we refer to the collection of these reads as the mapped reads. To normalize 5' and 3' tag read numbers to library size, we report the number of tag reads per million mapped reads (RPM).

LIT-seq proved highly effective in detecting introns. First, LIT-seq detected introns efficiently and specifically. Over 25% of total mapped reads from the 5'(rDbr1+) library mapped to intronic regions (Supplemental Table S1; Supplemental Fig. S1A,B). Note, in this trial of LIT-seq, without RNase R treatment (compare below), the 3'(rDbr1+) library was dominated by reads that mapped to multiple loci (i.e., rRNA reads, Supplemental Fig. S1C,D). Even so, when considering only the total uniquely mapped reads, over 85% of the reads from the 3'(rDbr1+) library mapped to intronic regions, comparable to the 5'(rDbr1+) library (93%, Fig. 2A), and roughly 100 reads mapped per intron, on average—more than enough for our analysis; indeed, the 3'(rDbr1+) library sufficed to yield findings qualitatively similar to those derived from the 5'(rDbr1+) library (see below and Supplemental Note 1). Importantly, these intronic reads are dependent on debranching by rDbr1, because debranched libraries yielded a 15-fold higher percentage of intronic reads than undebranched control libraries (Fig. 2A; Supplemental Fig. S1A–D; Supplemental Table S1; Supplemental Note 2). An example of this specificity and

rDbr1-dependence is illustrated for the *ARP2* gene (Fig. 2C). In the 5' and 3' debranched libraries, 2238 RPM and 22 RPM, respectively, mapped to the intron of *ARP2*, whereas no reads mapped to the exons. In the 5' and 3' control libraries, untreated with rDbr1, at least 20-fold fewer reads mapped to the intron—49 RPM and 0 RPM, respectively. Second, LIT-seq detected introns broadly. The 5' and 3' rDbr1-treated libraries revealed ~75% of known introns—225 and 238, respectively, out of the 298 currently annotated introns (Fig. 2B; Supplemental Table S2). This coverage is comparable to that observed in previous studies of the budding yeast intronome and reflects in part the repression of intron-containing genes (e.g., meiotic genes) during growth in rich media (Juneau et al. 2007; Zhang et al. 2007). Thus, LIT-seq identified the majority of annotated introns efficiently and specifically.

In addition, LIT-seq detected introns over a wide dynamic range; the number of reads mapping to introns varied by as much as 1000-fold for both the 5'(rDbr1+) and 3'(rDbr1+) libraries (Supplemental Fig. S2A). In addition, despite differing absolute numbers of reads, the number of 5' and 3' reads for a given intron correlated (Supplemental Fig. S2B).

FIGURE 2. LIT-seq detected introns broadly and specifically. (A) The majority of uniquely mapped 5' and 3' reads mapped to annotated introns. The percentage of uniquely mapped reads that mapped to annotated introns is shown for the experimental intron libraries, 5'(rDbr1+) and 3'(rDbr1+), and the control libraries, 5'(rDbr1-) and 3'(rDbr1-), derived from *dbr1Δ* cells and without RNase R treatment. (B) The 5'(rDbr1+) and 3'(rDbr1+) libraries each captured the majority of annotated introns. The Venn diagram illustrates the numbers of introns detected by both libraries, by one or the other library, or by neither library. (C) The 5'(rDbr1+) and 3'(rDbr1+) reads mapped specifically to 5' splice sites and branch sites, respectively, as illustrated for the annotated intron of the gene *ARP2*. The read numbers for the 5'(rDbr1+) library (red) and 3'(rDbr1+) library (blue) are shown, in comparison to read numbers for the control libraries (gray). Because the 3'(rDbr1+) libraries contained fewer intronic reads than the 5'(rDbr1+) library, a different scale was used. A zoomed-in track is shown for the 5'(rDbr1-) library (see Supplemental Note 3); no reads were observed for the corresponding track in the 3'(rDbr1-) library. Data are displayed in units of RPM, reads per million mapped reads.

LIT-seq defined 5' splice site and branch-point consensus sequences empirically

For the 5'(rDbr1+) library, we designed LIT-seq to capture the very 5' terminus of an intron, including the first six nucleotides, which define the 5' splice site. Indeed, for 5' reads mapping to an intron, the 5' end typically mapped to the 5' boundary of the intron (e.g., Fig. 2C). In a meta-analysis of intron-mapped 5' reads using intron windows normalized to the distance between the 5' and 3' splice sites, a vast majority of 5' reads (99%) mapped to the 5' splice site in the 5'(rDbr1+) library (Fig. 3A). In the control library, the number of intronic reads, scaled to total mapped reads (RPM), decreased substantially (Fig. 3B); note that, a significant portion of intronic reads mapped unexpectedly to the 5' splice site (Supplemental Note 3). Further, in a 20-bp window centered on each 5' splice site, the most upstream residue of the 5'(rDbr1+) reads mapped predominantly to the first nucleotide of the intron (Fig. 3C). In total, 5' reads reported on 75% of annotated 5' splice sites. Indeed, when we queried all unique 5' reads, mapping anywhere in the

FIGURE 3. Genome-wide identification of 5' splice sites and branch sites. (A–F) The 5'(rDbr1+) and 3'(rDbr1+) reads mapped specifically to 5' splice sites and branch sites, respectively, as illustrated genome-wide. (A,B) A meta-analysis of the number of 5'(rDbr1+) or 5'(rDbr1-) reads, respectively, that mapped to a particular position in an intronic window, in which the distance between the 5' and 3' splice sites was normalized. (C) A meta-analysis of the number of 5'(rDbr1+) reads mapping within a 20-nt window centered on each annotated 5' splice site. In A–C, the position of a read was determined using the first, 5' nucleotide of the read. (D,E) A meta-analysis of the number of 3'(rDbr1+) or 3'(rDbr1-) reads, respectively, mapping to a particular position in an intronic window in which the distance between the 5' splice site and branch point was normalized. (F) A meta-analysis of the number of 3'(rDbr1+) reads mapping within a 20-nt window centered on each annotated branch point. In D–F, the position of a read was determined using its most 3' genomically encoded nucleotide. The inset in B shows a zoomed in view. Note: The two highest peaks in the 5'(rDbr1-) library in B and the two highest peaks in the 3'(rDbr1-) library in E correspond to reads mapping to the 5' and 3' ends of potential stem loops residing in the introns of *RPL28*(*) and *RPS17A*(**), potentially reflecting endonucleolytic cleavage of these two introns (Supplemental Note 4). (G,H) LIT-seq successfully derived the 5' splice site and branch site consensus sequences. Panel G compares the consensus sequence for annotated 5' splice sites with the consensus sequences empirically derived from the uniquely mapped reads of the 5'(rDbr1+) and 5'(rDbr1-) libraries. Panel H compares the consensus sequence for annotated branch sites with the consensus sequences empirically derived from the uniquely mapped reads of the 3'(rDbr1+) and 3'(rDbr1-) libraries.

genome, for a sequence motif, we derived the motif sequence GUAUGU, which matches perfectly the consensus defined for annotated 5' splice sites in budding yeast (Fig. 3G; Spingola et al. 1999; Davis et al. 2000; Miura et al. 2006); in contrast, the control 5'(rDbr1-) library did not yield a 5' splice site consensus sequence (Fig. 3G), but instead yielded a motif that derived primarily from H/ACA box small nucleolar RNAs (e.g., snR9). Overall, this analysis establishes that LIT-seq can broadly define 5' splice sites empirically with single-nucleotide resolution.

Similarly, for the 3'(rDbr1+) library, we designed LIT-seq to capture the 3' terminus of an intron, which in the *dbr1Δ* strain corresponds to the branch site (Chapman and Boeke 1991; Zhang et al. 2007). Indeed, for 3' reads mapping to an intron, the last nucleotide typically mapped just downstream from the branch site (e.g., in *ARP2*; Fig. 2C). In a meta-analysis of intron-mapped 3' reads using intron windows normalized to the distance between the 5' splice site and branchpoint, the most 3' residue of a majority of reads (~60%) mapped just downstream from the branch site in the debranched but not the control library (Fig. 3D,E; Supplemental Note 3). Unlike the 5' reads, the 3' reads did not map precisely to the branch site but rather mapped with an increasing probability of close proximity to the branch site; most reads mapped 2-nt downstream from the annotated branchpoint (Fig. 3F). These 3' reads mapped to ~50% of annotated branch sites. Indeed, when we queried all 3' reads, uniquely mapping anywhere in the genome, for a sequence motif, we derived the motif sequence UACUAAC, which matches perfectly the consensus for annotated yeast branch sites defined largely through bioinformatics (Fig. 3H; Spingola et al. 1999; Davis et al. 2000; Kellis et al. 2003); in contrast, the control library did not yield a branch site consensus sequence (Fig. 3H), but instead yielded a motif that derived primarily from tRNAs (e.g., tRNA-Gln). Overall, this analysis establishes that LIT-seq can broadly define 5' splice sites and branch sites empirically.

LIT-seq revealed adenosine tailing of the 3' termini of introns

In pilot experiments, cloning and Sanger sequencing of a sampling of the rDbr1+ libraries, before MmeI digestion, revealed that lariat introns were frequently appended with non-templated adenosine at the 3' terminus, as illustrated in Figure 4A. Mapping of the A-tailed reads from deep sequencing of the 3'(rDbr1+) library confirmed that adenosine tailing occurred genome-wide, with >45% of the detected introns containing short A tails and with 99% of the A-tailed reads deriving from introns (Supplemental Table S1). Given that such adenosine tails are a hallmark of RNAs that have been marked by the TRAMP complex for degradation by the exosome (LaCava et al. 2005; Jia et al. 2011), adenosine tailing of introns implicates post-transcriptional modification of lariats for turnover (see Discussion).

A global analysis of the parameters of intronic adenosine tailing revealed several features. First, tailing is specific to adenosine; other nontemplated nucleotides were rare and occurred at the frequency expected for sequencing errors, with the exception of 1- and 2-nt cytosine tails consistent with tRNA 3' ends (Supplemental Table S3). Second, the length of adenosine tails is most frequently 1 nt and extends

up to 7 nt (Fig. 4B; Supplemental Table S3); note, although adenosine tails of longer than 7 nt are possible, we did not search for such, likely rare, tails for technical reasons (see Materials and Methods). This length distribution differs slightly from that in previously described TRAMP-dependent, adenosine tailed RNAs, which similarly contain short adenosine tails, but with 4 nt being the most frequent length (Jia et al. 2011; Wlotzka et al. 2011; see Discussion). Finally, intronic adenosine tailing occurred exclusively downstream from branch sites, with a frequency of >99%, in sharp contrast to the distribution of total 3' intronic reads (cf. Figs. 3D, 4C); significantly fewer A-tailed reads mapped to the intron in the undebranched control library (Fig. 4D). Thus, adenosine tailing uniquely marked branch site reads, which could be leveraged to improve branch site identification by LIT-seq. Note, adenosine tails were appended predominantly to the second genomically encoded nucleotide downstream from the branch point (Fig. 4E), corresponding precisely to the terminal nucleotide most frequently represented in the tailless 3' reads mapping to the branch site (Fig. 3F; Supplemental Fig. S3A). Overall, our discovery of intronic adenosine tails, missed by alternative approaches (Awan et al. 2013; Bitton et al. 2014; Mercer et al. 2015), highlights the utility of LIT-seq in targeting lariat termini.

Given that RNA ligase has a preference for RNAs with long unstructured 3' tails (Zhuang et al. 2012), the short 3' tail of lariats in *dbr1Δ* cells may inhibit ligation to the 3' adapter. In this regard, oligo(A) tails, especially the longer tails, would partially relieve the inhibition and reveal even shorter genomically encoded tail lengths (e.g., 1 nt). However, we observed that the genomically encoded tail length is most frequently 2 nt, regardless of the oligo(A) tail length (Supplemental Fig. S3A), implying that lariats with 2-nt tails predominate in *dbr1Δ* cells. Still, reads lacking any oligo(A) tail showed the lowest frequency of 2-nt tails, so we investigated the efficiency of 3' adapter ligation before and after debranching for two introns, from *U3A* and *ACT1*. Whereas the ligation of linear, debranched RNAs was efficient, approaching 90%, the ligation of lariat RNAs was less efficient and more variable (i.e., 20%–70%; Supplemental Fig. S3B, C). In contrast, debranching was efficient, at least 85%, for unligated or ligated lariats for both *U3A* and *ACT1*, as expected (Supplemental Fig. S3; Chapman and Boeke 1991; Khalid et al. 2005). Indeed, we observed increased overall levels of products when ligating a 3' adapter after, rather than before, debranching (Supplemental Fig. S3B,C, cf. lanes 4 and 5). Thus, for future applications of LIT-seq, we advise implementing 3' adapter ligation after debranching. Despite the potential for introducing bias when ligating a 3' adapter to a trimmed lariat, three introns, from *RPS24B*, *ACT1*, and *U3A*, revealed 2-nt tails regardless of whether we ligated the 3' adapter before or after debranching (Supplemental Fig. S3D–F), validating our conclusion that lariats with 2-nt tails predominate in *dbr1Δ* cells.

FIGURE 4. Adenosine tailing of lariat 3' termini. (A) The 3' termini of lariat introns were frequently appended just downstream from the branch site with short adenosine tails that were not encoded by the genome, as illustrated for the intron from *ACT1*. The adenosine tails are underlined. (B) Tails with one adenosine were most common, but longer tails with up to seven adenines were observed. The histogram shows the number of 3'(rDbr1+) reads having an A-tail of indicated length, from zero to seven; the zero-length tails correspond to the "tailless" class of reads. (C–E) Adenosine tailing occurs just downstream from branch site. Reads in C and D were plotted as in Figure 3D,E, respectively, except that only A-tailed reads were plotted. Reads in E were plotted as in Figure 3F except that only A-tailed reads were plotted.

RNase R enhanced enrichment for branch-site reads

Previous studies of lariat intron turnover in a *dbp1Δ* strain has implicated endonucleolytic cleavage of lariats that provides a slow but alternative mechanism for turnover, in the absence of debranching (Ooi et al. 1998; Danin-Kreiselman et al. 2003). Because LIT-seq does not inherently discriminate against the branching strand of such cleaved species, we expected that LIT-seq, as implemented above, would detect such branched “Y-shaped” species (Supplemental Fig. S4A). Indeed, for the 3′ intronic reads, a significant portion (~40%) did not map to an annotated branch site. Instead, these reads mapped within the intron lariat, between the 5′ splice site and branch point, with a bias toward the 5′ splice site (Fig. 3D), consistent with cleavage of a lariat upstream of the branch site. These upstream reads compromised the efficiency and specificity of LIT-seq in defining branch sites. Because such “Y-shaped” branched structures are sensitive to turnover by the 3′–5′ exonuclease RNase R (Suzuki et al. 2006), we modified LIT-seq to include digestion by RNase R, which also further discriminated against linear RNAs (Supplemental Fig. S4; Suzuki et al. 2006; Jeck et al. 2013; Mercer et al. 2015). As observed above with the RNase R untreated libraries, over 65% of 5′ and 3′ reads mapped to introns, 5′ and 3′ reads reported on ~90% of annotated introns on average, and over 96% of the intronic 5′ reads mapped to annotated 5′ splice sites (Supplemental Figs. S1E–H, S4B–E; Supplemental Table S1). As designed, in 3′ libraries treated with debranchase and RNase R, named 3′(rDbr1+, RNase R+), at least 94% of reads that mapped to introns mapped to branch sites (Supplemental Fig. S4F,G; Supplemental Table S1) and these reads covered 66% of annotated branch sites. Further, at least 30% of all mapped reads and at least 80% of uniquely mapped reads mapped to branch sites, an efficiency of defining branch sites that exceeds alternative protocols tested in other organisms by >10-fold (Awan et al. 2013; Bitton et al. 2014; Mercer et al. 2015; Stepankiw et al. 2015). Thus, LIT-seq establishes an unprecedented capacity to detect and define branch sites.

To judge the reproducibility of LIT-seq, in the above experiments evaluating the utility of RNase R, we performed LIT-seq on whole-cell RNA derived from three independent *dbp1Δ* yeast colonies to generate three experimental, replicate libraries with both 5′ and 3′ reads. Also, because LIT-seq yields splice site tags and many identical reads, to eliminate 3′ reads resulting from PCR duplicates, we uniquely labeled each RNA molecule by introducing a random barcode in the 3′ adapters (Shiroguchi et al. 2012). After removing PCR duplicates from the library of 3′ reads, the remaining reads still mapped predominately to branch sites (Supplemental Table S1, also cf. Supplemental Fig. S4F–I). Further, the total intronic 3′ reads and the corresponding intronic 3′ reads with PCR duplicates removed showed a tight correlation, with a Pearson r^2 value of ~0.8, suggesting that PCR amplification did not significantly bias the distribution of

reads (Supplemental Fig. S5D). Importantly, the 5′ reads, 3′ reads, and 3′ reads with PCR duplicates removed showed clear correlations between biological replicates, with Pearson r^2 values ranging from 0.4 to 0.64 (Supplemental Fig. S5A,B; data not shown). Thus, we conclude that LIT-seq is reproducible.

The reproducibility of LIT-seq implies that splice site tags could be utilized to quantify relative changes in intron levels. Still, neither the 5′ nor 3′ tags, derived with or without RNase R treatment, revealed correlations with nascent or mRNA transcripts levels, as determined by NET-seq (Supplemental Fig. S6A–F; Churchman and Weissman 2011) or RNA-seq (Supplemental Fig. S6G; Yassour et al. 2009), which do correlate with one another (S6H). Additionally, the number of intron reads from LIT-seq libraries treated with or without RNase R showed a poor correlation with one another, indicating bias in one approach, the other, or both (Supplemental Fig. S5E,F), such as degradation of Y-shaped branched structures in the RNase R-treated libraries. Still, the correlation between 5′ and 3′ reads (Supplemental Figs. S2B, S5C, S7G), between replicates (Supplemental Fig. S5A,B), and between reads dependent or not on PCR amplification (Supplemental Fig. S5D), implies that intron reads would reflect relative changes, rather than absolute, intron levels.

LIT-seq can be applied to wild-type cells

Given the specificity and sensitivity of LIT-seq, we next tested whether LIT-seq could be implemented effectively in wild-type cells without deleting debranchase to accumulate and enrich for lariats, as is commonly practiced in genome-wide intron studies (Juneau et al. 2007; Awan et al. 2013; Bitton et al. 2014; Stepankiw et al. 2015). An alternative to enriching for lariats by stabilization is to enrich for lariats by coimmunoprecipitation with the spliceosome; in extracts from wild-type cells, immunoprecipitation of endogenous spliceosomes coprecipitates lariats (e.g., Small et al. 2006; Chen et al. 2014). To test whether lariat enrichment by immunoprecipitation of the spliceosome enables the application of LIT-seq to wild-type cells, we immunoprecipitated spliceosomes using antibodies against Prp16, one of the factors that associates with the spliceosome at the lariat intermediate stage and through coimmunoprecipitation enriches for lariat intermediates (Schwer and Guthrie 1991). After extracting RNA from Prp16 immunoprecipitation, we implemented LIT-seq as above, with RNase R treatment to reveal branch point sequences; LIT-seq yielded two experimental libraries, namely Prp16-5′(rDbr1+, RNase R+) and Prp16-3′(rDbr1+, RNase R+), and two control libraries in which debranchase was omitted, namely, Prp16-5′(rDbr1–, RNase R+) and Prp16-3′(rDbr1–, RNase R+).

In the detection of introns and splice sites by LIT-seq, lariat enrichment by coimmunoprecipitation yielded results qualitatively similar to those derived from lariat enrichment by debranchase deletion. First, LIT-seq again proved sensitive

and efficient; >80% of mapped reads mapped to introns (Supplemental Table S1; Supplemental Figs. S1I–L, S7A). Second, LIT-seq again specifically defined splicing sites; >99% of the intronic reads mapped to 5' splice sites or branch-point sequences (Supplemental Table S1; Supplemental Fig. S7C–F). Third, the 5' and 3' libraries again correlated (Supplemental Fig. S7G). While the Prp16-IP library did not correlate with the rDbr1+ library (e.g., Supplemental Fig. S7J), the Prp16-IP library did show some correlation with the other RNase R-treated libraries (Supplemental Fig. S7H,I). Fourth, LIT-seq detected many annotated introns (~50%, Supplemental Fig. S7B); note, this coverage is somewhat lower than the 75% detected with enrichment from *dbp1Δ* cells, which could reflect a differential residency of lariat intermediates in Prp16-bound spliceosomes. Lastly, LIT-seq by this method also detected novel introns (see below). We conclude that LIT-seq, when coupled to spliceosome immunoprecipitation, can be applied broadly to wild-type cells to efficiently and specifically detect lariats and associated splice sites.

LIT-seq uncovered novel splicing events

Because in LIT-seq libraries a significant portion of uniquely mapped reads mapped to regions outside of annotated introns, we investigated whether any of these reads reflected novel splicing events. To query the reads for evidence of novel splicing events, we searched by inspection for transcribed regions of the genome in which at least two reads from both the 5' and 3' libraries (i) mapped with the 5' reads upstream of the 3' reads, (ii) mapped within established distances for annotated introns—50 bp to 1 kb (Spingola et al. 1999), and (iii) contained sequences that resemble the consensus sequences for the 5' splice site and branch site, allowing up to two mismatches (Supplemental Note 5). Note: For 5' reads, we required 5' splice site-like sequences to reside precisely at the 5' termini of the reads, whereas for 3' reads, we required branch site-like sequences but from 1 to 7 nt upstream of the 3' termini of the reads. Based on these criteria, we identified 45 novel splicing events not previously reported, with some of these novel introns appearing in more than one library (Supplemental Table S4).

Here, we focused on validating novel introns revealed by the 5'(rDbr1+) and 3'(rDbr1+) libraries. In these libraries, we detected novel splicing events reported in recent studies (e.g., *BDF2* and *GCR1*; Harigaya and Parker 2012; Volanakis et al. 2013; Kawashima et al. 2014). Further, we detected 12 novel splicing events not reported elsewhere. Because we recovered fewer uniquely mapped reads in the 3'(rDbr1+) libraries, we also relaxed the search criteria by requiring only 5' reads in which 5' splice site-like sequences resided at the 5' termini of the reads. By this criteria, we identified eight additional novel splicing events. Together, LIT-seq provided evidence for 20 novel splicing events. We were able to confirm

16 by independent methods (see below). These validated novel splicing events divided into three classes.

The first class of such novel splicing events was associated with annotated introns and reflected the use of alternative, upstream 5' splice sites. These alternative 5' splice sites reside in two genes—*MATa1* and *RPS14B* (Fig. 5A–D; Supplemental Fig. S8A–C). We verified these novel splicing events by a “lariat RT-PCR” assay that amplifies a branched junction and reveals the 5' splice site and branch site at single nucleotide resolution (Fig. 5B; Gao et al. 2008). In a complementary assay, we verified the novel exon–exon junctions by standard RT-PCR of the mRNA, which also verified that the 3' splice site corresponded to the first AG downstream from the branch point. For *MATa1* and *RPS14B*, we observed splicing at both the annotated and the novel 5' splice site (Fig. 5B–D, Supplemental Fig. S8B,C; data not shown), consistent with our recovery of reads from both sites. Importantly, these novel splicing events are not unique to the *dbp1Δ* strain, because by mRNA RT-PCR we also detected these events in a wild-type strain or strains expressing debranchase activity (e.g., *upf1Δ*; Fig. 5C).

Notably, there are several common features associated with these alternative 5' splice sites. First, these alternative 5' splice sites occurred upstream of the annotated 5' splice site. However, we did recover a recently reported alternative 5' splice site downstream from the annotated 5' splice site in *RPS22B* (data not shown; Kawashima et al. 2014). Second, the associated 5' splice site sequence deviated from the consensus (Supplemental Table S4), indicating inefficient splicing signals. Third, the mRNAs associated with the alternative 5' splice sites represent minor isoforms (Fig. 5C; Supplemental Fig. S8C), reflecting either inefficient splicing or biased degradation of these isoforms. Despite the fact that splicing at the alternative 5' splice sites in *MATa1* and *RPS14B* shifts the reading frame and introduces a premature termination codon that could in principle trigger NMD, splicing isoforms associated with these alternative splice sites remain as a minor isoform even when NMD is inhibited. Specifically, in strains deficient for NMD due to knockout of *UPF1* or *UPF2*, the *MATa1* novel isoform did not increase to any significant extent, and while the *RPS14B* novel isoform was detected, the level was still significantly less than the annotated isoform (Fig. 5C; Supplemental Fig. S8C). Previous studies have found that certain unspliced pre-mRNAs or exon-skipped mRNAs, despite encoding premature termination codons, are insensitive to NMD (Sayani et al. 2008; Eggecioglu et al. 2012); indeed, nuclear decay pathways have been implicated in the turnover of these species (Bousquet-Antonelli et al. 2000; Eggecioglu et al. 2012; Bitton et al. 2015). These observations highlight the utility of LIT-seq in enriching for splicing events that would not be significantly enriched by inactivation of NMD, an alternative approach for revealing novel splicing events (Kawashima et al. 2014).

The second class of novel splicing events were also associated with annotated introns but reflected the use of

FIGURE 5. (A–F) LIT-seq revealed novel alternative splicing sites, as illustrated by an alternative 5' splice site in *MATa1* (A) and an alternative branch site in *YBR255C-A* (E). (A,E) Gene annotation and relative read numbers are displayed as in Figure 2C. In E, conservation scores (Max = 1) between seven yeasts (*S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castelli*, and *S. kluyveri*) are shown in green. Note: In A, in the control track (rDbr1–), the reads mapping to the annotated 5' splice site show higher RPM than the experimental track, but this was not observed for other introns or libraries. (B,F) Verification of the novel 5' splice site and branch site via the intron product. The PAGE gels show lariat RT-PCR analysis of the novel branch junctions from WT and *dbr1Δ* cells. The traces show Sanger sequencing of lariat RT-PCR products, with the 5' splice site (red) and linked branch site (blue) highlighted. As observed previously (e.g., Gao et al. 2008), mutations, insertions, and deletions were frequently associated with lariat RT-PCR across a branch junction and are underlined here. (C,D) Verification of the novel 5' splice site via the mRNA product. In C, the PAGE gel shows mRNA RT-PCR analysis of the inferred, novel exon–exon junction, from WT, *dbr1Δ*, *upf1Δ*, and *upf2Δ* cells; in D, Sanger sequencing confirmed the novel junction resulting from the alternative 5' splice site and annotated 3' splice site.

alternative branch sites. These branch sites resided in two genes—*YBR255C-A* and *AML1*. In the case of *YBR255C-A*, the 3'(rDbr1+) reads did not map to the annotated branch site, AAUUAAC; instead, the reads mapped 15-nt upstream to a site that included the branch site-like sequence CGCUAAC, as verified by lariat RT-PCR (Fig. 5E,F). Notably, this novel branch site sequence along with several

others (Supplemental Table S4) have never before been reported in budding yeast. Importantly, we found that the new branch site is strongly conserved among different yeast species, whereas the annotated site is not (Fig. 5E), providing additional evidence that CGCUAAC is the authentic branch site for *YBR255C-A*. We also confirmed that *AML1* branches at a site that differs from the predicted branch site

(Supplemental Fig. S8D,E). Notably, in *AML1*, the annotated UACUAAC sequence, which matches the branch site consensus, is not used; instead, the overlapping AACUAAC sequence, which contains one mismatch, is used. Together, these observations highlight the limitations of bioinformatics and underscore the value of LIT-seq in empirically defining branch sites.

The third class of novel splicing events is not associated with annotated introns but rather with genes previously defined as “intronless.” We identified 12 such splicing events in ten protein-coding genes and two noncoding RNA genes (Fig. 6; Supplemental Fig. S8; Supplemental Table S4). Notably, several of the novel introns mapped to untranslated regions (UTRs), where introns are harder to detect bioinformatically. The intron in *GLE2* and *TEX1* mapped to the

5' UTR (Fig. 6A,B; data not shown), while the introns in *PTC3*, *MSL5*, *PRO3*, *QDR2*, and *YDR541C* mapped to the 3' UTR (Fig. 6C–E, Supplemental Fig. S8F–J; data not shown), a region in which introns are rarely documented in this organism. The novel introns in *OPT1*, *SFB2*, and *DIC1*, by contrast, mapped to the coding regions (data not shown). Notably, splicing in *OPT1* and *SFB2* has the potential to shift the reading frame. We were able to confirm these novel introns, plus those in the two ncRNA genes, by lariat RT-PCR (Fig. 6; Supplemental Fig. S8; data not shown). By mRNA RT-PCR, each of these newly identified splicing isoforms were expressed as minor species (Fig. 6; data not shown). Overall, our identification and validation of these three classes of novel splicing events establishes the utility of LIT-seq in uncovering novel, even rare, splicing events

FIGURE 6. (A–G) LIT-seq revealed novel introns, as illustrated by the unannotated introns found in the 5' UTR of *GLE2* (A), the 3' UTR of *PTC3* (C), and in the noncoding RNA that is antisense to *RGT2* (F). Gene annotations and read numbers are displayed as in Figure 2C. In F, chevrons show the direction of the *RGT2* sense transcript. (B,D,G) Verification of the novel introns, via the intron product. Lariat RT-PCR and Sanger sequencing of the branch junctions is illustrated as in Figure 5B. (E) Verification of the novel intron in *PTC3* via the mRNA product, as in Figure 5C, D. The asterisk in D marks a nonspecific band resulting from RT-PCR.

not predicted bioinformatically or documented by RNA-seq of the transcriptome or by alternative methods for interrogating the intronome.

DISCUSSION

Here, we establish a novel method, LIT-seq, for investigating pre-mRNA splicing genome-wide (Fig. 1). This method analyzes splicing from the perspective of the intron, rather than the mRNA, and focuses deep sequencing on the termini of lariat introns. In a test of LIT-seq on the intronome of the model organism budding yeast, we detected over 75% of the annotated introns (Fig. 2; Supplemental Figs. S4, S7). In addition, we were able to empirically and efficiently define 5' splice sites and, more importantly, branch sites, which have generally only been predicted genome wide (Fig 3; Supplemental Figs. S4, S7). Further, LIT-seq revealed A tailing at the 3' termini of excised introns, implicating these lariat introns as substrates for both the TRAMP complex and the exosome (Fig. 4). Importantly, we demonstrated that LIT-seq can be applied to wild-type cells, implying a general utility of LIT-seq (Supplemental Fig. S7). Finally, LIT-seq revealed novel splicing events, including alternative 5' splice sites and branch sites as well as entirely novel introns (Figs. 5, 6; Supplemental Fig. S8; Supplemental Table S4). Overall, our findings demonstrate that LIT-seq, with its focus on the termini of introns, provides an efficient, simple, and direct approach to interrogate splicing events and signals.

Comparison of LIT-seq with other sequencing approaches for analyzing splicing

Our study has demonstrated several benefits of investigating pre-mRNA splicing by focusing RNA-seq on the termini of lariats. First, LIT-seq reported on splicing efficiently—in RNase R-treated libraries, 34%–84% of total mapped reads mapped to introns (Supplemental Table S1)—an efficiency comparable to the highest efficiency of previous intron probing approaches (e.g., Awan et al. 2013), but significantly superior to typical mRNA sequencing (e.g., ~0.3%; Nagalakshmi et al. 2008; Kawashima et al. 2014). Second, LIT-seq efficiently revealed splicing sites—on average 60% of mapped reads revealed 5' splice sites and branch sites (Supplemental Table S1). This efficiency reflects a significant improvement over previous intronome studies, which detected splicing sites with an efficiency of <3% in total mapped reads (Awan et al. 2013; Bitton et al. 2014; Mercer et al. 2015; Stepankiw et al. 2015). Third, while LIT-seq may bias against long introns because of the current requirement to reverse transcribe the full length of the intron, LIT-seq efficiently detected short introns, including the shortest intron in budding yeast (*MATa1*, 52 nt). In contrast, other intron probing approaches that use random priming strategies have detected shorter introns less efficiently than longer ones (Awan et al. 2013; Bitton et al. 2014).

A limitation of LIT-seq approaches as implemented here is the loss of linkage between 5' and 3' reads due to the utilization of MmEI. Linkage can be inferred for genes with a single 5' splice site and branch point, alternative 5' splice sites and a single branch site (e.g., Fig. 5A), or a single 5' splice site and alternative branch points; however, linkage would be ambiguous for genes with multiple 5' splice sites and branch sites. In other intron approaches to splice site identification, the branched structure itself is exploited to link the 5' splice site and branch site, through RT reads that traverse the branch structure (Awan et al. 2013; Bitton et al. 2014; Mercer et al. 2015). LIT-seq could be adapted to maintain linkage between the 5' and 3' termini of an individual lariat. For example, the linkage between the 5' and 3' termini could be maintained by circularization of debranched lariats to join the ends (cf. Pelechano et al. 2013); alternatively, where intron length permits, paired-end sequencing could be utilized (Fullwood et al. 2009). Further adaptations, however, would be required to enable the detection of *trans*-splicing and back-splicing events (Spieth et al. 1993; Barrett et al. 2015).

Novel splicing events: splicing noise or hints of regulation?

In this study of budding yeast introns, we uncovered novel splicing events in 45 genes in total. The corresponding novel mRNA isoforms, of those we validated, were generally of low abundance, compared with the major species. The low abundance could in part result from inefficient splicing. Consistent with this possibility, these novel events, with four exceptions, are associated with weak splicing signals; the 5' splice site and branchpoint sequence of the novel introns differed from the consensus at one to two positions, as permitted by our criteria for selecting novel introns (Supplemental Table S4; cf. Kawashima et al. 2014). Indeed, these novel splicing events revealed two 5' splice site sequences and nine branch site sequences that have not been reported previously in budding yeast (Supplemental Table S4), indicating a broader specificity of the spliceosome than previously suspected. Because we generally could not detect conservation of these splice sites in closely related yeast species, we suspect that many of these inefficient splicing events reflect splicing noise that stems from off-target binding of the splicing machinery, binding that has no functional role (Melamud and Moulton 2009; Pickrell et al. 2010; Hon et al. 2013; Kawashima et al. 2014). However, we cannot rule out the possibility that these splicing events reflect *S. cerevisiae*-specific regulatory mechanisms, in which splicing becomes efficient under conditions that differ from those utilized in this study. Even if these splicing events reflect noise, they could reflect the tolerance of a pool of alternative isoforms that could be exploited during evolution.

Notably, we did observe conservation overlapping introns that reside in the 3' UTR of *MSL5* (Supplemental Fig. S8F), a gene encoding a splicing factor that functions in recognizing

the branchpoint sequence in early stages of spliceosome assembly (Berglund et al. 1997). Intriguingly, while the 5' splice site is optimal, the two alternative branch sites identified in *MSL5* are suboptimal, raising the possibility of a negative feedback loop in which high levels of Msl5p induce splicing in the 3' UTR to down-regulate expression, either through decreased translation or increased turnover. Further studies will be needed to distinguish between noise and function for this and the other inefficient splicing events.

Insight into lariat intron processing

In our study, the 3' termini of introns were appended with one or more nontemplated adenosines (Fig. 4). Notably, short, nontemplated oligo(A) tails are added by the TRAMP complex to a variety of RNAs for targeting by the exosome for turnover (LaCava et al. 2005), and introns have been reported to immunoprecipitate with the TRAMP complex (San Paolo et al. 2009). Thus, nontemplated adenosines at the 3' termini of lariat introns may target these introns to the exosome. Oligo(A) tailing of lariat introns just downstream from the branch site in a *dbp1Δ* strain likely results from the resistance of the branch to exonucleolytic digestion and likely reflects an effort to overcome the block through the recruitment of the exosome, as has been observed for stable but misprocessed RNAs (Houseley et al. 2006). However, the covalent nature of the branch would present a persistent barrier to the exonuclease activities of the exosome, resulting in a futile cycle of oligo(A) addition and degradation, potentially explaining why the dominant oligo(A) tails at the 3' termini of lariat introns at steady state extended only one nucleotide, as compared with other exosome substrates for which oligo(A) tails are longer and recruitment of the exosome to the oligo(A) tail would generally lead to processive degradation of the RNA (Jia et al. 2011; Wlotzka et al. 2011). Nevertheless, the exosome embodies not only exonucleolytic activities but also an endonucleolytic activity (Lebreton et al. 2008), which could potentially bypass the branched structure and account for the endonucleolytic cleavage of lariats stabilized in a *dbp1Δ* strain, cleavage that has been observed previously (Ooi et al. 1998; Garrey et al. 2014) and likely accounts for the population of 3' reads that mapped to the intron body, upstream of the branch (Fig. 3D).

Future applications of LIT-seq

Our current LIT-seq approaches restrict an analysis to the 3' termini of lariats that correspond to the branch point. Although not tested in this study, LIT-seq has the potential, through modifications to the protocol, to extend to lariat 3' termini that correspond to distinct features. First, LIT-seq, unlike other intron probing approaches, can be adapted to define 3' splice sites, by targeting the 3' termini of excised but intact introns. By defining 5' splice sites, branch sites, and

3' splice sites, LIT-seq would not only allow the empirical annotation of introns but would also have the potential to reveal recursive splicing events, in which the spliceosome excises large introns by sequentially removing nested introns (Hatton et al. 1998; Lopez and Séraphin 2000; Sibley et al. 2015; Duff et al. 2015). While the fully processed mRNA product cannot distinguish between intron excision in a single step from excision in multiple recursive steps, the intron products resulting from these two alternative pathways do distinguish between these mechanisms, because the lariats effectively record each splicing event.

Second, LIT-seq has the unique potential to reveal the lariat intermediate through the detection of a lariat terminus that maps downstream from a 3' splice site. Significantly, whereas a 3' terminus that maps to the polyadenylation site would indicate post-transcriptional splicing, a 3' terminus that maps to the 3' exon would indicate cotranscriptional splicing. Whereas most evidence suggests regulatory factors impact splicing early in assembly, previous studies have demonstrated regulation late in the pathway—after the first catalytic step (Bouck et al. 1995; Lallena et al. 2002; Volanakis et al. 2013); however, whether late regulatory events are more prevalent than currently appreciated remains to be determined. LIT-seq, through the relative quantification of lariat intermediates, could reveal such regulation.

Third, in addition to revealing reactive sites in splicing, LIT-seq has the potential to reveal rapid changes in splicing. Because the steady-state levels of lariats are low and lariats turn over quickly (Windhager et al. 2012), the levels of lariats are sensitive to changes in splicing—as compared with mRNAs, which exhibit higher steady-state levels and greater stability. Importantly, we show that LIT-seq is reproducible (Supplemental Fig. S5), and can efficiently capture introns (Supplemental Fig. S3B,C), implying that LIT-seq can be applied in the future for quantitative comparisons across conditions. Further, while the absolute levels of intronic reads detected by LIT-seq may not faithfully reflect the absolute levels of splicing (Supplemental Fig. S6; cf. Bitton et al. 2014), we expect that the relative levels of intronic reads under different conditions will reflect quantitatively changes in splicing, because any bias will cancel out. Indeed, preliminary data indicate that LIT-seq can detect a decrease of lariat levels after heat shock as effectively as a direct analysis of intron levels (data not shown).

Finally, LIT-seq can be adapted to include alternative approaches for intron enrichment, thereby allowing broad applications across organisms. While we and others have enriched for introns by exploiting the accumulation of introns in a *dbp1Δ* strain (Juneau et al. 2007; Zhang et al. 2007; Awan et al. 2013; Bitton et al. 2014; Stepankiw et al. 2015), this approach may be undesirable or impractical in some organisms. To enable LIT-seq in cells or organisms expressing debranchase, we have instead enriched for introns through immunoprecipitation of the spliceosome; by coupling LIT-seq to the coimmunoprecipitation (co-IP) of introns with

the spliceosome, we have succeeded in mapping 5' splices, branch sites, and novel introns in a wild-type strain (Supplemental Fig. S7). Because endogenous spliceosomes have been immunoprecipitated previously from other organisms, such as human cells (e.g., Girard et al. 2012), and transcripts bound to endogenous spliceosomes have already been identified (Volanakis et al. 2013; Chen et al. 2014; Nojima et al. 2015), we anticipate that LIT-seq can be applied broadly to other organisms to efficiently, specifically, and directly monitor splicing site choice genome-wide and with single-nucleotide resolution.

MATERIALS AND METHODS

Strains and growth conditions

To apply LIT-seq to total RNA from *dbp1Δ* cells, the wild-type yeast strain BY4741 (*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*) and the isogenic mutant strain *dbp1Δ* (GE Healthcare, clone ID 4999, genotype: *MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 dbp1Δ*) were grown at 30°C in rich medium (YPDA, yeast peptone dextrose adenine) and harvested in mid-log phase (OD₆₀₀ between 0.4 and 0.8). To apply LIT-seq to RNA extracted from endogenous spliceosomes immunoprecipitated via Prp16, the yeast strain yS107 (Burgess and Guthrie 1993) was transformed with the *LEU2*-marked *ACT1-CUP1* reporter pGAC24 (Lesser and Guthrie 1993) and grown at 30°C in -Leu yeast minimal media and harvested in mid-log phase.

Oligos

Oligos for constructing LIT-seq libraries and confirming novel splicing events were purchased from IDT and are listed in Supplemental Table S5.

Whole-cell RNA extraction

Whole-cell RNA was extracted by phenol/chloroform/isoamyl alcohol (25:24:1) at 65°C, as described (Stevens and Abelson 2002).

Lariat enrichment through spliceosome immunoprecipitation via Prp16

Whole-cell extracts were prepared using a liquid nitrogen method as described (Mayas et al. 2006). Then, 2 mL of whole-cell extract was mixed with 5 mL IPP150 (10 mL Tris [pH8.0], 150 mL NaCl, 0.1% NP-40, 2.5 mM MnCl₂) along with a protease inhibitor cocktail tablet (Roche) and DNase I (100 units, Life Technologies) and incubated on ice for 10 min. After spinning at 17,000 rpm in a Sorvall SS34 rotor for 20 min, the supernatant was incubated with 100 μg of Prp16 antibodies (purified from serum, from the Guthrie lab) at 4°C for 2 h. Next, 200 μL of a protein A sepharose bead slurry (1:1; Life Technologies) that was prewashed with IPP150 was added to the whole cell extract and incubated on a nutator at 4°C for another 1 h. The beads were washed five times with 10 volumes of IPP150, and the bound fractions were eluted with 50 mM glycine (pH 2.6); the elution was then phenol/chloroform extracted and ethanol precipitated to yield "Prp16-IP" RNA.

Lariat enrichment by exonuclease and phosphatase treatment

Whole-cell RNA (10 μg) or Prp16-IP RNA was incubated with 10 units of XRN-1 in buffer 3 (NEB) at 37°C for 1 h, followed by incubation with 20 units of CIP (NEB) in the same buffer at 37°C for another hour. Optionally, the XRN-1/CIP treated RNA was phenol/chloroform purified and further incubated with 20 units of RNase R in 1× RNase R buffer (Epicentre) at 37°C for 1 h.

Northern analysis of intronic RNA

Northern analysis was performed as described (Small et al. 2006), except that gene-specific probes against the intron body were used (Supplemental Table S5).

LIT-seq library construction

To construct LIT-seq libraries from *dbp1Δ* whole-cell RNA without RNase R treatment, we customized an adapter ligation strategy based on the small RNA seq protocol (version 1.5, Illumina). The preadenylated 3' oligo rDQ2, encoding the flow cell adapter (Supplemental Table S5), was ligated to the 3' terminus of RNA by incubation with 100 units of T4 RNA ligase 2 (truncated, K277Q) in T4 RNA ligase buffer (NEB) with 15% PEG-8000 at 20°C for 2 h. For the experimental rDbr1+ libraries, lariats were debranched to expose a free 5' phosphate by incubation with 25 ng rDbr1p in buffer (50 mM Tris-HCl [pH 7.0], 4 mM MnCl₂, 2.5 mM DTT, 25 mM NaCl, 0.01% Triton X-100, 0.1 mM EDTA, 0.15% glycerol) at 30°C for 15 min (Khalid et al. 2005); rDbr1 was prepared as described (Khalid et al. 2005). For the control libraries, rDbr1p treatment was omitted. Next, the oligo rDQ3, which encodes the sequencing adapter (Supplemental Table S5), was ligated to the 5' terminus of RNA, having a free 5' phosphate, by incubation with 10 units of T4 RNA ligase 1 (NEB) in T4 RNA ligase buffer and 10% DMSO at 37°C for 1 h. The ligation products were reverse transcribed by superscript III (Life Technologies) in 1× first strand buffer (Life Technologies) at 50°C for 1 h with primers oDQ56 or oDQ65B; Supplemental Table S5) that anneal to the 3' adapter and encode barcodes to distinguish experimental and control libraries and to distinguish the 3' library from the 5' library. Then, cDNA products were amplified by Phusion polymerase (Fisher) for 18 cycles using the forward primer oDQ49 and the reverse primer oDQ182 using following the parameters: 94°C, 10 sec; 60°C, 10 sec; 72°C, 2 min. To remove adapter-adapter dimers, the PCR products were run on a 6% native polyacrylamide gel and the DNA fragments that range from 110 bp, corresponding to an insert of 30 bp, up to the well, corresponding to inserts of at least 2 kb, were cut from the gel and electro-eluted. To capture the termini of lariats, the gel-purified PCR products were digested with MmeI (NEB) in buffer 2 plus 50 μM S-adenosylmethionine at 37°C for 30 min, yielding DNA fragments having a 2-nt 3' overhang and lengths of 63–64 bp or 49–50 bp for the 5' or 3' termini, respectively. The DNA fragments were then size selected on a 6% native polyacrylamide gel, cut from the gel, and electro-eluted. Next, for the 5' libraries, the flow cell adapter and library barcodes were introduced by ligating the preannealed DNA duplexes A or B, each with randomized, 2-nt 3' overhangs (Supplemental Table S5), to the MmeI-digested 5'(rDbr1+) or 5'(rDbr1-) libraries,

respectively, by T4 DNA ligase (NEB) at 16°C overnight. For the 3' libraries, the sequencing adapter was introduced by ligation of the Mme I-digested 3'(rDbr1+) or 3'(rDbr1-) libraries to a common DNA duplex C (Supplemental Table S5), also with a randomized, 2-nt 3' overhang. The ligation products were selectively amplified using oDQ49 and oDQ182 by Phusion polymerase following the same conditions as above. The PCR products of the expected size (90–91 bp for 5' libraries and 102–103 bp for 3' libraries) were separated on a 6% polyacrylamide gel, cut out, and electro-eluted. The samples were dissolved in H₂O before submitting for deep sequencing. Note: Mme I digestion generates a 20–21-bp tag; in the 3' libraries, the entire tag derived from the yeast transcriptome, because the Mme I site lies at the very 5' end of the flow cell adapter; however, in the 5' libraries, only 16–17 bp of the tag derived from the yeast transcriptome—the remaining 4 bp derived from the sequencing adapter, because in the 5' library, an Mme I site is coincidentally encoded in the sequencing adapter region and lies 4-nt upstream of the 3' end of the 5' sequencing adapter (Supplemental Table S5).

To construct LIT-seq libraries from *dbp1Δ* whole-cell RNA with RNase R treatment, we followed the same procedure with the following exceptions. The RT primers oDQ515, oDQ516, oDQ517, and oDQ518 (Supplemental Table S5) were used respectively to introduce a random 11 nt barcode (Shiroguchi et al. 2012) for the control library, named 3'(rDbr1-, RNase R+), and for the replicates 1, 2, and 3 of the experimental libraries, named 3'(rDbr1+, RNase R+), respectively, as described in the text; additionally, for the 5' libraries, the flow cell adapters were introduced by ligation to DNA duplexes D, E, F, and G (Supplemental Table S5), respectively, which encoded unique barcodes.

To construct LIT-seq libraries from Prp16-IP RNA (with RNase R treatment), we again followed the same procedure except that the RT primers oDQ157 and oDQ158 (Supplemental Table S5) were used, respectively, for the control library, named Prp16-3'(rDbr1-, RNase R+), and the experimental library, named Prp16-3'(rDbr1+, RNase R+), as described in the text; additionally, for the 5' libraries, the flow cell adapters were introduced by ligation to DNA duplexes H and I (Supplemental Table S5), respectively, which encoded unique barcodes.

Sequencing

After determining the relative concentration of each library sample, all barcoded samples were mixed in equal molar amounts and sequenced on a single lane of either an Illumina HiSeq 2000 (Beijing Genomic Institute, Shenzhen, China) or an Illumina HiSeq 2500 (Functional Genomics Facility, University of Chicago).

Bioinformatics

For LIT-seq libraries derived from *dbp1Δ* whole-cell RNA untreated with RNase R, sequencing reads were deconvoluted, using barcodes, into individual libraries, after removing low quality reads with a quality score <15. Then after adapter sequences were removed, reads with a length of 20–21 bp, for the 5' libraries, or 16–17 bp for the 3' libraries, were selected for subsequent analysis. Using the Novoalign program (v3.02.05, <http://www.novocraft.com/>), we mapped these reads to the *S. cerevisiae* genome (S288C reference genome_R64-1-1_20110203, downloaded from www.yeastgenome.org). Reads mapped to the genome with a perfect match

at each position were selected for subsequent analysis. Additionally, for the 3' library, due to adenosine tailing, unmapped reads terminating with an adenosine at the 3' end were trimmed by 1 nt and realigned to the genome for unique and perfect mapping; mapped reads were collected, whereas unmapped reads were subjected to another round of trimming and mapping, for up to seven rounds; additionally, the number of adenines requiring removal to allow mapping was recorded. Note that adenosine tails longer than seven nucleotides do exist; however, additional adenosine trimming after seven rounds renders reads shorter than 13 nt, which may not be uniquely mapped; therefore, we ignored further trimming. We defined the last genomically encoded nucleotide as the 3' termini for the A-tailed reads, after trimming and successful mapping. To assess the specificity of adenosine in tails, unmapped reads were tested in parallel for mapping after iteratively trimming U, C, or G from their 3' ends.

For *dbp1Δ* whole-cell RNA treated with RNase R and Prp16-IP RNA treated with RNase R, LIT-seq libraries were analyzed similarly, but with several modifications. First, after removal of the adapters, we also removed the two most 3' nucleotides in the 5' library and the two most 5' nucleotides in the 3' library, because the downstream analysis revealed that these nucleotides, which corresponded to the MmeI-derived overhang sequences were frequently mutated, likely due to the tolerance of mismatches in the ligation of a duplex by T4 DNA ligase (Alexander et al. 2003); for reasons that are not clear, such mutagenesis was not a significant problem for the RNase R-untreated libraries. After removal of these two nucleotides, reads with 14–15 nt for the 5' libraries and 18–19 nt for the 3' libraries were retained as tags. Second, to remove reads resulting from PCR amplification in the 3' libraries derived from *dbp1Δ* whole-cell RNA, before removal of the adapters, we also eliminated reads having identical, replicated barcodes derived from the randomized, 11-nt barcode incorporated into these libraries (Shiroguchi et al. 2012).

To analyze the specificity and efficiency of LIT-seq in detecting introns, reads were identified that overlapped with 298 annotated introns, defined by coordinates downloaded from the Ares lab yeast intron database, <http://intron.ucsc.edu/yeast4.3>. Detected introns were defined as those with at least 2 reads mapped. In the meta-analyses of read distribution within normalized intronic windows, the boundaries of the windows (5'SS, 3'SS and branch point) were based on the annotations in the Ares database. Note that we found that in the RNase R-untreated library, but not other libraries, reads from the intron of *PMI40* gene dominated the total intronic reads, accounting for ~50% of the reads; because this preponderance of reads was not observed in other libraries, we inferred that this excess was an artifact; therefore, to facilitate a balanced meta-analysis of the remaining introns in Figures 3, 4, Supplemental Figure S2 and Supplemental Table S3, we first filtered reads from the *PMI40* intron. For similar reasons, in Supplemental Figure S7, we removed reads from the plasmid-encoded *ACT1-CUP1* gene and reads from the intron of *QCR10* gene, which together accounted for >90% of intronic reads; note, however, that we have not ruled out that *QCR10* was enriched due to selection of endogenous spliceosomes at the Prp16 stage.

Sequence motifs were identified in the sequences of uniquely mapped 5' and 3' reads using the Multiple EM for motif Elicitation (MEME) tool version 4.10.0 with the default parameters (Bailey et al. 2009). Only the top hit in each case was selected for display in Figure 3.

Conservation tracks show phastCons scores for multiple alignments of six indicated yeast genomes relative to the *S. cerevisiae* genome (track downloaded from <http://hgdownload.cse.ucsc.edu/goldenPath/sacCer3/phastCons7way>).

To analyze the correlation between LIT-seq and NET-seq, the read number for a particular intron in LIT-seq was plotted against the read number for that intron-containing gene in NET-seq or RNA-seq, scaled to RPKM (reads per kilobase per million mapped reads). For RNA-seq, reported RPKM (data set: wild-type cells, rich medium) was used (Yassour et al. 2009); for NET-seq, raw reads (accession number: SRX031059; Churchman and Weissman 2011) were mapped to *S. cerevisiae* genome and the number of reads mapped to an intron containing gene was extracted and normalized to RPKM.

Validation of novel introns

To validate novel splicing events implicated by LIT-seq, we used RT-PCR. First, whole-cell RNA from either wild-type *DBR1* or mutant *dbi1Δ* strains was reverse transcribed by superscript II reverse transcriptase at 42°C for 1 h using a mixture of random hexamers as primers. Then, novel splicing events were confirmed by two complementary strategies: (i) lariat RT-PCR, to detect the branched structure of the lariat (Gao et al. 2008); or (ii) mRNA RT-PCR, to detect the exonic junction of the mRNA. In lariat RT-PCR, cDNAs were PCR amplified using a forward primer that anneals upstream of the putative branch site and a reverse primer that anneals downstream from the putative 5' splice site (Supplemental Table S5), using Phusion polymerase with the following parameters: 94°C, 10 sec; 53°C, 10 sec; and 72°C, 15 sec for 30 cycles. The putative bands corresponding to the branched junctions were excised, eluted, and subjected to TOPO cloning (Life Technologies); sample clones were sequenced by Sanger sequencing. In mRNA RT-PCR, cDNAs were PCR amplified using a forward primer that anneals upstream of the putative 5' splice site and a reverse primer that anneals downstream from the putative 3' splice site, using Phusion polymerase with the following parameters: 94°C, 10 sec; 53°C, 10 sec; 72°C, 15 sec for 30 cycles. Note: The putative 3' splice site was inferred as the first AG downstream from the branch site and confirmed in each case by mRNA RT-PCR coupled to TOPO cloning analysis.

DATA DEPOSITION

Raw read files were deposited in the Sequence Read Archive (SRA; accession number SRP059288 for *dbi1Δ* libraries and SRP063518 for Prp16-IP libraries).

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank A. Ruthenburg, G. Chanfreau, M. Long, and members of the Staley laboratory for helpful discussions and comments on the manuscript; B. Schwer for sharing the plasmid pET16b-Dbr1 for expressing rDbr1; C. Guthrie for sharing strain yS107 and Prp16 antiserum; Q. Dai and M. Werner for reagents; and Y.L. Chan for technical assistance. This work was supported by grants from the

National Institutes of Health (NIH) (R21HG5262; R01GM62264) to J.P.S.; D.Q. was supported by an American Heart Association Postdoctoral Fellowship (14POST18890040); L.H. and J.A. were supported by an NIH grant (UL1 RR024999) to Julian Solway; and sequencing costs were defrayed by an NIH grant (UL1 TR000430) to Julian Solway.

Received June 10, 2015; accepted November 4, 2015.

REFERENCES

- Alexander RC, Johnson AK, Thorpe JA, Gevedon T, Testa SM. 2003. Canonical nucleosides can be utilized by T4 DNA ligase as universal template bases at ligation junctions. *Nucleic Acids Res* **31**: 3208–3216.
- Awan AR, Manfredo A, Pleiss JA. 2013. Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. *Proc Natl Acad Sci* **110**: 12762–12767.
- Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202–W208.
- Barrett SP, Wang PL, Salzman J. 2015. Circular RNA biogenesis can proceed through an exon-containing lariat precursor. *eLife* **4**: e07540.
- Berglund JA, Chua K, Abovich N, Reed R, Rosbash M. 1997. The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAAC. *Cell* **89**: 781–787.
- Bitton DA, Rallis C, Jeffares DC, Smith GC, Chen YY, Codlin S, Marguerat S, Bähler J. 2014. LaSSO, a strategy for genome-wide mapping of intronic lariats and branch points using RNA-seq. *Genome Res* **24**: 1169–1179.
- Bitton DA, Atkinson SR, Rallis C, Smith GC, Ellis DA, Chen Y, Malecki M, Codlin S, Cotobal C, Lemay J-F, et al. 2015. Widespread exon skipping triggers degradation by nuclear RNA surveillance in fission yeast. *Genome Res* **25**: 884–896.
- Bouck J, Fu XD, Skalka AM, Katz RA. 1995. Genetic selection for balanced retroviral splicing: novel regulation involving the second step can be mediated by transitions in the polypyrimidine tract. *Mol Cell Biol* **15**: 2663–2671.
- Bousquet-Antonelli C, Presutti C, Tollervey D. 2000. Identification of a regulated pathway for nuclear pre-mRNA turnover. *Cell* **102**: 765–775.
- Burgess SM, Guthrie C. 1993. A mechanism to enhance mRNA splicing fidelity: the RNA-dependent ATPase Prp16 governs usage of a discard pathway for aberrant lariat intermediates. *Cell* **73**: 1377–1391.
- Chapman KB, Boeke JD. 1991. Isolation and characterization of the gene encoding yeast debranching enzyme. *Cell* **65**: 483–492.
- Chen W, Shulha HP, Ashar-Patel A, Yan J, Green KM, Query CC, Rhind N, Weng Z, Moore MJ. 2014. Endogenous U2-U5-U6 snRNA complexes in *S. pombe* are intron lariat spliceosomes. *RNA* **20**: 308–320.
- Churchman LS, Weissman JS. 2011. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**: 368–373.
- Clark TA, Sugnet CW, Ares M Jr. 2002. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* **296**: 907–910.
- Danin-Kreiselman M, Lee CY, Chanfreau G. 2003. RNase III-mediated degradation of unspliced pre-mRNAs and lariat introns. *Mol Cell* **11**: 1279–1289.
- Davis CA, Grate L, Spingola M, Ares M Jr. 2000. Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. *Nucleic Acids Res* **28**: 1700–1706.
- Deutsch M, Long M. 1999. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res* **27**: 3219–3228.
- Duff MO, Olson S, Wei X, Garrett SC, Osman A, Bolisetty M, Plocik A, Celniker SE, Graveley BR. 2015. Genome-wide identification of zero nucleotide recursive splicing in *Drosophila*. *Nature* **521**: 376–379.

- Egecioglu DE, Kawashima TR, Chanfreau GF. 2012. Quality control of MATa1 splicing and exon skipping by nuclear RNA degradation. *Nucleic Acids Res* **40**: 1787–1796.
- Fullwood MJ, Wei CL, Liu ET, Ruan Y. 2009. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res* **19**: 521–532.
- Gao K, Masuda A, Matsuura T, Ohno K. 2008. Human branch point consensus sequence is yUAnY. *Nucleic Acids Res* **36**: 2257–2267.
- Garrey SM, Katolik A, Prekeris M, Li X, York K, Bernards S, Fields S, Zhao R, Damha MJ, Hesselberth JR. 2014. A homolog of lariat-debranching enzyme modulates turnover of branched RNA. *RNA* **20**: 1337–1348.
- Girard C, Will CL, Peng J, Makarov EM, Kastner B, Lemm I, Urlaub H, Hartmuth K, Lührmann R. 2012. Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion. *Nat Commun* **3**: 994.
- Harigaya Y, Parker R. 2012. Global analysis of mRNA decay intermediates in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci* **109**: 11764–11769.
- Hatton AR, Subramaniam V, Lopez AJ. 1998. Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by reslicing at exon-exon junctions. *Mol Cell* **2**: 787–796.
- Hesselberth JR. 2014. Lives that introns lead after splicing. *Wiley Interdiscip Rev RNA* **4**: 677–691.
- Hon CC, Weber C, Sismeiro O, Proux C, Koutero M, Deloger M, Das S, Agrahari M, Dillies MA, Jagla B, et al. 2013. Quantification of stochastic noise of splicing and polyadenylation in *Entamoeba histolytica*. *Nucleic Acids Res* **41**: 1936–1952.
- Houseley J, LaCava J, Tollervey D. 2006. RNA-quality control by the exosome. *Nat Rev Mol Cell Biol* **7**: 529–539.
- Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, Marzluff WF, Sharpless NE. 2013. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19**: 141–157.
- Jia H, Wang X, Liu F, Guenther U-P, Srinivasan S, Anderson JT, Jankowsky E. 2011. The RNA helicase Mtr4p modulates polyadenylation in the TRAMP complex. *Cell* **145**: 890–901.
- Juneau K, Palm C, Miranda M, Davis RW. 2007. High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing. *Proc Natl Acad Sci* **104**: 1522–1527.
- Kalsotra A, Cooper TA. 2011. Functional consequences of developmentally regulated alternative splicing. *Nat Rev Genet* **12**: 715–729.
- Kawashima T, Douglass S, Gabunilas J, Pellegrini M, Chanfreau GF. 2014. Widespread use of non-productive alternative splice sites in *Saccharomyces cerevisiae*. *PLoS Genet* **10**: e1004249.
- Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**: 241–254.
- Khalid MF, Damha MJ, Shuman S, Schwer B. 2005. Structure-function analysis of yeast RNA debranching enzyme (Dbr1), a manganese-dependent phosphodiesterase. *Nucleic Acids Res* **33**: 6349–6360.
- LaCava J, Houseley J, Saveanu C, Petfalski E, Thompson E, Jacquier A, Tollervey D. 2005. RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. *Cell* **121**: 713–724.
- Lallena MJ, Chalmers KJ, Llamazares S, Lamond AI, Válcárcel J. 2002. Splicing regulation at the second catalytic step by Sex-lethal involves 3' splice site recognition by SPF45. *Cell* **109**: 285–296.
- Lebreton A, Tomecki R, Dziembowski A, Séraphin B. 2008. Endonucleolytic RNA cleavage by a eukaryotic exosome. *Nature* **456**: 993–996.
- Lesser CF, Guthrie C. 1993. Mutational analysis of pre-mRNA splicing in *Saccharomyces cerevisiae* using a sensitive new reporter gene, CUP1. *Genetics* **133**: 851–863.
- Lopez PJ, Séraphin B. 2000. Uncoupling yeast intron recognition from transcription with recursive splicing. *EMBO Rep* **1**: 334–339.
- Martin JA, Wang Z. 2011. Next-generation transcriptome assembly. *Nat Rev Genet* **12**: 671–682.
- Mayas RM, Maita H, Staley JP. 2006. Exon ligation is proofread by the DEXD/H-box ATPase Prp22p. *Nat Struct Mol Biol* **13**: 482–490.
- McGlinchy NJ, Smith CW. 2008. Alternative splicing resulting in non-sense-mediated mRNA decay: what is the meaning of nonsense? *Trends Biochem Sci* **33**: 385–393.
- Melamud E, Moult J. 2009. Stochastic noise in splicing machinery. *Nucleic Acids Res* **37**: 4873–4886.
- Mercer TR, Clark MB, Andersen SB, Brunck ME, Haerty W, Crawford J, Taft RJ, Nielsen LK, Dinger ME, Mattick JS. 2015. Genome-wide discovery of human splicing branchpoints. *Genome Res* **25**: 290–303.
- Miura F, Kawaguchi N, Sese J, Toyoda A, Hattori M, Morishita S, Ito T. 2006. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc Natl Acad Sci* **103**: 17846–17851.
- Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344–1349.
- Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**: 457–463.
- Nojima T, Gomes T, Grosso ARF, Kimura H, Dye MJ, Dhir S, Carmo-Fonseca M, Proudfoot NJ. 2015. Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* **161**: 526–540.
- Ooi SL, Samarsky DA, Fournier MJ, Boeke JD. 1998. Intronic snoRNA biosynthesis in *Saccharomyces cerevisiae* depends on the lariat-debranching enzyme: intron length effects and activity of a precursor snoRNA. *RNA* **4**: 1096–1110.
- Pelechano V, Wei W, Steinmetz LM. 2013. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**: 127–131.
- Pickrell JK, Pai AA, Gilad Y, Pritchard JK. 2010. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet* **6**: e1001236.
- Pleiss JA, Whitworth GB, Bergkessel M, Guthrie C. 2007a. Rapid, transcript-specific changes in splicing in response to environmental stress. *Mol Cell* **27**: 928–937.
- Pleiss JA, Whitworth GB, Bergkessel M, Guthrie C. 2007b. Transcript specificity in yeast pre-mRNA splicing revealed by mutations in core spliceosomal components. *PLoS Biol* **5**: e90.
- Roy SW, Gilbert W. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet* **7**: 211–221.
- San Paolo S, Vanacova S, Schenk L, Scherrer T, Blank D, Keller W, Gerber AP. 2009. Distinct roles of non-canonical poly(A) polymerases in RNA metabolism. *PLoS Genet* **5**: e1000555.
- Sayani S, Janis M, Lee CY, Toesca I, Chanfreau GF. 2008. Widespread impact of nonsense-mediated mRNA decay on the yeast intronome. *Mol Cell* **31**: 360–370.
- Schneider C, Kudla G, Wlotzka W, Tuck A, Tollervey D. 2012. Transcriptome-wide analysis of exosome targets. *Mol Cell* **48**: 422–433.
- Schwer B, Guthrie C. 1991. PRP16 is an RNA-dependent ATPase that interacts transiently with the spliceosome. *Nature* **349**: 494–499.
- Shiroguchi K, Jia TZ, Sims PA, Xie XS. 2012. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc Natl Acad Sci* **109**: 1347–1352.
- Sibley CR, Emmett W, Blazquez L, Faro A, Haberman N, Briese M, Trabzuni D, Ryten M, Weale ME, Hardy J, et al. 2015. Recursive splicing in long vertebrate genes. *Nature* **521**: 371–375.
- Small EC, Leggett SR, Winans AA, Staley JP. 2006. The EF-G-like GTPase Snu114p regulates spliceosome dynamics mediated by Brr2p, a DEXD/H box ATPase. *Mol Cell* **23**: 389–399.
- Spith J, Brooke G, Kuersten S, Lea K, Blumenthal T. 1993. Operons in *C. elegans*: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. *Cell* **73**: 521–532.
- Spingola M, Grate L, Haussler D, Ares M Jr. 1999. Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA* **5**: 221–234.

- Staley JP, Guthrie C. 1998. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* **92**: 315–326.
- Stepankiw N, Raghavan M, Fogarty EA, Grimson A, Pleiss JA. 2015. Widespread alternative and aberrant splicing revealed by lariat sequencing. *Nucleic Acids Res* **43**: 8488–8501.
- Stevens SW, Abelson J. 2002. Yeast pre-mRNA splicing: methods, mechanisms, and machinery. *Methods Enzymol* **351**: 200–220.
- Suzuki H, Zuo Y, Wang J, Zhang MQ, Malhotra A, Mayeda A. 2006. Characterization of RNase R-digested cellular RNA source that consists of lariat and circular RNAs from pre-mRNA splicing. *Nucleic Acids Res* **34**: e63.
- Taggart AJ, DeSimone AM, Shih JS, Filloux ME, Fairbrother WG. 2012. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nat Struct Mol Biol* **19**: 719–721.
- Talhouarne GJ, Gall JG. 2014. Lariat intronic RNAs in the cytoplasm of *Xenopus tropicalis* oocytes. *RNA* **20**: 1476–1487.
- Volanakis A, Passoni M, Hector RD, Shah S, Kilchert C, Granneman S, Vasiljeva L. 2013. Spliceosome-mediated decay (SMD) regulates expression of nonintronic genes in budding yeast. *Genes Dev* **27**: 2025–2038.
- Wahl MC, Will CL, Lührmann R. 2009. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**: 701–718.
- Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–476.
- Windhager L, Bonfert T, Burger K, Ruzsics Z, Krebs S, Kaufmann S, Malterer G, L'Hernault A, Schilhabel M, Schreiber S, et al. 2012. Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Res* **22**: 2031–2042.
- Wlotzka W, Kudla G, Granneman S, Tollervey D. 2011. The nuclear RNA polymerase II surveillance system targets polymerase III transcripts. *EMBO J* **30**: 1790–1803.
- Yassour M, Kaplan T, Fraser HB, Levin JZ, Pfiffner J, Adiconis X, Schroth G, Luo S, Khrebtkova I, Gnirke A, et al. 2009. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc Natl Acad Sci* **106**: 3264–3269.
- Zhang Z, Hesselberth JR, Fields S. 2007. Genome-wide identification of spliced introns using a tiling microarray. *Genome Res* **17**: 503–509.
- Zhang Y, Zhang XO, Chen T, Xiang JF, Yin QF, Xing YH, Zhu S, Yang L, Chen LL. 2013. Circular intronic long noncoding RNAs. *Mol Cell* **51**: 792–806.
- Zhuang F, Fuchs RT, Sun Z, Zheng Y, Robb GB. 2012. Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. *Nucleic Acids Res* **40**: e54.

RNA

A PUBLICATION OF THE RNA SOCIETY

Sequencing of lariat termini in *S. cerevisiae* reveals 5' splice sites, branch points, and novel splicing events

Daoming Qin, Lei Huang, Alissa Wlodaver, et al.

RNA 2016 22: 237-253 originally published online December 8, 2015

Access the most recent version at doi:[10.1261/rna.052829.115](https://doi.org/10.1261/rna.052829.115)

Supplemental Material

<http://rnajournal.cshlp.org/content/suppl/2015/11/30/rna.052829.115.DC1>

References

This article cites 79 articles, 25 of which can be accessed free at:
<http://rnajournal.cshlp.org/content/22/2/237.full.html#ref-list-1>

Creative Commons License

This article is distributed exclusively by the RNA Society for the first 12 months after the full-issue publication date (see <http://rnajournal.cshlp.org/site/misc/terms.xhtml>). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

To subscribe to *RNA* go to:

<http://rnajournal.cshlp.org/subscriptions>