Transcriptome-wide analysis of the functional intronome using spliceosome profiling

Weijun Chen* [1,4], Jill Moore* [2,4], Hakan Ozadam [3], Hennady P. Shulha [2,4,5], Nicholas Rhind [4], Zhiping Weng [2,4] and Melissa J. Moore[1,4]

[1] RNA Therapeutics Institute, [2] Bioinformatics and Integrative Biology, [3]Program in Systems Biology, [4] Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01655

[5] Current affilitation: British Columbia Centre on Substance Use, St. Paul's Hospital, Vancouver, BC.

* These authors contributed equally to this work.

Corresponding author:

Melissa J. Moore
RNA Therapeutics Institute
University of Massachusetts Medical School, ASC 4-2049
368 Plantation Street, Worcester, MA 01605
Tel: 1- 508-856-8014
Fax: 1-508-856-1002
Email: melissa.moore@umassmed.edu

SUMMARY

Full understanding of eukaryotic transcriptomes and how they respond to different conditions requires deep knowledge of all sites of intron excision. Although RNA-Seq provides much of this information, the low abundance of many spliced transcripts (often due to their rapid cytoplasmic decay) limits the ability of RNA-Seq alone to reveal the full repertoire of spliced species. Here we present "spliceosome profiling", a strategy based on deep sequencing of RNAs copurifying with late stage spliceosomes. Spliceosome profiling allows for unambiguous mapping of intron ends to single nucleotide resolution and branchpoint identification at unprecedented depths. Our data reveal hundreds of new introns in *S. pombe* and numerous others that were previously misannotated. By providing a means to directly interrogate sites of spliceosome assembly and catalysis genome-wide, spliceosome profiling may well transform our understanding of RNA processing in the nucleus much as ribosome profiling is transforming our understanding mRNA translation in the cytoplasm.

Highlights

- **Spliceosome profiling allows for in depth analysis of the functional "intronome".**
- **Many new introns and splice sites identified by 5'-P and 3'-OH Seq.**
- **New algorithm maps branchpoints for >86% of all *S. pombe* introns.**
- **Spliceosomes footprints cover >56 intronic nucleotides.**

Introduction

Current methods for eukaryotic transcriptome annotation rely heavily on RNA-Seq data from polyA+ or rRNA-depleted cellular RNA wherein intron positions are inferred from reads spanning exon junctions. Exon junction reads are also used to quantify alternative splicing patterns (Trapnell et al., 2012; Wang et al., 2008; 2009). However, due to biases inherent in all RNA-Seq library preparation methods (Hu et al., 2014; Li and Jiang, 2012) and any assumptions made during exon junction mapping (e.g., maximum intron length, number of allowed mismatches), not all intron positions can be accurately identified from exon junction reads alone (Tuerk et al., 2014). Further, because individual mRNA and lncRNA isoforms can have widely different decay rates (Clark et al., 2012; Geisberg et al., 2014; Wan et al., 2012), a steady-state snapshot of exon junction reads is unlikely to accurately reflect the flux through different alternative splicing pathways. Finally, if sequencing depth is insufficient and a spliced mRNA is highly unstable (e.g., AS-NMD and snoRNA host gene mRNAs)(McGlincy and Smith, 2008), traditional RNA-Seq may miss the exon-containing spliced product all together.

A more direct approach for mapping intron locations and measuring splicing flux would be to interrogate the introns themselves. One method is to analyze branched reads in RNA-Seq data (Awan et al., 2013; Bitton et al., 2014; Mercer et al., 2015; Taggart et al., 2012). Branches are generated during the first chemical step of splicing wherein the 2'-OH of an adenosine near the 3' end of the intron (the branchpoint; BP) attacks the 3'-5' phosphodiester at the 5' end of the intron (the 5' splice site; 5'SS) to liberate the 5' exon terminating with 3'-OH and form a 2'-5' bond between the BP and the 5'SS nucleotide. Subsequent attack of the 5' exon on the 3' end of the intron (the 3' splice site; 3'SS) during the second step of splicing frees the branched (aka lariat) intron and ligates the two exons. Subsequent debranching of the excised intron by debranching enzyme results in a linear intron beginning with a 5'-phosphate (5'-P) and ending with a 3'-OH.

Branches are incidentally captured during RNA-Seq library construction when reverse transcriptase initiates downstream of the 5'SS and traverses the 2'-5' bond to continue cDNA synthesis upstream of the BP. The low abundance of such inverted reads in traditional RNA-Seq datasets (ca. one in a million reads; Taggart et al., 2012), however, limits the utility of this approach. Although circular RNA species can be enriched by 2D electrophoresis (Awan et al., 2013) or exonucleolytic degradation of linear species (Mercer et al., 2015) these approaches work well only for relatively short introns whose lariats are less susceptible to breakage during sample workup. Thus, despite intensive bioinformatics analyses interrogating billions of RNA-Seq reads, <25% of human introns currently have mapped BPs (Mercer et al., 2015; Taggart et al., 2017). In *Schizosaccharomyces pombe*, where the longest known nuclear intron is just 1126 nts, BPs have been validated in only 31% of all introns (Awan et al., 2013; Bitton et al., 2014). With validated BPs for 75% of its 298 total introns, *Saccharomyces cerevisiae* is the current leader in terms of BP identification (Qin et al., 2016).

An alternate means to enrich for introns is to purify spliceosomes, the macromolecular machines responsible for identifying splice sites and catalyzing intron excision. Employing a split TAP-tag approach, we previously reported purification, complete proteomics, and preliminary RNA-Seq characterization of an endogenous *S. pombe* spliceosomal complex containing U2, U5 and U6 snRNAs and the Prp19 complex (NTC) (Chen et al., 2014). This endogenous U2·U5·U6·NTC species contains predominantly lariat intron product, so is often referred to as intron lariat spliceosome (ILS) complex (Chen et al., 2014). It was the first complete spliceosome for which a high-resolution structure was determined by cryo-EM (Hang et al., 2015; Yan et al., 2015).

Here we present multiple deep sequencing analyses of the transcripts contained within the *S. pombe* U2·U5·U6·NTC complex. These include intron RNA-Seq libraries to investigate overall coverage across introns and exons, 5'-P and 3'-OH libraries to unambiguously map

intron ends, spliceosome footprinting libraries to precisely map sites of spliceosome occupancy, and a new BP detection algorithm. Together, these analyses enabled us to annotate hundreds of new introns, correct the annotations of numerous others, detect new mRNA isoforms and sites of recursive splicing (zero-length exons), and identify BPs in 86% of all *S. pombe* introns. Thus deep sequencing of RNAs copurifying with spliceosomes, here dubbed "spliceosome profiling", allows for "intronome" (Qin et al., 2016) identification and analysis at unprecedented depths. By combining genome-wide intronome and "exonome" (i.e., "transcriptome") data, a significantly more comprehensive understanding of post-transcriptional gene regulation can be achieved than is currently obtainable with exonome data alone.

Results

Mapping of intron ends at single nucleotide resolution

ILS complex should contain a complete record of all sites of spliceosome catalysis. To access this record, we purified *S. pombe* U2·U5·U6·NTC complex from two independent batches of cells grown to late log in rich media (Chen et al., 2014) and isolated the associated RNA. This purified RNA was treated with recombinant debranching enzyme (DBE) and then used for intron RNA-Seq, 5'-P and 3'-OH library construction (see Methods; **Figure 1A and 2A**). Sequencing on the Illumina HiSeq 4000 platform yielded 25 to 97 million reads per library, 40-50% of which mapped uniquely to the *S. pombe* genome (ASM294v2.26) (**Table S1**). Unmapped reads were predominantly short (**Figure S1A**). As expected, intron RNA-Seq reads mapping to U2, U5 and U6 snRNAs accounted for a high fraction (49%) of mapped reads, with very few reads mapping to U1 and U4 snRNAs (0.07%) (**Figure 1B**). The next most abundant read classes were those mapping to annotated introns (36%) and exons (12%), with the average intron read density being 15-fold greater than the average exon read density. Consistent with this read coverage on individual genes was predominantly across introns (**Figure 1C**). In all, 4,499 introns were represented by more >10 reads in both biological replicates, which were highly correlated (**Table S1**). Thus, our intron RNA-Seq data cover 84% of all previously annotated *S. pombe* introns, with undetected introns generally residing in poorly expressed genes (**Figure S1B**).

The 5'-P and 3'-OH libraries allowed us to precisely map intron ends. For reads mapping to protein coding (CDS) genes, 20-30% of 5'-P reads mapped exactly to an annotated 5'SS and 39-43% of 3'-OH reads mapped to an annotated 3'SS. Aggregating the reads over all annotated introns revealed very strong peaks at intron ends **(Figure 2B and S1C**). In all, 4,691 introns had statistically significant (binomial test, FDR < 0.01) 5'SS 5'-P and 3'SS 3'-OH peaks in both biological replicates. Inclusion of introns with a statistically significant peak at just one end increased this number to 5,105, or 95.2% of previously annotated *S. pombe* introns. As with the

intron RNA-Seq, introns not observed in the 5'-P and 3'-OH libraries corresponded to low abundance mRNAs (**Figure S1B**). Thus our datasets cover the vast majority of intron-containing transcripts expressed during late log growth.

Correcting intron annotations, alternative 5'SS's and recursive splicing

In addition to the 5'-P and 3'-OH peaks discussed above, many other peaks met the <0.01 FDR threshold but did not correspond to current splice site annotations. Direct examination of all such peaks with mean >50 reads per million mapped (rpm; 328 peaks total) revealed numerous misannotated or previously unannotated spliced sites that we could confirm by parallel examination of existing RNA-Seq data (Rhind et al., 2011). In all we found sixteen 5'SS and 3'SS misannotations in CDS genes, all of which change the encoded protein (**Table S2A**). Of these sixteen, nine result in a small amino acid insertion or deletion (1-28 aa's) at an internal position, whereas five were due to transcription start site (TSS) or intron misannotations at the ends of CDS regions (where genome annotation is inherently more difficult). One particularly interesting example was an intron initiating with A and terminating with C (aka, an AT-AC intron) in SPBC1861.08c (Lea1) (**Figure 2C**). Although *S. pombe* does not contain the U11/U12-dependent minor spliceosome responsible for removing AT-AC introns in metazoans, the U1/U2-dependent major spliceosome can remove introns starting with A and ending with C in the context of otherwise strong 5'SS, BP and 3'SS consensus sequences (Wu and Krainer, 1997). Another interesting example was a 3'SS misannotation in SPAC27F1.09c (prp10/sap155/SF3B1) intron 2 (**Figure S2B**). Splicing of the corrected intron results in a frameshift and premature termination event early in the CDS. RNA-Seq data suggest the existence of two mRNA isoforms, one retaining the intron and encoding the full-length protein and another where the intron has been removed and the CDS frameshifted. SF3B1 is an essential spliceosomal protein that has recently gained notoriety as an oncogene (Jenkins and

Kielkopf, 2017). In metazoans, its levels are modulated by an intron retention mechanism (Jacob and Smith, 2017). Thus it seems possible that *S. pombe* employs an analogous strategy to modulate the levels of this crucial splicing factor.

Finally, unexpected peaks in the 336 nt intron in SPNCRNA.445 (snoRNA61) revealed an internal zero-length exon (i.e., a 3'SS immediately adjacent to a 5'SS) (**Figure 2D**). We found one other zero-length exon in the 167 nt first intron of SPBC1703.10 (**Figure S2A**). In multicellular organisms, zero-length exons are sites at which recursive splicing enables intron removal piecewise (Georgomanolis et al., 2016). To our knowledge, these are the first documented examples of recursive splicing in a yeast species. Sequence conservation in other *Schizosaccharomyces* species of both SPNCRNA.445 intron 1 and SPBC1703.10 intron 1 including the splice site sequences surrounding the zero-length exons, suggests that recursive splicing of these introns serves an important biological purpose.

New introns

Our data also revealed 230 new introns. Twenty of these occur in previously unannotated transcripts (**Table S2B**) and 110 interrupt ncRNAs of unknown function (**Table S2C**). Sixteen others harbor snoRNAs, including five new introns in the SPNCRNA659 host gene, each surrounding snoRNAs 2, 3, 4, 5 and 6 (**Figure S2B**). Whereas the majority (65 of 84) of novel introns in CDS genes lie fully within UTRs (**Table S2D)**, nineteen reside either partially or fully within the annotated ORF and so alter the predicted protein sequence. One particularly interesting example is a new 543 nt intron overlapping the 5'-UTR and ORF of Alcohol dehydrogenase 4, ADH4 (SPAC5H10.06c) (**Figure 2E**). Primer extension and RT-PCR analyses indicated that ADH4 has two mRNA isoforms produced from alternative transcription start sites (TSS1 and TSS2), with TSS2 residing within the newly detected intron. PolyA+ RNA-Seq tags indicative of both mRNA isoforms were readily identifiable in previous datasets (Rhind

et al., 2011). *S. pombe* Adh4 protein was previously shown to co-purify with mitochondria (Crichton et al., 2007), and transcripts initiating at TSS2 encode a protein isoform (II) containing an N-terminal mitochondrial targeting sequence (MTS). Initiation at TSS1, however, leads to intron excision and removal of the first 43 codons containing the MTS – thus protein isoform I is presumably cytoplasmic. Primer extension analysis revealed that relative abundance of these two mRNA isoforms depends on growth conditions. During log phase growth in rich media, mRNA isoform II predominates, but upon stress (e.g., inhibition of translation or glucose deprivation), mRNA isoform I predominates. Thus, in addition to identifying a new intron in the adh4 gene, our 5'-P and 3'-OH sequencing led us to discover a stress-responsive TSS and splicing switch that likely alters the subcellular targeting of a key metabolic enzyme.

Introns with high levels of splicing intermediates

In addition to the major 3'-OH peak at the 3'SS (3'SS 3'-OH reads) indicative of a fully excised intron, 3,528 introns had a second smaller 3'-OH peak (binomial test, FDR < 0.01, **Table S3A**) at the 3' end of the upstream exon (5'SS 3'-OH reads) indicative of splicing intermediate (**Figure 2A**). Across the 5,110 introns (old plus new annotations) having at least 25 3'SS 3'-OH reads in both biological replicates, 5'SS 3'-OH reads averaged 3.3% of total (3'SS + 5'SS) 3'-OH reads for that intron, indicating that endogenous U2·U5·U6·NTC complexes contain ~3% splicing intermediates on average, validating our previous estimate (Chen et al., 2014). Nonetheless, there was much variability among introns. For example, whereas the sole intron in mak10 (SPBC1861.03) had a total of 2,717 3'SS 3'-OH reads in the two biological replicates combined, no 5'SS 3'-OH reads were detected (**Figure 3A**). At the opposite extreme, 5'SS 3'-OH reads were ~2-fold more abundant than 3'SS 3'-OH reads for rps27 (SPBC1685.10) intron 1, indicating the presence of 65% splicing intermediate.

High levels of splicing intermediate can indicate inefficient exon ligation, the means by which the 3'-end of telomerase RNA is generated from the *S. pombe* TER1 transcript (Box et al., 2008). Consistent with this, we observed 29% splicing intermediate on the TER1 intron (**Figure 3A**), ranking it within the top 2% of all 5,110 introns with regard to percent splicing intermediate (**Figure 3B**). Other ncRNA transcripts included in the top 100 introns exhibiting high (>17%) splicing intermediate included SPNCRNA.445 (snoRNA 61), SPSNORNA.01 (snoRNA 40), SPSNORNA.30 (snoRNA 62), and three transcripts antisense to CDS genes (**Table S3A**). In the future it will be of interest to determine whether any of these generate functional RNA species through inefficient exon ligation.

All other high splicing intermediate introns were in CDS genes. Gene ontology analysis of the top 100, 150 or 200 introns ranked by fraction of splicing intermediate revealed them to disproportionately interrupt genes encoding ribosomal proteins (RPs) (**Table S3B-D**). In budding yeast, inefficient intron excision from RP transcripts serves to negatively regulate ribosome production (Juneau et al., 2006; Parenteau et al., 2011); our findings thus suggest that this may also be the case in *S. pombe*.

Examination of spliceosome footprinting reads (see below) across introns with high splicing intermediate revealed that many also had a high fraction of exon junction reads (**Figure S3A,B**). A high fraction of spliced exons could indicate inefficient spliced product release. Because the spliceosome can catalyze both forward and reverse splicing (Tseng and Cheng, 2008), inefficient spliced exon release could result in reverse second step, and therefore an increase in splicing intermediates. Consistent with this idea, introns with the highest fraction of splicing intermediates had statistically higher fractions of exon junction reads all other introns within the same genes (gene-matched intron set) (**Figure 3C**).

To identify intron and exon features associated with high splicing intermediate and/or high spliced exon junction reads, we compared the 100, 150 and 200 top intron sets defined

above against similarly sized sets of (1) introns with the lowest percent splicing intermediates or (2) the gene-matched intron set (**Figure 3D**). Among all features analyzed (including 5'SS, BP and 3'SS motifs, intron length, intron order, BP to 3'SS distance, sequence conservation among Schizosaccharomyces species, essential vs nonessential genes, intron and exon nucleotide content and intron and exon folding energies) the only statistically significant differences were: (1) intron nucleotide composition, with high splicing intermediate introns having lower GC and higher A content than either low splicing intermediate introns (**Figure S3C,D**) or the gene-matched intron set (**Figure 3E**, **Figure S3E**), (2) higher GC-content in the exons flanking high splicing intermediate introns (**Figure 3E, Figure S3D**), (3) and a stronger predicted folding energy of the exon downstream high splicing intermediate introns (**Figure 3F, Figure S3F**). In metazoans, a high GC content differential between an intron and its flanking exons is a signature of both higher exonic nucleosome occupancy and initial spliceosome assembly across exons (i.e., exon definition)(Amit et al., 2012). Although spliceosome assembly across introns (i.e., intron definition) is thought to predominate in *S. pombe* (Romfo et al., 2000), our data might suggest that this is not true of all *S. pombe* introns. We also note that strong secondary structure driven by high GC-content in a downstream exon could inhibit binding of Prp22, the DExH-box RNA helicase responsible for releasing the spliced exons (Schwer, 2008).

Testing the branch-only model for 5'SS cleavage

The currently accepted view of spliceosome catalysis is that 5'SS cleavage occurs solely by attack of the BP adenosine 2'-OH on the 5'SS phosphodiester to liberate the 5' exon and generate a lariat intron containing a 2'-5' bond between the BP and 5'SS. However, group II self-splicing introns, which are both evolutionarily and structurally related to the spliceosome (Pyle, 2016), can catalyze 5'SS cleavage via either branch formation or 5'SS hydrolysis, the latter resulting in a linear intron-3' exon intermediate (Jarrell et al., 1988; Podar et al., 1998). As

discussed above, the majority of *S. pombe* and metazoan introns still lack a positively identified BP. A potential explanation for this was that some spliceosomal 5'SS cleavage events occur by hydrolysis rather than branching.

To investigate the possibility of 5'SS hydrolysis, we made 5'-P and 3'-OH libraries as above, but without DBE treatment (DBE-). About 34% of introns (1,907) behaved as expected, with a 5'SS 5'-P peak only detectable after DBE treatment (**Figure 4A** and **Figure S4A**). The other 66% (3,686), however, had a significant (albeit reduced) 5'SS 5'-P peak in at least one replicate of the DBE- libraries (**Figure 4B** and **Figure S4A**). To investigate the possibilty that some introns are subject to debranching by endogenous debranching enzyme (DBR1) prior to U2·U5·U6·NTC complex disassembly, we isolated U2·U5·U6·NTC complex from dbr1Δ cells and prepared DBE+ and DBE- 5'-P and 3-OH libraries. About 8% of introns still retained significant 5'SS 5'-P peaks across both biological replicates even in the absence of both endogenous DBR1 and exogenous DBE (**Figure 4B** and **Figure S4B**). Incubation of ^{32}P-labeled branched oligos in cell extracts confirmed that, while debranching activity was clearly apparent in the parental (WT) extract, the dbr1Δ extract had no detectable debranching activity (**Figure 4C**). Thus a fraction of introns co-purifying with U2·U5·U6·NTC complex are linear and these linear species are not produced by DBR1.

Under certain salt conditions, purified human spliceosomes are capable of hydrolyzing the 2'-5' bond at the BP (Tseng and Cheng, 2013). Thus the possibility remained that the linear species in the dbr1Δ U2·U5·U6·NTC complex were generated during our spliceosome purification protocol. To test this, we made DBE+ and DBE- 5'-P libraries from total cellular RNA extracted from both the WT and dbr1Δ strains. In the DBE+ libraries, 4,705 introns had statistically significant (FDR < 0.01) 5'SS 5'-P peaks in both WT and dbr1Δ, and 3,640 5'SS 5'-P peaks were detectable in DBE- WT cellular RNA. Notably, however, no statistically significant

5'SS peaks were found in the DBE- dbr1Δ cellular RNA (**Figure 4D**). Thus the linear introns present in purified dbr1Δ U2·U5·U6·NTC complexes were most likely due to 2'-5' bond hydrolysis occurring during spliceosome purification and not initial 5'SS cleavage by hydrolysis. We conclude that, unlike group II introns which can activate either H_2O or the BP 2'-OH as the nucleophile for the first step of splicing, the spliceosome can only initiate 5'SS cleavage by activating the BP 2'-OH. Therefore the inability to positively identify BPs in many introns is most likely due to the inherent inefficiencies in previous transcriptome-wide mapping methodologies (see below).

Spliceosome footprints

Recent cryo-EM structures of spliceosomes have an unprecedented view of how the snRNAs and core spliceosomal proteins are arranged in 3-dimensions (Bertram et al., 2017; Yan et al., 2016b; 2016a; 2015; Zhang et al., 2017). The longest intron sequences modeled into the structures to date include the first 15 nts downstream of the 5'SS and 23 nts around the BP (14nts-UACUAAC-2nts) (Yan et al., 2016a). To investigate whether this represents the full extent of strong intron-spliceosome interactions, we performed footprinting experiments akin to ribosome footprinting. Briefly, during the U2·U5·U6·NTC complex purification protocol, we included a micrococcal nuclease (MNase) digestion step; we then treated the purified RNAs with DBE and made footprint libraries using a recently published lab protocol (Heyer et al., 2015) (**Figure 5A**). Single end sequencing (82 nts) of two biological replicates yielded 20 to 32 million reads (**Table S1A**) with captured fragment length ranging from 6 to 67 nts. As expected, a high percentage of reads mapped to U2, U5 and U6 snRNAs (~70%), but of the remaining 1.5 million reads that mapped uniquely to the genome, 82% mapped to introns and exon-exon junctions.

Examination of reads on individual genes consistently revealed one peak on shorter (<60 nts) introns, a broad peak on intermediate length (~60-150 nts) introns and two or three peaks on longer (>150 nts) introns (**Figure 5B**). To analyze these footprints in greater detail, we binned according to length all 4,746 annotated introns that were 35-200 nts long (excluding snRNA introns). Roughly half (2,627) are ≤56 nts long; for these, there is a near perfect correlation (r=0.96) between the average number of reads per intron and intron length (**Figure 5C**). For introns longer than 56 nts, the average number of MNase reads per intron was largely independent of intron length. This suggests that our purified splicing complexes protect a total of 56 intronic nts. For introns ≤56 nts, this protected region encompasses the entire intron.

To cleanly separate the 5'SS and 3'SS footprints and determine their lengths, we made aggregation plots of reads mapping to the 5' and 3' ends of introns that were >200 nts (**Figure 5D-F**). At both ends, the peaks were rather broad, indicating that the length of RNA protected by the spliceosome is somewhat variable. The peak summit downstream of the 5'SS occurs at +19 nts and is half maximal between +5 and +36 nts (**Figure 5D**). At intron 3' ends, footprints center on the TACTAAC branch site consensus, extending 14 nts upstream and 11 nts downstream for a total of 25 nts; in many introns, this protected region extends all the way to the 3'SS (**Figure 5E**). Thus, the spliceosome protects ~31 nts at the 5' ends of introns and ~ 25 nts around the BP.

On long introns, we also observed a third peak centered 48-49 nts upstream of the BP with a breadth of ~38 nts (**Figure 5B,F**). *De novo* motif analysis using the 314 introns in this set revealed no specific motifs; thus its presence is not due to alternative 5'SSs or BPs. More likely this peak represents the binding site of Cwf11 (aka IBP160, Aquarius), an SF1 RNA helicase known to interact in a sequence-independent manner with the region upstream of the BP (**Figure 5G**) (see Discussion).

BP identification: B reads and BAL algorithm

Although intron ends are readily defined from RNA-Seq data, BP locations are more challenging. If reverse transcriptase initiates cDNA synthesis near the 5' end of a lariat intron, it can cross the BP and generate a read wherein the region upstream of the BP is concatenated with the region downstream of the 5'SS to generate a branched ("B") read (**Figure 6B**). These B reads often contain a mismatch or indel at the BP nucleotide due to the presence of a 2'-5' bond (Gao et al., 2008; Vogel and Borner, 2002; Vogel et al., 1997). Since our purified spliceosomes should be highly enriched for BPs, we developed a Branch ALigner (BAL) algorithm similar to one previously described for mapping human BPs (Mercer et al., 2015) and applied it to intron RNA-Seq and spliceosome footprinting libraries constructed from RNAs copurifying with U2·U5·U6·NTC complexes from both WT and dbr1Δ strains (**Figure 6A**). In all, we identified 4,389 B reads from a total of 186,863,605 reads, enabling us to map 1,463 BPs in 1,420 introns. The vast majority (98.6%) of these B reads had a mismatch at the BP. While many of BAL-defined BPs overlapped with those previously identified in studies also employing B reads (Awan et al., 2013) (Bitton et al., 2014) (**Figure 6C**), 770 were new. All three studies combined resulted in positive identification of 3,759 BPs in 2,374 of *S. pombe's* 5,594 introns (~42% of total). However, the high fraction of non-overlapping BPs among the three studies indicated that none achieved saturation.

BP identification: G reads and random forest approach

As the above numbers illustrate, a major limitation to detecting branches via B reads is their extreme scarcity, even in libraries highly enriched for introns (one per 200,000 mapped reads in our datasets). This is likely due to the propensity of reverse transcriptase to stop at a BP rather than cross the 2'-5' bond and continue on. In our libraries the ratio of B reads to reads ending at the 5'SS was 0.0075, meaning that reverse transcriptase stops >99% of the time

when it encounters a 2'-5' bond. Much more common in our DBE- datasets were "G reads": genome-mapping reads either crossing or starting at the BP (**Figure 7A** and **B**). In the dbr1Δ intron RNA-Seq and 3'-OH libraries 8% of all genome-mapping reads (2,202,687 of 26,669,756 total) ended at a 3'SS (i.e., came from the intron product). For the introns with a BP identified by both BAL and Anwar *et al.* (2012), 15% of these crossed the BP, 44% began at the BP nucleotide and 32% began one nucleotide downstream of the BP. The remaining 9% began between the BP and 3'SS, so were uninformative as to BP location. Thus 7.3% (= 8% x 91%) of all genome-mapping reads in the dbr1Δ intron RNA-Seq and 3'-OH libraries contained information as to a BP location. Further, reverse transcriptase is substantially more efficient at crossing the 3'-5' bond at the BP (15% read through) than it is at crossing the 2'-5' bond (0.075% read through).

As with the B reads, we found that reverse transcriptase has a strong tendency to make a mistake (point mutation or indel 67% of the time) when traversing the 3'-5' bond at the BP (**Figure 7C**). It therefore seemed likely that we could use this information along with other features (e.g., BP sequence motifs, distance to 3'SS) to identify additional BP's via a supervised machine learning approach (**Figure 7D**). Using the combined reads from all U2·U5·U6·NTC complex dbr1Δ DBE- intron RNA-Seq and 3'-OH datasets (4 in total) and a set of 251 highly validated branch points (i.e., identified by both BAL and (Awan et al., 2013)), we generated and validated a model using the Random Forest algorithm (Breiman, 2001) (**Figure 7E)**. When challenged with a 50 BP test set, this model exhibited remarkably high performance (AUROC = 0.99; AUPR= 0.98; **Figure S5A**), outperforming a BP identification method using solely the BP consensus motif (AUROC= 0.98, AUPR= 0.82; **Figure S5B**). Among all features in our model, the most important was the fraction of 3'-OH reads terminating at the BP, followed by distance from the BP to the 3'SS (**Figure 7F**).

Using our Random Forest model, we were able to identify 5,421 BPs in 4,776 introns (86%) (**Figure S5C**). The previous branch point datasets (Pleiss and LaSSO) are largely included in our set (**Figure 7G**), with the combined set representing 5,045 of 5,594 introns (90%) **(Table S4**). In most aspects (e.g., BP motif, distance to 3'SS,) BPs identified by our Random Forest approach were indistinguishable from previously identified BPs (**Figure S5DE**). However, whereas all B read-only methods favored BP identification in longer introns, our Random Forest approach showed no such intron length bias (**Figure S5F**). Thus, taking advantage of the G reads contained in deep sequencing libraries prepared from purified, lariat-containing spliceosomes allows for much more comprehensive BP identification transcriptome-wide than approaches based solely on B reads.

Discussion

Recent implementation of "ribosome profiling" has greatly advanced the study of general and regulated translation. That technique, which involves deep sequencing of the mRNA footprints underlying translating ribosomes, is transforming our understanding of post-transcriptional gene regulation processes occurring in the cytoplasm (Brar and Weissman, 2015; Ingolia, 2014). By analogy, we hereby dub the deep sequencing of RNAs co-purifying with spliceosomes as "spliceosome profiling". Because it provides a means to directly interrogate all sites of spliceosome assembly and catalysis genome-wide, we expect that spliceosome profiling will similarly transform our understanding of post-transcriptional gene regulation in the nucleus.

Depending on the type of deep sequencing library constructed, spliceosome profiling can provide many different types of information. For example, intron RNA-Seq libraries (i.e., containing long random fragments of introns) enable visualization of entire introns (**Figure 1**). Exactly where on the introns spliceosomes reside can be revealed by including a nuclease digestion step to generate spliceosome footprints (**Figure 5**). Creation of 5'-P and 3'-OH libraries after debranching allows for the unambiguous identification of intron ends at nucleotide resolution, discovery of previously unannotated introns and alternative splice sites (**Figure 2**), and relative quantitation of splicing intermediates versus products for individual introns (**Figure 3**). Finally, a new random forest approach for detecting BPs applied to various libraries from samples not subjected to debranching allows for intronome-wide BP identification at unprecedented depths (**Figure 7**).

In the work reported here, we limited our analysis to RNAs co-purifying with a Protein A-tagged U2 component and a CBP-tagged U5 component in *S. pombe* during late log growth in rich media. This tandem affinity approach purifies endogenous U2·U5·U6·NTC complexes, which mainly consist of spliceosomes containing fully excised lariat introns (ILS complex).

Profiling of these complexes thus provides a transcriptome-wide snapshot for the very last stage of spliceosome action. However, many additional spliceosome assembly states exist in cells, including cross-exon complexes and cross-intron U1·U2 Complex A (aka, pre-spliceosomes), U1·U2·U4·U5·U6 pre-activation Complex B, U2·U5·U6·NTC post-activation Complex B^{act}, U2·U5·U6·NTC splicing intermediate-containing Complex C, and U2·U5·U6·NTC splice product-containing Complex P. Extensive proteomics analyses of these different complexes assembled in vitro have revealed that each has a unique protein signature (Fabrizio et al., 2009; Wahl et al., 2009). Therefore, by judicious choice of protein pairs it should be possible to query nearly every spliceosome assembly state genome-wide by tandem affinity purification or RNA immunoprecipitation in tandem (RIPiT) (Singh et al., 2012; 2014). Gene editing advances now make it possible to insert affinity tags into almost any organism (Lackner et al., 2015). Thus, our overall approach should enable generation of genome-wide occupancy maps for all major spliceosome assembly states in diverse cell types. We predict that spliceosome profiling will prove especially valuable for elucidating the mechanisms governing alternative pre-mRNA processing among different tissues of multicellular organisms or in cells subject to different growth or stress conditions.

New sites of spliceosome action

Our analysis via 5'-P-Seq and 3'-OH-Seq of intron ends copurifying with endogenous U2·U5·U6·NTC complexes enabled us to identify >200 new *S. pombe* introns and correct the annotation of numerous others (**Figure 2 and Table S2**). Many of the new or corrected introns were in UTRs or near the end of an ORF, where computational algorithms for predicting gene structure are least reliable. Importantly, many corrected intron annotations change the predicted protein sequence, which can have significant consequences for protein localization and/or function. For example, a new intron arising from an alternative TSS in the ADH4 gene encodes

a protein isoform lacking the N-terminal mitochondrial targeting sequence (MTS) present in the previously annotated coding region. We showed that relative abundance of the two mRNA isoforms depends on cellular growth and stress conditions. Thus, spliceosome profiling led us to discover a stress-responsive transcriptional and splicing switch that alters the subcellular targeting of a key metabolic enzyme. Our data also revealed two instances of recursive splicing via zero-length exons, an AT-AC intron, a regulatory intron retention event in the gene for a key spliceosomal protein, and scores of new intron-containing transcripts (**Table S2B**). While these latter transcripts were readily detectable in our purified U2·U5·U6·NTC complexes, their spliced exon products were not apparent in traditional RNA-Seq libraries likely due to their rapid decay by nonsense-mediated decay (NMD). These findings demonstrate the utility of spliceosome profiling both for improving genome annotation and for identifying all sites of spliceosome catalysis regardless of intragenic position, matches to expected splice site consensus sequences or subsequent spliced exon stability.

In addition to allowing for intron identification and more accurate splice site mapping, 3'-OH read ratios at exon and intron ends provides a new window for assessing second step splicing efficiencies. Although transcriptome-wide methods already exist for monitoring overall splicing kinetics (Barrass et al., 2015; Bhatt et al., 2012), none can specifically detect splicing intermediates. By using 3'-OH-Seq to specifically query exon and intron ends in catalytically active spliceosomes, our approach allows quantitative analysis of first and second step chemistries (**Figure 3**). Using this approach we found that many introns with high levels of splicing intermediate occur within ribosomal protein genes, consistent with previous findings that splicing efficiency regulates ribosomal protein expression in *S. cerevisiae* (Juneau et al., 2006; Parenteau et al., 2011).

Extensive spliceosomal footprints

Recent cryo-EM structure determination of late stage spliceosomes has enabled visualization of the 5'SS and BP spatial organization within the catalytic core (Galej et al., 2016; Wan et al., 2016a; Yan et al., 2015). To date the structures containing the most modeled intron nucleotides are the *S. cerevisiae* C and C* complexes (Wan et al., 2016a; Yan et al., 2016a). In all, these models contain 15 nts downstream of the 5'SS and 23 nts surrounding the BP for a total of 38 nts. Consistent with this, our analysis of protected RNA fragments co-purifying with endogenous *S. pombe* U2·U5·U6·NTC complexes revealed a ~31 nt footprint downstream of the 5'SS (**Figure 5D**) and a ~25 nt footprint surrounding the BP (**Figure 5E**). Because the 5'SS and BP consensus sequences (GUAGA and UACAAC, respectively) are much shorter than the footprints, spliceosomal contacts with adjacent regions are presumably driven by sugar-phosphate backbone interactions (Allain et al., 1997; Jovine et al., 1996; Wan et al., 2016b). These structural contacts likely strengthen the association between the splicing machinery and intron so as to prevent consensus sequence dissociation during the many structural rearrangements required for spliceosome assembly and activation.

In addition to the 5'SS and BP footprints, we detected a third ~38 nt footprint on long introns centered ~38-39 nts upstream of the BP (**Figure 5F**). In human C complex, up to 44 nts 5' of the BP are protected from RNase digestion (Wolf et al., 2012), with nts -33 to -40 interacting with Aquarius (AQR; aka IBP160/KIAA0560 in humans) (Hirose et al., 2006). A stoichiometric component of our purified U2·U5·U6·NTC complexes (Chen et al., 2014), the *S. pombe* AQR homolog, Cwf11, is thus likely responsible for this third footprint. Consistent with the lack of any detectable sequence motifs under the third footprint, AQR is an SF1 RNA helicase containing a sequence-independent RNA binding site spanning ~ 7 nts (De et al., 2015; Hirose et al., 2006). In the recent cryo-EM structures of late stage *S. pombe* and human spliceosomes (Bertram et al., 2017; Yan et al., 2015), this RNA binding site is perfectly

positioned relative to the spliceosomal catalytic core (~14 nm away) to account for the observed footprint (**Figure 5G**) .

Given the positioning of the Cwf11 footprint ~50 nts upstream of the 3'SS (**Figure 5F**), an unresolved question is where does Cwf11 bind on introns not long enough to accommodate it (e.g., on the vast majority of S. pombe introns that are <60 nts)? In these cases does it simply not bind, or does it interact with the 5' exon upstream of the 5'SS? In human C complex, the last 27 nts of the 5' exon are fully protected from nuclease digestion, with a previously unidentified ~170 kDa protein crosslinking in the region between the exon junction complex (EJC) deposition site and the 5'SS (Jurica et al., 2002; Reichert et al., 2002; Wolf et al., 2012). Intriguingly, in mammalian spliceosomes the 160 kDa AQR serves as a recruitment site for EJC components prior to their deposition on the 5' exon (Ideue et al., 2007). Thus the ~170 kDa crosslinked protein is most likely AQR.

New methods for BP mapping

Although the BP consensus motif has high information content in budding and fission yeast, such is not the case in most metazoans (Lim and Burge, 2001). Thus, human BPs cannot be accurately predicted from sequence information alone. However, recent realization that mutations in spliceosomal proteins involved in BP selection are major drivers of cancer (Carrocci et al., 2017; Darman et al., 2015; Je et al.; Tang et al., 2017) has greatly increased the need for tools to precisely map human BPs. Our work here demonstrates that deep sequencing of RNA contained in late stage spliceosomes can increase BP identification to unprecedented depth. We used two different methods. The first, BAL, relied on identification of concatenated, branched (B) reads in our intron RNA-Seq and spliceosome footprinting libraries. Although spliceosome purification should greatly increase the prevalence of such reads, we found that even in the footprinting libraries where all sequences not stably protected by the splicing

machinery were eliminated, only 4,389 of 186,863,605 (frequency = $2x10^{-5}$) reads corresponded to a B read. In comparison, 369,746 reads in the DBE- libraries ended exactly at the 5'SS indicating that the read-through efficiency of reverse transcriptase across a 2'-5' bond is 0.75%. Consistent with B reads being insufficiently abundant to saturate coverage, we observed only partial overlap between the set of BPs we identified with BAL and previously identified sets (**Figure 6C**).

A much more successful approach was to use genome-mapping (G) reads that crossed or ended at the BP from the 3' to 5' direction. In our 3'-OH DBE- libraries, such reads represented 8% of all mapped reads. Our data indicate that when reverse transcriptase initiates downstream of the BP, its read-through efficiency across the BP 3'-5' bond (15%) is substantially higher its read-through efficiency across a 2'-5' bond (0.75%). Further, when the BP 3'-5' bond is traversed, there is a very high rate of mutation (67% of reads contain a point mutation or indel). By incorporating these features into a random forest algorithm (**Figure 7**), we were able to positively identify BPs on >86% of all introns in *S. pombe*. Importantly, the majority of all previously identified *S. pombe* BPs were included in this set. Thus our random forest approach using G reads from RNAs copurifying with late stage spliceosomes is a powerful new approach for BP identification. We anticipate that high-depth BP identification will be another useful application of spliceosome profiling.

AUTHOR CONTRIBUTIONS

M.J.M. conceived and designed the experiments. W.C. performed the biochemical experiments. J.M., H.O. and H.P.S. performed all computational analyses. W.C., J.M., H.O., N.R., Z.W., and M.J.M. interpreted the data. W.C., J.M., and M.J.M. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Declaration of Interests

The authors declare no competing interests.

ACKNOWLEDGMENTS

We thank Beate Schwer and Masad Damha for their generous gifts of DBE and bRNA, respectively. We also thank Harleen Saini and Carrie Kovalak for critical reading of the manuscript. This work was supported by funding from HHMI, NIH R01-GM53007 (M.J.M.) and NSF DBI-0850008 (Z.W.). M.J.M. was a HHMI Investigator at the time this study was conducted.

REFERENCES

Allain, F.H.T., Howe, P.W.A., Neuhaus, D., and Varani, G. (1997). Structural basis of the RNA-binding specificity of human U1A protein. The EMBO Journal *16*, 5764–5774.

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., et al. (2012). Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition. Cell Reports *1*, 543–556.

Awan, A.R., Manfredo, A., and Pleiss, J.A. (2013). Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. Proc. Natl. Acad. Sci. U.S.a. *110*, 12762–12767.

Barrass, J.D., Reid, J.E.A., Huang, Y., Hector, R.D., Sanguinetti, G., Beggs, J.D., and Granneman, S. (2015). Transcriptome-wide RNA processing kinetics revealed using extremely short 4tU labeling. Genome Biology 2014 15:1 *16*, 282.

Bertram, K., Agafonov, D.E., Liu, W.-T., Dybkov, O., Will, C.L., Hartmuth, K., Urlaub, H., Kastner, B., Stark, H., and LUhrmann, R. (2017). Cryo-EM structure of a human spliceosome activated for step 2 of splicing. Nat *542*, 318–323.

Bhatt, D.M., Pandya-Jones, A., Tong, A.-J., Barozzi, I., Lissner, M.M., Natoli, G., Black, D.L., and Smale, S.T. (2012). Transcript Dynamics of Proinflammatory Genes Revealed by Sequence Analysis of Subcellular RNA Fractions. Cell *150*, 279–290.

Bitton, D.A., Rallis, C., Jeffares, D.C., Smith, G.C., Chen, Y.Y.C., Codlin, S., Marguerat, S., and Bahler, J. (2014). LaSSO, a strategy for genome-wide mapping of intronic lariats and branch points using RNA-seq. Genome Research *24*, 1169–1179.

Box, J.A., Bunch, J.T., Tang, W., and Baumann, P. (2008). Spliceosomal cleavage generates the 3′ end of telomerase RNA. Nature *456*, 910–914.

Brar, G.A., and Weissman, J.S. (2015). Ribosome profiling reveals the what, when, where and how of protein synthesis. Nat Rev Mol Cell Biol *16*, 651–664.

Carrocci, T.J., Zoerner, D.M., Paulson, J.C., and Hoskins, A.A. (2017). SF3b1 mutations associated with myelodysplastic syndromes alter the fidelity of branchsite selection in yeast. Nucleic Acids Research gkw1349.

Chapman, K.B., and Boeke, J.D. (1991). Isolation and characterization of the gene encoding yeast debranching enzyme. Cell *65*, 483–492.

Chen, W., Shulha, H.P., Ashar-Patel, A., Yan, J., Green, K.M., Query, C.C., Rhind, N., Weng, Z., and Moore, M.J. (2014). Endogenous U2.U5.U6 snRNA complexes in S. pombe are intron lariat spliceosomes. Rna *20*, 308–320.

Clark, M.B., Johnston, R.L., Inostroza-Ponta, M., Fox, A.H., Fortini, E., Moscato, P., Dinger, M.E., and Mattick, J.S. (2012). Genome-wide analysis of long noncoding RNA stability. Genome Research *22*, 885–898.

Clark, N.E., Katolik, A., Roberts, K.M., Taylor, A.B., Holloway, S.P., Schuermann, J.P., Montemayor, E.J., Stevens, S.W., Fitzpatrick, P.F., Damha, M.J., et al. (2016). Metal dependence and branched RNA cocrystal structures of the RNA lariat debranching enzyme Dbr1. Proc. Natl. Acad. Sci. U.S.a. *113*, 14727–14732.

Claros, M.G., and Vincens, P. (1996). Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur. J. Biochem. *241*, 779–786.

Crichton, P.G., Affourtit, C., and Moore, A.L. (2007). Identification of a mitochondrial alcohol dehydrogenase in Schizosaccharomyces pombe: new insights into energy metabolism. Biochem. J. *401*, 459.

Darman, R.B., Seiler, M., Agrawal, A.A., Lim, K.H., Peng, S., Aird, D., Bailey, S.L., Bhavsar, E.B., Chan, B., Colla, S., et al. (2015). Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point. Cell Reports *13*, 1033–1045.

De, I., Bessonov, S., Hofele, R., Santos, dos, K., Will, C.L., Urlaub, H., LUhrmann, R., and Pena, V. (2015). The RNA helicase Aquarius exhibits structural adaptations mediating its recruitment to spliceosomes. Nature Structural & Molecular Biology *22*, 138–144.

Fabrizio, P., Dannenberg, J., Dube, P., Kastner, B., Stark, H., Urlaub, H., and Luhrmann, R. (2009). The evolutionarily conserved core design of the catalytic activation step of the yeast spliceosome. Molecular Cell *36*, 593–608.

Galej, W.P., Wilkinson, M.E., Fica, S.M., Oubridge, C., Newman, A.J., and Nagai, K. (2016). Cryo-EM structure of the spliceosome immediately after branching. Nature *537*, 197–201.

Gao, K., Masuda, A., Matsuura, T., and Ohno, K. (2008). Human branch point consensus sequence is yUnAy. Nucleic Acids Research *36*, 2257–2267.

Geisberg, J.V., Moqtaderi, Z., Fan, X., Ozsolak, F., and Struhl, K. (2014). Global Analysis of mRNA Isoform Half-Lives Reveals Stabilizing and Destabilizing Elements in Yeast. Cell *156*, 812–824.

Georgomanolis, T., Sofiadis, K., and Papantonis, A. (2016). Cutting a Long Intron Short: Recursive Splicing and Its Implications. Front. Physiol. *7*, e1002215.

Hang, J., Wan, R., Yan, C., and Shi, Y. (2015). Structural basis of pre-mRNA splicing. Science *349*, 1191–1198.

Heyer, E.E., Ozadam, H., Ricci, E.P., Cenik, C., and Moore, M.J. (2015). An optimized kit-free method for making strand-specific deep sequencing libraries from RNA fragments. Nucleic Acids Research, 2015, Vol. 43, No.1.

Hirose, T., Ideue, T., Nagai, M., Hagiwara, M., Shu, M.-D., and Steitz, J.A. (2006). A Spliceosomal Intron Binding Protein, IBP160, Links Position-Dependent Assembly of Intron-Encoded Box C/D snoRNP to Pre-mRNA Splicing. Molecular Cell *23*, 673–684.

Hu, Y., Liu, Y., Mao, X., Jia, C., Ferguson, J.F., Xue, C., Reilly, M.P., Li, H., and Li, M. (2014).

PennSeq: accurate isoform-specific gene expression quantification in RNA-Seq by modeling non-uniform read distribution. Nucleic Acids Research *42*, e20–e20.

Ideue, T., Sasaki, Y.T.F., Hagiwara, M., and Hirose, T. (2007). Introns play an essential role in splicing-dependent formation of the exon junction complex. Genes &Amp; Development *21*, 1993–1998.

Ingolia, N.T. (2014). Ribosome profiling: new views of translation, from single codons to genome scale. Nat. Rev. Genet. *15*, 205–213.

Jacob, A.G., and Smith, C.W.J. (2017). Intron retention as a component of regulated gene expression programs. Hum Genet *136*, 1043–1057.

Jarrell, K.A., Peebles, C.L., Dietrich, R.C., Romiti, S.L., and Perlman, P.S. (1988). Group II intron self-splicing. Alternative reaction conditions yield novel products. J. Biol. Chem. *263*, 3432–3439.

Je, E.M., Yoo, N.J., Kim, Y.J., Kim, M.S., and Lee, S.H. Mutational analysis of splicing machinery genes SF3B1, U2AF1 and SRSF2 in myelodysplasia and other common tumors. International Journal of Cancer *133*, 260–265.

Jenkins, J.L., and Kielkopf, C.L. (2017). Splicing Factor Mutations in Myelodysplasias: Insights from Spliceosome Structures. Trends in Genetics *33*, 336–348.

Jovine, L., Oubridge, C., Avis, J.M., and Nagai, K. (1996). Two structurally different RNA molecules are bound by the spliceosomal protein U1A using the same recognition strategy. Structure *4*, 621–631.

Juneau, K., Miranda, M., Hillenmeyer, M.E., Nislow, C., and Davis, R.W. (2006). Introns Regulate RNA and Protein Abundance in Yeast. Genetics *174*, 511–518.

Jurica, M.S., Licklider, L.J., Gygi, S.R., Grigorieff, N., and Moore, M.J. (2002). Purification and characterization of native spliceosomes suitable for three-dimensional structural analysis. Rna *8*, 426–439.

Khalid, M.F. (2005). Structure-function analysis of yeast RNA debranching enzyme (Dbr1), a manganese-dependent phosphodiesterase. Nucleic Acids Research *33*, 6349–6360.

Khalid, M.F., Damha, M.J., Shuman, S., and Schwer, B. (2005). Structure-function analysis of yeast RNA debranching enzyme (Dbr1), a manganese-dependent phosphodiesterase. Nucleic Acids Research *33*, 6349–6360.

Kim, D., Ben Langmead, and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat *12*, 357–360.

Lackner, D.H., Carré, A., Guzzardo, P.M., Banning, C., Mangena, R., Henley, T., Oberndorfer, S., Gapp, B.V., Nijman, S.M.B., Brummelkamp, T.R., et al. (2015). A generic strategy for CRISPR-Cas9-mediated gene tagging. Nat Comms *6*, ncomms10237.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat *9*,

357–359.

Li, W., and Jiang, T. (2012). Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads. Bioinformatics *28*, 2914–2921.

Lim, L.P., and Burge, C.B. (2001). A computational analysis of sequence features involved in recognition of short introns. Proc. Natl. Acad. Sci. U.S.a. *98*, 11193–11198.

McGlincy, N.J., and Smith, C.W.J. (2008). Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? Trends Biochem. Sci. *33*, 385–393.

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

Moran, Y., Fredman, D., Praher, D., Li, X.Z., Wee, L.M., Rentzsch, F., Zamore, P.D., Technau, U., and Seitz, H. (2014). Cnidarian microRNAs frequently regulate targets by cleavage. Genome Research *24*, 651–663.

Nilsen, T.W. (2014). 3′-End Labeling of RNA with [5′-32P]Cytidine 3′,5′-Bis(Phosphate) and T4 RNA Ligase 1. Cold Spring Harb Protoc *2014*, pdb.prot080713–pdb.prot080713.

Parenteau, J., Durand, M., Morin, G., Gagnon, J., Lucier, J.-F., Wellinger, R.J., Chabot, B., and Abou Elela, S. (2011). Introns within Ribosomal Protein Genes Regulate the Production and Function of Yeast Ribosomes. Cell *147*, 320–331.

Podar, M., Chu, V.T., Pyle, A.M., and Perlman, P.S. (1998). Group II intron splicing in vivo by first-step hydrolysis. Nature *391*, 915–918.

Pyle, A.M. (2016). Group II Intron Self-Splicing. Annu. Rev. Biophys. *45*, 183–205.

Qin, D., Huang, L., Wlodaver, A., Andrade, J., and Staley, J.P. (2016). Sequencing of lariat termini in S. cerevisiaereveals 5′ splice sites, branch points, and novel splicing events. Rna *22*, 237–253.

Reichert, V.L., Hir, H.L., Moore, S.J., and Moore, M.J. (2002). 5' exon interactions within the human spliceosome establish a framework for exon junction complex structure and assembly. Genes & Development *16*, 2778–2791.

Rhind, N., Chen, Z., Yassour, M., Thompson, D.A., Haas, B.J., Habib, N., Wapinski, I., Roy, S., Lin, M.F., Heiman, D.I., et al. (2011). Comparative Functional Genomics of the Fission Yeasts. Science *332*, 930–936.

Romfo, C.M., Alvarez, C.J., van Heeckeren, W.J., Webb, C.J., and Wise, J.A. (2000). Evidence for splice site pairing via intron definition in Schizosaccharomyces pombe. Molecular and Cellular Biology *20*, 7955–7970.

Schwer, B. (2008). A Conformational Rearrangement in the Spliceosome Sets the Stage for Prp22-Dependent mRNA Release. Molecular Cell *30*, 743–754.

Singh, G., Kucukural, A., Cenik, C., Leszyk, J.D., and Shaffer, S.A. (2012). The Cellular EJC Interactome Reveals Higher-Order mRNP Structure and an EJC-SR Protein Nexus. Cell.

Singh, G., Ricci, E.P., and Moore, M.J. (2014). RIPiT-Seq: A high-throughput approach for footprinting RNA:protein complexes. Methods *65*, 320–332.

Taggart, A.J., DeSimone, A.M., Shih, J.S., Filloux, M.E., and Fairbrother, W.G. (2012). Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. Nature Structural & Molecular Biology *19*, 719–721.

Taggart, A.J., Lin, C.-L., Shrestha, B., Heintzelman, C., Kim, S., and Fairbrother, W.G. (2017). Large-scale analysis of branchpoint usage across species and cell lines. Genome Research *27*, 639–649.

Tang, Q., Rodriguez-Santiago, S., Wang, J., Pu, J., Yuste, A., Gupta, V., Moldón, A., Xu, Y.-Z., and Query, C.C. (2017). SF3B1/Hsh155 HEAT motif mutations affect interaction with the spliceosomal ATPase Prp5, resulting in altered branch site selectivity in pre-mRNA splicing. Genes & Development *30*, 2710–2723.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc *7*, 562–578.

Tseng, C.K., and Cheng, S.C. (2008). Both Catalytic Steps of Nuclear Pre-mRNA Splicing Are Reversible. Science *320*, 1782–1784.

Tseng, C.K., and Cheng, S.C. (2013). The spliceosome catalyzes debranching in competition with reverse of the first chemical reaction. Rna *19*, 971–981.

Tuerk, A., Wiktorin, G., and Güler, S. (2014). Mixture models reveal multiple positional bias types in RNA-Seq data and lead to accurate transcript concentration estimates. bioRxiv 011767.

Vogel, J., and Borner, T. (2002). Lariat formation and a hydrolytic pathway in plant chloroplast group II intron splicing. The EMBO Journal *21*, 3794–3803.

Vogel, J., Hess, W., and Borner, T. (1997). Precise branch point mapping and quantification of splicing intermediates. Nucleic Acids Research *25*, 2030–2031.

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

Wan, L., Yan, X., Chen, T., and Sun, F. (2012). Modeling RNA degradation for RNA-Seq with applications. Biostatistics *13*, 734–747.

Wan, R., Yan, C., Bai, R., Huang, G., and Shi, Y. (2016a). Structure of a yeast catalytic step I spliceosome at 3.4 A resolution. Science *353*, 895–904.

Wan, R., Yan, C., Bai, R., Wang, L., Huang, M., Wong, C.C.L., and Shi, Y. (2016b). The 3.8 Å structure of the U4/U6.U5 tri-snRNP: Insights into spliceosome assembly and catalysis. Science *351*, 466–475.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Wolf, E., Kastner, B., and LUhrmann, R. (2012). Antisense-targeted immuno-EM localization of the pre-mRNA path in the spliceosomal C complex. Rna *18*, 1347–1357.

Wu, Q., and Krainer, A.R. (1997). Splicing of a divergent subclass of AT-AC introns requires the major spliceosomal snRNAs. Rna *3*, 586–601.

Yan, C., Wan, R., Bai, R., Huang, G., and Shi, Y. (2016a). Structure of a yeast step II catalytically activated spliceosome. Science.

Yan, C., Wan, R., Bai, R., Huang, G., and Shi, Y. (2016b). Structure of a yeast activated spliceosome at 3.5 A resolution. Science *353*, 904–911.

Yan, C., Hang, J., Wan, R., Huang, M., Wong, C.C.L., and Shi, Y. (2015). Structure of a yeast spliceosome at 3.6-angstrom resolution. Science *349*, 1182–1191.

Zhang, X., Yan, C., Hang, J., Finci, L.I., Lei, J., and Shi, Y. (2017). An Atomic Structure of the Human Spliceosome. Cell *169*, 918–929.e14.

Figure legends

Figure 1. Overview of RNA species contained in U2·U5·U6·NTC complexes

(A) Schematic of intron RNA-Seq library preparation from tandem affinity-purified *S. pombe* U2·U5·U6·NTC complexes tagged with Protein A and calmodulin binding peptide (CBP). Extracted RNAs were treated with debranching enzyme (DBE), randomly fragmented by hydrolysis, 3'-dephosphorylated and then used for intron RNA-Seq library preparation via 3' adaptor ligation.

(B) Pie chart showing genome-mapping read distributions across indicated RNA species and classes.

(C) Genome browser screenshot of individual genes containing one, two, three, or fifteen previously-annotated introns showing intron RNA-Seq read coverage (numbers in brackets) in a single biological replicate. Thick gray rectangles: coding regions; medium black lines: UTRs; thin yellow lines: introns.

Figure 2. Mapping sites of spliceosome catalysis with 5'-P and 3'-OH libraries

(A) Schematic of 5'-P and 3'-OH library preparation from *S. pombe* U2·U5·U6·NTC complexes. Following RNA extraction and DBE treatment, samples were divided and used directly for either 5' or 3' adaptor ligation.

(B) Aggregation plot of all WT DBE+ 5'-P 5' read ends (red) and 3'-OH 3' read ends (blue) across all annotated 5' and 3' splice sites in CDS genes.

(C) An ATAC intron in lea1 (SPBC1861.08c). Top: intron RNA-Seq read distribution on lea1 intron 1. Middle: 5'-P (red) and 3'-OH (blue) read ends from a single biological replicate (read numbers in brackets). Bottom: previous intron annotation.

(D) Zero-length exon and recursive splicing in SPNCRNA.445 gene. Top: SPNCRNA.445 and flanking genes showing 5'-P and 3'-OH read ends, polyA+ RNA-Seq reads (Rhind et al., 2011) and sequence conservation across four *Schizosaccharomyces* species (Rhind et al., 2011). Bottom: Expanded view of SPNCRNA.445 intron showing zero-length exon and polyA+ RNA-Seq reads supporting recursive splicing.

(E) A new transcription start site (TSS1) and intron in ADH4 gene. Top: ADH4 gene in ASM294v2.35 *S. pombe* genome annotation, 5'-P (red) and 3'-OH (blue) read ends from a single biological replicate (read numbers in brackets) and position of new 543 nt intron. Middle: SYBR gold-stained polyacryamide gel of PCR products (U: unspliced; S: spliced) from genomic DNA (G) and cDNA (C). Bottom left: Adh4 mRNA isoforms. Isoform II initiates at TSS2 and encodes a 380 aa protein with a predicted mitrochondrial targeting sequence (MTS; Mitoprot score = 0.9971; (Claros and Vincens, 1996). Isoform I initiates at TSS1 and after splicing encodes a 337 aa protein lacking the MTS. Bottom right: Primer extension (primer location indicated by red splotch) reveals that isoform II predominates during log phase growth, but isoform I appears upon stress (cycloheximide (CHX) or late log/glucose deprivation).

Figure 3. High splicing intermediate introns

(A) Examples of introns with different levels of splicing intermediates. Numbers in brackets: 3'-OH 3' read ends terminating at the 5'SS or 3'SS.

(B) Histogram of percent of splicing intermediate (calculated as fraction of 5'SS 3'-OH to all intron 3'-OH reads) across all introns with $\geq$ 25 3'SS 3'-OH reads in both replicates. Dashed vertical lines indicate introns shown in A.

(C) Violin plots of fraction of exon junction reads for introns with high (>17%) splicing intermediate to all other introns in the same gene (gene matched introns). P-values: Wilcoxon rank sum test.

(D) Example of a gene with a high splicing intermediate intron flanked by "gene matched" introns with low splicing intermediate.

(E) Violin plots comparing GC content of introns and exons (upstream, -1 to -40 nts from 5'SS; downstream; +1 to +40 nts from 3'SS) for sets described in C. P-values: Wilcoxon rank sum test.

(F) Violin plots comparing folding free energy (RNAFold) for upstream and downstream exon sequences for sets described in C. P-values: Wilcoxon rank sum test.

Figure 4. Dbr1-independent debranching

(A) 5'-P and 3'-OH read ends on SPACBE11.07c intron 1 copurifying with U2·U5·U6·NTC complexes from WT or *dbr1Δ* cells with (+) or without (-) DBE treatment. Observation of a 5'SS 5'P peak only upon DBE treatment indicates a high proportion of lariat species.

(B) Same as A, except SPBC1734.14c intron 1. Observation of 5'-P peak in all samples indicates that a fraction of intron-containing species copurifying with U2·U5·U6·NTC complexes are linear and are not due to Dbr1 activity.

(C) Polyacrylamide gel showing migration of 3'-end labeled branched oligo (structure at right) and cleavage products without (-) or following DBE treatment (+) or incubation in WT or *dbr1Δ* cell extract. Lack of cleavage products in the *dbr1Δ* cell extract indicates absence of another cellular debranching activity.

(D) Same as B, except 5'-P reads from cellular RNA.

Figure 5. Spliceosome footprints

(A) Schematic of footprint library preparation from purified U2·U5·U6·NTC complexes treated with micrococcal nuclease (MNase).

(B) Examples of footprints on short (51 nt), medium (96 and 150 nt), and long (210 and 216 nt) introns. Intron identities are as indicated, with scales in brackets.

(C) Scatter plot showing near perfect correlation between intron length and average number of footprint reads for introns up to 56 nts.

(D) Aggregation plot of normalized read counts (number of reads at position divided by total number of intron reads) from both replicates of footprinting libraries mapping to the end of the 5' exon and/or within the first 150 nts for 314 introns with length >200 nts. Grey: Counts from reads mapping fully within the intron; Green: Counts from reads spanning the 5'SS; Red: Counts from reads mapping fully within the exon; Purple: Counts from exon junction spanning reads. Inset: sequence logo for 5'SS motif found *de novo* using MEME.

(E) Same as D, except numbers indicate nucleotide positions relative to 3'SS. Inset: sequence logo for 3'SS motif found *de novo* using MEME.

(F) Same as E, except numbers indicate nucleotide positions relative to BP identified using FIMO and consensus motif. Whereas MEME identified a strong motif over the BP, no motifs were detectable in the footprint centered 48-49 nts upstream.

(G) Partial structure of the S. pombe U2·U5·U6·NTC complex (Yan et al., 2015) showing distance between BP and RNA binding motif in Cwf11.

Figure 6. BP identification by B reads and BAL algorithm

(A) Schematic of libraries used to identify B reads.

(B) Top: Schematic of BAL pipeline showing 20 nt seed (light green) downstream of 5'SS, seed extension (dark green) and 5' remainder mapping upstream of the BP (purple). Bottom:

Examples of BPs supported by B reads.

(C) Overlap of the number of introns with BPs identified by BAL (blue), LaSSO (pink; (Bitton et al., 2014)), and 2D gel lariat purification (yellow;(Awan et al., 2013)).

Figure 7. BP identification by G reads and random forest approach

(A) Examples of intron RNA-seq reads used to identify BPs. Upper left: When reads in DBE- libraries traverse the BP, they often contain a mutation at BP (black). Upper right: Other DBE- reads terminate at or one nucleotide before the BP. Lower: These patterns are not observed in DBE+ libraries.

(B) Schematic of genome mapping "G" reads that can be used to identify BPs. Such reads can derive from lariat product (left) or lariat intermediate (right).

(C) Mismatch rates of G reads crossing BPs identified by both BAL and 2D gel lariat purification.

(D) Features used in Random Forest Model.

(E) Overview of Random Forest Model. Using the features in D, we trained the model on 200 BPs identified by both BAL and 2D gel lariat purification. We tuned parameters using a 50 BP validation set and evaluated model performance with a 51 BP test set.

(F) Importance of features from Random Forest Model. Features are colored based on whether they are derived from the intron RNA-Seq data (green), 3'-OH data (blue), or from the sequence itself (gray).

(G) Overlap of the number of introns with BPs identified by BAL or Random Forest Model (blue), LaSSO (pink), and 2D gel lariat purification (yellow).

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
Micrococcal nuclease	Sigma	Cat# N3755
IgG Sepharose Fast Flow	Fisher Scientific	Cat# 45000173
AcTEV protease	Life Technologies	Cat # 12575023
Calmodulin Affinity Resin	Agilent	Cat# 214303
DBE	A gift from Dr. Schwer	N/A
bRNA	A gift from Dr. Damha	N/A
SuperScript III Reverse Transcriptase	ThermoFisher	Cat# 18080-044
T4 RNL2Tr.K227Q	NEB	Cat # M0351L
CircLigase	Illumina, Inc	Cat# CL4115K
2x KAPA HIFi Kit	Kapa Biosystems	Cat# KK2611
Experimental Models: Organisms/Strains		
S. pombe: MJM01 (Cdc5-GFP, Lea1-Protein A, Snu114-CBP, ura4-D18leu1-32 h-)	Previous study	(Chen et al., 2014)
S. pombe: MJM04 (dbr1Δ, Cdc5-GFP, Lea1-Protein A, Snu114-CBP, ura4-D18leu1-32 h-)	This study	N/A
Oligonucleotides		
Primer: SPAC5H10.06C Forward: 5'-GTGACCATTCATATTGCTTGG-3'	This study	N/A
Primer: SPAC5H10.06C Reverse: 5'- TTGGCGGCTTCAGCAAG-3'	This study	N/A
Primer: SPBC3E7.04C Forward: 5'-GTACAAATTGTACACAATGAG-3'	This study	N/A
Primer: SPBC3E7.04C Reverse: 5'- GCAAATATCATCGGTTTGTG -3'	This study	N/A
Primer: SPSNORNA.02-0.6 Forward: 5'-GCCAGTATACGGCGTTGTTAG-3'	This study	N/A
Primer: SPSNORNA.02-0.6 Reverse: 5'- GTTGATCGTAAATCCCTTCTGTTC -3'	This study	N/A
Software and Algorithms		
Cutadapt V1.71	Martin (2011)	https://github.com/marcelm/cutadapt/
Fastx Toolkit V0.0.14	http://hannonlab.cshl.edu/fastx_toolkit/	https://github.com/agordon/fastx_toolkit
Bowtie2 V2-2.1.0	Langmead *et al.* (2012)	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
BP Identification by Random Forest Algorithm	This Study	https://github.com/Jill-Moore/SPombe-Splicing/
BAL	This Study	https://github.com/hakanozadam/bal
Other		
Ball Mill	Retsch	PM 100

STAR★METHODS

KEY RESOURCES TABLE (See separate document)

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Melissa Moore (melissa.moore@umassmed.edu).

METHOD DETAILS

S. pombe strains and U2·U5·U6·NTC complex purification

Strain MJM01 carrying Protein A-tagged Lea1 (a U2 snRNP protein), calmodulin binding peptide- (CBP) tagged Snu114 (a U5 snRNP protein) and GFP-tagged Cdc5 (an NTC protein) were as previously described (Chen et al., 2014). Strain MJM04 (Lea1-Protein A, Snu114-CBP, Cdc5-GFP, dbr1Δ, ura4-D18leu1-32 h-) was generated by mating of MJM01 with a dbr1Δ strain from the Bioneer knockout collection, followed by sporulation and PCR validation of haploid derivatives. Large scale (6 L initial cell culture) U2·U5·U6·NTC complex purification under High Salt (HS) conditions and MNase digestions (footprinting) were carried out as previously described (Chen et al., 2014).

RNA purification and deep sequencing library preparation

For all library preparations from purified U2·U5·U6·NTC complex, we isolated RNA from ~40 μg complex (as measured by total protein) using hot acidic phenol, followed by chloroform extraction and ethanol precipitation. Cellular RNA was purified similarly, starting with 10 mL 1xYES culture grown to OD_{600} = 1.5.

Intron RNA-Seq and MNase footprinting libraries

For intron RNA-Seq libraries, 250 ng total RNA from purified U2·U5·U6·NTC complex first subjected to random hydrolysis at 94°C for 3 min in 5× first strand buffer (250 mM Tris-HCl, pH8.3, 375 mM KCl, 15 mM $MgCl_2$) to generate a fragment distribution centered around 60 nts. For MNase footprinting libraries, this initial hydrolysis step was omitted because the RNA had already been fragmented will still associated with U2·U5·U6·NTC complex. Following ethanol precipitation, debranched (DBE+) samples were generated by incubating 250 ng RNA with 100 ng recombinant debranching enzyme (DBE) in 100 μL DBE buffer (50 mM Tris-HCl pH 7.0, 8 mM $MnCl_2$, 2 mM $MgCl_2$, 25 mM NaCl, 0.01% Triton X-100, 2.5 mM DTT, 0.15% glycerol) for 1 hour at 30°C. Reactions were stopped by adding phenol/chloroform, and the debranched RNA was ethanol precipitated. For library preparation, 250 ng of branched (DBE-) or debranched (DBE+) RNA was loaded onto a 1.5 mm thick 8% denaturing polyacrylamide gel. After excision of the region corresponding to the desired fragment length (~10-100 nts), RNA was eluted, phenol extracted and ethanol precipitated. Recovered fragments were then treated with T4 polynucleotide kinase in the absence of ATP to remove 3'-phosphates and subjected to 3'-adaptor ligation, reverse transcription (RT) and cDNA circularization as previously described (Heyer et al., 2015). Barcodes, 3' adaptor and RT primer sequences are shown in **Table S1**.

3′-OH libraries

3'-OH library preparation was similar to intron RNA-Seq library preparation except that samples were not subjected to random hydrolysis and both the size selection and 3'-dephosphorylation steps were omitted.

5'-P libraries

For 5'-P library preparation, unfragmented total RNA samples from cells or purified U2·U5·U6·NTC complex were treated (DBE+) or not (DBE-) with recombinant DBE as above

and then used directly for library construction employing a protocol similar to that described for making degradome libraries (Moran et al., 2014), except that 5' adapter (aka, RNA linker) ligations were carried out for 2 hours at 25°C and ethanol precipitations were substituted for cleanup steps calling for Agencourt RNAClean XP beads. Also the RNA linker (**Table S1**) was modified to contain a region of nine random nucleotides in order to label each captured RNA fragment with a unique molecular identifier (UMI). RT and PCR primer sequences (**Table S1**) were exactly as in (Moran et al., 2014).

Deep sequencing and read mapping

Sequencing specifics (e.g., sequencing platform used, bar codes, read lengths, etc.) and read count statistics are given in **Table S1**. All libraries will be deposited in the NCBI short read archive upon manuscript acceptance. Scripts for all bioinformatics analyses available at: https://github.com/weng-lab/Jill_Moore/tree/master/Moore/Scripts

Using Cutadapt (V1.7.1), we trimmed 5' barcode and 3' adaptor sequences from read pairs. Then using custom python scripts (available on GitHub) or fastx collapse (fastx toolkit V0.0.14), we collapsed reads into non-redundant species and removed the random mers from each species. We then aligned reads to the *S. pombe* genome using Bowtie2 (V2-2.1.0) and default parameters. We selected all uniquely mapping, concordant read pairs for additional analysis. We then mapped the remaining reads to exon-exon junction sequences, selecting those that uniquely map and combining them with the uniquely mapping genomic reads.

Binomial test for significant 5'-P and 3'-OH peaks

Annotated Introns

We tested for significant 5'-P and 3'-OH read-end peaks at annotated intron ends using the binomial test. At annotated 5'SSs, we counted the number of 5'-P reads that started exactly at the 5'SS compared to those staring at any position in a 25 nt downstream window. At annotated 3'SSs, we counted the number of 3'-OH reads that ended exactly at the 3'SS compared to those ending in a 25 nt upstream window. To quantify the amount of splicing intermediate, we also analyzed the number of 3'-OH reads ending one nt upstream of the 5'SS and compared it to the number of reads ending in a 25 nt window centered at the 5'SS.

In each case we tested for significance using a binomial distribution:

$$\mathrm{p}(x) = \binom{n}{x} p^x 1 - p^{n-x}$$

where n is the number of reads in the window, p is 1/length of window, and x is the number of reads at the splice site or -1 position. We applied an FDR correction and selected all peaks with FDR < 0.01.

Genome Wide

To identify significant peaks 5'-P and 3'-OH elsewhere in the genome we counted the number of 5'-P reads starting at and 3'-OH reads ending at every genomic location not previously identified as a 5'SS or 3'SS and compared this number to the number of reads ending in a 25 nt window centered around that position. We calculated the significance of each peak using the aforementioned binomial test, applied an FDR correction, and reported all peaks with FDR < 0.01.

BAL Algorithm

We first mapped libraries against the *S. pombe* genome using the HiSat exon junction aligner to identify genome- and exon junction-mapping reads (Kim et al., 2015). Unmapped reads were then locally aligned using Bowtie2 (Langmead and Salzberg, 2012) against the first 20 nucleotides downstream of every 5'SS. Mapping to a particular 5'SS was confirmed if the 3' remainder (dark green region in **Figure 6B**) mapped to the intron sequence immediately downstream. The 5' remainder (purple region in **Figure 6B**) was then mapped to the downstream intron to identify the BP. To increase depth by allowing for inclusion of shorter reads, we then generated a BP sequence reference list by concatenating the first 100 nts of each intron with the 100 nts upstream of each identified BP. All non-mapped reads were then mapped to this reference using Bowtie2 in the end-to-end alignment mode. B reads were defined as those with read length >22 and mapping quality >5; BP confirmation required a >70% mismatch rate at the BP among all reads mapping to that reference sequence.

Random Forest

Random Forrest Method relies on reads terminating at or crossing BPs in 3'-OH DBE- libraries. For some introns, however, we did not initially observe such reads from the initial genome-mapping pipeline because the BP was so close to the 3'SS (< 15 nts) that these short reads did not map to the genome. To capture these additional short reads that cannot map uniquely when compared to the entire genome, we remapped previously unmapped reads using only introns as our reference. This yielded an additional 1.3 million (increase of ~10%) mapped reads for the 3' OH DBE- libraries.

To create our training, validation, and testing sets, we used 300 BPs from 291 introns that were identified by both BAL and the Pleiss lab (Awan et al., 2013) (**Table S4B**). The identified BPs were our positives and all other "A" nts within these 291 introns were our negatives. We included 192, 49, and 50 introns in the training, validation, and testing sets respectively (which represented 200, 50 and 51 BPs, respectively).

For training, we included eight features in our Random Forest model:

1) Distance from BP to 3'SS
2) % of total intron reads starting at 3'SS and stopping at BP
3) % of total intron reads starting at 3'SS and stopping one nucleotide downstream of BP
4) % of total intron reads starting one nucleotide 5' to BP
5) Mismatch % (mutation and/or indel) at BP for reads starting at 3'SS (i.e., from lariat product)
6) Mismatch % (mutation and/or indel) at BP for reads starting in following exon (i.e. from lariate intermediate)
7) Nucleotide immediately 5' to BP (A, C, G or T)
8) Nucleotide immediately 3' to BP (A, C, G or T)

We ran each model ten times, generating a Random Forest with 100 trees each time. For prediction, we averaged the score across all ten iterations. We then applied the model genome wide, we reported all predictions with probability > 0.5.

To assess the validity of our Random Forest Model, we compared its performance to an algorithm that relied solely on the BP consensus motif. For each BP we calculated its log-odds

score match to the consensus motif using FIMO and ranked predictions from high to low score. To evaluate each method, we calculated the area under the receiver operator characteristic (ROC) and precision recall (PR) curves using the ROCR and PROCR R packages. We also calculated precision and recall using FIMO with Q-value cutoffs of 0.01 and 0.05.

PCR confirmation of new introns

New introns in the SPNCRNA.659 and ADH4 (SPAC5H10.06c) gene loci were confirmed by comparing PCR products derived from genomic DNA or by reverse transcription (RT) of total cellular RNA. Primer sequences are in Key Resources Table. PCR from genomic DNA was performed with GoTaq Green (Promega, Cat. #M7112) following the manufacturer's instructions. RT-PCR was performed using SuperScript III One-Step RT-PCR System with Platinum Taq High Fidelity (Invitrogen, Cat. #12574-035). All detected bands were confirmed by direct Sanger sequencing of gel-purified PCR products.

Primer extension assays

For primer extension analysis of adh4 mRNA isoforms, we used as input 40 μg total RNA isolated from cells grown at 30°C to: (1) OD_{600} = 1.5 in 1×YES medium and harvested immediately (log phase cells); (2) OD_{600} = 1.5 in 1×YES medium, then supplemented with 100 μg/mL cycloheximide and incubated for an additional 2 hours at 30°C (log phase +CHX); or (3) OD_{600} = 23 in 6×YES medium (late log glucose depletion). Samples were annealed with 1 ng 5'-^{32}P labeled Primer A: 5'-GGCTTTCGAAGAGGTTGATTGATG-3' (~300,000 cpm) in Annealing Buffer (5 mM Tris pH 8.3, 75 mM KCl, 1 mM EDTA) by first heating for 1 minute in a boiling water bath and then incubating at 48°C for 45 minutes. Primer extension was performed with SuperScript III RT (Invitrogen) according to manufacturer's instructions. Following ethanol precipitation, pellets were resuspended in 3.5 μL 40 μg/mL RNase A and incubated at room

temp for 3 min to degrade residual RNA. Labeled extension products were then separated on an 8% denaturing polyacrylamide gel.

Debranching assay using labeled branched oligo

Cell extracts and debranching assays were based on previous publications (Chapman and Boeke, 1991) (Khalid, 2005). Briefly, 50 ml of WT or dbr1Δ cells grown in 1xYES to mid-log (OD at 600 nm ~1.0) were harvested by centrifugation, washed once with cold water and resuspended in 400 ul cold buffer A (1 M sorbitol, 50 mM Tris pH 8.0, 10 mM $MgCl_2$, 3 mM DTT) plus 300 ul glass beads. Cells were broken by vigorous shaking on a small ball mill for 4 min at 4°C. After addition of 20 ul Buffer B (300 mM HEPES pH 7.8, 1.4 M KCl, 30 mM MgCl2), extracts were incubated on ice for 20 min and then cell debris removed by centrifugation (13,000 rpm for 5 min at 4°C in a bench top centrifuge). After transferring the supernatant to a 3 ml tube (Beckman Cat. # 362305), a second high speed spin (58,000 rpm for 60 min at 4°C in a Beckman TLA 110 rotor) yielded clarified extract, which was divided into 100 ul aliquots, quick frozen in liquid N_2 and stored at -80°C.

The branched oligo (25 pmol bRNA-AK86 ((Clark et al., 2016); sequence shown in **Figure 4C**) was 3'-end-labeled with T4 RNA ligase (NEB Cat. # M0204S) and 20 μL ^{32}P-pCp (Perkin Elmer # NEG019A250UC) as previously described in (Nilsen, 2014). Preliminary labeling experiments revealed that the ^{32}P-pCp as supplied by the manufacturer contained residual γ-^{32}P-ATP and T4 polynucleotide kinase activity, which led to inadvertent 5'-end labeling of bRNA-AK86 and circle formation during the overnight T4 RNA ligase incubation. To prevent these undesired side reactions, residual T4 polynucleotide kinase was inactivated by incubating ^{32}P-pCp at 65°C for 20 min prior to addition to the T4 RNA ligase reaction. Ligation reactions were carried out

overnight at 4°C and then unincorporated ^{32}P-pCp removed by spinning the reaction through an AM10070 NucAway Spin Column.

For detection of debranching activity, control reactions (20 μl) contained Buffer A (50 mM Tris–HCl pH 7.0, 8 mM $MnCl_2$, 2.5 mM DTT, 25 mM NaCl, 2mM $MgCl_2$, 0.01% Triton X-100, 0.15% glycerol) alone (**Figure 4C**, lane 1) or plus 100 ng purified *S. cerevisae* DBE (Khalid et al., 2005) (lane 2). Cell extract reactions (20 μl) contained Buffer B (50 mM Tris–HCl pH 7.0, 8 mM $MnCl_2$, 2.5 mM DTT, 0.01% Triton X-100, 0.15% glycerol), 700 ng purified total *S. pombe* cellular RNA, 5-7 μl WT (lane 3) or dbr1Δ (lane 4) cell extract (300 μg total protein), and 3 fmol 3'-^{32}P-pCp-labeled bRNA-AK86. All reactions were incubated at 30°C for 1 hour and then an aliquot resolved by 20% denaturing PAGE.

SUPPLEMENTAL INFORMATION

Figure S1. Mapped and unmapped read lengths, expression levels of intron-containing genes with or without detectable 5'-P and 3'-OH peaks, and 5'-P and 3-OH coverage on U6 snRNA, Related to Figure 1

Figure S2. Examples of corrected intron annotations resulting from 5'-P and 3'-OH libraries, Related to Figure 2

Figure S3. Comparison of A content, exon junction reads, GC content and minimum folding energy for high and low splicing intermediate introns, Related to Figure 3

Figure S4. 5'-P peak correlation between DBE- and DBE+ 5'-P libraries, Related to Figure 4

Figure S5. Comparison of Random Forest Method to other BP identification methods, Related to Figure 7

Table S1: Libraries used in this study, Related to All Figures

Table S2: New and previously misannotated introns, Related to Figure 2

Table S3: High splicing intermediate introns, Related to Figure 3

Table S4: Identified branch points, Related to Figure 6 and 7

Supplemental figure legends

Figure S1. Mapped and unmapped read lengths, expression levels of intron-containing genes with or without detectable 5'-P and 3'-OH peaks, and 5'-P and 3-OH coverage on U6 snRNA, Related to Figure 1

(A) Distribution of read lengths for mapped (green) and unmapped (gray) reads from the intron RNA-Seq libraries.

(B) Violin plots comparing gene expression of genes with introns covered by intron RNA-Seq (green), 5'-P (red), or 3'-OH (blue) libraries. Gene expression was calculated from Broad polyA+ RNA-Seq libraries (ERR135906 and ERR135907; https://www.ncbi.nlm.nih.gov/pubmed/23101633) (Rhind et al., 2011)

(C) Aggregation plot of all WT DBE+ 5'-P 5' read ends (red) and 3'-OH 3' read ends (blue) across 5' and 3' splice sites of the U6 intron.

Figure S2. Examples of corrected intron annotations resulting from 5'-P and 3'-OH libraries, Related to Figure 2

(A) Zero-length exon and recursive splicing in SPBC1703.10 gene. Top: SPBC1703.10 introns showing 5'-P and 3'-OH read ends, polyA+ RNA-Seq reads and sequence conservation across four *Schizosaccharomyces* species (Rhind et al., 2011). Bottom: Expanded view of SPBC1703.10 intron 1 showing zero-length exon and polyA+ RNA-Seq reads supporting recursive splicing.

(B) Corrected intron on prp10 (SPAC27F1.09c). Top: intron RNA-Seq read distribution across prp10 intron 2. Middle: 5'-P (red) and 3'-OH (blue) read ends from a single biological replicate (read numbers in brackets) revealed a shorter intron (46 nt instead of 66 nt) also

supported by RNA-Seq data (Rhind et al., 2011). Bottom: Previous ASM294v2.35 intron annotation.

(C) Five new introns in snoRNA host gene SPNCRNA.659. Top: ASM294v2.35 annotation of SPNCRNA.659 gene, 5'-P (red) and 3'-OH (blue) read ends from a single biological replicate (read numbers in brackets) and positions of five new introns. Bottom left: SYBR gold-stained polyacryamide gel of PCR products (U: unspliced; S: spliced) from genomic DNA (G) and cDNA (C). Bottom right: New intron lengths and their of 5'SS, BS, and 3'SS consensus sequences.

Figure S3. Comparison of A content, exon junction reads, GC content and minimum folding energy for high and low splicing intermediate introns, Related to Figure 3

(A) Histograms and density plot comparing the fraction of exon junction reads from spliceosome footprinting data between introns with high (>17%) splicing intermediate to all other introns. (P-values: Wilcoxon rank sum test).

(B) Violin plots comparing of fraction of exon junction reads for the top 200 introns with highest level of splicing intermediate to the 200 introns with lowest splicing intermediate. P-values: Wilcoxon rank sum test.

(C) Histograms and density plot comparing A and T content of top 200 introns with highest levels of splicing intermediate to 200 introns with lowest levels of splicing (P-values: Wilcoxon rank sum test).

(D) Violin plots comparing GC content of introns and exons (upstream, -1 to -40 nts from 5'SS; downstream; +1 to +40 nts from 3'SS) for sets described in B. P-values: Wilcoxon rank sum test.

(E) Histograms and density plot comparing A and T content of introns with high (>17%) splicing intermediate to all other introns in the same gene (gene matched introns) (P-values:

Wilcoxon rank sum test).

(F) Violin plots comparing folding free energy (RNAFold) for upstream and downstream exon sequences for sets described in B. P-values: Wilcoxon rank sum test.

Figure S4. 5'-P peak correlation between DBE- and DBE+ 5'-P libraries, Related to Figure 4

(A) Scatterplot of normalized (reads per million averaged between replicates) read numbers of 5'-P reads at the 5'SS of introns between WT DBE- and DBE+ 5'-P libraries. In red (N=2,483) are introns with a significant peak in both replicates.

(B) Scatterplot of normalized (reads per million averaged between replicates) read numbers of 5'-P reads at the 5'SS of introns between dbr1Δ DBE- and DBE+ 5'-P libraries. In red (N=472) are introns with a significant peak in both replicates.

Figure S5. Comparison of Random Forest Method to other BP identification methods, Related to Figure 7

(A) Receiver Operator Characteristic (ROC) curve for Random Forest Model (black) and motif consensus method (gray).

(B) Precision Recall (PR) curve for Random Forest Model (black) and motif consensus method (gray).

(C) Histograms of the number of BPs identified per intron from 2D intron lariat purification (yellow), LaSSO (pink), BAL (blue) and the Random Forest Model (blue).

(D) Box plots of distance from BPs identified by each method indicated to the overlapping intron's annotated 3'SS.

(E) Sequence logos for 11 nt windows centered at BPs identified by each method.

(F) Violin plots of intron length for introns with BPs identified by each method indicated. All methods except for the Random Forest Model are biased toward identifying BPs in longer introns (p-values shown from Wilcoxon rank sum test compared to background of all introns).

Supplemental table legends

Table S1: Libraries used in this study, Related to All Figures

A. Library specifics and raw read numbers
B. Transcript class read mapping statistics for intron RNA-Seq and footprinting libraries (Round I).
C. Transcript class read mapping statistics for intron RNA-Seq and footprinting libraries (Round II).
D. Transcript class read mapping statistics for 5'-P libraries.
E. Transcript class read mapping statistics for 3'-OH libraries.

Table S2: New and previously misannotated introns, Related to Figure 2

A. Previously misannotated introns in CDS genes (17)
B. New intron-containing transcripts (20)
C. New introns in noncoding RNA genes (126)
D. New introns in CDS genes (85)

Table S3: High splicing intermediate introns, Related to Figure 3

A. Introns with high splicing intermediate
B. Gene ontology (Top 100)
C. Gene ontology (Top 150)

D. Gene ontology (Top 200)

Table S4: Identified branch points, Related to Figure 6 and 7

A. BPs identified by BAL

B. BPs included in Training, Validation, and Test Sets for Random Forest Model

C. BPs identified by Random Forest Model

Figure 1

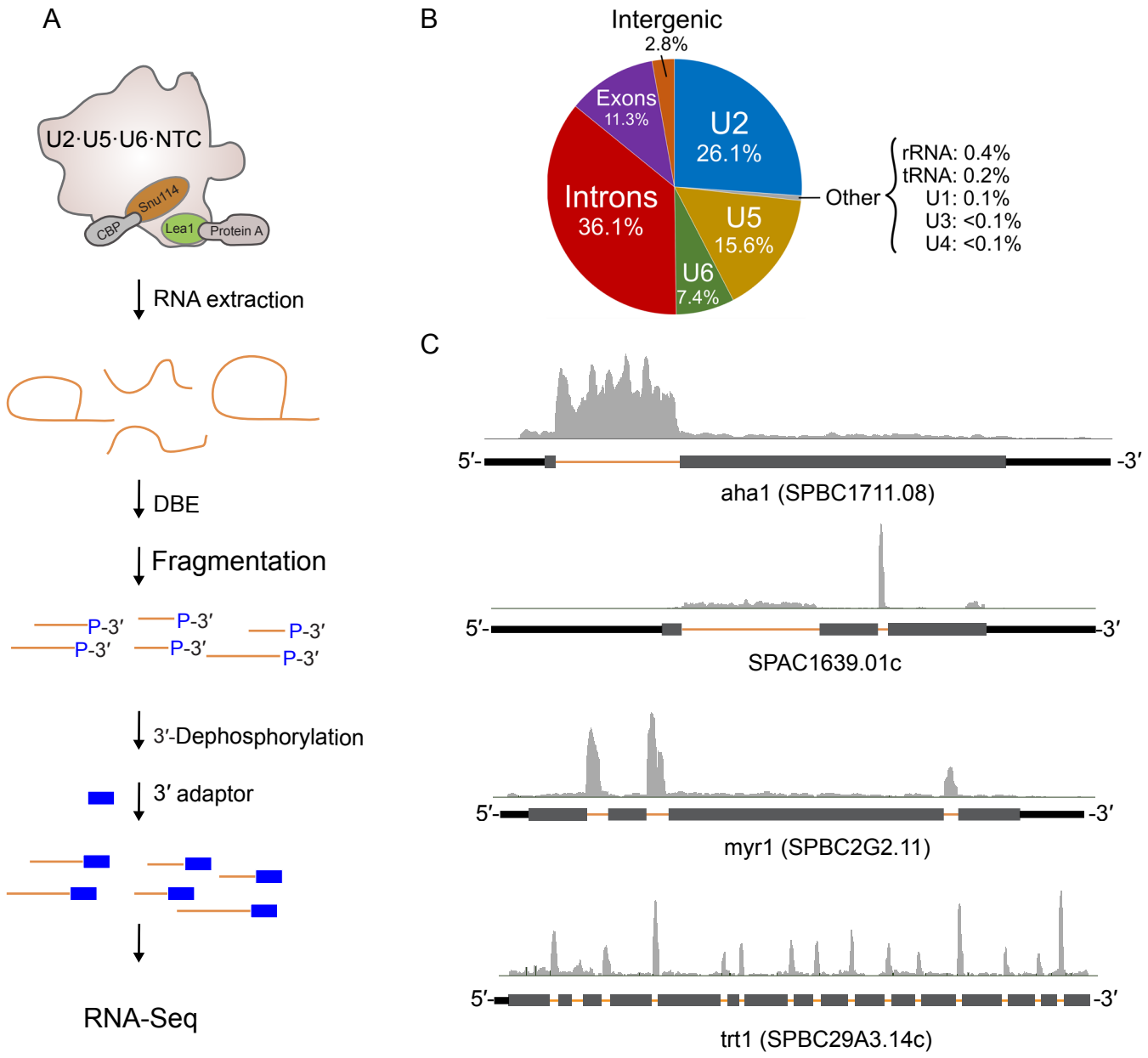

A
U2·U5·U6·NTC
Snu114
CBP
Lea1
Protein A
RNA extraction
DBE
Fragmentation
P-3′
3′-Dephosphorylation
3′ adaptor
RNA-Seq
B
Intergenic
2.8%
Exons
11.3%
U2
26.1%
Introns
36.1%
U5
15.6%
U6
7.4%
Other
rRNA: 0.4%
tRNA: 0.2%
U1: 0.1%
U3: <0.1%
U4: <0.1%
C
5′-
-3′
aha1 (SPBC1711.08)
SPAC1639.01c
myr1 (SPBC2G2.11)
trt1 (SPBC29A3.14c)

Figure 2

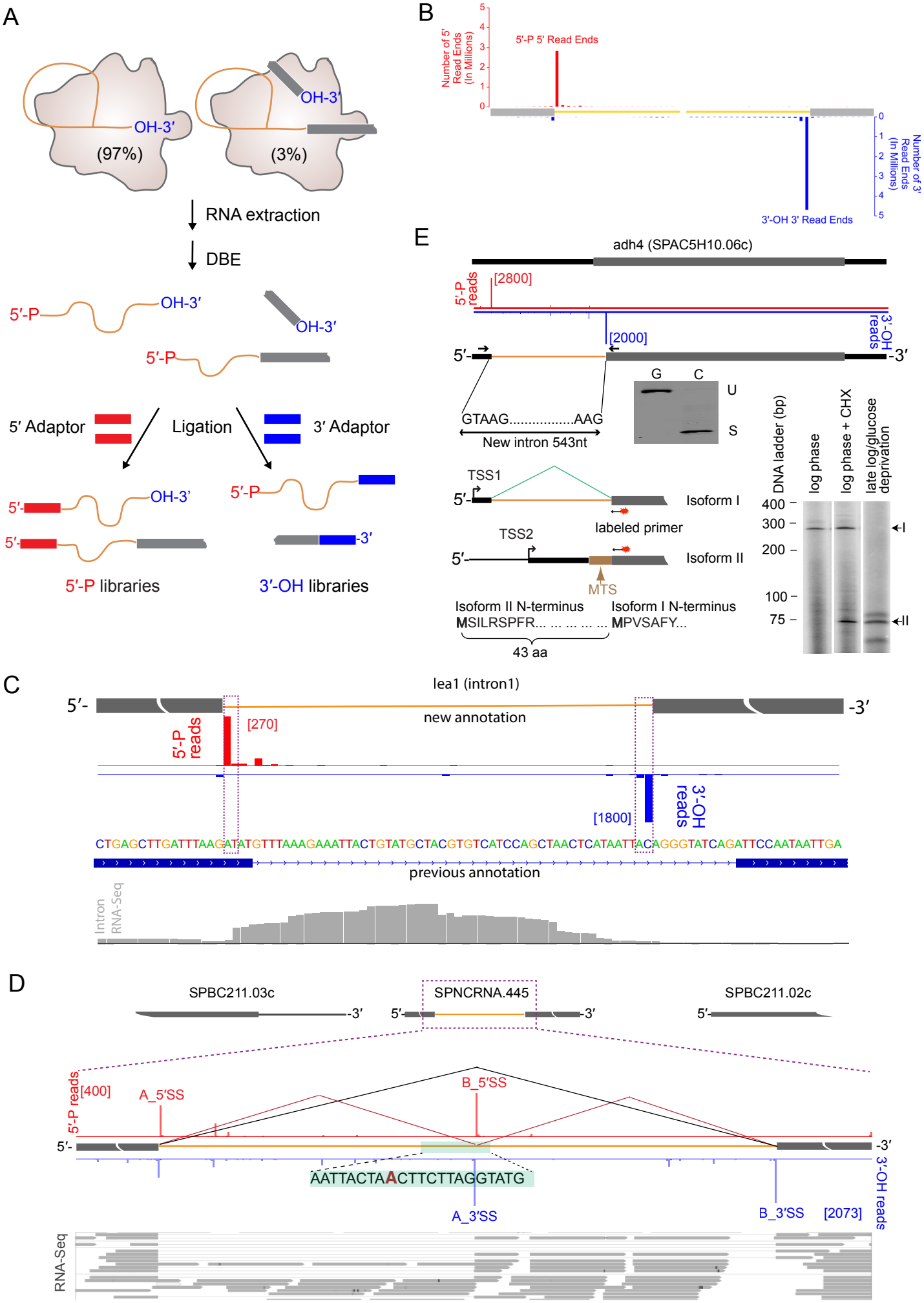

A
OH-3′
(97%)
OH-3′
(3%)
RNA extraction
DBE
5′-P
OH-3′
OH-3′
5′-P
5′ Adaptor
Ligation
3′ Adaptor
5′
OH-3′
5′-P
5′
-3′
5′-P libraries
3′-OH libraries
B
Number of 5′ Read Ends (In Millions)
5′-P 5′ Read Ends
3′-OH 3′ Read Ends
Number of 3′ Read Ends (In Millions)
E
adh4 (SPAC5H10.06c)
5′-P reads
[2800]
[2000]
3′-OH reads
5′-
-3′
GTAAG.....AAG
New intron 543nt
G
C
U
S
TSS1
5′-
Isoform I
labeled primer
TSS2
5′-
Isoform II
MTS
Isoform II N-terminus
MSILRSPFR... ... ...
Isoform I N-terminus
MPVSAFY...
43 aa
DNA ladder (bp)
log phase
log phase + CHX
late log/glucose deprivation
400
300
200
100
75
I
II
C
lea1 (intron1)
5′-
-3′
new annotation
5′-P reads
[270]
[1800]
3′-OH reads
CTGAGCTTGATTTAAGATATGTTTAAAGAAATTACTGTATGCTACGTGTCATCCAGCTAACTCATAATTACAGGGTATCAGATTCCAATAATTGA
previous annotation
Intron RNA-Seq
D
SPBC211.03c
-3′
SPNCRNA.445
5′-
-3′
SPBC211.02c
5′-
5′-P reads
[400]
A_5′SS
B_5′SS
5′-
-3′
AATTACTAACTTCTTAGGTATG
A_3′SS
B_3′SS
[2073]
3′-OH reads
RNA-Seq

Figure 3

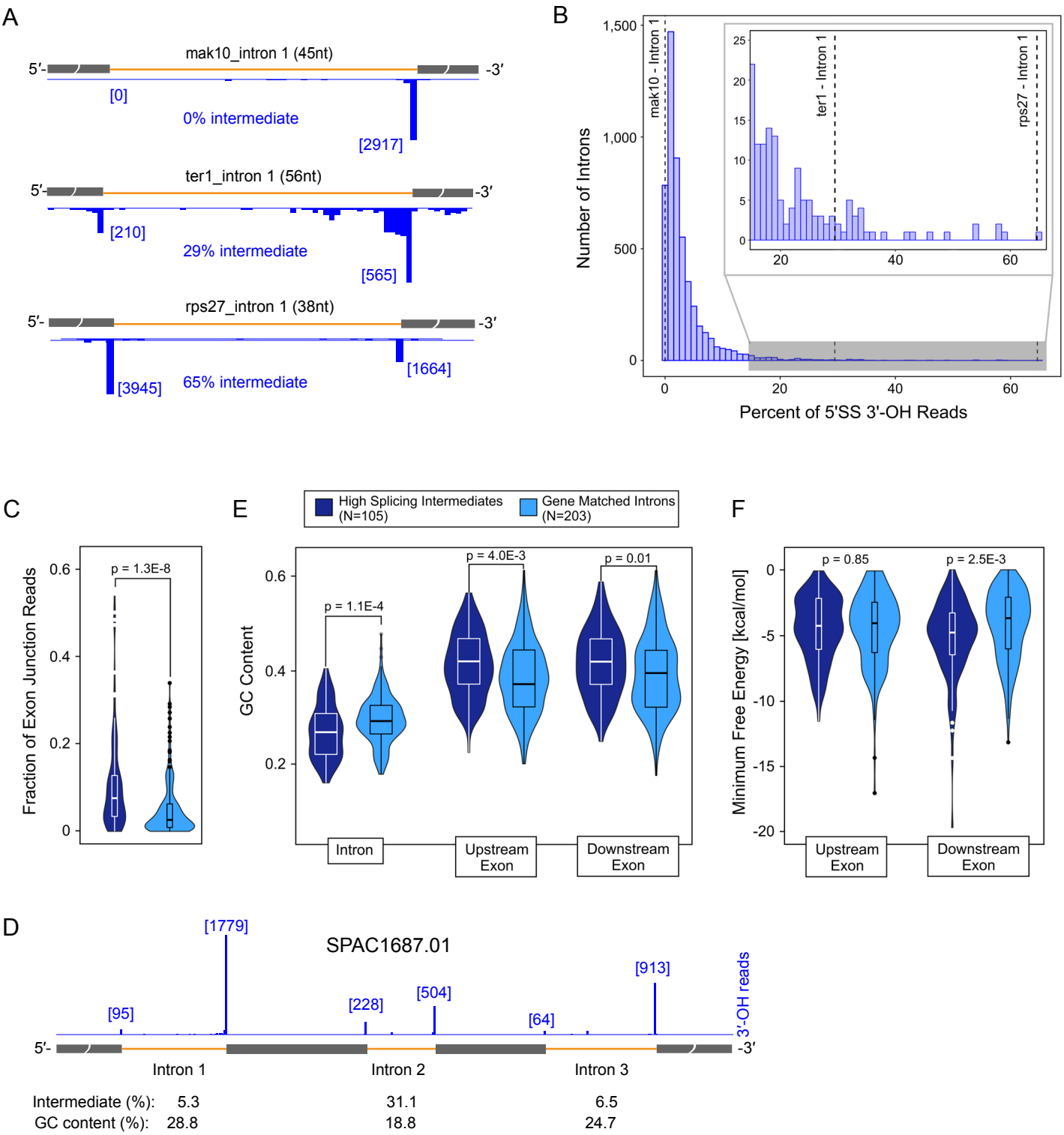

A
mak10_intron 1 (45nt)
5′-
-3′
[0]
0% intermediate
[2917]
ter1_intron 1 (56nt)
[210]
29% intermediate
[565]
rps27_intron 1 (38nt)
[3945]
65% intermediate
[1664]
B
Number of Introns
Percent of 5'SS 3'-OH Reads
mak10 - Intron 1
ter1 - Intron 1
rps27 - Intron 1
C
Fraction of Exon Junction Reads
p = 1.3E-8
E
High Splicing Intermediates (N=105)
Gene Matched Introns (N=203)
GC Content
p = 1.1E-4
p = 4.0E-3
p = 0.01
Intron
Upstream Exon
Downstream Exon
F
Minimum Free Energy [kcal/mol]
p = 0.85
p = 2.5E-3
Upstream Exon
Downstream Exon
D
SPAC1687.01
[95]
[1779]
[228]
[504]
[64]
[913]
3′-OH reads
Intron 1
Intron 2
Intron 3
Intermediate (%): 5.3 31.1 6.5
GC content (%): 28.8 18.8 24.7

Figure 4

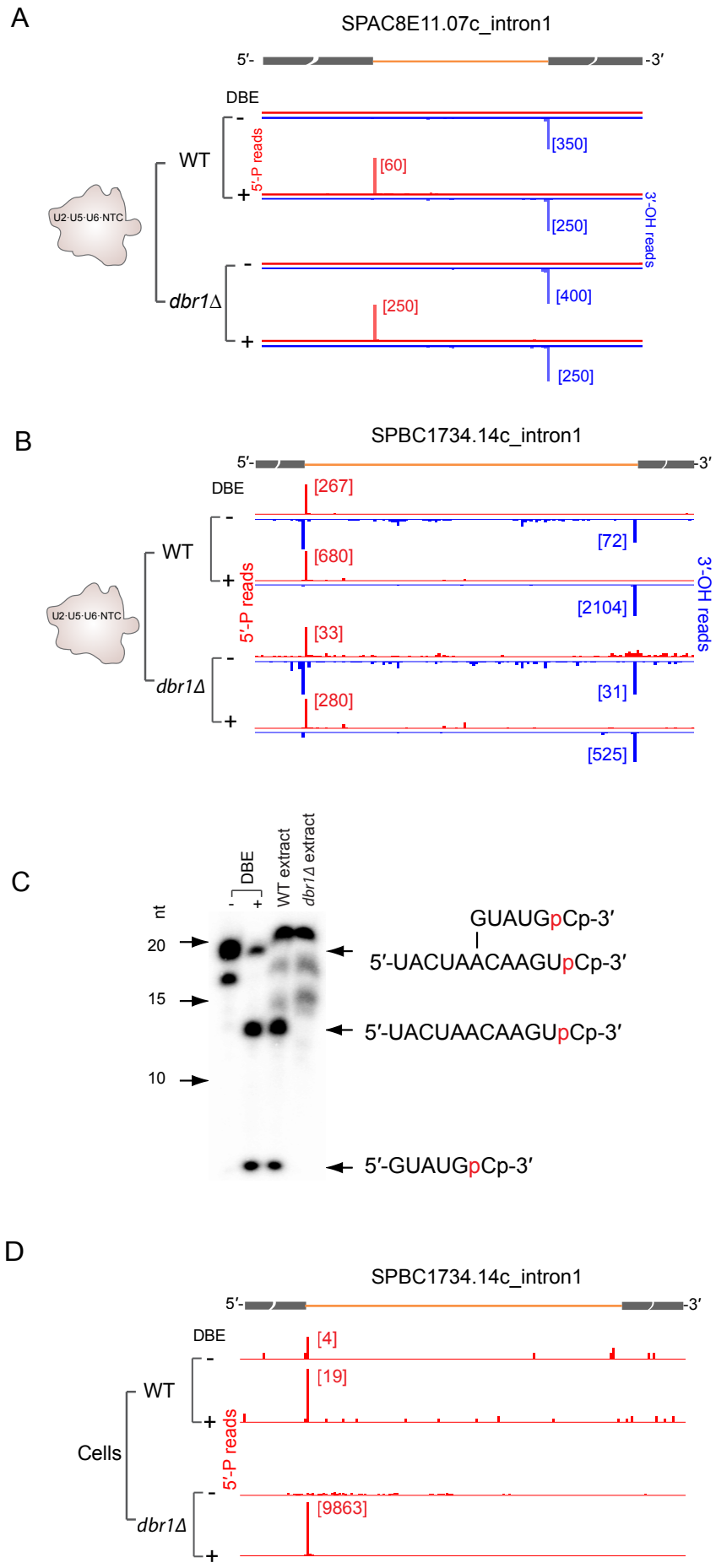

A
SPAC8E11.07c_intron1
5′-
-3′
DBE
WT
dbr1Δ
U2·U5·U6·NTC
5′-P reads
3′-OH reads
[350]
[60]
[250]
[400]
[250]
[250]
B
SPBC1734.14c_intron1
[267]
[72]
[680]
[2104]
[33]
[31]
[280]
[525]
C
DBE
WT extract
dbr1Δ extract
nt
20
15
10
GUAUGpCp-3′
5′-UACUAACAAGUpCp-3′
5′-UACUAACAAGUpCp-3′
5′-GUAUGpCp-3′
D
SPBC1734.14c_intron1
Cells
[4]
[19]
[9863]

Figure 5

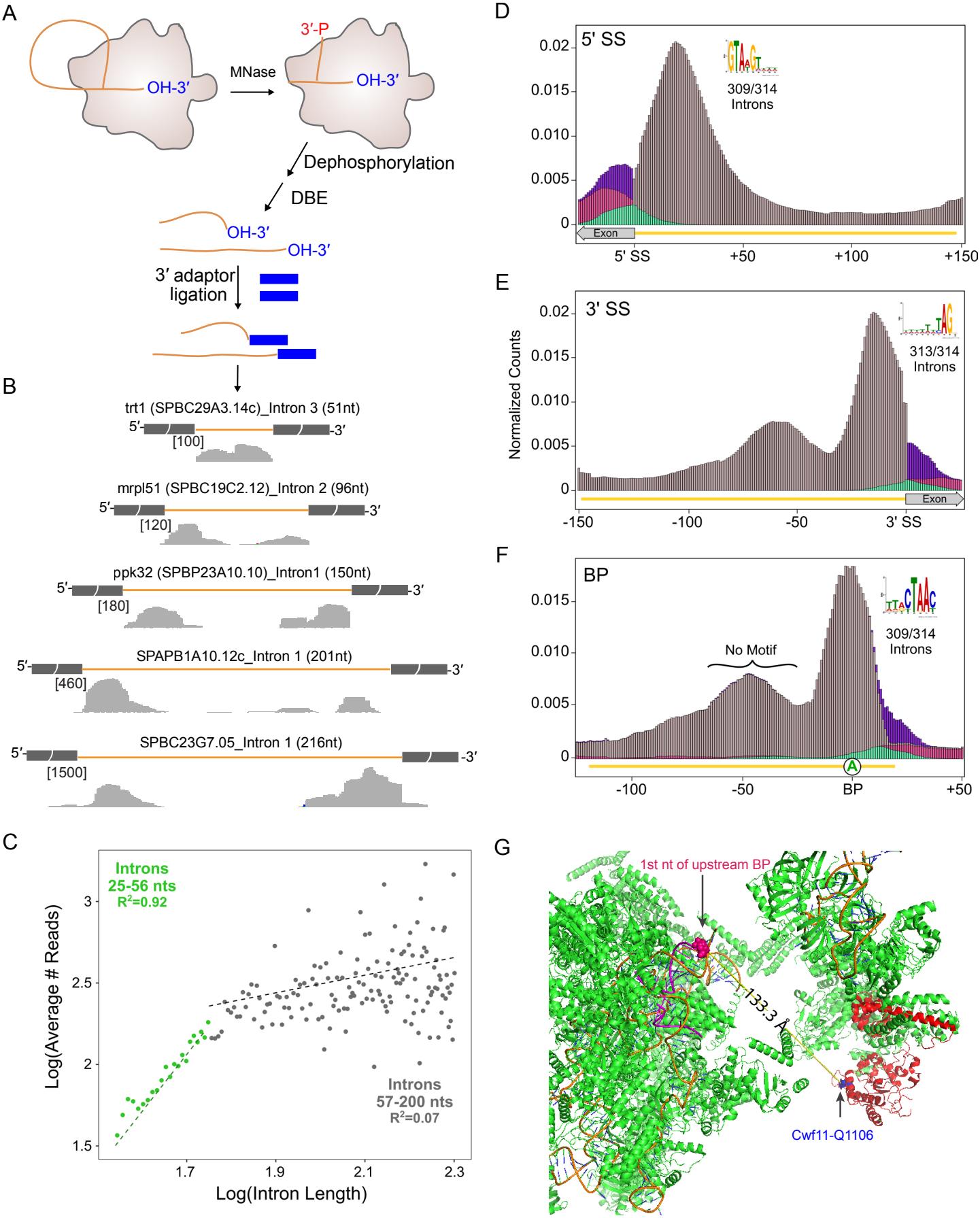

A
3′-P
OH-3′
MNase
OH-3′
Dephosphorylation
DBE
OH-3′
OH-3′
3′ adaptor ligation
B
trt1 (SPBC29A3.14c)_Intron 3 (51nt)
5′
-3′
[100]
mrpl51 (SPBC19C2.12)_Intron 2 (96nt)
5′
-3′
[120]
ppk32 (SPBP23A10.10)_Intron1 (150nt)
5′
-3′
[180]
SPAPB1A10.12c_Intron 1 (201nt)
5′
-3′
[460]
SPBC23G7.05_Intron 1 (216nt)
5′
-3′
[1500]
C
Introns
25-56 nts
R²=0.92
Introns
57-200 nts
R²=0.07
Log(Average # Reads)
Log(Intron Length)
D
5' SS
309/314
Introns
Exon
5' SS
+50
+100
+150
E
3' SS
313/314
Introns
Normalized Counts
Exon
-150
-100
-50
3' SS
F
BP
No Motif
309/314
Introns
A
-100
-50
BP
+50
G
1st nt of upstream BP
133.3 Å
Cwf11-Q1106

Figure 6

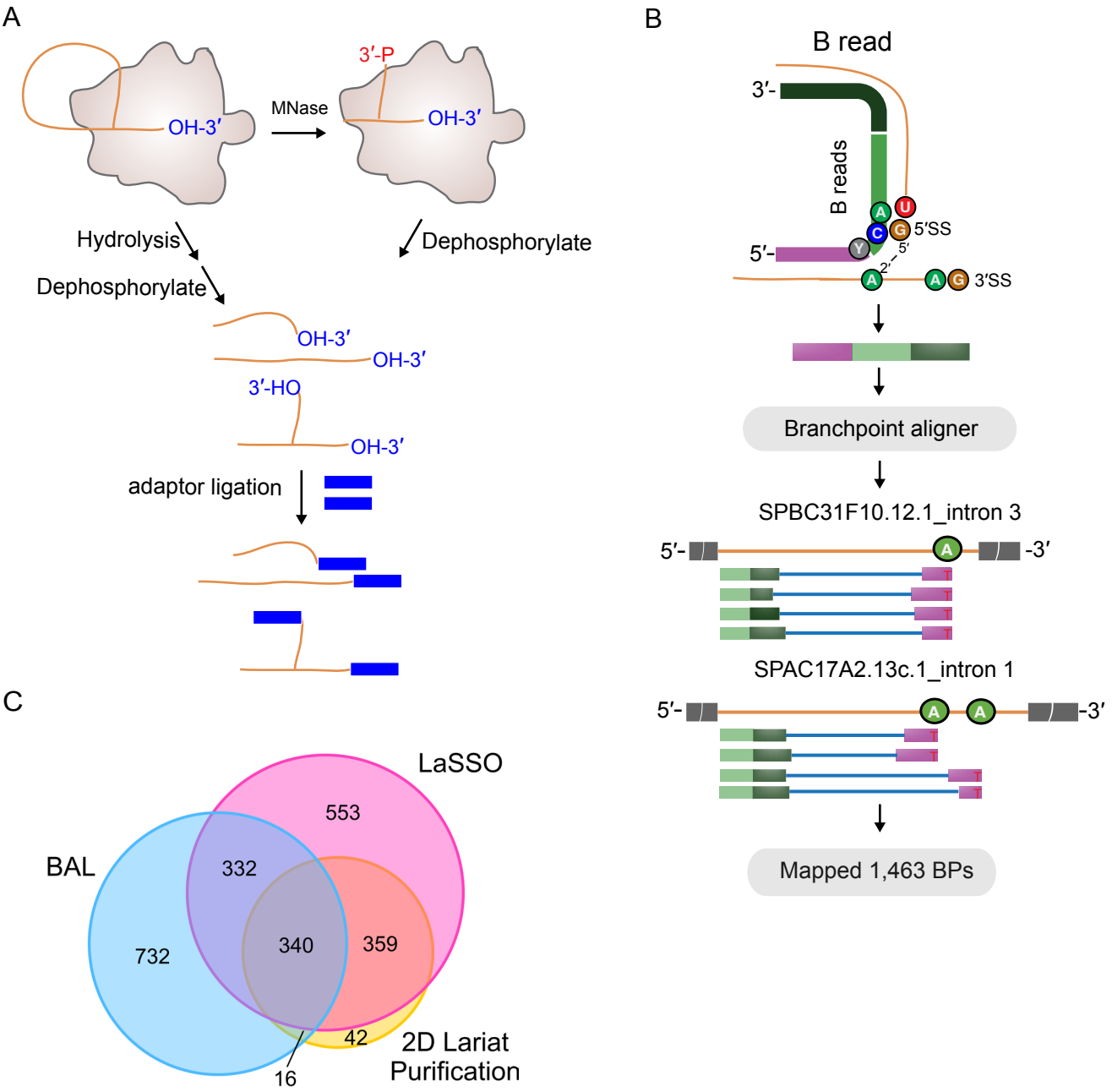

A
OH-3′
MNase
3′-P
OH-3′
Hydrolysis
Dephosphorylate
Dephosphorylate
OH-3′
OH-3′
3′-HO
OH-3′
adaptor ligation
B
B read
3′-
B reads
5′SS
5′-
5′
2′
3′SS
Branchpoint aligner
SPBC31F10.12.1_intron 3
5′-
-3′
SPAC17A2.13c.1_intron 1
5′-
-3′
Mapped 1,463 BPs
C
LaSSO
553
BAL
332
732
340
359
42
16
2D Lariat
Purification

Figure 7

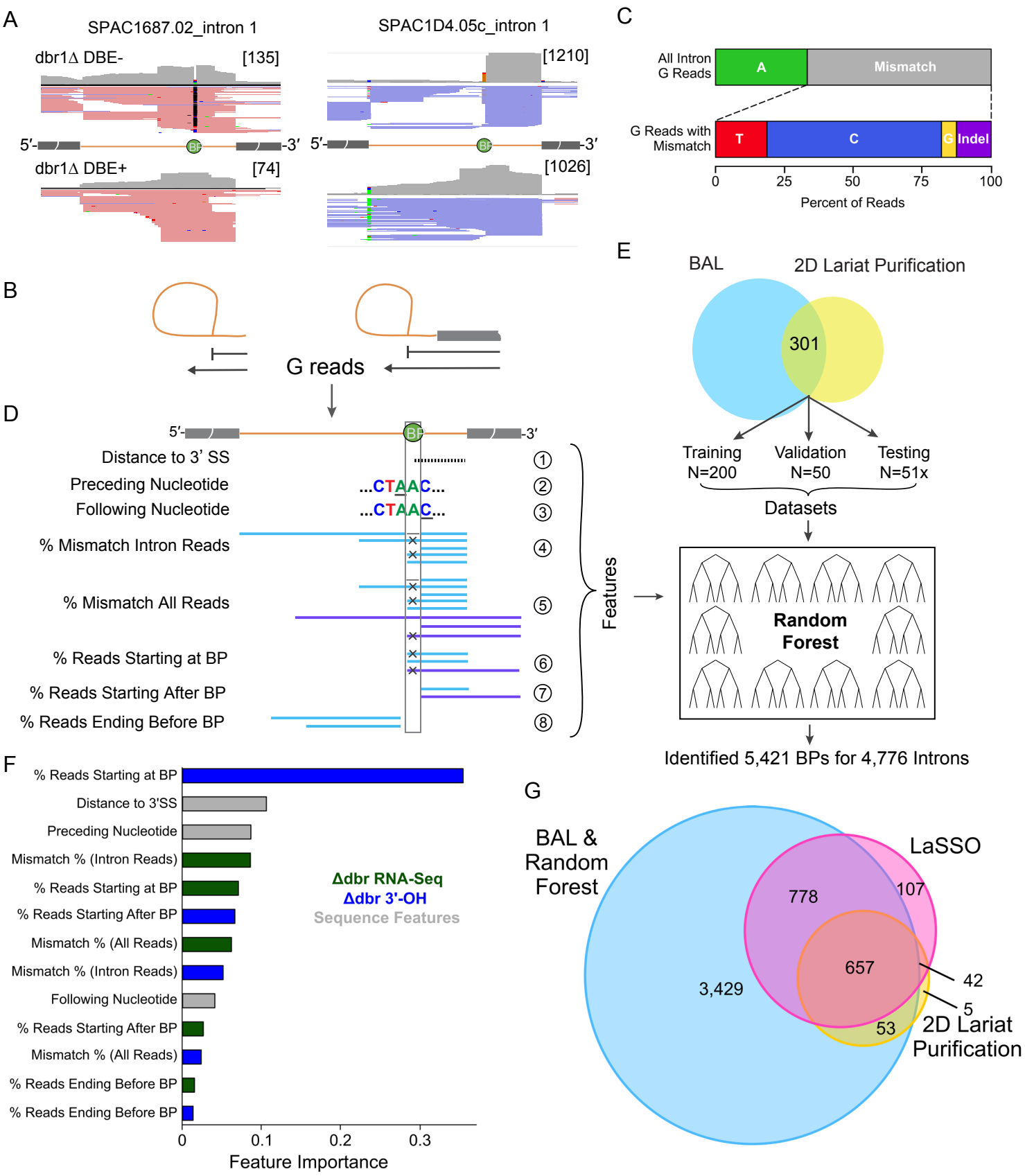

A
SPAC1687.02_intron 1
SPAC1D4.05c_intron 1
dbr1Δ DBE-
[135]
[1210]
5′-
BP
-3′
dbr1Δ DBE+
[74]
[1026]
B
G reads
C
All Intron G Reads
A
Mismatch
G Reads with Mismatch
T
C
G
Indel
0
25
50
75
100
Percent of Reads
D
Distance to 3' SS
Preceding Nucleotide
Following Nucleotide
...CTAAC...
% Mismatch Intron Reads
% Mismatch All Reads
% Reads Starting at BP
% Reads Starting After BP
% Reads Ending Before BP
Features
E
BAL
2D Lariat Purification
301
Training N=200
Validation N=50
Testing N=51x
Datasets
Random Forest
Identified 5,421 BPs for 4,776 Introns
F
Δdbr RNA-Seq
Δdbr 3'-OH
Sequence Features
Mismatch % (Intron Reads)
Mismatch % (All Reads)
Following Nucleotide
0
0.1
0.2
0.3
Feature Importance
G
BAL & Random Forest
LaSSO
2D Lariat Purification
778
107
657
42
5
53
3,429

Figure S1

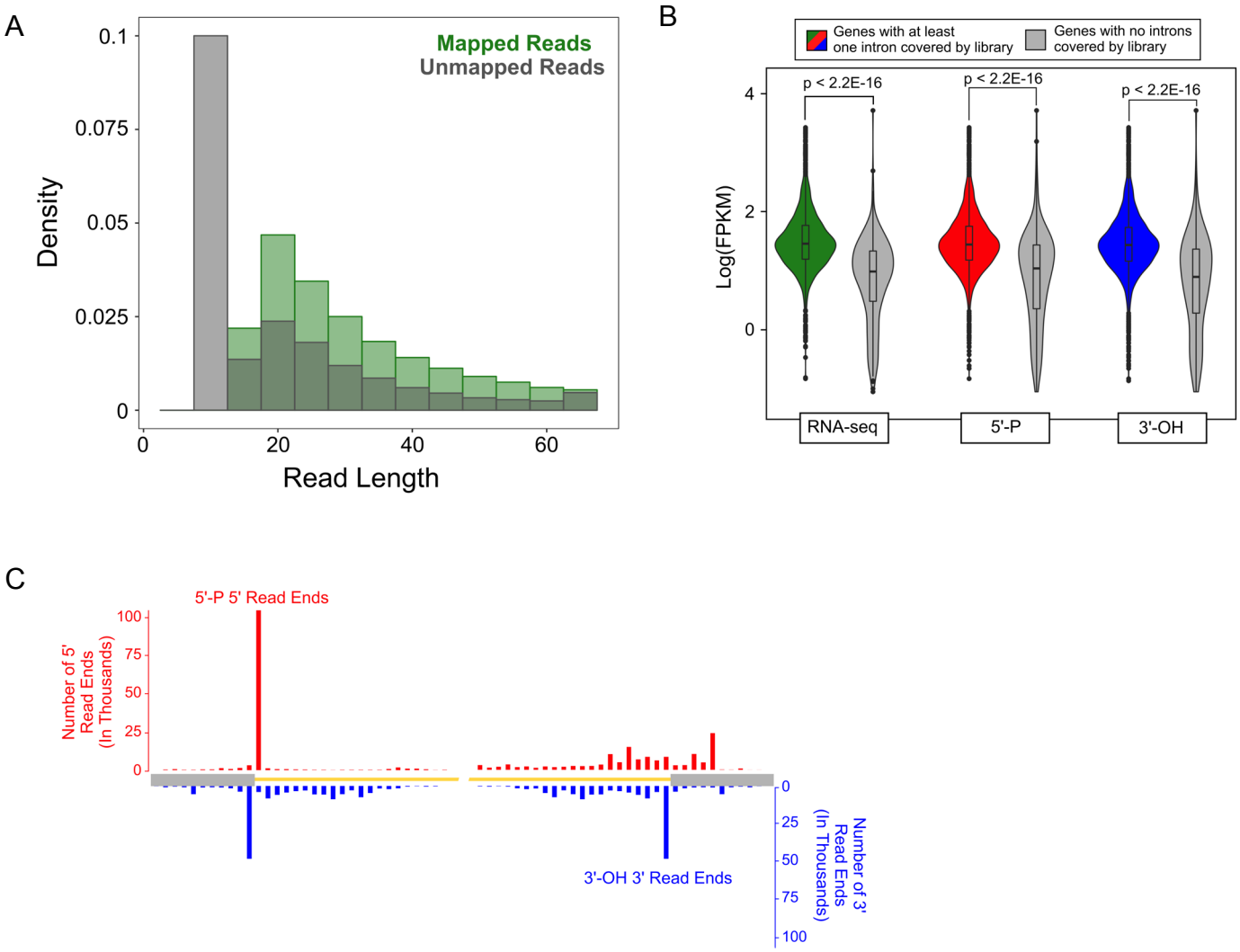
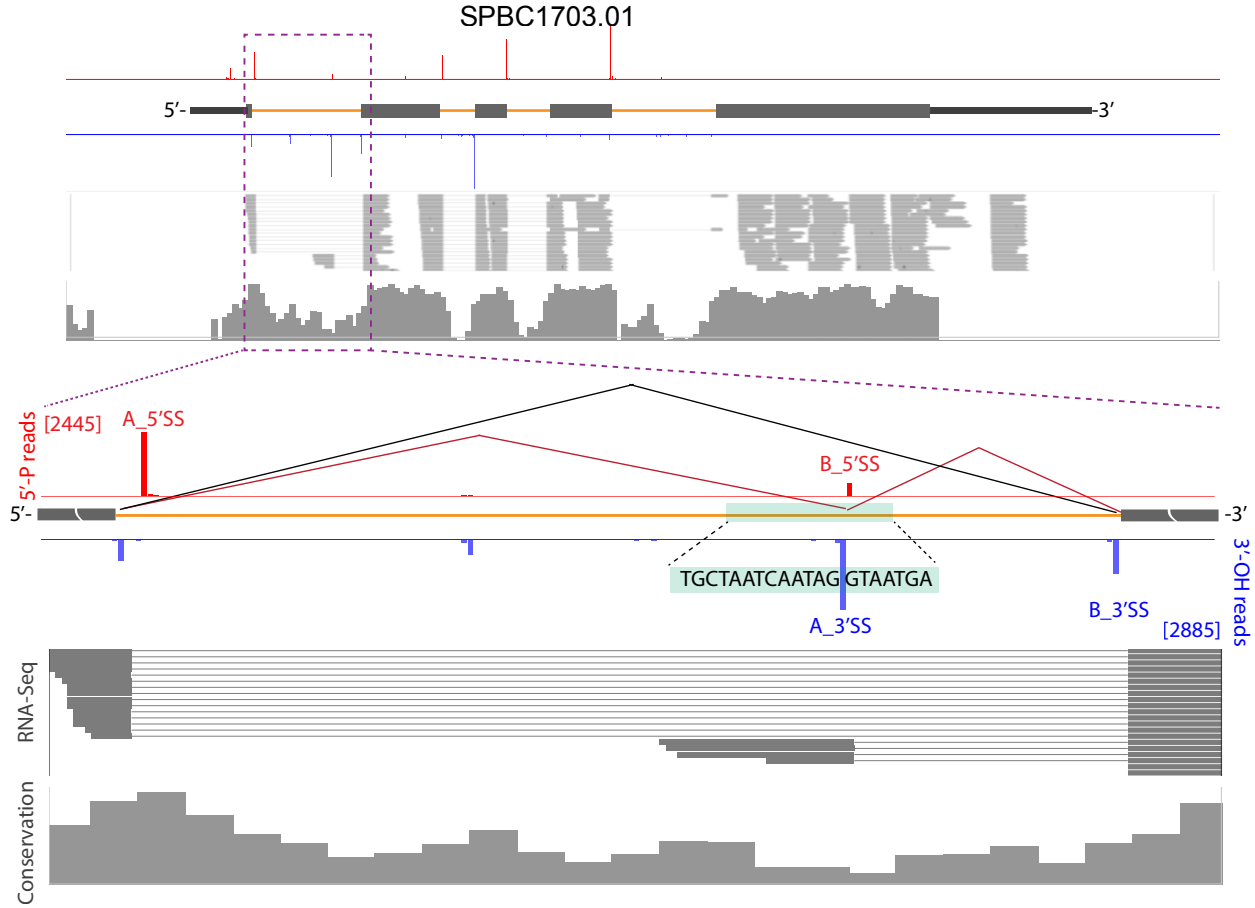
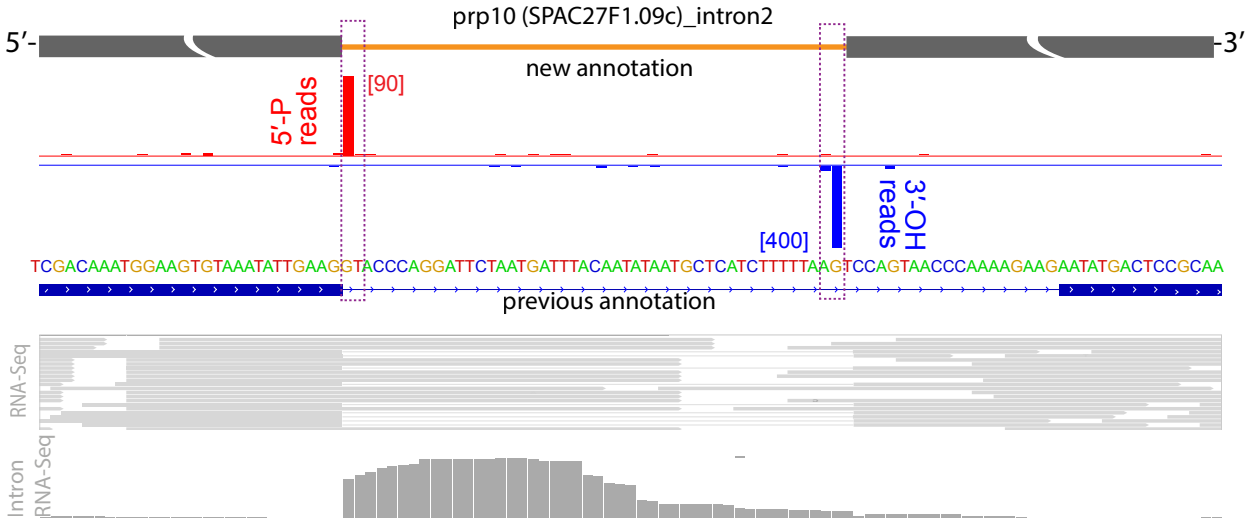

A
Mapped Reads
Unmapped Reads
Density
0.1
0.075
0.05
0.025
0
0
20
40
60
Read Length
B
Genes with at least one intron covered by library
Genes with no introns covered by library
p < 2.2E-16
p < 2.2E-16
p < 2.2E-16
Log(FPKM)
4
2
0
RNA-seq
5'-P
3'-OH
C
5'-P 5' Read Ends
Number of 5' Read Ends (In Thousands)
100
75
50
25
0
3'-OH 3' Read Ends
Number of 3' Read Ends (In Thousands)
0
25
50
75
100

Figure S2

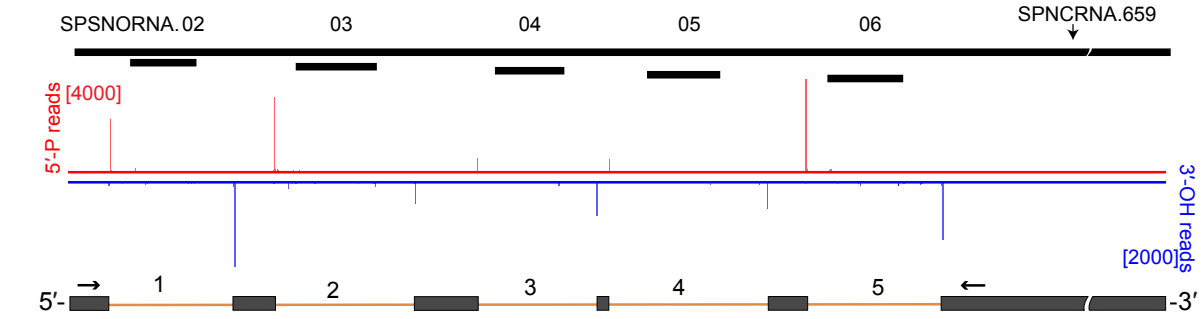
A

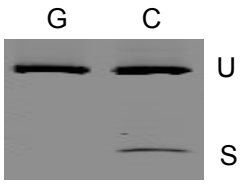


B



C





Intron #:	nt	5′ SS	BS	3′ SS
1:	172	GTATG...	CTAAC...	TAG
2:	183	GTAAG...	TTAAC...	TAG
3:	156	GTATG...	CTAAC...	TAG
4:	206	GTATG...	CTAAC...	AAG
5:	177	GTAAG...	TTAAC...	TAG

Figure S3

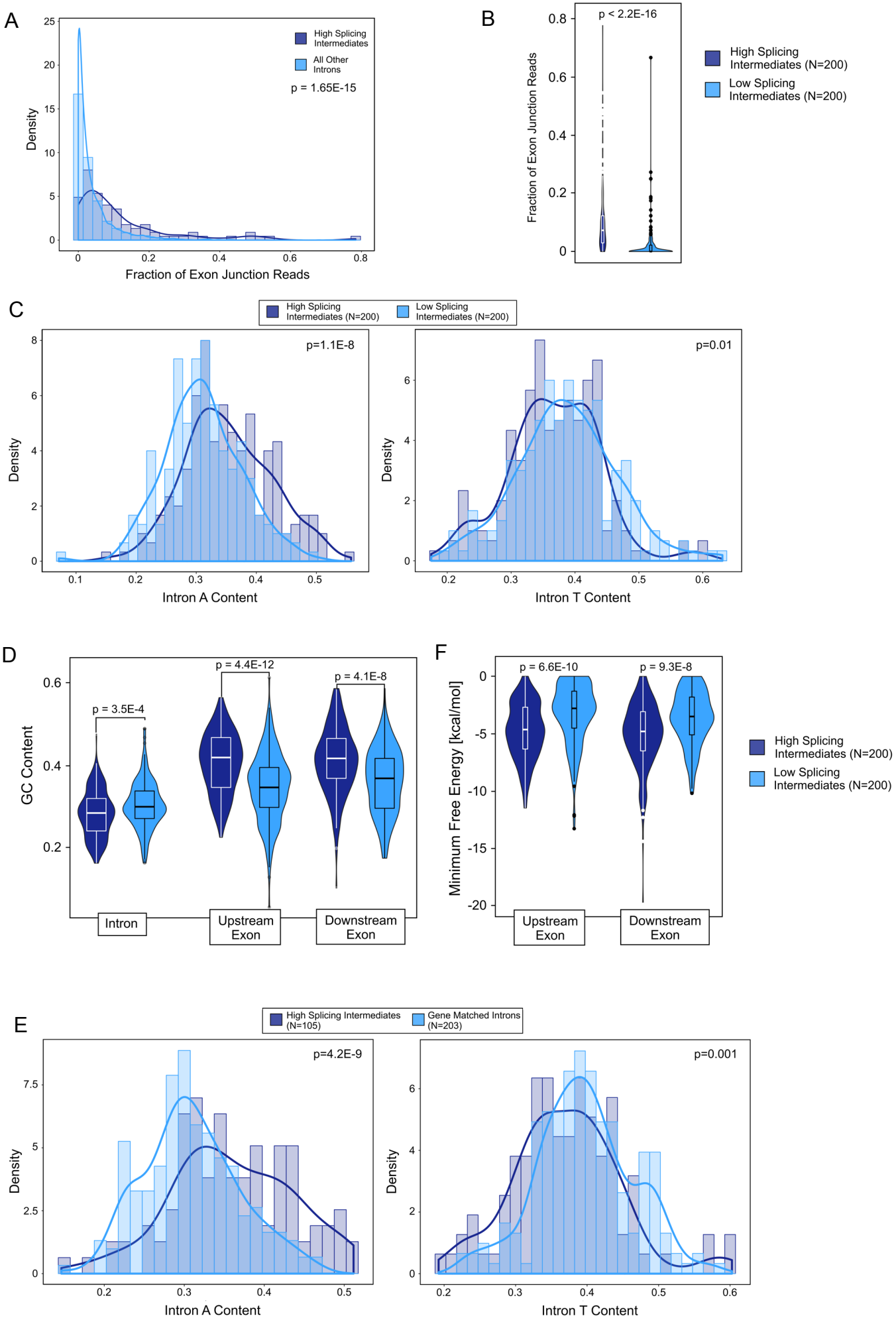
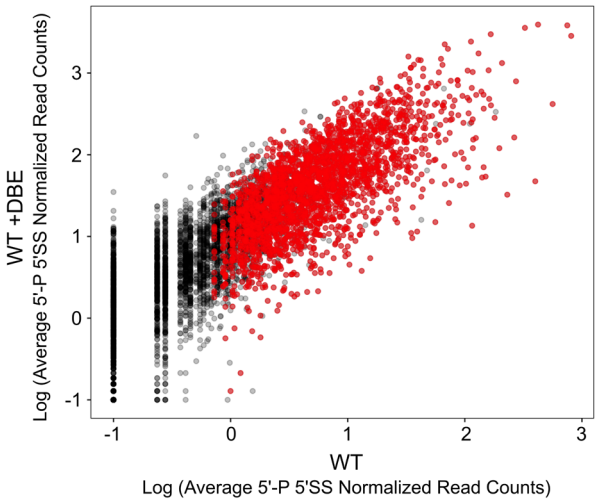

A
High Splicing Intermediates
All Other Introns
p = 1.65E-15
Density
Fraction of Exon Junction Reads
B
p < 2.2E-16
Fraction of Exon Junction Reads
High Splicing Intermediates (N=200)
Low Splicing Intermediates (N=200)
C
High Splicing Intermediates (N=200)
Low Splicing Intermediates (N=200)
p=1.1E-8
p=0.01
Density
Intron A Content
Intron T Content
D
p = 3.5E-4
p = 4.4E-12
p = 4.1E-8
GC Content
Intron
Upstream Exon
Downstream Exon
F
p = 6.6E-10
p = 9.3E-8
Minimum Free Energy [kcal/mol]
Upstream Exon
Downstream Exon
High Splicing Intermediates (N=200)
Low Splicing Intermediates (N=200)
E
High Splicing Intermediates (N=105)
Gene Matched Introns (N=203)
p=4.2E-9
p=0.001
Density
Intron A Content
Intron T Content

Figure S4

A


WT +DBE
Log (Average 5'-P 5'SS Normalized Read Counts)
WT
Log (Average 5'-P 5'SS Normalized Read Counts)
-1
0
1
2
3

B

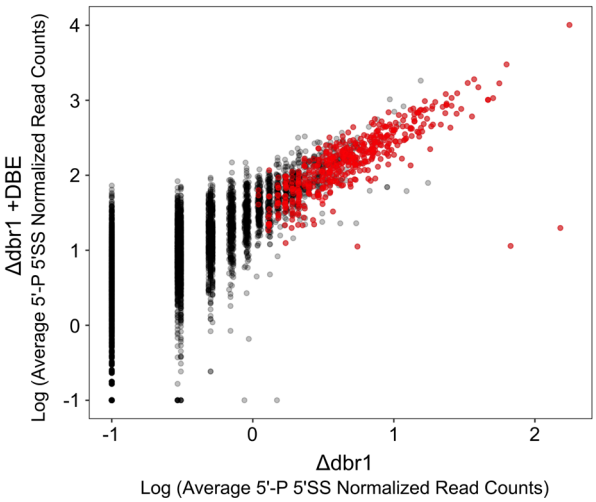
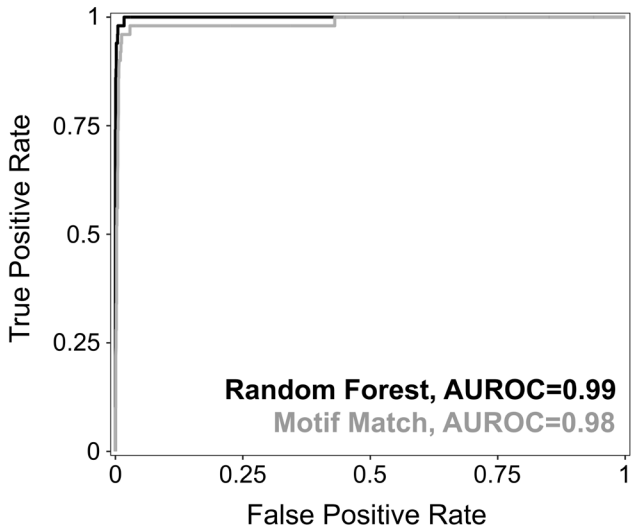
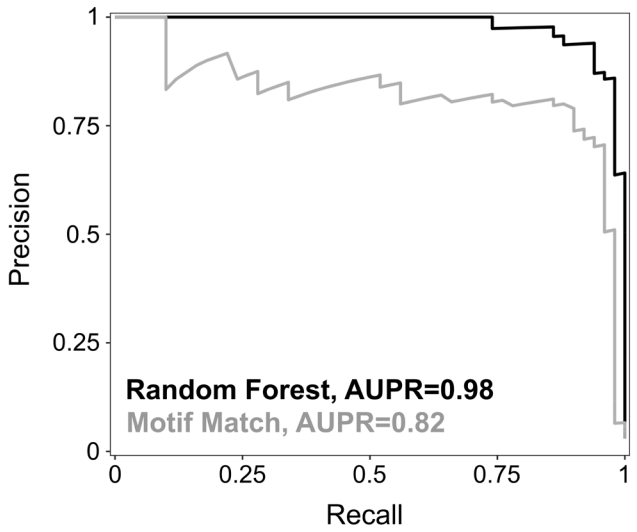

Δdbr1 +DBE
Log (Average 5'-P 5'SS Normalized Read Counts)
Δdbr1
Log (Average 5'-P 5'SS Normalized Read Counts)
-1
0
1
2
3
4

Figure S5

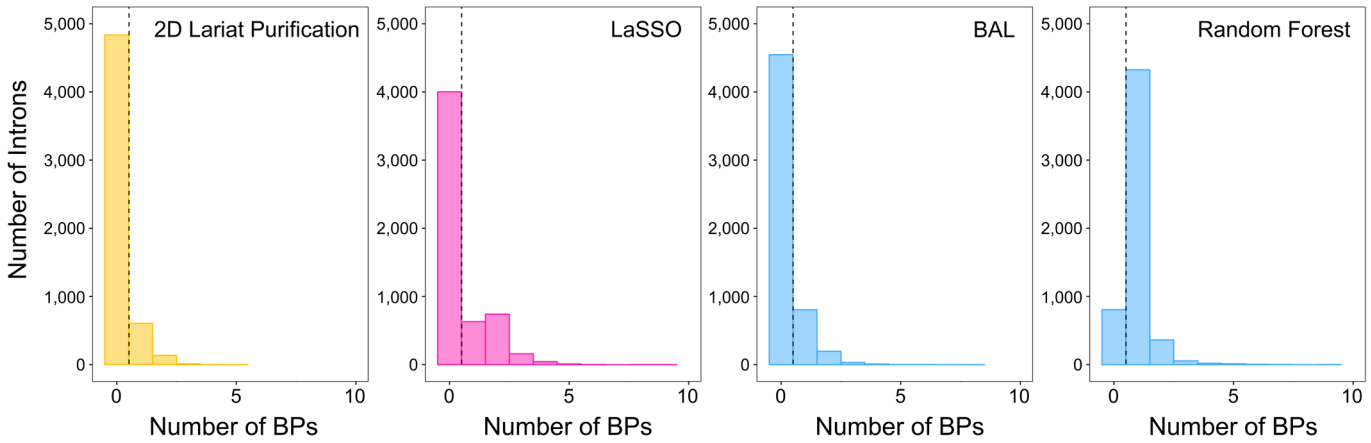
A



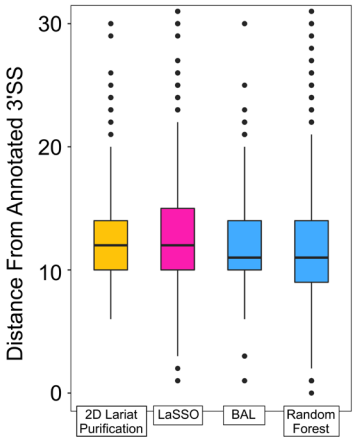
B



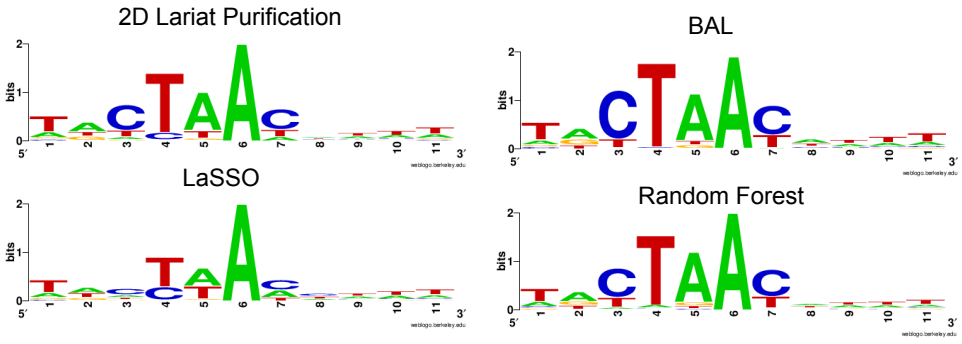
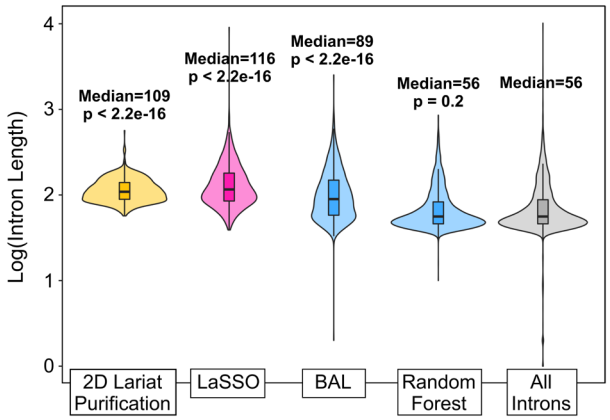
C



D



E



F