Evolution of Ultrasmall Spliceosomal Introns in Highly Reduced Nuclear Genomes

Claudio H. Slamovits and Patrick J. Keeling

Department of Botany, University of British Columbia, Vancouver, BC, Canada

Intron reduction and loss is a significant component of genome compaction in many eukaryotic lineages, including yeasts, microsporidia, and some nucleomorphs. Nucleomorphs are the extremely reduced relicts of algal endosymbiont nuclei found in two lineages, cryptomonads and chlorarachniophytes. In cryptomonads, introns are rare or even lost altogether. In contrast, the nucleomorph of the chlorarachniophyte *Bigelowiella natans* contains the smallest nuclear genome known but paradoxically also retained over 800 tiny spliceosomal introns, ranging from 18 to 21 nt in length. Because introns have not been described in any other chlorarachniophyte nucleomorph, we do not know when these introns were reduced or whether they have been lost in other lineages. To gain insight into the evolution of these unique introns, we sequenced more than 150 spliceosomal introns in the nucleomorph of the chlorarachniophyte *Gymnochlora stellata* and compared size distribution, sequence features, and patterns of gain/loss. To clarify the possible relationship between intron size and splicing efficiency, we also analyzed the outcome of 580 splicing events. Overall, these data indicate that the radical intron size reduction took place in the ancestor of all extant chlorarachniophytes and that although most introns have been retained through this reductive process, intron loss has also occurred. We also show that intron size is not static, and splicing is not determined strictly by size, but that size does play a strong role in splicing efficiency, likely as part of a combination of sequence features and size.

Introduction

Spliceosomal introns are, with two exceptions, present in every known eukaryotic nuclear genome, but the density of introns in a genome and the nature of those introns both vary widely. Humans have an average of eight introns per gene, whereas in *Saccharomyces*, 95% of the genes are completely intronless (Bon et al. 2003). Similarly, human introns are quite large on average, as opposed to yeast introns that are only 264 on average (Bon et al. 2003). Indeed, genomes with low intron densities also tend to have small introns and tend to be reduced in other ways as well, as seen in the generally streamlined genomes such as that of yeast, and the more extremely compacted genomes such as those of microsporidia (Katinka et al. 2001).

Another kind of extremely compacted eukaryotic genome is found in the nucleomorphs, but here, the evolution of introns has taken two very different tracks (Douglas et al. 2001; Gilson and McFadden 2002; Archibald 2007). Nucleomorphs are the relict nuclei of endosymbiotic algae that have formed by the process of secondary endosymbiosis that led to the acquisition of plastids in many groups of algae (for review see Keeling 2004). They are only found in two groups of unicellular algae today, cryptomonads and chlorarachniophytes, where they are known to have originated independently from red and green algae, respectively (Keeling 2004). In both groups, the nucleomorph genome is highly reduced, and the overall structure of their genomes share a lot in common, which is thought to be due to convergence given that they evolved from distinct lineages of nonreduced algae. The reduction in chlorarachniophytes is generally more pronounced: Their genomes are smaller on average and the smallest nuclear genome known is that of the recently sequenced *Bigelowiella natans* nucleomorph (Gilson et al. 2006). This 373-kb genome consists of three linear chromosomes possessing only 285 genes in high-density chromosomal cores flanked by subtelomeric ribosomal RNA genes. Many overall characteristics of this genome are also found in the nucleomorph of cryptophytes, suggesting that severe reduction and compaction have converged in strikingly similar genomic organization (Gilson and McFadden 2002; Archibald 2007). Indeed, the major difference between the reduction of these two genomes is in how their introns have evolved. In the *G. theta* genome, there is nothing noteworthy about the introns themselves, but with only 18 known, they are very few in number (Douglas et al. 2001), and in the related species *Hemiselmis andersonii*, introns and splicing machinery have been completely eliminated (Lane et al. 2007), altogether suggesting that intron reduction by loss has been a theme in this lineage. This trend is also seen in the hyper-compacted genomes of microsporidia (Keeling and Slamovits 2005). As in cryptomonad nucleomorphs, microsporidian introns are not extremely small, but are rare in general (Katinka et al. 2001), and have perhaps even been completely lost in one species (Akiyoshi et al. 2009). By contrast, the nucleomorph of the chlorarachniophyte *B. natans* contains over 800 introns, resulting in an intron density comparable to that of plants and other intron-rich eukaryotes (Gilson et al. 2006). But these introns have been severely reduced in size. The majority of them are 19 bp, whereas a few are 18 or 20 bp and two are 21 bp. These are the smallest introns known, and it has been suggested that the remarkably limited range reflects a functional limit for spliceosomal recognition (Cavalier-Smith 2006).

Overall, therefore, *B. natans* nucleomorph introns are not only remarkable for their small size but also for their perseverance: It has even been estimated that chlorarachniophyte nucleomorphs lose introns more slowly than plants and green algae (Roy and Penny 2007). This unusual situation raises questions about what drove these introns to be severely reduced in size, but not be eliminated as it happened in other hyper-compact nuclear genomes? One possibility is that they have been "trapped" in the genome and shrunk because they cannot be lost (Gilson et al. 2006). This may or may not explain the overall reduced size of these introns, but there is also the issue of their reduced

Key words: splicing, introns, endosymbiosis, genome evolution.

E-mail: pkeeling@interchange.ubc.ca.

Mol. Biol. Evol. 26(8):1699–1705. 2009
doi:10.1093/molbev/msp081
Advance Access publication April 20, 2009

range of sizes. Here, it has been suggested that the size range is a key to how the spliceosome recognizes these tiny introns: A spliceosomal "caliper" that measures distances between GT–AG splice sites would be consistent with a narrow distribution of sizes (Gilson et al. 2006).

Our ability to address these questions is restricted by the fact that nucleomorph introns have only ever been sequenced from a single chlorarachniophyte species, *B. natans*. The size and density of introns in other chlorarachniophyte nucleomorphs are completely unknown: Indeed, it is unknown if other chlorarachniophytes even have the same tiny introns, which is obviously an important consideration that affects all previous hypotheses about their origin and function. To provide a useful comparison to *B. natans* and reveal trends in intron gain and loss, and size reduction in chlorarachniophytes in general, we have carried out an expressed sequence tag (EST) project on the chlorarachniophyte *Gymnochlora stellata*, which is distantly related to *B. natans* within the chlorarachniophyte lineage (but still evolved from a common ancestor with a nucleomorph) and has a slightly larger nucleomorph genome (Ishida 1999; Silver et al. 2007). We identified all nucleomorph-derived cDNAs and characterized the genomic copy of 57 genes, yielding over 150 *G. stellata* nucleomorph introns. Overall, the *G. stellata* introns are found to be reduced to a similar size as those of *B. natans*. Because the highly reduced nucleomorph genome was already a feature of the common ancestor of these species, we take this to conclude that nucleomorph intron reduction must also have occurred early in the chlorarachniophyte lineages. We also found that the distribution of intron sizes is similar between *G. stellata* and *B. natans*, but that there is no correlation in the size of any particular intron, further supporting the notion that there are functional constraints on intron size, specifically favoring 19-bp introns. In addition, comparing all intron positions shows that some ancient introns have been lost since the reduction in intron size and that introns larger than the canonical size range can be recognized by the *G. stellata* nucleomorph spliceosome.

Materials and Methods

Strains, Cultivation, and EST Library Construction

Gymnochlora stellata strain CCMP 2057 was cultivated in F/2-Si medium. Approximately 8 l of culture was harvested by centrifugation, and total RNA was prepared in 20-ml batches with Trizol (Invitrogen, Eugene, OR) according to the manufacturer's directions, resulting in 1 μg of RNA. A directional cDNA library was constructed in pBluescript II SK using *Eco*R1 and *Xho*I sites (Amplicon Express, Pullman, WA). Seven thousand two hundred sixty-six clones were picked and 5′-end sequenced, resulting in 6,526 EST sequences that assembled into 3,138 unique clusters using tbESTdb (O'Brien et al. 2007).

Identification and Sequencing of Nucleomorph Genes

We used netblast (http://www.ncbi.nlm.nih.gov/BLAST/download.shtml) to compare the 3,138 clusters assembled by tbESTdb (O'Brien et al. 2007) against the nonredundant NCBI database. Candidate nucleomorph sequences were identified by closest sequence similarity to nucleomorph genes from *B. natans*. Potential nucleomorph genes absent from the *B. natans* nucleomorph were identified by measuring AT content and/or by looking for ultrashort introns.

Analysis of Intron Positions and Sequence

Of all nucleomorph genes identified, only those corresponding to intron-containing genes in *B. natans* were analyzed further. Specific primers were designed to polymerase chain reaction (PCR) amplify potential intron-containing genes from *G. stellata* genomic DNA. Primer positions were chosen to maximize the number of introns recovered. PCR products were cloned into pCR-TOPO vectors (Invitrogen) and sequenced. DNA and EST sequences were aligned manually using Sequencher (Genecodes, Ann Arbor, MI), and intron sequences and positions were recorded. Intron positions and phases were determined and compared with homologous genes in *B. natans* using Artemis 2(http://www.sanger.ac.uk/Software/Artemis/). Presence and position of introns in *Arabidopsis thaliana* and *Chlamydomonas reinhardtii* were investigated using Blast to NCBI databases and the *Chlamydomonas* genome project at Joint Genome Institute (http://genome.jgi-psf.org).

Results and Discussion

Gymnochlora stellata Nucleomorph Genes

We sequenced 7,266 ESTs from *G. stellata* CCMP 2057, resulting in 1,421,101 bp of sequence assembling into 3,138 unique clusters. To identify putative nucleomorph genes, we used BlastX against NCBI nr database and selected sequences with the highest similarity to *B. natans* nucleomorph–encoded proteins. An initial survey generated a list of about 70 candidate genes. *Bigelowiella natans* nucleomorph genes also have a characteristically high AT content (75%), and this feature was also found to be shared by all 70 candidate *G. stellata* nucleomorph genes. Accordingly, we also ranked all clusters by AT content and examined those with the highest proportion of AT, thereby identifying three genes that are apparently nucleomorph encoded in *G. stellata* but not *B. natans*. Two of these are unidentified open reading frame (ORF)s, whereas the third is similar to RPL21, which is absent from the *B. natans* nucleomorph genome (see supplementary table S1, Supplementary Material online). From this pool of cDNAs, those that correspond to intron-containing genes in *B. natans*, plus the three new genes (54 genes in supplementary table S1, Supplementary Material online) were chosen for characterization at the DNA level.

Gymnochlora stellata Nucleomorph Introns

Introns were identified by comparing homologous genomic and cDNA sequences, and in some cases, introns were also observed directly in EST clusters where mRNAs read through a gene into a downstream gene in the opposite

Table 1
Overall Characteristics of *Bigelowiella natans* and *Gymnochlora stellata* Nucleomorph Introns in 54 Homologous Genes

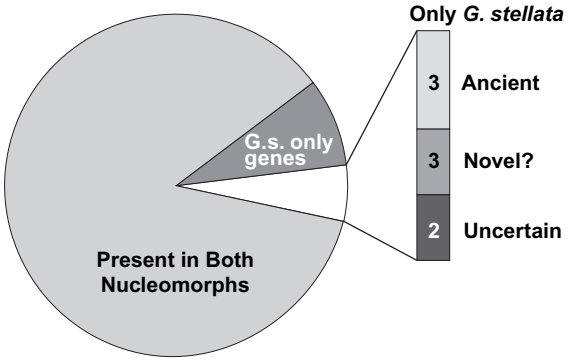
Intron Size	*B. natans* Number (%)	*G. stelata* Number (%)
18	15 (11.7)	2 (1.4)
19	96 (75.0)	112 (78.3)
20	15 (11.7)	23 (16.1)
21	2 (1.6)	5 (3.5)
24	0	1 (0.7)
	128	143
Intron content	2,436 bp	2,753 bp
Genome size	373 kbp	385 kbp

strand, a situation also known from *B. natans* (Williams et al. 2005). Overall, 153 introns were identified (supplementary table S1, Supplementary Material online). Of these, six were found in the three genes with no similarity to *B. natans* nucleomorph sequences (two unidentified ORFs and *rpl21*). Six additional introns were found in antisense strands of genes (on messages of adjacent genes), and another four appear to be the products of spurious splicing of coding sequence as their removal disrupt otherwise meaningful ORFs. The six introns spliced from untranslated regions (UTRs) likely have no functional significance and are simply processed because they contain sequences resembling functional splicing signals. Conversely, the spurious removal of nonintronic sequences from bona fide coding regions would presumably have a deleterious effect but evidently occur at a low enough level to be tolerated in this system. This is an interesting difference between *G. stellata* and *B. natans*, where the splicing of such "pseudointrons" was sought in exon sequences with the greatest overall similarity to the main characteristics of introns, but no evidence of spurious splicing could be found (Gilson et al. 2006). All of the spuriously spliced "introns" in *G. stellata* are peculiar in one way or another. One is 27 nt, the longest recorded splicing event in either of the species (see below), and except for the length, it shows a typical intronic composition. Interestingly, the sequence contains two internal AG dinucleotides 20 and 24 nt from the GT, respectively. We did not find any case where these borders are used, but our sample is too small to confirm that mis-splicing might also involve these sites. The other three sense spurious introns are also interesting. All three are 20 nt in length, and all occur consecutively in the *dbp1* gene, separated by 88 and 29 nt (within the size range of most exons in the nucleomorph). The first two are removed at a relatively low frequency (once and twice of six transcripts, respectively) but the third one is spliced in four of six transcripts. Considered in combination, only a single transcript of six sampled appears to be intact. This raises the question, why three of the four spuriously spliced exon sequences were found within just over 100 bases of the same gene? It suggests that mis-splicing of exons may be rare overall, as it is in *B. natans*, but that this gene is unusual, and perhaps a pseudogene.

As in *B. natans*, spliceosomal introns in the nucleomorph of *G. stellata* are exceptionally short and exhibit a restricted size range, both of which are uniquely extreme in these two species. A number of properties of nucleomorphs of cryptomonads and chlorarachniophytes have been shown to have arisen convergently, in particular overall genome structure features such as chromosome number, and telomeric rRNA repeats, suggesting that in cases of extreme reduction, the same solution can be found independently in different lineages. However, these characters are thought to be convergent because there is strong evidence that the cryptomonad and chlorarachniophyte nucleomorphs do not share a common ancestry. In contrast, there is no question that *G. stellata* and *B. natans* are both members of the chlorarachniophyte lineage and that their common ancestor already had a reduced nucleomorph genome, so it seems highly unlikely that the shared characteristics of their introns arise from convergence. Instead, both the reduced size and highly restricted range of sizes (i.e., 18–21 bp) of introns in the nucleomorph lineage appear to have already been established prior to the divergence of the major chlorarachniophyte lineages. The recently sequenced *C. reinhardtii* genome shows an average intron size of 373 bp (Merchant et al. 2007). Molecular evidence indicates that the endosymbiont of chlorarachniophytes arose from a green alga belonging to the same group (the UTC green algae) as *Chlamydomonas* (Rogers et al. 2007). Assuming that the green alga that gave rise to the plastid had introns more or less like those of *Chlamydomonas*, the process of intron reduction in the chlorarachniophyte nucleomorph was strong enough to achieve a 20-fold reduction, perhaps in a relatively short time.

However, there are also subtle differences in both the range and distribution of intron sizes between the two nucleomorph genomes for homologous intron positions. *Gymnochlora stellata* introns are slightly biased toward larger sizes (table 1): There are fewer 18-bp introns and the 19–21-bp classes are proportionally more abundant than in *B. natans*. Furthermore, two *G. stellata* introns lie outside the known range of *B. natans*. A 24-bp intron that was found in the *tfIIg* gene and the 27-bp sequence discussed above that was spuriously spliced in some transcripts of the *rpb6* gene (see also below for more on this sequence). Because our sample accounts for about 15% of the expected number of introns in the genome, it is likely that additional introns larger than 21 bp were not sampled, but from the existing data, it is clear that intron size in *G. stellata* is less constrained than in *B. natans* where 21 bp is the largest intron, and even that size is only found in two cases of over 800 introns (Gilson et al. 2006).

Interestingly, if the sizes of introns at homologous positions are compared between *B. natans* and *G. stellata*, no correlation (or even tendency to short vs. long) is observed (supplementary table S1, Supplementary Material online). Because homologous introns do not tend to be of the same size but the distribution of sizes is similar (i.e., favoring 19-bp introns), the size distribution most likely reflects some functional constraints on intron sizes that are common to both genomes. Likewise, there is no detectable similarity between homologous introns at the nucleotide sequence level, except for the conserved borders (data not shown), which is expected given that introns generally evolved very rapidly and in these genomes are also highly biased to AT nucleotides.



Gene/Intron	G. stellata	B. natans	A. thaliana	C. reinhardtii
RPS27a/1b	+	-	-	-
RPC1/17b	+	-	+	+
RPS11/1b	+	-	-	+
RPS3a/1b	+	-	+ (1)	+
MYB/3b	+	-	-	? (3)
TFIID/5b	+	-	? (2)	? (4)
RPS23*	+	-	-	-
SMD3*	+	-	-	-

FIG. 1.—Proportion of shared and unique *Gymnochlora stellata* introns. (*A*) *Gymnochlora stellata* introns that have homologs in *Bigelowiella natans* are shown in light gray. Dark gray slice corresponds to introns occurring in genes not present in the *B. natans* nucleomorph. The white slice corresponds to *G. stellata* introns that are absent from homolog positions in *B. natans*. Of these, three are deduced to be ancient introns (those from the RPC1, RPS11, and RPS3 genes), three are potentially novel introns (although they may also be ancient introns retained in *G. stellata* but lost in *B. natans*, *Arabidopsis thaliana*, and *Chlamydomonas reinhardtii*) and two are uncertain as they cannot be accurately aligned. (*B*) Distribution of *G. stellata* introns absent from *B. natans* in a green alga (*C. reinhardtii*) and a land plant (*A. thaliana*). Asterisk: introns in the 3′ UTR region. 1) Present in *Oryza sativa* (but absent from *A. thaliana*). 2) An intron is present 7 nt away. 3) An intron is present 18 nt away. 4) Unalignable region.

Intron Gain and Loss in Nucleomorph Evolution

Because *G. stellata* and *B. natans* represent the deepest split in known chlorarachniophyte diversity (Ishida 1999; Silver et al. 2007), comparing the two is appropriate to assess intron gain, loss, and conservation throughout the whole of chlorarachniophyte evolution. Of the 137 introns that are comparable between *B. natans* and *G. stellata* (i.e., not the six introns that are located in sequences with no *B. natans* counterpart, nor the four spuriously spliced exons we found), 129 (94%) were found to be unambiguously shared between both species (fig. 1). We define this as the two introns being found in the same phase of the same codon in a region of the protein that is clearly alignable. Overall, therefore, the level of intron conservation is very high and consistent with the notion that introns may be difficult to eliminate from these genomes.

However, we also identified several clear cases of intron loss, predominantly in the *B. natans* lineage. Intron loss was inferred by identifying ancient introns in one lineage but not the other. Ancient introns were defined as those where homologous introns were also found in representative genomes of plants and green algae (*A. thaliana* and *C. reinhardtii*, respectively). This indicates that the intron was in the ancestor of nucleomorphs and therefore provides a reasonable indication of loss when an ancient intron is found in only one of the nucleomorph genomes. Three of the eight *G. stellata* introns that are absent from *B. natans* have homologs in *A. thaliana*, *C. reinhardtii*, or both, indicating that these are ancient introns that were present in the ancestor of the green algae and nucleomorphs, but lost in *B. natans* since its divergence from *G. stellata* (fig. 1). Two other cases (*myb* and *tfIIg2* in fig. 1) are uncertain because they are located in regions of insufficient sequence conservation to confidently compare nucleomorphs, *A. thaliana* and *C. reinhardtii*. Three additional *G. stellata* introns that are absent from *B. natans* have no identifiable counterpart in *A. thaliana* or *C. reinhardtii*, so we cannot distinguish between gain or loss in these cases either, although two introns in *rps23* and s*md3* occur in the 3′ UTR region of the *G. stellata* gene, making them likely candidates for intron gain (i.e., UTR sequence that happened to match the requirements for splicing). These data show unambiguously that, although the nucleomorph introns are relatively stable, intron loss can and does occur. At least three, and possibly six, instances of intron loss were identified, corresponding to about 2–5% of the introns examined. Because the sharing of similar, small, and narrow distribution of sizes strongly suggests that intron-size reduction took place in the common ancestor of chlorarachniophytes, this means that already miniaturized introns can be and are eliminated, and therefore that introns have not been retained in these genomes because they are impossible to lose (Cavalier-Smith 2002). The data also quite strongly suggest that intron gain can also occur, at least in the UTR flanking genes.

Nucleomorph Intron and Pseudointron Recognition and Splicing Efficiency

Spliceosomal intron removal requires a series of interactions between the transcript and many protein and ribonucleoprotein complexes making up the spliceosome (Rappsilber et al. 2002), all of which are dependent on sometimes poorly understood sequence characteristics located throughout the intron and surrounding exons. The 5′ and 3′ splice sites and the branching point have all been recognized to be relevant for recognition and removal, as have several elements that are important in some but not all organisms, such as the polypyrimidine tract (Hastings and Krainer 2001). How the extremely small nucleomorph introns are recognized and processed by the spliceosomal machinery is not clear. Not only are the introns too small to contain much in the way of sequence information, but the over 800 *B. natans* introns have also been shown to lack recognizably conserved sequence motifs other than being relatively high in AT and a tendency for an A at position −2 (Gilson et al. 2006).

Aligning the introns of the *G. stellata* nucleomorph reveals a similar situation. The most prominent feature is a high AT: Intronic sequence (excluding GT–AG borders) is 85% AT, contrasting with 71% AT for exons (fig. 2). In addition, a preponderance of A at position −2 is another feature common to the introns of both species (Gilson et al. 2006, fig. 2). In other lineages, the nucleotide preceding the 3′ splice site is of great importance in acceptor site selection (Smith et al. 1993). In mammalian introns, this position is predominantly

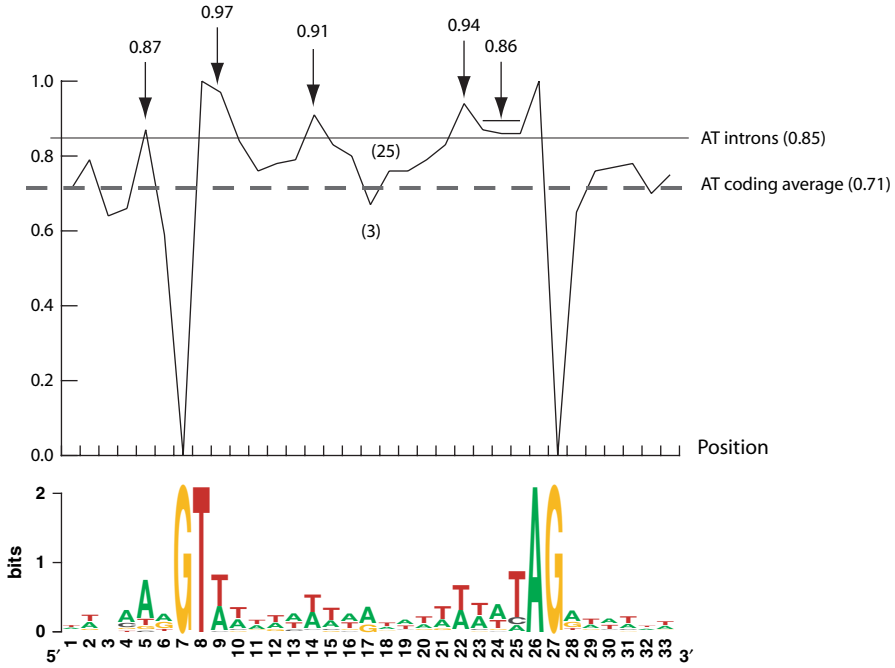


FIG. 2.—Sequence analysis of 151 spliceosomal introns from the nucleomorph of *Gymnochlora stellata*. The upper graph shows A/T fraction at each position in the alignment starting from 6 nt before the 5′ splice site and including 6 nt after the 3′ border. Arrows indicate peaks of AT content. Broken line marks the average fraction of A and T for coding sequence (excludes introns), and the continuous line marks the average fraction of A and T of intronic sequence (excluding GT and AG borders). Numbers in parentheses indicate the number of times those points are represented and correspond to the 21- and 20-nt introns. The lower graph is a WebLogo representation of base composition and information content at each position of the aligned introns. Absolute height of each column indicates the information content (or sequence conservation) in bits at that position, and the relative size of each letter reflects the relative frequency of that nucleotide at such position (Crooks et al. 2004).

C or U, A is less frequent, and G is very rare (Smith et al. 1993). In *G. stellata*, we found only two cases with a G in this position, whereas U is the most prevalent nucleotide (106 introns) followed by C (20 cases). Interestingly, one of the two introns where a G was found at this position failed to be spliced in three of the five sampled transcripts, suggesting this is a key position in nucleomorph introns as well. Three other positions (marked by asterisks in fig. 2) are especially AT-rich with a slight excess of T over A. There is no obvious branch point recognition motif or polypyrimidine tract.

The extremely limited range of intron sizes in *B. natans* led to the suggestion that intron size has evolved to be a critical factor for recognition (Cavalier-Smith 2006; Gilson et al. 2006). It is possible, for example, that as introns shrunk under reductive pressure or due to some deletion-favoring ratchet, the spliceosomal machinery accordingly coevolved to splice smaller and smaller introns. At some point, the reduction in intron size would begin to limit the information content of introns, so the size itself might become a necessary signal for splicing, and from that point, the spliceosome might be limited in its ability to recognize introns over a certain size and in turn lead to a restricted size range. One the other hand, it is also possible that there are no mechanistic restrictions on the upper limit of intron size, and it is instead maintained by the reductive pressure or deletion-favoring ratchet that led to genome compaction in the first place. The former hypothesis would suggest that introns larger than about 21 bp are not only rare but also hard or impossible to splice efficiently. The latter hypothesis would suggest splicing efficiency of such introns would be no different that that of smaller ones. In *G. stellata*, sequences of 24 and 27 bases are spliced, indicating that the spliceosomal complex can remove introns larger than 21 nt. However, these introns have some unusual features.

The 27-nt spliced sequence appears to be a case of spurious splicing in an exon of the *rpb6* gene (see above). Its removal disrupts the predicted protein, and it is a low-efficiency reaction because it was removed in one of four transcripts. The 24-base intron is a genuine intronic sequence corresponding to the second intron in the *tfIIg* gene. Removal is required to maintain the conserved protein sequence and an 18-base intron is present at exactly the same position and phase in *B. natans*. However, removal of this intron also appears to be problematic. We sampled this locus by sequencing 14 reverse transcriptase-PCR products and found several versions of the transcript with four different variants at this intron (fig. 3). In one variant, the 24-base intron is spliced, and this is the only variant where the ORF is maintained in the resulting mRNA. Another variant still encodes the unspliced intron, and the two additional variants use alternative splice sites. In one of these, a 20-nt fragment is removed using an alternative 3′ splice site, whereas in the other, a 25-nt fragment is removed using an alternative 5′ splice site. In other variant transcripts, the first intron is unspliced, which is separated from the second intron by a miniexon of only three codons (fig. 3). Alternative splicing is used in plant and mammalian mRNAs in a highly regulated way resulting in an increased diversity of the proteome (Maniatis and Tasic 2002). However, it is unlikely that the splicing products observed for the

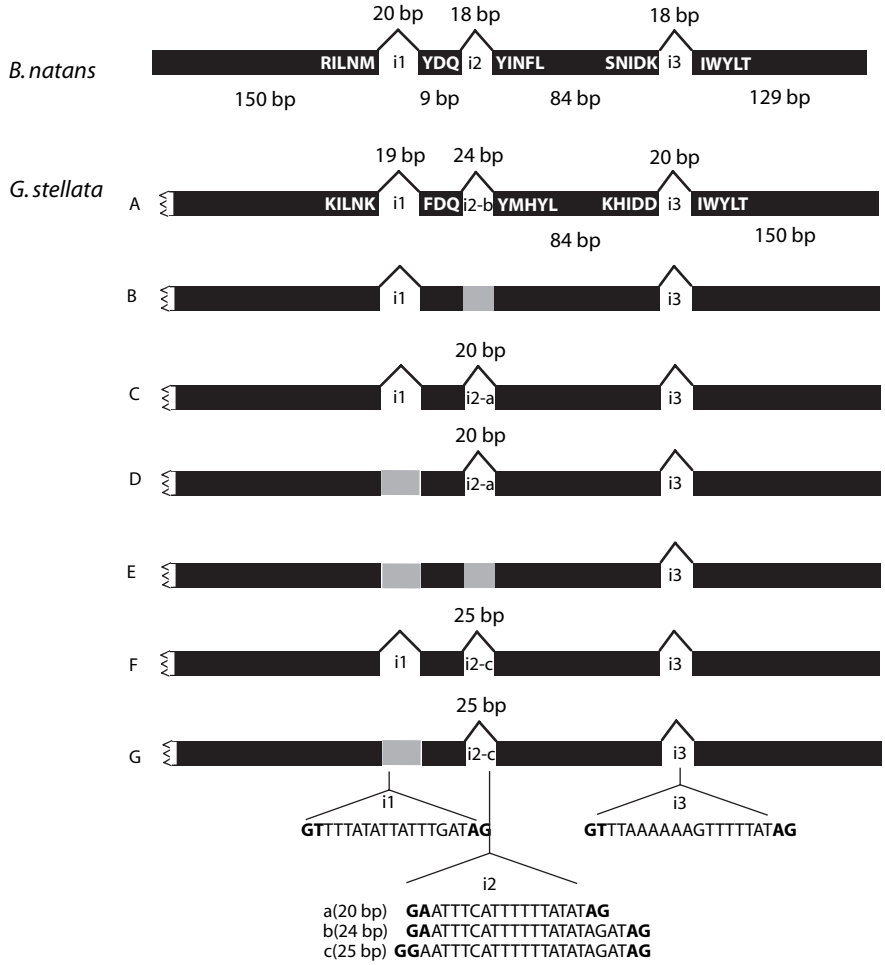


FIG. 3.—Splicing products of the *Gymnochlora stellata tfIIg* gene. The figure shows the intron–exon structure of the *tfIIg* gene from *Bigelowiella natans* (top) and *G. stellata* (*A–G*). i1, i2, and i3 indicate first, second, and third intron positions, respectively. Black boxes represent exons and white letters show translated amino acids surrounding introns. (*A–G*) Correspond to all distinct configurations obtained from EST and RT-PCR sequences. Scheme *A* is the only case that maintains a continuous translated amino acid sequence throughout the gene. Gray rectangles indicate unspliced introns. Nucleotide sequences of all *G. stellata* introns are shown at the bottom (bold letters correspond to intron borders): a, b, and c are three distinct variants of the second intron.

tfIIg gene in *G. stellata* are due to deliberate alternative splicing because every variant except removal of the 24-bp intron results in an immediate termination codon within this otherwise highly conserved transcription factor. Interestingly, in addition to being large, this is one of the few introns in either nucleomorph to use a noncanonical 5′ splice site (GA or GG, see fig. 3). This could also play an obvious role in splicing efficiency, so it is impossible to distinguish between intron size versus noncanonical splice sites when interpreting the apparent inefficiency of this intron. Indeed, the combination of these two rare and apparently deleterious characteristics leads one to wonder if this gene is in the process of disappearing, perhaps because a functional copy has moved to the nucleus in *G. stellata* but not *B. natans*, although no additional copy of this gene was found in our ESTs.

That the nucleomorph spliceosome can recognize a 24-nt intron and mis-splice a 27-nt fragment both suggest that the caliper model for recognition of 18- to 21-bp introns may not be completely accurate. However, it is also clear that intron size and splicing efficiency are closely related. From our EST data, we calculated the total number of splicing events for each intron size class across all introns on all cDNAs. Of the 580 splicing events we recorded, there is a strong correlation between small size and splicing frequency (fig. 4). In general, introns in the "canonical" size range of 18–20 nt are spliced with high efficiency (83–100%), but even in 21-nt introns, splicing efficiency appears to drop, altogether suggesting that the narrow size distribution may be at least partially maintained by spliceosome recognition.

Concluding Remarks

The extreme miniaturization of introns in the chlorarachniophyte nucleomorph is a remarkable departure from the more commonly observed wholesale loss of introns in heavily reduced genomes such as those of cryptomonad nucleomorphs and microsporidia (Lane et al. 2007; Akiyoshi et al. 2009). In these lineages, there are little data to conclude much about the fine-scale distribution of those introns that have been retained, but the lack of data in chlorarachniophytes is even more acute because up to now, these tiny introns were only known from one species, *B. natans*. By

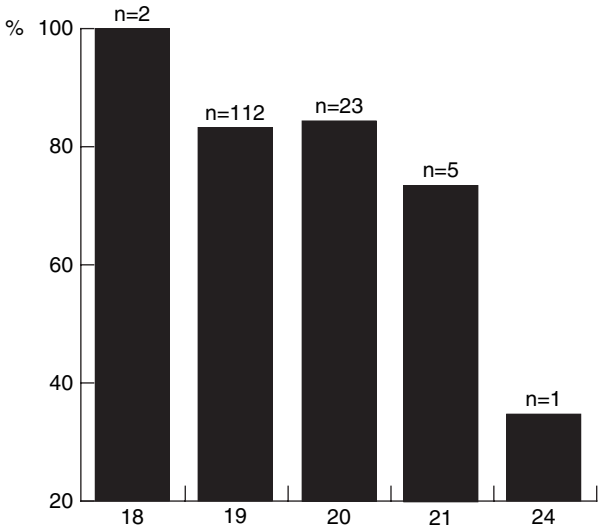


FIG. 4.—Efficiency of splicing for each intron size class. Percentage of observed splicing events over the total transcripts recorded for a given position are plotted for each size. *N*: number of introns in each class.

broadly surveying the nucleomorph genome of *G. stellata*, we can now conclude that the pattern of widespread retention but severe miniaturization of introns took place early in the evolution of chlorarachniophytes and not just in one lineage. In general, *B. natans* and *G. stellata* both tend to have retained most ancestral, ancient introns, but some clear cases of intron loss were also identified, undermining the hypothesis that the size reduction is due to conditions making intron loss impossible. It is also now clear that, although the vast majority of introns are between 18 and 21 bp in length, the spliceosome of *G. stellata* at least is capable of splicing larger introns, albeit inefficiently. Overall, it is likely that a mixture of factors are involved in intron recognition: A few sequence motifs (e.g., GT–AG boundaries, an A at the −2 position, and a T preceding the 3′ splice site), a generally high AT content, and the size itself all likely play a role in spliceosome recognition, but the relative importance of these factors will await direct experimental evidence, in addition to more comparative data from other nucleomorph genomes.

Supplementary Material

Supplementary table S1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Literature Cited

Akiyoshi DE, Morrison HG, Lei S, et al. (11 co-authors). 2009. Genomic survey of the non-cultivatable opportunistic human pathogen, *Enterocytozoon bieneusi*. PLoS Pathog. 5:e1000261.

Archibald JM. 2007. Nucleomorph genomes: structure, function, origin and evolution. Bioessays. 29:392–402.

Bon E, Casaregola S, Blandin G, et al. (11 co-authors). 2003. Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. Nucleic Acids Res. 31:1121–1135.

Cavalier-Smith T. 2002. Nucleomorphs: enslaved algal nuclei. Curr Opin Microbiol. 5:612–619.

Cavalier-Smith T. 2006. The tiny enslaved genome of a rhizarian alga. Proc Natl Acad Sci USA. 103:9379–9380.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Res. 14:1188–1190.

Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu X, Reith M, Cavalier-Smith T, Maier UG. 2001. The highly reduced genome of an enslaved algal nucleus. Nature. 410:1091–1096.

Gilson PR, McFadden GI. 2002. Jam packed genomes—a preliminary, comparative analysis of nucleomorphs. Genetica. 115:13–28.

Gilson PR, Su V, Slamovits CH, Reith ME, Keeling PJ, McFadden GI. 2006. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. Proc Natl Acad Sci USA. 103:9566–9571.

Hastings ML, Krainer AR. 2001. Pre-mRNA splicing in the new millennium. Curr Opin Cell Biol. 13:302–309.

Ishida K. 1999. Diversification of a chimaeric algal group, the Chlorarachniophytes: phylogeny of nuclear and nucleomorph small-subunit rRNA genes. Mol Biol Evol. 16:321–331.

Katinka MD, Duprat S, Cornillot E, et al. (17 co-authors). 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature. 414:450–453.

Keeling P. 2004. A brief history of plastids and their hosts. Protist. 155:3–7.

Keeling PJ, Slamovits CH. 2005. Causes and effects of nuclear genome reduction. Curr Opin Genet Dev. 15:601–608.

Lane CE, van den Heuvel K, Kozera C, Curtis BA, Parsons BJ, Bowman S, Archibald JM. 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. 104:19908–19913.

Maniatis T, Tasic B. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature. 418:236–243.

Merchant SS, Prochnik SE, Vallon O, et al. (117 co-authors). 2007. The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science. 318:245–250.

O'Brien EA, Koski LB, Zhang Y, Yang L, Wang E, Gray MW, Burger G, Lang BF. 2007. TBestDB: a taxonomically broad database of expressed sequence tags (ESTs). Nucleic Acids Res. 35:D445–D451.

Rappsilber J, Ryder U, Lamond AI, Mann M. 2002. Large-scale proteomic analysis of the human spliceosome. Genome Res. 12:1231–1245.

Rogers MB, Gilson PR, Su V, McFadden GI, Keeling PJ. 2007. The complete chloroplast genome of the chlorarachniophyte *Bigelowiella natans*: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. Mol Biol Evol. 24:54–62.

Roy SW, Penny D. 2007. Patterns of intron loss and gain in plants: intron loss-dominated evolution and genome-wide comparison of *O. sativa* and *A. thaliana*. Mol Biol Evol. 24:171–181.

Silver TD, Koike S, Yabuki A, Kofuji R, Archibald JM, Ishida K. 2007. Phylogeny and nucleomorph karyotype diversity of chlorarachniophyte algae. J Eukaryot Microbiol. 54:403–410.

Smith CW, Chu TT, Nadal-Ginard B. 1993. Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. Mol Cell Biol. 13:4939–4952.

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. 102:10936–10941.

Manolo Gouy, Associate Editor

Accepted April 8, 2009