DOI: 10.1111/jeu.12927

ORIGINAL ARTICLE

The evolution of pre-mRNA splicing and its machinery revealed by reduced extremophilic red algae

Donald K. Wong[1] | **Cameron J. Grisdale**[1,2] | **Viktor A. Slat**[3] | **Stephen D. Rader**[3] | **Naomi M. Fast**[1]

[1]Biodiversity Research Centre and Department of Botany, University of British Columbia, Vancouver, BC, Canada

[2]Michael Smith Genome Sciences Centre, Vancouver, BC, Canada

[3]Department of Chemistry, University of Northern British Columbia, Prince George, BC, Canada

Correspondence
Naomi M. Fast, Department of Botany, University of British Columbia, 3156-6270 University Boulevard, Vancouver, BC, Canada.
Email: naomi.fast@ubc.ca

Funding information
Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: 298521 and 262988; Tula Foundation

Abstract

The Cyanidiales are a group of mostly thermophilic and acidophilic red algae that thrive near volcanic vents. Despite their phylogenetic relationship, the reduced genomes of *Cyanidioschyzon merolae* and *Galdieria sulphuraria* are strikingly different with respect to pre-mRNA splicing, a ubiquitous eukaryotic feature. Introns are rare and spliceosomal machinery is extremely reduced in *C. merolae*, in contrast to *G. sulphuraria*. Previous studies also revealed divergent spliceosomes in the mesophilic red alga *Porphyridium purpureum* and the red algal derived plastid of *Guillardia theta* (Cryptophyta), along with unusually high levels of unspliced transcripts. To further examine the evolution of splicing in red algae, we compared *C. merolae* and *G. sulphuraria*, investigating splicing levels, intron position, intron sequence features, and the composition of the spliceosome. In addition to identifying 11 additional introns in *C. merolae*, our transcriptomic analysis also revealed typical eukaryotic splicing in *G. sulphuraria*, whereas most transcripts in *C. merolae* remain unspliced. The distribution of intron positions within their host genes was examined to provide insight into patterns of intron loss in red algae. We observed increasing variability of 5′ splice sites and branch donor regions with increasing intron richness. We also found these relationships to be connected to reductions in and losses of corresponding parts of the spliceosome. Our findings highlight patterns of intron and spliceosome evolution in related red algae under the pressures of genome reduction.

KEYWORDS
Cyanidiales, intron, spliceosome, transcriptome

THE Rhodophyta, or red algae, are a large and diverse group of organisms whose plastids derive from the primary endosymbiotic event that brought photosynthesis to the eukaryotic domain. Despite their name, an early-diverging branch of the red algae includes species that are not red. These are the Cyanidiales, whose members are mostly thermophilic and acidophilic, and are often found around volcanic vents. They lack the red pigments found in most rhodophytes and are unicellular. Regardless, the Cyanidiales and red algae "proper" both share a lack of chlorophylls *b* and *c*, flagella and centrioles in any life stage (Seckbach et al., 2010). Due to their general resilience to a variety of growth conditions, Cyanidiales have been proposed as ideal candidates to use in the development of biofuels. The ability of some of these species to uptake heavy metals with little ill effect

could also be exploited in bioremediation processes (Schönknecht et al., 2013). The genomes of two members of the Cyanidiales, *Cyanidioschyzon merolae* and *Galdieria sulphuraria*, have been sequenced (Matsuzaki et al., 2004; Nozaki et al., 2007; Schönknecht et al., 2013). Although these two species are ecologically similar to extremophiles, large phylogenetic distances separate them (Yoon et al., 2006).

In addition to the absence of flagella and centrioles across all red algae, many other gene families common to all eukaryotes are also reduced or lost in red algae. These absences have led to hypotheses that the ancestor of Cyanidiales and other red algae was anciently reduced (Qiu et al., 2015). Some species of red algae also exhibit reductions in noncoding RNA, such as spliceosomal introns. The genome of the macroalga *Chondrus crispus* is intron-poor with an average of 0.32 introns per gene (Collen et al., 2013), and intron scarcity was thought to be common across all red algae. Other sequenced macroalgal genomes, such as those of *Gracilaria changii* or *Pyropia yezoensis*, are also intron-poor (Ho et al., 2018; Wang et al., 2020). However, more recently sequenced genomes have many more introns, although not at the density seen in green algae or metazoans.

The most intron-poor red alga sequenced is *Cyanidioschyzon merolae*, a tiny unicellular extremophile that thrives in acidic hot springs (De Luca et al., 1978). Cells of *C. merolae* have a very simple architecture with a single nucleus, mitochondrion, and plastid—all of which can be made to divide synchronously under 12 h:12 h light:dark cycles (Terui et al., 1995). The genome of *C. merolae* has been completely sequenced (Matsuzaki et al., 2004; Nozaki et al., 2007), and is the best-studied genome of any red alga. Its 16.5 Mbp genome is among the most reduced of the red algae, with 4775 protein-coding genes on its 20 chromosomes (Matsuzaki et al., 2004; Nozaki et al., 2007). The reduction has also shaped the *C. merolae* genome in a number of ways that make it unique among eukaryotes. It has three nonrepeat rRNA (18S-5.8S-28S) units, instead of the typical tandem repeat rRNA units of most eukaryotes (Matsuzaki et al., 2004). Along with three nearly identical copies of the 5S rRNA, this makes the rRNA complement of *C. merolae* the smallest known (Matsuzaki et al., 2004). Widespread intron loss has also occurred, leaving just 27 introns in 26 genes (Matsuzaki et al., 2004). Why *C. merolae* has retained so few introns, requiring dozens of spliceosome components for their removal, remains a mystery.

Galdieria sulphuraria possesses an even smaller genome than *C. merolae* at 13.7 Mbp (Schönknecht et al., 2013). However, *G. sulphuraria* has nearly 2000 more protein-coding genes and much shorter intergenic regions than *C. merolae*, with an extensive complement of genes thought to be derived by horizontal gene transfer from prokaryotes (Schönknecht et al., 2013). In a dramatic contrast with *C. merolae*, most *G. sulphuraria* genes are interrupted by introns—on average, each gene is split by two. In total, *G. sulphuraria* possesses more than 13,000 annotated introns, making it the most intron-rich red alga sequenced thus far. This extreme difference in intron density between *C. merolae* and *G. sulphuraria* is also reflected in their spliceosomal machinery.

The spliceosome is comprised of five core small nuclear ribonucleoprotein complexes (snRNPs)—each of these a complex of proteins and its corresponding small nuclear RNA (snRNA). These assemble around the intron, with some snRNPs recognizing key sequence motifs of the intron through sequence complementarity to their snRNAs. A number of other proteinaceous complexes are involved in accessory functions, such as activating the spliceosome into its catalytic form, and in debranching the excised intron (Will & Lührmann, 2011). These ultimately result in two transesterification reactions that ligate the two flanking exons and remove the intron. Spliceosomal components are conserved across most eukaryotic lineages, with over 100 spliceosomal protein-coding genes present in yeast and more than 200 in mammals (Jurica & Moore, 2003; Wahl et al., 2009). However, some reduced systems with extremely low intron densities, such as microsporidian parasites and eukaryotic endosymbionts, have extensive reductions in spliceosomal components (Cuomo et al., 2012; Douglas et al., 2001; Katinka et al., 2001; Lane et al., 2007). As mentioned previously, the *C. merolae* genome is intron-poor, and investigations into the *C. merolae* spliceosome found 45 core splicing proteins, far fewer than are found in the yeast genome (Hudson et al., 2019; Reimer et al., 2017; Stark et al., 2015). Surprisingly, no U1 snRNA or U1-associated proteins were found in *C. merolae*, suggesting a complete loss of the U1 snRNP (Stark et al., 2015). In contrast, a complex spliceosome of 150 proteins was identified in the intron-rich *G. sulphuraria* (Qiu et al., 2018), suggesting the importance of a complex spliceosome for efficient recognition of many introns in an intron-rich genome.

Previous studies have linked spliceosomal complexity to splicing efficiency. An analysis of the highly reduced microsporidian parasite *Encephalitozoon cuniculi* revealed high levels of unspliced transcripts, or intron retention, in this extremely intron-poor species with a reduced spliceosome, suggesting that inefficient splicing (or mis-splicing) could be the result of spliceosomal component losses (Grisdale et al., 2013). High levels of unspliced transcripts were also reported in the red algal-derived nucleomorph of the cryptomonad *Guillardia theta* and the unicellular mesophilic red alga *Porphyridium purpureum* (Wong et al., 2018, 2021). As in *E. cuniculi*, both the *G. theta* nucleomorph and *P. purpureum* are intron-poor (Douglas et al., 2001; Lee et al., 2019; Wong et al., 2021), although the spliceosome of *P. purpureum* is more divergent rather than reduced (Wong et al., 2021). Since both *P. purpureum* and the *G. theta* nucleomorph represent red algal genomes with unusually high levels of unspliced transcripts (and divergent or reduced spliceosomes), we

examined these aspects of pre-mRNA splicing in the extremophilic Cyanidiales to provide insight into the evolution of splicing, transcription and gene regulation among red algae and, more broadly, in eukaryotes with reduced genomes.

In this study, we generated transcriptomes from the extremophiles *C. merolae* and *G. sulphuraria* to compare various aspects of pre-mRNA splicing. Our transcriptomic data uncovered 11 additional introns not yet annotated in the genome of *C. merolae*, increasing its total number of introns from 27 to 38. We also showed a dramatic contrast in the levels of pre-mRNA splicing between the two Cyanidiales species. Very strict sequences for the 5′ splice site and branch donor regions were observed for *C. merolae* introns, supporting the trend of decreased sequence variability in intron-poor organisms. Finally, a comparison of spliceosome content and snRNA structures highlighted the stark differences in pre-mRNA splicing between *C. merolae* and *G. sulphuraria*.

MATERIALS AND METHODS

Cell culture and RNA preparation

Cultures of *C. merolae* 10D (NIES strain 1332) were obtained from the NIES Microbial Culture Collection (Japan), while cultures of *G. sulphuraria* 074 W were generously provided by the Weber Lab (Heinrich-Heine-University, Düsseldorf, Germany). Cells from each species then seeded at least two replicate cultures that were then subcultured separately. Both *C. merolae* and *G. sulphuraria* were grown in 50 ml volumes of modified Allen's (MA2, pH 2.5) media (Minoda et al., 2004) within 250 ml Erlenmeyer flasks atop an orbital shaker rotating at 120 rpm. These were subjected to 90 µmol photons/m^2/s of light on a 12 h:12 h light: dark cycle at 40 °C. Cells from dense, late exponential phase *C. merolae* cultures were collected by centrifugation midway through both the light and dark cycles after 3 weeks of growth. Cultures of exponential phase *G. sulphuraria* were similarly pelleted midway through the light cycle after 6 weeks of growth. All cultures were grown without media exchange until extraction. Total RNA was subsequently extracted from two biological replicates of *C. merolae* for the light and dark conditions, using the Ambion RNAqueous kit as per the manufacturer's protocols (Ambion, Thermo Fisher Scientific). The tough cell wall of *G. sulphuraria* necessitated the use of more vigorous methods of cellular disruption. Pellets of *G. sulphuraria* were resuspended in 1 ml TRIzol reagent (Ambion, Thermo Fisher Scientific) and 200 µl of Lysing Matrix Y beads (MP Biomedicals), and then homogenized with the use of a bead mill (VWR). Total RNA extracts were quantified using a NanoDrop spectrophotometer (Thermo Fisher Scientific). To eliminate any contaminating DNA, extracted RNA was treated with Invitrogen DNA-free DNA Removal Kit (Thermo Fisher Scientific) according to the manufacturer's protocols. A QuBit 2.0 fluorimeter (Invitrogen, Thermo Fisher Scientific) was then used to more accurately quantify RNA, and to ensure samples were free of DNA. The DNA-free RNA was then enriched for mRNA through poly-A purification using NEXTflex Poly(A) Beads (BioO Scientific) without modification to manufacturer protocols.

RNA-Seq library preparation and sequencing

A total of eight Illumina libraries were prepared for *C. merolae*. Four libraries—two biological replicates of both light and dark condition cultures to provide statistical confidence for our analyses—were prepared according to the TruSeq library preparation protocol (Illumina), using a total of 4 µg of DNA-depleted RNA. As the Illumina TruSeq libraries do not retain strand information, four additional libraries were made using the Illumina-compatible NEXTflex Directional RNA-Seq kit (BioO Scientific), resulting in a total of 12 libraries for *C. merolae*. A total of 2 µg of RNA was used as input for the directional RNA-Seq libraries. Strand specificity is maintained using the dUTP method, and library preparation was performed according to the kit's protocols without modification. Two stranded libraries representing biological replicates of *G. sulphuraria* were also similarly prepared from poly-A selected RNA. Paired-end sequencing on all libraries was performed on an in-house Illumina HiSeq 2000 (Illumina) at the Biodiversity Research Centre (University of British Columbia, Vancouver, Canada).

Analysis of RNA-Seq data

Data from all sequencing runs were, respectively, mapped to either the *C. merolae* 10D (GenBank accession GCA_000091205.1) or the *G. sulphuraria* 074 W (GenBank accession GCA_000341285.1) reference genomes and annotations. Mapping was performed using the splice-aware STAR mapper (Dobin et al., 2013), and each read was restricted to only a single alignment. Read counts of annotated genes were used to calculate expression level in fragments per kilobase per million mapped reads (FPKM) (Mortazavi et al., 2008). All biological replicates showed high levels of correlation (>0.98); thus, RNA-Seq datasets were appropriately pooled for downstream analyses.

Potential new introns in *C. merolae* were identified through split reads that mapped beyond annotated junction boundaries. We began with a file generated from the STAR mapper (Dobin et al., 2013) that summarizes mapped split reads, which identified a total of 47,932 potential splice junctions. We then filtered these for

splice junctions that have split read support from every mapped *C. merolae* library, which reduced the number of candidate junctions to 5034. Only 760 of these are supported with 10 or more split reads, and because introns in *C. merolae* have an extremely conserved branch donor region, we excluded those that lack the consensus sequence, leaving a total of 48 splice junctions. Finally, after removing already annotated introns, the remaining candidate splice junctions were scrutinized further through visualization using the Integrative Genomics Viewer (Thorvaldsdóttir et al., 2013), taking into consideration aspects such as the location of the potential intron in relation to reading frame. Following this final curation by eye, we identified a total of 11 potential new introns.

RT-PCR confirmation of new introns in *Cyanidioschyzon merolae*

The 27 introns originally annotated in *C. merolae* had previously been confirmed via RT-PCR (Stark et al., 2015). To confirm the novel introns annotated in this study, we similarly amplified the genes containing these new introns. Five genes could not be amplified because of the difficulty of designing primers in short exons and regions of high A/T content. Primer sequences used can be found in Table S1.

Intron sequence feature analysis

Intronic sequences from *C. merolae* and *G. sulphuraria* were retrieved from the reference genome using the coordinates of annotated boundaries. These sequences were then used to generate sequence logos of 5′ splice sites of each species with WebLogo (Crooks et al., 2004). The largely invariable branch donor regions were identified by eye in the extremely intron-sparse *C. merolae*. Branch donor regions of *G. sulphuraria* introns were predicted using BPP (Zhang et al., 2017).

To determine the extent of positional bias of introns, we calculated the fraction of first exon length against the total coding length (total length of exons) of a gene for all single intron protein-coding genes in *C. merolae* and *G. sulphuraria*. We excluded genes with multiple introns from this analysis, as they are more likely to have introns closer to the start and stop codons of its genes.

Assessing splicing efficiency

Mapped reads from both species located in the vicinity (within 90 nucleotides) of each annotated junction were assessed as per Wong et al. (2018, 2021) for the type of splicing event they might represent, such as splicing at annotated boundaries or intron retention. To determine splicing efficiency of any particular junction, splicing levels were calculated by dividing the number of reads split across a splice junction at its annotated boundaries, normalized to the length of the intron in question, by the total number of reads in its vicinity. The resulting percentage provides a measure of the proportion of spliced transcripts vs. intron-retaining transcripts.

For *C. merolae*, splicing values were initially calculated from the stranded data and unstranded data separately. However, there is not a significant difference between the two datasets, as the correlations (Pearson's *r*) between strand-specific and nonstranded splicing levels in light or dark conditions were both 0.94. Thus, splicing levels were combined from the two datasets in our results.

Determining the structures of snRNAs in *Galdieria sulphuraria*

The snRNAs of *G. sulphuraria* have been previously identified by Qiu et al. (2018). Infernal (Nawrocki & Eddy, 2013) provides some prediction of an RNA's secondary structure through an alignment of the candidate to a consensus structure generated from a covariance model. However, certain difficult regions or extremely large inserts may result in a less than ideal structure prediction. To refine these preliminary predictions, we used mfold (Zuker, 2003) to generate additional structures. We then scrutinized these further predictions by eye, taking into account other available structures from related species to generate our final predictions for *G. sulphuraria* snRNA secondary structures.

RESULTS AND DISCUSSION

Novel introns in *Cyanidioschyzon merolae*

To investigate aspects of transcription and splicing in *C. merolae* and *G. sulphuraria*, we sequenced poly-A enriched transcriptomes from each. Twelve RNA-Seq libraries of *C. merolae* were sequenced, with six from RNA sampled during the day and six from the night. Each of these six transcriptomes was comprised of four unstranded and two stranded datasets. Two replicate stranded libraries from RNA extracted during the day were sequenced for *G. sulphuraria*. Transcriptomes were mapped to their respective reference genomes and biologically equivalent datasets were pooled for downstream analyses (see Materials and Methods).

We used our transcriptomic data to assess splicing of the annotated junctions of *C. merolae* and *G. sulphuraria*. Only 27 introns were annotated in the original *C. merolae* assembly, leaving 99.5% of protein-coding genes without introns (Matsuzaki et al., 2004; Nozaki et al., 2007). This is an exceptionally small number of

introns, as higher intron counts can be found in organisms with even smaller genomes; *Saccharomyces cerevisiae* (12.5 Mbp), *Cryptosporidium parvum* (9.1 Mbp), *Ostreococcus tauri* (12.5 Mbp), and *Plasmodium falciparum* (22.8 Mbp) all contain introns in at least 5% of their protein-coding genes (Abrahamsen et al., 2004; Blanc-Mathieu et al., 2014; Derelle et al., 2006; Engel et al., 2013; Gardner et al., 2002; Goffeau et al., 1996; Spingola et al., 1999).

We found evidence for splicing of all 27 of the originally annotated *C. merolae* introns through split reads. Furthermore, we found split reads that mapped elsewhere in the genome, suggesting the presence of additional introns that were missed in the original annotations. Following filtering and additional manual scrutiny (see Materials and Methods), we identified 11 new introns, bringing the total to 38 spliceosomal introns in *C. merolae* (Table 1). We were also able to design primers and confirm splicing for six of these 11 new introns in *C. merolae* through RT-PCR (Figure S1).

With respect to the locations of the 11 new introns, six of these interrupt previously annotated ORFs. Three of these six split the gene in a way that moves the start codon upstream from what was originally annotated, and introns were found within the 3′ untranslated regions of CME196C and CMG006C. Three novel introns were found to interrupt transcripts currently annotated as noncoding, although removal of the intron for CMP330T appears to result in a potential open reading frame. Finally, these new introns bear canonical splicing motifs and are supported by split reads like the original 27 introns, as discussed later. Taken together with evidence provided by RT-PCR, this strongly suggests that these new introns are not resulting from sequencing artifacts or transcriptional "noise," and that these are real introns missed by early annotation pipelines. Regardless, 38 introns across 4775 genes in *C. merolae* is still extremely low.

Applying the same scrutiny to mapped transcriptomic data for the intron-rich *G. sulphuraria*, we identified a very small number of regions where spliced reads did not correspond to annotated junction boundaries. However, these did not appear to represent new introns, as visualization of read depth coverage in these regions indicates simple misannotation of a nearby intron.

Intron positions in *Cyanidioschyzon merolae* and *Galdieria sulphuraria*

Genomic evidence from sequenced red algae suggests that the entire group was ancestrally intron-poor (Qiu et al., 2015). Because the last common ancestor of extant eukaryotes was likely intron-rich with a complex spliceosome (Csuros et al., 2011; Fedorov et al., 2002; Koonin, 2006), red algae represent one of the instances across eukaryotes where massive intron loss has taken place. However, the presence of intron-rich red algae such as *G. sulphuraria* or *Gracilaria changii* (Brawley et al., 2017) confounds this conclusion. Given the phylogenetic position of *G. sulphuraria* with respect to *C. merolae*, either extensive intron gain took place in *Galdieria*, or red algae are in fact not ancestrally intron-poor.

Intron loss has been suggested to occur through a number of different mechanisms, with evidence for each of these provided by the presence of specific sequence "signatures" when compared to homologous regions of closely related organisms. One such mechanism is homologous recombination with the reverse transcript of a gene, which predicts that introns are more likely to be lost at the 3′ end (Fink, 1987; Roy & Gilbert, 2005). This is supported by the preponderance of introns biased toward the 5′ end of a gene in intron-sparse organisms, with some extreme cases where they immediately follow the start codon (Douglas et al., 2001; Katinka et al., 2001; Moore et al., 2012). Another proposed mechanism of intron loss is through genomic deletion, which should not result in any positional bias for the remaining introns, but may be accompanied by evidence of further insertions or deletions in a missing intron's vicinity (Ma et al., 2015). Finally, introns have been shown to be lost during end repair of double-stranded DNA breaks; this mechanism also does not result in positional bias of introns (Farlow et al., 2011). Whether any particular mechanism of intron loss is more likely to occur remains unknown, though perhaps this is related to the ease of detecting the specific genomic signatures of each mechanism. To determine the patterns of intron loss among Cyanidiales, we investigated the positional bias of introns between the intron-dense *G. sulphuraria* (2 introns per gene) and the extremely intron-poor *C. merolae* (0.008 introns per gene). The mesophilic *P. purpureum* was also included for comparison, as it is a non-Cyanidiales red alga of intermediate intron richness, with an average of 0.02 introns per gene.

We analyzed the positions of introns within single intron genes and expressed them as a fraction of the first exon length to the entire length of the spliced gene (see Materials and Methods). These calculated fractions are summarized as violin and box plots in Figure 1 to highlight the distribution of intron positions across genes. In *C. merolae*, its introns are present at a wide range of positions within their host genes, but are most likely to be found in the 5′ half of the coding region (Figure 1). However, the most frequent values for *C. merolae* intron positions lie between 0.3 and 0.4, which indicate only a very slight 5′ bias, in contrast with intron-poor red algal derived genomes (Douglas et al., 2001; Moore et al., 2012). Confirmed introns in *P. purpureum* show a slight bimodal distribution of positions, although introns are most frequently found between 0.3 and 0.4 as in *C. merolae*. In the intron-rich *G. sulphuraria*, the distribution of positions of 1673 introns in single-intron genes is

TABLE 1 Properties of introns of *C. merolae*

Gene	Coordinates	5′ Splice site	Branch point	3′ Splice site	Length	Phase
CMC008C[a]	18,397–18,609	GTAAGTTAGA	AACTAAC	AATACTAG	213	2
CMC053C	122,765–122,933	GTAAGTTGTT	AACTAAC	ACTTGTAG	169	2
CMD067C	170,067–170,151	GTAAGCATAA	AACTAAC	GAACGCAG	85	1
CME034C	95,612–95,761	GTAAGTGTCA	AACTAAC	AACTGCAG	150	1
CME196C[a]	512,966–513,104	GTAAGTTCGC	TACTAAC	TATCCCAG	139	1
CMF072C	188,541–188,625	GCAAGTTGAC	AACTAAC	GAGTCTAG	85	0
CMF136C	383,596–383,777	GTAAGTTTTA	TACTAAC	CGAGTCAG	182	0
CMG006C[a]	16,275–16,413	GTAAGTTCGC	TACTAAC	TATCCCAG	139	0
CMG136C[a]	344,397–344,494	GTAAGTTTTG	GACTAAC	ACGCTCAG	98	1
CMI315Z[a]	583,166–583,408	GTAAGTGAAG	CACTAAC	TCGGACAG	243	ncRNA
CMJ116C[a]	331,315–331,609	GTAAGTCTTC	CACTAAC	TCGGTTAG	295	2
CMJ129C	362,444–362,685	GTAAGTTTTC	AACTAAC	TCAATTAG	242	1
CMK142T[a]	385,008–385,712	GTAAGTAAAC	TACTAAC	TTTTTTAG	705	ncRNA
CMK219C[a]	598,707–599,146	GTAAGTTGAA	GACTAAC	GGCTGTAG	440	2
CMK245C	673,778–675,086	GTAAGTTTGT	AACTAAC	CTATTCAG	1309	0
CMK260C	713,501–713,871	GTAAGTTGTG	TACTAAC	TTATCTAG	371	1
CML049C[a]	105,245–105,389	GTAAGTTGGT	AACTAAC	GGAATCAG	145	0
CMM175C[a]	415,787–415,870	GTATGTTCTT	AACTAAC	TTCTTCAG	84	2
CMN285C	707,826–707,902	GTAAGTTTGG	CACTAAC	AAATCTAG	77	2
CMO094C	239,451–239,626	GTAAGTTTTG	AACTAAC	CAAATTAG	176	2
CMO159C	409,685–409,918	GTAAGCTGGT	AACTAAC	GAGCACAG	234	2
CMO267C	680,930–681,230	GTAAGTTATG	AACTAAC	GAGTTTAG	301	0
CMP330T[a]	828,107–828,298	GTAAGTACGC	GACTAAC	TTAAACAG	192	1
CMQ117C	275,562–275,818	GTAAGTGACT	AACTAAC	GGTGACAG	257	1
CMQ270C	699,519–699,783	GTAAGTTTAC	AACTAAC	CCTTGTAG	265	0
CMQ382C	992,043–992,350	GTAAGTTTCT	AACTAAC	CATTGTAG	308	2
CMR289C	697,226–697,295	GTAAGTTTTC	TACTAAC	CTTTATAG	70	1
CMR350C-1	848,946–849,072	GTAGGTTTGG	AACTAAC	GTTTCCAG	127	2
CMR350C-2	849,249–849,352	GTAGGTTTTG	GACTAAC	TGTCACAG	104	2
CMS262C	662,724–662,960	GTAAGTTTCA	AACTAAC	GACTTTAG	237	1
CMS270C	681,713–682,078	GTAAGTTTGA	AACTAAC	ACGAAAAG	366	0
CMS311C	802,563–802,678	GTAAGTTGAC	GACTAAC	AATCGCAG	116	2
CMS315C	807,778–808,022	GTAAGTGAAG	AACTAAC	ATACCCAG	245	0
CMS342C	876,128–876,372	GTAAGTTTCT	AACTAAC	GTGAATAG	245	2
CMT222C	560,390–560,782	GTAAGTTTAT	CACTAAC	ATGTACAG	393	0
CMT275C	686,878–687,045	GTAAGTTTCG	AACTAAC	GCCTGTAG	168	2
CMT476C	1,198,768–1,198,865	GTAAGTATAC	TACTAAC	ACGTCCAG	98	1
CMT570C	1,445,351–1,445,627	GTAAGTTTCT	AACTAAC	CACGCCAG	277	2

[a]Novel introns identified in this study.

more uniform (Figure 1), yet introns are still less likely to be found in the 3′ half of the coding region. Interestingly, these *G. sulphuraria* introns are most frequently found in the interval between 0.1 and 0.2 (Figure 1), indicating a pronounced 5′ bias for these introns. Overall, the general 5′ bias of introns across the red algae we investigated provides support for the reverse transcription mechanism of intron loss. More sequencing of red algal genomes, especially red algae outside of the Cyanidiales, will provide further opportunities for comparison of homologous introns and investigations into the mechanisms and evolution of intron loss across the red algal tree.

One key evolutionary mystery of intron-poor organisms is that complete intron loss appears to be very rare in eukaryotes. However, the presence of even a single spliceosomal intron would require the maintenance of a spliceosome in its genome. Introns that still exist in intron-poor organisms must then either be performing

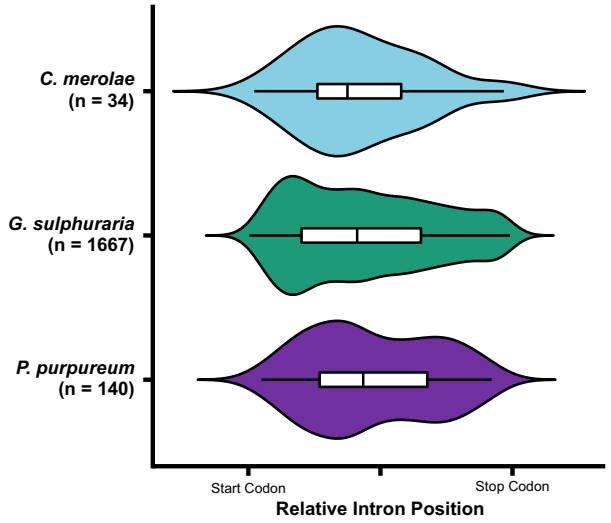

FIGURE 1 Distribution of intron locations along the length of single-intron genes in *C. merolae*, *G. sulphuraria,* and *P. purpureum*. The frequency of an intron being present at any particular interval of a gene's coding region is represented by violin and box plots for each species. *N* represents the number of single-introns genes analyzed

some sort of function in gene regulation, or be in the process of being lost, perhaps from positions in the genome where deletion is improbable. There is a propensity of remaining introns in intron-poor organisms in genes encoding ribosomal components, where they might provide a functional role in gene regulation (Parenteau et al., 2019; Spingola et al., 1999). These introns also have an extreme 5′ bias (Douglas et al., 2001; Lee et al., 2010; Moore et al., 2012; Spingola et al., 1999), although the reason for this relationship is not clear. However, only two *C. merolae* introns interrupt genes encoding ribosomal components, and they are not among the most 5′ biased; no other apparent pattern exists in the functions of its other intron-containing genes. Similarly, only 59 of the 1667 single-intron genes in *G. sulphuraria* and just one single-intron gene in *P. purpureum* encode ribosomal components, and their introns do not have an extreme 5′ bias.

An intron could provide regulatory functions within its host gene depending on its position with respect to the reading frame, whether its length is a multiple of three nucleotides, and whether it contains in-frame stop codons. Investigating these aspects of *C. merolae* introns could provide additional insight into their persistence in its genome. Intron phase describes its position relative to the host gene reading frame—a phase-0 intron lies between each triplet, while phase-1 and -2 introns interrupt codons. Phase-0 introns are most common across eukaryotes, while phase-2 introns are rarest (Fedorov et al., 1992; Long et al., 1995, 1998). However, there are more phase-2 (39.5%) introns than phase-0 (26.3%) introns in *C. merolae* (Table 1), although there are limitations in drawing conclusions from such a small number of introns. Only 11 introns in protein-coding genes are a multiple of three nucleotides (3n introns) in length (Table 1), which could allow read-through if the intron does not contain a stop codon. Six introns lack premature stop codons; interestingly, three of these are 3n, allowing potential read-through. These three introns have diverged 5′ splice site motifs and are all shorter than the average intron size in *C. merolae*. Two of these show increased read coverage across the intron, supporting read-through of these introns and possibly resulting in the favoring of shorter intron length and relaxed selection at the 5′ splice site motif. This suggests that these introns could be in the process of being lost from the genome. Regardless, the increased number of phase-1 and -2, or frameshift-inducing, introns containing premature stop codons may be important for *C. merolae* gene regulation. This may be especially relevant for *C. merolae*, as its transcripts are often unspliced, as discussed later. Previous analyses of introns in the yeast *Yarrowia lipolytica* and ciliate *Paramecium tetraurelia* showed an over-representation of introns that cause frame-shifts, and an under-representation of 3n introns (Jaillon et al., 2008; Mekouar et al., 2010). The under-represented 3n introns in *P. tetraurelia* are significantly enriched for stop codons, suggesting that introns in *P. tetraurelia* are under selective pressure to introduce premature stop codons into intron-retaining transcripts (Jaillon et al., 2008). The similar pattern in *C. merolae* suggests that selective pressures may be favoring introns that result in premature stop codons, potentially as a means of targeting them for degradation.

Intron sequence features of *Cyanidioschyzon merolae* and *Galdieria sulphuraria*

The difference in intron richness between *C. merolae* (38 total introns) and *G. sulphuraria* (over 13,000 introns) is extremely stark (Matsuzaki et al., 2004; Nozaki et al., 2007; Schönknecht et al., 2013), highlighting very divergent paths in the genome evolution of these two related species after an ancient genome reduction event (Qiu et al., 2015). Because intron sequence motifs are important for the splicing mechanism, these may have also diverged between the two Cyanidiales.

Spliceosome assembly in most eukaryotes is initiated upon the recognition of the 5′ splice site by the U1 snRNP, which depends on a base-pairing interaction between the 5′ splice site and the 5′ open region of the U1 snRNA. The consensus 5′ splice site across eukaryotes is GUAAG (Irimia et al., 2009); however, variability from this consensus is observed to correlate with intron richness. In intron-rich organisms such as humans or *A. thaliana*, the 5′ splice site consensus is simply reduced to GU, while an intron-poor organism such as *S. cerevisiae* has a lengthier 5′ splice

site consensus (Irimia et al., 2007). Invariability of intron sequence motifs has previously been proposed to be important if that organism has a reduced spliceosome, where it would allow for increased interactions with the remaining components of the spliceosome. Likewise, the branch donor region is recognized by the U2 snRNP, facilitated by a base-pairing interaction between the branch donor site and the U2 snRNA. As with 5′ splice sites, branch donor regions are difficult to identify in intron-rich organisms like metazoans or land plants (Gao et al., 2008; Irimia & Roy, 2008; Tolstrup et al., 1997), while a pattern of increasing invariability in branch donor regions was seen in intron-sparse organisms (Berglund et al., 1997; Irimia & Roy, 2008; Whelan et al., 2019).

We thus analyzed intron sequences from *C. merolae* and *G. sulphuraria* for patterns in sequence motifs (Figures 2, 3). The 5′ splice site and branch donor region consensus of the mesophilic red alga *P. purpureum* was also included for comparison. As expected for an intron-rich organism, *G. sulphuraria* exhibits more variability in both its 5′ splice sites (Figure 2) and its branch donor regions (Figure 3). In contrast, branch donor region sequences are highly constrained in *C. merolae* (Figure 3). As a point of comparison, variability of branch donor regions in *P. purpureum* is intermediate to the extremophilic *C. merolae* and *G. sulphuraria* (Figure 3). Although the exact number of introns in *P. purpureum* is not known (Wong et al., 2021), its intron richness lies between *C. merolae* and *G. sulphuraria*, highlighting the correlation between intron richness and splice motif variability. Thus, selection on these noncoding sequences to conform to a consensus is likely heightened with a reduced spliceosome.

As previously mentioned by Matsuzaki et al. (2004), the 5′ splice sites are largely unvarying in the extremely intron-sparse *C. merolae* (Figure 2), although the consensus sequence itself is quite typical of eukaryotes. An extended stretch of uracil is often found after the fifth nucleotide of *C. merolae* introns—this was also seen in *P. purpureum*, where the uracils are able to extend a base-pairing interaction with its U1 snRNA (Wong et al., 2021). While the lack of deviation from a strong consensus 5′ splice site in *C. merolae* might also suggest an extended interaction with its U1 snRNP, this particle of the spliceosome is completely absent in *C. merolae* (Stark et al., 2015). However, the U6 snRNP also interacts with the 5′ splice site (Sawa & Shimura, 1992), displacing the U1 snRNP in the formation of the precatalytic B complex (Will & Lührmann, 2011). The tendency for additional uracils in the sixth position and beyond in *C. merolae* introns (Figure 2) increases potential base pairing interactions with its U6 snRNA (Stark et al., 2015), suggesting that

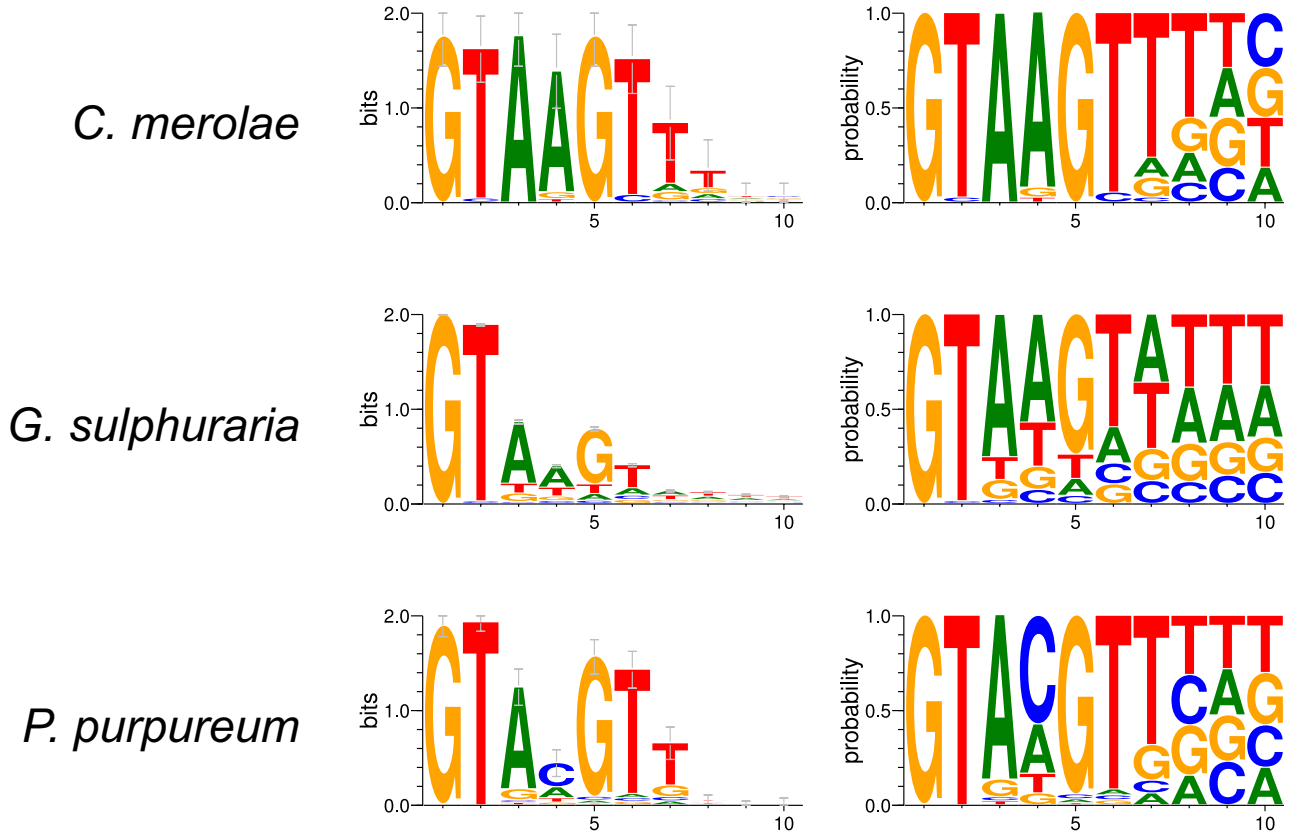

FIGURE 2 Variability in the 5′ splice sites of red algae. Sequence logos were generated for 5′ splice sites from all introns in *C. merolae* and *G. sulphuraria*, and confirmed introns of *P. purpureum*

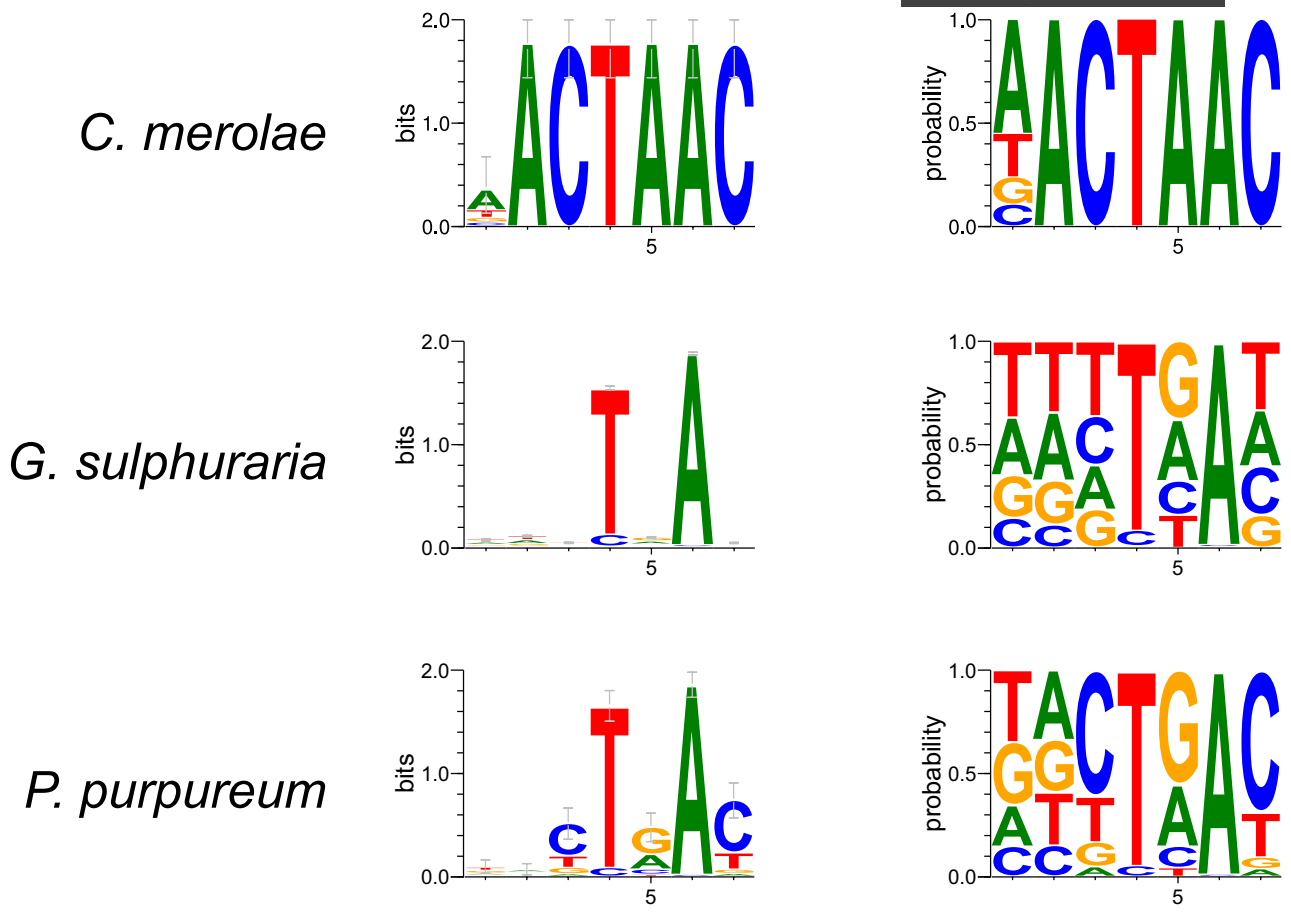

FIGURE 3 Consensus sequence of branch donor regions in the introns of red algae. Sequence logos were generated from predicted branch donor regions from *C. merolae*, *G. sulphuraria*, and *P. purpureum*

the U6 snRNP may have replaced the U1 snRNP in 5′ splice site recognition in *C. merolae*. Interestingly, both Cyanidiales species generally prefer adenosine in the fourth position of the 5′ splice site, which may be related to the absence of METTL16, a methyltransferase responsible for modifying the 5′ splice site interaction region of the U6 snRNA. A recent study in fission yeast *Schizosaccharomyces pombe* showed that by knocking out METTL16, the splicing efficiency of introns that do not contain adenosine in the fourth position is highly reduced (Ishigami et al., 2021). Its loss in intron-poor species was suggested to contribute to constraints in their 5′ splice site sequences, especially at the fourth nucleotide. While METTL16 is absent in both *C. merolae* and *G. sulphuraria*, adenosine is only slightly preferred in the fourth position of *G. sulphuraria* introns. Furthermore, METTL16 is also absent in *P. purpureum*, yet cytosine is slightly preferred in the same position in its introns, and the fourth nucleotide of its 5′ splice sites is actually the most variable. Regardless, the proclivity for specific nucleotides at 5′ splice sites could perhaps be helpful in identifying introns in genomes of non-Cyanidiales red algae.

Striking differences in the levels of intron retention in *Cyanidioschyzon merolae* and *Galdieria sulphuraria*

Analysis of key sequence motifs within introns has revealed largely invariant splice sites in the intron-poor *C. merolae*, while the introns of *G. sulphuraria* generally do not follow a strict consensus with their splicing motifs. In light of the vastly different complexities of the spliceosomes of *C. merolae* (Matsuzaki et al., 2004; Reimer et al., 2017; Stark et al., 2015) and *G. sulphuraria* (Qiu et al., 2018; Schönknecht et al., 2013) that must recognize these motifs, we investigated how strong a contrast might exist between pre-mRNA splicing levels between the two. We determined levels of pre-mRNA splicing by calculating the proportion of reads split across junctions (spliced) and reads mapping to intronic regions (unspliced). For *C. merolae*, the strand-specific and non-stranded datasets were combined following statistical analysis confirming that they are not significantly different (see Materials and Methods).

As expected, rather typical levels of pre-mRNA splicing were observed in the intron-rich *G. sulphuraria*

(Figure 4). While our data show that introns in *G. sulphuraria* are spliced at an extremely wide range of levels, the vast majority of introns were spliced at their predicted splice sites more than 80% of the time, with an average percent of spliced reads at 87.1%. This level of splicing is typical across most eukaryotes, as alternative splicing (including intron retention) is generally considered a rare event. On the other hand, we found levels of pre-mRNA splicing to be very low in the extremely intron-sparse *C. merolae* (Figure 5). No intron was spliced at more than 60%, and the average splicing levels of the normalized, combined datasets were 21.5% during the day, and 21.6% during the night. This similarity in splicing levels between day and night shows that very limited differential splicing takes place in *C. merolae*. Although certain junctions do exhibit some diurnal differences in splicing (Figure 5), these are not statistically significant.

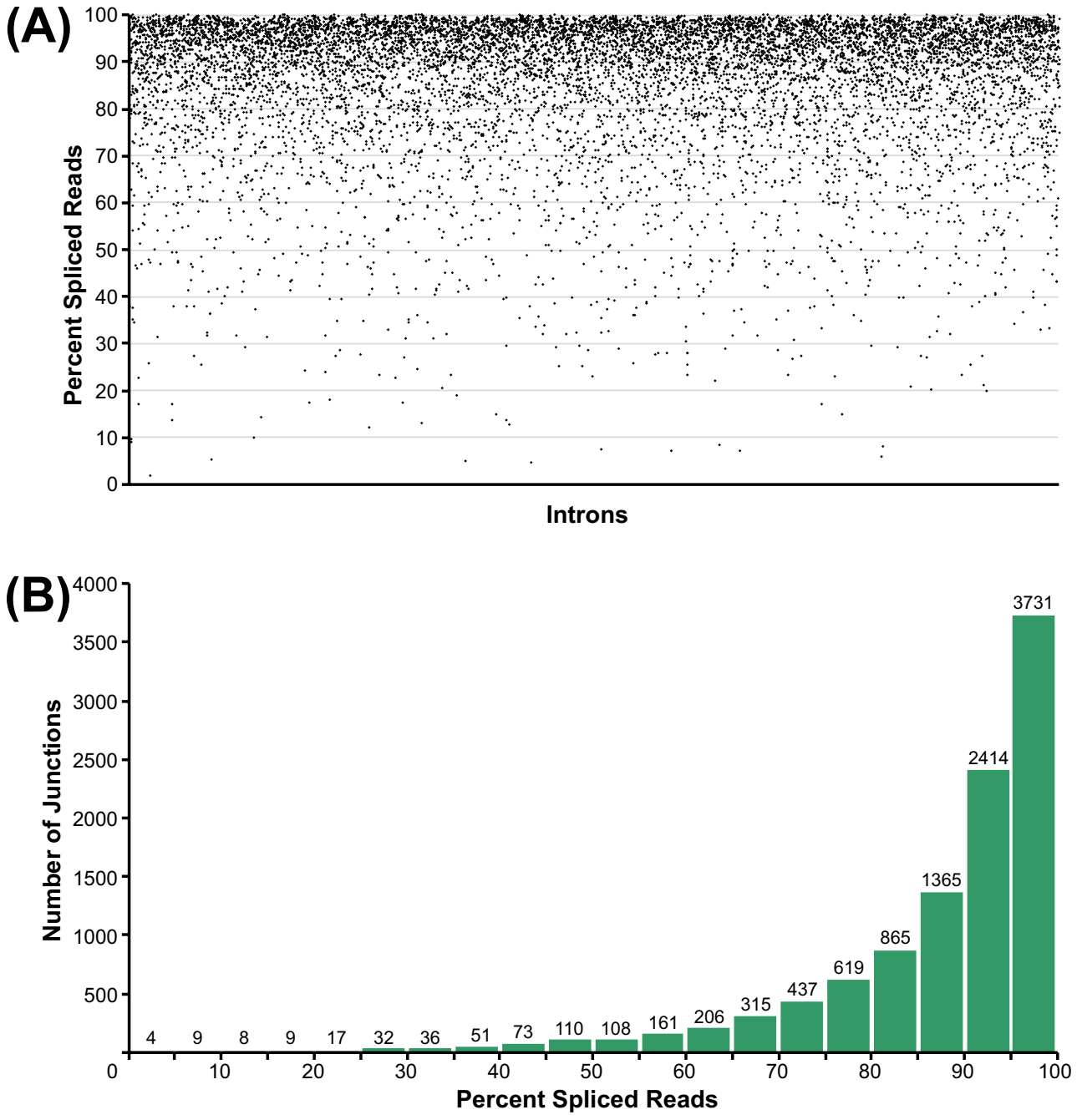

FIGURE 4 Typical eukaryotic levels of splicing in *G. sulphuraria*. (A) each point represents the percent of spliced reads in the vicinity of a *G. sulphuraria* intron. While introns are spliced at a large range of splicing levels, the average percent of spliced reads is 87.1%, and points are clustered at the top of the graph. (B) the splicing levels of *G. sulphuraria* introns are summarized as a histogram. Most introns are spliced at higher than 95%

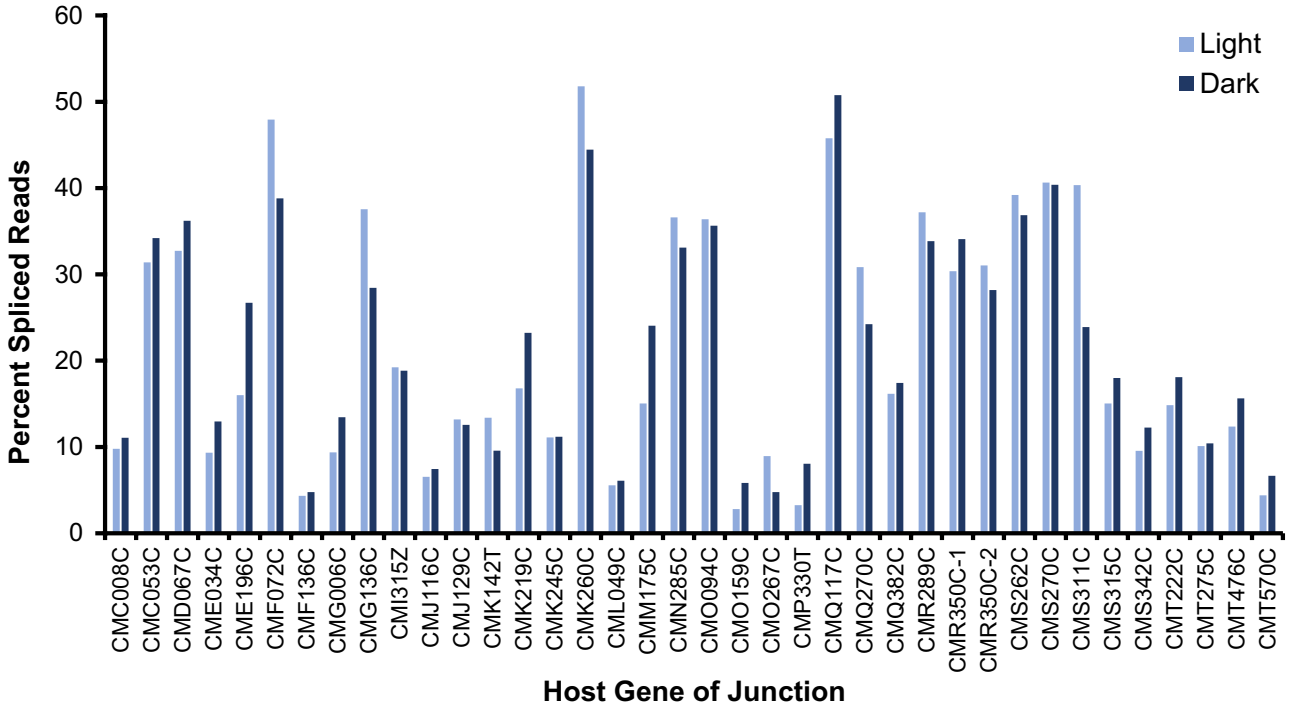

FIGURE 5 High prevalence of unspliced transcripts in *C. merolae*. Splicing levels from light and dark transcriptomes are present as separate bars for each of the 38 introns. No intron was spliced higher than 60%

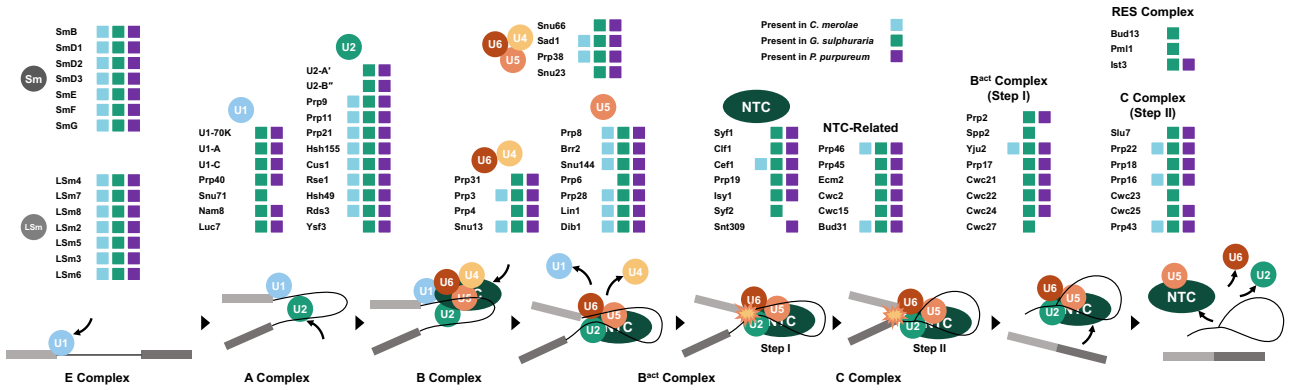

FIGURE 6 Comparison of spliceosome composition across *C. merolae*, *G. sulphuraria,* and *P. purpureum*. Spliceosomal proteins from *C. merolae*, *G. sulphuraria*, and *P. purpureum* are presented by their associated snRNP, accessory complex or step of the reaction, as per Fabrizio et al. (2009). Shaded boxes indicate the presence of that protein in each of the species in comparison

Low levels of splicing were previously reported in the red algal derived nucleomorph of the cryptophyte *Guillardia theta* (Wong et al., 2018) and the mesophilic red alga *Porphyridium purpureum* (Wong et al., 2021), both of which are also intron-sparse and relatively reduced (Douglas et al., 2001; Lee et al., 2019). However, introns in these two red algal genomes are spliced on average at greater than 50% (Wong et al., 2018, 2021), a level intermediate to *C. merolae* and *G. sulphuraria*. If *P. purpureum* and the *G. theta* nucleomorph are representative of red algae, the extremophiles *C. merolae* and *G. sulphuraria*—with such deep contrasts in their respective spliceosome compositions and intron densities—could represent two extremes of pre-mRNA splicing across red algae.

While splicing in *G. sulphuraria* is quite similar to that of most other eukaryotes, the prevalence of unspliced transcripts in *C. merolae* is highly unusual—its average percentage of spliced reads is far lower than any red alga observed to date. Such low levels of splicing have only been observed in the unrelated microsporidian parasite *E. cuniculi*, where it was shown to have less than 50% splicing efficiency in 30 of its 37 introns (Grisdale et al., 2013). These similar levels of splicing observed in *C. merolae* and *E. cuniculi* are thus strikingly low, especially when compared with other unicellular eukaryotes. An analysis of splicing in the somewhat intron-rich malaria parasite *Plasmodium falciparum* found that only a small proportion of intron-containing genes (~5%) had

splicing levels below 50% (Sorber et al., 2011). Similarly, a comparative analysis of splicing in two fungal species, *Saccharomyces cerevisiae* and *Candida albicans,* found that only 10–15% of introns analyzed have low (<50%) splicing levels (Grisdale et al., 2013). Regardless, low splicing efficiencies have only been shown to occur in some reduced systems, where intron densities and spliceosome complexity are low. The reduction and loss of spliceosomal components likely result in fewer interactions taking place during splicing. This has the potential to result in decreased splicing fidelity, which could affect splicing efficiency, and which is taken to the extreme in *C. merolae*. Comparing spliceosome composition across red algae could highlight losses that explain the prevalence of unspliced transcripts in *C. merolae*.

Spliceosomal composition of red algae and the snRNAs of *Galdieria sulphuraria*

We investigated the relationship between spliceosome complexity, intron richness, and prevalence of unspliced transcripts by comparing predicted spliceosomes of *C. merolae* and *G. sulphuraria*. We also compared these spliceosomes to that of the mesophile *P. purpureum*, and the results are summarized in Figure 6. The spliceosomes of *C. merolae* and *G. sulphuraria* have each been characterized (Qiu et al., 2018; Stark et al., 2015). We also explored relationships between snRNA divergence, splicing motif sequence variability, and splicing levels. We predicted structures (Figure 7) for the snRNAs previously identified in *G. sulphuraria* (Qiu et al., 2018) to compare with those of *C. merolae* (Stark et al., 2015). This further highlights the divergence of snRNAs in *C. merolae*.

As expected for the intron-rich *G. sulphuraria*, many spliceosomal proteins can be found in its genome, and these generally have high similarity to homologs in other eukaryotes (Qiu et al., 2018), Furthermore, all five snRNAs of *G. sulphuraria* form structures similar to those in other eukaryotes (Figure 7), indicating a complex and canonical spliceosome in *G. sulphuraria* despite genome reduction. On the other hand, *C. merolae* was shown to have an exceptionally reduced spliceosome (Stark et al., 2015). However, it was noted that spliceosomal proteins with homologs in *C. merolae* were not particularly divergent from homologs of other eukaryotes such as humans and *S. cerevisiae*; this trend extends to the canonical U4 and U6 snRNAs found in *C. merolae* (Stark et al., 2015). Spliceosomal components that are present in *C. merolae* generally belong to the core snRNPs (especially the U4/U6·U5 tri-snRNP), highlighting their functional importance. Unusually, the *C. merolae* U5 snRNA is greatly expanded (Stark et al., 2015), and forms a unique structure not seen in *G. sulphuraria* (Figure 7C) or any other eukaryote. The function of these additional loops in *C. merolae* remains unclear given its rather canonical complement of U5 snRNP proteins. Most missing proteins in *C. merolae* are instead part of accessory complexes or are other transient proteins in the spliceosome (Figure 6). In eukaryotes that are richer in introns, peripheral or non-essential proteins of the spliceosome appear to be associated with efficient splicing of particular populations of introns divergent in sequence or position of splicing motifs, as highlighted in a comparison of splicing in yeasts (Sales-Lee et al., 2021). As *C. merolae* possesses extremely strict splicing motifs (Figures 2 and 3), its missing spliceosomal proteins are in line with this trend.

Reductions, divergences, and losses of specific parts of the spliceosome have been associated with poor splicing efficiency and high levels of unspliced transcripts in *P. purpureum* and the microsporidian parasite *E. cuniculi* (Grisdale et al., 2013; Wong et al., 2021). In *C. merolae*, the most striking absence is the entire U1 snRNP (Stark et al., 2015), in contrast to the canonical U1 snRNP found in *G. sulphuraria*, whose U1 snRNA is dubiously "unique" for being so canonical relative to known red algal U1 snRNA structures. No such snRNA has been found in *C. merolae* (Stark et al., 2015), and the *P. purpureum* homolog has a greatly expanded SL II with unknown functional consequence (Wong et al., 2021). Divergence and loss of the U1 snRNA and its associated snRNP impact 5′ splice site recognition and assembly of the spliceosome, and correlate with increased intron retention in *C. merolae* (this study) and *P. purpureum* (Wong et al., 2021). While the invariance of *C. merolae* 5′ splice sites (Figure 2) and extended capability for base pairing with its U6 snRNA could provide an alternative mechanism of 5′ splice site recognition (Stark et al., 2015), this alternative mechanism of 5′ splice site recognition and initiation of spliceosome assembly, if it is taking place, is apparently not efficient.

The Prp19 complex ("Nineteen Complex," or NTC), the RES complex, and many Step I- and Step II-associated proteins are either highly reduced or lost entirely in *C. merolae* (Figure 6). Only a few NTC proteins remain in *C. merolae*, and Prp19, for which the complex is named, is absent (Figure 6). In contrast, *G. sulphuraria* and *P. purpureum* both have largely conserved NTCs, suggesting that the last common red algal ancestor possessed a typical eukaryotic NTC. Homologs of RES complex, Step I and Step II proteins are also found in *G. sulphuraria*, while an intermediate level of protein loss is seen in the more distantly related *P. purpureum* (Figure 6). Shared absences of Step I and Step II proteins and the RES complex between *C. merolae* and *P. purpureum* suggest that, like the U1 snRNP, accessory proteins are readily lost in conjunction with intron loss in red algae. The NTC is crucial for stabilizing the interaction between the U5 and U6 snRNPs after promoting the dissociation of the U4/U6 snRNPs, and is generally important for the transition to a catalytically active spliceosome (Chan et al., 2003; Chan & Cheng, 2005). The RES complex is important for efficient transition into the B^{act} complex (Bao et al., 2017); previously, we

Eukaryotic Microbiology ISOP Published by the International Society of Protistologists

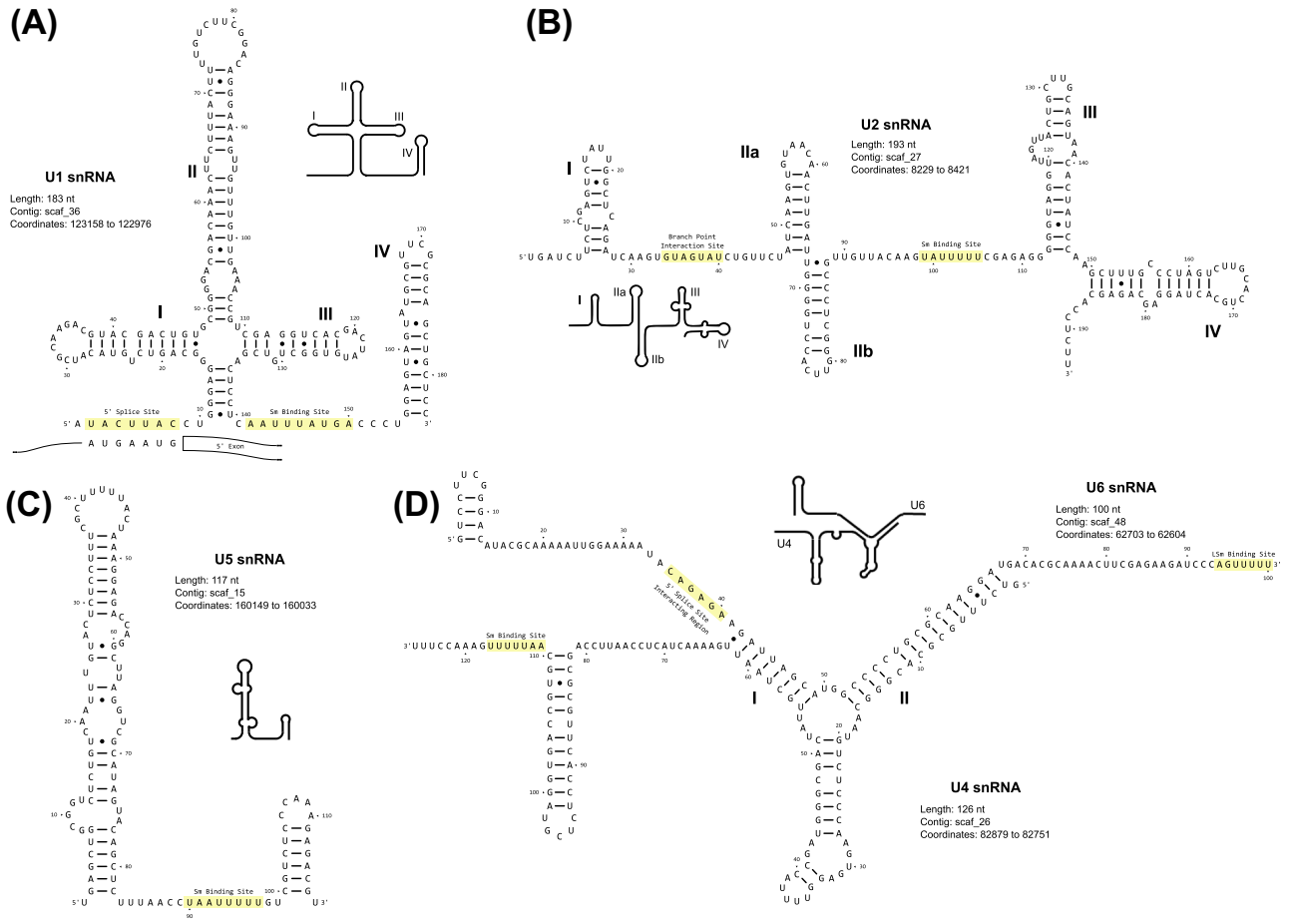

FIGURE 7 Predicted structures of *G. sulphuraria* snRNAs. Lengths and genomic coordinates of predicted *G. sulphuraria* snRNAs are provided for reference, along with schematic line drawings of consensus structures of each of the (A) U1 snRNA, (B) U2 snRNA, (C) U5 snRNA, (D) U4/U6 snRNAs

suggested that reduction of the RES complex in *P. purpureum* might be correlated with its increased levels of intron retention (Wong et al., 2021). Reduction of the NTC and loss of the RES complex in *C. merolae* likely result in unstable B and B^{act} complexes, contributing to its extremely inefficient splicing.

Peculiarities of *C. merolae* intronic splicing sequence motifs also appear to be associated with loss of spliceosomal proteins. While the U2 snRNP of *C. merolae* is the most conserved and complex snRNP in its highly reduced spliceosome (Stark et al., 2015), many U2-related proteins are found in other eukaryotes, such as the U2AF (U2 auxiliary factors), are absent (Figure 6). These encourage recruitment of the U2 snRNP to the branch donor region, and are themselves recruited by the U1 snRNP interacting with the 5′ splice site (Ruskin et al., 1988; Wu & Fu, 2015). Without U2AF or the U1 snRNP present in *C. merolae*, dependence on sequence complementarity between the branch donor region and the U2 snRNA presumably increases greatly, resulting in every intron having the exact same branch donor region (Figure 3). Furthermore, the U2 snRNA of *C. merolae* (Stark et al., 2015) is shorter than that of *G. sulphuraria* (Figure 7B) and most other eukaryotes, and the shortened portions correspond to stem-loops III and IV in other eukaryotes. In humans and *S. cerevisiae*, SL IV is tightly bound by U2-A′ and U2-B″ (Caspary & Séraphin, 1998; Price et al., 1998)—both of which are missing from the spliceosome of *C. merolae*. While this part of the U2 snRNP is not directly involved in branch donor region recognition, the absence of Lea1 (U2-A′) and Msl1 (U2-B″) in *S. cerevisiae* results in reduced levels of splicing (Caspary & Séraphin, 1998). The loss of these two proteins in *C. merolae*, along with their associated region of the U2 snRNA, very likely contributes to its extremely low levels of splicing and the convergence of all its introns' branch donor regions to consensus.

The differing spliceosomes of *C. merolae* and *G. sulphuraria* highlight patterns of coevolution involving intron richness, intron sequence motifs and the loss of specific components of the spliceosome. These patterns could also help explain the profound difference in levels of pre-mRNA splicing between these two related red algae. Finally, contrasting intron sequence motifs, splicing levels, and spliceosomal complexity between these two extremophiles emphasizes their vastly different evolutionary paths under the pressures of genome reduction.

ACKNOWLEDGMENTS

This work was supported by a grant to the Centre for Microbial Diversity and Evolution from the Tula Foundation, and the Natural Sciences and Engineering Research Council of Canada Discovery Grants [262988 to N.M.F., 298521 to S.D.R.].

DATA DEPOSITION

Raw reads generated from all sequencing runs have been deposited to NCBI's Sequence Read Archive (SRA) under BioProject accession PRJNA791314.

ORCID

Donald K. Wong https://orcid.org/0000-0003-3123-2309
Stephen D. Rader https://orcid.org/0000-0001-7242-1785

REFERENCES

Abrahamsen, M.S., Templeton, T.J., Enomoto, S., Abrahante, J.E., Zhu, G., Lancto, C.A. et al. (2004) Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science*, 304, 441–445.

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 Bact complex. *Genes & Development*, 31, 2416–2429.

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, 781–787.

Blanc-Mathieu, R., Verhelst, B., Derelle, E., Rombauts, S., Bouget, F.-Y., Carré, I. et al. (2014) An improved genome of the model marine alga *Ostreococcus tauri* unfolds by assessing Illumina de novo assemblies. *BMC Genomics*, 15, 1103.

Brawley, S.H., Blouin, N.A., Ficko-Blean, E., Wheeler, G.L., Lohr, M., Goodson, H.V. et al. (2017) Insights into the red algae and eukaryotic evolution from the genome of *Porphyra umbilicalis* (Bangiophyceae, Rhodophyta). *Proceedings of the National Academy of Sciences*, 114, E6361–E6370.

Caspary, F. & Séraphin, B. (1998) The yeast U2A'/U2B complex is required for pre-spliceosome formation. *The EMBO Journal*, 17, 6348–6358.

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 *. *The Journal of Biological Chemistry*, 280, 31190–31199.

Chan, S.-P., Kao, D.-I., Tsai, W.-Y. & Cheng, S.-C. (2003) The Prp19p-associated complex in spliceosome activation. *Science*, 302, 279–282.

Collen, J., Porcel, B., Carre, W., Ball, S.G., Chaparro, C., Tonon, T. et al. (2013) Genome structure and metabolic features in the red seaweed *Chondrus crispus* shed light on evolution of the Archaeplastida. *Proceedings of the National Academy of Sciences*, 110, 5247–5252.

Crooks, G.E., Hon, G., Chandonia, J.-M. & Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Research*, 14, 1188–1190.

Csuros, M., Rogozin, I.B. & Koonin, E.V. (2011) A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Computational Biology*, 7, e1002150.

Cuomo, C.A., Desjardins, C.A., Bakowski, M.A., Goldberg, J., Ma, A.T., Becnel, J.J. et al. (2012) Microsporidian genome analysis reveals evolutionary strategies for obligate intracellular growth. *Genome Research*, 22, 2478–2488.

De Luca, P., Taddei, R. & Varano, L. (1978) « Cyanidioschyzon merolae »: a new alga of thermal acidic environments. *Webbia*, 33, 37–44.

Derelle, E., Ferraz, C., Rombauts, S., Rouzé, P., Worden, A.Z., Robbens, S. et al. (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 11647–11652.

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

Douglas, S., Zauner, S., Fraunholz, M., Beaton, M., Penny, S., Deng, L.-T. et al. (2001) The highly reduced genome of an enslaved algal nucleus. *Nature*, 410, 1091–1096.

Engel, S.R., Dietrich, F.S., Fisk, D.G., Binkley, G., Balakrishnan, R., Costanzo, M.C. et al. (2013) The reference genome sequence of *Saccharomyces cerevisiae*: then and now. *G3 GenesGenomesGenetics*, 4, 389–398.

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

Farlow, A., Meduri, E. & Schlötterer, C. (2011) DNA double-strand break repair and the evolution of intron density. *Trends in Genetics*, 27, 1–6.

Fedorov, A., Merican, A.F. & Gilbert, W. (2002) Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proceedings of the National Academy of Sciences of the United States of America*, 99, 16128–16133.

Fedorov, A., Suboch, G., Bujakov, M. & Fedorova, L. (1992) Analysis of nonuniformity in intron phase distribution. *Nucleic Acids Research*, 20, 2553–2557.

Fink, G.R. (1987) Pseudogenes in yeast? *Cell*, 49, 5–6.

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

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W. et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419, 498–511.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H. et al. (1996) Life with 6000 genes. *Science*, 274, 546–567.

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 pre-mRNA splicing in a reduced eukaryote. *BMC Genomics*, 14, 207.

Ho, C.-L., Lee, W.-K. & Lim, E.-L. (2018) Unraveling the nuclear and chloroplast genomes of an agar producing red macroalga, *Gracilaria changii* (Rhodophyta, Gracilariales). *Genomics*, 110, 124–133.

Hudson, A.J., McWatters, D.C., Bowser, B.A., Moore, A.N., Larue, G.E., Roy, S.W. et al. (2019) Patterns of conservation of spliceosomal intron structures and spliceosome divergence in representatives of the diplomonad and parabasalid lineages. *BMC Evolutionary Biology*, 19, 162.

Irimia, M., Penny, D. & Roy, S.W. (2007) Coevolution of genomic intron number and splice sites. *Trends in Genetics*, 23, 321–325.

Irimia, M. & Roy, S.W. (2008) Evolutionary convergence on highly-conserved 3′ intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. *PLoS Genetics*, 4, e1000148.

Irimia, M., Roy, S.W., Neafsey, D.E., Abril, J.F., Garcia-Fernandez, J. & Koonin, E.V. (2009) Complex selection on 5′ splice sites in intron-rich organisms. *Genome Research*, 19, 2021–2027.

Ishigami, Y., Ohira, T., Isokawa, Y., Suzuki, Y. & Suzuki, T. (2021) A single m6A modification in U6 snRNA diversifies exon sequence at the 5′ splice site. *Nature Communications*, 12, 3244.

Jaillon, O., Bouhouche, K., Gout, J.-F., Aury, J.-M., Noel, B., Saudemont, B. et al. (2008) Translational control of intron splicing in eukaryotes. *Nature*, 451, 359–362.

Jurica, M.S. & Moore, M.J. (2003) Pre-mRNA splicing: awash in a sea of proteins. *Molecular Cell*, 12, 5–14.

Katinka, M.D., Duprat, S., Cornillot, E., Méténier, G., Thomarat, F., Prensier, G. et al. (2001) Genome sequence and gene compaction of the eukaryote parasite encephalitozoon cuniculi. *Nature*, 414, 450–453.

Koonin, E.V. (2006) The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biology Direct*, 1, 22.

Lane, C.E., van den Heuvel, K., Kozera, C., Curtis, B.A., Parsons, B.J., Bowman, S. et al. (2007) Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 19908–19913.

Lee, J., Kim, D., Bhattacharya, D. & Yoon, H.S. (2019) Expansion of phycobilisome linker gene families in mesophilic red algae. *Nature Communications*, 10, 4823.

Lee, R.C.H., Gill, E.E., Roy, S.W. & Fast, N.M. (2010) Constrained intron structures in a microsporidian. *Molecular Biology and Evolution*, 27, 1979–1982.

Long, M., Rosenberg, C. & Gilbert, W. (1995) Intron phase correlations and the evolution of the intron/exon structure of genes. *Proceedings of the National Academy of Sciences of the United States of America*, 92, 12495–12499.

Long, M., de Souza, S.J., Rosenberg, C. & Gilbert, W. (1998) Relationship between "proto-splice sites" and intron phases: evidence from dicodon analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 95, 219–223.

Ma, M.-Y., Che, X.-R., Porceddu, A. & Niu, D.-K. (2015) Evaluation of the mechanisms of intron loss and gain in the social amoebae Dictyostelium. *BMC Evolutionary Biology*, 15, 286.

Matsuzaki, M., Misumi, O., Shin-i, T., Maruyama, S., Takahara, M., Miyagishima, S. et al. (2004) Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature*, 428, 653–657.

Mekouar, M., Blanc-Lenfle, I., Ozanne, C., Da Silva, C., Cruaud, C., Wincker, P. et al. (2010) Detection and analysis of alternative splicing in Yarrowia lipolytica reveal structural constraints facilitating nonsense-mediated decay of intron-retaining transcripts. *Genome Biology*, 11, R65.

Minoda, A., Sakagami, R., Yagisawa, F., Kuroiwa, T. & Tanaka, K. (2004) Improvement of culture conditions and evidence for nuclear transformation by homologous recombination in a red alga, *Cyanidioschyzon merolae* 10D. *Plant & Cell Physiology*, 45, 667–671.

Moore, C.E., Curtis, B., Mills, T., Tanifuji, G. & Archibald, J.M. (2012) Nucleomorph genome sequence of the Cryptophyte alga *Chroomonas mesostigmatica* CCMP1168 reveals lineage-specific gene loss and genome complexity. *Genome Biology and Evolution*, 4, 1162–1175.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, 5, 621–628.

Nawrocki, E.P. & Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29, 2933–2935.

Nozaki, H., Takano, H., Misumi, O., Terasawa, K., Matsuzaki, M., Maruyama, S. et al. (2007) A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga *Cyanidioschyzon merolae*. *BMC Biology*, 5, 28.

Parenteau, J., Maignon, L., Berthoumieux, M., Catala, M., Gagnon, V. & Abou, E.S. (2019) Introns are mediators of cell response to starvation. *Nature*, 565, 612–617.

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*, 394, 645–650.

Qiu, H., Price, D.C., Yang, E.C., Yoon, H.S. & Bhattacharya, D. (2015) Evidence of ancient genome reduction in red algae (Rhodophyta).Valentin K. (ed.). *Journal of Phycology*, 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 Evolutionary Biology*, 18, 41.

Reimer, K.A., Stark, M.R., Aguilar, L.-C., Stark, S.R., Burke, R.D., Moore, J. et al. (2017) The sole LSm complex in *Cyanidioschyzon merolae* associates with pre-mRNA splicing and mRNA degradation factors. *RNA*, 23, 952–967.

Roy, S.W. & Gilbert, W. (2005) The pattern of intron loss. *Proceedings of the National Academy of Sciences*, 102, 713–718.

Ruskin, B., Zamore, P.D. & Green, M.R. (1988) A factor, U2AF, is required for U2 snRNP binding and splicing complex assembly. *Cell*, 52, 207–219.

Sales-Lee, J., Perry, D.S., Bowser, B.A., Diedrich, J.K., Rao, B., Beusch, I. et al. (2021) Coupling of spliceosome complexity to intron diversity. *Current Biology*, 31, 4898–4910.e4.

Sawa, H. & Shimura, Y. (1992) Association of U6 snRNA with the 5′-splice site region of pre-mRNA in the spliceosome. *Genes & Development*, 6, 244–254.

Schönknecht, G., Chen, W.-H., Ternes, C.M., Barbier, G.G., Shrestha, R.P., Stanke, M. et al. (2013) Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science*, 339, 1207–1210.

Seckbach, J., Chapman, D.J. & Oren, A. (Eds.). (2010) *Red algae in the genomic age*. Dordrecht; New York: Springer.

Sorber, K., Dimon, M.T. & DeRisi, J.L. (2011) RNA-seq analysis of splicing in *Plasmodium falciparum* uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. *Nucleic Acids Research*, 39, 3820–3835.

Spingola, M., Grate, L., Haussler, D. & Ares, M. (1999) Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA*, 5, 221–234.

Stark, M.R., Dunn, E.A., Dunn, W.S.C., Grisdale, C.J., Daniele, A.R., Halstead, M.R.G. et al. (2015) Dramatically reduced spliceosome in *Cyanidioschyzon merolae*. *Proceedings of the National Academy of Sciences*, 112, E1191–E1200.

Terui, S., Suzuki, K., Takahashi, H., Itoh, R. & Kuroiwa, T. (1995) Synchronization of chloroplast division in the Ultramicroalga *Cyanidioschyzon merolae* (rhodophyta) by treatment with light and aphidicolin. *Journal of Phycology*, 31, 958–961.

Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P. (2013) Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14, 178–192.

Tolstrup, N., Rouzé, P. & Brunak, S. (1997) A branch point consensus from Arabidopsis found by non-circular analysis allows for better prediction of acceptor sites. *Nucleic Acids Research*, 25, 3159–3163.

Wahl, M.C., Will, C.L. & Lührmann, R. (2009) The spliceosome: design principles of a dynamic RNP machine. *Cell*, 136, 701–718.

Wang, D., Yu, X., Xu, K., Bi, G., Cao, M., Zelzion, E. et al. (2020) *Pyropia yezoensis* genome reveals diverse mechanisms of carbon acquisition in the intertidal environment. *Nature Communications*, 11, 4028.

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 Biology and Evolution*, 11, 263–269.

Will, C.L. & Lührmann, R. (2011) Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology*, 3, a003707.

Wong, D.K., Grisdale, C.J. & Fast, N.M. (2018) Evolution and diversity of pre-mRNA splicing in highly reduced nucleomorph Genomes. Meyer M. (ed.). *Genome biol Evol.*, 10, 1573–1583.

Wong D. K., Stark M. S., Rader S. D. & Fast N. M. 2021. Characterization of pre-mRNA splicing and Spliceosomal machinery in *Porphyridium purpureum* and evolutionary implications for red algae. *The Journal of Eukaryotic Microbiology*, 68. Available from: https://onlinelibrary.wiley.com/doi/10.1111/jeu.12844, e12844

Wu, T. & Fu, X.-D. (2015) Genomic functions of U2AF in constitutive and regulated splicing. *RNA Biology*, 12, 479–485.

Yoon, H.S., Müller, K.M., Sheath, R.G., Ott, F.D. & Bhattacharya, D. (2006) Defining the major lineages of red algae (Rhodophyta). *Journal of Phycology*, 42, 482–492.

Zhang, Q., Fan, X., Wang, Y., Sun, M., Shao, J. & Guo, D. (2017) BPP: a sequence-based algorithm for branch point prediction. Hancock J. (ed.). *Bioinformatics*, 33, 3166–3172.

Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31, 3406–3415.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Wong, D.K., Grisdale, C.J., Slat, V.A., Rader, S.D. & Fast, N.M. (2022) The evolution of pre-mRNA splicing and its machinery revealed by reduced extremophilic red algae. *Journal of Eukaryotic Microbiology*, 00, e12927. Available from: https://doi.org/10.1111/jeu.12927