Nucleomorph Genome Sequence of the Cryptophyte Alga *Chroomonas mesostigmatica* CCMP1168 Reveals Lineage-Specific Gene Loss and Genome Complexity

Christa E. Moore, Bruce Curtis, Tyler Mills, Goro Tanifuji, and John M. Archibald*

Integrated Microbial Biodiversity Program, Canadian Institute for Advanced Research, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada

*Corresponding author: E-mail: jmarchib@dal.ca.

Accepted: September 28, 2012

Data deposition: Accession numbers for the three nucleomorph chromosome sequences of *C. mesostigmatica* are reported in the article: CP003680, CP003681, and CP003682. Accession numbers for three *Rhodomonas sp*. CCMP1178 gene sequences, *prsA7*, *ebi*, and *prp8* are also reported in the article: JX515791, JX515790, and JX515789.

Abstract

Cryptophytes are a diverse lineage of marine and freshwater, photosynthetic and secondarily nonphotosynthetic algae that acquired their plastids (chloroplasts) by "secondary" (i.e., eukaryote–eukaryote) endosymbiosis. Consequently, they are among the most genetically complex cells known and have four genomes: a mitochondrial, plastid, "master" nuclear, and residual nuclear genome of secondary endosymbiotic origin, the so-called "nucleomorph" genome. Sequenced nucleomorph genomes are ~1,000-kilobase pairs (Kbp) or less in size and are comprised of three linear, compositionally biased chromosomes. Although most functionally annotated nucleomorph genes encode proteins involved in core eukaryotic processes, up to 40% of the genes in these genomes remain unidentifiable. To gain insight into the function and evolutionary fate of nucleomorph genomes, we used 454 and Illumina technologies to completely sequence the nucleomorph genome of the cryptophyte *Chroomonas mesostigmatica* CCMP1168. At 702.9 Kbp in size, the *C. mesostigmatica* nucleomorph genome is the largest and the most complex nucleomorph genome sequenced to date. Our comparative analyses reveal the existence of a highly conserved core set of genes required for maintenance of the cryptophyte nucleomorph and plastid, as well as examples of lineage-specific gene loss resulting in differential loss of typical eukaryotic functions, e.g., proteasome-mediated protein degradation, in the four cryptophyte lineages examined.

Key words: nucleomorph, cryptophyte, endosymbiosis, genome reduction, comparative genomics.

Introduction

Endosymbiosis is a fundamental process that has shaped the evolution and diversity of modern-day eukaryotes. The engulfment and retention of a photosynthetic cyanobacterium by a primitive nonphotosynthetic eukaryote, known as the "primary endosymbiosis," gave rise to the double-membrane-bound plastids of glaucophytes, red algae, and green algae (and their land plant descendents) (Cavalier-Smith 2000; Reyes-Prieto et al. 2007). The subsequent engulfment and retention of these "primary plastid"-containing cells by nonphotosynthetic eukaryotic hosts has in turn resulted in the evolution of a myriad of biologically, economically, and medically important photosynthetic (and secondarily nonphotosynthetic) unicellular organisms. In most lineages containing these "secondary" plastids, reduction of the endosymbiont nucleus has reached completion; however, in two lineages, the cryptophytes (excluding the aplastidic *Goniomonas*) and chlorarachniophytes, a miniature endosymbiont nucleus, the "nucleomorph," persists (Cavalier-Smith 2002). The nucleomorphs in these two groups are of independent origin: the cryptophyte plastid and nucleomorph are of red algal ancestry (Douglas and Penny 1999; Douglas et al. 2001; Yoon et al. 2002), whereas in chlorarachniophytes, the endosymbiont was a green alga (Gilson et al. 2006; Rogers et al. 2007).

Nucleomorph genomes are the smallest nuclear genomes known and range in size from ~330 to 1,030 kilobase pairs (Kbp) (Silver et al. 2007; Phipps et al. 2008; Tanifuji et al. 2010;

Ishida et al. 2011), orders of magnitude smaller than even the most reduced genomes of eukaryotic parasites, such as the 2.9 megabase pair (Mbp) genome of the microsporidian *Enchephalitozoon cuniculi* (Katinka et al. 2001). The genomes of these miniature nuclei have shrunk dramatically in size and content over millions of years to ~1 Mbp or less and with only several hundred genes. The process of genome reduction has resulted in most of the genes being lost or transferred to the host nucleus, streamlining of the intergenic spacers, and almost complete elimination of repetitive sequence. To date, three cryptophyte nucleomorph genomes have been sequenced, those of *Guillardia theta* (Douglas et al. 2001), *Hemiselmis andersenii* (Lane et al. 2007), and the secondarily nonphotosynthetic *Cryptomonas paramecium* (Tanifuji et al. 2011), which are 550.5 Kbp, 571.4 Kbp, and 485.9 Kbp in size, respectively. A single chlorarachniophyte nucleomorph genome has also been sequenced, the 373 Kbp nucleomorph genome of *Bigelowiella natans* (Gilson et al. 2006). With this limited sampling, nucleomorph genome comparisons within the chlorarachniophyte lineage are impossible and between the chlorarachniophyte and cryptophyte lineages limited. Consequently, we know little about the evolutionary forces that have shaped these genomes and why nucleomorphs persist in chlorarachniophytes and cryptophytes but have been lost in other secondary plastid-bearing algae (reviewed by Moore and Archibald 2009).

Comparative studies of the three sequenced cryptophyte nucleomorph genomes reveal striking similarities with respect to genome architecture and composition. All three genomes (and in fact all nucleomorph genomes examined to date) have three small chromosomes with ribosomal DNA (rDNA) operons on the chromosome ends and with one of two types of unusual telomere sequences: GA_n (GA_{17} for *H. andersenii* and GA_9 for *C. paramecium*) and $[AG]_7AAG_6A$ for *G. theta* (Douglas et al. 2001; Lane et al. 2007; Silver et al. 2007; Tanifuji et al. 2010, 2011). These genomes display a similar degree of nucleotide composition bias (~75% A+T) and have similar coding capacities (518–548 genes). This latter point is interesting given that their total genome sizes differ by up to 64 Kbp, yet they have very similar gene densities (0.98–1.09 gene/Kbp). Approximately 60% of the genes annotated in these genomes encode proteins involved in core eukaryotic processes, such as transcription, translation, and protein folding, but the remaining ~40% cannot be ascribed a particular function based on sequence similarity as they either show homology only to other cryptophyte nucleomorph genes of unknown function or they show no similarity whatsoever to any known gene in current databases. Essentially nothing is known about the latter "ORFan" genes except that their transcripts have been observed in EST surveys (e.g., Patron et al. 2006), and they tend to encode proteins rich in amino acids specified by A+T-rich codons (Lane et al. 2007).

Nucleomorph gene sequences are notoriously divergent compared with their homologs in free-living organisms and are often shorter as a result of internal deletions and the whittling away of amino and carboxy terminal-coding regions (Lane et al. 2007). In addition, spliceosomal introns are rare in cryptophyte nucleomorph genomes and are in fact completely absent in the case of *H. andersenii*, the first described instance of complete spliceosomal intron loss from a nuclear genome (Lane et al. 2007). Of the known genes present in nucleomorph genomes, there are very few whose protein products function in the plastid, which is surprising given that nucleomorph-encoded "plastid" genes are often touted as the primary reason for nucleomorph persistence (Zauner et al. 2000; Gilson and McFadden 2002; Archibald 2007). To gain a better understanding of the evolution and ultimate fate of nucleomorphs in cryptophyte algae, we completely sequenced the nucleomorph genome of *Chroomonas mesostigmatica* CCMP1168. Members of the genus *Chroomonas* have predicted nucleomorph genome sizes that are >200 Kbp larger than those currently sequenced (Lane et al. 2006; Tanifuji et al. 2010). The nucleomorph genome of *C. mesostigmatica* is the largest and the most complex of its kind, with numerous repetitive regions and multicopy genes, features that are rare in nucleomorph genomes sequenced to date. Comparative analyses provide insight into the identity of some of the mysterious ORFan genes, evidence for a more highly conserved core set of genes than previously thought, and further support for the notion that nucleomorphs have yet to reach an end point in their reductive evolution.

Materials and Methods

Cell Culture and DNA Isolation

Chroomonas mesostigmatica CCMP1168 was grown in f/2 media at room temperature on a 12-h light/dark cycle. Total DNA was extracted from 120 l of dense culture using a standard phenol/chloroform extraction procedure. Total DNA was fractionated by Hoechst dye (No 33258, Sigma-Aldrich, St. Louis, MO, USA)-cesium chloride density gradient centrifugation, and the resulting fractions were analyzed by Southern blot hybridizations using organelle genome-specific probes to identify fractions of predominantly mitochondrial (*cox*1), plastid (16S rDNA), nuclear (host 18S rDNA), and nucleomorph (endosymbiont 18S rDNA) origin. Hoechst dye-cesium chloride density gradient centrifugation and fraction purification, and Southern blot hybridizations were performed as described in Lane and Archibald (2006) and Lane et al. (2006).

RNA Extraction, Reverse-Transcriptase Polymerase Chain Reaction, and Transcriptome Sequencing

RNA for reverse-transcription (RT)-polymerase chain reaction (PCR) experiments was obtained using the RNeasy® Mini and

RNeasy® MinElute™ Cleanup kits (Qiagen, Toronto, ON, Canada). The quality and quantity of RNA extracted was assessed by gel electrophoresis. Two site-specific primers per gene, one upstream of the 5′-intron splice site and one downstream of the 3′-intron splice site, were designed for RT-PCR verification of predicted spliceosomal introns in the following genes: *rps*16 (Cmeso_rps16_F1: 5′-CATAGTCCAAGTATTCGGAAAAA-3′ and Cmeso_rps16_R1: 5′-GCTCCTTTTCCTCCTGCTTT-3′), *rps*23 (Cmeso_rps23_F1: 5′-TGTTTAAATAAAAGAATGGGATCAG-3′ and Cmeso_rps23_R1: 5′-TCAAAGAAACCCCTGCTACC-3′), *rps*24 (Cmeso_rps24_F1: 5′-GGAAGAAATCAAAATTACCACCA-3′ and Cmeso_rps24_R1: 5′-TGTTAAAACGAGTTTTTCCTCTGA-3′), and *rpl*9 (Cmeso_rpl9_F1: 5′-GCATGAAACCAATTCTAACAAACA-3′ and Cmeso_rpl9_R1: 5′-CGGCACCAGCAGATACAAG-3′). RT reactions using the site-specific primers were performed using the Omniscript™ RT kit (Qiagen, Toronto, ON, Canada), followed by PCR amplification of the resulting cDNA. The same site-specific primers were also used in PCR reactions with total genomic DNA template. Amplicon sizes were determined and compared using gel electrophoresis. Purified cDNA PCR products for each of the genes were then cloned using the TOPO-XL vector (Invitrogen, Burlington, ON, Canada) and Sanger sequenced using a Beckman-Coulter CEQ 8000 capillary DNA sequencer.

For transcriptome sequencing, total RNA was extracted from ~3.3×10^7 cells using Trizol (Invitrogen, Burlington, ON, Canada), followed by standard phenol/chloroform precipitation and subsequent precipitation using lithium chloride. A 10 µg sample of RNA was sequenced using Illumina RNA-Seq (National Center for Genome Resources, Santa Fe, NM, USA), generating 2.37 gigabase pairs of raw sequence data.

Pulsed-Field Gel Electrophoresis

Agarose plugs for pulsed-field gel electrophoresis (PFGE) were created using 400 ml of log-phase culture following the method described in Eschbach et al. (1991) for three approximate cell counts of 5×10^6, 1×10^7, or 5×10^7 per plug. The agarose plugs were run in a CHEF-DR III Pulsed-Field Electrophoresis System (BioRad Laboratories, Hercules, CA, USA) on a 1% agarose gel dissolved in 0.5% TBE buffer (TRIS, boric acid, and ethylenediaminetetraacetic acid) at 14 °C for 22 h using a voltage of 6.0 V/cm and a 0.2–22 s switch time. The PFGE was repeated using plugs with 5×10^7 cells, a voltage of 4.1 V/cm, and a 30–10 s switch time for 60 h.

Genome Sequencing, Assembly, and Annotation

A sample containing approximately 5 µg of nucleomorph genome-enriched DNA was 454 pyrosequenced (GS FLX titanium series, McGill University and Génome Québec Innovation Centre, Montréal, QC, Canada) to a depth of ~100× coverage, generating over 110 Mbp of raw sequence data. Reads over 300 base pairs (bp) in length were assembled using the GAP4 program (Staden package, v4.11; Bonfield et al. 1995), and contigs were manually refined. An additional 1 µg of nucleomorph-enriched DNA was sequenced on an Illumina GAIIx sequencer (Cofactor Genomics, St. Louis, MO, USA), producing 833,000 reads, 90% of which were integrated into the 454-based assembly to aid in frameshift and homopolymer correction. Contigs were assigned to their respective chromosomes using Southern blot hybridizations, and the remaining gaps were closed using PCR and Sanger sequencing. Raw reads from the RNA-Seq data were mapped onto the contigs using the Burrows–Wheeler Aligner (v0.5.9 with default settings; Li and Durbin 2009) to further verify the assembly and aid in spliceosomal intron boundary identification. Open reading frames (ORFs) larger than 50 amino acids (aa) in size were predicted using Artemis (v13.0; Rutherford et al. 2000), and genes were manually annotated based on blastx and blastp (*e* value < 0.001; Altschul et al. 1990) searches against the GenBank nr database (National Center for Biotechnology Information, Bethesda, MD, USA) and a local database of red algal (*Cyanidioschyzon merolae*) and cryptophyte nucleomorph genomic data (*G. theta, H. andersenii,* and *C. paramecium*). Pfam searches were also performed (Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK). Annotations followed the conventions of Douglas et al. (2001), Lane et al. (2007), and Tanifuji et al. (2011). For ease of comparison with other cryptophyte nucleomorph ORFs, we categorized each of the *C. mesostigmatica* nucleomorph ORFs as a "conserved ORF," a "nORF," or an "ORFan." A conserved ORF is a protein-coding gene with annotated homologs in other nuclear (and in most cases, other cryptophyte nucleomorph) genomes. A nORF is a hypothetical protein-coding gene with annotated homologs only in other cryptophyte nucleomorph genomes. An ORFan is a hypothetical protein-coding gene with no significant sequence similarity to any gene in any other known genome.

Spliceosomal introns were identified manually using canonical GT/AG intron boundary searches and alignments of homologous protein sequences. Transfer RNAs (tRNA) were identified using tRNAScan-SE (v1.21; Lowe and Eddy 1997) (http://lowelab.ucsc.edu/tRNAscan-SE/, last accessed June 2, 2011), and ribosomal RNAs (rRNA) were identified using blastn. One small nuclear RNA (snRNA) was identified using a blastn search against our local database. The complete nucleomorph genome sequence of *C. mesostigmatica* CCMP1168 has been deposited in GenBank using the following accession numbers: CP003680, CP003681, and CP003682. One-way analysis of variance tests for significance were performed for protein size and intergenic spacer size comparisons between *C. mesostigmatica*, *H. andersenii*, *C. paramecium*, and *G. theta* nucleomorph data using AnalystSoft Inc., StatPlus:mac—statistical analysis program for Mac OS, version 2009 (www.analystsoft.com/en/, last accessed August 24, 2012).

Results and Discussion

Genome Architecture and Size Variation

The nucleomorph genome of *C. mesostigmatica* CCMP1168 is comprised of three linear chromosomes of ~244 Kbp, 233 Kbp, and 226 Kbp, with a total genome size of 702.9 Kbp (fig. 1). A previous karyotyping analysis of the *C. mesostigmatica* nucleomorph showed three similarly sized chromosomes totaling approximately 805 Kbp (Lane et al. 2006). We performed an independent analysis, which suggests that the three nucleomorph chromosomes are indeed smaller than the original size estimates, supporting the genome size determined by sequencing (fig. 2). Although smaller than initial PFGE-based estimates, the *C. mesostigmatica* nucleomorph genome is still the largest nucleomorph genome sequenced to date. The G+C content is 25.94%, similar to that seen in other cryptophyte nucleomorph genomes (table 1). The telomere sequence is GA_{13}, similar to the GA_{17} and GA_9 nucleomorph telomere sequences of *H. andersenii* (Lane et al. 2007) and *C. paramecium* (Tanifuji et al. 2011), respectively. Subtelomeric rDNA operons exist on all six chromosome ends, followed by a long stretch (up to 13 Kbp) of repeated sequence consisting of several ORFs (for both hypothetical proteins and proteins of known function) and repetitive sequence that has presumably been homogenized through recombination. There are several other multicopy genes that appear on more than one chromosome (two copies of ubiquitin, *cpeT*-like and *tfIIA-S* genes, and three copies of *orf266*), and in each case, the duplicates are essentially identical to one another. The majority of the genes are, however, present in single copy. Unlike all other nucleomorph genomes sequenced to date, the *C. mesostigmatica* nucleomorph genome is rich in simple, repetitive sequence, consisting primarily of innumerable A-T homopolymer runs of varying length (in coding and noncoding sequence), short sequence repeats (in the intergenic regions and in the internal transcribed spacer [ITS] regions of the rDNA operon), and a 12-bp repeat ($TA_2GA_2TA_5$, 4–25 copies) on five of the six chromosome ends. In addition to repetitive sequence in the intergenic spacers, there is also repetitive sequence present within both protein genes and rRNA genes. The variable regions of the 28S large subunit rRNA gene contain lengthy A-T homopolymer runs (up to 37 bp) and short sequence repeats. Repetitive elements in the variable regions of rDNA genes have been observed in other organisms and can mimic the base composition of the ITS sequences (see Gray and Schnare 1990 and references therein).

The *C. mesostigmatica* nucleomorph genome harbors 580 genes: 505 protein-coding genes (453 unique genes), 50 tRNAs (all tRNAs present), and 25 other nonmessenger RNA genes (rRNAs and a U6 snRNA) (table 1). At 703 Kbp, the genome is more than 100 Kbp larger than any of the other cryptophyte nucleomorph genomes sequenced to date, yet the total number of genes encoded is remarkably similar. There are only 61 more genes in *C. mesostigmatica* than in *C. paramecium* (whose genome is 217 Kbp smaller), 32 more genes than in *G. theta* (which is 152.4 Kbp smaller), and 55 more genes than *H. andersenii* (which is 131.5 Kbp smaller). The gene density of the *C. mesostigmatica* nucleomorph genome is 0.83 genes/Kbp, which is notably lower than *H. andersenii*, *G. theta*, or *C. paramecium* at 1.09, 0.98, and 1.07 genes/Kbp, respectively. Overall, however, although the number of genes in the *C. mesostigmatica* nucleomorph genome is higher than in the other sequenced nucleomorph genomes, gene number is not strictly correlated with nucleomorph genome size. The larger size of the *C. mesostigmatica* nucleomorph genome cannot be attributed solely to the presence of more genes.

Previous work (Lane et al. 2007; Tanifuji et al. 2011) observed size differences for homologous nucleomorph-encoded proteins, which could account for some of the genome size variation seen thus far. We tested this hypothesis with a four-way comparison. A trend toward a decrease in gene/protein size with decreasing nucleomorph genome size was observed, but this trend is not strict and not statistically significant, whether we compare the average size of all the proteins encoded in each of the cryptophyte nucleomorph genomes ($P > 0.05$) or a 227-protein subset that is shared among all four cryptophyte nucleomorphs ($P > 0.05$) (table 1). If the sizes of these 227 homologous proteins are examined individually, we do see a net gain in amino acids when compared with the total number of amino acids for these proteins from the smallest nucleomorph genome to the largest (74,684, 75,098, 79,289, and 80,142 for *C. paramecium*, *G. theta*, *H. andersenii*, and *C. mesostigmatica*, respectively). An increase of 853 amino acids for the 227 proteins in *C. mesostigmatica* compared with *H. andersenii* does increase the genome size, but there is a 130 Kbp size difference between these two genomes, so the increase due to protein size alone is minimal.

Interestingly, a significant difference in protein size is observed ($P < 0.01$) when the average sizes of the ORFan genes are compared across all four genomes (table 1). It was previously observed that the smallest sequenced cryptophyte nucleomorph genome, that of *C. paramecium*, encodes ORFan proteins that are on average much smaller in size compared with those in the other nucleomorph genomes, and so it was hypothesized that nucleomorph genome size diversity may be largely influenced by size variation in these ORFan genes (Tanifuji et al. 2011). However, the *C. mesostigmatica* nucleomorph genome does not encode larger ORFan proteins on average; ORFan gene size is thus not a contributing factor to increased nucleomorph genome size in *C. mesostigmatica* and cannot account for the nucleomorph genome size variation observed within the cryptophytes. It is also not the case that the *C. mesostigmatica* nucleomorph genome is significantly enriched in longer genes. If we compare the distribution of ORFs according to their size across the four cryptophyte

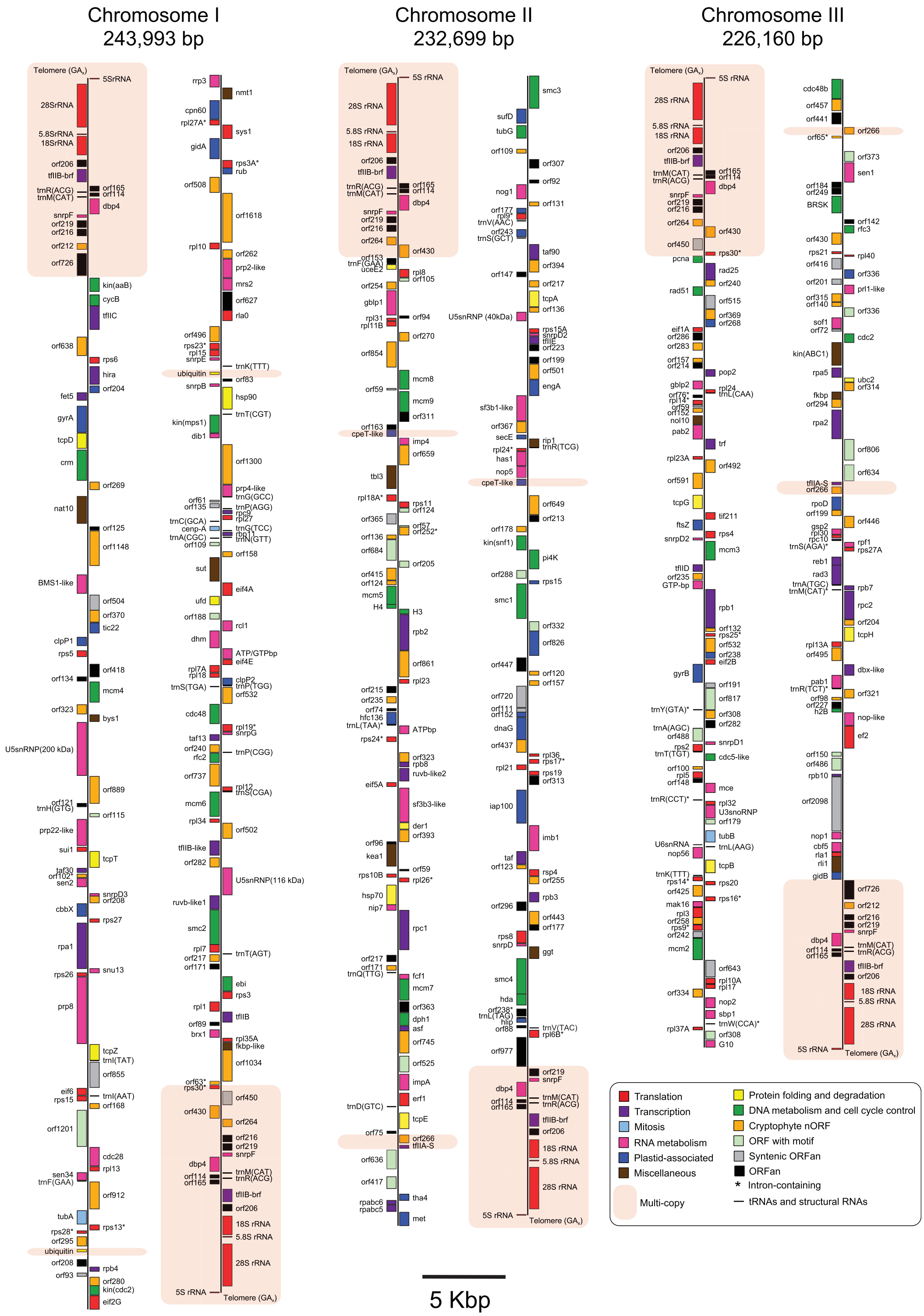

FIG. 1.—*Chroomonas mesostigmatica* CCMP1168 nucleomorph genome map. The genome is comprised of three linear chromosomes, shown broken artificially at their midpoints, with genes on the left indicating transcription from bottom to top, and genes on the right indicating transcription from top to bottom. Colors of the blocks correspond to assigned functional categories, and multicopy genes are highlighted in pink. An asterisk beside the gene name indicates the gene contains an intron. Genes for which there are currently no known homologs (ORFans) are shown in black, genes that have homologs only in other cryptophyte nucleomorphs (nORFs) are shown in orange, and motif-containing genes whose identity cannot be determined with confidence are shown in light green. ORFan genes that retain conserved positions within syntenic regions between one or more other cryptophyte nucleomorphs are shown in gray (syntenic ORFans).

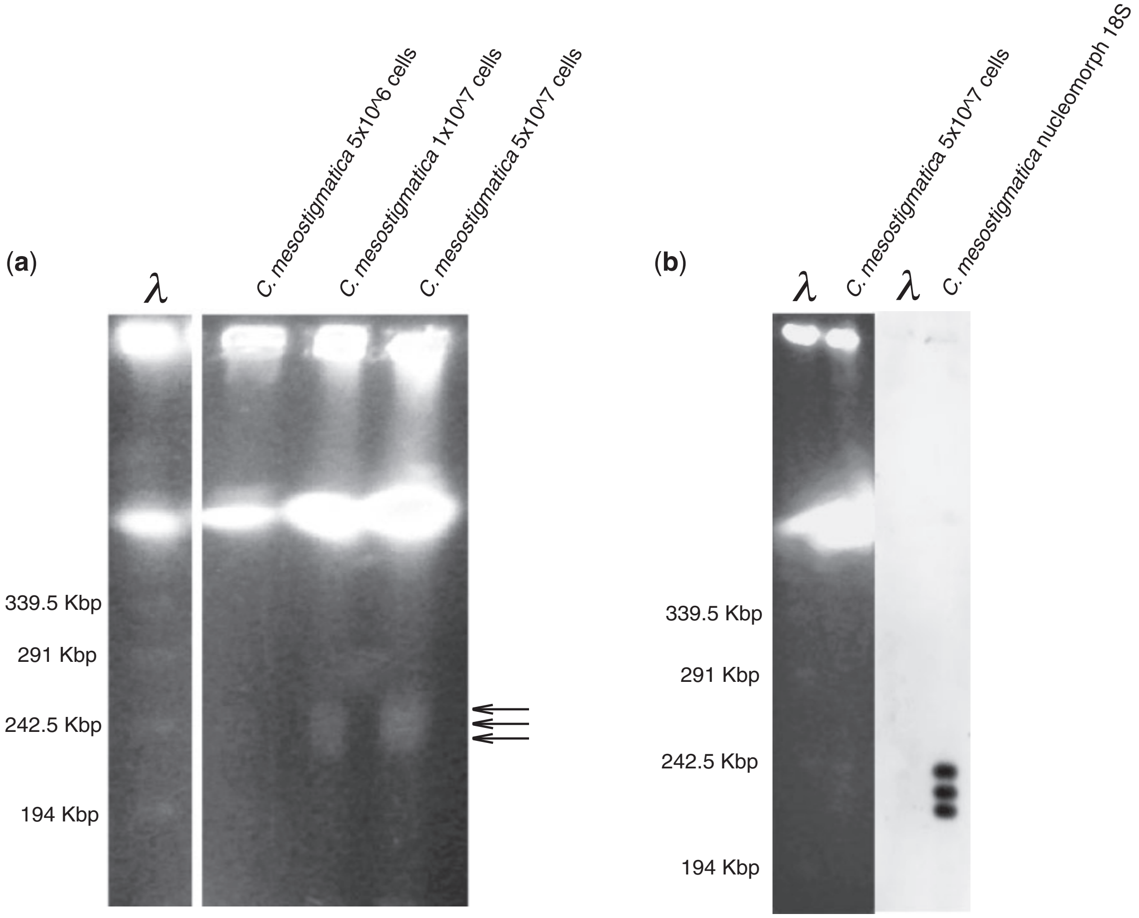
(a)
λ
C. mesostigmatica 5x10^6 cells
C. mesostigmatica 1x10^7 cells
C. mesostigmatica 5x10^7 cells
339.5 Kbp
291 Kbp
242.5 Kbp
194 Kbp
(b)
λ
C. mesostigmatica 5x10^7 cells
λ
C. mesostigmatica nucleomorph 18S
339.5 Kbp
291 Kbp
242.5 Kbp
194 Kbp

Fig. 2.—Karyotype analysis of the *C. mesostigmatica* nucleomorph. (*a*) Ethidium-bromide-stained *C. mesostigmatica* chromosomes separated by PFGE for 24 h for three approximate total cell counts: 1) 5 × 10^6 cells, 2) 1 × 10^7 cells, and 3) 5 × 10^7 cells. Nucleomorph chromosomes are indicated by arrowheads. Lambda DNA is used for size markers. (*b*) Ethidium-bromide-stained *C. mesostigmatica* nucleomorph chromosomes separated by PFGE for 60 h (left) and Southern-blot hybridization using a *C. mesostigmatica* nucleomorph-specific 18S probe.

nucleomorph genomes, we see a striking trend when the shortest ORFs (<150 aa) are considered, that is, the percentage of ORFs of this size is negatively correlated with genome size, such that the smallest genome contains the highest percentage of small ORFs, and as genome size increases, the percentage of ORFs of that size decreases (fig. 3). As ORF size increases, the trend is not perfectly linear. Nevertheless, we do observe that for ORFs longer than 550 amino acids, there is a slight trend toward the larger genomes containing more of these longer genes than the smaller genomes. This trend is, however, not enough to account for most of the genome size variation observed.

The single largest contributing factor to the larger nucleomorph genome size in *C. mesostigmatica* is the amount of noncoding sequence. The average intergenic spacer size is significantly larger ($P < 0.01$) in the *C. mesostigmatica* nucleomorph genome compared with the nucleomorph genomes of *H. andersenii*, *C. paramecium*, and *G. theta* when the average size of all the intergenic spacers are compared (200 bp for *C. mesostigmatica*, 132 bp for *H. andersenii*, 93 bp for *G. theta*, and 102 bp for *C. paramecium*) and when the average size of the intergenic spacers within syntenic regions are compared (91 bp for *C. mesostigmatica*, 77 bp for *H. andersenii*, 41 bp for *G. theta*, and 62 bp for *C. paramecium*) (table 1). Interestingly, differences were observed in intergenic spacer sizes depending on the relative orientation of the bounding genes. For all four species examined, the intergenic spacers are the smallest when the bounding genes are oriented tail-to-tail (table 2). The size difference is statistically significant when compared with intergenic spacers bounded by genes that are oriented head-to-head, or head-to-tail, for *H. andersenii*, *G. theta*, and *C. paramecium*. In fact, the highest prevalence of overlapping genes, or genes that have no spacer between them, occurs when the genes are oriented tail-to-tail (table 2). In sum, the observed diversity in cryptophyte nucleomorph genome size can be attributed primarily to differences in the

Table 1
Comparison of Cryptophyte Nucleomorph Genome Features

Genome Feature	*Chroomonas mesostigmatica*	*Hemiselmis andersenii*	*Guillardia theta*	*Cryptomonas paramecium*
Genome size (Kbp)[a]	702.9	571.4	550.5	485.9
G+C content (%)	25.94	25.18	26.43	26.05
Number of genes[b]				
Protein coding	505	472	487	466
Total	580	525	548	519
Gene density (genes/Kbp)	0.83	1.09	0.977	1.07
Number of overlapping genes	20	44	11	33
Average protein length (aa)				
All proteins	357	338	312	289
227 shared proteins	353	349	329	331
ORFans	264	190	268	190
Average intergenic spacer (bp)				
Syntenic spacers	91	77	41	62
All spacers	200	132	93	102
Number of ORFan genes (% of protein-coding genes)	94 (19)	74 (16)	155 (32)	133 (29)
Number of spliceosomal introns	24	0	17	2

[a] Telomere sequences are not included in the total genome size.

[b] Includes current data from GenBank, the gene analysis of Tanifuji et al. (2011), and the previously unannotated *gidB* in *G. theta*.

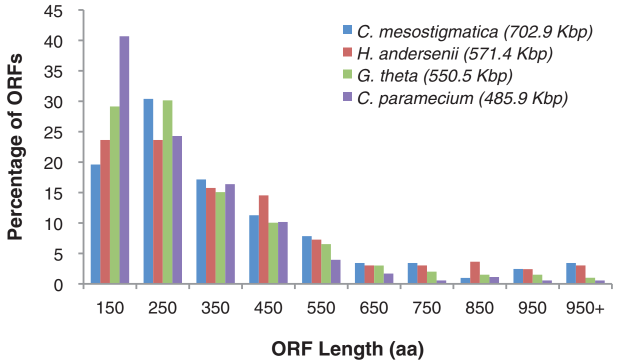

Fig. 3.—Percentage of all cryptophyte nucleomorph ORFs per genome as a function of length. Each of the four nucleomorph genomes examined in this study has a different distribution of ORF sizes. The smaller nucleomorph genomes are enriched in shorter ORFs, and as the size of the ORF increases, the percentage of those ORFs decreases. Larger nucleomorph genomes are slightly enriched in longer ORFs.

amount of noncoding DNA, as well as the number of genes (in particular the presence/absence of multicopy genes), together with minor variation in the length of homologous genes.

Proteins of Known Function

Of the 505 putative protein-coding genes in the *C. mesostigmatica* nucleomorph genome, 235 encode proteins predicted to have core eukaryotic "housekeeping" functions. These include transcription, translation, DNA metabolism and cell cycle control, RNA metabolism, protein folding, protein degradation, and mitosis (supplementary table S1, Supplementary Material online). The *C. mesostigmatica* nucleomorph genome contains the identical set of 31 plastid-associated genes found in the nucleomorph genomes of both *H. andersenii* and *G. theta*. The gene for the plastid-targeted glucose-inhibited division protein B, *gidB*, was initially presumed missing from the plastid and nucleomorph genomes of *G. theta* (Douglas et al. 2001). However, our comparative analyses have identified a single copy of *gidB* in the nucleomorph genome of *G. theta*. Apart from the nonphotosynthetic species *C. paramecium*, which has lost many photosynthesis-related genes, it is not clear why precisely the same set of 31 genes for plastid-associated proteins are retained in *C. mesostigmatica*, *G. theta,* and *H. andersenii* and have not been differentially transferred to the host nuclear genome. Nonetheless, their presence in the four species examined (a subset of which is present in the nonphotosynthetic species *C. paramecium*) suggests that this particular suite of plastid-associated genes was "locked in" before the radiation of the major cryptophyte lineages.

A comparison of gene content for biological functions conserved across the four species reveals a high degree of overlap (fig. 4). Out of the 311 genes examined, 216 (69.5%) are present in the nucleomorph genomes of all four species. In the case of genes that are missing from two or three species, there are no clear patterns to account for their loss; a punctate distribution of gene loss is observed across all the functional categories examined, which include transcription, translation, mitosis, cell cycle control, protein folding, protein degradation, DNA metabolism, and RNA metabolism (supplementary table S1, Supplementary Material online). The only notable exception is that the nucleomorph genome of *C. mesostigmatica*

Table 2

Comparison of Average Intergenic Spacer Size for Different Gene Orientations

	Average Intergenic Spacer Size (bp)			
Feature	***Chroomonas mesostigmatica***	***Hemiselmis andersenii***	***Guillardia theta***	***Cryptomonas paramecium***
Gene orientation				
Head–Head	203.4	130.4	103.0	119.8
Head–Tail	217.9	152.2	106.3	115.5
Tail–Tail	166.5	95.7	64.3	63.7
One-way ANOVA *P* level	0.07216	0.00562	0.04407	0.00006
No. of 0 bp spacers/total				
Head–Head	8/151 (5.3%)	2/135 (1.5%)	7/137 (5.1%)	10/127 (7.9%)
Head–Tail	6/271 (2.2%)	4/249 (1.6%)	20/273 (7.3%)	7/259 (2.8%)
Tail–Tail	10/154 (6.5%)	8/138 (5.8%)	31/140 (22.1%)	18/130 (13.8%)

Note.—ANOVA, analysis of variance.

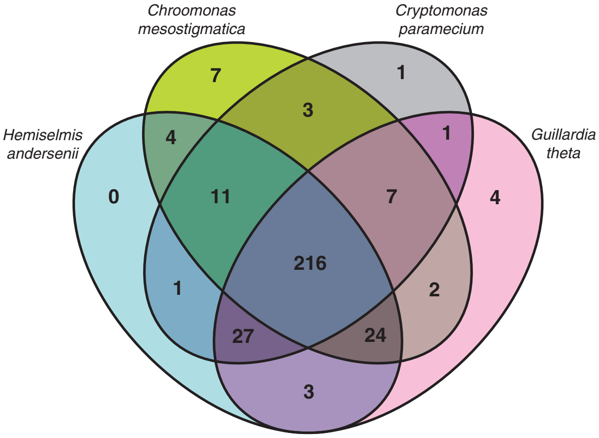

Fig. 4.—Four-way cryptophyte nucleomorph gene content comparison. There are 311 genes of known or predicted function annotated in cryptophyte nucleomorph genomes, 216 of which (~70%) are present in all four cryptophyte nucleomorph genomes presently sequenced, forming a highly conserved core gene set. Aside from the lineage-specific photosynthesis-related, spliceosome, and proteasome gene loss, the distribution of missing genes appears to be random with respect to each species and functional gene category.

contains more genes whose protein products function in spliceosomal intron removal (discussed later). There are, however, three very distinct patterns of gene loss for those genes lost from only a single species. As previously reported, *C. paramecium* has lost photosynthesis capability and as a result has a reduced set of nucleomorph-encoded photosynthesis-related genes (Tanifuji et al. 2011). As expected, the set of 24 genes that are present in *C. mesostigmatica*, *H. andersenii,* and *G. theta* but missing from *C. paramecium* contains primarily plastid-associated genes. Similarly, the nucleomorph genome of *H. andersenii* has been shown to be completely devoid of spliceosomal introns and deficient in splicing-related genes, thus it is unsurprising that genes required for spliceosomal intron removal make up the set of seven genes missing from *H. andersenii*. There is no obvious functional explanation to account for the 11 genes that are absent from the nucleomorph genome of *G. theta*. Previous comparative studies of cryptophyte nucleomorph genes have shown that *G. theta* genes tend to be more divergent compared with their homologs in other nucleomorph genomes (Tanifuji et al. 2011). Our data support this observation, and having additional nucleomorph genome data from a close relative of *G. theta* would help in determining whether these genes are indeed missing or are present but have diverged beyond detection by sequence similarity.

The most surprising observation from the four-way comparison is that in the set of 27 genes that are shared between *G. theta*, *H. andersenii,* and *C. paramecium* but absent from *C. mesostigmatica*, 22 are involved in protein degradation. All 21 genes encoding subunits of the proteasome are missing from the nucleomorph genome of *C. mesostigmatica*, as well as the E2 ubiquitin conjugating enzyme gene *ubc4* (supplementary table S1, Supplementary Material online). The significance of this observation is unclear. We examined RNA-Seq data from *C. mesostigmatica* for nuclear genes that encode proteasome subunits that could potentially be targeted into the periplastidial compartment, that is, the residual cytoplasm of the endosymbiont in which proteasome-mediated protein degradation would presumably take place. However, only a single, apparently host-derived copy of each proteasome subunit gene was found, each of which encodes a protein with no obvious amino-terminal extensions reminiscent of the bipartite leader sequences required for such targeting (Gould et al. 2006a, 2006b). It is thus unclear whether canonical protein degradation pathways exist within the periplastidial compartment of *C. mesostigmatica* and if so, which proteins are involved. It is entirely possible that some of the mysterious ORFan genes, which constitute 20% of the protein-coding genes in the *C. mesostigmatica* nucleomorph genome, and of which we know nothing about, play a role in protein degradation. Interestingly, some hallmark genes of the ubiquitin–proteasome degradation pathway are present in the *C. mesostigmatica* nucleomorph genome, such as ubiquitin (two copies), the ubiquitin-fusion degradation protein (*ufd*), and two ubiquitin-conjugating enzymes (*uceE2* and *ubc2*). However, most of these genes are known to be involved in other cellular processes besides protein degradation, such as protein import by SELMA, the preprotein translocator located in the second outermost membrane of complex plastids in cryptophytes (*ufd* as well as *cdc28* and *der1*, which are also nucleomorph-encoded in *C. mesostigmatica*) (Bolte et al. 2011); DNA damage repair (*ubc2*) (Jentsch et al. 1987); disassembly of the mitotic spindle (*ufd*) (Cao et al. 2003); and stress response, gene expression, and chromatin structural maintenance (ubiquitin) (Finley and Chau 1991, Conaway et al. 2002, and Gardner et al. 2005, respectively). Furthermore, in the absence of a complete *C. mesostigmatica* nuclear genome sequence, it is presently not possible to conclude with certainty that the host nuclear genome does not possess at least some of the missing proteasome genes. However, it is interesting that the nucleomorph genome of the chlorarachniophyte *B. natans* is also devoid of obvious proteasome subunit genes (Gilson et al. 2006). Analysis of the nuclear genome sequence of *B. natans* (http://genome.jgi.doe.gov/Bigna1/Bigna1.home.html, last accessed July 19, 2012) should provide more insight into the mechanism of protein degradation for nucleomorph-derived proteins in chlorarachniophytes, insight that could prove useful in characterizing the protein degradation system for nucleomorph-derived proteins in *C. mesostigmatica.*

Proteins of Unknown Function

A substantial proportion of the protein-coding genes in the *C. mesostigmatica* nucleomorph genome (∼30%) are hypothetical in nature. One-third of these genes are cryptophyte nucleomorph-specific ORFs, or nORFs, meaning they have clear homologs in other cryptophyte nucleomorph genomes but not in other known genomes. The remaining two-thirds, however, are true ORFan genes, meaning they show no obvious sequence-based homology to any gene in known databases, nucleomorph derived or otherwise. Some of these genes are present in syntenic blocks, that is, they occupy the same position within a block of genes conserved between different nucleomorph genomes. These syntenic ORFans, as first described by Lane et al. (2007), not only exhibit positional conservation but often their size is also conserved. Despite showing no detectable sequence similarity, based on their size and positional information, they are presumed to be homologous.

Interestingly, many of the *C. mesostigmatica* nORFs show significant sequence similarity to those of *H. andersenii* but are noticeably less similar to those of *C. paramecium* or *G. theta*, a pattern that is also seen for genes of known function. This gives further support to phylogenies inferred from host and nucleomorph 18S rDNA, which suggest that members of the genus *Chroomonas* and the genus *Hemiselmis* are more closely related to each other than they are to other known cryptophytes (e.g., Hoef-Emden 2008). Upon close inspection, many of the hypothetical protein-coding genes in the *C. mesostigmatica* and *H. andersenii* nucleomorph genomes do in fact show sequence similarity to each other. The high degree of sequence similarity and the more conservative nature of the *C. mesostigmatica* ORFs allowed us to ascribe predicted functions to nine previously annotated hypothetical protein-coding genes based on homology and synteny. These include the plastid-associated gene *gidB*, the DNA-directed RNA polymerase II subunit gene *rpb4*, and the highly conserved mini-chromosome maintenance genes *mcm8* and *mcm9* in *G. theta*; the messenger RNA (mRNA) splicing factor gene *prl1-like* in *H. andersenii*; the acidic ribosomal protein gene *rla1* in both *H. andersenii* and *C. paramecium*; and the nucleolar protein gene *nop-like* and spliceosomal genes *prp4-like* and U5 snRNP (40 kDa) in *C. paramecium*. In addition, the observed sequence similarity between hypothetical *C. mesostigmatica* ORFs to previously annotated ORFan genes in *H. andersenii, C. paramecium*, and *G. theta* allowed us to reclassify 58 of these ORFan genes as nORFs (we were also able to reclassify 73 ORFan genes as syntenic ORFs). In terms of the total proportion of nORF genes relative to ORFans, the addition of *C. mesostigmatica* nucleomorph protein genes resulted in a reduction from 76% ORFan genes in a three-way analysis (i.e., *G. theta*, *H. andersenii,* and *C. paramecium*) to 55% in a four-way comparison (fig. 5). This significant reduction can be mostly attributed to the close relatedness of *C. mesostigmatica* to *H. andersenii*. Despite these improvements, many questions remain surrounding the functions of the nORF and ORFan genes that can only be answered with additional nucleomorph genomic data from other closely related cryptophyte species, nuclear genomic data from red algae, and more detailed biochemical knowledge of the processes taking place within the periplastidial compartment.

Spliceosomal Introns

The number of genes related to spliceosomal introns varies across the four cryptophyte nucleomorph genomes. Most notably, the nucleomorph genome of *H. andersenii* is completely devoid of spliceosomal introns and genes dedicated to intron removal (Lane et al. 2007). Nevertheless, the *H. andersenii* genome retains divergent homologs of a few genes (*prl1-like*, *snu13*, *cdc28*, *snrpD*, and *snrpD2*) whose protein products function in mRNA splicing in other organisms. The nucleomorph genomes of *C. paramecium*, which contains only two spliceosomal introns (62 bp and 100 bp in length), and *G. theta*, which contains 17 spliceosomal introns (42–52 bp in length) possess 17 and 15 "spliceosomal" genes, respectively (supplementary table S1, Supplementary Material online). In comparison, the *C. mesostigmatica* nucleomorph genome is predicted to possess 24 spliceosomal introns (fig. 1 and supplementary table S2, Supplementary Material online), seven more than *G. theta*, and possesses 28 splicing-related protein genes, almost double the number found in *G. theta*. The lengths of the *C. mesostigmatica* nucleomorph spliceosomal introns are longer on average, ranging from 50 to 211 bp in size (supplementary table S2, Supplementary Material online).

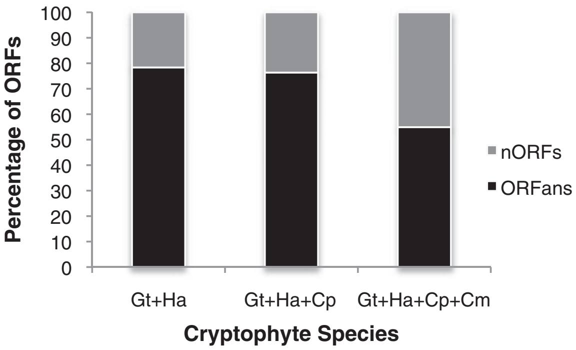

Fig. 5.—Hypothetical proteins inferred from complete cryptophyte nucleomorph genomes. The graph shows the proportion of cryptophyte nucleomorph-specific hypothetical protein-coding genes (nORFs) and hypothetical protein-coding genes unique to each individual nucleomorph genome (ORFans) relative to the total number of hypothetical protein-coding genes as additional nucleomorph genomic sequences become available. The leftmost bar compares the proportions of the two types of hypothetical proteins for *Guillardia theta* and *Hemiselmis andersenii*. The second bar compares these proportions with the addition of *Chroomonas paramecium*. The proportions do not change substantially until the addition of the fourth genome, *C. mesostigmatica*, shown in the third bar, where the proportion of ORFan genes drops by 21%.

We used RT-PCR, cloning, and cDNA sequencing to confirm the splicing of four nucleomorph introns in *C. mesostigmatica* and to verify their predicted intron–exon boundaries: *rps23*, *rps24*, *rpl9*, and *rps16* were shown to have introns of 114, 114, 77, and 57 bp, respectively (fig. 6). In addition, we used RNA-Seq data to verify the correct boundaries and removal of spliceosomal introns from *rpl18A*, *rpl19*, *rpl24*, *rpl26*, *rpl27A*, *rps9*, *rps13*, *rps25*, *rps28*, *orf65*, and *orf102*. Interestingly, we observed a case of mis splicing of a 57 bp intron in the ribosomal protein gene *rpl26* by an alternate 3′-intron boundary. Use of the alternate 3′-splice site results in a substantially truncated, and presumably nonfunctional, protein of only 53 amino acid residues (the full-length protein is predicted to be 124 amino acids). A second RNA-Seq-derived contig shows the correct removal of the intron using the predicted GT/AG intron boundaries producing a full-length *rpl26* mRNA. Similar to the nucleomorph spliceosomal introns of *G. theta* (Douglas et al. 2001), most of the introns are present at the extreme 5′-end of the genes, several immediately following the start codon, and contain a 5′-GTAA GT consensus motif.

Examination of the distribution of nucleomorph spliceosomal introns across the tree of cryptophytes reveals a very clear pattern. The cryptophytes branch into five major clades: 1) *Chroomonas, Hemiselmis,* and *Komma*, with the *Hemiselmis*

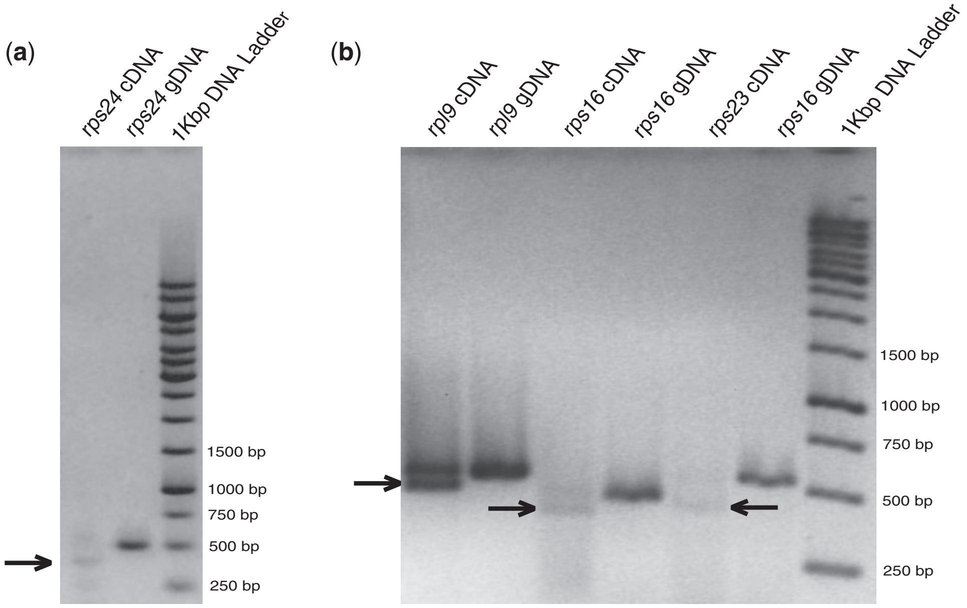

Fig. 6.—Verification of spliceosomal intron removal in the *Chroomonas mesostigmatica* nucleomorph genome. The figure shows agarose gel electrophoresis of PCR amplicons generated using cDNA and genomic DNA template and site-specific primers. Genes examined were (*a*) *rps24* and (*b*) *rpl9*, *rps16*, and *rps23*. The cDNA amplicons (indicated by arrowheads) for each gene are shorter in length compared with their respective PCR-generated genomic DNA amplicons. Intron removal was verified by sequencing.

species branching from within the *Chroomonas* clade; 2) *Guillardia* and *Hanusia*; 3) *Cryptomonas*; 4) *Geminigera*, *Plagioselmis*, and *Teleaulax*; and 5) *Rhinomonas*, *Rhodomonas*, and *Storeatula* (Hoef-Emden 2008). Complete nucleomorph genome sequences are now available for members of clades 1, 2, and 3, and so we can infer that spliceosomal introns are present in all three clades. There are no nucleomorph genome sequences available for members of clade 4, which is a poorly understood group. Most sequences in current databases for members of this group represent rRNA gene sequences amplified from environmental samples; few members are available in culture. Partial nucleomorph genomic data for a member of clade 5, *Rhodomonas* sp. CCMP1178, show that spliceosomal introns and the machinery required for their splicing are indeed present. These include a 51-bp spliceosomal intron present in a gene for one of the proteasome subunits, *prsA7* (JX515791), and a 76-bp spliceosomal intron present in the regulator of epidermal growth factor gene, *ebi* (JX515790). The 5′-splice sites for these two introns have the 5′-GTAAGT consensus motif observed in the spliceosomal introns in the *G. theta, C. paramecium,* and *C. mesostigmatica* nucleomorph genomes, and additionally, these two genes in the *G. theta* nucleomorph genome also contain spliceosomal introns. Also present in the *Rhodomonas* sp. CCMP1178 nucleomorph genomic data is the large and highly conserved spliceosomal protein gene *prp8* (JX515789), whose protein product performs a key role in the catalytic core of the spliceosome (Grainger and Beggs 2005) and is present in all spliceosomal intron-containing nucleomorph genomes sequenced to date (Douglas et al. 2001; Gilson et al. 2006; Lane et al. 2007; Tanifuji et al. 2011). Given that spliceosomal introns and the machinery required for their removal are present in the nucleomorphs of all clades examined thus far, coupled with the fact that spliceosomal introns are absent in the nucleomorph genome of *H. andersenii* yet present in the nucleomorph genome of its close relative, *C. mesostigmatica*, it seems likely that the complete loss of spliceosomal introns occurred somewhere within the *Hemiselmis* clade.

Synteny and Recombination

We have shown that gene order conservation, or synteny, can be a helpful feature for annotating nucleomorph genomes. For example, the existence of an ORFan or nORF in one genome in the same position as an evolutionarily conserved ORF in another genome allows us to predict the identity of the ORFan/nORF. As in other reduced eukaryotic genomes such as those of microsporidian parasites (e.g., Slamovits et al. 2004), the highly compact nature of nucleomorph genomes has been suggested to represent a barrier to the frequent recombination seen in "typical" nuclear genomes; intergenic regions are very short, and most genes are single copy, making recombination-mediated disruption of an ORF probable and likely to be deleterious (Archibald and Lane 2009). We have found that although the degree of synteny is certainly highest between the *C. mesostigmatica* nucleomorph genome and that of its closest relative, *H. andersenii*, the lengths of the syntenic blocks are noticeably shorter compared with those shared between *H. andersenii*, *C. paramecium*, and *G. theta*. The average number of genes within a syntenic region, defined as a stretch of four or more homologous genes (not including nORFs), between *C. mesostigmatica* and *H. andersenii, C. paramecium,* and *G. theta* is 9.0 ($n = 36$), 7.1 ($n = 34$), and 6.7 ($n = 29$), respectively. In comparison, the average number of genes in syntenic regions between *H. andersenii* and *C. paramecium* is 19.4 ($n = 18$), 9.4 ($n = 34$) between *H. andersenii* and *G. theta*, and 10.1 ($n = 27$) between *C. paramecium* and *G. theta*. The more highly "scrambled" nature of the *C. mesostigmatica* nucleomorph genome is presumably due to the fact that it contains longer intergenic regions, making viable intra- and interchromosomal recombination events more likely. Interestingly, many apparent disruptions of syntenic blocks in the *C. mesostigmatica* genome occur where proteasome subunit genes are presumed to have been present, based on their conserved gene order in the other nucleomorph genomes. This observation suggests a recombination-mediated mechanism for gene loss, at least in the case of the missing proteasome subunit genes. There is one instance where a small hypothetical ORF of 59 aa is present in *C. mesostigmatica* within a syntenic region whose counterpart in the other three cryptophyte nucleomorph genomes is a much larger ORF encoding the prsB5 proteasome subunit: 219 aa in *H. andersenii* (fig. 7*a*), 234 aa in *C. paramecium*, and 205 aa *in G. theta*.

Even more prevalent are instances of a single large ORF in the *C. mesostigmatica* nucleomorph genome showing positional synteny with one or more small ORFs in *H. andersenii, C. paramecium,* and *G. theta*. For example, there is one ORF of 42 amino acids in length and two small ORFs of 77 and 65 amino acids in length in the *H. andersenii* genome that occupy the same positions as the putative spliceosomal genes *prp2-like* and *prp4-like*, respectively, in the *C. mesostigmatica* nucleomorph genome (fig. 7*b*). Similarly, there are small ORFs of 55 aa, 61 aa, 62 aa, and 57 aa in *C. paramecium* in syntenic position with the conserved plastid-associated gene *orf268*, hypothetical protein-coding gene *orf425*, the histone chaperone gene *hira*, and the thylakoid assembly protein gene *tha4* (there is also a 75 aa ORF occupying this position in the *G. theta* nucleomorph genome), respectively, in *C. mesostigmatica* (data not shown).

There are also a few instances where the corresponding syntenic position in one of the other cryptophyte nucleomorph genomes to that of an ORF in *C. mesostigmatica* is occupied by several ORFs whose sizes sum to be similar in length to the single syntenic ORF in *C. mesostigmatica*. For example, there are three ORFan genes that total 1,122 aa (orf450, orf271, and orf401) in *C. paramecium* that occupy the same syntenic position of the 1,156 aa mRNA splicing

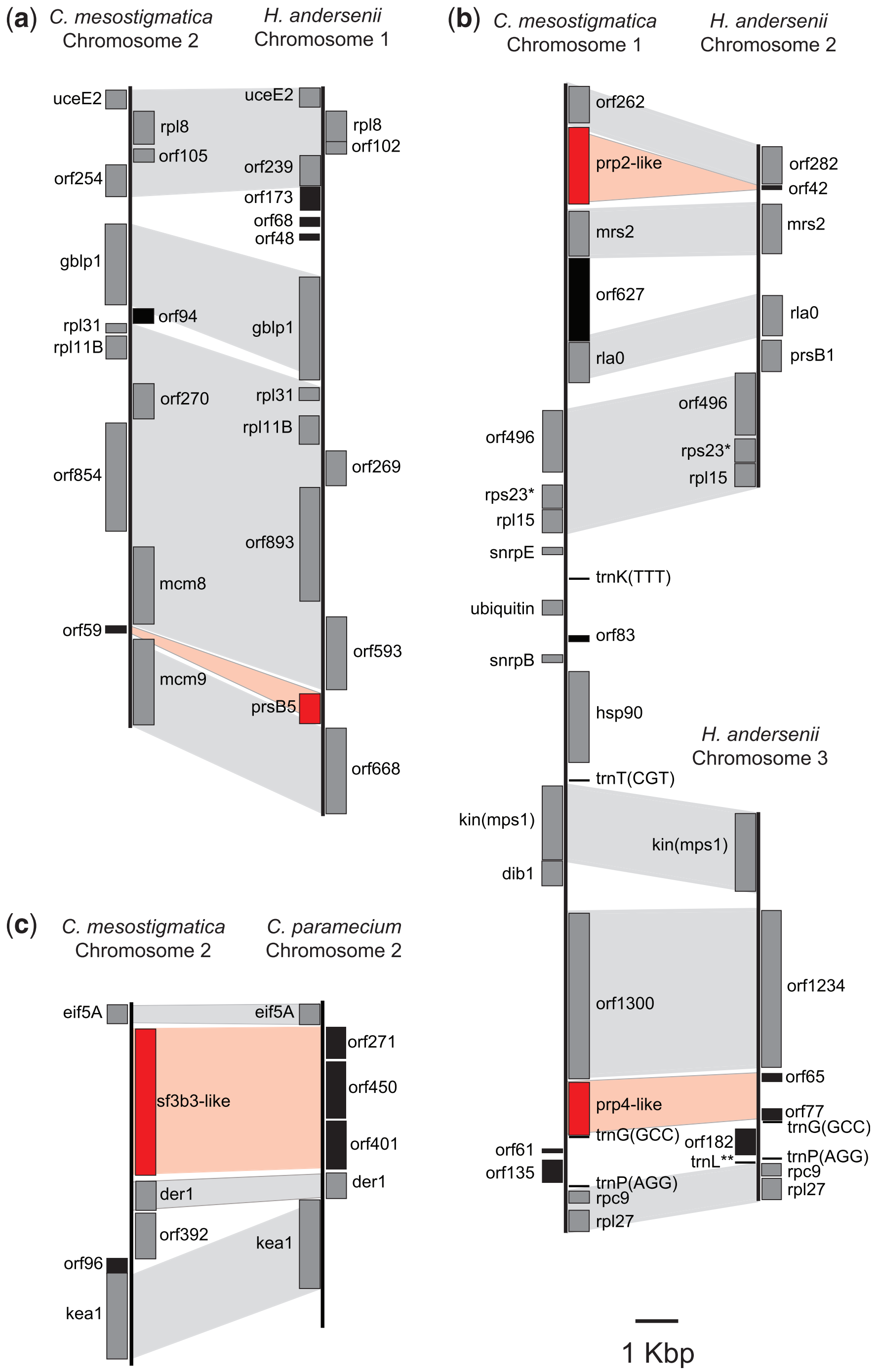

FIG. 7.—ORF degradation in cryptophyte nucleomorph genomes. Schematic shows degenerating ORFs in syntenic regions between *C. mesostigmatica*, *H. andersenii,* and *C. paramecium*. Homologous genes are shown in gray, with gray highlights indicating the syntenic positions of the genes on the chromosome of each species. Genes shown in black are ORFan genes. Genes shown in red are those where one or more ORFan genes occupy the same syntenic position in a stretch of genes that have conserved order in another nucleomorph genome, which are highlighted in red. (*a*) An ORFan gene of 59 amino acids in *C. mesostigmatica* occupies the same syntenic position as the proteasome subunit gene *prsB5* in *H. andersenii*. (*b*) ORFan genes occupy the same syntenic positions in *H. andersenii* as the spliceosomal genes *prp2-like* and *prp4-like* in *C. mesostigmatica*. (*c*) Three ORFan genes on chromosome two in *C. paramecium* occupy the same syntenic position and sum to be a similar size as the splicing factor gene *sf3b3-like* in *C. mesostigmatica*.

factor *sf3b3-like* in *C. mesostigmatica* (fig. 7c). Similarly, there are some smaller ORFs occupying the same syntenic position in *C. paramecium* as the splicing factor gene *sf3b1-like* in *C. mesostigmatica* (data not shown). The *C. paramecium* nucleomorph genome only contains two spliceosomal introns, and it appears as though the splicing factor genes are deteriorating: the small syntenic hypothetical protein-coding genes are presumably remnants of genes that are no longer functional, as they have been reduced through mutation and purging of intergenic sequence. Whether these genes have been transferred to the host nucleus or simply lost can only be determined with complete nuclear genomes for these organisms. The *G. theta* nuclear genome sequence (http://www.jgi.doe.gov/sequencing/why/50026.html, last accessed June 18, 2012) will help to answer questions about the tempo and mode of endosymbiotic gene transfer.

In conclusion, the complete nucleomorph genome sequence of *C. mesostigmatica* has provided valuable insight into the factors contributing to cryptophyte nucleomorph size diversity, genome biology, and their evolutionary fate. We have identified several factors that contribute to the size variation in the nucleomorph genomes observed, including slight differences in the lengths of protein-coding genes and the total number of genes, but most notably differences in the lengths of the intergenic spacers. In contrast to the other cryptophyte nucleomorph genomes, the nucleomorph genome of *C. mesostigmatica* contains numerous (and larger) spliceosomal introns, more multicopy genes, a lower degree of synteny, and repetitive regions. These features make the *C. mesostigmatica* nucleomorph genome the most complex nucleomorph genome studied to date, exhibiting features more characteristic of its presumed free-living red algal relatives. The presence of spliceosomal introns in the *C. mesostigmatica* nucleomorph genome, yet their absence in the nucleomorph genome of its close relative *H. andersenii*, means that we can now pinpoint the complete loss of nucleomorph spliceosomal introns to somewhere within the genus *Hemiselmis*. Additional nucleomorph genome data from *Chroomonas* and *Hemiselmis* species will help to determine the "when" and "how" of spliceosomal intron loss; there have been no other reports of spliceosomal intron loss in any other nuclear genome to date.

Although nucleomorph genes tend to be highly divergent compared with their counterparts in free-living relatives, the nucleomorph genes of *C. mesostigmatica* are more conservative in nature than those in the other cryptophyte nucleomorph genomes sequenced thus far, allowing for additional functional annotation of previously annotated hypothetical protein-coding genes. Furthermore, our comparative analyses show that there is a more highly conserved core of genes present in cryptophyte nucleomorph genomes than previously thought, including an ultraconserved set of plastid-associated genes. Beyond this highly conserved core, however, there is lineage-specific gene loss: spliceosomal genes in *Hemiselmis,* plastid-associated genes in *Cryptomonas*, and proteasome genes in *Chroomonas*. Our synteny analysis has shown apparent genome decay of some of these genes, indicating that nucleomorph genome reduction in the cryptophytes has not yet reached an endpoint. Nucleomorph genomes are valuable models for studying genome reduction and compaction in eukaryotes. As we have shown, much can be learned from nucleomorph comparative genomics studies of more closely related cryptophyte species.

Supplementary Material

Supplementary tables S1 and S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Naoko Onodera for technical assistance with PFGE protocols and genome sequence submission. This work was supported by an operating grant from the Canadian Institutes of Health Research (CIHR) and the Nova Scotia Regional Partnership Program [grant number ROP85016 to J.M.A.]. J.M.A. also acknowledges support from the CIHR New Investigator Award Program, the Canadian Institute for Advanced Research (Program in Integrated Microbial Biodiversity), the Tula Foundation, and the Centre for Comparative Genomics and Evolutionary Bioinformatics at Dalhousie University. C.E.M. is holder of a Doctoral Student Award from the Natural Sciences and Engineering Research Council of Canada.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

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

Archibald JM, Lane CE. 2009. Going, going, not quite gone: nucleomorphs as a case study in nuclear genome reduction. J Hered. 100: 582–590.

Bolte K, et al. 2011. Making new out of old: Recycling and modification of an ancient protein translocation system during eukaryotic evolution. Bioessays 33:368–376.

Bonfield JK, Smith KF, Staden R. 1995. A new DNA sequence assembly program. Nucleic Acids Res. 23:4992–4999.

Cao K, Nakajima R, Meyer HH, Zheng Y. 2003. The AAA-ATPase Cdc48/ p97 regulates spindle disassembly at the end of mitosis. Cell 115: 355–367.

Cavalier-Smith T. 2000. Membrane heredity and early chloroplast evolution. Trends Plant Sci. 5:174–182.

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

Conaway RC, Brower CS, Conaway JW. 2002. Emerging roles of ubiquitin in transcription regulation. Science 296:1254–1258.

Douglas SE, Penny SL. 1999. The plastid genome of the cryptophyte alga, *Guillardia theta*: complete sequence and conserved synteny groups confirm its common ancestry with red algae. J Mol Evol. 48:236–244.

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

Eschbach S, Hofmann CJ, Maier UG, Sitte P, Hansmann P. 1991. A eukaryotic genome of 660 kb: electrophoretic karyotype of nucleomorph and cell nucleus of the cryptomonad alga *Pyrenomonas salina*. Nucleic Acids Res. 19:1779–1781.

Finley D, Chau V. 1991. Ubiquitination. Annu Rev Cell Biol. 7:25–69.

Gardner RG, Nelson ZW, Gottschling DE. 2005. Ubp10/Dot4p regulates the persistence of ubiquitinated histone H2B: distinct roles in telomeric silencing and general chromatin. Mol Cell Biol. 25:6123–6139.

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

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

Gould SB, et al. 2006a. Protein targeting into the complex plastids of cryptophytes. J Mol Evol. 62:674–681.

Gould SB, et al. 2006b. Nucleus-to-nucleus gene transfer and protein retargeting into a remnant cytoplasm of cryptophytes and diatoms. Mol Biol Evol. 23:2413–2422.

Grainger RJ, Beggs JD. 2005. Prp8 protein: at the heart of the spliceosome. RNA 11:533–557.

Gray MW, Schnare MN. 1990. Evolution of the modular structure of rRNA. In: Hill WE, et al. editors. The ribosome: structure, function and evolution. Washington: American Society for Microbiology. p. 589–597.

Hoef-Emden K. 2008. Molecular phylogeny of the phycocyanin-containing cryptophytes: evolution of biliproteins and geographical distribution. J Phycol. 44:985–993.

Ishida K, Endo H, Koike S. 2011. *Partenskyella glossopodia* (Chlorarachniophyceae) possesses a nucleomorph genome of approximately 1 Mbp. Phycol Res. 59:120–122.

Jentsch S, McGrath JP, Varshavsky A. 1987. The yeast DNA repair gene RAD6 encodes a ubiquitin-conjugating enzyme. Nature 329:131–134.

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

Lane CE, Archibald JM. 2006. Novel nucleomorph genome architecture in the cryptomonad genus *Hemiselmis*. J Eukaryot Microbiol. 53: 515–521.

Lane CE, et al. 2006. Insight into the diversity and evolution of the cryptomonad nucleomorph genome. Mol Biol Evol. 23:856–865.

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 U S A. 104:19908–19913.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25: 955–964.

Moore CE, Archibald JM. 2009. Nucleomorph genomes. Annu Rev Genet. 43:251–264.

Patron NJ, Rogers MB, Keeling PJ. 2006. Comparative rates of evolution in endosymbiotic nuclear genomes. BMC Evol Biol. 6:46.

Phipps KD, Donaher NA, Lane CE, Archibald JM. 2008. Nucleomorph karyotype diversity in the freshwater cryptophyte genus *Cryptomonas*. J Phycol. 44:11–14.

Reyes-Prieto A, Weber APM, Bhattacharya D. 2007. The origin and establishment of the plastid in algae and plants. Annu Rev Genet. 41: 147–168.

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.

Rutherford K, et al. 2000. Artemis: sequence visualization and annotation. Bioinformatics 16:944–945.

Silver TD, et al. 2007. Phylogeny and nucleomorph karyotype diversity of chlorarachniophyte algae. J Eukaryot Microbiol. 54:403–410.

Slamovits CH, Fast NM, Law JS, Keeling PJ. 2004. Genome compaction and stability in microsporidian intracellular parasites. Curr Biol. 14: 891–896.

Tanifuji G, Onodera NT, Hara Y. 2010. Nucleomorph genome diversity and its phylogenetic implications in cryptomonad algae. Phycol Res. 58: 230–237.

Tanifuji G, et al. 2011. Complete nucleomorph genome sequences of the nonphotosynthetic alga *Cryptomonas paramecium* reveals a core nucleomorph gene set. Genome Biol Evol. 3:44–54.

Yoon HS, Hackett JD, Pinto G, Bhattacharya D. 2002. The single, ancient origin of chromist plastids. Proc Natl Acad Sci U S A. 99: 15507–15512.

Zauner S, et al. 2000. Chloroplast protein and centrosomal genes, a tRNA intron, and odd telomeres in an unusually compact eukaryotic genome, the cryptomonad nucleomorph. Proc Natl Acad Sci U S A. 97:200–205.

Associate editor: Geoff McFadden