# Chloroplast protein and centrosomal genes, a tRNA intron, and odd telomeres in an unusually compact eukaryotic genome, the cryptomonad nucleomorph

**Stefan Zauner*, Martin Fraunholz*, Jürgen Wastl*, Susanne Penny†, Margaret Beaton‡§, Thomas Cavalier-Smith‡¶, Uwe-G. Maier*‖, and Susan Douglas†**

*Cell Biology and Applied Botany, Philipps-University Marburg, Karl-von-Frisch-Strasse, D-35032 Marburg, Germany; †National Research Council of Canada Institute for Marine Biosciences and Program in Evolutionary Biology, Canadian Institute of Advanced Research, 1411 Oxford Street, Halifax, NS, Canada B3H 3ZI; and ‡Department of Botany, University of British Columbia and Program in Evolutionary Biology, Canadian Institute for Advanced Research, #3529–6270 University Boulevard, Vancouver, BC, Canada V6T 1Z4

Communicated by Michael Smith, University of British Columbia, Vancouver, Canada, October 29, 1999 (received for review January 4, 1999)

**Cells of several major algal groups are evolutionary chimeras of two radically different eukaryotic cells. Most of these "cells within cells" lost the nucleus of the former algal endosymbiont. But after hundreds of millions of years cryptomonads still retain the nucleus of their former red algal endosymbiont as a tiny relict organelle, the nucleomorph, which has three minute linear chromosomes, but their function and the nature of their ends have been unclear. We report extensive cryptomonad nucleomorph sequences (68.5 kb), from one end of each of the three chromosomes of *Guillardia theta.* Telomeres of the nucleomorph chromosomes differ dramatically from those of other eukaryotes, being repeats of the 23-mer sequence $(AG)_7AAG_6A$, not a typical hexamer (commonly TTAGGG). The subterminal regions comprising the rRNA cistrons and one protein-coding gene are exactly repeated at all three chromosome ends. Gene density (one per 0.8 kb) is the highest for any cellular genome. None of the 38 protein-coding genes has spliceosomal introns, in marked contrast to the chlorarachniophyte nucleomorph. Most identified nucleomorph genes are for gene expression or protein degradation; histone, tubulin, and putatively centrosomal *ranbpm* genes are probably important for chromosome segregation. No genes for primary or secondary metabolism have been found. Two of the three tRNA genes have introns, one in a hitherto undescribed location. Intergenic regions are exceptionally short; three genes transcribed by two different RNA polymerases overlap their neighbors. The reported sequences encode two essential chloroplast proteins, FtsZ and rubredoxin, thus explaining why cryptomonad nucleomorphs persist.**

Nucleomorphs are vestigial nuclei of eukaryotic endosymbionts that were sequestered by phagocytic host cells in a process known as secondary endosymbiosis and retained for the purpose of photosynthesis (1–3). They have been reported in only two groups of organisms: cryptomonads (4) and chlorarachniophytes (5–7). Phylogenetic analyses of rRNA genes (8–11) have demonstrated that cryptomonads and chlorarachniophytes arose by two independent secondary endosymbiotic events, from red and green algae, respectively. Despite their separate origins, nucleomorphs of both groups contain three small chromosomes each encoding rRNA (7, 12). These two natural experiments in eukaryotic genome miniaturization potentially offer important insights into many basic features of nuclear genome organization and function (13). Comparisons of their gene complement, architecture, and expression with each other, and with the much larger genomes of other eukaryotes, also will provide fascinating insights into the evolution of these complex cells, including the raison d'être for the nucleomorph in only two of the groups of organisms that acquired their plastids by secondary endosymbiosis (3).

Sequence analysis of 13 kb of the nucleomorph genome of an unnamed chlorarachniophyte revealed a highly compact structure with overlapping genes (some cotranscribed) and tiny spliceosomal introns in all seven protein genes (14). However, few details of the larger cryptomonad nucleomorph genome are known (15). Nucleomorph DNA of cryptomonads coexists in the cell with three other genomes and is only 0.1% of total cellular DNA, making it extremely difficult to recover nucleomorph DNA-containing recombinants from libraries constructed from unfractionated cellular DNA. Purification of nucleomorph chromosomes is possible on a small scale by pulse-field gel electrophoresis (9), allowing the recovery of small amounts of DNA from which we have cloned and sequenced some genes from the approximately 555-kb nucleomorph genome of the cryptomonad *Guillardia theta*. By using these authentic nucleomorph clones to screen a total cell DNA library we have obtained additional nucleomorph sequences by chromosome walking and by random sequencing of nucleomorph-enriched libraries. We report here the unique features revealed by long contiguous sequences from one end of all three nucleomorph chromosomes (20.600 kb from chromosome I; 30.968 kb from chromosome II; 16.932 kb from chromosome III; about 68.5 kb in all, over 12% of the genome).

Like the chlorarachniophyte nucleomorph (14), this vestigial eukaryotic genome proves to be exceptionally compact and gene-rich, with occasional overlapping genes and short inverted repeats containing rRNA cistrons at its chromosome ends. The most striking difference is a total absence of spliceosomal introns in the cryptomonad nucleomorph sequences reported here, which makes gene identification much easier than in the chlorarachniophyte nucleomorph genome that is peppered with minute introns (14). Compared with the yeast genome the fraction of protein genes with no assigned function is low, implying that well-conserved genes of established function were selectively retained during genome miniaturization, and confirming that the cryptomonad nucleomorph genome is indeed a concentrated repository of genes fundamentally important for eukaryotic gene expression (13).

## Materials and Methods

**Libraries.** Total *G. theta* (CCMP327) genomic DNA was prepared by using the cetyltrimethylammonium bromide method (16). Genomic DNA libraries were generated by partial digestion with *Sau*3AI and cloning into λEMBL3, or by partial digestion and fill-in reaction to generate *Xho*I-compatible ends for successive ligation into the *Xho*I site of λFIXII (Stratagene). cDNA libraries were produced by using the Amersham Pharmacia TimeSaver cDNA-synthesis kit accord-

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AF083031, AJ010592, and AF165818).

§Present address: Biology Department, Mount Allison University, Sackville, NB, Canada, E4L 1G7.

¶Present address: Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, United Kingdom.

‖To whom reprint requests should be addressed. E-mail: maier@mailer.uni-marburg.de.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

ing to their instructions. cDNA molecules were ligated in Stratagene's λZAPII vector and subsequently packaged with GigaPack-Gold III packaging extracts. A 300-bp segment unique to the SSU rDNA of the nucleomorph genome (17) was used as a probe to identify a phage containing nucleomorph DNA. This phage served as a starting point for chromosome walking.

Pulse-field gel electrophoresis, shotgun cloning, and hybridizations were performed as described (18).

**Telomere Cloning and Mapping.** Nucleomorph DNA isolated from pulse-field gels was digested with *Eco*RV, and the ends were filled in and cloned into *Eco*RV-digested pBluescript (Stratagene) as the manufacturer recommended. Random sequencing of clones from the *Eco*RV library yielded a clone that overlapped the rDNA cistron and contained 4½ telomere repeats. PCR amplifications of nucleomorph DNA used an oligonucleotide primer based on the telomere repeat sequence (T; see Fig. 2*a*) and two different primers internal to the rDNA cluster (A: 5′-GTCCATCCCAACATGCT-G-3′ and B: 5′-GGATAAACGGGAGGG-3′). Amplification conditions were: 1 cycle of 30 s at 94°C; 30 cycles of 30 s at 94°C, 30 s at 50°C, 3 min at 72°C. Amplification products were resolved on a 4% NuSieve gel by using a 1-kb ladder (Amersham Pharmacia) as a marker. Southern hybridization was at 58°C using the PCR product TB as probe.

**Intron Mapping and Reverse Transcriptase–PCR.** Reverse transcription of total *G. theta* RNA (7.5 μg) used Superscript II reverse transcriptase (Bethesda Research Laboratories) as the manufacturer recommended and 20 pmol of the $tRNA^{Ser}$ RT primer (5′-ACGGCAAGATTCGAACT-3′). PCR was performed by using 2 μl of the resulting cDNA in a 100 μl reaction containing 5 units of *Taq* polymerase, 0.2 mM dNTPs, and 10 pmol each of the RT primer and the 5′ PCR primer (5′-GCACACGTGGCCGAGTG-3′). Amplification conditions were: 1 cycle of 3 min at 94°C; 30 cycles of 1 min at 94°C, 1 min at 50°C, 1 min at 72°C; 1 cycle of 10 min at 72°C. Amplification products were ligated into a pCR-TOPO vector (Invitrogen) according to the manufacturer's recommendations, and recombinant plasmids were sequenced as described below.

**Sequence Analysis.** Automated sequencing was performed on an ALFexpress sequenator (Amersham Pharmacia) by using CY5-labeled primers and Amersham's Thermosequenase sequencing kit and on an ABI 373A (Perkin–Elmer) by using the AmpliTaqFS dye terminator cycle sequencing ready reaction kit (Perkin–Elmer). Contigs were assembled with the GCG Wisconsin package or SEQUENCHER (Gene Codes, Ann Arbor, MI). Database searches used the BLASTP, BLASTX, and BLASTN algorithms on the National Institutes of Health web page (http://www.ncbi.nih.gov/BLAST), and tRNA searches used tRNAscan-SE at http://genome.wustl.edu/eddy/tRNAscan-SE.

## Results

**Chromosome Organization.** By sequencing clones obtained from nucleomorph DNA eluted from pulse-field gels, as well as others found by chromosome walking using these as probes, we assembled physical maps of one end of each chromosome of *G. theta* (Fig. 1). The terminal region comprising the rRNA cistrons and the genes for the ubiquitin conjugation enzyme (*ubc4*) and the TATA-binding protein (a subunit of the general transcription factor TFIID) is present at the end of chromosomes II and III; chromosome I has the same terminal genes except for the TATA-binding protein. The chromosomal locations of these genes and several nonrepeated genes from each contig were determined by hybridization to blots of pulse-field gels where the three chromosomes are well resolved: chromosome I ≈200 kb, chromosome II ≈180 kb, chromosome III ≈175 kb. Within the terminal repeats, the 4-kb segment lacking identified genes has eight ORFs sized from 50 to 210 aa, not shown on Fig. 1 or included in the overall gene tally as several overlap and we are unsure which are real genes. Repeated structures common in regions of RNA-polymerase I promoters are not detectable adjacent to the rRNA genes. The rest of the sequenced DNA is very densely packed indeed with genes; putative functions for 30 of the 53 ORFs were assigned by BLAST analysis (Table 1). Partial sequences of the other ends of chromosomes II and III (unpublished data) reveal identical repeats (with a *tfIId* gene on chromosome III but not II).

The *G. theta* nucleomorph genome has a very high A/T content (up to 90% in spacers) and, outside the terminal repeats, very short intergenic spacers (average 75 bp). Gene density is one in every 0.8 kb, in agreement with all other internal regions of the nucleomorph genome scanned so far (data not shown). Coding capacity is maximized, as shown by an overlapping gene structure where the 5′ and 3′ ends of *rpc10* appear to be used as coding regions for $tRNA^{Ser}$ and $tRNA^{Gln}$, respectively. Moreover, no spliceosomal introns were detected in any genes.

**Unusual Telomeric Repeats.** The telomere consists of the repeated structure $[(AG)_7AAG_6A]_n$ and is unlike most eukaryotic telomeres (which are usually variations on the core structure TTAGGG, ref. 19) or a chlorarachniophyte nucleomorph telomere (20), TCTAGGG. A separate clone from a random *Bam*HI library of total *G. theta* DNA (containing 10 repeats of the sequence, TTTAGGG) is presumed to be of nuclear origin because it hybridizes only to nuclear chromosomes upon Southern analysis (data not shown).

PCR amplifications of total *G. theta* DNA using a 23-nt telomere primer and two different internal primers (A and B) yielded the expected products of approximately 450 and 1,250 bp, respectively (Fig. 2*b*). A weaker band below that of 450 bp probably represents a product arising from priming one telomere unit in from the chromosome end. Southern analysis of pulse-field gel electrophoresis-resolved nucleomorph DNA using the 1,250-bp telomere-containing PCR product (TB) as probe shows hybridization to all three chromosomes (Fig. 2*c*). Some hybridization to the unresolved nuclear chromosomes is also evident, but it is probably background hybridization to 5S rRNA in the huge quantity of DNA present in this region of the gel. In addition, *Eco*RV, *Sac*I, and *Bam*HI (which cut the chromosomes progressively toward the middle of the rDNA cluster; Fig. 2*a*) yielded bands of increasing sizes (1.65, 2.1 and 3.0 kb, respectively) in Southern analyses with the telomere probe (data not shown). This indicates that there is an average of 14 repeats in the telomeres. A separate experiment with a more specific telomere probe from which all except 32 nt of the 5S rRNA were excluded gave the same results. It is therefore probable that this telomere sequence is present on all three chromosomes.

**A Unique tRNA Intron.** Two of the three tRNA genes, $tRNA^{Ser}$ and $tRNA^{Arg}$, contain introns adjacent to the anticodon loop; they are 7 and 8 nt long, respectively. However, the $tRNA^{Ser}$ gene contains an additional insertion of 10 nt in a unique position in the D loop (Fig. 3*a*). Although introns have been reported in the extra arm and anticodon stem of some archaebacteria (21), we report on an intron in the D loop of any tRNA, which emphasizes the unique nature of some nucleomorph-encoded genes. Reverse transcription–PCR experiments were performed to determine whether this latter insertion is excised correctly or whether this tRNA gene is a pseudogene. Four products at various stages of processing were obtained (Fig. 3*b*), including one containing both introns, intermediates with a single or partial intron, and one containing the mature tRNA with both introns removed. This shows that the $tRNA^{Ser}$ gene indeed contains two introns and that at least some transcripts are accurately spliced yielding functional tRNAs.

Introns in eukaryotic tRNA genes have been reported only in the anticodon loop where they typically form an extended anticodon stem and a 3-nt bulge containing the 3′ splice site. Such a structure

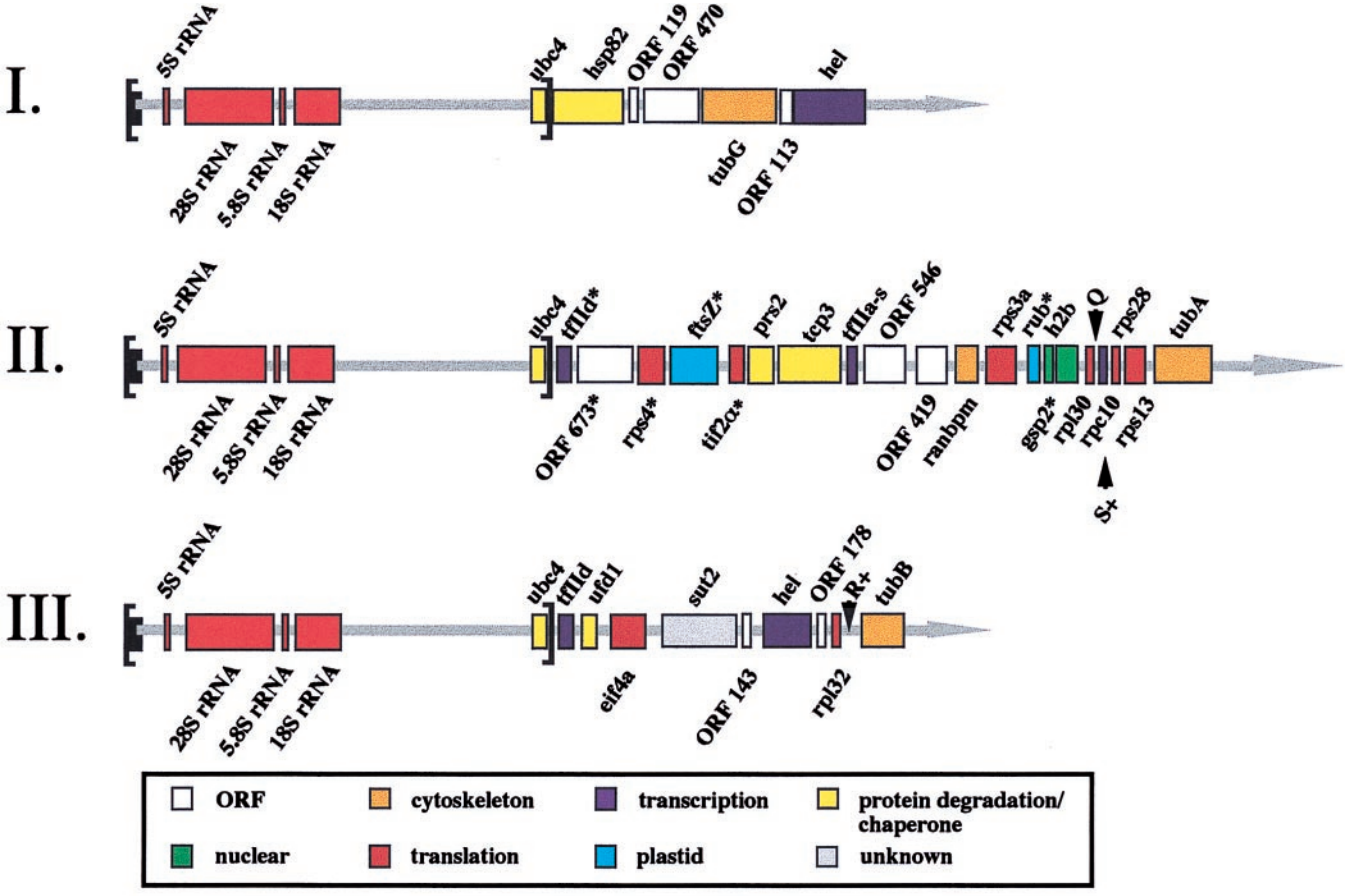


**Fig. 1.** Physical map of terminal regions of nucleomorph chromosomes I, II, and III of the cryptomonad *G. theta*. Genes transcribed from the + strand are indicated above the line and those from the − strand below. ubc4, ubiquitin-conjugating enzyme E2; hsp82, heat shock protein 82; tubG, γ-tubulin; hel, RNA helicase; tfIId, transcription initiation factor TFIID; rps4, 40S ribosomal protein S4; ftsZ, cell division protein FtsZ; tif211, hypothetical translational initiation factor 2 α-subunit; prs2, proteasome IOTA subunit; tcp3, T-complex protein 1, TCP-1-γ; tfIIα, transcription initiation factor IIA γ-chain; ranbpm, centrosomal RAN-binding protein; rps3a, 40S ribosomal protein S3a; rub, electron carrier rubredoxin; h2b, histone H2b; gsp2, GTP-binding nuclear protein RAN; rpl30, 60S ribosomal protein L30; rpc10, 7.7-kDa subunit of DNA-directed RNA polymerases I, II and III; rps28, 40S ribosomal protein S28; rps13, 40S ribosomal protein S13; tubA, α-tubulin; ufd, ubiquitin fusion degradation protein; eif4a, eukaryotic initiation factor 4a;. sut2, sulfate permease; rpl32, 60S ribosomal protein L32; tubB, β-tubulin. Arrowheads show positions of $tRNA^{Gln}$(CTG), $tRNA^{Ser}$(AGA), and $tRNA^{Arg}$(CCT) and + signs indicate the presence of introns. * mark genes from which cDNAs have been isolated.

is also possible in the $tRNA^{Ser}$ intron (Fig. 3*c*). Interestingly, extended base pairing of the D stem is conceivable (Fig. 3*a*, dotted lines), possibly indicating a similar recognition system for intron splicing at this location. Removal of this intron is necessary for the formation of the canonical fourth base pair in the D stem.

**Nuclear Housekeeping Genes.** Protein-coding genes identified include many nucleomorph homologues of nuclear-located proteins: Ran (a GTP-binding protein involved in nuclear import and export: Gsp2 in yeast nomenclature); a subunit common to all three RNA polymerases (Rpc10); two RNA helicases; the TATA-binding protein (a subunit of the general transcription factor TFIID); and the γ subunit of TFIIa (*tfIIa-s*). Interestingly, a histone H2b gene is present (in contrast to immunological studies, refs. 22–24), suggesting that nucleomorph DNA is arranged in nucleosome structures. Other genes involved in nucleosome formation (H3, histone acetyl transferase and histone deacetylase) were detected in other nucleomorph contigs (data not shown).

Surprisingly, genes for tubulins (*tub*A, *tub*B, and *tub*G) are present, although microtubules have never been seen in dividing nucleomorphs (25, 26) or in the periplastid space (the residual cytoplasm of the former red alga that surrounds them, in which only starch, 80S ribosomes and small vesicles are microscopically detectable). If tubulin is involved in nucleomorph chromosome segregation, as we suspect, it is likely to be located within the nucleomorph, because mitotic spindles are largely intranuclear in red algae (27). The presence of a gene for RanBPM, a protein that colocalizes with γ-tubulin (encoded by *tub*G) in centrosomes (28), also implies that nucleomorphs retain a functional centrosome.

**Periplastid Space Proteins.** Elements of the translational machinery, including tRNAs (*trnS*, *trnQ*, and *trnR*), ribosomal proteins (*rps3a*, *rps4*, *rps28*, *rpl13*, *rpl30*, and *rpl32*), and initiation factors (*tif211* and *eif4a*), also are encoded by the nucleomorph, as was expected from the presence of 80S ribosomes in the periplastid space. Additional periplastid processes revealed here are protein degradation, evidenced by components of the proteasome (*prs2*) and the ubiquitin pathway (*ubc4* and *ufd1*), and protein folding involving the T-complex protein (*tcp3*) and a homologue of cytosolic (not ER) chaperonin Hsp 82. Clearly both depend on nucleomorph-encoded products.

**Putative Chloroplast Proteins.** We expected our nucleomorph sequencing project to identify at least one plastid-located but nucleomorph-encoded gene (13). This prediction is borne out: two genes are identified with plastid functions yet neither is encoded on the plastid genome, which is now fully sequenced (29). One encodes FtsZ (30), a prokaryotic cell division protein that is chloroplast-

**Table 1. Protein gene identification by BLASTP analysis**

| Gene name | Chromosome and GenBank coordinate, bp | Gene with best match, species; accession no | Identity, % |
|---|---|---|---|
| ubc4 | I:12264–12707<br>II:12171–12611<br>III:7219–7662 | *Schizosaccharomyces pombe* P46595 | 70 |
| *hsp82* | I:12774–14828 | *Ipomoea nil* P51819 | 69 |
| *tubG* | I:16800–18074 | *Plasmodium* P34787 | 27 |
| *hel* | I:18427–18450 | *Arabidopsis* AAD39319 | 35 |
| *tfIId* | II:13078–13824<br>III:8126–8875 | *Acanthamoeba castellanii* P26354 | 75 |
| *rps4* | II:16832–16062 | *Solanum tuberosum* P46300 | 53 |
| *ftsZ* | II:16900–18096 | *Anabaena sp.* P45482 | 67 |
| *tif211* | II:18910–18362 | *Schizosaccharomyces* P56286 | 32 |
| *prs2* | II:18953–19663 | *Glycine max* AF034572 | 25 |
| *tcp3* | II:19795–21300 | *Mus musculus* P80318 | 33 |
| *tfIIa-s* | II:21330–21713 | *Homo sapiens* P52657 | 25 |
| *ranbpm* | II:25839–25000 | *H. sapiens* NP_005484 | 41 |
| *rps3a* | II:25949–26602 | *H. sapiens* L13802 | 35 |
| *rub* | II:26818–27126 | *Anabaena variabilis* CAB45645 | 48 |
| *h2b* | II:27165–27470 | *Drosophila hydei* S21939 | 66 |
| *gsp2* | II:28138–27500 | *Mus musculus* P28746 | 77 |
| rp130 | II:28489–28169 | *Schizosaccharomyces* P52808 | 58 |
| *rpc10* | II:28790–28572 | *Saccharomyces* P40422 | 35 |
| *rps28* | II:28867–29058 | *Schizosaccharomyces* Q10421 | 75 |
| *rps13* | II:29535–29083 | *Arabidopsis* P49203 | 59 |
| *tubA* | II:29625–30968 | *Chlamydomonas* P09204 | 79 |
| *ufd1* | II:8963–9490 | *Arabidopsis* CAB38813 | 44 |
| *eif4a* | III:10698–9541 | *Schizosaccharomyces* P47934 | 51 |
| *sut2* | III:10763–13015 | *Arabidopsis* AB008782 | 34 |
| *hel* | III:13534–14679 | *Arabidopsis* CAA09199 | 50 |
| *rp132* | III:15603–15244 | *Candida albicans* CAA21942 | 40 |
| *tubB* | III:15727–16932 | *Physarum* P07436 | 78 |

located in *Arabidopsis thaliana* (31) and the moss *Physcomitrella* (32). The second gene encodes rubredoxin (*rub*), an iron-containing electron carrier without acid-labile sulfur. Both genes encode N-terminal extensions of 36 and 46 aa, respectively, relative to their cyanobacterial homologues, which presumably act as transit peptides to direct the polypeptides into the plastid. In addition to *fts*Z and *rub*, a gene for sulfate permease (*sut2*) has been identified that may participate in the transport of sulfate into the plastid. Whether the permease is located in the plastid envelope or the periplastid membrane (the former plasma membrane of the red alga, which originally would have had a sulfate importer) has not been established.

**Gene Expression.** Genes transcribed by RNA polymerase I (three rRNAs), by RNA polymerase II (35 unique protein-encoding genes or ORFs over 50 aa plus extra copies in the terminal duplications) and by RNA polymerase III (5S rRNA, three tRNAs) have been identified.

Whereas genes are very economically arranged in the cryptomonad nucleomorph, the transcriptional system appears to be inaccurate, yielding families of transcripts of varying lengths. By analyzing a set of cDNAs for several genes (indicated by *, Fig. 1), we detected some cDNAs that initiate within the coding sequence of the adjacent upstream gene. In addition, some cDNAs terminate within the spacers between genes, whereas other cDNAs originating from the same gene show transcription continuing into the next gene. Although all analyzed transcripts have a poly(A) tail, termination at the 3′ end appears to be inaccurate; a possible explanation is that the highly AT-rich sequences in nucleomorph DNA might inevitably form several sites resembling poly(A) addition signals.

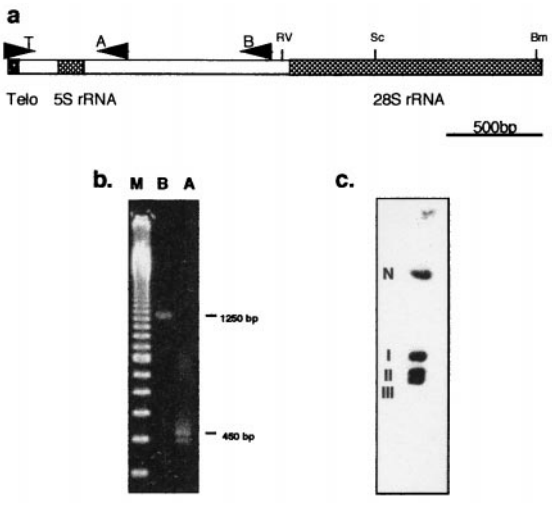


**Fig. 2.** Mapping *G. theta* telomeres. (*a*) Graphic representation of the telomere-containing end of nucleomorph chromosomes. rRNA genes are shaded and restriction sites are marked (RV, *Eco*RV; Sc, *Sac*I; Bm, *Bam*HI). Primers used in PCRs are represented by lettered arrows (T, A, and B). (*b*) Resolution of PCR-amplification products obtained by using primers T/A (A) and T/B (B). Markers (M) are a 1-kb ladder. (*c*) Southern analysis of total *G. theta* DNA resolved by pulse-field gel electrophoresis and hybridized with a telomere probe. The positions of nuclear (N) and nucleomorph chromosomes (I, II, and III) are shown.

## Discussion

The relatively low fraction of ORFs with unassigned functions (32% if the eight ORFs in the terminal repeats are all genes; 21% if none are) compared with other genomes emphasizes the highly conserved, and thus functionally fundamental, character of most genes retained in the cryptomonad nucleomorph genome. Therefore, identifying the gene products and functions of these nucleomorph ORFs is potentially important for understanding basic eukaryotic functions. The identified genes reveal several nucleomorph functions, while the high degree of genomic compaction and the

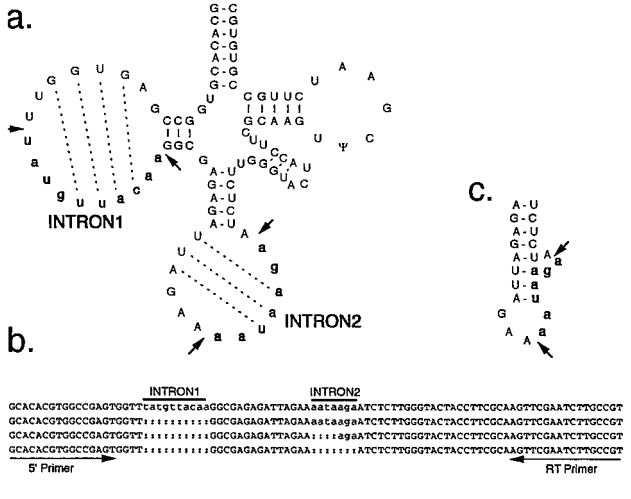


**Fig. 3.** Mapping introns of *G. theta* $tRNA^{Ser}$ gene. (*a*) Secondary structure of *G. theta* $tRNA^{Ser}$ showing introns 1 and 2 in bold lowercase. Intron splice sites are marked by arrows; dotted lines show potential base pairing between the intron and exon sequences. (*b*) Alignment of unspliced, partly spliced, and completely spliced reverse transcription–PCR products. Intron positions are indicated by black bars and the primers used for reverse transcription and PCR by black arrows. (*c*) Putative secondary structure of the extended anticodon stem and bulge structure containing the 3′ splice site.

EVOLUTION

identically repeated chromosome ends are evolutionarily significant.

**Nucleomorph Division and Centrosomal Functions.** The presence of histone genes shows that, even though the nucleomorph is highly miniaturized compared with a typical nucleus, its chromatin probably has the same basic organization; previous attempts to demonstrate histones in nucleomorphs were inconclusive (22–24). The presence of $\alpha$- and $\beta$-tubulin genes, despite the absence of microscopically detectable microtubules during nucleomorph division (25, 26), suggests that this dwarfed nucleus may have some very short and transient microtubules and a form of mitosis, and thus be a paradigm in miniature for understanding chromosome segregation in ordinary nuclei. This is reinforced by our finding a gene for a homologue of RanBPM, a microtubule-nucleating protein only recently discovered in animal centrosomes in association with $\gamma$-tubulin (28); as $\gamma$-tubulin is also nucleomorph-encoded, we suggest that the two proteins interact to form nucleomorph centrosomes, the existence of which was previously unsuspected. Elsewhere (33) it is shown that nucleomorph $\gamma$-tubulin is the most divergent known. This high divergence is much greater than for the $\alpha$- and $\beta$-tubulin genes (33), implying a looser functional constraint on nucleomorph $\gamma$-tubulin, suggesting that the nucleomorph centrosome may be simplified and have fewer interacting proteins than in other eukaryotes; however, the nucleomorph RanBPM is less highly divergent. Phylogenetic analysis shows that nucleomorph $\alpha$- and $\beta$-tubulin genes are indeed derived from the red algal ancestor and are entirely distinct from those of cryptomonad nuclei (33); no molecular information for red algal centrosomal proteins or homologues of the other nucleomorph proteins is available for comparison.

**Genomic Compaction in Nucleomorphs.** Initial data from the nucleomorph genome of a chlorarachniophyte (14), which obtained its chloroplasts and nucleomorph from a green alga in a separate endosymbiosis (10, 11, 34), presents interesting similarities and striking contrasts. Chlorarachniophyte nucleomorphs also have three small chromosomes with rRNA repeats adjacent to the telomeres, but the telomere sequence is distinctly different (20) and the telomere is much further from the rRNA cistron than in the cryptomonad (over 1 kb compared with 150 bp). The absence so far of introns in the protein-encoding genes of the cryptomonad nucleomorph contrasts with the numerous small (18–20 bp) introns in the chlorarachniophyte nucleomorph genome, which also encodes spliceosome components. This distinction may reflect the situation in the nuclear genomes of their ancestral endosymbionts: intron-poor in red algae, intron-rich in green algae. In other respects both nucleomorph genomes are compacted to comparable degrees. In both, gene density is much higher than in other eukaryotic or even bacterial genomes, so selection has successfully reduced noncoding DNA to exceptionally low levels. This ability of selection to eliminate most nonfunctional DNA has important implications for the function of the large amount of noncoding DNA in typical nuclei, as discussed elsewhere (35). The presence of tRNA introns in this highly miniaturized genome emphasizes the extreme difficulty of losing them, even with such strong selection for small genome size, and implies that their splicing proteins are encoded by or imported into the nucleomorph. Whether introns also have been retained in chlorarachniophyte tRNAs is unknown.

Overlapping genes are very rare in eukaryotic genomes and further emphasize the effectiveness of selection for genome reduction and the elimination of almost all noncoding DNA from nucleomorphs. One case of overlapping genes was previously found in the chlorarachniophyte nucleomorph genome (14), but that overlap was in the 3′ untranslated region of the gene not the coding sequence as in the cryptomonad. Transcripts in both organisms are polyadenylated; however, transcription strategies differ. Whereas in the chlorarachniophyte cotranscription occurs (14), cryptomonad mRNAs are inaccurately terminated, leading to transcripts harboring parts of downstream-located genes.

Nucleomorph genome size is variable in different species of cryptomonads (12) and chlorarachniophytes (36), but is on average greater in cryptomonads. The virtual absence of noncoding DNA in the sequences reported here, outside the terminal repeats, makes it probable that cryptomonad nucleomorphs encode more proteins than chlorarachniophyte nucleomorphs. Further comparisons of the nucleomorph genomes of these independently evolved eukaryote/eukaryote chimaeras may reveal general principles underlying genomic miniaturization after secondary symbiogenesis (13), but our initial impression is that the detailed course of evolution has been quite different in each case.

**Why Nucleomorphs Are Kept.** Why have nucleomorphs been retained in cryptomonads and chlorarachniophytes, unlike in the other chimeric products of secondary symbiogenesis, where the nucleus of the algal endosymbiont was lost (3, 37)? As we stressed previously (13), retention of nucleomorphs and their gene expression system is a functional necessity if at least one chloroplast protein gene, originally present in the algal endosymbiont nucleus, was never successfully transferred to the host nucleus and retargeted to the periplastid space. We have now identified two such chloroplast proteins, FtsZ (30) and rubredoxin; apart from these plastid proteins and the sulfate permease, the putative cellular roles of all proteins encoded by cryptomonad nucleomorph genes so far sequenced are limited to nuclear maintenance and transport, translation, protein degradation and folding, and microtubule/centrosome functions.

This spectrum of gene functions strongly supports our view that the cryptomonad nucleomorph and periplastid space are retained for one reason only: to provide a minimal eukaryotic expression apparatus for a remarkably small number of nucleomorph-encoded chloroplast proteins. Partial sequence information from over 90% of the rest of the genome (unpublished data) is consistent with this as it has not revealed any genes for enzymes of primary or secondary metabolism. It is likely that this also will prove true of chlorarachniophyte nucleomorphs; as in cryptomonads, the genes so far identified encode components of the transcription and translation machinery, but no metabolic enzymes. It is likely that the chlorarachniophyte nucleomorph is also kept solely to allow the expression of a very few chloroplast proteins: one nucleomorph gene was suggested to be active in the plastid (the catalytic subunit of ClpP protease; ref. 14); being far more divergent from homologues than are the two genes for putative chloroplast proteins identified here, its identity was less certain.

This rarity of genes for metabolic enzymes contrasts with the reduced genomes of obligate symbionts like *Mycoplasma genitalium* (38) (580 kb versus the 555-kb *G. theta* nucleomorph genome) and *Rickettsia prowazekii* (39) (1,111 kb), in which genes encoding enzymes of primary metabolism are frequent. This fundamental difference arises because these bacteria, though obligate parasites, remain independent organisms, whereas the red alga ancestral to the cryptomonad nucleomorph (3, 8–10) long ago ceased to be a distinct organism. It became a fully integrated part of the novel chimaeric cryptomonad cell by evolving protein-import and metabolite exchange machinery (3, 37). This allowed it to dispense with its own metabolism and lose most of its coding capacity, apart from genes for an expression machinery to make plastid-located proteins encoded by the nucleomorph.

If every chloroplast gene had been transferred to the nucleus, then the nucleomorph itself would have been completely lost, as happened in heterokont, haptophyte, and dinoflagellate algae, which also obtained former red algal plastids by secondary symbiogenesis (3, 10, 37, 40). If all four groups obtained their plastid in the same symbiogenetic event from the same red alga (37), the cryptomonad nucleomorph would have originated more than 500 million years ago, not long after the symbiotic origin of chloroplasts,

dated at about 600 million years ago by comparing the divergence times of these groups on molecular trees with the fossil record (3); even if the cryptomonads obtained their plastids separately (1), their phyletic depth on molecular trees (10) implies an age of over 200 million years. Therefore, although nucleomorphs and the surrounding periplastid compartment with its bounding periplastid membrane originated as eukaryotic endosymbionts, it is inappropriate to view them still as endosymbionts. Instead they are fully integrated cell organelles that were parts of an endosymbiont hundreds of millions of years ago. An organelle differs from an obligate endosymbiont by possessing a protein-import mechanism enabling it to import proteins encoded by the nucleus (41), and thereby dispense with many genes necessary for an autonomous organism. Genomes of mitochondria and chloroplasts, also formerly endosymbionts but now true organelles like nucleomorphs, similarly lost most or all genes for metabolic enzymes, many being transferred to the nucleus and their proteins reimported from the cytosol.

**Concerted Evolution of Chromosome Ends.** The fact that all three nucleomorph chromosomes have identical terminal regions of over 13 kb (at least two at both ends) indicates that they must regularly exchange genetic information, by either physical exchange or gene conversion. The rRNA genes, the internal transcribed spacers, the 4-kb region without identified genes, and the *ubc4* genes are identical in sequence on all three chromosomes. As intergenic regions always evolve rapidly, their identity must be actively maintained in the face of independent mutations on the different chromosomes. In chromosome I the repeated region extends to and includes the termination codon of *ubc4*; in chromosome II and III it extends into the spacer proximal to *tfIId*. The sequences on either side of the boundary between repeated and unique DNA are thus not the same in each case. Physical duplications originally must have created the terminal repeats, but reciprocal exchanges would not maintain homogeneity. The simplest explanation for their virtual identity is frequent gene conversion that normally extends as far as it can toward the unique region; that such gene conversion does not always extend right up to the unique region is suggested by the fact that alignments of all three terminal regions reveal a few single base changes in the most proximal intergenic spacer of each repeat, but none in the more distal spacers. Concerted evolution by gene conversion between separate chromosomes also occurs in dinoflagellate plastid minicircles (40).

The orientation of the rRNA cistrons is opposite to that in chlorarachniophyte nucleomorphs, and a 5S rRNA gene is present only in the cryptomonad rRNA cistron. It will be interesting to determine the structure of chromosome ends in extant algae related to nucleomorphs, namely rhodellophyte red algae (10, 42) for cryptomonads and ulvophycean green algae (34) for chlorarachniophytes, to see whether these differences and similarities simply reflect those found in the ancestral symbionts or whether, instead, the chromosome structure of the two kinds of nucleomorph converged during miniaturization. Cryptomonad nucleomorphs are a good system for comparative studies of the concerted evolution of chromosome termini.

The presence of identical subtelomeric repeats of the same genes on different chromosomes is widespread in eukaryotes as diverse as *Drosophila* (43) and yeast (44). It is not uncommon to have the same genes repeated at both ends of a single chromosome (44), but this is not invariable; for example, in *Giardia duodenalis* (45) rRNA genes are at only one end of the chromosomes.

**Concluding Remark.** A basic question for the future is whether the nucleomorph genome is self-sufficient for its replication, segregation, transcription, translation and structure or also requires, like mitochondria and chloroplasts, the import of some nuclear-encoded proteins.

We thank Dr. Claudia Hofmann for much help at the beginning of the project, Dr. Mark Ragan for critically reading the manuscript, Lang-Tuo Deng and Dr. Xiaonan Wu for technical assistance, and the Canadian Institute for Advanced Research for a Fellowship for T.C.-S. Our research is supported by the Deutsche Forschungsgemeinschaft (Germany) and the Natural Sciences and Engineering Research Council (Canada). This is National Research Council of Canada publication 42273.

1. Palmer, J. D. & Delwiche, C. F. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 7432–7435.
2. McFadden, G. I. & Gilson, P. R. (1995) *Trends Ecol. Evol.* **10,** 12–17.
3. Cavalier-Smith, T. (1995) in *Biodiversity and Evolution*, eds. Arai, R., Kato, M. & Doi, Y. (The National Science Museum Foundation, Tokyo), pp. 75–114.
4. Greenwood, A. D. (1974) in *Electron Microscopy 1974*, eds. Sanders, J. V. & Goodchild, D. J. (Australian Academy of Sciences, Canberra), pp. 566–567.
5. Hibberd, D. J. & Norris, R. E. (1984) *J. Phycol.* **20,** 310–330.
6. Ludwig, M. & Gibbs, S. (1989) *J. Phycol.* **25,** 385–394.
7. McFadden, G. I., Gilson, P. R., Hofmann, C. J. B., Adcock, G. A. & Maier, U.-G. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 3690–3694.
8. Douglas, S. E., Murphy, C. A., Spencer, D. F. & Gray, M. W. (1991) *Nature (London)* **350,** 148–151.
9. Maier, U.-G., Hoffmann, C. J. B., Eschbach, S., Wolters, J. & Igloi, G. (1991) *Mol. Gen. Genet.* **230,** 155–160.
10. Cavalier-Smith, T., Couch, J. A., Thorsteinsen, K. E., Gilson, P. R., Deane, J. A., Hill, D. R. A. & McFadden, G. I. (1996) *Eur. J. Phycol.* **31,** 315–328.
11. Van de Peer, Y., Rensing, S. A., Maier, U.-G. & De Wachter, R. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 7732–7736.
12. Rensing, S. A., Goddemeier, M., Hofmann, C. J. B. & Maier, U.-G. (1994) *Curr. Genet.* **26,** 451–455.
13. McFadden, G. I., Gilson, P. R., Douglas, S. E., Cavalier-Smith, T., Hofmann, C. J. B. & Maier, U.-G. (1997) *Trends Genet.* **13,** 46–49.
14. Gilson, P. R. & McFadden, G. I. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 7737–7742.
15. Gilson, P. R., Maier, U-G. & McFadden, G. I. (1997) *Curr. Opin. Genet. Dev.* **7,** 800–806.
16. Rogers, S. O. & Bendich, A. J. (1985) *Plant Mol. Biol.* **5,** 69–73.
17. McFadden, G. I., Gilson, P. R. & Douglas, S. E. (1994) *J. Cell Sci.* **107,** 649–657.
18. Hofmann, C. J. B., Rensing, S. A., Häuber, M. M., Martin, W. F., Muller, S. B., Couch, J., McFadden, G. I., Igloi, G. L. & Maier, U.-G. (1994) *Mol. Gen. Genet.* **243,** 600–604.
19. König, P. & Rhodes, D. (1997) *Trends Biochem. Sci*. **22,** 42–47.
20. Gilson, P. R. & McFadden, G. I. (1995) *Chromosoma* **103,** 635–641.
21. Belfort, M. & Weiner, A. (1997) *Cell* **89,** 1003–1006.
22. Hansmann, P., Maerz, M. & Sitte, P. (1987) *Endocyt. Cell Res.* **4,** 289–295.
23. Gibbs, S. P. (1990) in *Experimental Phycology 1: Cell Walls and Surfaces, Reproduction, Photosynthesis,* eds. Wiessner, W., Robinson, D. G. & Starr, R. C. (Springer, Heidelberg), pp. 145–157.
24. Müller, S. B., Rensing, S. A. & Maier, U.-G. (1994) *Gene* **150,** 299–302.
25. Morrall, S. & Greenwood, A. D. (1982) *J. Cell Sci.* **54,** 311–318.
26. McKerracher, L. & Gibbs, S. P. (1982) *Can. J. Bot.* **60,** 2440–2452.
27. Scott, J. (1983) *Protoplasma* **118,** 56–70.
28. Nakamura, M., Masuda, H., Horii, J., Kuma, K., Yokoyama, N., Ohba, T., Nishitani, H., Miyata, T., Tanaka, M. & Nishimoto, T. (1998) *J. Cell Biol.* **143,** 1041–1052.
29. Douglas, S. E. & Penny, S. L. (1999) *J. Mol. Evol.* **48,** 236–244.
30. Fraunholz, M. J., Moerschel, E. & Maier, U.-G. (1999) *Mol. Gen. Genet.* **260,** 207–211.
31. Osteryoung, K. W. & Vierling, E. (1995) *Nature (London)* **376,** 473–474.
32. Strepp, R., Scholz, S., Kruse, S., Speth, V. & Reski, R. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 4368–4373.
33. Keeling, P. J., Deane, J. A., Hink-Schauer, C., Douglas, S. E., Maier, U.-G. & McFadden, G. I. (1999) *Mol. Biol. Evol.* **16**, 1308–1313.
34. Ishida, K., Green, B. R. & Cavalier-Smith, T. (1999) *Mol. Biol. Evol.* **16,** 321–331.
35. Cavalier-Smith, T. & Beaton, M. J. (1999) in *Structural Biology and Functional Genomics*, eds. Bradbury, E. M. & Pangoe, S. (Kluwer, Dordrecht, The Netherlands), pp. 1–18.
36. Gilson, P. R. & McFadden, G. I. (1999) *Phycol. Res.,* in press.
37. Cavalier-Smith, T. (1999) *J. Euk. Microbiol.* **46,** 347–366.
38. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G. & Kelley, J. M. (1995) *Science* **270,** 397–403.
39. Andersson, S. V., Zomorodipour, A., Andersson, J. O, Sicheritz-Pontén, T., Alsmark, U. C. M., Podowski, R. M., Näslund, A. K., Eriksson, A., Winkler, H. H. & Kurland, C. G. (1998) *Nature (London)* **396,** 133–140.
40. Zhang, Z. Green, B. R. & Cavalier-Smith, T. (1999) *Nature (London)* **400,** 155–159.
41. Cavalier-Smith, T. & Lee, J. J. (1985) *J. Protozool.* **32,** 376–379.
42. Cavalier-Smith, T. (1998) *Biol. Rev.* **73,** 203–266.
43. Mason, J. M. & Biessmann, H. (1995) *Trends Genet.* **11,** 58–62.
44. Bussey H., Kaback, D. B. & Stroms, R. K. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 3809–3813.
45. Upcroft, P., Chen, N. & Upcroft, J. A. (1997) *Genome Res.* **7,** 37–46.