

## ORIGINAL ARTICLE

**Characterization of Pre-mRNA Splicing and Spliceosomal Machinery in *Porphyridium purpureum* and Evolutionary Implications for Red Algae**Donald K. Wong<sup>a</sup> , Martha S. Stark<sup>b</sup>, Stephen D. Rader<sup>b</sup>  & Naomi M. Fast<sup>\*a</sup><sup>a</sup> Department of Botany, University of British Columbia, 3156-6270 University Boulevard, Vancouver, BC, Canada<sup>b</sup> Department of Chemistry, University of Northern British Columbia, 3333 University Way, Prince George, BC, Canada**Keywords**

Intron retention; rhodophyte; snRNA; spliceosome; transcriptome; U1 snRNP.

**\*Correspondence**N.M. Fast, Department of Botany, University of British Columbia, 3156-6270 University Boulevard, Vancouver, BC, Canada  
Telephone number: +1-604-822-1630; e-mail: naomi.fast@ubc.caReceived: 23 November 2020; revised 28 January 2021; accepted February 5, 2021.  
Early View publication 12 March, 2021

doi:10.1111/jeu.12844

**ABSTRACT**

Pre-mRNA splicing is a highly conserved eukaryotic process, but our understanding of it is limited by a historical focus on well-studied organisms such as humans and yeast. There is considerable diversity in mechanisms and components of pre-mRNA splicing, especially in lineages that have evolved under the pressures of genome reduction. The ancestor of red algae is thought to have undergone genome reduction prior to the lineage's radiation, resulting in overall gene and intron loss in extant groups. Previous studies on the extremophilic red alga *Cyanidioschyzon merolae* revealed an intron-sparse genome with a highly reduced spliceosome. To determine whether these features applied to other red algae, we investigated multiple aspects of pre-mRNA splicing in the mesophilic red alga *Porphyridium purpureum*. Through strand-specific RNA-Seq, we observed high levels of intron retention across a large number of its introns, and nearly half of the transcripts for these genes are not spliced at all. We also discovered a relationship between variability of 5' splice site sequences and levels of splicing. To further investigate the connections between intron retention and splicing machinery, we bioinformatically assembled the *P. purpureum* spliceosome, and biochemically verified the presence of snRNAs. While most other core spliceosomal components are present, our results suggest highly divergent or missing U1 snRNP proteins, despite the presence of an uncharacteristically long U1 snRNA. These unusual aspects highlight the diverse nature of pre-mRNA splicing that can be seen in lesser-studied eukaryotes, raising the importance of investigating fundamental eukaryotic processes outside of model organisms.

IN eukaryotic gene expression, messenger RNA (mRNA) often undergoes an extensive maturation process: pre-mRNA can be capped, a poly-adenosine tail added, and noncoding regions (introns) removed through a process called pre-mRNA splicing. This process is often viewed as ubiquitous and highly conserved across eukaryotes; almost every eukaryotic genome possesses introns and encodes the conserved spliceosomal machinery responsible for their removal.

The process of pre-mRNA splicing has been most extensively characterized in humans and the budding yeast *Saccharomyces cerevisiae*, where the process and machinery are well conserved (reviewed in Will and

Lührmann 2011). At the core of the spliceosome are the five small nuclear ribonucleoproteins (snRNPs)—the U1, U2, U4, U5, and U6 snRNPs—which are each comprised of a single small nuclear RNA (snRNA) and associated proteins. These dynamically assemble around an intron by recognizing conserved motifs such as the 5' and 3' splice sites and the region surrounding the branch donor adenosine; then, through two transesterification reactions, the intron is removed, and the flanking exons are ligated together. A variety of additional complexes assist in identifying and defining introns and exons, remodeling the spliceosome, and debranching and degrading the excised intron.

Comparing eukaryotes that have evolved under the pressures of genome reduction provides unique opportunities to further our understanding of pre-mRNA splicing, as they highlight minimal systems that remain functional, thereby distinguishing essential structural and catalytic components from those that might only regulate efficiency or timing of splicing. Previous studies on the evolutionary implications of genome reduction on pre-mRNA splicing have revealed intriguing diversity in an otherwise highly conserved process (Hudson et al. 2015). In a number of lineages, we see a dramatic reduction in intron density, where their genomes contain as few as 27 introns, along with a stark reduction in spliceosomal components (Katinka et al. 2001; Matsuzaki et al. 2004; Stark et al. 2015). Transcriptomic studies of intron-sparse organisms reveal that mature transcripts of intron-containing genes often remain unspliced under the conditions tested (Grisdale et al. 2013; Wong et al. 2018). In other reduced genomes, intron density remains similar to close relatives, but intron lengths have been shortened to as tiny as 15 nucleotides (Gilson et al. 2006; Slabodnick et al. 2017; Suzuki et al. 2015; Tanifuji et al. 2014).

The advent of genomics has greatly improved our understanding of some of the less-studied eukaryotic lineages, such as red algae. Recently, it has been reported that the ancestor of all red algae may have undergone genome reduction prior to that group's radiation (Qiu et al. 2015). The relationship between pre-mRNA splicing and genome reduction has been studied in some detail in two of the most well-sequenced red algal genomes to date—those of unicellular extremophiles in the early-branching lineage Cyanidiales, *Cyanidioschyzon merolae* and *Galidieria sulphuraria*. Both organisms were originally found in volcanic vents, thriving in acidic, high-temperature environments. Although the two are ecologically similar and more closely related to each other than to the mesophilic rhodophytes, their genomes have diverged significantly. At 16.5 Mbp, *C. merolae* has only 27 annotated introns scattered across 5,331 genes (Matsuzaki et al. 2004). The spliceosome of *C. merolae* is highly reduced—a bioinformatic inventory and biochemical analysis revealed only 45 core spliceosomal proteins (Reimer et al. 2017; Stark et al. 2015). Indeed, the spliceosome is so reduced that the entire U1 snRNP is missing, including the gene for the associated snRNA, suggesting an alternate mechanism for 5' splice site recognition (Stark et al. 2015). In contrast, while the genome of *G. sulphuraria* is even smaller—at 13.7 Mbp, it harbors 6,623 protein-coding genes and more than 13,000 introns, a substantial difference in gene content and intron density from *C. merolae* (Schönknecht et al. 2013). A recent study on pre-mRNA splicing in *G. sulphuraria* noted remarkable conservation of spliceosomal components (Qiu et al. 2018). Finally, we previously reported unusually high levels of unspliced transcripts in the nucleomorph of the cryptophyte *Guillardia theta* (Wong et al. 2018), whose plastid is derived from an endosymbiotic event where a red alga of unknown phylogenetic position was taken up and reduced to become the current organelle. The remnant red algal nucleus persists as the

highly reduced nucleomorph, and our results provide an additional perspective on the patterns of pre-mRNA splicing in red algae.

While pre-mRNA splicing has been studied closely in extremophilic red algae, less is known about this process among the mesophiles. Research into biofuels has garnered interest in several species of *Porphyridium*, a genus of unicellular mesophiles. One such species, *Porphyridium purpureum*, has had its genome sequenced (Bhattacharya et al. 2013; Lee et al. 2019). A recent reassembly of the genome estimates its genome size to be slightly larger than that of *C. merolae* at 22.1 Mbp, with 9,898 annotated protein-coding genes (Lee et al. 2019). Most intriguing to us is the relative paucity of introns—only 235 introns were predicted in the original genome annotation of *P. purpureum* (Bhattacharya et al. 2013). Although the most recent genome reassembly and reannotation increases the number of predicted introns (Lee et al. 2019), as we will discuss below, our analyses still suggest that *P. purpureum* is among the most intron-poor red algae sequenced to date.

In this study, we investigated multiple aspects of pre-mRNA splicing in *P. purpureum*. Our data reveal extensive intron retention in its mature transcripts, and an intimate connection between the similarity of an intron's 5' splice site to the consensus sequence and splicing levels. We also annotated the spliceosome of *P. purpureum* through comprehensive bioinformatic searches of its genome, highlighting the presence of core spliceosomal components. We discovered an unusual U1 snRNP with an uncharacteristically long U1 snRNA and a smaller protein complement, including highly divergent proteins. Our study contributes to increasing our understanding of the immense diversity within the process of pre-mRNA splicing, providing additional insight into the evolution of this ubiquitous eukaryotic process in red algae, and also the effects of genome reduction on splicing.

## METHODS

Detailed methods can be found in Methods S1.

### Culturing of *Porphyridium purpureum*

Monoclonal cultures of *P. purpureum* CCMP 1328 were obtained from the National Center for Microbiota and Algae (NCMA, formerly CCMP, East Boothbay, ME). Organisms were cultured in 250 ml Erlenmeyer flasks using 50 ml of f/2-Si media without agitation, and were exposed to 30  $\mu\text{mol photons/m}^2/\text{s}$  of light in a 12-h light/dark cycle at a constant temperature of 14 °C.

### RNA extraction and poly(A) purification

Dense cultures of late exponential phase of *P. purpureum* were pelleted 6 h into the light cycle. Total RNA for RNA-Seq was extracted from these pellets using the TRIzol reagent (Ambion, Thermo Fisher Scientific, Waltham, MA) following the manufacturer's protocol, and quantified using a NanoDrop spectrophotometer. RNA was treated with

DNase using the Invitrogen DNA-free DNA Removal Kit (Life Technologies, Thermo Fisher Scientific). Further quantification of RNA and DNA was performed using QuBit fluorimetry. Poly-A purification was performed using NEXTflex Poly(A) Beads (BioO Scientific, Austin, TX) to enrich samples for mRNA. For northern blotting, RNA was extracted under non-denaturing conditions using cold phenol in order to preserve RNA–RNA base pairing (Methods S1).

### Strand-specific library preparation and second-generation sequencing

Two strand-specific libraries of *P. purpureum* were prepared as replicates using the NEXTflex Directional RNA-Seq Kit (BioO Scientific), which uses the dUTP method to maintain strand specificity (Parkhomchuk et al. 2009; Wang et al. 2011), and the libraries were prepared according to the manufacturer's protocol. The two libraries were sequenced on an in-house Illumina HiSeq 2000, generating a total of 54,558,016 and 53,151,718 paired-end reads, respectively.

### Transcriptome analysis of sequence data

The resulting libraries were mapped using STAR (Dobin et al. 2013) to available genome assemblies and annotations of *P. purpureum* (GenBank accession GCA\_008690995.1), followed by determining the number of mapped reads per gene. These raw counts were then normalized to determine relative expression levels using the FPKM (fragments per kilobase of exon per million reads mapped) method (Mortazavi et al. 2008). We determined that these replicate libraries showed similar gene expression levels for any particular gene between the two libraries (Pearson's  $r = 0.9533$ ), so the datasets were pooled for the remaining analyses.

The most recent assembly (Lee et al. 2019) of the *P. purpureum* genome predicts 1,856 introns, a number not explicitly stated in the text, but one we inferred from exonic coordinates in publicly available annotations. This is a marked increase over the 235 in the original assembly (Bhattacharya et al. 2013). To assess splicing efficiency in *P. purpureum*, we first enumerated the mapped read pairs in the vicinity of all 1,856 junctions we inferred from Lee et al. (2019) for the type of splicing event they represent, such as spliced transcripts, intron retention, and other alternative splicing events. We then calculated the splicing levels, or percent of spliced reads, for each annotated junction by dividing the number of spliced reads by total reads mapped in its vicinity. There were 889 introns that did not have adequate read coverage (597 of these had no coverage at all), preventing further analysis. Of the remaining 967 introns, a large number of these had extremely high read coverage with only a few split reads, and 657 of these lacked any evidence of splicing through representation by split reads (File S1). Further inspection showed that these introns with poor coverage or a lack of evidence of splicing have no discernible branch point region or other expected sequence features, suggesting widespread misannotation of introns. To avoid issues with false

positives stemming from below-threshold junctions and potentially false annotations, we then focused the remainder of our analyses to junctions with coverage lower than 25 split reads from further analyses, leaving a total of 192 introns, a number more similar to the 235 introns of the original assembly (Bhattacharya et al. 2013).

### 5' splice site analysis

We retrieved the sequences of the 192 introns from our splicing level analysis as described above and generated sequence logos using WebLogo (Crooks et al. 2004) to determine the overall *P. purpureum* 5' splice site consensus. We also generated additional sequence logos by binning introns based on our calculated splicing levels in nine intervals each spanning 10% (no introns we analyzed were spliced > 90%) to determine each interval's consensus 5' splice site.

### Bioinformatic reconstruction of the spliceosome in *Porphyridium purpureum*

We searched the *P. purpureum* genome assembly, as well as an assembled transcriptome from our RNA-Seq data, for snRNAs and known spliceosomal proteins (Methods S1) to reconstruct the spliceosome of *P. purpureum*. To identify spliceosomal (U1, U2, U4, U5, and U6) snRNAs, we retrieved curated alignments for each of those from Rfam (Kalvari et al. 2018). We used the Infernal (Nawrocki and Eddy 2013) suite of scripts for this search, as it makes use of RNA secondary structure to increase sensitivity. In brief, covariance models were generated and used to search the *P. purpureum* genome assembly for the snRNAs. We predicted secondary structures using mfold (Zuker 2003). We searched for spliceosomal proteins in the genome of *P. purpureum* through reciprocal BLAST similar to Ward and Moreno-Hagelsieb (2014) and Stark et al. (2015), using homologs from *Homo sapiens*, *S. cerevisiae*, *Arabidopsis thaliana*, and *C. merolae* (File S2).

### Biochemical detection of snRNAs in *Porphyridium purpureum*

To biochemically verify the snRNAs, we designed probes (Methods S1) for four snRNAs using candidate sequences (U1, U2, U4 identified through Infernal, U6 sequence predicted based on complementarity to U4), followed by a denaturing northern blot on total RNA. A native northern blot was performed to verify the U4-U6 snRNA interaction by electrophoresing untreated and heat-treated total RNA on a non-denaturing gel, followed by northern blot analysis with U4 and U6 probes.

## RESULTS AND DISCUSSION

### Many transcripts in *Porphyridium purpureum* are not spliced

The low intron density of *P. purpureum* could lead to profound impacts on the patterns of pre-mRNA splicing, as

extensive intron retention has previously been reported in organisms with intron-sparse genomes (Grisdale et al. 2013; Wong et al. 2018). To assess pre-mRNA splicing across all transcripts of *P. purpureum*, we generated transcriptomic data from duplicate strand-specific RNA-Seq libraries. Strand specificity of our transcriptomic reads allowed us to distinguish between intron retention and antisense transcription in the vicinity of an intron, as both types of transcriptional events would appear the same under traditional RNA-Seq methods. In total, 98,810,814 pooled read pairs mapped uniquely to the existing genome assembly of *P. purpureum*, which represent 91.74% of our entire dataset.

We analyzed 192 *P. purpureum* introns (see Materials and Methods) by calculating the proportion of split reads in the vicinity of the intron. We observed that many of these introns in *P. purpureum* are not removed from transcripts, as summarized in Fig. 1. None of these introns had more than 90% spliced reads—across all the *P. purpureum* introns we analyzed, the average percent of spliced reads is 53% (Fig. 1). This suggests that just under half of the population of transcripts of these genes is not spliced and that intron retention is prevalent in *P. purpureum*.

We excluded introns that are likely misannotated in the most recent assembly (see Materials and Methods). Not only do a large number of these introns show no evidence of splicing in our transcriptomic data as they completely lack split reads, most of the excluded introns also lack typical intronic sequence features such as a discernible 5' splice site or branch donor site. Taken together, these observations suggest that the true number of introns in *P. purpureum* is closer to the 235 introns predicted in the original assembly (Bhattacharya et al. 2013). In the unlikely event that these sequences are actually noncanonical introns for which we could find no split reads, our conclusions regarding the extent of intron retention across the *P. purpureum* genome would not be invalid—they would instead be a vast underestimation.

Previous reports of extensive genome-wide intron retention are from organisms that have extremely low intron density. The microsporidian *Encephalitozoon cuniculi* has only 37 annotated introns (Grisdale et al. 2013; Katinka et al. 2001), while the red algal-derived nucleomorph of the cryptophyte *G. theta* has 17 introns in its tiny genome (Douglas et al. 2001). In these and other organisms with low intron density, extensive intron retention could possibly be more easily tolerated by the reduced need for proteins encoded by intron-containing genes. Another possibility is introns that remain within a genome with low intron densities tend to lie in genes with drastically increased expression, where higher levels of transcripts can compensate for poor processing. In a previous study, we compared splicing levels between nucleomorphs with vastly different intron densities (Wong et al. 2018). While the intron-sparse and highly reduced *G. theta* nucleomorph exhibited intron retention in its transcripts, the even more reduced—yet intron-dense—nucleomorph of the chlorarachniophyte alga *Bigeloviella natans* did not

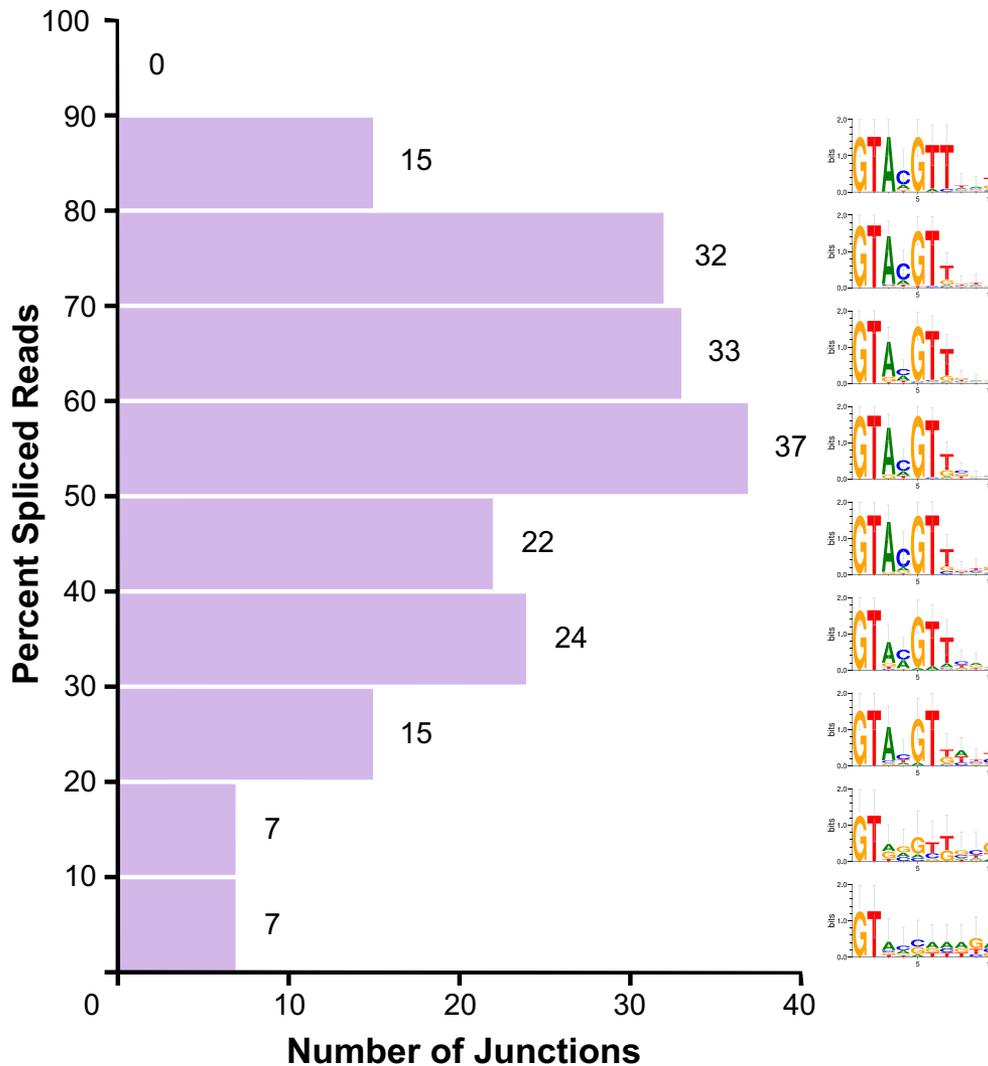
(Wong et al. 2018). In cases of high intron density, there is also likely much stronger pressure to efficiently remove introns, as extensive intron retention would be highly detrimental to normal gene expression in a cell.

However, while the intron density of *P. purpureum* is low, current estimates still place it higher than that of *C. merolae*, or even the budding yeast *Saccharomyces cerevisiae*. As a point of comparison, *S. cerevisiae* has ~300 annotated introns in ~6,500 protein-coding genes in a smaller genome than that of *C. merolae* or *P. purpureum*, yet *S. cerevisiae* genes are well spliced. Thus, intron density of any given reduced genome might not be so straightforwardly correlated with splicing levels. In *P. purpureum*, we determined that the percent of spliced reads across the introns we analyzed spans a very broad range. We were unable to find any correlation between the splicing level of a particular intron and the expression level of the gene in which it resides, ruling out the possibility that observed intron retention could be an artifact of poor read coverage (File S3). We also could not find any clear correlation linking splicing levels to the functions of these intron-containing genes (File S3), suggesting that intron retention might be related to some property of the intron itself. Finally, without finding connections between intron length and splicing level (File S3), we queried key sequences of these introns for their relation to intron retention.

### Increased variability of the 5' splice site sequence in poorly spliced introns

Intron recognition involves spliceosomal proteins binding to the intron, or base pairing between key regions of an snRNA and a corresponding intron motif. Pre-mRNA splicing is initiated through an interaction between the unpaired 5' end of U1 snRNA and the 5' splice site of the intron. The strength of this interaction has been observed to affect the splicing level of an intron, providing selective pressure to establish a strong sequence consensus and complementarity between the 5' splice site and the U1 snRNA (Freund et al. 2005). Indeed, organisms with a low density of introns, which often also possess a reduced spliceosome, tend to have stronger intron motif consensus sequences allowing for stronger interactions between the introns and the remaining components of the spliceosome (Freund et al. 2005; Irimia et al. 2007). With this in mind, we examined *P. purpureum* intron sequences and compared the extent of similarity to its 5' splice site consensus relative to splicing levels (Fig. 1).

Across the 192 introns we analyzed, the 5' splice site shows a relatively strong sequence consensus, with the canonical 5'-GTANG motif, and variability in the fourth nucleotide—though it is a cytosine in just over half of these introns (Fig. 1). Furthermore, extended conservation of two additional thymidines beyond the typical eukaryotic consensus is observed. This supports the hypothesis that introns in an intron-sparse reduced genome reduce the amount of sequence variation in the 5' splice site to better interact with a reduced spliceosome. Further examination



**Figure 1** Pre-mRNA splicing levels in *Porphyridium purpureum* and their association with variability of the 5' splice site. The splicing levels of 192 junctions are summarized in this histogram showing the frequency of junctions with any particular splicing level, along with consensus 5' splice site sequences from introns at each bin. The average intron has 52.8% of mapping reads spliced, while the median percent spliced reads is 54.9%. The variability of the 5' splice site sequence generally increases as the percent of spliced reads decreases.

of 5' splice site sequences with respect to intron splicing levels reveals a clear trend. Introns spliced at the highest levels (e.g. those with 60–90% spliced reads) more often conform to a strong and extended consensus, with a 5' splice site consensus seven nucleotides long of 5'-GTACGTT (Fig. 1). As splicing levels decrease, the strength of this consensus diminishes—the fourth and seventh nucleotides notably increase in variability (Fig. 1). The 5' splice site consensus of the most poorly spliced introns (those between 0% and 10% spliced) weakens even further to just GT. On the other hand, we did not observe such a connection between branch point regions or 3' splice sites and the level of splicing. Therefore, the splicing level of a *P. purpureum* intron is strongly related to the sequence of its 5' splice site, suggesting selective pressure toward a strong 5' splice site consensus for a functional interaction with the U1 snRNA.

In *C. merolae*, the 5' splice sites of its 27 annotated introns also have a strong consensus sequence of 5'-GTAAGTT (Matsuzaki et al. 2004). Although it is not currently known how the 5' splice sites of *C. merolae* introns are recognized, due to the complete absence of the U1 snRNP, such conservation of the 5' splice site must be important. Given similar trends of a strong 5' splice site consensus in the more highly spliced introns of *P. purpureum*, we searched for the U1 snRNA in order to learn more about the mechanism of 5' splice site recognition in *P. purpureum*.

#### snRNAs in the *Porphyridium purpureum* genome and the unusual structure of the U1 snRNA

Our observation of a strong 5' splice site consensus sequence among highly spliced introns suggested to us

the possibility of a reduced spliceosome—there is selection toward the consensus to better interact with remaining components of the spliceosome. We searched for all five snRNAs in the *P. purpureum* genome using Infernal under its default settings (Nawrocki and Eddy 2013), using covariance models generated from alignments available from Rfam (Kalvari et al. 2018).

We were able to retrieve a single candidate sequence of the U1 snRNA from the *P. purpureum* genome (Fig. 2; File S4). Based on Infernal's inclusion thresholds, this was a marginal match, and the candidate sequence was not able to retrieve U1 snRNA when used to reciprocally search back to Rfam. There appeared to be two substantial insertions within Stem Loops II and III relative to the Rfam consensus model (File S4). While these two insertions must impact how this U1 snRNA candidate will fold, we nevertheless compared its predicted structures to those of known U1 snRNAs from other eukaryotes.

The canonical U1 snRNA folds into a cloverleaf structure that comprises four major stem loops (SL I through IV), of which SL I and II directly interact with the proteins U1-70K and U1-A, respectively (Fig. 2A; Kondo et al. 2015; Nagai et al. 1990; Pomeranz Krummel et al. 2009; Oubridge et al. 1994; Scherly et al. 1989; Stark et al. 2001). A 5' single-stranded region interacts with the 5' splice site, while a 3' unpaired region between SL III and SL IV interacts with the Sm complex (Kondo et al. 2015; Stark et al. 2001). Despite the insertions, the predicted structure of the *P. purpureum* U1 snRNA bears strong similarity to existing consensus structures (Fig. 2A), especially with respect to stem loops I, III, and IV. Although Infernal predicted the insertions to interrupt both SL II and III, structural predictions show that the insert region forms an appropriately sized SL III and a greatly expanded SL II (Fig. 2A). Variability is not unheard of, as the *S. cerevisiae* U1 snRNA is also greatly expanded (568 nt), with an enlarged SL III (Kretzner et al. 1990; Li et al. 2017).

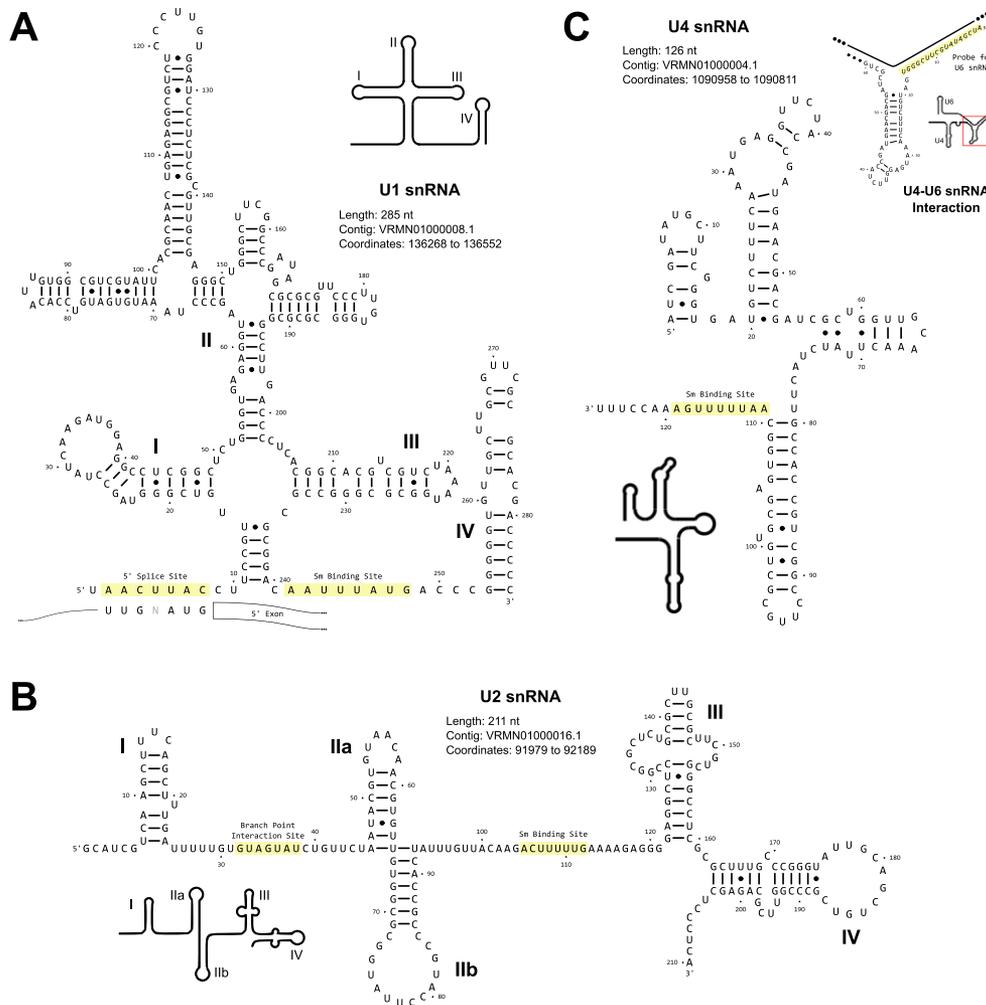
Studies investigating the diversity of snRNA sequences across metazoans and yeasts conclude unsurprisingly that in the U1 snRNA the loops and other unpaired regions where spliceosomal proteins bind are more conserved in sequence than the stems (Marz et al. 2008; Mitrovich and Guthrie 2007). Indeed, the loop in SL I (Fig. 2A, nucleotides 27–40) of the *P. purpureum* U1 snRNA candidate bears some level of sequence conservation. The protein U1-70K interacts with a conserved GAUCA motif in this loop in humans, where a tyrosine and leucine in an RNA binding motif in U1-70K could be cross-linked to the G and U nucleotides, respectively (Kondo et al. 2015; Urlaub et al. 2000). In *S. cerevisiae*, a similar interaction is seen between its U1-70K (Snp1) homolog and SL I of its U1 snRNA (Li et al. 2017). In *P. purpureum*, the loop contains UAUCA (Fig. 2A, nucleotides 28 to 32) instead. As discussed later, we identified a divergent homolog of U1-70K in *P. purpureum*—this single nucleotide change in the U1 snRNA might be related to its divergence. The 3' single-stranded region has the conserved Sm complex binding site (Stark et al. 2001), and the sizes of SL III and IV also closely match consensus structures of the U1 snRNA

from Rfam. Most importantly, the 5' unpaired region of our U1 snRNA candidate is complementary to the consensus 5' splice site of highly spliced introns (Fig. 1, 2A).

Extended sequence interactions between snRNAs and complementary intron features have been previously observed. Of note is the extended base-pairing interaction between the region of the branch point adenosine of *S. cerevisiae* introns and its U2 snRNA (Berglund et al. 1997; Langford and Gallwitz 1983). In microsporidians, a group of intracellular fungal parasites where extensive intron retention has been observed, this extended base pairing is taken even further, where introns that possess branch point regions with high complementarity to U2 snRNA have been linked to distinctively high levels of splicing (Grisdale et al. 2013; Whelan et al. 2019). Our results suggest that a similar relationship between the U1 snRNA and the 5' splice site could be taking place—introns in *P. purpureum* that bear a 5' splice site with potential for extended base pairing with the U1 snRNA are associated with higher levels of splicing (Fig. 1, 2A).

Strong candidates were found for both U2 and U4 snRNAs, and these sequences from *P. purpureum* could reciprocally identify their respective snRNAs from Rfam (Fig. 2; File S4). The U2 snRNA, which recognizes and interacts directly with nucleotides adjacent to the branch point adenosine, is highly conserved in structure across diverse eukaryotes (Fast et al. 1998; Hudson et al. 2019; Madhani and Guthrie 1992; Mitrovich and Guthrie 2007; Stark et al. 2015), and the U2 snRNA candidate from *P. purpureum* is no exception (Fig. 2B). When compared to the Rfam consensus structure, the U2 snRNA candidate from *P. purpureum* can form the same stem loops and open regions that allow it to interact with associated U2 snRNP components, Sm proteins and the branch point adenosine. Likewise, the 126 nt U4 snRNA candidate from *P. purpureum* is predicted to form a structure (Fig. 2C) conserved across a broad range of eukaryotes (Krol et al. 1988; Mysliński and Branlant 1991; Stark et al. 2015).

Infernal was unable to identify candidates for U5 and U6 snRNAs above the program's default threshold (File S4), regardless of the *P. purpureum* genome assembly used for our searches. In fact, Qiu et al. (2018) reported the presence of U5 and U6 snRNAs in *P. purpureum* based only on these identical below-threshold e-values and Infernal scores, suggesting our studies identified the same dubious candidates. Indeed, close inspection of these candidate sequences showed that they did not resemble consensus U5 and U6 snRNAs (File S4), nor did they match any known noncoding RNA from Rfam. Furthermore, the U6 snRNA candidate did not possess sequence capable of interacting with the U2 and U4 snRNAs. Despite the lack of recognizable U5 and U6 snRNAs, it is unlikely that they are absent in *P. purpureum*, as they are intimately associated with the U4 snRNA as part of the U4/U6-U5 tri-snRNP. Instead, this might point to high sequence divergence, and to difficulty in identifying the U5 snRNA in the *P. purpureum* genome through bioinformatics alone. This is not particularly surprising for the U5 snRNA, as its nucleotide sequence has



**Figure 2** The snRNAs identified in *Porphyridium purpureum* and their predicted structures. Lengths and genomic coordinates of each snRNA, and schematic line drawings of Rfam consensus structures for U1 and U2 snRNA, along with the proposed structure of *Saccharomyces cerevisiae* free U4 snRNA and U4-U6 snRNA interaction, are provided for reference. **(A)** An unusually long U1 snRNA was found in the *P. purpureum* genome, which is still predicted to fold with three of the four stem loops of canonical U1 snRNAs from other eukaryotes. **(B)** A canonical U2 snRNA, with a highlighted branch point interacting site, was identified in the *P. purpureum* genome. **(C)** The U4 snRNA candidate identified in *P. purpureum* can form a similar structure to the *S. cerevisiae* free U4 snRNA. A proposed structure highlighting the region of U4-U6 snRNA interaction in *P. purpureum* is provided here, and was used to guide probe design for northern blotting. The shaded sequence is the probe used to detect *P. purpureum* U6 snRNA.

been noted to be extremely variable between different eukaryotes (Frank et al. 1994; Hinz et al. 1996).

In contrast, this is quite unusual for the U6 snRNA, as it is the most conserved of all the snRNAs (Brow and Guthrie 1980). The U6 snRNA, at the core of the U6 snRNP, plays a major role throughout pre-mRNA splicing, as its nucleotides directly interact in turn with the U4 snRNA, the U2 snRNA and the 5' splice site (Brow and Guthrie 1980; Kandels-Lewis and Séraphin 1993; Sawa and Shimura 1992; Wu and Manley 1991). Considering its central role within the spliceosome, and the presence of canonical U2 and U4 snRNAs in *P. purpureum*, it seems highly unlikely that it would not be present. Because the interaction between the U4 and U6 snRNAs

is well characterized, we used the 5' end of the candidate U4 snRNA sequence to exhaustively search both assemblies (Bhattacharya et al. 2013; Lee et al. 2019) of the *P. purpureum* genome, as well as the raw reads used for either assembly, in the unlikely event that the region containing the U6 snRNA gene somehow did not get assembled. However, no attempts resulted in identification of anything resembling a U6 snRNA. It is possible that the region containing the U6 snRNA gene evaded sequencing or that its sequence has diverged to such an extent as to not be recognizable. Given our results, it seemed highly unlikely that *P. purpureum* is actually missing the U6 snRNA; thus, we attempted to verify its presence biochemically.

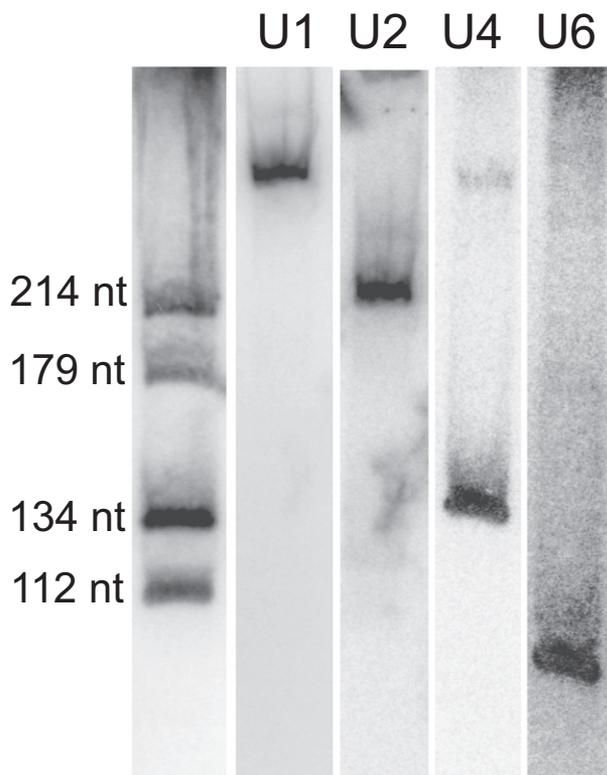
### Biochemical verification of the snRNAs and detection of U6 snRNA

To confirm the presence of our candidate snRNAs (Fig. 2; File S4), we probed for them by northern blotting (Fig. 3). Complementary probes for U1, U2, and U4 snRNAs (see Methods S1) were designed using their respective candidate sequences and their expression, and predicted size was verified (Fig. 3). The probe for U1 snRNA interacts strongly with an RNA of approximately 285 nucleotides in length (Fig. 3), confirming that the *P. purpureum* U1 snRNA is expressed with the inserts as predicted.

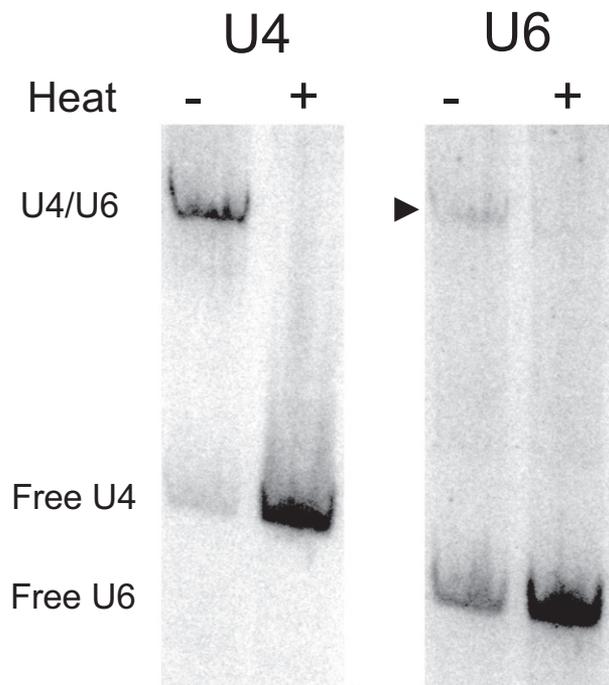
Though we could not directly design a probe for the U6 snRNA, as no *P. purpureum* gene for it could be identified, we used the aforementioned U4 and U6 snRNA interaction (Fig. 2) to design a probe to target the U6 snRNA. Indeed, our northern blots revealed a product ~100 nucleotides in length (Fig. 3), which is of the size expected for a U6 snRNA (Kalvari et al. 2018). To further test its identity as the U6 snRNA, we determined if it interacts with the U4 snRNA using a native northern blot. (Fig. 4). RNA from *P. purpureum* was either briefly heated

or left untreated to respectively dissociate or preserve the interaction between the U4 and U6 snRNAs. In the untreated sample, both probes bound to a higher molecular weight complex that we interpret as base-paired U4 and U6 snRNAs, along with a lower molecular weight band that corresponds to the free snRNA (Fig. 4). Furthermore, as previously described in *S. cerevisiae* and the frog *Xenopus laevis* (Hamm and Mattaj 1989; Siliciano et al. 1987), the U6 snRNA appears to be present in excess, as shown by the difference in relative intensities of the signal of free vs. complexed U6 snRNA (Fig. 4). Upon heat treatment, this complex dissociates and only free snRNAs are detected (Fig. 4). Our results strongly support the presence of U6 snRNA in *P. purpureum*, and that its gene is somehow missing from the current genome assembly. Because of the close association of U4 and U6 snRNAs with the U5 snRNA, and additional evidence from our reconstruction of the spliceosome discussed later, we predict that the U5 snRNA is also present.

Although this might raise concerns of an incomplete *P. purpureum* genome assembly, this assembly's BUSCO score, a measure of conservation of expected eukaryotic gene sets as a proxy for genome completeness (Simão et al. 2015), was reported to be 90.40% (Lee et al. 2019). These scores are in line with those reported from genomes of other red algae such as *C. merolae* (93.4%) and *G. sulphuraria* (91.1%) (Qiu et al. 2018), and are comparable to the more frequently studied genomes of humans



**Figure 3** Biochemical verification of snRNAs through denaturing northern blotting. Four of the five (U1, U2, U4, and U6) spliceosomal snRNAs could be detected from probes designed to candidate sequences. The U1 snRNA probe detected a fragment at relatively high molecular weight, confirming that the U1 snRNA in *Porphyridium purpureum* exists in its full length of 285 nt. We also probed for small RNA of known sizes in *Saccharomyces cerevisiae* and *Cyanidioschyzon merolae* as markers (see Methods S1).



**Figure 4** The U4/U6 snRNA interaction under native northern blotting. Lanes representing the heat-treated (denatured) sample are marked with (+), while the untreated (native) sample is represented with (-). The same blot was probed sequentially for U4 and U6 snRNAs.



fied in *P. purpureum* likely reflect lineage-specific losses, given that these proteins are present in *G. sulphuraria* and in land plants. As expected, the *P. purpureum* spliceosome is less complex than the 200+ component spliceosomes of intron-rich metazoans or land plants, although many of these “extra” components are thought to be mostly involved in alternative splicing (Fabrizio et al. 2009; Hudson et al. 2015; Kalyna and Barta 2004; Lorković et al. 2000; Wahl et al. 2009). More surprisingly, the *P. purpureum* spliceosome appears more complex than that of *S. cerevisiae*, which has a comparable intron density to *P. purpureum*. We identified a number of additional *P. purpureum* components found in human and *A. thaliana* spliceosomes that are not found in *S. cerevisiae* (Table 1 and Table S1).

The five small nuclear ribonucleoproteins (U1, U2, U4, U5, and U6 snRNPs) are at the heart of the spliceosome, directly interacting with sequence features of introns, and with one another, to catalyze intron removal. As expected, both the Sm and LSm complexes, which are key components of all five snRNPs (Achsel et al. 2001), are present (Table S1). Similarly, almost all protein components of the U2 snRNP, which recognizes the region of the branch donor adenosine, could be unambiguously identified in *P. purpureum* (Table 1). Together with our previous identification of a highly conserved U2 snRNA (Fig. 2), this suggests a rather canonical U2 snRNP. Of the *C. merolae* snRNPs, U2 snRNP was found to be the most complex amid extreme spliceosome reduction, highlighting the importance of branch point recognition in pre-mRNA splicing (Hudson et al. 2015; Stark et al. 2015). Similarly, this particle is also present with all known components in *G. sulphuraria* (Qiu et al. 2018). As for U2-B”, the most recent assembly of the *P. purpureum* genome identifies a gene encoding this protein (Lee et al. 2019). We discuss later that this gene’s product is likely shared between the U1 and U2 snRNPs.

Likewise, proteins of the U4/U6-U5 tri-snRNP are present in *P. purpureum* (Table 1). While we were not able to identify the U5 and U6 snRNAs in the genome, our biochemical analyses provide extremely strong evidence that the U6 snRNA is present (Fig. 3). By association with the U4 and U6 snRNA, the U5 snRNA is likely also present, as it is improbable for the proteins of the U4/U6-U5 tri-snRNP to exist without their associated snRNAs. The recruitment of this complex to the spliceosome leads to a coordinated rearrangement where the U6 snRNP dissociates from the U4 snRNP, and displaces the U1 snRNP to interact with the 5’ splice site (Will and Lührmann 2011). An association then forms between the U6 and U2 snRNPs to bring the branch point adenosine and 5’ splice site into close proximity for the first catalytic step. Some of the large and highly conserved proteins of the U5 snRNP, such as Prp8, play central roles throughout this process (Grainger and Beggs 2005). These roles emphasize the importance of the U4/U6-U5 tri-snRNP and their conservation across eukaryotes. Although some of the tri-snRNP proteins are missing in *C. merolae* (Stark et al. 2015), the presence of all associated proteins in *P.*

*purpureum* (Table 1) and *G. sulphuraria* (Qiu et al. 2018) suggests that these proteins were also already present in the rhodophyte ancestor and perhaps secondarily lost in *C. merolae*.

We identified an unusually long candidate for the U1 snRNA in the *P. purpureum* genome, and confirmed its expression at its predicted length (Fig. 2A, 3). Its divergence from the canonical U1 snRNA, along with the strong 5’ splice site consensus in *P. purpureum* introns (Fig. 1), suggested the possible divergence of the protein components of the U1 snRNP (Table 1). In the U1 snRNP of humans and *S. cerevisiae*, U1-70K and U1-A interact respectively with stem loops I and II of the U1 snRNA, while U1-C associates with U1-70K to round out the core of the U1 snRNP (Scherly et al. 1989; Nagai et al. 1990; Oubridge et al. 1994; Stark et al. 2001, Pomeranz Krummel et al. 2009; Kondo et al. 2015). Despite its unusual U1 snRNA, all three of these proteins (U1-A, U1-C and U1-70K) appear to be present in the *P. purpureum* genome. Unlike proteins of the other four snRNPs, identification of these three proteins required reducing our stringent BLAST thresholds (see Methods S1). The identified U1 snRNP components in *P. purpureum* were often truncated or divergent in amino acid sequence with respect to the query sequences, which is further supported by lower percent identities and higher e-values of their best BLAST hits especially when compared to those of other snRNPs (Table 1 and Table S1; File S5). This presents a propensity toward divergent, yet recognizable, proteins in the U1 snRNP of *P. purpureum*, in contrast with the generally highly conserved homologs identified in the other four snRNPs.

Inspection of a U1-70K alignment shows increased sequence divergence of the RNA recognition regions in the *P. purpureum* homolog even compared to other red algal U1-70K homologs (File S6). Human U1-70K binds SL I of U1 snRNA, where tyrosine and leucine in positions 112 and 175 respectively can be cross-linked to G and U in a conserved GAUCA sequence (Urlaub et al. 2000). In *P. purpureum*, both homologous positions of its U1-70K homolog bear isoleucine instead (filled triangles on columns 188 and 267 of the U1-70K alignment in File S6), and the predicted structure of its U1 snRNA has UAUCA (Fig. 2A)—what effects these divergences could have on their interaction will require further investigation. The U1-C homolog in *P. purpureum* bears closer resemblance in sequence to other identified U1-C proteins but is highly truncated—at only 90 amino acids long, it is predicted to be reduced to just the N-terminal zinc finger domain (File S6). Despite this truncation, *in vitro* experiments on human U1-C show that only this N-terminal domain is critical for its interaction with U1-70K and the other U1 snRNP components (Muto et al. 2004; Nelissen et al. 1991).

As discussed previously, the insertions within the *P. purpureum* U1 snRNA result in a greatly expanded SL II (Fig. 2A). Since this is the binding region for U1-A, we were curious if the *P. purpureum* homolog might be highly divergent, or even present at all. We identified a single ambiguous homolog in *P. purpureum* in a gene currently

annotated as U2-B" (Lee et al. 2019). While this is not particularly surprising given their relatedness (Delaney et al. 2014; Polycarpou-Schwarz et al. 1996; Price et al. 1998; Saldi et al. 2007; Simpson et al. 1995; Williams and Hall 2011; Williams et al. 2013), we were able to identify only one region in the *P. purpureum* genome that could encode either one of these proteins. A shared protein between the U1 and U2 snRNPs is not unexpected for *P. purpureum*, as no other red algae appear to have separate U1-A and U2-B" proteins (Collén et al. 2013; Qiu et al. 2018; Schönknecht et al. 2013). Thus, we have labeled this particular protein in *P. purpureum* as U1-A/U2-B" (Table 1 and Table S1). Phylogenetic evidence suggests that a single protein shared by the U1 and U2 snRNPs was the ancestral state in metazoans (Williams et al. 2013). Duplication and subfunctionalization of this protein into U1-A and U2-B" have occurred repeatedly in several metazoan lineages, as well as in land plants and fungi, including *S. cerevisiae* (Delaney et al. 2014; Polycarpou-Schwarz et al. 1996; Saldi et al. 2007; Simpson et al. 1995; Williams and Hall 2011). Future investigations into the interactions of this protein with the U1 or U2 snRNAs will provide insight into the evolution of this protein family in other lineages of eukaryotes.

Although we identified U1-C, U1-70K, and the shared U1-A/U2-B", other U1 snRNP proteins known to play roles in the early spliceosome, such as Prp39, Prp42, Snu56, and Snu71 (Gottschalk et al. 1998; Li et al. 2017; Lockhart and Rymond 1994), could not be identified. This might not be particularly unusual, as these proteins have only been shown to be stably associated with the U1 snRNP in *S. cerevisiae* (Fortes et al. 1999; Gottschalk et al. 1998), while being more loosely associated in humans and *A. thaliana*, where they are thought to be involved in alternative splicing (de Francisco Amorim et al. 2018; Puig et al. 2007). However, we did identify homologs of Prp40, Nam8, and Luc7-Like in the *P. purpureum* genome, although Luc7-Like was only a marginal hit and is currently annotated as a hypothetical protein (Table 1, File S5). In yeast, the proteins Luc7-Like and Nam8 function by binding upstream (Luc7-Like) or downstream (Nam8) of the 5' splice site to enhance its interaction with the U1 snRNP (Fortes et al. 1999; Puig et al. 2007; Puig et al. 1999). Therefore, these two proteins and the aforementioned strengthening of the 5' splice site consensus (Fig. 1) are probably crucial for promoting successful 5' splice site recognition and spliceosome assembly in *P. purpureum*, given the divergent U1 snRNA.

When considering the search statistics such as e-value and percent identity for all snRNP components (Table 1) and the alignments of U1 snRNP proteins (File S6), we observe a trend toward divergent, but recognizable, proteins associated with the U1 snRNP of the intron-sparse *P. purpureum*. The extremophilic *C. merolae* is even more intron-poor than *P. purpureum*, and is missing its entire U1 snRNP without extreme reduction of other snRNPs of the spliceosome (Stark et al. 2015). Taken together, this supports the intriguing possibility that this complex might be the first snRNP to diverge rapidly or be lost during

genome and intron reduction, as suggested to have occurred in *C. merolae* and a number of diverse protist species (Hudson et al. 2015). This is further supported by the apparent reduction of the U1 snRNP in the microsporidian *E. cuniculi*—only a few associated proteins have ever been identified, although the U1 snRNA has since been found (Belkorchia et al. 2017; Dávila López et al., 2008; Katinka et al. 2001). Further biochemical analysis of this particle in these organisms could add to our understanding of the evolution of the U1 snRNP and recognition of the 5' splice site.

Aside from the five snRNPs, numerous other complexes and transient proteins are also associated with the spliceosome both during and after the two transesterification steps of pre-mRNA splicing. The Prp19 complex (NTC) is a large agglomeration of proteins that is recruited to the active splice site when the U4/U6-U5 tri-snRNP joins the other two snRNPs to form the catalytically active B complex of the spliceosome (Chan and Cheng 2005; Chan et al. 2003; Hogg et al. 2010). There, it plays crucial roles in dissociating the U4 and U6 snRNPs, as well as in stabilizing the interaction between U5 and U6 snRNP (Chan and Cheng 2005; Chan et al. 2003; Hogg et al. 2010). We were able to identify homologs in *P. purpureum* for almost the entire NTC (Table 1), including some proteins that are not considered essential in *S. cerevisiae* (Hogg et al. 2010) such as Isy1, Bud31, and Snt309. The protein Syf2 was notably absent from our searches in *P. purpureum*; however, its homolog in *S. cerevisiae* is not considered essential (Hogg et al. 2010), and is also absent in *C. merolae* (Stark et al. 2015).

Another of these spliceosomal complexes is the RES complex, which is implicated in retaining unspliced mRNA within the nucleus (Dziembowski et al. 2004), as well as in ensuring efficient transition from the B to the catalytically active B<sup>act</sup> complex (Bao et al. 2017). It is comprised of three proteins: Bud13, Pml1, and Snu17/Ist3. Interestingly, only Snu17/Ist3 could be identified in *P. purpureum* (Table 1). In *S. cerevisiae*, although the homologs of these genes are not essential, deletions of Bud13 or Snu17 resulted in an increased level of unspliced transcripts (Dziembowski et al. 2004). Furthermore, knockouts of homologs of RES complex proteins in *A. thaliana* and zebrafish all show general impairment in splicing, though not all introns are equally affected (Fernandez et al. 2018; Xiong et al. 2019). Considering the role the RES complex plays in efficient splicing, this apparent reduction of the RES complex in *P. purpureum* might correlate with the high levels of intron retention observed in our polyadenylated transcriptome data.

The remaining spliceosomal proteins include other splicing factors, many of which are helicases, that take part in activating the spliceosome, as well as disassembling it when the process is complete so that the complexes may be recycled. We were able to identify most of these in *P. purpureum* (Table 1). A few notable absences from our searches were Cwc23 and Spp2. The Cwc23 protein is associated with Cef1 and the rest of the NTC (Ohi et al. 2002), and its depletion in *S. cerevisiae* results in

accumulation of unspliced mRNA (Sahi et al. 2010). The Spp2 protein associates with Prp2 and is necessary for promoting the spliceosome into its catalytically active state (Roy et al. 1995; Warkocki et al. 2015; Zang et al. 2014). Because the interacting partners of Cwc23 and Spp2 could be found in *P. purpureum* (Table 1), it is possible that Cwc23 and Spp2 are also present but too divergent to detect with our methods. Alternatively, their absences could be contributing to a spliceosome that cannot effectively splice mRNA.

Overall, our results paint a picture of a spliceosome where much of the core complex is conserved, despite its reduced genome, relatively low intron density, and high levels of intron retention (Fig. 1). Additionally, with the marked exception of U1 snRNP proteins, many of the proteins we did identify are very similar to homologs from other eukaryotes (Table 1 and Table S1; File S6). In *C. merolae* and *G. sulphuraria*, despite the expectation that their spliceosomes might have diverged to reflect their physiology as extremophiles, it was noted that the identified spliceosomal components are not particularly different from homologs found in other eukaryotic lineages (Qiu et al. 2018; Stark et al. 2015). These results speak to the strong conservation of the spliceosomal machinery, and of the selective pressures to maintain this essential eukaryotic process despite immense evolutionary distances between eukaryotes. Although most of the five snRNPs and larger accessory complexes are rather well conserved, we observed a reduction in the RES complex and other accessory proteins, and significant divergence exists in the homologs of U1 snRNP components. Reductions in these particles may explain the high levels of intron retention we observed in *P. purpureum*, and collectively these absences could contribute to reduced spliceosomal function in this organism. Further investigations into pre-mRNA splicing in this red alga will help us understand which parts of the spliceosome are essential and which ones are dispensable—thereby revealing mechanistic details about splicing—as well as how organisms can tolerate such high levels of intron retention.

## ACKNOWLEDGMENTS

This work was supported by National Sciences and Engineering Research Council of Canada Discovery Grants [262988 to N.M.F., 298521 to S.D.R.] and a grant to the Centre for Microbial Diversity and Evolution from the Tula Foundation. The authors also acknowledge D. Tack (formerly University of British Columbia, currently Penn State) for his custom scripts used in the analysis of our RNA-Seq data, and T. Whelan (University of British Columbia) for helpful discussions and advice.

## DATA AVAILABILITY STATEMENT

Raw reads generated from all sequencing runs have been deposited to NCBI's Sequence Read Archive (SRA) under BioProject accession PRJNA668394.

## LITERATURE CITED

- Achsel, T., Stark, H. & Lührmann, R. 2001. The Sm domain is an ancient RNA-binding motif with oligo(U) specificity. *Proc. Natl. Acad. Sci. USA*, 98(7):3685–3689.
- Bao, P., Will, C. L., Urlaub, H., Boon, K. L. & Lührmann, R. 2017. The RES complex is required for efficient transformation of the precatalytic B spliceosome into an activated B<sup>act</sup> complex. *Genes Dev.*, 31(23–24):2416–2429.
- Belkorchia, A., Pombert, J. F., Polonais, V., Parisot, N., Delbac, F., Brugère, J. F., Peyret, P., Gaspin, C. & Peyretailade, E. 2017. Comparative genomics of microsporidian genomes reveals a minimal non-coding RNA set and new insights for transcription in minimal eukaryotic genomes. *DNA Res*, 24(3):251–260.
- Berglund, J. A., Chua, K., Abovich, N., Reed, R. & Rosbash, M. 1997. The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell*, 89(5):781–787.
- Bhattacharya, D., Price, D. C., Chan, C. X., Qiu, H., Rose, N., Ball, S., Weber, A. P., Arias, M. C., Henrissat, B., Coutinho, P. M., Krishnan, A., Zäuner, S., Morath, S., Hilliou, F., Egizi, A., Perrineau, M. M. & Yoon, H. S. 2013. Genome of the red alga *Porphyridium purpureum*. *Nat Commun.*, 4:1941.
- Brow, D. A. & Guthrie, C. 1980. Spliceosomal RNA U6 is remarkably conserved from yeast to mammals. *Nature*, 334(6179):213–218.
- Chan, S. P. & Cheng, S. C. 2005. The Prp19-associated complex is required for specifying interactions of U5 and U6 with pre-mRNA during spliceosome activation. *J. Biol. Chem.*, 280(35):31190–31199.
- Chan, S. P., Kao, D. I., Tsai, W. Y. & Cheng, S. C. 2003. The Prp19p-associated complex in spliceosome activation. *Science*, 302(5643):279–282.
- Collén, J., Porcel, B., Carré, W., Ball, S. G., Chaparro, C., Tonon, T., Barbeyron, T., Michel, G., Noel, B., Valentin, K., Elias, M., Artiguenave, F., Arun, A., Aury, J. M., Barbosa-Neto, J. F., Bothwell, J. H., Bouget, F. Y., Brillet, L., Cabello-Hurtado, F., Capella-Gutiérrez, S., Charrier, B., Cladière, L., Cock, J. M., Coelho, S. M., Colleoni, C., Czjzek, M., Da Silva, C., Delage, L., Denoeud, F., Deschamps, P., Dittami, S. M., Gabaldón, T., Gachon, C. M., Groisillier, A., Hervé, C., Jabbari, K., Katinka, M., Kloareg, B., Kowalczyk, N., Labadie, K., Leblanc, C., Lopez, P. J., McLachlan, D. H., Meslet-Cladiere, L., Moustafa, A., Nehr, Z., Nyvall Collén, P., Panaud, O., Partensky, F., Poulain, J., Rensing, S. A., Rousvoal, S., Samson, G., Symeonidi, A., Weissenbach, J., Zambounis, A., Wincker, P. & Boyen, C. 2013. Genome structure and metabolic features in the red seaweed *Chondrus crispus* shed light on evolution of the Archaeplastida. *Proc. Natl. Acad. Sci. USA*, 110(13):5247–5252.
- Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. 2004. WebLogo: a sequence logo generator. *Genome Res.*, 14:1188–1190.
- Dávila López, M., Alm Rosenblad, M. & Samuelsson, T. 2008. Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. *Nucleic Acids Res.*, 36(9):3001–3010.
- de Francisco Amorim, M., Willing, E. M., Szabo, E. X., Francisco-Mangilet, A. G., Droste-Borel, I., Maček, B., Schneeberger, K. & Laubinger, S. 2018. The U1 snRNP subunit LUC7 modulates plant development and stress responses via regulation of alternative splicing. *Plant Cell*, 30(11):2838–2854.
- Delaney, K. J., Williams, S. G., Lawler, M. & Hall, K. B. 2014. Climbing the vertebrate branch of U1A/U2B" protein evolution. *RNA*, 20(7):1035–1045.

- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.
- Douglas, S., Zauner, S., Fraunholz, M., Beaton, M., Penny, S., Deng, L. T., Wu, X., Reith, M., Cavalier-Smith, T. & Maier, U. G. 2001. The highly reduced genome of an enslaved algal nucleus. *Nature*, 410(6832):1091–1096.
- Dziembowski, A., Ventura, A. P., Rutz, B., Caspary, F., Faux, C., Halgand, F., Lapr evote, O. & S eraphin, B. 2004. Proteomic analysis identifies a new complex required for nuclear pre-mRNA retention and splicing. *EMBO J.*, 23(24):4847–4856.
- Fabrizio, P., Dannenberg, J., Dube, P., Kastner, B., Stark, H., Urlaub, H. & L uhrmann, R. 2009. The evolutionarily conserved core design of the catalytic activation step of the yeast spliceosome. *Mol. Cell*, 36(4):593–608.
- Fast, N. M., Roger, A. J., Richardson, C. A. & Doolittle, W. F. 1998. U2 and U6 snRNA genes in the microsporidian *Nosema locustae*: evidence for a functional spliceosome. *Nucleic Acids Res.*, 26(13):3202–3207.
- Fernandez, J. P., Moreno-Mateos, M. A., Gohr, A., Miao, L., Chan, S. H., Irimia, M. & Giraldez, A. J. 2018. RES complex is associated with intron definition and required for zebrafish early embryogenesis. *PLoS Genet.*, 14(7):e1007473.
- Fortes, P., Bilbao-Cort es, D., Fornerod, M., Rigaut, G., Raymond, W., S eraphin, B. & Mattaj, I. W. 1999. Luc7p, a novel yeast U1 snRNP protein with a role in 5' splice site recognition. *Genes Dev.*, 13(18):2425–2438.
- Frank, D. N., Roiha, H. & Guthrie, C. 1994. Architecture of the U5 small nuclear RNA. *Mol. Cell. Biol.*, 14(3):2180–2190.
- Freund, M., Hicks, M. J., Konermann, C., Otte, M., Hertel, K. J. & Schaal, H. 2005. Extended base pair complementarity between U1 snRNA and the 5' splice site does not inhibit splicing in higher eukaryotes, but rather increases 5' splice site recognition. *Nucleic Acids Res.*, 33(16):5112–5119.
- Gilson, P. R., Su, V., Slamovits, C. H., Reith, M. E., Keeling, P. J. & McFadden, G. I. 2006. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. *Proc. Natl. Acad. Sci. USA*, 103(25):9566–9571.
- Gottschalk, A., Tang, J., Puig, O., Salgado, J., Neubauer, G., Colot, H. V., Mann, M., S eraphin, B., Rosbash, M., L uhrmann, R. & Fabrizio, P. 1998. A comprehensive biochemical and genetic analysis of the yeast U1 snRNP reveals five novel proteins. *RNA*, 4(4):374–393.
- Grainger, R. J. & Beggs, J. D. 2005. Prp8 protein: at the heart of the spliceosome. *RNA*, 11(5):533–557.
- Grisdale, C. J., Bowers, L. C., Didier, E. S. & Fast, N. M. 2013. Transcriptome analysis of the parasite *Encephalitozoon cuniculi*: an in-depth examination of the pre-mRNA splicing in a reduced eukaryote. *BMC Genom.*, 14:207–215.
- Hamm, J. & Mattaj, I. W. 1989. An abundant U6 snRNP found in germ cells and embryos of *Xenopus laevis*. *EMBO J.*, 8(13):4179–4187.
- Hinz, M., Moore, M. J. & Bindereif, A. 1996. Domain analysis of human U5 RNA. Cap trimethylation, protein binding, and spliceosome assembly. *J Biol Chem.*, 271(31):19001–19007.
- Hogg, R., McGrail, J. C. & O'Keefe, R. T. 2010. The function of the NineTeen Complex (NTC) in regulating spliceosome conformations and fidelity during pre-mRNA splicing. *Biochem. Soc. Trans.*, 38(4):1110–1115.
- Hudson, A. J., McWatters, D. C., Bowser, B. A., Moore, A. N., Larue, G. E., Roy, S. W. & Russell, A. G. 2019. Patterns of conservation of spliceosomal intron structures and spliceosome divergence in representatives of the diplomonad and parabasalid lineages. *BMC Evol. Biol.*, 19:162.
- Hudson, A. J., Stark, M. R., Fast, N. M., Russell, A. G. & Rader, S. D. 2015. Splicing diversity revealed by reduced spliceosomes in *C. merolae* and other organisms. *RNA Biol.*, 12(11):1–8.
- Irimia, M., Penny, D. & Roy, S. W. 2007. Coevolution of genomic intron number and splice sites. *Trends Genet.*, 23(7):321–325.
- Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., Bateman, A., Finn, R. D. & Petrov, A. I. 2018. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.*, 46(D1):D335–D342.
- Kalyna, M. & Barta, A. 2004. A plethora of plant serine/arginine-rich proteins: redundancy or evolution of novel gene functions? *Biochem. Soc. Trans.*, 32(4):561–564.
- Kandels-Lewis, S. & S eraphin, B. 1993. Involvement of U6 snRNA in 5' splice site selection. *Science*, 262(5142):2035–2039.
- Katinka, M. D., Duprat, S., Cornillot, E., M et enier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretailade, E., Brottier, P., Wincker, P., Delbac, F., El Alaoui, H., Peyret, P., Saurin, W., Gouy, M., Weissenbach, J. & Vivar es, C. P. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature*, 414(6862):450–453.
- Kondo, Y., Oubridge, C., van Roon, A. M. & Nagai, K. 2015. Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. *eLife*, 4:e04986.
- Kretzner, L., Krol, A. & Rosbash, M. 1990. *Saccharomyces cerevisiae* U1 small nuclear RNA secondary structure contains both universal and yeast-specific domains. *Proc. Natl. Acad. Sci. USA*, 87(2):851–855.
- Krol, A., Branlant, C., Lazar, E., Gallinaro, H. & Jacob, M. 1988. Primary and secondary structures of chicken, rat and man nuclear U4 RNAs. Homologies with U1 and U5 RNAs. *Nucleic Acids Res.*, 9:2699–2716.
- Langford, C. J. & Gallwitz, D. 1983. Evidence for an intron-contained sequence required for the splicing of yeast RNA polymerase II transcripts. *Cell*, 33:519–527.
- Lee, J., Kim, D., Bhattacharya, D. & Yoon, H. S. 2019. Expansion of phycobilisome linker gene families in mesophilic red algae. *Nat. Commun.*, 10(1):4823.
- Li, X., Liu, S., Jiang, J., Zhang, L., Espinosa, S., Hill, R. C., Hansen, K. C., Zhou, H. & Zhao, R. 2017. CryoEM structure of *Saccharomyces cerevisiae* U1 snRNP offers insight into alternative splicing. *Nat. Commun.*, 8:1035.
- Lockhart, S. R. & Rymond, B. C. 1994. Commitment of yeast pre-mRNA to the splicing pathway requires a novel U1 small nuclear ribonucleoprotein polypeptide, Prp39p. *Mol. Cell Biol.*, 14(6):3623–3633.
- Lorkovi c, Z. J., Wieczorek Kirk, D. A., Lambermon, M. H. L. & Filipowicz, W. 2000. Pre-mRNA splicing in higher plants. *Trends Plant Sci.*, 5(4):160–167.
- Madhani, H. D. & Guthrie, C. 1992. A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome. *Cell*, 71(5):803–817.
- Marz, M., Kirsten, T. & Stadler, P. F. 2008. Evolution of spliceosomal snRNA genes in metazoan animals. *J. Mol. Evol.*, 67(6):594–607.
- Matsuzaki, M., Misumi, O., Shin-I, T., Maruyama, S., Takahara, M., Miyagishima, S., Mori, T., Nishida, K., Yagisawa, F., Yoshida, Y., Nishimura, Y., Nakao, S., Kobayashi, T., Momoyama, Y., Higashiyama, T., Minoda, A., Sano, M., Nomoto, H., Oishi, K., Hayashi, H., Ohta, F., Nishizaka, S., Haga, S., Miura, S., Morishita, T., Kabeya, Y., Terasawa, K., Suzuki, Y., Ishii, Y., Asakawa, S., Takano, H., Ohta, N., Kuroiwa, H., Tanaka, K., Shimizu, N., Sugano, S., Sato, N., Nozaki, H., Ogasawara, N., Kohara, Y. & Kuroiwa, T. 2004. Genome

- sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature*, 428(6983):653–657.
- Mitrovich, Q. M. & Guthrie, C. 2007. Evolution of small nuclear RNAs in *S. cerevisiae*, *C. albicans*, and other hemiascomycetous yeasts. *RNA*, 13(12):2066–2080.
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(1):621–628.
- Muto, Y., Pomeranz Krummel, D., Oubridge, C., Hernandez, H., Robinson, C. V., Neuhaus, D. & Nagai, K. 2004. The structure and biochemical properties of the human spliceosomal protein U1C. *J. Mol. Biol.*, 341(1):185–198.
- Myslinski, E. & Branlant, C. 1991. A phylogenetic study of U4 snRNA reveals the existence of an evolutionarily conserved secondary structure corresponding to 'free' U4 snRNA. *Biochimie*, 73(1):17–28.
- Nagai, K., Oubridge, C., Jessen, T. H., Li, J. & Evans, P. R. 1990. Crystal structure of the RNA-binding domain of the U1 small nuclear ribonucleoprotein A. *Nature*, 348(6301):515–520.
- Nawrocki, E. P. & Eddy, S. R. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935.
- Nelissen, R. L. H., Heinrichs, V., Habets, W. J., Simons, F., Lührmann, R. & van Venrooij, W. J. 1991. Zinc finger-like structure in U1-specific protein C is essential for specific binding to U1 snRNP. *Nucleic Acids Res.*, 19(3):449–454.
- Ohi, M. D., Link, A. J., Ren, L., Jennings, J. L., McDonald, W. H. & Gould, K. L. 2002. Proteomics analysis reveals stable multi-protein complexes in both fission and budding yeasts containing Myb-related Cdc5p/Cef1p, novel pre-mRNA splicing factors, and snRNAs. *Mol. Cell Biol.*, 22(7):2011–2024.
- Oubridge, C., Ito, N., Evans, P. R., Teo, C. H. & Nagai, K. 1994. Crystal structure at 1.92 Å resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature*, 372:432–438.
- Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Halen, L., Krobtsch, S., Lehrach, H. & Soldatov, A. 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.*, 37(18):e123.
- Polycarpou-Schwarz, M., Gunderson, S. I., Kandels-Lewis, S., Séraphin, B. & Mattaj, I. W. 1996. Drosophila SNF/D25 combines the functions of the two snRNP proteins U1A and U2B' that are encoded separately in human, potato, and yeast. *RNA*, 2(1):11–23.
- Pomeranz Krummel, D. A., Oubridge, C., Leung, A. K., Li, J. & Nagai, K. 2009. Crystal structure of human spliceosomal U1 snRNP at 5.5 Å resolution. *Nature*, 458(7237):475–480.
- Price, S. R., Evans, P. R. & Nagai, K. 1998. Crystal structure of the spliceosomal U2B''–U2A' protein complex bound to a fragment of U2 small nuclear RNA. *Nature*, 397:645–650.
- Puig, O., Bragado-Nilsson, E., Koski, T. & Séraphin, B. 2007. The U1 snRNP-associated factor Luc7p affects 5' splice site selection in yeast and human. *Nucleic Acids Res.*, 35(17):5874–5885.
- Puig, O., Gottschalk, A., Fabrizio, P. & Séraphin, B. 1999. Interaction of the U1 snRNP with nonconserved intronic sequences affects 5' splice site selection. *Genes Dev.*, 13(5):569–580.
- Qiu, H., Price, D. C., Yang, E. C., Yoon, H. S. & Bhattacharya, D. 2015. Evidence of ancient genome reduction in red algae (*Rhodophyta*). *J. Phycol.*, 51:624–636.
- Qiu, H., Rossoni, A. W., Weber, A. P. M., Yoon, H. S. & Bhattacharya, D. 2018. Unexpected conservation of the RNA splicing apparatus in the highly streamlined genome of *Galdieria sulphuraria*. *BMC Evol. Biol.*, 18(1):41.
- Reimer, K. A., Stark, M. R., Aguilar, L. C., Stark, S. R., Burke, R. D., Moore, J., Fahlman, R. P., Yip, C. K., Kuroiwa, H., Oeffinger, M. & Rader, S. D. 2017. The sole LSm complex in *Cyanidioschyzon merolae* associates with pre-mRNA splicing and mRNA degradation factors. *RNA*, 23(6):952–967.
- Roy, J., Kim, K., Maddock, J. R., Anthony, J. G. & Woolford Jr, J. L. 1995. The final stages of spliceosome maturation require Spp2p that can interact with the DEAH box protein Prp2p and promote step 1 of splicing. *RNA*, 1(4):375–390.
- Sahi, C., Lee, T., Inada, M., Pleiss, J. A. & Craig, E. A. 2010. Cwc23, an essential J protein critical for pre-mRNA splicing with a dispensable J domain. *Mol. Cell Biol.*, 30(1):33–42.
- Saldi, T., Wilusz, C., MacMorris, M. & Blumenthal, T. 2007. Functional redundancy of worm spliceosomal proteins U1A and U2B''. *Proc. Natl. Acad. Sci. USA*, 104(23):9753–9757.
- Sawa, H. & Shimura, Y. 1992. Association of U6 snRNA with the 5'-splice site region of pre-mRNA in the spliceosome. *Genes Dev.*, 6(2):244–254.
- Scherly, D., Boelens, W., van Venrooij, W. J., Dathan, N. A., Hamm, J. & Mattaj, I. W. 1989. Identification of the RNA binding segment of human U1 A protein and definition of its binding site on U1 snRNA. *EMBO J.*, 8(13):4163–4170.
- Schönknecht, G., Chen, W. H., Ternes, C. M., Barbier, G. G., Shrestha, R. P., Stanke, M., Bräutigam, A., Baker, B. J., Banfield, J. F., Garavito, R. M., Carr, K., Wilkerson, C., Rensing, S. A., Gagneul, D., Dickenson, N. E., Oesterhelt, C., Lercher, M. J. & Weber, A. P. 2013. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science*, 339(6124):1207–1210.
- Siliciano, P. G., Brow, D. A., Roiha, H. & Guthrie, C. 1987. An essential snRNA from *S. cerevisiae* has properties predicted for U4, including interaction with a U6-like snRNA. *Cell*, 50:585–592.
- Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212.
- Simpson, G. G., Clark, G. P., Rothnie, H. M., Boelens, W., van Venrooij, W. & Brown, J. W. 1995. Molecular characterization of the spliceosomal proteins U1A and U2B'' from higher plants. *EMBO J.*, 14(18):4540–4550.
- Slabodnick, M. M., Ruby, J. G., Reiff, S. B., Swart, E. C., Gosai, S., Prabakaran, S., Witkowska, E., Larue, G. E., Fisher, S., Freeman Jr, R. M., Gunawardena, J., Chu, W., Stover, N. A., Gregory, B. D., Nowacki, M., Derisi, J., Roy, S. W., Marshall, W. F. & Sood, P. 2017. The macronuclear genome of *Stentor coeruleus* reveals tiny introns in a giant cell. *Curr. Biol.*, 27(4):569–575.
- Stark, H., Dube, P., Lührmann, R. & Kastner, B. 2001. Arrangement of RNA and proteins in the spliceosomal U1 small nuclear ribonucleoprotein particle. *Nature*, 409:539–542.
- Stark, M. R., Dunn, E. A., Dunn, W. S. C., Grisdale, C. J., Daniele, A. R., Halstead, M. R. G., Fast, N. M. & Rader, S. D. 2015. Dramatically reduced spliceosome in *Cyanidioschyzon merolae*. *Proc. Natl. Acad. Sci. USA*, 112(11):E1191–E1200.
- Suzuki, S., Shirato, S., Hirakawa, Y. & Ishida, K.-I. 2015. Nucleomorph genome sequences of two chlorarachniophytes, *Amorphochlora amoebiformis* and *Lotharella vacuolata*. *Genome Biol. Evol.*, 7(6):1533–1545.
- Tanifuji, G., Onodera, N. T., Brown, M. W., Curtis, B. A., Roger, A. J., Wong, G. K., Melkonian, M. & Archibald, J. M. 2014. Nucleomorph and plastid genome sequences of the chlorarachniophyte *Lotharella oceanica*: convergent reductive evolution and frequent recombination in nucleomorph-bearing algae. *BMC Genom.*, 15(1):374.

- Urlaub, H., Hartmuth, K., Kostka, S., Grelle, G. & Lührmann, R. 2000. A general approach for identification of RNA-protein cross-linking sites within native human spliceosomal small nuclear ribonucleoproteins (snRNPs). Analysis of RNA-protein contacts in native U1 and U4/U6.U5 snRNPs. *J. Biol. Chem.*, 275(52):41458–41468.
- Wahl, M. C., Will, C. L. & Lührmann, R. 2009. The spliceosome: design principles of a dynamic RNP machine. *Cell*, 136(4):701–718.
- Wang, L., Si, Y., Dedow, L. K., Shao, Y., Liu, P. & Brutnell, T. P. 2011. A low-cost library construction protocol and data analysis pipeline for Illumina-based strand-specific multiplex RNA-seq. *PLoS One*, 6(10):e26426.
- Ward, N. & Moreno-Hagelsieb, G. 2014. Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss? *PLoS One*, 9(7):e101850.
- Warkocki, Z., Schneider, C., Mozaffari-Jovin, S., Schmitzová, J., Höbartner, C., Fabrizio, P. & Lührmann, R. 2015. The G-patch protein Spp2 couples the spliceosome-stimulated ATPase activity of the DEAH-box protein Prp2 to catalytic activation of the spliceosome. *Genes Dev.*, 29(1):94–107.
- Whelan, T. A., Lee, N. T., Lee, R. C. H. & Fast, N. M. 2019. Microsporidian introns retained against a background of genome reduction: characterization of an unusual set of introns. *Genome Biol. Evol.*, 11(1):263–269.
- Will, C. L. & Lührmann, R. 2011. Spliceosome structure and function. *Acc. Chem. Res.*, 44(12):1292–1301.
- Williams, S. G. & Hall, K. B. 2011. Human U2B<sup>''</sup> protein binding to snRNA stemloops. *Biophys. Chem.*, 159(1):82–89.
- Williams, S. G., Harms, M. J. & Hall, K. B. 2013. Resurrection of an Urbilaterian U1A/U2B<sup>''</sup>/SNF protein. *J. Mol. Biol.*, 425(20):3846–3862.
- Wong, D. K., Gridale, C. J. & Fast, N. M. 2018. Evolution and diversity of pre-mRNA splicing in highly reduced nucleomorph genomes. *Genome Biol. Evol.*, 10(6):1573–1583.
- Wu, J. A. & Manley, J. L. 1991. Base pairing between U2 and U6 snRNAs is necessary for splicing of a mammalian pre-mRNA. *Nature*, 352(6338):818–821.
- Xiong, F., Ren, J.-J., Yu, Q., Wang, Y.-Y., Kong, L.-J., Otegui, M. S. & Wang, X.-L. 2019. *AtBUD13* affects pre-mRNA splicing and is essential for embryo development in Arabidopsis. *Plant J.*, 98:714–726.
- Zang, S., Lin, T. Y., Chen, X., Gencheva, M., Newo, A. N., Yang, L., Rossi, D., Hu, J., Lin, S. B., Huang, A. & Lin, R. J. 2014. GPKOW is essential for pre-mRNA splicing in vitro and suppresses splicing defect caused by dominant-negative DHX16 mutation in vivo. *Biosci. Rep.*, 34(6):e00163.
- Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–3415.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**File S1.** Read counts within the vicinity of 967 *Porphyridium purpureum* introns with adequate read coverage, and their corresponding percent of spliced reads.

**File S2.** Database for BLAST searches containing spliceosomal protein homologs from *Homo sapiens*, *Saccharomyces cerevisiae*, *Cyanidioschyzon merolae*, and *Arabidopsis thaliana*.

**File S3.** Read counts and percent of spliced reads for the focused set of 192 *P. purpureum* introns, and their corresponding host gene identity and expression levels.

**File S4.** Best Infernal candidates for each of the five snRNAs (U1, U2, U4, U5, and U6) in *P. purpureum*.

**File S5.** Unfiltered reciprocal BLASTp and tBLASTn/BLASTx results for spliceosomal proteins in *P. purpureum*.

**File S6.** Alignments of all identified U1 snRNP proteins in *P. purpureum*, and two selected proteins from each U2, U4/U6, and U5 snRNPs.

**Table S1.** Complete table of confirmed spliceosomal components in *P. purpureum*, including Sm, LSm, and other miscellaneous splicing factors.

**Methods S1.** Additional details on methods used in this study.