Current Biology 21, 328–333, February 22, 2011 ©2011 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2011.01.037

Report

Red and Green Algal Monophyly and Extensive Gene Sharing Found in a Rich Repertoire of Red Algal Genes

Cheong Xin Chan,[1,5] Eun Chan Yang,[2,5] Titas Banerjee,[1] Hwan Su Yoon,[2,*] Patrick T. Martone,[3] José M. Estevez,[4] and Debashish Bhattacharya[1,*]
[1]Department of Ecology, Evolution, and Natural Resources and Institute of Marine and Coastal Sciences, Rutgers University, New Brunswick, NJ 08901, USA
[2]Bigelow Laboratory for Ocean Sciences, West Boothbay Harbor, ME 04575, USA
[3]Department of Botany, University of British Columbia, 6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada
[4]Instituto de Fisiología, Biología Molecular y Neurociencias (IFIBYNE UBA-CONICET), Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, 1428 Buenos Aires, Argentina

Summary

The Plantae comprising red, green (including land plants), and glaucophyte algae are postulated to have a single common ancestor that is the founding lineage of photosynthetic eukaryotes [1, 2]. However, recent multiprotein phylogenies provide little [3, 4] or no [5, 6] support for this hypothesis. This may reflect limited complete genome data available for red algae, currently only the highly reduced genome of *Cyanidioschyzon merolae* [7], a reticulate gene ancestry [5], or variable gene divergence rates that mislead phylogenetic inference [8]. Here, using novel genome data from the mesophilic *Porphyridium cruentum* and *Calliarthron tuberculosum*, we analyze 60,000 novel red algal genes to test the monophyly of red + green (RG) algae and their extent of gene sharing with other lineages. Using a gene-by-gene approach, we find an emerging signal of RG monophyly (supported by ~50% of the examined protein phylogenies) that increases with the number of distinct phyla and terminal taxa in the analysis. A total of 1,808 phylogenies show evidence of gene sharing between Plantae and other lineages. We demonstrate that a rich mesophilic red algal gene repertoire is crucial for testing controversial issues in eukaryote evolution and for understanding the complex patterns of gene inheritance in protists.

Results and Discussion

Assessing Red and Green Algal Monophyly Based on Exclusive Gene Sharing

Here, with 36,167 expressed sequence-tagged (EST) unigenes from *Porphyridium cruentum* and 23,961 predicted proteins from *Calliarthron tuberculosum*, we report analyses of >60,000 novel genes from mesophilic red algae. Of the 36,167 *P. cruentum* unigenes (6.7-fold greater than the gene number [5,331] from *Cyanidioschyzon merolae* [7]), 13,632 encode proteins with significant BLASTp hits (e value $\leq 10^{-10}$) to sequences in our local database, in which we included the 23,961 predicted proteins from *C. tuberculosum* (see Table S1 available online). Of these hits, 9,822 proteins (72.1%, including many *P. cruentum* paralogs) were present in *C. tuberculosum* and/or other red algae, 6,392 (46.9%) were shared with *C. merolae*, and 1,609 were found only in red algae. A total of 1,409 proteins had hits only to red algae and one other phylum. Using this repertoire, we adopted a simplified reciprocal BLAST best-hits approach to study the pattern of exclusive gene sharing between red algae and other phyla (see Experimental Procedures). We found that 644 proteins showed evidence of exclusive gene sharing with red algae. Of these, 145 (23%) were found only in red + green algae (hereafter, RG) and 139 (22%) only in red + Alveolata (Figure 1A). In comparison, we found only 34 (5%) proteins in red + Glaucophyta, likely as a result of the limited availability of glaucophyte data in the database. As we restricted this search by requiring a larger number of hits per query (x) from both phyla, the proportion of RG proteins increased relative to other taxa. For instance, the number of red + Alveolata and red + Metazoa proteins was reduced from 139→1→0 and from 55→3→0 when $x \geq 2$ (644 proteins), $x \geq 10$ (96 proteins), and $x \geq 20$ (22 proteins), respectively (Figures 1A–1C). This BLASTp analysis is based on the implicit assumption that significant similarity among a group of sequences indicates a putative homologous relationship (i.e., a shared common ancestry). This approach could potentially be misled by convergence at the amino acid level that results in high similarity among non-homologous sequences (i.e., homoplasy [9, 10]). Alternatively, because RG are primarily photoautotrophs, exclusive gene sharing could be explained by these lineages having retained a common set of ancestral genes that were lost in other eukaryotes. With these potential issues in mind, we suggest that exclusive gene sharing (as defined by significant reciprocal BLASTp hits) provisionally favors the RG grouping.

Gene Sharing between RG and Other Lineages

Using a phylogenomic approach, we generated maximum-likelihood (ML) trees for each of the 13,632 *P. cruentum* proteins with significant hits to the local database. One of the major confounding issues in phylogenomic analysis is inadequate and/or biased taxon sampling. To reduce such biases in our inference of gene phylogeny, we restricted our analysis to trees that contain ≥ 3 phyla (per tree) and analyzed these phylogenies based on the minimum number of terminal taxa per tree (n), ranging from 4 to 40 (Figure 1D). The expectation was that the impact of inadequate taxon sampling on our interpretation of the data would be minimal in trees with large n. Applying these restrictions, $n \geq 4$ returned 1,367 trees that contained red algae positioned within a strongly supported (bootstrap $\geq 90\%$) monophyletic clade (Figure 1D); the majority of these trees (1,129 of 1,367; 83%) had $n \geq 10$. Among the 1,367 trees, 329 showed exclusive RG monophyly, of which 53 trees defined RG + glaucophytes (i.e., were putatively Plantae-specific). The number of trees that recovered the RG remained similar between cases of $n \geq 4$ and $n \geq 10$, with only 71 trees having n between 4 and 9. As n increased, the proportion of RG groupings remained

*Correspondence: hsyoon@bigelow.org (H.S.Y.), bhattacharya@aesop.rutgers.edu (D.B.)
[5]These authors contributed equally to this work

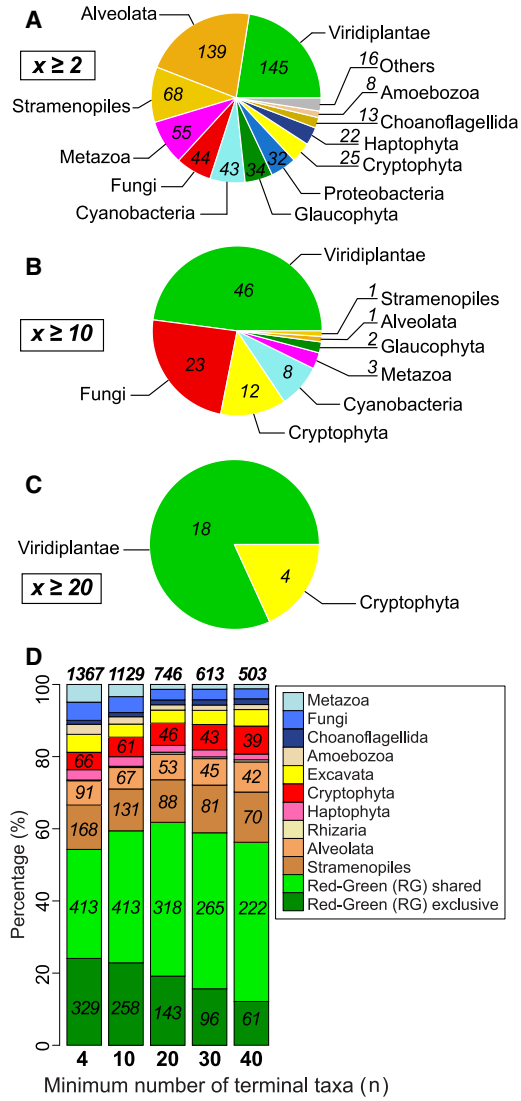

Figure 1. Analysis of Predicted Proteins from the Red Alga *Porphyridium cruentum*

(A–C) The distribution of phyla with exclusive BLASTp hits to *P. cruentum* proteins where the number of hits per query (x) is as follows, (A) x ≥ 2, (B) x ≥ 10, and (C) x ≥ 20. The colors indicate the different phyla that share proteins exclusively with *P. cruentum*.

(D) The percentage of maximum-likelihood (ML) protein trees (raw numbers shown in the bars for the five most frequently found groupings) that support the monophyly of red algae with other eukaryote phyla (bootstrap ≥ 90%). The impact of increasing the number of terminal taxa in each tree (n) on these proportions is shown for the progression from 4→10→20→30→40. The total number of trees for each category is shown on top of each bar. The category "Red-Green (RG) exclusive" refers to trees in which these two phyla form an exclusive clade, whereas "Red-Green (RG) shared" refers to trees in which red-green monophyly is well supported but other phyla are found within this clade (i.e., due to gene sharing). See also Figure S1.

similar across all categories, although the number of trees supporting this clade gradually decreased. These estimates reflect our current database and will change as more genome data become available. Figure 2A shows the phylogeny of a putative Plantae-specific gene (of unknown function) that appears to have undergone an ancient gene duplication in the Plantae ancestor followed by subsequent duplications (particularly among land plants).

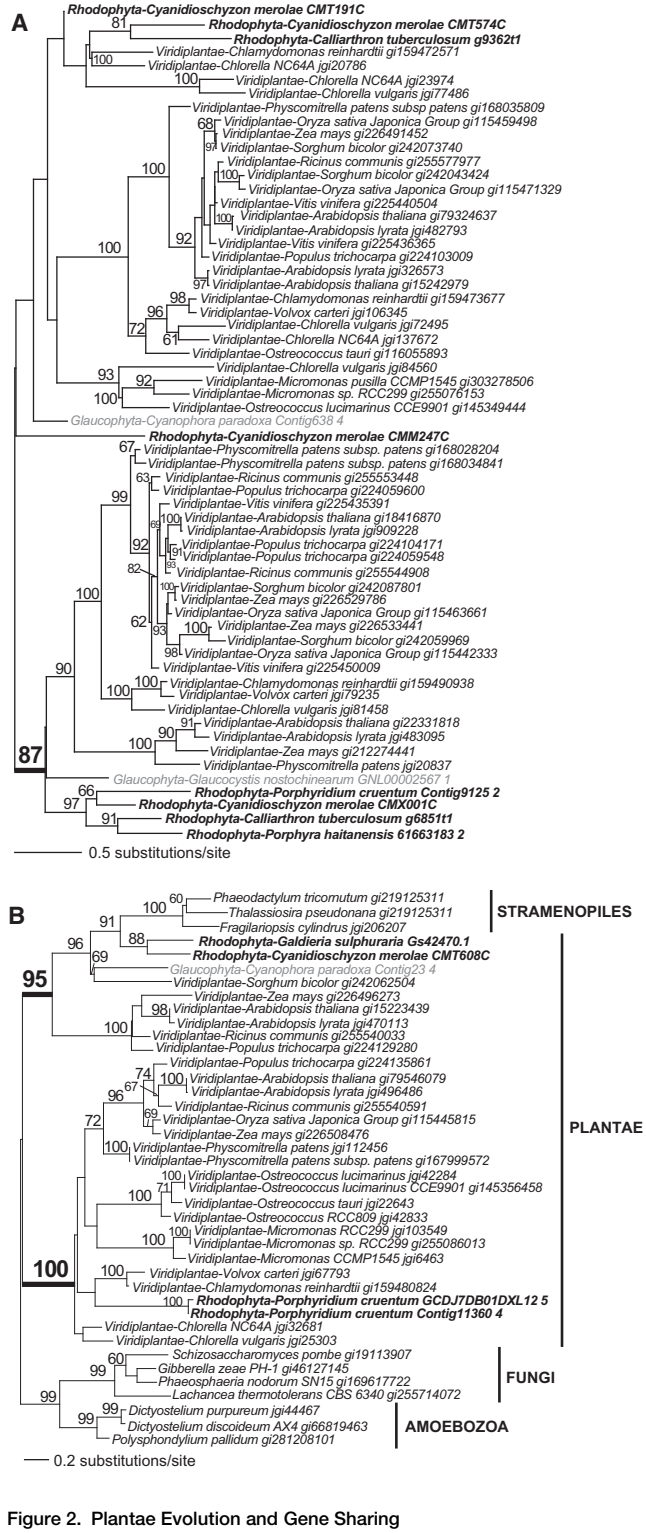

Figure 2. Plantae Evolution and Gene Sharing

(A) Phylogeny of a gene of unknown function that is putatively specific to Plantae.

(B) Phylogeny of a gene encoding a putative phosphoglyceride transfer protein, SEC14, with a well-supported monophyly (bootstrap 95%) of plants, red algae, the glaucophytes, and diatoms and a monophyly (bootstrap 100%) between *Porphyridium cruentum* and green algae (including other plants). RAxML [30] bootstrap support values ≥ 60% based on 100 nonparametric replicates are shown at the nodes. Red algae are shown in boldface and glaucophytes in gray. The unit of branch length is the number of substitutions per site. See also Figure S2.

In these analyses, we also examined instances of RG monophyly in which other taxa interrupted this clade, e.g., Stramenopiles, presumably resulting from endosymbiotic/horizontal gene transfer (E/HGT). We referred to such instances as "RG shared" (Figure 1D), whereby there was a strongly supported monophyly (i.e., bootstrap $\geq$ 90%) of RG algae with other non-Plantae lineages. We applied the condition that RG shared clades include $\geq$75% of all terminal taxa in a tree, and within this clade, a majority (>50%) of the tips defined red and green algae. Using this definition, we found an additional 413 trees that support RG monophyly (Figure 1D). Therefore, at n $\geq$ 4, a total of 742 (54%) of 1,367 trees returned by our pipeline supported the RG union (bootstrap $\geq$ 90%). At a less stringent bootstrap threshold of $\geq$70%, 997 (46%) of 2,167 trees showed support for RG monophyly (Figure S1). An example of a phylogeny showing nonexclusive gene sharing with Plantae lineages is shown in Figure 2B for a putative phosphoglyceride transfer protein, SEC14. The phylogeny shows a well-supported monophyly (bootstrap 95%) of plants, red algae (*Galdieria sulphuraria* and *C. merolae*), the glaucophyte *Cyanophora paradoxa*, and diatoms. The diatom gene likely arose via secondary endosymbiotic gene transfer from a red algal donor [11]. In addition, a divergent red algal-derived gene copy is present in *P. cruentum* that groups with green algae and other gene copies found in plants (bootstrap 100%). Although complete genome data from glaucophytes and other red algae are required to delineate the extent of gene duplication and convergence between these two lineages, this phylogeny illustrates two key properties of protist gene and genome evolution that pose challenges to the inference of lineage relationships: ancient gene duplication (e.g., multiple copies in plants) and loss (i.e., putatively of a gene copy in green algae, e.g., *Ostreococcus* spp.), and nonlineal gene sharing involving algal lineages.

The next most frequently found positions of red algae in these trees were as sister to Stramenopiles (168, 12%), Alveolata (91, 7%), Excavata (68, 5%), and Cryptophyta (66, 5%). Increasing the minimum number of terminal taxa per tree (n) from 4 through 40 (while maintaining $\geq$3 phyla) did not affect the relative proportion of trees that support RG monophyly, but the number of cases with other well-supported phylogenetic affiliations (e.g., red + Metazoa, red + Fungi) fell sharply (Figure 1D). When we relaxed the bootstrap threshold to $\geq$70%, the patterns reported here generally remained unchanged (Figure S1) but allowed the identification of single-protein markers that may prove useful for delineating the eukaryote tree of life (e.g., V-type ATPase I 116 kDa subunit family; Figure S2; see also [12]).

We found 1,808 trees that showed strong support (at bootstrap $\geq$90%) for the monophyly of RG with other "foreign" taxa. Figure 3 shows the number of these trees that contain different foreign phyla within the well-supported RG clade. The sources of the foreign genes are depicted in a schematic representation of the putative tree of life. The most common partners of gene sharing with RG (i.e., barring significant phylogenetic artifacts in our approach) are Stramenopiles (e.g., the diatoms; 1,264 proteins), bacteria other than Cyanobacteria (1,108), Haptophyta (839), Cyanobacteria (827), Alveolata (622), and Metazoa (473). The majority of these proteins (1,322 of 1,808) are shared between RG and two or more other phyla, demonstrating the complex evolutionary history of the algal genes. We recognize that our results are biased by the unbalanced contribution of available genome data from microbial eukaryotes in our database (e.g., diatoms

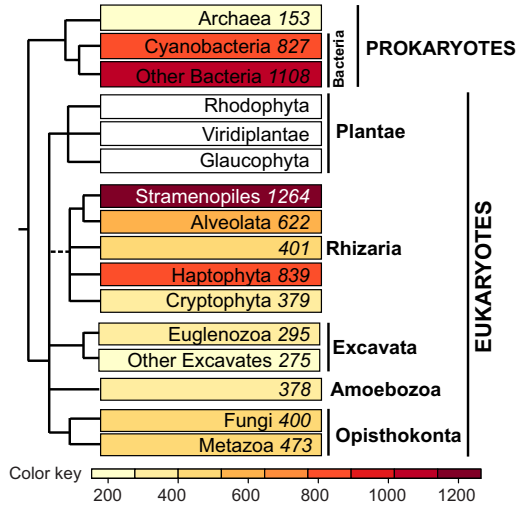

Figure 3. Schematic Representation of the Putative Tree of Life Showing the Extent of RG Gene Sharing with Other Eukaryote and Prokaryote Phyla

The branch shown as a dashed line represents ambiguous relationships among the lineages to the right. The color key indicates the number of trees found for each "foreign" (non-RG) phylum.

are gene rich, cryptophytes are gene poor). In addition, the detection of gene transfer using phylogenetic approaches is susceptible to a number of technical limitations such as modularity [13, 14] and amelioration [15, 16] of the transferred genes, which result in underestimation of the extent of HGT in gene-by-gene surveys. Nevertheless, our findings indicate that single-gene or multigene analysis of Plantae should take into account extensive gene sharing vis-à-vis other eukaryote lineages (e.g., nongreen affiliation in nearly one-half of *P. cruentum* proteins shown in Figure 1D).

Lastly, we examined whether the observed signal of RG monophyly was contributed primarily or solely by nuclear-encoded plastid-targeted proteins (i.e., whether they reflect the evolution of the organelle rather than the host cell). To do this, we analyzed all RG-exclusive and RG-shared proteins (Figure 1D) with n ranging from 4 through 40. Using an integrated pipeline that incorporates three independent target-prediction approaches, we found that circa 40% of the proteins that support RG monophyly at bootstrap $\geq$ 90% may be plastid targeted (253 of 742, 34.1% at n $\geq$ 4; 119 of 283, 42.1% at n $\geq$ 40; see Supplemental Experimental Procedures). Although bioinformatic predictions of organelle targeting are clearly provisional, these results suggest that in addition to the expected significant contribution to plastid function by proteins that unite the RG (i.e., the vast majority of these taxa are photoautotrophs), over one-half of them may not be destined for the plastid.

Enrichment of Red Algal Genes Enhances Our Understanding of Eukaryote Evolution

To investigate the impact of increasing the number of genes available from mesophilic red algae in comparison to use of the genes of *C. merolae* alone, we applied the reciprocal BLASTp best-hits approach using *C. merolae* proteins as the query against our database. In this case, however, we excluded *P. cruentum* and *C. tuberculosum* from the database. With this approach, we found 127 proteins that showed exclusive gene sharing with red algae, of which 39 (31%)

provided evidence for the RG grouping. Therefore, inclusion of our novel red algal genome data results in a nearly 4-fold increase in the number of red algal genes (145 versus 39) that support exclusive gene sharing among RG taxa.

Our findings also show that red algal genes are distributed among diverse eukaryote lineages that in many instances (e.g., Stramenopiles, Cryptophyta, Haptophyta) are most certainly explained by endosymbiotic gene transfer because these taxa contain a red algal-derived plastid [5, 17, 18]. Of the 474 proteins that show strong support for RG monophyly (145 found only in RG [Figure 1A]; 329 with RG showing exclusive monophyly at bootstrap $\geq$ 90% [Figure 1D]), only 129 (27.2%) have homologs in *C. merolae*. Therefore, with respect to testing the RG or Plantae hypothesis, the red algal gene repertoire from *P. cruentum* and *C. tuberculosum* contributes an almost 4-fold increase in the number of red algal genes useful for phylogenomic analysis as compared with *C. merolae* alone. In addition, only 1,207 (67% of 1,808) genes with a history of gene sharing include homologs from *C. merolae*, suggesting that the extent of gene transfer in eukaryotes has been significantly underestimated in previous phylogenetic analyses that relied on a more limited sample of red algal genes.

In summary, we have uncovered clear evidence of RG monophyly in our analysis of reciprocal BLASTp hits and individual protein trees. No competing hypothesis rises to the level of support that we found for the RG clade. Testing the coherence of the Plantae hypothesis will require the addition of complete genome data from *Cyanophora paradoxa* and other glaucophytes. This is of great interest, because the Plantae lineages provide an important opportunity to advance our knowledge of the tree of life, the intricacies of genome evolution among protists, and the origin of photosynthesis in eukaryotes. For example, it has been known for some time that Plantae share key traits that are usually, but not exclusively, associated with photosynthesis and other plastid functions, which strongly supports their union [17–20]. However, reliance on plastid characters (e.g., trees inferred from organelle genes or nuclear-encoded plastid-targeted proteins) may mislead phylogenetic inference if there has been a complex gain-and-loss pattern of plastids (with associated intracellular gene transfers) among Plantae lineages [6, 8]. Therefore, finding evidence of RG and ultimately Plantae monophyly could greatly improve our understanding of plastid endosymbiosis by tying together the lineages that share a primary plastid, and therefore the innovations underlying organellogenesis [21]. In contrast, Plantae polyphyly will lead to more complex explanations of how primary plastids and their supporting nuclear genes have been distributed among algal lineages. In either case, what has become clear is that concatenated protein data sets often fail to provide resolution of "deep" nodes in the tree of life, including the Plantae (e.g., [3–6, 22]). We suggest that in light of our data, reliance on the standard vertical inheritance model of gene evolution to infer the eukaryote tree of life (e.g., [6, 22]) may need to be critically reassessed on a gene-by-gene basis using an expanded collection of protist genomes. For instance, although providing support for phylum-level relationships, the V-type ATPase I tree (Figure S2) reveals a complex history of gene duplications that makes it a poor marker for species relationships. More problematic is the recent finding of hundreds of green algal-derived genes (that likely arose via ancient gene transfers) in diatoms and other chromalveolates [23] that play key roles in the cell [24]. These studies demonstrate how much still remains to be understood about the evolutionary history of protist genomes. Once a comprehensive knowledge of gene history is gained, then the rapidly accumulating genome data can be incorporated with more confidence into multigene tree-of-life analyses. In summary, our work demonstrates the importance of a rich mesophilic red algal gene repertoire in testing controversial aspects of eukaryote evolution and in enhancing our understanding of the complex patterns of gene inheritance among protists.

Experimental Procedures

Generation of Expressed Sequence Tags from *Porphyridium cruentum*

Total RNA from *Porphyridium cruentum* CCMP1328 (Provasoli-Guillard National Center for Culture of Marine Phytoplankton, Boothbay Harbor, ME) was extracted (TRIzol, Invitrogen) and purified (QIAquick PCR Purification Kit, QIAGEN) according to the manufacturer's instructions. The cDNAs were generated (Mint cDNA Synthesis Kit, Evrogen) from 2 μg of total RNA and normalized (Trimmer cDNA Normalization Kit, Evrogen). The normalized cDNAs were sequenced (GS FLX Titanium, Roche/454 Life Sciences) at the University of Iowa (Iowa City, IA), resulting in 386,903 EST reads. We found no obvious evidence of contamination in the data set from other algal sources or from bacteria based on a sequence similarity search aimed at nontarget taxa (BLAST e value $\leq 10^{-5}$ [25]). We assembled the ESTs into a total of 56,490 sequences with CAP3 [26] using the default settings, yielding 16,651 contigs and 39,839 singlets. To ensure that the phylogenetic signal derived from these sequences was significant and biologically meaningful, we excluded contigs of length < 150 bases and singlets of length < 296 bases (median length for singlets) from subsequent analysis, resulting in 36,167 unigenes for phylogenomic analysis. The assembled ESTs are available at http://dblab.rutgers.edu/home/downloads/. We generated six-frame translations for each of these EST unigenes for the phylogenomic analysis.

Partial Genome Data from *Calliarthron tuberculosum*

Fresh thalli of the coralline red alga *Calliarthron tuberculosum* were collected from the low intertidal zone at Botanical Beach Provincial Park on Vancouver Island, British Columbia, Canada (48° 31′ 43.468″ N, 124° 27′ 12.485″ W). Genomic DNA from the algal cells was extracted (DNeasy Plant Mini Kit, QIAGEN) and sequenced (GS FLX Titanium, Roche/454 Life Sciences) at the McGill University and Génome Québec Innovation Centre (Montréal), resulting in circa 750 Mbp of data. These reads were assembled using gsAssembler (Newbler) version 2.3 (Roche/454 Life Sciences) with default parameters, resulting in 169,975 contigs totaling 51.1 Mbp. The assembled mitochondrial (25,515 bases) and plastid (178,624 bases) DNAs were removed prior to phylogenomic analysis. Proteins encoded by these genome contigs were predicted using a machine learning approach under a generalized hidden Markov model as implemented in AUGUSTUS [27], in which protein models of *Arabidopsis thaliana* were used as the training set. These predicted proteins were incorporated into our in-house sequence database for subsequent phylogenomic analysis. All genome contigs and predicted proteins of *C. tuberculosum* used in this work are available at http://dblab.rutgers.edu/home/downloads/.

Analysis of Exclusive Gene Sharing

For this and all following phylogenomic analyses, we used an in-house database consisting of all annotated protein sequences from RefSeq release 37 at GenBank (http://www.ncbi.nlm.nih.gov/RefSeq/), predicted protein models available from the Joint Genome Institute (ftp://ftp.jgi-psf.org/pub/JGI_data/), and six-frame translated proteins from EST data sets of all publicly available algal and unicellular eukaryote sources, i.e., dbEST at GenBank (http://www.ncbi.nlm.nih.gov/projects/dbEST/) and TBestDB (http://tbestdb.bcm.umontreal.ca/), as well as data from *P. cruentum* and *C. tuberculosum* (see above), totaling 10,469,787 sequences (Table S1). Using 36,167 unigenes of *P. cruentum* as a query platform against the database (BLASTp, e value $\leq 10^{-10}$), we found 1,409 genes to have hits only in red algae and one other phylum. For each of the top five BLASTp hits (or fewer, if there were fewer than five hits) for a *P. cruetum* protein (among 1,409), we generated a list of hits via BLASTp searches against our database. The sequence hits that were found in all of these lists (including the *P. cruentum* protein) were grouped into a set. A protein set consisting only of red algae and one other phylum represented a putative case of exclusive gene sharing between the two phyla.

Phylogenomic Analysis

Of the 13,632 *P. cruentum* genes (37.7% of 36,167) that had significant matches in our database (e value $\leq 10^{-10}$), 1,609 had matches restricted to other rhodophytes, whereas the remainder had hits with diverse prokaryote and/or eukaryote sources; these constituted the homologous protein sets. We applied two sampling criteria to ensure a reasonable representation of the diverse groupings, i.e., $\leq$5 bacterial subgroups, and no single species, strain, or genome was represented >4 times. Sequence alignments were generated with MUSCLE version 3.8.31 [28], and noninformative sites within the alignments were removed with Gblocks version 0.91b [29], with the options b3 = 200, b4 = 2, and b5 = h. We used a strict set of criteria to ensure that the results obtained were phylogenetically meaningful: (1) each protein family (hence alignment) had $\geq$4 members but were limited to $\leq$100 members, and (2) phylogenetically informative sites in each set of aligned protein sequences were $\geq$75 amino acid positions. Under these criteria, 1,635 protein families were excluded. The phylogenies for each of the remaining 12,003 protein alignments were reconstructed using a maximum-likelihood approach [30] using the WAG amino acid substitution model [31] with a discrete gamma distribution [32] and nonparametric bootstrap of 100 replicates. All sequence alignments and the resulting phylogenetic trees used in this study, including the trees sorted by monophyletic relationships (distinct phyla $\geq$ 3, number of terminal taxa $\geq$ 30) based on bootstrap support 70% and 90%, are available at http://dblab.rutgers.edu/home/downloads/.

Accession Numbers

The sequenced EST reads from the normalized cDNAs of *P. cruentum* are available at NCBI GenBank (http://www.ncbi.nlm.nih.gov/) under the accession numbers HS588189–HS975091. The 454 sequence data of *C. tuberculosum* are available at the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under the project accession number SRP005182.

Supplemental Information

Supplemental Information includes two figures, one table, and Supplemental Experimental Procedures and can be found with this article online at doi:10.1016/j.cub.2011.01.037.

Acknowledgments

This research was supported by National Science Foundation (NSF) grants EF 08-27023 and DEB 09-36884 (H.S.Y. and D.B.), NSF OCE-821374 (H.S.Y.), and MCB 09-46528 (D.B.). D.B. acknowledges generous support from Rutgers University. T.B. was supported by NSF Research Experiences for Undergraduates award DEB 10-26425. P.T.M. and J.M.E. would like to thank Chris Somerville for financial and intellectual support of the *Calliarthron* genome sequencing project. P.T.M. acknowledges funding provided by the Natural Sciences and Engineering Research Council of Canada. J.M.E. thanks funding provided by Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT) and Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET) (Argentina).

Received: October 22, 2010
Revised: December 17, 2010
Accepted: January 12, 2011
Published online: February 10, 2011

References

1. Rodríguez-Ezpeleta, N., Brinkmann, H., Burey, S.C., Roure, B., Burger, G., Löffelhardt, W., Bohnert, H.J., Philippe, H., and Lang, B.F. (2005). Monophyly of primary photosynthetic eukaryotes: Green plants, red algae, and glaucophytes. Curr. Biol. *15*, 1325–1330.
2. Weber, A.P., Linka, M., and Bhattacharya, D. (2006). Single, ancient origin of a plastid metabolite translocator family in Plantae from an endomembrane-derived ancestor. Eukaryot. Cell *5*, 609–612.
3. Patron, N.J., Inagaki, Y., and Keeling, P.J. (2007). Multiple gene phylogenies support the monophyly of cryptomonad and haptophyte host lineages. Curr. Biol. *17*, 887–891.
4. Burki, F., Shalchian-Tabrizi, K., Minge, M., Skjaeveland, Å., Nikolaev, S.I., Jakobsen, K.S., and Pawlowski, J. (2007). Phylogenomics reshuffles the eukaryotic supergroups. PLoS ONE *2*, e790.
5. Baurain, D., Brinkmann, H., Petersen, J., Rodríguez-Ezpeleta, N., Stechmann, A., Demoulin, V., Roger, A.J., Burger, G., Lang, B.F., and Philippe, H. (2010). Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. Mol. Biol. Evol. *27*, 1698–1709.
6. Nozaki, H., Maruyama, S., Matsuzaki, M., Nakada, T., Kato, S., and Misawa, K. (2009). Phylogenetic positions of Glaucophyta, green plants (Archaeplastida) and Haptophyta (Chromalveolata) as deduced from slowly evolving nuclear genes. Mol. Phylogenet. Evol. *53*, 872–880.
7. Matsuzaki, M., Misumi, O., Shin-I, T., Maruyama, S., Takahara, M., Miyagishima, S.Y., Mori, T., Nishida, K., Yagisawa, F., Nishida, K., et al. (2004). Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature *428*, 653–657.
8. Stiller, J.W. (2007). Plastid endosymbiosis, genome evolution and the origin of green plants. Trends Plant Sci. *12*, 391–396.
9. Brandley, M.C., Warren, D.L., Leaché, A.D., and McGuire, J.A. (2009). Homoplasy and clade support. Syst. Biol. *58*, 184–198.
10. Sanderson, M.J., and Donoghue, M.J. (1989). Patterns of variation in levels of homoplasy. Evolution *43*, 1781–1795.
11. Li, S., Nosenko, T., Hackett, J.D., and Bhattacharya, D. (2006). Phylogenomic analysis identifies red algal genes of endosymbiotic origin in the chromalveolates. Mol. Biol. Evol. *23*, 663–674.
12. Gogarten, J.P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E.J., Bowman, B.J., Manolson, M.F., Poole, R.J., Date, T., Oshima, T., et al. (1989). Evolution of the vacuolar H^+-ATPase: Implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. USA *86*, 6661–6665.
13. Chan, C.X., Beiko, R.G., Darling, A.E., and Ragan, M.A. (2009). Lateral transfer of genes and gene fragments in prokaryotes. Genome Biol. Evol. *1*, 429–438.
14. Chan, C.X., Darling, A.E., Beiko, R.G., and Ragan, M.A. (2009). Are protein domains modules of lateral genetic transfer? PLoS ONE *4*, e4524.
15. Lawrence, J.G., and Ochman, H. (1997). Amelioration of bacterial genomes: Rates of change and exchange. J. Mol. Evol. *44*, 383–397.
16. Chan, C.X., Beiko, R.G., and Ragan, M.A. (2006). Detecting recombination in evolving nucleotide sequences. BMC Bioinformatics *7*, 412.
17. Reyes-Prieto, A., and Bhattacharya, D. (2007). Phylogeny of Calvin cycle enzymes supports Plantae monophyly. Mol. Phylogenet. Evol. *45*, 384–391.
18. Tyra, H.M., Linka, M., Weber, A.P., and Bhattacharya, D. (2007). Host origin of plastid solute transporters in the first photosynthetic eukaryotes. Genome Biol. *8*, R212.
19. Colleoni, C., Linka, M., Deschamps, P., Handford, M.G., Dupree, P., Weber, A.P.M., and Ball, S.G. (2010). Phylogenetic and biochemical evidence supports the recruitment of an ADP-glucose translocator for the export of photosynthate during plastid endosymbiosis. Mol. Biol. Evol. *27*, 2691–2701.
20. Huang, J.L., and Gogarten, J.P. (2007). Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biol. *8*, R99.
21. Gross, J., and Bhattacharya, D. (2009). Mitochondrial and plastid evolution in eukaryotes: An outsiders' perspective. Nat. Rev. Genet. *10*, 495–505.
22. Parfrey, L.W., Grant, J., Tekle, Y.I., Lasek-Nesselquist, E., Morrison, H.G., Sogin, M.L., Patterson, D.J., and Katz, L.A. (2010). Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. Syst. Biol. *59*, 518–533.
23. Moustafa, A., Beszteri, B., Maier, U.G., Bowler, C., Valentin, K., and Bhattacharya, D. (2009). Genomic footprints of a cryptic plastid endosymbiosis in diatoms. Science *324*, 1724–1726.
24. Frommolt, R., Werner, S., Paulsen, H., Goss, R., Wilhelm, C., Zauner, S., Maier, U.G., Grossman, A.R., Bhattacharya, D., and Lohr, M. (2008). Ancient recruitment by chromists of green algal genes encoding enzymes for carotenoid biosynthesis. Mol. Biol. Evol. *25*, 2653–2667.
25. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: Architecture and applications. BMC Bioinformatics *10*, 421.
26. Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Res. *9*, 868–877.
27. Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucleic Acids Res. *34* (Web Server issue), W435–W439.
28. Edgar, R.C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

29. Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. *56*, 564–577.
30. Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics *22*, 2688–2690.
31. Whelan, S., and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. *18*, 691–699.
32. Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. *39*, 306–314.