ORIGINAL RESEARCH
published: 28 January 2022
doi: 10.3389/fgene.2021.818697

Identification of Alternative Polyadenylation in *Cyanidioschyzon merolae* Through Long-Read Sequencing of mRNA

Leonard Schärfen[1]†, Dagmar Zigackova[1]†, Kirsten A. Reimer[1], Martha R. Stark[2], Viktor A. Slat[2], Nancy J. Francoeur[3], Melissa L. Wells[4], Lecong Zhou[5], Perry J. Blackshear[4], Karla M. Neugebauer[1] and Stephen D. Rader[2]**

[1]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, United States, [2]Department of Chemistry, University of Northern British Columbia, Prince George, BC, Canada, [3]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [4]The Signal Transduction Laboratory, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, Durham, NC, United States, [5]Integrative Bioinformatics Support Group, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, Durham, NC, United States

OPEN ACCESS

Edited by:
Athma A Pai,
University of Massachusetts Medical School, United States

Reviewed by:
Yongsheng Shi,
University of California, Irvine, United States
Congting Ye,
Xiamen University, China
Matthew Taliaferro,
University of Colorado Anschutz Medical Campus, United States

Correspondence:
Stephen D. Rader
Stephen.Rader@unbc.ca
Karla M. Neugebauer
karla.neugebauer@yale.edu

†These authors have contributed equally to this work

Specialty section:
This article was submitted to RNA,
a section of the journal Frontiers in Genetics

Received: *19 November 2021*
Accepted: *22 December 2021*
Published: *28 January 2022*

Citation:
Schärfen L, Zigackova D, Reimer KA, Stark MR, Slat VA, Francoeur NJ, Wells ML, Zhou L, Blackshear PJ, Neugebauer KM and Rader SD (2022) Identification of Alternative Polyadenylation in Cyanidioschyzon merolae *Through Long-Read Sequencing of mRNA.*
Front. Genet. 12:818697.
doi: 10.3389/fgene.2021.818697

Alternative polyadenylation (APA) is widespread among metazoans and has been shown to have important impacts on mRNA stability and protein expression. Beyond a handful of well-studied organisms, however, its existence and consequences have not been well investigated. We therefore turned to the deep-branching red alga, *Cyanidioschyzon merolae*, to study the biology of polyadenylation in an organism highly diverged from humans and yeast. *C. merolae* is an acidothermophilic alga that lives in volcanic hot springs. It has a highly reduced genome (16.5 Mbp) and has lost all but 27 of its introns and much of its splicing machinery, suggesting that it has been under substantial pressure to simplify its RNA processing pathways. We used long-read sequencing to assess the key features of *C. merolae* mRNAs, including splicing status and polyadenylation cleavage site (PAS) usage. Splicing appears to be less efficient in *C. merolae* compared with yeast, flies, and mammalian cells. A high proportion of transcripts (63%) have at least two distinct PAS's, and 34% appear to utilize three or more sites. The apparent polyadenylation signal UAAA is used in more than 90% of cases, in cells grown in both rich media or limiting nitrogen. Our documentation of APA for the first time in this non-model organism highlights its conservation and likely biological importance of this regulatory step in gene expression.

Keywords: *Cyanidioschyzon merolae*, alternative polyadenylation, splicing, intron retention, cleavage and polyadenylation, polyadenylation (polyA) signal, polyadenylation site (PAS), nitrogen stress

INTRODUCTION

Splicing and polyadenylation are key events required for the maturation of precursor protein coding RNAs to functional mRNAs. Both processes have the potential to generate diverse mRNA isoforms from the same RNA precursor, altering their protein-coding potential and/or their stability (Mayr, 2019; Zhang et al., 2021). Together with transcriptional regulation, these major RNA processing mechanisms thereby underlie differences between the transcriptomes and proteomes expressed during cell differentiation and development in all eukaryotic species. Splicing greatly influences the gene output and is tightly regulated by other pre-mRNA processing steps, such as the transcription elongation and 5′ end capping (Carrocci & Neugebauer, 2019). Coupling between co-transcriptional splicing and polyadenylation cleavage has recently been demonstrated in budding and fission yeasts as well as mouse cells (Herzel 2018; Alpert et al.,2020; Reimer et al., 2021). In fission yeast, transcriptional read-through is antagonized by the nuclear exosome, which degrades transcripts that are not spliced co-transcriptionally (Zhou et al., 2015; Herzel et al., 2018). Thus, splicing can alter the mRNA levels and is coupled to polyadenylation, including cleavage at the polyA site (PAS) and the addition of a 3′ polyA tail. In most organisms, protein coding transcripts are only exported to the cytoplasm when splicing is completed and initial polyA tails are installed.

The critical roles of mRNA polyadenylation include protection from degradation, export of the mature mRNA to the cytoplasm, and facilitation of translation (Jalkanen et al., 2014). Most eukaryotic genes studied so far contain more than one PAS and therefore produce isoforms of mRNAs in a process called alternative polyadenylation (APA). Systematic mapping of PAS's estimates that up to 80% of mammalian, 75% of *Drosophila*, and more than 80% of yeast genes harbor alternative PAS and display APA (Derti et al., 2012; Hoque et al., 2013; Liu, Freitas, et al., 2017; Liu, Hoque, et al., 2017). APA sites, sometimes also called cryptic sites, can be divided into two categories based on their location relative to the stop codon (Rehfeld et al., 2013). The majority of APA sites are located in the 3′ untranslated region (3′UTR), and the selection of a particular PAS proximal or distal to the stop codon determines the length of the 3′UTR. The length of the 3′UTR and the presence or absence of regulatory sequence affects the stability of the transcripts and protein translation (Ren et al., 2020). However, APA sites can also be found in the introns or exons within the coding region and, in these cases, the PAS selection generates shortened protein isoforms (Ren et al., 2020).

Despite the pivotal role of splicing and APA in gene regulation, these processes have been studied in depth in only in a handful of species. Investigating other organisms could bring important insights, especially from species with fewer components involved in these mechanisms. Species adapted to live in vastly different conditions, such as extreme temperatures, are inherently interesting as their study contributes to our understanding of how basic mechanisms are tuned to the environment. Therefore, we explored the transcriptome of *Cyanidioschyzon merolae (C. merolae),* a unicellular red microalga. Extremophilic *C. merolae* grows in a sulfate-rich, acidic (pH 0.2–4) volcanic springs with temperatures as high as 57°C. The organellar and nuclear genomes of *C. merolae* were the first eukaryotic algal genomes to be completely sequenced (Ohta et al., 1998; Ohta et al., 2003; Matsuzaki et al., 2004; Nozaki et al., 2007). With a compact genome of 16.5 Mbp and 5,331 protein-coding genes (4,775 nuclear) and low genetic redundancy, it is considered one of the simplest eukaryotic organisms (Matsuzaki et al., 2004).

The published *C. merolae* genome reported only 27 introns in 26 genes (Matsuzaki et al., 2004). Interestingly, *C. merolae* contains a simplified spliceosome compared to other organisms. The spliceosomal small nuclear ribonucleoproteins (snRNPs) are comprised of five snRNAs (U1, U2, U4, U5 and U6) and about 100 associated spliceosomal proteins in *Saccharomyces cerevisiae* or over 200 associated proteins in humans (Fabrizio et al., 2009; Agafonov et al., 2011). In contrast, only four snRNAs and 45 core splicing proteins were identified in *C. merolae* (Stark et al., 2015; Reimer et al., 2017). The missing spliceosome component is the U1 snRNA and its associated proteins. U1 snRNA is involved in binding to the 5′ splice site prior to splicing and in telescripting, in which U1 snRNA prevents premature mRNA cleavage at cryptic PAS's (Berg et al., 2012; Oh et al., 2017; Singh & Singh, 2019). This implies the presence of alternative mechanisms mediating 5′ splice site recognition and protection from premature cleavage in *C. merolae*. It also raises the question of whether *C. merolae* displays altered patterns of PAS recognition and cleavage in the absence of U1. The unique simplicity of the *C. merolae* genome and spliceosome provide an advantage over highly complex systems for studies of mRNA processing. Moreover, the crystal structure of the first *C. merolae* core spliceosome factor, CmSnu13, revealed a near identity to its yeast and human orthologs, further emphasizing the suitability of *C. merolae* in studies of the function of the core spliceosome (Black et al., 2016).

Here, we have employed long-read sequencing (LRS) to enable characterization of individual *C. merolae* transcripts, from their 5′ to 3′ ends, allowing us to investigate splicing efficiency and PAS selection for the first time in this organism. *C. merolae* cells were cultivated either in rich media (RM) or in low nitrogen (LN) media to induce nitrogen starvation stress to study the regulation of splicing and APA under stress. Our results indicate a wide range of splicing efficiencies of individual intron-containing genes and overall increased splicing in the LN conditions. We discovered that a majority of *C. merolae* genes undergo APA under the control of highly conserved regulatory sequences.

METHODS

C. merolae Culturing

The 10D strain of *C. merolae* (NIES-3377) was cultured in modified Allen's autotrophic media with double the amount of trace elements (MA2 = rich media/RM) (Ohnuma et al., 2008) or in MA2 media in which the 40 mM $(NH_4)_2SO_4$ was replaced with 40 mM Na_2SO_4 (low nitrogen media/LN). Cultures were grown in graduated glass cylinders at 42°C in a CO_2 incubator, with 2% CO_2 bubbled directly into the cylinders. Cells were grown under continuous light at 90 μmol photons m^{-2} s^{-1}. Cells were grown in RM until they reached an OD_{750} of 0.8–1.0. Cultures were centrifuged at 2000xg for 10 min and then washed in either RM or LN before being resuspended in the corresponding media and grown for 1 h (from the time of the wash) before being harvested.

Total RNA Isolation

C. merolae cultures (10 ml) were harvested after 1 h of growth in the selected media. Cell pellets were resuspended in 500 μL cold phenol RNA lysis buffer (0.5 M NaCl, 0.2 M TrisHCl (pH 7.5), 0.01 M EDTA, 1% SDS) and were sonicated briefly to shear the genomic DNA. Total RNA was extracted two times with acid phenol/chloroform (pH 4.5, Ambion), followed by chloroform extraction. RNA was EtOH precipitated and resuspended in dH2O and then treated with Turbo DNase (Ambion) following the manufacturer's instructions.

Standard IsoSeq Library Preparation and Sequencing

IsoSeq SMRTbell libraries were prepared as recommended by the manufacturer (Pacific Biosciences). Briefly, 500 nanograms of total RNA from each sample was used as input for cDNA synthesis using the NEBNext Single Cell/Low Input cDNA Synthesis & Amplification Module (NEB, #E6421L), which employs a modified oligodT primer and template switching technology to reverse-transcribe full-length polyadenylated transcripts. Following double-stranded cDNA amplification and purification, the full-length cDNA was used as input into SMRTbell library preparation, using SMRTbell Express Template Preparation Kit v2.0. Briefly, a minimum of 100 ng of cDNA from each sample was treated with a DNA Damage Repair enzyme mix to repair nicked DNA, followed by an End Repair and A-tailing reaction to repair blunt ends and polyadenylate each template. Next, barcoded overhang SMRTbell adapters were ligated onto each template and purified using 0.6X AMPure PB beads to remove small fragments and excess reagents (Pacific Biosciences). The completed SMRTbell libraries were further treated with the SMRTbell Enzyme Clean Up Kit to remove unligated templates and equimolar pooled together. The final libraries were then annealed to sequencing primer v4 and bound to sequencing polymerase 3.0 before being sequenced on one SMRTcell 8M on the Sequel II system with a 30 h movie.

Custom IsoSeq Library Preparation and Sequencing to Capture Non-polyadenylated Transcripts

To profile non-polyadenylated, full-length RNA molecules, a custom adaptation of the standard IsoSeq protocol was performed. First, polyA polymerase (*E. coli* PolyA Polymerase, NEB, #M0276S) was used to catalyze polyA tailing of the 3′ ends of all RNA templates using 3 μg of total RNA as input and by following the manufacturer's instructions. Next, to avoid sequencing predominant rRNA molecules, the NEBNext Custom RNA Depletion Tool (NEB) was used to design hybridization probes tiled against known *C. merolae* rRNA sequences. The custom hybridization probes were then synthesized as a ssDNA oligo pool supplied at a concentration of 10 pmol per oligo (Integrated DNA Technologies) and were used in combination with the NEBNext RNA Depletion Core Reagent Set (NEB, #E7865S) to deplete rRNA molecules. Briefly, 1 μg of polyA-tailed RNA per sample was supplied as input to a hybridization reaction with the ssDNA depletion oligos and the targeted molecules were subsequently degraded by RNaseH digestion. Surplus DNA probes were then digested with DNase I treatment and the samples were subsequently washed using 1.8X Agencourt RNAClean XP beads (Beckman Coulter) to remove digested nucleotides and excess reagents. The purified polyA-treated, rRNA-depleted RNA samples were then subjected to SMRTbell library preparation and sequencing as described above.

IsoSeq Data Primary Processing

After data collection, the raw sequencing subreads were imported to the SMRTLink analysis suite, version 10.1 for processing. The latest version of PacBio's IsoSeq3 (v.3.4.0) pipeline was used for analysis. First, intramolecular error correcting was performed using the circular consensus sequencing (CCS) algorithm (v.6.0.0) to produce highly accurate (>Q20) CCS reads, each requiring a minimum of three polymerase passes. The CCS reads were then passed to the *lima* (v.2.0.0) tool first to demultiplex samples by barcoded adapters, and second, using the specialized "-isoseq" mode, to remove IsoSeq template-switching oligo sequences, which also orients the isoforms into the correct 5′ to 3′ direction. Reads were aligned to the *C. merolae* genome using minimap2 (Li, 2018) with the long-read/Pacbio-CCS preset and a maximum intron length of 1 kb.

Calculating Fraction Spliced of PacBio Long Reads

First, PacBio long reads that spanned an intron-containing gene were extracted from mapped data: Mapped reads in BAM format were converted to BED format with bedtools, then intron-spanning reads were identified by overlapping an annotation of *C. merolae* intron coordinates using bedtools intersect with parameters: *wo*, *F* = 0.3, *s* (Quinlan & Hall, 2010; Dale et al., 2011). Spliced reads in BED format were then classified as intron-spanning reads which had a "blockSize" greater than one, indicating the read mapped to two separate regions in the genome. For each intron, the total number of reads spanning the intron and the number of reads which were classified as spliced was tabulated. Fraction spliced was calculated as the number of spliced reads divided by the total number of reads for each intron. Coverage for each intron was defined as the total number of intron-spanning reads.

Analysis of Alternative Polyadenylation

Individual reads were assigned to a gene by intersecting a list of gene coordinates generated from the annotation in BED format using bedtools intersect with settings *wo*, *s*, *f* = 0.25. Next, the 3′-end coordinate for each aligned read was extracted by parsing BAM files, and assigned a gene based on the previously generated BED file. Read 3′-ends were mapped for each individual transcript by drawing histograms of the 3′-end positions. The find_peaks algorithm from scipy.signal was used with empirically determined height threshold of 4 and minimum peak distance threshold of 18 nt. Peaks were sorted according to their height and the number of peaks was recorded for every transcript. For meta-analysis, signals for individual transcripts were normalized by dividing by the total number of 3′-ends and then summed after aligning at the highest peak. Sequences of the 500 most expressed genes 40 nt upstream of the first highest, second highest and third highest peak were extracted and exported as fasta files for motif analysis with XTREME using the RNA library and default parameters except for the minimum motif length, which was decreased to 4. The length between the polyadenylation signal and the cleavage site was determined by searching for “UAAA” within 40 nt upstream of the highest peak and calculating the distance between the peak coordinate and the search result coordinate.

Correlating Intron Features With Fraction Spliced

For each previously identified intron in the *C. merolae* genome, intron length, GC content, and gene 3′UTR length were calculated directly from the gene annotation. Branchpoint sequences for each intron were estimated by eye based on sequence homology to known branchpoint sequences, and the distance from the branchpoint “A” to the 3′SS was calculated. For splice site sequence analysis, the MaxEnt webserver was used to calculate a splice site score for the 5′SS (using the sequence 3 nt upstream to 5 nt downstream) and for the 3′SS (using the sequence 20 nt upstream to 3 nt downstream) (Yeo and Burge, 2004). For secondary structure prediction, intron sequences were input into the RNAfold webserver and the output minimum free energy (kcal/mol) was tabulated for each intron (http://rna.tbi.univie.ac.at//cgi-bin/RNAWebSuite/RNAfold.cgi).

All features for each intron were tabulated and plotted against the fraction spliced calculated as above.

Illumina RNA Sequencing and Data Analysis

Total RNA was quantitated with a Qubit fluorometer, and 1 μg of each sample was transcribed to generate cDNA libraries using the TruSeq Stranded mRNA Sample Prep Kit (Illumina Inc, San Diego, CA) following the manufacturers protocol with the following modifications: 10 cycles were used to enrich DNA fragments to minimize the risk of over-amplification and dual-indexed barcode adapters were applied to each library. Libraries were pooled in an equimolar ratio for sequencing on the Illumina HiSeq 4000 System (Illumina Inc., San Diego, CA). The raw Illumina sequencing reads in FASTQ format were evaluated for quality using FastQC v0.11.9 (www.bioinformatics.babraham.ac.uk/projects/fastqc). The Illumina reads for three rich media replicates were mapped to the *C. merolae* 10D reference genome (Matsuzaki et al., 2004; Nozaki et al., 2007), which was obtained from release 46 of Ensembl Plants (Dobin et al., 2012), using STAR v2.7.1a (Dobin et al., 2012). IRFinder v1.2.5 (Middleton et al., 2017) was used to determine the average intron retention ratio for each intron, which was then subtracted from 1 to calculate the fraction of spliced reads.

RESULTS

Long-Read Sequencing Reveals Splicing and Alternative Polyadenylation Status of *C. merolae* mRNAs

To prepare libraries for LRS, *C. merolae* was cultivated either in RM or LN condition for 60 min. Total RNA was then isolated from both samples, and library preparation proceeded in two different ways (**Figure 1.A**). First, we used part of the total RNA for reverse transcription using an anchored oligo dT reverse primer, enriching for transcripts with polyA tails on mature mRNAs. Since splicing and polyadenylation are primarily co-transcriptional events, we aimed to also detect precursor RNAs that do not possess polyA tails. Therefore, in the second type of LRS library total RNA was polyadenylated prior to reverse transcription by anchored oligo dT. Thus, we captured both polyA+ and polyA− transcripts. Importantly, ribosomal RNA was removed before reverse transcription. As commercially available kits for the ribosomal depletion did not efficiently remove *C. merolae* rRNA, we designed our own set of probes against *C. merolae* rRNA (**Supplementary Table S1**). The sequencing on a Pacific Biosciences (PacBio) SMRT cell 8M with the Sequel II system resulted in 905,977 and 522,782 mapped reads for the polyA + libraries and 727,053 and 654,955 reads for the polyA+/− libraries for RM and LN samples respectively (**Figure 1A**). The reads were mapped to the updated *C. merolae* genome annotation 10D (unpublished). **Figure 1B** shows that our LRS experiment captures different mRNA isoforms. Comparison of the fraction of spliced reads resulted in a high correlation between the two library preparation strategies (**Figure 1C**), which indicates that their merge will likely not distort our data analysis. To improve the gene coverage for the downstream data analysis, we therefore combined both polyA+ and polyA ± libraries.

Alternative Polyadenylation Is Frequent in *C. merolae*

To assess the frequency of APA in *C. merolae* we performed a transcriptome-wide analysis of all protein-coding genes. PAS’s were identified by peak calling of read 3′ end positions after trimming of the polyA tails. A peak was considered a PAS when at least four reads ended at the same nucleotide. **Figures 2A,B** show example of reads mapped to a gene with at least four PAS’s and of 3′end peaks of genes with one or two PAS’s. The metagene analysis of the most prominent peak for each transcript showed cleavage approximately 20 nucleotides to both sides of the primary peak (**Figure 2C**), indicating that the 3′ end cleavage downstream of a particular polyA signal can occur within a sequence spanning 30–40 nucleotides. Based on this observation the peaks were considered as two different PAS’s if they were at least 18 nucleotides apart. Afterwards, the number of detected peaks with these parameters were counted for each gene. Surprisingly, we detected 63% of genes with more than one PAS and 34% with more than two PAS’s in RM conditions (**Figure 2D**). Our results indicate that APA is a frequent event in *C. merolae.*

C. merolae Uses UAAA Tetranucleotide as a Polyadenylation Signal

Cleavage of pre-mRNA at the PAS is stimulated by the polyadenylation signal (polyA signal) typically located 10–30 nt upstream of the PAS (Proudfoot, 1991; Colgan & Manley, 1997). Therefore, as our next step, we aimed to identify the polyA signals used in *C. merolae.* To find the motif for the preferentially used PAS’s with the most prominent peaks,

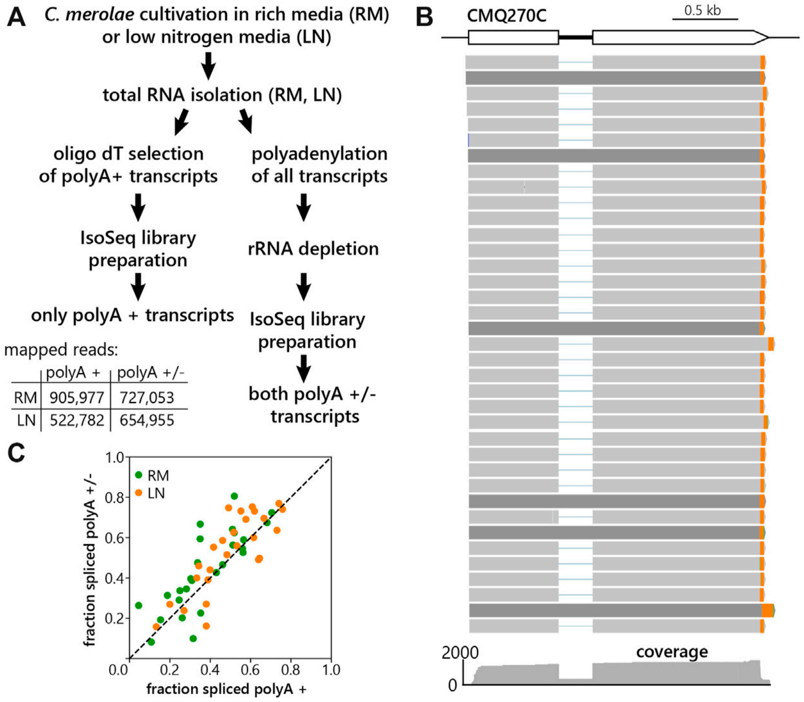

FIGURE 1 | Long-read sequencing of *C. merolae* polyA+ and polyA- RNA. **(A)** Workflow of the experiment with indicated numbers of mapped reads resulting from individual library preparations. **(B)** Long reads aligned to the reference genome, using the intron-containing gene CMQ270 as an example. 3′-ends are marked in orange. **(C)** Correlation of fraction of spliced reads obtained from the polyA+ and polyA+/– libraries prepared from the RM (rich media, green) and LN (low nitrogen, orange) samples of *C. merolae*. Fraction spliced was calculated as the number of intron-spanning reads which were spliced divided by the total number of intron-spanning reads for each gene shown. Pearson's correlation = 0.80 (RM), 0.82 (LN).

upstream sequences were extracted and the motif discovery tool XSTREME was used (Grant and Bailey, 2021). The polyA signal motif for the primary, i.e. most frequently occurring, PAS was found to be UAAA in 93% of genes (**Figure 3A**). Similarly, the UAAA tetranucleotide was recognized as a polyA signal motif also for the secondary and tertiary PAS's, in 63 and 52% of genes, respectively (**Figure 3A**). We did not identify any other significant consensus motifs among the sequences that did not contain UAAA. The distance between the polyA signal and PAS is in the range of 3–35 nucleotides with an average of 17 nucleotides (**Figure 3B**). These results indicate that the UAAA tetranucleotide serves as a polyA signal in *C. merolae*.

Intron-Containing Genes in *C. merolae* Display a Wide Variety of Splicing Efficiencies.

We calculated the fraction of spliced reads for each of the 26 genes annotated as intron-containing (Matsuzaki et al., 2004). The number of spliced reads was divided by the total number of intron-spanning reads for each gene. Surprisingly, *C. merolae* genes display a wide spectrum of splicing efficiencies from about 10 to 70% spliced reads (**Figure 4A**). This indicates the presence of a large proportion of transcripts with retained introns that contain polyA tails. In addition, splicing is almost uniformly increased in the LN condition, suggesting stress-induced splicing up-regulation (**Figures 4A,B**). Though subtle in most cases, the increase in the overall spliced fraction in the LN condition is statistically significant with a *p*-value of 0.014 (**Figure 4B**). Importantly, the comparison between our LRS data and Illumina short-read sequencing data obtained from the same *C.merolae* samples showed good correlation (**Figure 4C**). Thus, the splicing efficiency was confirmed by two independent approaches.

Next, we aimed to analyze individual intron features (**Figure 5A**) to see if they predict splicing efficiency (**Figures 5B,C**). We calculated 3′ splice site (using nucleotides -20:+3) and 5′ splice site (using nucleotides -3:+5) scores, reflecting the similarity with the human consensus sequences; the splicing branchpoint to 3′ splice site distance; GC content; predicted secondary structure and lengths of intron and 3′UTR of the gene. Overall, 5′ splice sites display higher similarity to the human consensus sequence over the 3′ splice site (**Figure 5B**). However, we did not observe a strong correlation between the fraction of spliced reads and any of the tested features, and thus none of the features analyzed thus far proved to be a reliable predictor(s) of splicing efficiency (**Figure 5C**).

DISCUSSION

In this study, we present the first analysis of pre-mRNA processing on a global scale in the thermotolerant alga, *C.*

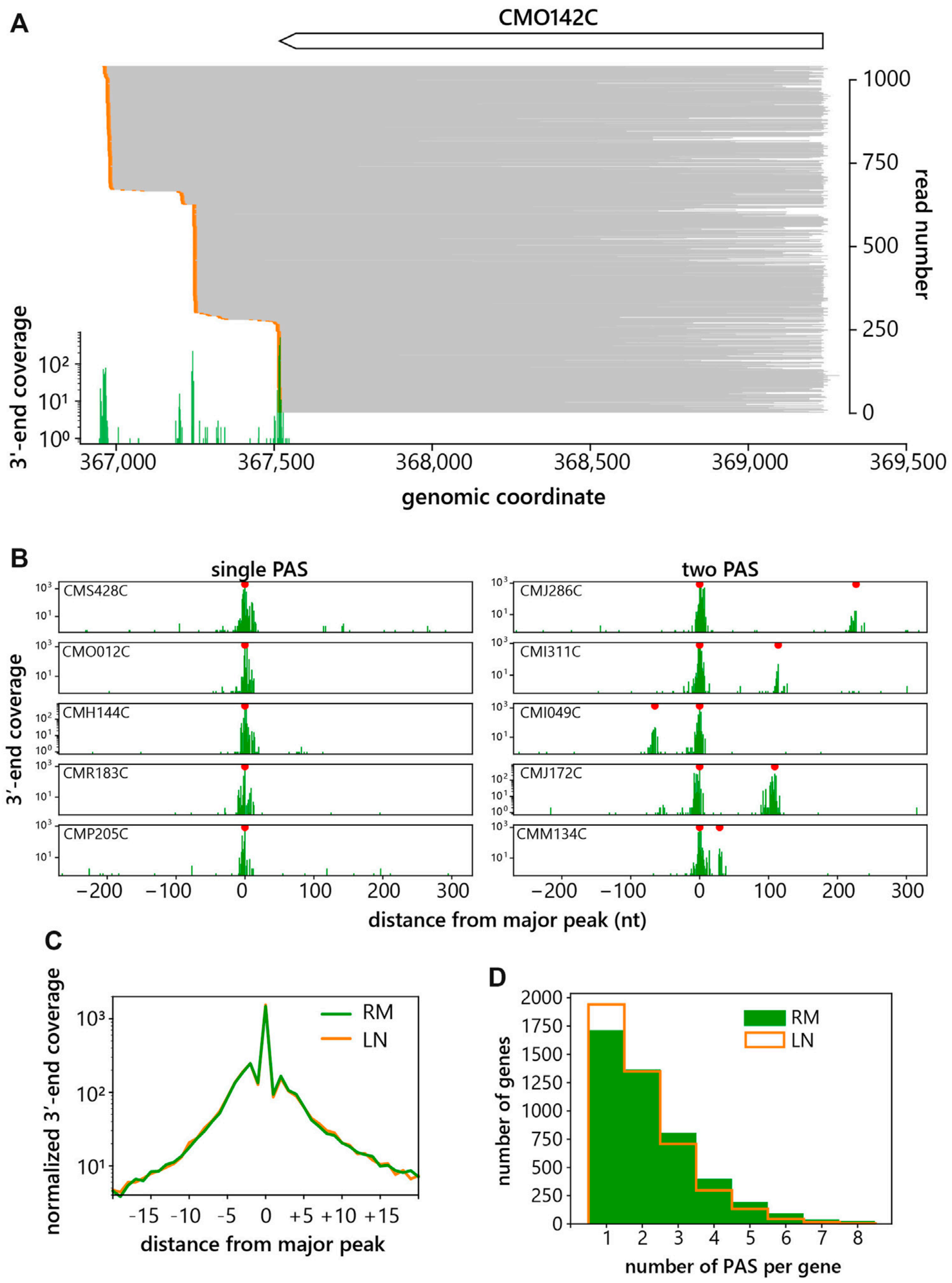

FIGURE 2 | Alternative polyadenylation in *C. merolae.* **(A)** Example of a gene CMO142C with at least four PAS's showing aligned reads in gray, 3′-ends in orange and a histogram of 3′-end coverage in green. **(B)** Examples of 3′-end coverage distributions for genes with one or two PAS's. Red dots indicate called peaks that passed the thresholds. **(C)** Metagene analysis of primary PAS peaks. The major peak is at the position of nucleotide 0, RM = rich media, LN = low nitrogen. **(D)** Histogram of the number of PAS's per gene. RM = rich media, LN = low nitrogen.

merolae, using long-read sequencing of individual RNA molecules. Our datasets yield information about *C. merolae* pre-mRNA splicing and 3′ end cleavage—two steps that have the potential to generate diverse mRNA isoforms. Using LRS, we were able to quantify the frequency of APA in *C. merolae* and the splicing efficiency of its intron-containing genes. Our results indicate frequent APA, because 63% of protein-coding genes have more than one PAS (**Figure 2B**). The UAAA tetranucleotide was identified as the major motif defining the polyA signal. Interestingly, nitrogen starvation slightly improved otherwise overall poor splicing efficiencies of the 27 originally identified *C. merolae* introns (**Figure 4B**).

Our comprehensive transcriptome-wide analysis of APA in this red alga indicate similarities with the green alga, *Chlamydomonas reinhardtii* that was found to have 40% or 68% of genes with APA in two different studies (Bell et al.,

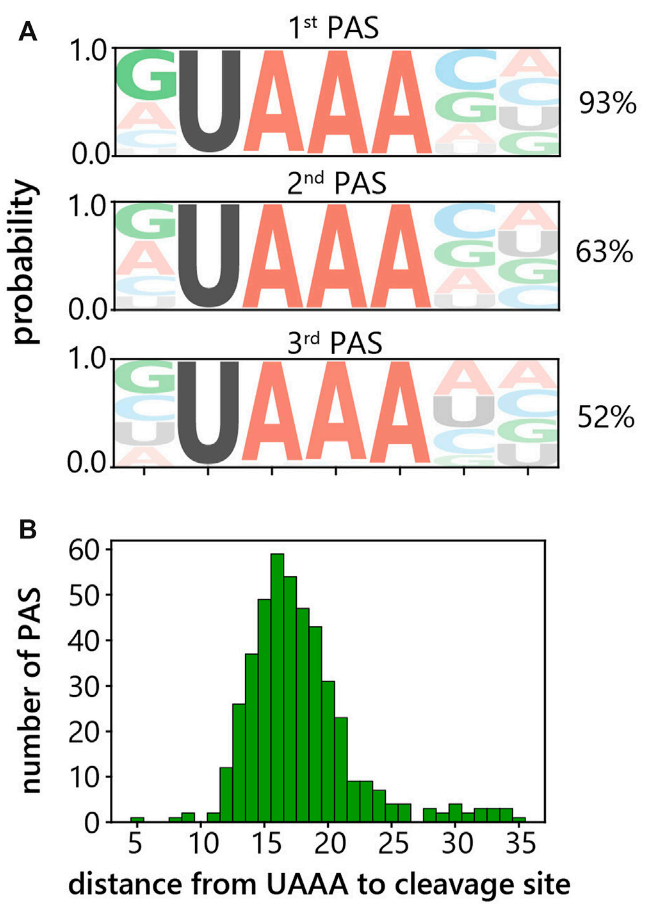

FIGURE 3 | Identification of polyadenylation signal in *C. merolae.* **(A)** Motif analysis of 500 sequences each upstream of primary, secondary and tertiary 3′-end peak. Percentages indicate how many sequences out of the input contained the motif. **(B)** Histogram of the distance between upstream UAAA tetranucleotide and primary cleavage site.

2016; Zhao et al., 2014). UGUAA was recognized to serve as the major polyA signal motif in *C. reinhardtii* and was used with a frequency of about 95% (Bell et al., 2016; Li and Du, 2014; Shen et al., 2008; Wodniok et al., 2007; Zhao et al., 2014). Our findings identified the UAAA motif as the primary polyA signal in *C. merolae* (**Figure 3A**), resembling the major canonical polyA signal AAUAAA used in a wide range of species, including human, mice, *Caenorhabditis elegans*, *Drosophila melanogaster* as well as budding and fission yeasts (Graber et al., 1999; Tian et al., 2005; Kamasawa & Horiuchi, 2008; Mangone et al., 2010; Ozsolak et al., 2010). In the first study attempting to predict *C. merolae* PASs based on genomic sequence and expressed sequence tags, the authors concluded that UAAA motif is present in 86.55% of transcripts and GUAA in 43.23% of transcripts (Zhao et al., 2019). This is partially consistent with our findings; we did not identify GUAA as a polyA signal motif.

The frequency of intron retention in 27 *C. merolae* mRNAs (30–90% on a per intron basis, see **Figure 4A**) suggests relatively inefficient pre-mRNA splicing compared with other species where introns are removed co-transcriptionally (Carrocci & Neugebauer, 2019). The significant correlation of fraction of spliced reads between our polyA+ and polyA+/− libraries (**Figure 1C**) suggests that *C. merolae* splicing is post-transcriptional. Thus, the long-read data agree with the first study of splicing in this organism, which revealed a large unspliced fraction by RT-PCR (Stark et al., 2015). One of the most plausible explanations for the abundant intron retention in *C. merolae* is the missing U1 snRNP and overall reduced spliceosome (Stark et al., 2015). Our 5′splice site motif analysis identified the sequence GUAAGU (**Figure 5B**), which is almost identical with the vertebrate 5′splice site consensus sequence that is recognized by U1 snRNP (Rosbash and Seraphin, 1991). It remains to be elucidated which factor recognizes the 5′splice site motif in *C. merolae.* Although we cannot exclude the possibility that a U1 snRNA gene has eluded sequencing or annotation, a distinct possibility is that the sequence conservation is due to base-pairing with U5 and U6 snRNAs, which exhibit unusual complementarity to 5′splice sites in *C. merolae* (Stark et al., 2015).

Intriguingly, other intron-sparse organisms, such as *Encephalitozoon cuniculi*, *Guillardia theta* and *Porphyridium purpureum*, also display low splicing efficiency (Grisdale et al.,

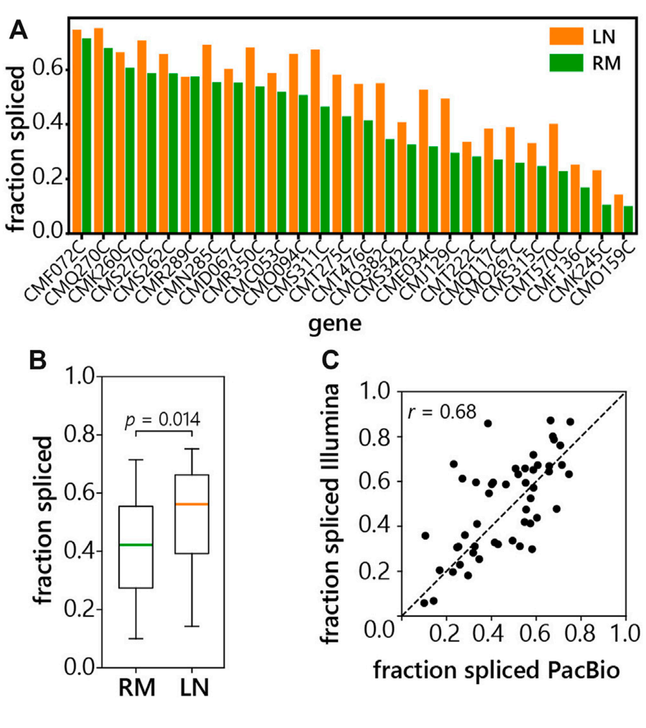

FIGURE 4 | Splicing efficiency of *C. merolae* intron-containing genes. **(A)** LRS was used to measure the fraction of reads that were spliced at each intron-containing gene in *C. merolae*, fraction spliced was calculated as in **Figure 1B** (see above), LN (low-nitrogen, orange), RM (rich media, green). **(B)** Box plot of a fraction of spliced reads grown in RM or LN. Significance of the difference between fraction of spliced reads was tested by two-sided Mann-Whitney U test. **(C)** Correlation of fraction of spliced reads obtained from the short-read (Illumina) and long-read (PacBio) libraries prepared from samples of *C. merolae.* Fraction spliced was calculated as the number of intron-spanning reads which were spliced divided by the total number of intron-spanning reads for each gene shown. Pearson's correlation = 0.68.

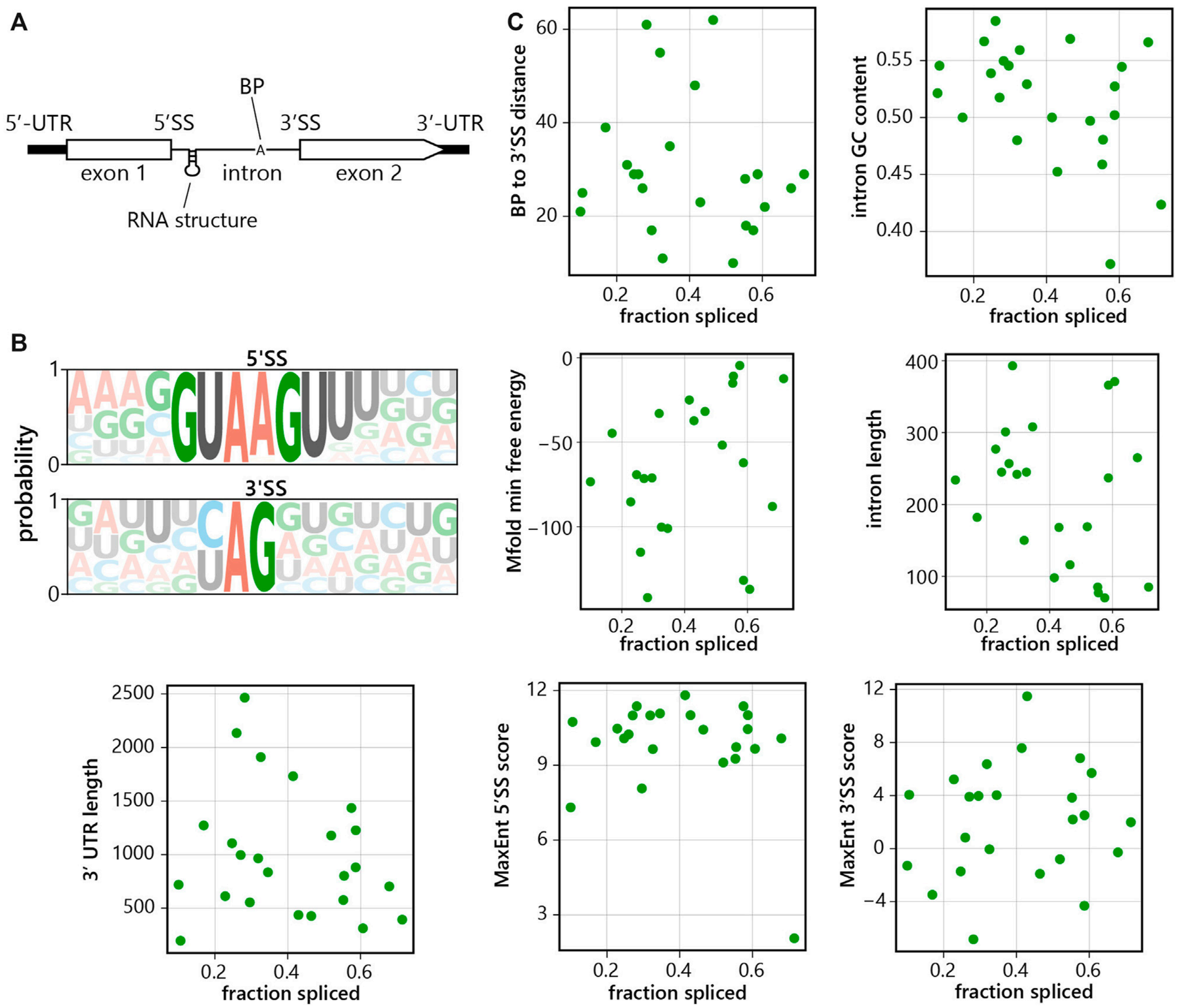

FIGURE 5 | *C. merolae* intron features do not reliably predict splicing efficiency. **(A)** Scheme of intron-containing gene, UTR = untranslated region, SS = splice site, BP = branch point. **(B)** 5′splice site (5′SS) and 3′splice site (3′SS) consensus sequences of *C. merolae* introns. **(C)** Intrinsic features of each intron in *C. merolae* were analyzed to determine their correlation with splicing efficiency (as calculated above, **Figure 1B**). Each point represents a single intron. Features analyzed included: branchpoint to 3′SS distance (Pearson's correlation (*r*) = –0.148, *p* = 0.489), 3′SS score (*r* = 0.118, *p* = 0.582), 5′SS score (*r* = -0.237, *p* = 0.264), GC content (*r* = –0.413, *p* = 0.045), predicted secondary structure (*r* = 0.408, *p* = 0.048), intron length (*r* = -0.421, *p* = 0.040), and gene 3′UTR length (*r* = –0.191, *p* = 0.371).

2013; Wong et al., 2018; Wong et al., 2021). Transcriptome-wide sequencing analysis of the microsporidium *E. cuniculi,* with 37 annotated introns in a 2.9 Mb genome, revealed a low splicing rate as more than half of its introns remained unspliced in >80% of reads (Katinka et al., 2001; Grisdale et al., 2013). Comparably to in *C. merolae,* the spliceosome of *E. cuniculi* is reduced and no U1 snRNA has been identified (Katinka et al., 2001; Davila Lopez et al., 2008). Another example of an extensive intron retention is found in the 551 kb nucleomorph genome of the cryptomonad alga *Guillardia theta* that contains 17 introns (Wong et al., 2018). Finally, the 22.1 Mb genome of the red alga *P. purpureum* was predicted to contain only around 200 introns (Bhattacharya et al., 2013; Wong et al., 2021). Approximately half of its transcripts were found to remain unspliced, and the predicted U1 snRNA in *P. purpureum* is unusually long with an atypical folding pattern (Wong et al., 2021). Similar to *C. merolae,* some of the spliceosome components are reduced in *P. purpureum* (Wong et al., 2021). Therefore, reduced spliceosomes may to contribute to the poor splicing efficiencies in these organisms. An alternative possibility is that the high level of intron-containing transcripts is due to a decreased degradation rate, even though many RNA degradation factors were bioinformatically predicted to be encoded by the *C. merolae* genome (Reimer et al., 2017).

Based on our observations, nitrogen starvation improves splicing in *C. merolae* (**Figures 4A,B**) suggesting that genes with improved splicing may code for proteins contributing to cell survival in these stress conditions. Studies in another intron-sparse organism, the budding yeast *S. cerevisiae*, revealed that the accumulation of introns of specific genes helps the organism to cope with nutrient deprivation (Morgan et al., 2019; Parenteau et al., 2019). Even though the trend is opposite to *C. merolae* it implies the possibility that introns themselves might have a role in growth under normal or stress conditions. For example, the Johnson lab showed in budding yeast that some retained introns can provide alternative start codons to produce alternative protein products in response to changing environment (Hossain et al., 2016). Interestingly, *C. merolae* genes have intron-encoded AUG's, some with the potential to serve as alternative start codons and produce shortened protein isoforms.

Altogether, *C. merolae* provides a valuable platform for understanding the molecular basis of these complex and essential processes in a simplified system. It would be interesting to use *C. merolae* as a model system to address basic mechanisms of coupling between pre-mRNA processing events, such as splicing and cleavage. A prime example is the failure of cleavage at PAS that lead to transcriptional read-through and the production of 5–45 kb long Downstream of Gene (DoG) transcripts (Hangauer et al., 2013; Vilborg et al., 2015; Vilborg et al., 2017). DoGs are transcripts produced by RNA polymerase II under stress conditions and may function as first responders that help the cell to survive. Further investigation of splicing in *C. merolae* might help to elucidate unknown details of the mechanism of splicing since its simplified spliceosome reveals the necessity or dispensability of individual spliceosome factors. Similarities in the polyA cleavage machinery are expected, based on sequence similarities in polyA signals revealed here. However, significant differences could also be expected in *C. merolae* when this machinery is more deeply investigated in future studies.

DATA AVAILABILITY STATEMENT

The sequencing data generated in this study are deposited in NCBI SRA under the BioProject accession number PRJNA785740 (https://dataview.ncbi.nlm.nih.gov/object/PRJNA785740?reviewer=n4m3m499ajm2shl4hsm9l7e 6fc.).

AUTHOR CONTRIBUTIONS

LS analyzed the PacBio data, prepared figures and edited the article. DZ wrote the article and edited figures. KAR analyzed the PacBio data, MRS, SDR, and PJB designed the nitrogen depletion experiment. MRS grew the cultures, and prepared total RNA. NJF developed methods for ribodepletion, prepared libraries for PacBio sequencing, and sequenced the RNA. MLW prepared libraries for Illumina sequencing and sequenced the RNA. LZ processed the Illumina data. VAS optimized splicing measurements for the Illumina data. KMN edited the article.

FUNDING

This work was supported by the National Institutes of Health (R01 GM112766, R01 GM140735, and R21 HL150642 to KN), by GenomeBC Sector Innovation Program grant RC18-3517 and NSERC Discovery Grant 298521 (SR) and by the Intramural Research Program of the NIEHS, a component of the National Institutes of Health (MW, LZ, and PB).

ACKNOWLEDGMENTS

We are grateful to members of the Rader, Blackshear and Neugebauer labs for helpful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.818697/full#supplementary-material

REFERENCES

Agafonov, D. E., Deckert, J., Wolf, E., Odenwälder, P., Bessonov, S., Will, C. L., et al. (2011). Semiquantitative Proteomic Analysis of the Human Spliceosome via a Novel Two-Dimensional Gel Electrophoresis Method. *Mol. Cel Biol* 31 (13), 2667–2682. doi:10.1128/Mcb.05266-11

Alpert, T., Straube, K., Oesterreich, F. C., Herzel, L., and Neugebauer, K. M., (2020). Widespread Transcriptional Readthrough Caused by Nab2 Depletion Leads to Chimeric Transcripts with Retained Introns. *Cell Rep.* 33 (13), 108324–108331. doi:10.1016/j.celrep.2020.108496

Bell, S. A., Shen, C., Brown, A., and Hunt, A. G. (2016). Experimental Genome-wide Determination of RNA Polyadenylation in Chlamydomonas Reinhardtii. *PLoS One* 11 (1), e0146107. doi:10.1371/journal.pone.0146107

Berg, M. G., Singh, L. N., Younis, I., Liu, Q., Pinto, A. M., Kaida, D., et al. (2012). U1 snRNP Determines mRNA Length and Regulates Isoform Expression. *Cell* 150 (1), 53–64. doi:10.1016/j.cell.2012.05.029

Bhattacharya, D., Price, D. C., Chan, C. X., Qiu, H., Rose, N., Ball, S., et al. (2013). Genome of the Red Alga Porphyridium Purpureum. *Nat. Commun.* 4, 1941. doi:10.1038/ncomms2931

Black, C. S., Garside, E. L., MacMillan, A. M., and Rader, S. D. (2016). Conserved Structure of Snu13 from the Highly Reduced Spliceosome of Cyanidioschyzon Merolae. *Protein Sci.* 25 (4), 911–916. doi:10.1002/pro.2894

Carrocci, T. J., and Neugebauer, K. M. (2019). Pre-mRNA Splicing in the Nuclear Landscape. *Cold Spring Harb Symp. Quant Biol.* 84, 11–20. doi:10.1101/sqb.2019.84.040402

Colgan, D. F., and Manley, J. L. (1997). Mechanism and Regulation of mRNA Polyadenylation. *Genes Dev.* 11 (21), 2755–2766. doi:10.1101/gad.11.21.2755

Dale, R. K., Pedersen, B. S., and Quinlan, A. R. (2011). Pybedtools: a Flexible Python Library for Manipulating Genomic Datasets and Annotations. *Bioinformatics* 27 (24), 3423–3424. doi:10.1093/bioinformatics/btr539

Derti, A., Garrett-Engele, P., Macisaac, K. D., Stevens, R. C., Sriram, S., Chen, R., et al. (2012). A Quantitative Atlas of Polyadenylation in Five Mammals. *Genome Res.* 22 (6), 1173–1183. doi:10.1101/gr.132563.111

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2012). STAR: Ultrafast Universal RNA-Seq Aligner. *Bioinformatics* 29 (1), 15–21. doi:10.1093/bioinformatics/bts635

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. *Mol. Cel* 36 (4), 593–608. doi:10.1016/j.molcel.2009.09.040

Graber, J. H., Cantor, C. R., Mohr, S. C., and Smith, T. F. (1999). In Silico detection of Control Signals: mRNA 3'-End-Processing Sequences in Diverse Species. *Proc. Natl. Acad. Sci.* 96 (24), 14055–14060. doi:10.1073/pnas.96.24.14055

Grant, C. E., and Bailey, T. L. (2021). XSTREME: Comprehensive Motif Analysis of Biological Sequence Datasets. *BioRxiv*. doi:10.1101/2021.09.02.458722

Grisdale, C. J., Bowers, L. C., Didier, E. S., and 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. doi:10.1186/1471-2164-14-207

Hangauer, M. J., Vaughn, I. W., and McManus, M. T. (2013). Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *Plos Genet.* 9 (6), e1003569. doi:10.1371/journal.pgen.1003569

Herzel, L., Straube, K., and Neugebauer, K. M. (2018). Long-read Sequencing of Nascent RNA Reveals Coupling Among RNA Processing Events. *Genome Res.* 28 (7), 1008–1019. doi:10.1101/gr.232025.117

Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., et al. (2013). Analysis of Alternative Cleavage and Polyadenylation by 3′ Region Extraction and Deep Sequencing. *Nat. Methods* 10 (2), 133–139. doi:10.1038/nmeth.2288

Hossain, M. A., Claggett, J. M., Edwards, S. R., Shi, A., Pennebaker, S. L., Cheng, M. Y., et al. (2016). Posttranscriptional Regulation of Gcr1 Expression and Activity Is Crucial for Metabolic Adjustment in Response to Glucose Availability. *Mol. Cel* 62 (3), 346–358. doi:10.1016/j.molcel.2016.04.012

Jalkanen, A. L., Coleman, S. J., and Wilusz, J. (2014). Determinants and Implications of mRNA Poly(A) Tail Size - Does This Protein Make My Tail Look Big? *Semin. Cel Develop. Biol.* 34, 24–32. doi:10.1016/j.semcdb.2014.05.018

Kamasawa, M., and Horiuchi, J. (2008). Identification and Characterization of Polyadenylation Signal (PAS) Variants in Human Genomic Sequences Based on Modified EST Clustering. *Silico Biol.* 8 (3-4), 347–361. https://www.ncbi.nlm.nih.gov/pubmed/19032167.

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 (6862), 450–453. doi:10.1038/35106579

Li, H. (2018). Minimap2: Pairwise Alignment for Nucleotide Sequences. *Bioinformatics* 34 (18), 3094–3100. doi:10.1093/bioinformatics/bty191

Li, X.-Q., and Du, D. (2014). Motif Types, Motif Locations and Base Composition Patterns Around the RNA Polyadenylation Site in Microorganisms, Plants and Animals. *BMC Evol. Biol.* 14, 162. doi:10.1186/s12862-014-0162-7

Liu, X., Freitas, J., Zheng, D., Oliveira, M. S., Hoque, M., Martins, T., et al. (2017). Transcription Elongation Rate Has a Tissue-specific Impact on Alternative Cleavage and Polyadenylation in *Drosophila melanogaster*. *RNA* 23 (12), 1807–1816. doi:10.1261/rna.062661.117

Liu, X., Hoque, M., Larochelle, M., Lemay, J.-F., Yurko, N., Manley, J. L., et al. (2017). Comparative Analysis of Alternative Polyadenylation in *S. cerevisiae* and *S. pombe*. *Genome Res.* 27 (10), 1685–1695. doi:10.1101/gr.222331.117

López, M. D., Alm Rosenblad, M., and 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. doi:10.1093/nar/gkn142

Mangone, M., Manoharan, A. P., Thierry-Mieg, D., Thierry-Mieg, J., Han, T., Mackowiak, S. D., et al. (2010). The Landscape of *C. elegans* 3′UTRsThe Landscape of *C. elegans* 3'UTRs. *Science* 329 (5990), 432–435. doi:10.1126/science.1191244

Matsuzaki, M., Misumi, O., Shin-i, T., Maruyama, S., Takahara, M., Miyagishima, S.-y., et al. (2004). Genome Sequence of the Ultrasmall Unicellular Red Alga Cyanidioschyzon Merolae 10D. *Nature* 428 (6983), 653–657. doi:10.1038/nature02398

Mayr, C. (2019). What Are 3′ UTRs Doing?What Are 3' UTRs Doing? *Cold Spring Harb Perspect. Biol.* 11 (10), a034728. doi:10.1101/cshperspect.a034728

Middleton, R., Gao, D., Thomas, A., Singh, B., Au, A., Wong, J. J.-L., et al. (2017). IRFinder: Assessing the Impact of Intron Retention on Mammalian Gene Expression. *Genome Biol.* 18 (1), 51. doi:10.1186/s13059-017-1184-4

Morgan, J. T., Fink, G. R., and Bartel, D. P. (2019). Excised Linear Introns Regulate Growth in Yeast. *Nature* 565 (7741), 606–611. doi:10.1038/s41586-018-0828-1

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 Biol.*, 5, 28. doi:10.1186/1741-7007-5-28

Oh, J.-M., Di, C., Venters, C. C., Guo, J., Arai, C., So, B. R., et al. (2017). U1 snRNP Telescripting Regulates a Size-Function-Stratified Human Genome. *Nat. Struct. Mol. Biol.* 24 (11), 993–999. doi:10.1038/nsmb.3473

Ohnuma, M., Yokoyama, T., Inouye, T., Sekine, Y., and Tanaka, K. (2008). Polyethylene Glycol (PEG)-Mediated Transient Gene Expression in a Red Alga, Cyanidioschyzon Merolae 10D. *Plant Cel Physiol.* 49 (1), 117–120. doi:10.1093/pcp/pcm157

Ohta, N., Matsuzaki, M., Misumi, O., Miyagishima, S. Y., Nozaki, H., Tanaka, K., et al. (2003). Apr 30)Complete Sequence and Analysis of the Plastid Genome of the Unicellular Red Alga Cyanidioschyzon Merolae. *DNA Res.* 10 (2), 67–77. doi:10.1093/dnares/10.2.67

Ohta, N., Sato, N., and Kuroiwa, T. (1998). Structure and Organization of the Mitochondrial Genome of the Unicellular Red Alga Cyanidioschyzon Merolae Deduced from the Complete Nucleotide Sequence. *Nucleic Acids Res.* 26 (22), 5190–5198. doi:10.1093/nar/26.22.5190

Ozsolak, F., Kapranov, P., Foissac, S., Kim, S. W., Fishilevich, E., Monaghan, A. P., et al. (2010). Comprehensive Polyadenylation Site Maps in Yeast and Human Reveal Pervasive Alternative Polyadenylation. *Cell* 143 (6), 1018–1029. doi:10.1016/j.cell.2010.11.020

Parenteau, J., Maignon, L., Berthoumieux, M., Catala, M., Gagnon, V., and Abou Elela, S. (2019). Introns Are Mediators of Cell Response to Starvation. *Nature* 565 (7741), 612–617. doi:10.1038/s41586-018-0859-7

Proudfoot, N. (1991). Poly(A) Signals. *Cell* 64 (4), 671–674. doi:10.1016/0092-8674(91)90495-k

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a Flexible Suite of Utilities for Comparing Genomic Features. *Bioinformatics* 26 (6), 841–842. doi:10.1093/bioinformatics/btq033

Rehfeld, A., Plass, M., Krogh, A., and Friis-Hansen, L. (2013). Alterations in Polyadenylation and its Implications for Endocrine Disease. *Front. Endocrinol.* 4, 53. doi:10.3389/fendo.2013.00053

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 (6), 952–967. doi:10.1261/rna.058487.116

Reimer, K. A., Mimoso, C. A., Adelman, K., and Neugebauer, K. M., (2021). Co-Transcriptional Splicing Regulates 3′ End Cleavage During Mammalian Erythropoiesis. *Mol. Cell.* 81 (5), 998–1012.e7. doi:10.1016/j.molcel.2020.12.018

Ren, F., Zhang, N., Zhang, L., Miller, E., and Pu, J. J. (2020). Alternative Polyadenylation: a New Frontier in post Transcriptional Regulation. *Biomark Res.* 8 (1), 67. doi:10.1186/s40364-020-00249-6

Rosbash, M., and Se´raphin, B. (1991). Who's on First? the U1 snRNP-5 Splice Site Interaction and Splicing. *Trends Biochem. Sci.* 16 (5), 187–190. doi:10.1016/0968-0004(91)90073-5

Shen, Y., Liu, Y., Liu, L., Liang, C., and Li, Q. Q. (2008). Unique Features of Nuclear mRNA Poly(A) Signals and Alternative Polyadenylation in Chlamydomonas Reinhardtii. *Genetics* 179 (1), 167–176. doi:10.1534/genetics.108.088971

Singh, R. N., and Singh, N. N. (2019). A Novel Role of U1 snRNP: Splice Site Selection from a Distance. *Biochim. Biophys. Acta (Bba) - Gene Regul. Mech.* 1862 (6), 634–642. doi:10.1016/j.bbagrm.2019.04.004

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. *Proc. Natl. Acad. Sci. USA* 112 (11), E1191–E1200. doi:10.1073/pnas.1416879112

Tian, B., Hu, J., Zhang, H., and Lutz, C. S. (2005). A Large-Scale Analysis of mRNA Polyadenylation of Human and Mouse Genes. *Nucleic Acids Res.* 33 (1), 201–212. doi:10.1093/nar/gki158

Vilborg, A., Passarelli, M. C., Yario, T. A., Tycowski, K. T., and Steitz, J. A. (2015). Widespread Inducible Transcription Downstream of Human Genes. *Mol. Cel* 59 (3), 449–461. doi:10.1016/j.molcel.2015.06.016

Vilborg, A., Sabath, N., Wiesel, Y., Nathans, J., Levy-Adam, F., Yario, T. A., et al. (2017). Comparative Analysis Reveals Genomic Features of Stress-Induced Transcriptional Readthrough. *Proc. Natl. Acad. Sci. USA* 114 (40), E8362–E8371. doi:10.1073/pnas.1711120114

Wodniok, S., Simon, A., Glöckner, G., and Becker, B. (2007). Gain and Loss of Polyadenylation Signals during Evolution of green Algae. *BMC Evol. Biol.* 7, 65. doi:10.1186/1471-2148-7-65)

Wong, D. K., Grisdale, C. J., and Fast, N. M. (2018). Evolution and Diversity of Pre-mRNA Splicing in Highly Reduced Nucleomorph Genomes. *Genome Biol. Evol.* 10 (6), 1573–1583. doi:10.1093/gbe/evy111

Wong, D. K., Stark, M. S., Rader, S. D., and Fast, N. M. (2021). Characterization of Pre-mRNA Splicing and Spliceosomal Machinery in Porphyridium Purpureum and Evolutionary Implications for Red Algae. *J. Eukaryot. Microbiol.* 68 (3), e12844. doi:10.1111/jeu.12844

Yeo, G., and Burge, C. B. (2004). Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *J. Comput. Biol.* 11 (2-3), 377–394. doi:10.1089/1066527041410418

Zhang, Y., Qian, J., Gu, C., and Yang, Y. (2021). Alternative Splicing and Cancer: a Systematic Review. *Sig Transduct Target. Ther.* 6 (1), 78. doi:10.1038/s41392-021-00486-7

Zhao, Z., Wu, X., Ji, G., Liang, C., and Li, Q. (2019). Genome-Wide Comparative Analyses of Polyadenylation Signals in Eukaryotes Suggest a Possible Origin of the AAUAAA Signal. *Ijms* 20 (4), 958. doi:10.3390/ijms20040958

Zhao, Z., Wu, X., Kumar, P. K. R., Dong, M., Ji, G., Li, Q. Q., et al. (2014). Bioinformatics Analysis of Alternative Polyadenylation in Green Alga Chlamydomonas Reinhardtii Using Transcriptome Sequences from Three Different Sequencing Platforms. *G3 (Bethesda)* 4 (5), 871–883. doi:10.1534/g3.114.010249

Zhou, Y., Zhu, J., Schermann, G., Ohle, C., Bendrin, K., Sugioka-Sugiyama, R., et al. (2015). The Fission Yeast MTREC Complex Targets CUTs and Unspliced Pre-mRNAs to the Nuclear Exosome. *Nat. Commun.* 6, 7050. doi:10.1038/ncomms8050

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.