Evolution and Diversity of Pre-mRNA Splicing in Highly Reduced Nucleomorph Genomes

Donald K. Wong[1], Cameron J. Grisdale[1,2], and Naomi M. Fast[1,*]

[1]Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada

[2]Present address: Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, Canada

*Corresponding author: E-mail: naomi.fast@ubc.ca.

Accepted: May 30, 2018

Data deposition: Sequenced reads from this project have been deposited at NCBI's Sequence Read Archive (SRA) under the accession SRP129823.

Abstract

Eukaryotic genes are interrupted by introns that are removed in a conserved process known as pre-mRNA splicing. Though well-studied in select model organisms, we are only beginning to understand the variation and diversity of this process across the tree of eukaryotes. We explored pre-mRNA splicing and other features of transcription in nucleomorphs, the highly reduced remnant nuclei of secondary endosymbionts. Strand-specific transcriptomes were sequenced from the cryptophyte *Guillardia theta* and the chlorarachniophyte *Bigelowiella natans*, whose plastids are derived from red and green algae, respectively. Both organisms exhibited elevated nucleomorph antisense transcription and gene expression relative to their respective nuclei, suggesting unique properties of gene regulation and transcriptional control in nucleomorphs. Marked differences in splicing were observed between the two nucleomorphs: the few introns of the *G. theta* nucleomorph were largely retained in mature transcripts, whereas the many short introns of the *B. natans* nucleomorph are spliced at typical eukaryotic levels (>90%). These differences in splicing levels could be reflecting the ancestries of the respective plastids, the different intron densities due to independent genome reduction events, or a combination of both. In addition to extending our understanding of the diversity of pre-mRNA splicing across eukaryotes, our study also indicates potential links between splicing, antisense transcription, and gene regulation in reduced genomes.

Key words: RNA-Seq, transcriptome, intron retention, cryptophyte, cryptomonad, chlorarachniophyte.

Introduction

The regulation and flow of information within a cell are vital processes. Many genes in eukaryotes are interrupted with intervening sequences known as introns, which are removed from transcripts via a ubiquitous process known as pre-mRNA splicing. A large complex of proteins and small nuclear RNAs (snRNAs) known as the spliceosome mediates this process. The proper assembly of the spliceosome and subsequent intron removal require conserved intronic sequence signals such as the 5′ splice site (most often "GU"), the 3′ splice site (most often "AG"), and a biochemically important branch point adenosine residue (Will and Lührmann 2011).

The presence of introns and conserved spliceosomal components across eukaryotes suggests that splicing is a mechanistically conserved process present in the last common ancestor of eukaryotes. Even organisms that have highly reduced or highly derived genomes can still have introns. Typically, these organisms have introns that are few in number, are short (~30 bp or less), or are both. In organisms with intron-sparse genomes, their spliceosomes often possess a reduced set of components (Katinka et al. 2001; Grisdale et al. 2013; Stark et al. 2015), and studying such reduced systems could provide insight into the core mechanisms of splicing. Although rare, there are examples of genomes that have lost all introns (Lane et al. 2007; Cuomo et al. 2012).

Although splicing has been studied extensively in budding yeast and humans, it is assumed that this process occurs with little variation across eukaryotes. However, splicing has recently been analyzed in detail in two very different organisms with reduced genomes, extending our understanding of the diversity of pre-mRNA splicing. The microsporidian intracellular parasite *Encephalitozoon cuniculi* has an extremely tiny genome of 2.9 Mbp, with only 37 annotated introns (Katinka et al. 2001; Lee et al. 2010). The extremophilic red

alga *Cyanidioschyzon merolae* also has a reduced genome (16.5 Mbp) and only 27 annotated introns (Matsuzaki et al. 2004; Stark et al. 2015). Transcriptomic studies of *E. cuniculi* have revealed a high frequency of intron retention—that is, a large proportion of mature mRNA still retain introns (Grisdale et al. 2013). Also, both of these species have only a limited complement of spliceosomal proteins. Indeed, in *C. merolae*, one of the five major subcomplexes of the spliceosome is missing (Stark et al. 2015).

Antisense transcription is a relatively rare process in eukaryotes where transcripts complementary to a "sense" gene are generated (Pelechano and Steinmetz 2013). Whereas most antisense transcripts are noncoding, some "antisense" transcripts are mRNA of oppositely oriented, neighboring genes. Such transcripts are inherently more common when genes are especially close to each other, as in reduced genomes (Williams et al. 2005; Pelechano and Steinmetz 2013). Long thought to only be errant transcriptional noise, it is now clear that antisense transcription can be another source of gene regulation (Wagner and Simons 1994; Vanhee-Brossollet and Vaquero 1998; Pelechano and Steinmetz 2013). This property could be especially relevant for reduced genomes, especially if antisense transcription could compensate for a reduced set of transcription factors and regulatory elements in the genome. Antisense transcripts bound to complementary mRNA can target it for degradation, for example by microRNAs (miRNA) or small interfering RNAs (siRNA; Ghildiyal and Zamore 2009; Moazed 2009; Pelechano and Steinmetz 2013). A bound antisense transcript can also mask splicing signals, leading to intron retention and other alternative splicing events (Morrissy et al. 2011).

The most reduced eukaryotic genomes in nature are the nucleomorphs, relict nuclei found within secondary plastids of two distant lineages of algae (Keeling 2004). While there are a number of algal lineages that have acquired photosynthesis by taking up an already-photosynthetic eukaryote as an endosymbiont, in only two independent lineages does the remnant endosymbiont nucleus (with an associated tiny genome) persist. The cryptophytes bear a plastid derived from a red alga, while the chlorarachniophytes acquired photosynthesis from a green alga. As the nucleomorph is derived from a eukaryotic nucleus, the genome has typical eukaryotic features, such as linear chromosomes, genes containing introns and pre-mRNA splicing. Nucleomorphs provide insight into the reduction of a eukaryotic genome through endosymbiosis (as opposed to parasitism, for example). The first two nucleomorphs sequenced were those of the cryptophyte *Guillardia theta* and the chlorarachniophyte *Bigelowiella natans* (Douglas et al. 2001; Gilson et al. 2006). Since then, additional nucleomorphs from both lineages have been sequenced (Lane et al. 2007; Tanifuji et al. 2011; Moore et al. 2012; Tanifuji, Onodera, Brown, et al. 2014; Suzuki et al. 2015).

Nucleomorph genomes vary in size, but none has been found to be >1 Mbp. Interestingly, all nucleomorphs observed so far (from both cryptophytes and chlorarachniophytes) carry their tiny genomes on three short linear chromosomes (Douglas et al. 2001; Gilson et al. 2006; Lane et al. 2007; Tanifuji et al. 2011; Moore et al. 2012; Tanifuji, Onodera, Brown, et al. 2014; Suzuki et al. 2015). Whether or not there is a functional significance to this convergence, rather than a mere coincidence, remains to be seen. As with any organellar genome, very few genes remain; many have been transferred to the host nucleus or lost. The vast majority of remaining nucleomorph genes is housekeeping genes, such as chaperone proteins, ribosomal proteins and those involved in DNA replication, along with the genes for rRNAs and an incomplete set of tRNAs (Douglas et al. 2001; Gilson et al. 2006; Lane et al. 2007; Tanifuji et al. 2011; Moore et al. 2012; Tanifuji, Onodera, Brown, et al. 2014; Suzuki et al. 2015).

Introns have been found in all but one of the nucleomorph genomes sequenced to date (Douglas et al. 2001; Gilson et al. 2006; Lane et al. 2007; Tanifuji et al. 2011; Moore et al. 2012; Tanifuji, Onodera, Brown, et al. 2014; Suzuki et al. 2015). However, the density of introns in the nucleomorph genomes of cryptophytes and chlorarachniophytes is quite different. For example, the cryptophyte *G. theta* has 485 tightly packed protein-coding genes in its 550 kbp nucleomorph genome, and only 17 of these genes are interrupted by introns that range from 42 bp to 52 bp in length (Douglas et al. 2001). In contrast, the smaller (370 kbp) nucleomorph genome of *B. natans* has almost 900 extremely short (18–21 bp) introns that interrupt a majority of the 283 protein-coding nucleomorph genes (Gilson et al. 2006; Tanifuji, Onodera, Brown, et al. 2014). Whereas canonical 5′ and 3′ splice sites are present in these tiny introns, other commonly conserved splicing motifs such as the branch donor adenosine are not discernible (Gilson et al. 2006; Tanifuji, Onodera, Brown, et al. 2014).

There have been a number of studies on the peculiarities of transcription in nucleomorphs (Williams et al. 2005; Hirakawa et al. 2011; Hirakawa et al. 2014; Tanifuji, Onodera, Moore, et al. 2014; Suzuki *et al.* 2016; Sanitá Lima and Smith 2017), although studies about the unique introns and pre-mRNA splicing of nucleomorphs are lacking. Here, we seek to understand pre-mRNA splicing in the highly reduced nucleomorphs through strand-specific RNA-Seq on mRNA extracted from both the cryptophyte *G. theta* and the chlorarachniophyte *B. natans*, and conducting an in-depth analysis and comparison of pre-mRNA splicing and antisense transcription. Our data reveal that although there are similarities between the nucleomorphs of *G. theta* and *B. natans* with respect to gene expression and antisense transcription, there are marked differences in proportions of transcripts that remain unspliced. Whereas the intron-sparse nucleomorph of *G. theta* exhibits much intron retention, the many tiny introns of the *B. natans* nucleomorph are spliced at high levels, highlighting contrasting and possibly lineage-specific differences in the evolutionary outcomes of pre-mRNA splicing due to genome reduction. Furthermore, our study provides

insight into the diversity of evolutionary trajectories due to genome reduction, and to the diverse nature of what is considered a ubiquitous and conserved eukaryotic process.

Materials and Methods

Culturing of *G. theta* and *B. natans*

Monoclonal cultures of *G. theta* strain CCMP 2712 and *B. natans* strain CCMP 621 were obtained from the National Center for Microbiota and Algae (NCMA, formerly CCMP). Both organisms were cultured in 250 ml Erlenmeyer flasks using 50 ml of f/2-Si media for *B. natans*, and the same volume of media supplemented with 50 mM of NH_4Cl for *G. theta*, as it is unusual in its requirement of ammonium as its nitrogen source (Hill and Wetherbee 1990). Cultures were agitated on a shaking platform rotating at 120 rpm, and were exposed to 30 μmol photons/m^2/s of light for 12 h per day.

RNA Extraction

Guillardia theta cells were pelleted from 10 ml of culture spun down 6 h, or halfway, into the "daylight" phase of the light cycle. Total RNA was extracted from these pellets using the Ambion RNaqueous Kit (Life Technologies) with the manufacturer's recommended protocol. The same volume of *B. natans* culture was spun down at the same time-point as *G. theta* cultures, and total RNA extracts were prepared from the cell pellets using the TRIzol reagent (Ambion) under the manufacturer's recommended conditions. Eluted total RNA samples were quantified using a NanoDrop spectrophotometer (Thermo Scientific).

RNA Cleanup and Poly-A Purification

Total RNA extracts were cleaned of gross DNA contamination using the Invitrogen DNA-free DNA Removal Kit (Life Technologies) and further quantified for RNA and DNA using Invitrogen Qubit fluorometry (Life Technologies) for both macromolecules. Poly-A purification was performed using NEXTflex Poly(A) Beads (BioO Scientific) to enrich samples for mRNA and reduce the relative proportion of rRNA.

Strand-Specific Library Preparation and Second-Generation Sequencing

Two strand-specific libraries of *G. theta* were prepared as replicates using the NEXTflex Directional RNA-Seq Kit (BioO Scientific), which uses the dUTP method to maintain strand specificity. The libraries were prepared without modification to the manufacturer's protocol. This method employs the addition of deoxyuracil triphosphate (dUTP) in place of deoxythymidine triphosphate (dTTP) during second-strand synthesis of reverse transcription, and subsequent digestion of uracil using uracil-DNA glycosylase (UDG) introduces breaks in the second strand, ensuring strand-specificity of the resultant sequenced fragments (Parkhomchuk et al. 2009; Wang et al. 2011). These two libraries were sequenced on an in-house Illumina HiSeq 2500, generating a total of 17,056,967 and 31,687,031 paired-end reads. The same process was repeated to generate two replicate libraries of *B. natans*, resulting in 65,432,233 and 62,505,229 paired-end reads.

Bioinformatics Analysis of Sequence Data

The resulting sets of reads were mapped using TopHat2 (Kim et al. 2013) to concatenated reference genomes of *G. theta* (GenBank accession AEIE00000000) and *B. natans* (GenBank accession ADNK00000000) that include all nucleomorph chromosomes, plastid and nuclear genomic sequences. Mapped read pairs in SAM format alignment files were then processed with custom Python scripts to sort them into sense or antisense read pairs based on existing gene annotations of *G. theta* and *B. natans*. Raw read counts were summed for each gene for use in downstream calculations of gene expression and antisense transcription levels. These counts were then normalized to determine relative expression levels using the FPKM (fragments per kilobase of exon per million reads mapped) method (Mortazavi et al. 2008).

For each annotated junction in *G. theta* and *B. natans*, further custom Python scripts were employed to enumerate the mapped read pairs in that vicinity for the type of splicing event it represents: spliced transcripts, intron retention and other alternative splicing events. To ensure that the splicing events are real and not represented by spuriously mapped reads, a junction with <25 reads mapped to its vicinity was excluded from further analysis. Using read counts for spliced transcripts and intron-retained transcripts, we calculated the percent spliced reads for each annotated junction by dividing the number of canonically spliced reads by the total number of reads. We also performed a similar calculation for each annotated junction by dividing the number of intron-retained reads by total reads to generate percent intron retention, which totals to 100% when summed with percent spliced reads. Because the length of introns in both the *G. theta* and *B. natans* nucleomorphs are much shorter than read length, our calculated percent intron retention is a proxy for the percent spliced in (PSI) value used in alternative splicing studies. Percent intron retention for nucleomorph junctions are provided as supplementary figures S1 and S2, Supplementary Material online.

A discrepancy exists between the number of introns we analyzed versus the latest number of introns found in the nucleomorph of *B. natans*. Whereas 99 additional introns were presented in a recent analysis of the *B. natans* nucleomorph genome (Tanifuji, Onodera, Brown, et al. 2014), the available annotation files contain only the 852 introns annotated from the original genome sequencing project

(Gilson et al. 2006), and these are the 852 *B. natans* nucleomorph introns analyzed here.

Results and Discussion

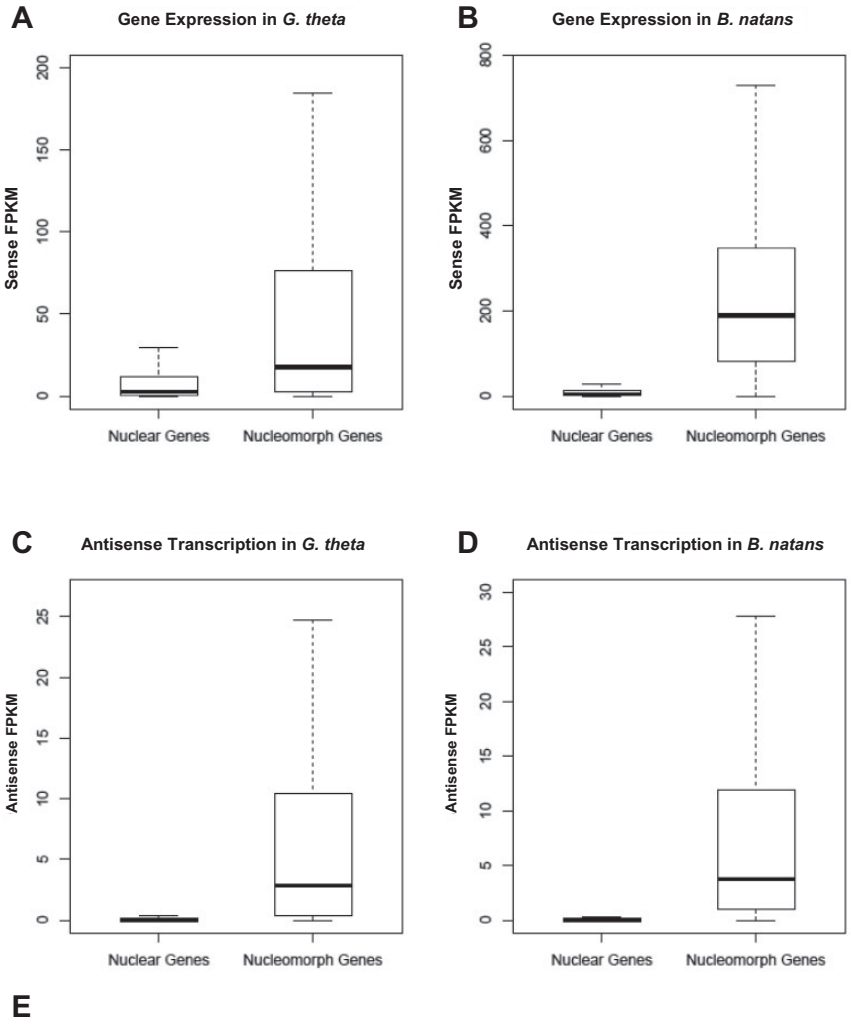
Increased Levels of Transcription in Both Cryptophyte and Chlorarachniophyte Nucleomorphs

The nucleomorphs of cryptophytes and chlorarachniophytes represent evolutionarily convergent structures, and our comparison of pre-mRNA splicing between nucleomorphs of both lineages offers insight into the evolution of a conserved eukaryotic process in two independent cases of genome reduction. To examine the process of pre-mRNA splicing in nucleomorphs, we performed strand-specific RNA-Seq on the cryptophyte *G. theta* and the chlorarachniophyte *B. natans*, and mapped the reads to their respective nucleomorph genomes. Because both *G. theta* and *B. natans* also have sequenced nuclear genomes (Curtis et al. 2012), we simultaneously mapped our RNA-Seq data to their respective nuclei to be used as examples of typical eukaryotic genomes. Using available genome annotations, we totaled the number of mapped reads for each gene representing expression, antisense transcription, intron excision or retention, alternative splicing, and so forth. We determined with these counts that the replicate RNA-Seq libraries were statistically similar (Pearson's $r = 0.95$ for *G. theta*; $r = 0.99$ for *B. natans*), and all mapped reads from each species were pooled for our final transcriptome analyses. In total, 26,561,411 pooled reads were mapped to all *G. theta* genomic sequences, with 1,091,066 of those (4.1%) mapping to the nucleomorph. For *B. natans*, 62,768,143 pooled reads were mapped, and 6,690,653 of those (10.7%) mapped to the nucleomorph.

Previous transcriptomic studies on *G. theta* and *B. natans* have focused on gene expression. In those studies, researchers noted the high gene expression levels in nucleomorphs (Hirakawa et al. 2014; Tanifuji, Onodera, Moore, et al. 2014), and observed that virtually the entire nucleomorph genome is transcribed (Williams et al. 2005; Sanitá Lima and Smith 2017). An assessment of relative gene expression levels was performed to compare our data to those studies to ensure that our data and methodology were sound. We determined relative gene expression levels from our RNA-Seq data by normalizing counts of mapped reads across all (nuclear and nucleomorph) annotated genes in *G. theta* and *B. natans* using FPKM (fragments per kilobase of exon per million reads mapped). Tanifuji, Onodera, Moore, et al. (2014) showed that gene expression in *G. theta* nucleomorph genes was on average 2.6 times higher than in its nuclear genes. The nucleomorph genes of *B. natans* were shown to be expressed almost 15 times higher than its nuclear genes (Tanifuji, Onodera, Moore, et al. 2014). Our RNA-Seq data showed very similar increases in nucleomorph gene expression (fig. 1)—*G. theta* nuclear genes have an average FPKM of 24.9, whereas *G. theta* nucleomorph genes have an ~3-fold higher relative expression with an average FPKM of 75.4. Likewise, an average FPKM of 317 in nucleomorph genes of *B. natans* versus 17.1 in nuclear genes suggest a 19-fold increase in relative gene expression (fig. 1). Taken together with results from *G. theta*, our expression data correspond well with previous studies, and we proceeded to examine further aspects of transcription only allowed by our use of strand-specific methodologies.

Antisense transcription has not been previously analyzed in nucleomorphs, and we used our strand-specific RNA-Seq data to investigate the extent of antisense transcription in both the nucleomorph and nuclear genomes of *G. theta* and *B. natans*. Reduced genomes are thought to exhibit more antisense transcription, as intergenic spaces are small and transcripts of neighboring and oppositely oriented genes are likely to overlap (Williams et al. 2005; Pelechano and Steinmetz 2013). Therefore, we predicted that the nucleomorphs of *G. theta* and *B. natans* would have more antisense transcription taking place than in their respective nuclei. To determine relative antisense transcription levels, FPKM was calculated from mapped antisense reads for all annotated genes. Indeed, more antisense transcription occurs in the nucleomorphs of both *G. theta* and *B. natans* than in their respective nuclei (fig. 1).

The increased levels of antisense transcription in nucleomorphs compared with their host nuclei suggest the provocative possibility that these transcripts could be playing a functional role in gene regulation in the nucleomorph. While Tanifuji, Onodera, Moore, et al. (2014) suggest that increased nucleomorph gene expression could be a compensatory mechanism against high levels of errant antisense transcription, the converse could also be true—antisense transcripts could down-regulate the massively increased levels of gene expression. Given the extremely reduced set of genes within the nucleomorph, the nucleomorph genome could have dispensed with many of the factors necessary for finer transcriptional regulation. Indeed, based on sequence similarity to known transcription-related proteins, the nucleomorph genome of *G. theta* is predicted to harbor only a small complement of 30 such proteins, while the *B. natans* nucleomorph genome has even fewer at 22 (Douglas et al. 2001; Gilson et al. 2006). Although some putative nucleomorph-targeted transcription factors have been identified for both *G. theta* and *B. natans*, the full contribution of nuclear-encoded transcription-associated proteins is unknown (Curtis et al. 2012). Regardless, the extremely short intergenic regions of the *G. theta* and *B. natans* nucleomorph present considerable limitations on the position and sequences of regulatory motifs, especially if those motifs are involved in suppressing transcription, resulting in poorly controlled and uniformly high expression of nucleomorph genes and a possible reliance on antisense transcription as an alternate mechanism of gene regulation. This is supported by a recent study revealing that only two nucleomorph genes of *B. natans* are



E

	Guillardia theta		*Bigelowiella natans*	
	Nuclear genes	**Nucleomorph genes**	**Nuclear genes**	**Nucleomorph genes**
Sense FPKM	Mean: 24.9 Median: 2.93	Mean: 75.4 Median: 17.8	Mean: 17.1 Median: 5.24	Mean: 317 Median: 207
Antisense FPKM	Mean: 1.00 Median: 0.0181	Mean: 11.5 Median: 2.87	Mean: 0.618 Median: 0	Mean: 19.1 Median: 3.88

Fig. 1.—Increased gene expression and antisense transcription is observed in nucleomorph genes of both *Guillardia theta* and *Bigelowiella natans* relative to their respective nuclear genes. (*A*) Box plot representing gene expression level (normalized using the FPKM formula) differences between nuclear and nucleomorph genes of *G. theta*. (*B*) Box plot representing gene expression level differences between nuclear and nucleomorph genes of *B. natans*. (*C*) Box plot representing levels of antisense transcription (normalized using the FPKM formula) differences between nuclear and nucleomorph genes of *G. theta*. (*D*) Box plot representing antisense transcription level differences between nuclear and nucleomorph genes of *B. natans*. (*E*) Summary table of mean and median FPKM and antisense FPKM of both nuclear and nucleomorph genes in *G. theta* and *B. natans*.

differentially expressed between light and dark cycles, suggesting a lack of transcriptional regulation (Suzuki *et al.* 2016). Others have suggested that overexpression of nucleomorph genes could be compensating for missplicing, or increased intron retention, which may be more common in a reduced genome (Grisdale et al. 2013; Tanifuji, Onodera, Moore, et al. 2014). As discussed later, our splicing analyses partially support this idea, raising questions about the

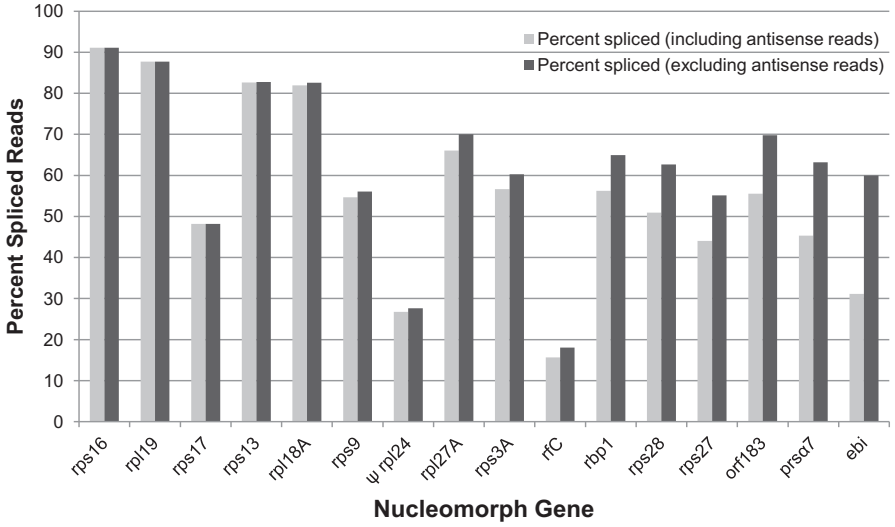


Fig. 2.—Pre-mRNA splicing levels in *Guillardia theta* nucleomorph genes. For each intron in the nucleomorph, the percentage of spliced reads was calculated in two different ways—one including reads mapped to the antisense strand to simulate a traditional nonstranded RNA-Seq (light bars), and one excluding these antisense reads (dark bars). The genes are arranged left to right in order of increasing effect of antisense reads on the calculated percent spliced reads, highlighted by the difference in heights of the bars. Nucleomorph gene *orf183* had <50 reads mapped across its junction, while the junction from *orf263* was excluded from this analysis due to poor coverage.

potential interplay of overexpression, antisense transcription and pre-mRNA splicing within reduced genomes.

Increased Intron Retention in *G. theta* Nucleomorph Genes

Our study provides the first analysis of pre-mRNA splicing in a cryptophyte nucleomorph, and the first comparison of splicing between the independently reduced cryptophyte and chlorarachniophyte nucleomorphs. While our RNA-Seq data showed that high expression is a common feature to both the *G. theta* and *B. natans* nucleomorphs, we observe significant differences in the patterns of pre-mRNA splicing between the two. Only 17 introns have been annotated in the nucleomorph genome of *G. theta*, each one located in a different gene, all with a noticeable bias towards the 5′ end of the gene (Douglas et al. 2001). Previous transcriptomic studies on another reduced genome have shown that relatively few introns remain in the genome, and these introns are often retained in mature transcripts (Grisdale *et al.* 2013). Considering the low intron density and overall level of genome reduction in the nucleomorph of *G. theta*, we expected that high levels of intron retention would also be observed in these 17 introns.

We examined the splicing levels for 16 of the 17 annotated nucleomorph introns—the remaining intron was located on a gene with poor read coverage and did not meet our cut-off criterion. As shown in figure 2, intron retention is prevalent—on average, only around 60% of the reads mapped to the vicinity of introns are spliced. This is significant, as the only other report of unusual intron retention is from an RNA-Seq study on the highly reduced microsporidian *E. cuniculi* where intron retention was extensive (Grisdale et al. 2013). To rule out the possibility that increased intron retention seen in the *G. theta* nucleomorph was reflecting technical issues, we also calculated the percent of spliced reads at *G. theta* nuclear junctions (supplementary fig. S3, Supplementary Material online). As expected for a typical eukaryotic genome, nuclear gene transcripts of *G. theta* are spliced at high levels with little intron retention. In the *G. theta* nucleomorph, the percent of spliced reads varies between the introns, from the intron for the replication factor *rfC* being spliced at <20%, to a number of other genes where intron retention is not extensive and nearly 90% of reads are spliced, such as the ribosomal proteins *rps16*, *rpl19*, and *rps13* (fig. 2). This range of splicing levels might suggest that introns of certain genes are spliced at higher levels than others; however, there does not appear to be any correlation between the extent of intron retention and the function of the gene in which the intron lies.

The low splicing levels seen in the highly reduced *G. theta* nucleomorph are similar to patterns seen in a previous transcriptomic study of the microsporidian parasite *E. cuniculi* (Grisdale et al. 2013), consistent with the hypothesis that one of the effects genome and spliceosome reduction is a reduction in splicing levels of the few remaining introns. It is possible that increased intron retention represents missplicing due to poor recognition of reduced introns (with reduced, divergent, or missing splicing motifs) by the spliceosome. In a study by Jaillon et al. (2008), they found that the short introns (20–34 bp) of the ciliate *Paramecium tetraurelia* possess reduced splicing signals, and suggested with some support that these introns are not removed efficiently from transcripts. However, the 17 annotated introns in the *G. theta* nucleomorph have clearly defined splice sites and a

well-defined branch point adenosine (Douglas et al. 2001), and we did not find any discernible correlations between features of the intron sequence and the propensity for its removal. In fact, the 5′ splice site is well-conserved across all 17 *G. theta* nucleomorph introns (Douglas et al. 2001), in line with other studies showing that genomes with low intron density tend towards stronger, rather than weaker, splicing motifs (Irimia et al. 2007, 2009; Irimia and Roy 2008; Lee et al. 2010). It has also been suggested that an incomplete set of spliceosomal components in reduced genomes result in increased missplicing and intron retention (Grisdale et al. 2013). Although a small number of spliceosomal components and one of the snRNAs have been identified within the nucleomorph genome (Douglas et al. 2001; López *et al.* 2008), the full contribution of spliceosomal components from nuclear encoded genes is unclear (Curtis et al. 2012). Regardless, because our data show that not all of the 17 introns exhibit extensive intron retention, a functional spliceosome must be present. With so few, short introns remaining, one wonders why they have not been dispensed with altogether given the high level of genome reduction. Although it is possible that their positional bias within the gene prevents their loss, it is also possible that they play a regulatory role that contributes to their persistence. For example, retained introns in mature transcripts, especially if splicing levels are actively regulated by the organism, could act as a form of post-transcriptional regulation, down-regulating overly abundant transcripts from being translated (Lewis et al. 2003; Lareau et al. 2007). Another possibility could be interplay between antisense transcription and splicing, as has been previously documented (Morrissy *et al.* 2011; Pelechano and Steinmetz 2013).

Our strand-specific methodology allows us to explore these potential links between antisense transcription and pre-mRNA splicing. We have shown that transcription occurs at high levels in nucleomorphs, in line with previous studies exploring nucleomorph gene expression (Hirakawa et al. 2014; Tanifuji, Onodera, Moore, et al. 2014). However, in those previous studies, there was no way to reliably discern how many of the transcripts were antisense. Conventional methods of reverse transcription do not preserve the strandedness of the original transcript, and the lack of strand specificity of the sequence data renders antisense transcripts indistinguishable from "sense" mRNA to read mapping programs—antisense reads mapping to an annotated intron would thus appear as intron retention. Our strand-specific RNA-Seq data allow us to compare methodologies to determine if this is a legitimate concern, in addition to providing opportunities to assess any potential effects of antisense transcription on pre-mRNA splicing. We were able to make two calculations of percent transcripts spliced: one simulated results from a conventional RNA-Seq experiment by calculating the percentage of spliced transcripts from all sense and antisense reads, and a second calculation excluding the antisense reads. The differences between the calculations of percent spliced transcripts for each junction are represented in figure 2. A number of introns (on the right side of fig. 2), such as those in the genes for ribosomal protein *rps27* or growth factor *ebi*, have high levels of antisense transcripts that would confound calculations of percent spliced transcripts. The intron in *ebi* is most affected; antisense reads comprise nearly half of the mapped reads in the vicinity of its intron. Without strand data for the reads, the percent of spliced reads would be 31% as opposed to 60%, severely overestimating the occurrence of intron retention for this gene. However, most other introns do not appear to have enough antisense reads that would cause an overestimation of the extent of intron retention—the proportion of antisense reads mapping to regions within the intron is $<10\%$ for more than half of the *G. theta* nucleomorph introns. In fact, no antisense reads were mapped to the vicinity of the introns in *rps16*, *rpl19*, or *rps17* (fig. 2). However, because reduced genomes tend to have very few introns, even a few overestimations can skew the overall impression of the extent of intron retention. Thus, it is worthwhile to consider strand-specific libraries for any RNA-Seq experiment for organisms with reduced genomes.

Our transcriptomic data from the *G. theta* nucleomorph revealed that despite high levels of antisense transcription across the entire nucleomorph genome (see above), antisense transcription occurs significantly less frequently (Welch's unequal variances *t*-test, $P < 0.001$) in the 17 intron-containing genes than in all the other single-exon genes. Although there are limitations to drawing robust statistical conclusions from only 17 introns, this observation could suggest that in the *G. theta* nucleomorph, antisense transcripts complementary to multi-exon genes are actively down-regulated to allow for proper splicing to occur. This could highlight a potential link between antisense transcription and pre-mRNA splicing, requiring further investigation. It has previously been suggested that antisense transcripts occlude splicing signals, leading to intron retention, and indeed a positive correlation has been observed in animals between antisense transcription and alternatively spliced genes (Morrissy *et al.* 2011; Pelechano and Steinmetz 2013). Although untangling the links between gene expression, antisense transcription and pre-mRNA splicing in such a reduced genome presents many technical challenges, this interplay within the *G. theta* nucleomorph could not only represent an unusual network of regulatory mechanisms shared with other reduced genomes, but also provide an evolutionary explanation for the persistence of introns, and their associated spliceosome.

Near-typical Eukaryotic Splicing Levels in *B. natans* Nucleomorph Genes

The independently reduced nucleomorph of the chlorarachniophyte *B. natans* bears remarkable contrasts to that of the cryptophyte *G. theta*. To investigate pre-mRNA splicing patterns in *B. natans* further, we calculated the percent of spliced

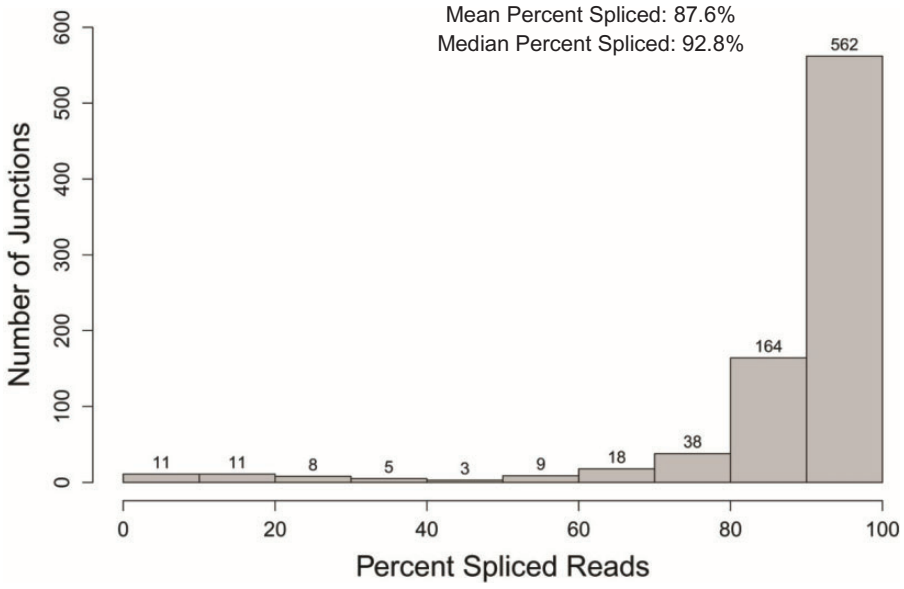


Fig. 3.—*Bigelowiella natans* nucleomorph introns are spliced at high levels. Histogram showing the proportion of junctions at different levels of percent spliced reads. The vast majority of junctions has >80% spliced reads.

reads at all annotated introns of both the nuclear and nucleomorph genomes of *B. natans* using the same methods and criteria as *G. theta*. As with *G. theta* nuclear genes, reads mapped to the *B. natans* nucleus are spliced with a median percent spliced reads higher than 90%, indicating little intron retention (supplementary fig. S4, Supplementary Material online).

With only 283 annotated genes and nearly 900 introns (see Materials and Methods for clarification) in the *B. natans* nucleomorph genome, the genes are densely populated with extremely short (18–21 bp) introns (Gilson et al. 2006; Tanifuji, Onodera, Brown, et al. 2014). The average intron density is 3.1 introns per gene, and only 43 of the 283 annotated genes lack introns (Gilson et al. 2006). However, based on our current understanding of conserved mechanisms of splicing, these short intron lengths place constraints on this process. The introns of *B. natans* nucleomorph genes do not have a discernible branch point adenosine (Gilson et al. 2006), and considering the intron lengths, it is difficult to imagine the formation of lariat structures typical of spliceosomal intron removal. Also, aside from the GU-AG boundaries, no other motifs are apparent (Gilson et al. 2006). Finally, the typical eukaryotic spliceosome (or even a reduced version) is a very large conglomerate of proteins that likely dwarfs these tiny introns. On the basis of these unusual aspects, and extensive intron retention in other reduced genomes, we expected to observe low levels of splicing in the *B. natans* nucleomorph.

Using our minimum read coverage criterion as described in the Materials and Methods, we analyzed splicing levels of 829 introns from the *B. natans* nucleomorph. Surprisingly, the calculated splicing levels of *B. natans* nucleomorph introns follow a pattern more like its respective nucleus than other reduced genomes (fig. 3). More than half of all annotated nucleomorph introns have 90–100% of junction-mapped reads spliced, and only 12% of all annotated introns have <80% of junction-mapped reads spliced. Furthermore, for the vast majority of these introns, antisense reads comprise <10% of the mapped reads, and do not contribute to any significant overestimation of intron retention (supplementary File S5, Supplementary Material online). As with the *G. theta* nucleomorph, we were unable to correlate splicing levels of a particular gene's introns and that gene's proposed function. Furthermore, given the high levels of splicing across all introns, it is less clear if these tiny introns could play a functional role. However, the possibility remains that under certain conditions, such as stress, nucleomorph splicing could be regulated.

This "proficiency" in pre-mRNA splicing we observed means that the relatively high nucleomorph gene expression in *B. natans* is not compensating for excessive missplicing, as suggested by Tanifuji, Onodera, Moore, et al. (2014). Instead, near-typical eukaryotic splicing levels in the highly reduced *B. natans* nucleomorph genome could be a consequence of having a relatively high intron density similar to other free-living green algae (Slamovits and Keeling 2009), where extensive intron retention would likely be deleterious. While there are 24 predicted spliceosomal components in the *B. natans* nucleomorph (Gilson et al. 2006; Curtis et al. 2012), this is a small fraction of the number of components required to form the familiar spliceosome that is conserved amongst other eukaryotes, even with contributions from nuclear-encoded products. Taking our data together with the unusual sequence features of their introns suggest that the *B. natans* nucleomorph must use a novel or highly divergent mechanism to effectively remove its extremely short introns with great

accuracy. It has been suggested that the length and sequence of the *B. natans* nucleomorph introns may play a role in efficient identification and removal, as the range of lengths is very narrow (18–21 bp), and intronic sequences have a marked AU bias (Gilson et al. 2006). Whereas an EST-based study of the chlorarachniophyte *Gymnochlora stellata* suggested that splicing levels of nucleomorph introns are correlated with their lengths (Slamovits and Keeling 2009), we find no significant differences in splicing levels across *B. natans* nucleomorph introns (supplementary File S5, Supplementary Material online). Furthermore, the only identifiable splicing sequence signals in *B. natans* nucleomorph introns (the 5′ and 3′ splice sites) are essentially invariable (Gilson et al. 2006). Indeed, Gilson et al. (2006) propose that the *B. natans* nucleomorph spliceosome could operate as a molecular "caliper," excising 18–21 base-pair-long AU-rich regions bound by canonical splice sites. Having very typical eukaryotic splicing levels in a highly reduced genome with such short introns, pre-mRNA splicing in the *B. natans* nucleomorph merits further biochemical and proteomic studies to elucidate this process and allow comparison with canonical splicing.

Evolution of Pre-mRNA Splicing in Reduced Eukaryotes

Our analyses of splicing in the nucleomorphs of *G. theta* and *B. natans* highlight major differences in the patterns of pre-mRNA splicing in highly reduced genomes. Although the trend of increased intron retention in reduced genomes was seen in the highly reduced (but intron-sparse) nucleomorph of *G. theta*, transcripts from the even tinier (but intron-dense) *B. natans* nucleomorph were spliced at levels seen in most other eukaryotes. This stark contrast between the two nucleomorphs could simply indicate that splicing levels are influenced by intron density. However, the difference in splicing levels could also reflect the evolutionary ancestries of the secondary plastids. These two possibilities are not mutually exclusive, and their resolution would be useful for generalizing the patterns of pre-mRNA splicing across cryptophyte and chlorarachniophyte nucleomorphs and other reduced eukaryotic genomes. As discussed previously, it would be deleterious for an intron-dense organism to exhibit extensive intron retention, even under the evolutionary pressures of genome reduction. This is supported not only by our splicing analysis of the *B. natans* nucleomorph, but also by existing EST analyses on nucleomorph transcripts from another chlorarachniophyte *G. stellata* (Slamovits and Keeling 2009). In that study, a number of *G. stellata* nucleomorph genes were found to have similar densities of very short introns as the *B. natans* nucleomorph, and most of those introns were removed in 80–100% of transcripts (Slamovits and Keeling 2009). However, there are genomes where intron density is rather low, yet intron retention is not widespread. The most notable of these is in the budding yeast *Saccharomyces cerevisiae*, whose genome is relatively small, and from where most research on the biochemistry of pre-mRNA splicing has been conducted. No reports exist of widespread intron retention in *S. cerevisiae*, and Grisdale et al. (2013) showed in a splicing analysis using existing *S. cerevisiae* RNA-Seq data that the vast majority of the 253 introns is spliced at high levels. Thus, while intron density could be a very strong determinant of intron retention in reduced genomes, other biological or evolutionary factors are also involved.

Annotations from other sequenced nucleomorph genomes show that all cryptophyte nucleomorphs are intron-sparse, while chlorarachniophyte nucleomorphs have very many tiny introns (Lane et al. 2007; Tanifuji et al. 2011; Moore et al. 2012; Tanifuji, Onodera, Brown, et al. 2014; Suzuki et al. 2015). The chlorarachniophyte plastid was derived from a green alga, and its nucleomorph's intron density is similar to those of *Arabidopsis thaliana* and *Chlamydomonas reinhardtii* (Gilson et al. 2006; Slamovits and Keeling 2009). There are also clear indications that some of the annotated chlorarachniophyte introns bear positional homology to those in extant green algae (Gilson et al. 2006; Slamovits and Keeling 2009). The cryptophyte plastid, on the other hand, is derived from a red alga. Current genomic data repeatedly suggest that free-living red algae are generally gene- and intron-poor, hinting at a possible ancient genome reduction event before the radiation of rhodophytes (Qiu et al. 2015). Should this be true, the red algal endosymbiont ancestor of the cryptophyte plastid would have already been reduced with respect to intron density and the number of spliceosomal components. Consequently, extant red algae may also exhibit pre-mRNA splicing patterns similar to what we observed in the *G. theta* nucleomorph. Indeed, in the extremophilic red alga *C. merolae*, extensive intron retention is observed in its 27 annotated introns (Grisdale CJ, unpublished data), and only a relatively small complement of spliceosomal components remain: the U1 snRNP, the subunit of the spliceosome responsible for the recognition of the 5′ splice site, is entirely missing from the genome (Matsuzaki et al. 2004; Stark et al. 2015). However, *C. merolae* could represent a highly derived lineage of red algae, and pre-mRNA splicing remains to be studied in detail in mesophilic rhodophytes. Further sampling of red algae and cryptophyte nucleomorphs will be required to determine if evolutionary ancestry is responsible for the differences in pre-mRNA splicing in the highly reduced nucleomorphs of cryptophytes and chlorarachniophytes.

Conclusions

The nucleomorphs of cryptophytes and chlorarachniophytes provide unique opportunities to compare the evolution and diversity of conserved eukaryotic processes within the context of genome reduction. Our data reveal similar patterns of high gene expression and high antisense transcription in both nucleomorphs. We also observed differences in levels of

antisense transcription around junctions in *G. theta*, suggesting potential links between antisense transcription and pre-mRNA splicing. The marked differences observed in pre-mRNA splicing between the nucleomorphs highlights the diversity of what is considered to be a conserved process across eukaryotes, and raises awareness of the value of investigating splicing in the lesser-studied branches of the eukaryotic tree. Further investigations of the nature and mechanisms of pre-mRNA splicing in reduced genomes will provide valuable insight and improve our understanding of this key eukaryotic process.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

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

Literature Cited

Cuomo CA, et al. 2012. Microsporidian genome analysis reveals evolutionary strategies for obligate intracellular growth. Genome Res. 22(12):2478–2488.

Curtis BA, et al. 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature 492(7427):59–65.

Douglas S, et al. 2001. The highly reduced genome of an enslaved algal nucleus. Nature 410(6832):1091–1096.

Gilson PR, et al. 2006. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. Proc Natl Acad Sci USA. 103(25):9566–9571.

Grisdale CJ, Bowers LC, Didier ES, Fast NM. 2013. Transcriptome analysis of the parasite *Encephalitozoon cuniculi*: an in-depth examination of the pre-mRNA splicing in a reduced eukaryote. BMC Genomics. 14(1):207–215.

Ghildiyal M, Zamore PD. 2009. Small silencing RNAs: an expanding universe. Nat Rev Genet. 10(2):94–108.

Hill DRA, Wetherbee R. 1990. *Guillardia theta* gen. et sp.nov. (Cryptophyceae). Can J Bot. 68(9):1873–1876.

Hirakawa Y, Burki F, Keeling PJ. 2011. Nucleus- and nucleomorph-targeted histone proteins in a chlorarachniophyte alga. Mol Microbiol. 80(6):1439–1449.

Hirakawa Y, Suzuki S, Archibald JM, Keeling PJ, Ishida K-I. 2014. Overexpression of molecular chaperone genes in nucleomorph genomes. Mol Biol Evol. 31(6):1437–1443.

Irimia M, Penny D, Roy SW. 2007. Coevolution of genomic intron number and splice sites. Trends Genet. 23(7):321–325.

Irimia M, Roy SW. 2008. Evolutionary convergence on highly-conserved 3′ intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. PLoS Genet. 4(8):e1000148.

Irimia M, et al. 2009. Complex selection on 5′ splice sites in intron-rich organisms. Genome Res. 19(11):2021–2027.

Jaillon O, et al. 2008. Translational control of intron splicing in eukaryotes. Nature 451(7176):359–362.

Katinka MD, et al. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature 414(6862):450–453.

Keeling PJ. 2004. Diversity and evolutionary history of plastids and their hosts. Am J Bot. 91(10):1481–1493.

Kim D, et al. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14(4):R36.

Lane CE, et al. 2007. Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function. Proc Natl Acad Sci USA. 104(50):19908–19913.

Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. 2007. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature 446(7138):926–929.

Lee RCH, Gill EE, Roy SW, Fast NM. 2010. Constrained intron structures in a microsporidian. Mol Biol Evol. 27(9):1979–1982.

Lewis BP, Green RE, Brenner SE. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci USA. 100(1):189–192.

López MD, Alm Rosenblad M, Samuelsson T. 2008. Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. Nucleic Acids Res. 36(9):3001–3010.

Matsuzaki M, et al. 2004. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature 428(6983):653–657.

Moazed D. 2009. Small RNAs in transcriptional gene silencing and genome defence. Nature 457(7228):413–420.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 5(7):621–628.

Moore CE, Curtis B, Mills T, Tanifuji G, Archibald JM. 2012. Nucleomorph genome sequence of the cryptophyte alga *Chroomonas mesostigmatica* CCMP1168 reveals lineage-specific gene loss and genome complexity. Genome Biol Evol. 4(11):1162–1175.

Morrissy AS, Griffith M, Marra MA. 2011. Extensive relationship between antisense transcription and alternative splicing in the human genome. Genome Res. 21(8):1203–1212.

Parkhomchuk D, et al. 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. 37(18):e123.

Pelechano V, Steinmetz LM. 2013. Gene regulation by antisense transcription. Nat Rev Genet. 14(12):880–893.

Qiu H, Price DC, Yang EC, Yoon HS, Bhattacharya D. 2015. Evidence of ancient genome reduction in red algae (*Rhodophyta*). J Phycol. 51(4):624–636.

Sanitá Lima M, Smith DR. 2017. Pervasive transcription of mitochondrial, plastid, and nucleomorph genomes across diverse plastid-bearing species. Genome Biol Evol. 9(10):2650–2657.

Slamovits CH, Keeling PJ. 2009. Evolution of ultrasmall spliceosomal introns in highly reduced nuclear genomes. Mol Biol Evol. 26(8):1699–1705.

Stark MR, et al. 2015. Dramatically reduced spliceosome in *Cyanidioschyzon merolae*. Proc Natl Acad Sci USA. 112(11):E1191–E1200.

Suzuki S, Shirato S, Hirakawa Y, Ishida K-I. 2015. Nucleomorph genome sequences of two chlorarachniophytes, *Amorphochlora amoebiformis* and *Lotharella vacuolata*. Genome Biol Evol. 7(6):1533–1545.

Suzuki S, Ishida K-I, Hirakawa Y. 2016. Diurnal transcriptional regulation of endosymbiotically derived genes in the chlorarachniophyte *Bigelowiella natans*. Genome Biol Evol. 8(9):2672–2682.
Tanifuji G, et al. 2011. Complete nucleomorph genome sequence of the nonphotosynthetic alga *Cryptomonas paramecium* reveals a core nucleomorph gene set. Genome Biol Evol. 3:44–54.
Tanifuji G, Onodera NT, Moore CE, et al. 2014. Reduced nuclear genomes maintain high gene transcription levels. Mol Biol Evol. 31(3):625–635.
Tanifuji G, Onodera NT, Brown MW, et al. 2014. Nucleomorph and plastid genome sequences of the chlorarachniophyte *Lotharella oceanica*: convergent reductive evolution and frequent recombination in nucleomorph-bearing algae. BMC Genomics 15(1):374.
Vanhee-Brossollet C, Vaquero C. 1998. Do natural antisense transcripts make sense in eukaryotes? Gene 211(1):1–9.
Wagner EG, Simons RW. 1994. Antisense RNA control in bacteria, phages, and plasmids. Annu Rev Microbiol. 48:713–742.
Wang L, et al. 2011. A low-cost library construction protocol and data analysis pipeline for Illumina-based strand-specific multiplex RNA-seq. PLoS One 6(10):e26426.
Will CL, Lührmann R, et al. 2011. Spliceosome structure and function. Acc Chem Res. 44(12):1292–1301.
Williams BA, Slamovits CH, Patron NJ, Fast NM, Keeling PJ. 2005. A high frequency of overlapping gene expression in compacted eukaryotic genomes. Proc Natl Acad Sci USA. 102(31): 10936–10941.

Associate editor: Michelle Meyer