Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function

Christopher E. Lane*, Krystal van den Heuvel*, Catherine Kozera†, Bruce A. Curtis†, Byron J. Parsons†‡, Sharen Bowman†§, and John M. Archibald*¶

*Canadian Institute for Advanced Research, Integrated Microbial Biodiversity Program, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS, Canada B3H 1X5; §Department of Process Engineering and Applied Science, Dalhousie University, Halifax, NS, Canada B3H 3J5; and †Atlantic Genome Centre, Halifax, NS, Canada B3J 1S5

Edited by Jeffrey D. Palmer, Indiana University, Bloomington, IN, and approved October 29, 2007 (received for review August 6, 2007)

Nucleomorphs are the remnant nuclei of algal endosymbionts that took up residence inside a nonphotosynthetic eukaryotic host. The nucleomorphs of cryptophytes and chlorarachniophytes are derived from red and green algal endosymbionts, respectively, and represent a stunning example of convergent evolution: their genomes have independently been reduced and compacted to <1 megabase pairs (Mbp) in size (the smallest nuclear genomes known) and to a similar three-chromosome architecture. The molecular processes underlying genome reduction and compaction in eukaryotes are largely unknown, as is the impact of reduction/compaction on protein structure and function. Here, we present the complete 0.572-Mbp nucleomorph genome of the cryptophyte *Hemiselmis andersenii* and show that it is completely devoid of spliceosomal introns and genes for splicing RNAs—a case of complete intron loss in a nuclear genome. Comparison of *H. andersenii* proteins to those encoded in the slightly smaller (0.551-Mbp) nucleomorph genome of another cryptophyte, *Guillardia theta*, and to their homologs in the unicellular red alga *Cyanidioschyzon merolae* reveal that (*i*) cryptophyte nucleomorph genomes encode proteins that are significantly smaller than those in their free-living algal ancestors, and (*ii*) the smaller, more compact *G. theta* nucleomorph genome encodes significantly smaller proteins than that of *H. andersenii*. These results indicate that genome compaction can eliminate both coding and noncoding DNA and, consequently, drive the evolution of protein structure and function. Nucleomorph proteins have the potential to reveal the minimal functional units required for basic eukaryotic cellular processes.

endosymbiosis | genome evolution | genome reduction

Nuclear genome size in eukaryotes varies ≈200,000-fold (1). Toward the lower end of this spectrum are the reduced genomes of microorganisms that have become symbionts or intracellular pathogens, such as apicomplexans (e.g., *Plasmodium*, the causative agent of malaria) and microsporidian parasites (e.g., *Encephalitozoon*, an opportunistic pathogen of AIDS patients). The nuclear genomes of these organisms are smaller and more compact than those of their free-living relatives and contain little in the way of repetitive DNA (2). Far and away the most extreme examples of eukaryotic genome reduction are the "nucleomorph" genomes of cryptophytes and chlorarachniophytes. Nucleomorphs are the relic nuclei of algal endosymbionts that became permanent inhabitants of nonphotosynthetic eukaryotic host cells (3–5). Through the combined effects of genome compaction and intracellular gene transfer, the nucleomorph genomes of cryptophytes and chlorarachniophytes have shrunk to a fraction of the size of the algal nuclear genomes from which they are derived and, thus, represent a fascinating system for studying the process of genome evolution.

The first nucleomorph genome to be sequenced was the 551-kilobase pair (kbp) genome of the model cryptophyte, *Guillardia theta* (6). The *G. theta* genome contains 513 genes, primarily with "housekeeping" functions such as transcription, translation, and protein folding/degradation (6). Recently, the nucleomorph genome of the chlorarachniophyte alga *Bigelowiella natans* was completely sequenced and, at 373 kbp (7), is even smaller than that of *G. theta*. Like *G. theta*, the *B. natans* nucleomorph genome is largely composed of genes whose function is to perform core eukaryotic cellular processes and to maintain the expression of a small number of essential genes/proteins involved in photosynthesis (3, 5, 7). A striking similarity between the *G. theta* and *B. natans* nucleomorph genomes is that both are composed of three chromosomes, each with subtelomeric ribosomal DNA (rDNA) cistrons (3, 5). This is intriguing, considering the independent evolutionary history of these organisms: the algal endosymbiont that gave rise to the cryptophyte nucleomorph and plastid (chloroplast) is derived from an ancestor of modern-day red algae, whereas in chlorarachniophytes, the endosymbiont was a green alga (reviewed in refs 8, 9). The observed similarities in basic karyotype and genome structure between the two nucleomorphs are, thus, the result of convergent evolution, the biological significance of which is unknown (3). Importantly, the gene content of the *G. theta* and *B. natans* nucleomorph genomes, in particular the complement of genes for plastid-targeted proteins, are very different from one another, emphasizing the independent evolutionary trajectories taken by the two genomes since their enslavement.

Beyond the cryptophyte *G. theta* and the chlorarachniophyte *B. natans*, very little is known about nucleomorph genome diversity within members of each lineage. Preliminary karyotype diversity studies have revealed considerable size variation, with estimated nucleomorph genome sizes ranging from ≈450 to 845 kbp in cryptophytes and ≈330–610 kbp in chlorarachniophytes (4, 10–14). The presence of three chromosomes is thus far a universal feature of nucleomorph genomes (3, 12, 14), as is the existence of subtelomeric rDNA repeats. An interesting exception was recently discovered within members of the cryptophyte genus *Hemiselmis*, where only three of the six nucleomorph chromosome ends contain intact repeats, the other three containing only the 5S rDNA locus (11). To better understand the

Author contributions: J.M.A. designed research; C.E.L., K.v.d.H., C.K., B.A.C., B.J.P., and S.B. performed research; C.K., B.A.C., B.J.P., and S.B. contributed new reagents/analytic tools; C.E.L., K.v.d.H., B.A.C., and J.M.A. analyzed data; and C.E.L. and J.M.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. CP000881, CP000882, and CP000883).

‡Present address: Ocean Nutrition Canada, 101 Research Drive, Dartmouth, NS, Canada B2Y 4T6.

¶To whom correspondence should be addressed. E-mail: john.archibald@dal.ca.

This article contains supporting information online at www.pnas.org/cgi/content/full/0707419104/DC1.

sequence and structural diversity of nucleomorph genomes and, more generally, the causes and consequences of genome reduction and compaction in eukaryotes, we have completely sequenced the nucleomorph genome of a newly described species, *Hemiselmis andersenii* (15). Detailed comparison of the *H. andersenii* genome to that of *G. theta* (6) provides the first glimpse into the tempo and mode of nucleomorph genome evolution and highlights the significant impact of genome compaction on gene and protein structure.

Results and Discussion

Chromosome and Genome Structure. *H. andersenii* CCMP644 nucleomorph DNA was isolated by using cesium chloride-Hoechst dye density gradient centrifugation, cloned and shotgun sequenced to ≈9× coverage. After the use of long-range PCR to link contigs and fill remaining gaps, three chromosome-sized contigs (207.5, 184.7, and 179.6 kbp) were produced, in agreement with genome size estimates based on pulsed-field gel electrophoresis (11). The complete *H. andersenii* genome is 571,872 bp in size (Figs. 1 and 2*A*) with an overall G+C content of 25.2% (24.7% in single-copy regions, 39.0% in rDNA repeats). The *H. andersenii* nucleomorph chromosome ends are highly unusual: telomeres are composed of a never-before-seen $(GA_{17})_{4-7}$ repeat, in contrast to those in *G. theta* $(([AG]_7AAG_6A)_{11})$. Consistent with previous observations (11), intact subtelomeric rDNA operons are present only on both ends of chromosome I and one end of chromosome III (5S rDNA exists in isolation on chromosome II and one end of chromosome III; Fig. 1).

Loss of Introns and Splicing Machinery. The *G. theta* nucleomorph genome possesses 17 small (42- to 52-bp) spliceosomal introns with standard GT/AG boundaries, primarily in ribosomal protein genes and invariably located at their 5′ ends (6). With the exception of orf183 and orf263, all of the *G. theta* intron-containing genes have homologs in *H. andersenii*. Unexpectedly, none of these contain introns nor do any of the other predicted genes (Fig. 1). Introns are widely considered to be a universal feature of nuclear genomes (16) and are removed by the spliceosome, a large, evolutionarily conserved ribonucleoprotein complex consisting of five small nuclear (sn) RNAs and >50 proteins (17, 18). Although intron density varies greatly, even the most reduced and compacted nuclear genomes examined thus far retain at least a few introns. For example, the genomes of the parasites *Giardia lamblia* (19) and *Encephalitozoon cuniculi* (20) possess 4 and 13 introns, respectively, and encode snRNAs and dozens of core spliceosomal protein components necessary for their removal (19–21).

To gain further insight into the significance of intron loss in the *H. andersenii* nucleomorph genome, we performed a detailed analysis of 51 *G. theta* and/or *H. andersenii* nucleomorph genes with predicted roles in RNA metabolism [supporting information (SI) Fig. 4]. Nineteen of 51 genes have clear functions in ribosome biogenesis (e.g., cbf5, nop56), 17 of which are present in both genomes (U3 snoRNP and brx1 are missing in *G. theta*). Both genomes encode an mRNA capping enzyme (mce), two polyadenylate-binding proteins (pab1,2) and several DExD/H box RNA helicases (e.g., has1, dbp4), which participate in a wide range of RNA-related processes (22). In stark contrast, whereas the *G. theta* genome encodes 13 proteins with known or predicted spliceosomal functions, most notably two U5 snRNP subunits and the large, highly conserved and spliceosome-specific protein prp8, all but four of these are absent in *H. andersenii* (SI Fig. 4). The remaining four proteins are highly divergent snrpD and D2 homologs with weak similarity to two of the seven snRNP-associated protein genes in *G. theta*, cdc28, a DExD/H box helicase whose yeast counterpart (prp2) functions in spliceosome activation (23) and snu13, a protein that functions in both the spliceosome and as part of the rRNA processing machinery (24). Significantly, we were also unable to detect *H. andersenii* genes for any of the five spliceosome-specific snRNAs (U1, U2, U4, U5, and U6; SI Fig. 4), all of which are found in *G. theta* (6). Collectively, these results provide strong evidence for the hypothesis of complete loss of introns and splicing in the *H. andersenii* nucleomorph. Nevertheless, it is formally possible that the missing splicing factors in *H. andersenii* are, in fact, nucleus-encoded and imported to the organelle posttranslationally, as must be the case for many nucleomorph and plastid proteins in cryptophytes (3), although it is not clear what their present functions would be. The *G. theta* nuclear genome is being completely sequenced (www.jgi.doe.gov/sequencing/why/CSP2007/guillardia.html) and it will be possible to assemble a complete "parts list" for the nucleomorph spliceosome in this organism. Assuming that there is indeed no spliceosome in the *H. andersenii* nucleomorph, comparing and contrasting the suite of nucleomorph-localized proteins involved in RNA metabolism in *G. theta* and *H. andersenii* should provide key insight into eukaryotic nuclear proteins whose functions are restricted to splicing and those that are multifunctional.

Genome Synteny and Recombination. A comparison of gene order between the *H. andersenii* and *G. theta* nucleomorph genomes reveals an exceptional degree of synteny. Ninety-four percent of homologous genes (see below) reside within syntenic blocks (Fig. 2*b*), with a relatively small number of intra- and interchromosomal recombinations and inversions having scrambled the two genomes since they diverged from one another. For example, a significant fraction of *H. andersenii* chromosome I corresponds to *G. theta* chromosome III, whereas chromosomes II and III of *H. andersenii* share large blocks of synteny with *G. theta* chromosome II (Fig. 2*b*). Several blocks of synteny are as large as 30 kbp in size and most differ only in organism-specific ORF content (Fig. 1; below). In some cases, such as one end of chromosome I, large portions of the chromosome share gene content with a portion of a *G. theta* chromosome, but these regions are broken into syntenic blocks that have been inverted since the common ancestor of the two genomes.

Compared with prokaryotic and organellar genomes (25–27), gene order in nuclear genomes is typically only conserved between closely related species (28–30). An interesting exception occurs in microsporidian parasites where a recent genomic investigation (31) revealed that their reduced and compacted genomes are unexpectedly stable relative to the fungal genomes from which they evolved, presumably because of a decrease in recombination frequency. Our data suggest that the extreme reduction and compaction that has occurred during cryptophyte evolution has led to an even greater degree of genomic stability in nucleomorphs, on par with that seen in reduced prokaryotes and organellar genomes. Nonhomologous recombination events are likely to disrupt coding sequences in gene-dense nucleomorph genomes, reducing the rate of viable genomic rearrangements and resulting in the retention of large blocks of synteny observed between distantly related cryptophytes. The amount of time since *H. andersenii* and *G. theta* diverged from a common ancestor is not known, but molecular phylogenies reveal that they are not closely related (12, 32), their nucleus- and nucleomorph-encoded rDNAs being only ≈90% and ≈80% identical, respectively.

The nucleomorph chromosomes of both *Guillardia theta* and *Bigellowiella natans* encode substantial subtelomeric repeats, characterized by the presence of rDNA cistrons (6, 7). These repeats are presumably undergoing recombination/conversion at rates high enough to maintain nearly identical sequence. Interestingly, differences in gene content exist at the most internal portions of the repeats in both genomes. The *B. natans* genome encodes a complete copy of the heat shock protein gene dnaK

EVOLUTION

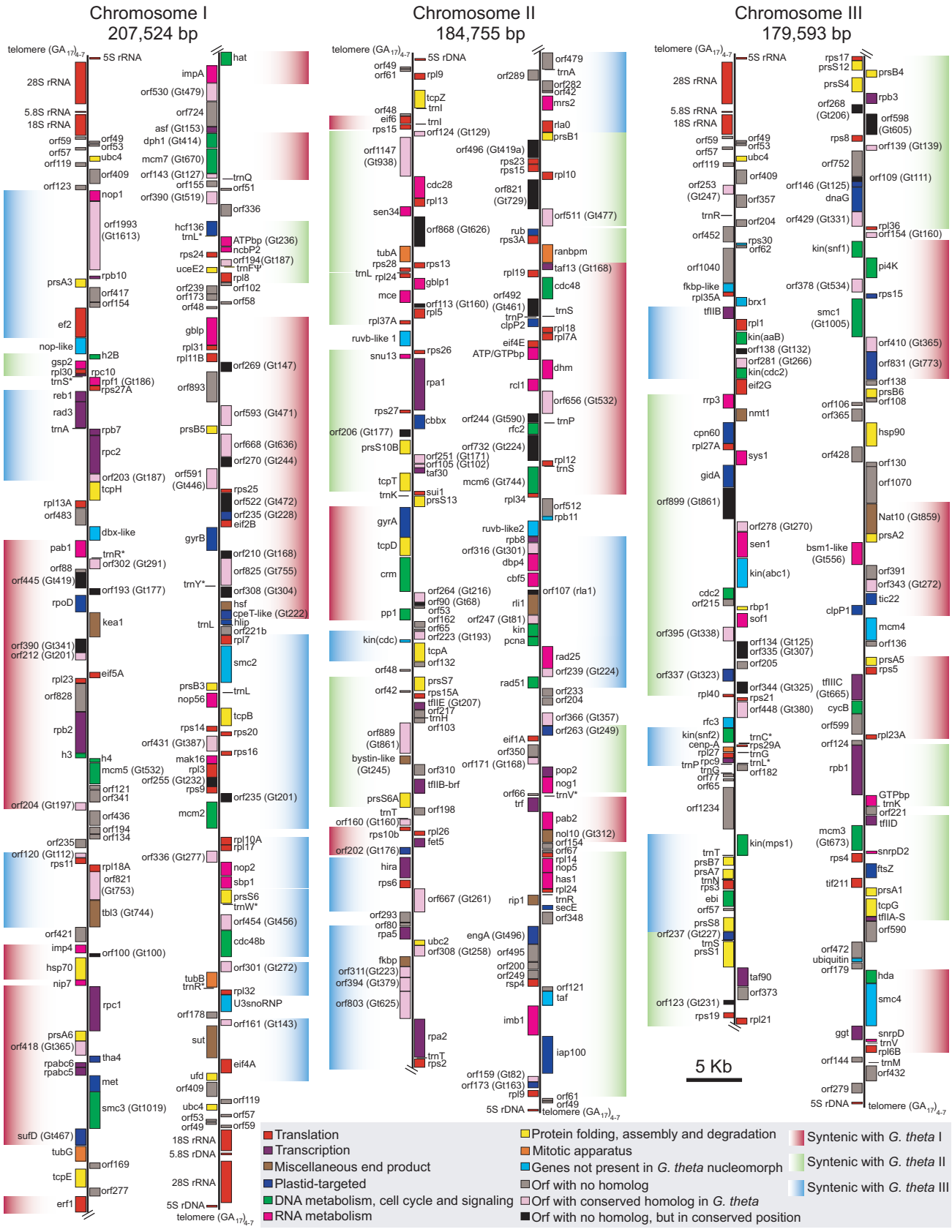

Fig. 1. *H. andersenii* nucleomorph genome. The genome is 571,872 bp in size with three chromosomes, shown artificially broken at their midpoint. Colors correspond to predicted functional categories, and shaded bars indicate regions of synteny with the nucleomorph genome of *G. theta*. Unidentified ORFs in *H. andersenii* with names followed by brackets including ''Gt'' correspond to demonstrably homologous unidentified ORFs in *G. theta* (pink) or ORFs with no obvious sequence similarity residing in a syntenic genomic position (black). Genes mapped on the left side of each chromosome are transcribed bottom-to-top and those on the right, top-to-bottom.

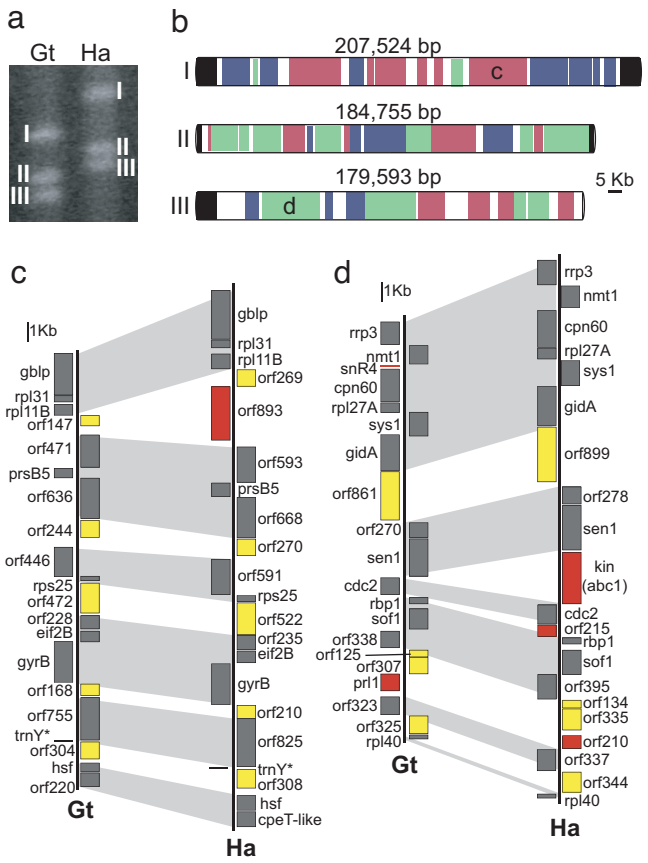

Fig. 2. A high degree of synteny between the *H. andersenii* (Ha) and *G. theta* (Gt) nucleomorph genomes. (*a*) Pulsed-field gel electrophoresis showing the relative sizes of the three nucleomorph chromosomes in the two species. (*b*) Schematic of the three *H. andersenii* chromosomes with syntenic blocks highlighted in color: black, multicopy DNA; red, syntenic with *G. theta* chromosome I; green, syntenic with *G. theta* II, blue, syntenic with *G. theta* III. ''c'' and ''d'' indicate portions of the genome shown in greater detail in the lower portion of the figure. (*c* and *d*) Comparisons of two selected syntenic regions shown on chromosome I and III, respectively, in *b*. Genes in gray are shared between the two genomes, whereas those in red are not. Yellow ORFs show no detectable sequence homology to each other or any known gene/protein, despite occupying the same position in each genome.

on the internal side of one chromosome II repeat and truncated pseudogene copies in the same location on the other five chromosome ends (7). In *G. theta*, the repeats encode more proteins and are more variable in content. The ubc4 gene resides within five of the six *G. theta* repeats, and tfIID exists on three of the chromosome ends. However, the region outside of ubc4 is essentially identical on all ends (6). A similar pattern has been shown to exist in two other cryptophytes, *Hanusia phi* and *Proteomonas sulcata* (12).

In the case of the cryptophyte genus *Hemiselmis*, exploratory Southern blot hybridizations (11) revealed that all members of this genus appear to lack the majority of the rDNA cistron on chromosome II. The complete nucleomorph genome presented here confirms this for *H. andersenii* and demonstrates that this situation also exists on one end of chromosome III (Fig. 1). The immediately subtelomeric 5S rDNA is the only remnant of the rDNA cistron on all three of these ends. Interestingly, rpl9 occurs on both ends of chromosome II and is part of a large syntenic block shared with *G. theta* on one of these ends. This suggests that the original rDNA cistron was replaced by a portion of the chromosome normally located internal to the repeat. The rpl9 locus was presumably then propagated to the opposite end of the chromosome. Whether chromosome II or III was the first to lose most of the rDNA cistron from one of its ends is unclear.

Subtelomeric rDNA cistrons appear to be widespread in cryptophyte and chlorarachniophyte nucleomorph genomes (6, 12) and, curiously, are also found in the reduced nuclear genomes of the microsporidian *E. cuniculi* (20) and the diplomonad *G. lamblia* (33). It is possible that the elevated G+C content of the rDNA loci in such genomes serves a role in the maintenance of chromosome ends in genomes with a high average A+T content. However, the lack of full rDNA cistrons on half of the *H. andersenii* chromosome ends, combined with their extraordinarily A+T-rich telomeric sequences, suggest that this is unlikely. More plausibly, the presence of subtelomeric rDNA cistrons is related to the high rates of recombination and gene conversion that are typical for chromosome ends, paradoxically, in a region of the genome often associated with moderate levels of gene expression. The consequences of having half of the ''typical'' nucleomorph complement of rDNA cistrons in *Hemiselmis* is not immediately obvious, but the different complement of multicopy protein genes associated with the chromosome ends in nucleomorph genomes examined thus far suggests that such genes are likely the beneficiaries of serendipity rather than selection.

Analysis of *Hemiselmis* and *Guillardia* ORFs. The *H. andersenii* nucleomorph genome possesses 472 predicted protein genes, compared with 465 in *G. theta* (SI Fig. 4). We used two interrelated criteria to infer gene homology between the two genomes, standard DNA/protein sequence similarity and consideration of synteny. Based on sequence similarity alone, >50% (254 of 472) of the *H. andersenii* genes are present in *G. theta* and have identifiable homologs in canonical nuclear genomes. Eighteen *H. andersenii* genes with clear eukaryotic homologs are not found in *G. theta*, whereas 16 ''conserved'' *G. theta* genes are absent in *H. andersenii* (Fig. 1 and SI Fig. 4). Intriguingly, both nucleomorphs harbor an identical set of 30 genes predicted to encode plastid-targeted proteins. The retention of essential genes for photosynthesis (and the machinery to express them) in the nucleomorph has been touted as the raison d'être of these enigmatic organelles (6). Clearly the complement of cryptophyte nucleomorph-encoded plastid protein genes was established very early in the evolution of this lineage, before the divergence of *G. theta* and *H. andersenii*. The functional significance of this observation, in terms of the migration of nucleomorph genes to the host nucleus, is unknown.

The remaining *H. andersenii* ORFs can be classified as follows: (*i*) ORFs that are demonstrably homologous between *H. andersenii* and *G. theta* but that show no obvious similarity to genes in other genomes (60 of 472 = 13%), (*ii*) ORFs showing no homology to any gene in *G. theta* or elsewhere (110 of 472 = 23%), and (*iii*) *H. andersenii* ORFs with no detectable homology to known genes but that reside within regions of syntenic conservation (30 of 472 = 6.4%). Remarkably, pairwise comparisons of the *H. andersenii* and *G. theta* ORFs in the third category (Fig. 2 *c* and *d*) reveal that they are almost always similar in size and, despite sharing no obvious amino acid sequence similarity, have similar pI values and number of predicted transmembrane helices (if present). Given that proteins encoded in nucleomorph genomes are often divergent in sequence (3, 34–36), it appears likely that these genes share a common origin but have rapidly diverged in sequence. Together with the *H. andersenii*-specific unidentified ORFs, the ''syntentic ORFs'' encode predicted proteins with highly biased amino acid compositions, much more so than ORFs with demonstrable homologs in other eukaryotes and/or in both nucleomorph genomes. Remarkably, the proportion of phenylalanine and asparagine residues (which are encoded by A+T-rich codons) in many of the *H. andersenii*- and *G. theta*-specific proteins exceeds

EVOLUTION

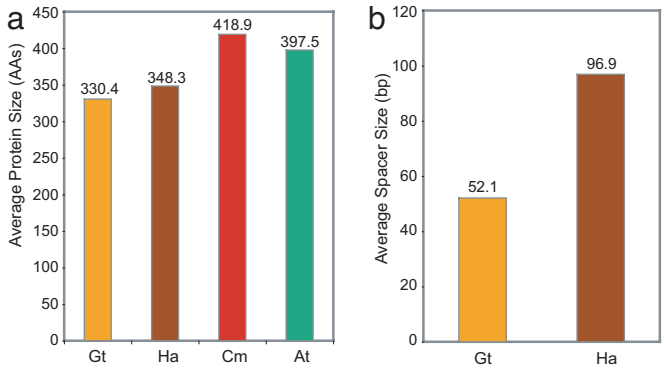

Fig. 3. Impact of genome reduction and compaction on gene density and gene/protein size. (*a*) Histogram showing average protein sizes for a set of 198 homologous proteins in the *H. andersenii* and *G. theta* nucleomorph genomes and the nuclear genomes of the red alga *C. merolae* and the land plant *A. thaliana*. All pairwise comparisons are significant when paired *t* test and binomial test statistics ($P < 0.0005$) are used. (*b*) Histograms showing average sizes of 164 homologous intergenic spacers in the *H. andersenii* and *G. theta* nucleomorph genomes.

25%. Despite their unusual composition, these proteins are very likely bona fide: 83 are >150 aa in length, 45 are >300 aa long, and 8 *H. andersenii* ORFs with no sequence homology to known proteins are >800 aa in length. It would appear that through the combined effects of increased mutation rate and/or reduced selective constraint, a significant fraction of cryptophyte nucleomorph-encoded proteins are evolving extraordinarily quickly, yet for unknown reasons, are retained in both *H. andersenii* and *G. theta*.

Protein Size Reduction. We next sought to test whether the process of genome reduction/compaction has influenced the size of nucleomorph-encoded proteins as well as their composition. We compared the sizes of 198 proteins found in both *H. andersenii* and *G. theta* with their homologs in the red alga *C. merolae* (37) and the land plant *Arabidopsis thaliana* (38) (the genome of *B. natans* (7) could not be analyzed because of the small number of genes its green-algal-derived nucleomorph shares with cryptophytes). Ninety-two percent of these proteins were smaller in nucleomorphs than in both *C. merolae* and *A. thaliana* (Fig. 3*a* and SI Table 1). All pairwise size comparisons were significant when paired *t* test and binomial test statistics ($P < 0.0005$) were used. No functional bias was observed, because the trend was apparent in proteins involved in a wide range of cellular processes, including protein folding and degradation, transcription, translation, and RNA metabolism.

To determine where nucleomorph protein shortening had occurred, we examined 50 protein sequence alignments assembled to include homologs from diverse eukaryotes. Although the amino and carboxyl termini were almost always shorter than their homologs in algae and other eukaryotes (SI Fig. 5 *a* and *b*), numerous internal deletions were also apparent (SI Fig. 5 *c*–*e*). Deletions were often localized to regions of the proteins that were variable in length, presumably corresponding to surface loops in protein structure. However, in many cases, the cryptophyte nucleomorph-encoded proteins were >100 aa shorter than their homologs in other eukaryotes, suggesting that entire protein domains have been removed. A striking example is a transcription factor involved in the regulation of heat shock protein gene expression: in *H. andersenii* and *G. theta* (6, 34) the HSF protein is 236 and 185 aa long, respectively, compared with 467 in *C. merolae*, 476 in *A. thaliana* (SI Table 1) and 833 in the yeast *Saccharomyces cerevisae*. Although the amino-terminal DNA-binding domain remains intact, the transactivation domain at the carboxyl terminus has been deleted (data not shown), suggesting a fundamentally different mode of action for this transcription factor in the cryptophyte nucleomorph. Another example is the largest subunit of RNA polymerase II (RPB1). The C-terminal domain (CTD) of RPB1 in most eukaryotes contains an evolutionarily conserved, tandemly arrayed heptapeptide repeat that serves as a platform for interactions with a variety of proteins involved in transcription (39). The nucleomorph-encoded RPB1 proteins are >300 aa shorter than those in *C. merolae* and *A. thaliana* (SI Table 1) and completely lack a CTD (*C. merolae* contains a CTD with atypical repeats). A host of other transcription-related proteins are shorter as well (e.g., RPA1, RPA2, RPC1). The 76-aa ubiquitin monomer, which is typically encoded as part of a polyubiquitin tract, has been lost in *G. theta* but is retained in the *H. andersenii* nucleomorph genome as a single stand-alone ORF, as in the reduced genomes of *G. lamblia* (40) and *E. cuniculi* (20).

Unexpectedly, not only are the cryptophyte nucleomorph-encoded proteins shorter than their homologs in other eukaryotes, the sizes of *H. andersenii* and *G. theta* proteins differ significantly from one another. Eighty-one percent of 290 comparable homologs in the 0.572-Mbp *H. andersenii* genome are larger than their counterparts in *G. theta* (Fig. 3*a* and SI Table 1), whose genome is smaller (0.551 Mbp) and more compact: comparison of homologous gene spacers reveals a mean intergenic distance of 52 bp in the *G. theta* genome versus 97 bp in *H. andersenii* (Fig. 3*b*). The difference in both protein and intergenic spacer size is significant at $P < 0.0005$ when both binomial and *t* test statistics were used. An interesting comparison of the 2.9-Mbp genome of the microsporidian *E. cuniculi* to the 12-Mbp *S. cerevisiae* genome revealed that 85% of its proteins were smaller than their homologs in yeast (20). Based on the assumption that in eukaryotes large proteins facilitate complex regulatory networks (41), it was suggested (20) that this discrepancy reflects a decreased requirement for protein–protein interactions in a highly simplified intracellular parasite with fewer proteins and a simplified "interactome." In the case of nucleomorphs, the significantly different sizes of proteins in *H. andersenii* and *G. theta*, whose genomes encode approximately the same number of proteins (and whose endosymbiont compartments presumably import approximately the same number of nucleus-encoded proteins), suggest that genome compaction can play a direct role in the process of protein shortening, beyond simply providing the mechanism for the eventual elimination of genes (or parts of genes) that are no longer essential. We hypothesize that a deletion bias accounts for the smaller, more compact nucleomorph genome of *G. theta* as well as its smaller proteins. This can be tested by comparing the nucleomorph genomes of very closely related cryptophytes and, when present, pseudogenes, as has been done to demonstrate the existence of a deletion bias in species of *Drosophila* (42, 43), once these data become available.

Methods

DNA Isolation, Genome Sequencing, and Genome Annotation. By using density gradient-purified DNA as starting material (11), nucleomorph DNA from *H. andersenii* CCMP644 was nebulized and electrophoretically separated on a 1% agarose gel, cloned into pUC19 vector, and shotgun sequenced to ≈9× coverage by using ET terminator chemistry (GE Healthcare) and MegaBace capillary DNA sequencers. Assembly and editing of the ≈16,500 end reads was performed by using Staden (44) and resulted in 28 nonoverlapping contigs. Contigs were mapped to specific chromosomes by using Southern blot hybridization and the remaining gaps were filled by using long-range PCR. PCR products were cloned and sequenced as described (11). ORFs >40 aa in size were identified in Artemis (45) and examined for their coding potential by using BLASTX (46). tRNA genes were identified

by using tRNA-scan (http://lowelab.ucsc.edu/tRNAscan-SE/). Ribosomal RNA and snRNA genes were identified by using BLASTN and by comparison to the *G. theta* nucleomorph genome. The *H. andersenii* nucleomorph genome sequence has been deposited in GenBank under the following accession numbers: CP000881, CP000882, and CP000883.

Identification of Syntenic Regions. Portions of the genome were considered syntenic if they shared identifiable ORFs in the same order and orientation as *G. theta*. ORFs showing no detectible homology to known ORFs were not considered interruptions of a syntenic block, nor were tRNAs or genes present in only one of the genomes. Only blocks that included three or more conserved genes (or ORFs shared between the nucleomorphs) were described as syntenic. Genes encoding rRNAs were not included.

Statistical Analysis. Protein length and intergenic spacer size was compared between the two genomes by using two test statistics. A paired *t* test was implemented by using the following equation: t = mean(length difference)/[s/sqrt(n)] where s is the square root of the sample variance. Infinite degrees of freedom were used when calculating P values. A binomial test was also used with the following formula: $P(Z > = [\text{phat} - 1/2]/(0.5/\text{sqrt}(n)))$, where Z has a normal distribution, and phat is the proportion of proteins where the length is shorter in *G. theta* than *H. andersenii* or the nucleomorph representative, in cases where a nucleomorph genome was being compared with a nuclear genome.

We thank O. Zhaxybayeva, F. Doolittle, A. Roger, E. Bapteste, H. Philippe, and M. Gray for comments and discussion; A. Fong for DNA sequencing; and E. Susko for assistance with statistical analyses. This work was supported by Genome Atlantic and a Natural Sciences and Engineering Research Council (Canada) Discovery grant (to J.M.A.). J.M.A. is a Scholar of the Canadian Institute for Advanced Research, Program in Integrated Microbial Biodiversity.

1. Gregory TR (2001) *Biol Rev* 76:65–101.
2. Keeling PJ, Slamovits CH (2005) *Curr Opin Gen Dev* 15:601–608.
3. Archibald JM (2007) *BioEssays* 29:392–402.
4. Eschbach S, Hofmann CJ, Maier UG, Sitte P, Hansmann P (1991) *Nucleic Acids Res* 19:1779–1781.
5. Gilson PR, McFadden GI (2002) *Genetica* 115:13–28.
6. Douglas SE, Zauner S, Fraunholz M, Beaton M, Penny S, Deng L, Wu X, Reith M, Cavalier-Smith T, Maier U-G (2001) *Nature* 410:1091–1096.
7. Gilson PR, Su V, Slamovits CH, Reith ME, Keeling PJ, McFadden GI (2006) *Proc Natl Acad Sci USA* 103:9566–9571.
8. Archibald JM (2005) *Int Union Biochem Mol Biol Life* 57:539–547.
9. Palmer JD (2003) *J Phycol* 39:4–11.
10. Gilson PR, McFadden GI (1999) *Phycol Res* 47:7–19.
11. Lane CE, Archibald JM (2006) *J Eukaryot Microbiol* 53:515–521.
12. Lane CE, Khan H, MacKinnon M, Fong A, Theophilou S, Archibald JM (2006) *Mol Biol Evol* 23:856–865.
13. Rensing SA, Goddemeier M, Hofmann CJ, Maier UG (1994) *Curr Genet* 26:451–455.
14. Silver T, Koike S, Yabuki A, Kofuji R, Archibald JM, Ishida K-I (2007) *J Eukaryot Microbiol* 54:403–410.
15. Lane CE, Archibald JM (2008) *J Phycol*, in press.
16. Martin W, Koonin EV (2006) *Nature* 440:41–45.
17. Collins CA, Guthrie C (2000) *Nat Struct Biol* 7:850–854.
18. Nilsen TW (2003) *BioEssays* 25:1147–1149.
19. Morrison HG, McArthur AG, Gillin FD, Aley SB, Adam RD, Olsen GJ, Best AA, Cande WZ, Chen F, Cipriano MJ, *et al.* (2007) *Science* 317:1921–1926.
20. Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P, *et al.* (2001) *Nature* 414:450–453.
21. Nixon JE, Wang A, Morrison HG, McArthur AG, Sogin ML, Loftus BJ, Samuelson J (2002) *Proc Natl Acad Sci USA* 99:3701–3705.
22. Linder P (2006) *Nucleic Acids Res* 34:4168–4180.
23. Edwalds-Gilbert G, Kim DH, Silverman E, Lin RJ (2004) *RNA* 10:210–220.
24. Watkins NJ, Segault V, Charpentier B, Nottrott S, Fabrizio P, Bachi A, Wilm M, Rosbash M, Branlant C, Luhrmann R (2000) *Cell* 103:457–466.
25. Saccone C, Gissi C, Reyes A, Larizza A, Sbisa E, Pesole G (2002) *Gene* 286:3–12.
26. Suyama M, Bork P (2001) *Trends Genet* 17:10–13.
27. Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ, Sandstrom JP, Moran NA, Andersson SG (2002) *Science* 296:2376–2379.
28. Huynen MA, Bork P (1998) *Proc Natl Acad Sci USA* 95:5849–5856.
29. Ranz JM, Casals F, Ruiz A (2001) *Genome Res* 11:230–239.
30. Sharakhov IV, Serazin AC, Grushko OG, Dana A, Lobo N, Hillenmeyer ME, Westerman R, Romero-Severson J, Costantini C, Sagnon N, *et al.* (2002) *Science* 298:182–185.
31. Slamovits CH, Fast NM, Law JS, Keeling PJ (2004) *Curr Biol* 14:891–896.
32. Deane JA, Strachan IM, Saunders GW, Hill DRA, McFadden GI (2002) *J Phycol* 38:1236–1244.
33. Le Blancq SM, Adam RD (1998) *Mol Biochem Parasitol* 97:199–208.
34. Archibald JM, Cavalier-Smith T, Maier U, Douglas S (2001) *J Mol Evol* 52:490–501.
35. Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H (2005) *Syst Biol* 54:743–757.
36. Keeling PJ, Deane JA, Hink-Schauer C, Douglas SE, Maier UG, McFadden GI (1999) *Mol Biol Evol* 16:1308–1313.
37. Matsuzaki M, Misumi O, Shin IT, Maruyama S, Takahara M, Miyagishima SY, Mori T, Nishida K, Yagisawa F, Nishida K, *et al.* (2004) *Nature* 428:653–657.
38. The *Arabidopsis* Genome Initiative (2000) *Nature* 408:796–815.
39. Stiller JW, Hall BD (2002) *Proc Natl Acad Sci USA* 99:6091–6096.
40. Krebber H, Wostmann C, Bakker-Grunwald T (1994) *FEBS Lett* 343:234–236.
41. Zhang J (2000) *Trends Genet* 16:107–109.
42. Petrov DA, Lozovskaya ER, Hartl DL (1996) *Nature* 384:346–349.
43. Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL (2000) *Science* 287:1060–1062.
44. Bonfield JK, Smith KF, Staden R (1995) *Nucleic Acids Res* 23:4992–4999.
45. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B (2000) *Bioinformatics* 16:944–945.
46. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) *Nucleic Acids Res* 25:3389–3402.