Phylogenomic Analysis Identifies Red Algal Genes of Endosymbiotic Origin in
the Chromalveolates

Shenglan Li,¹ Tetyana Nosenko,¹ Jeremiah D. Hackett,² and Debashish Bhattacharya
Department of Biological Sciences and Roy J. Carver Center for Comparative Genomics, University of Iowa

Endosymbiosis has spread photosynthesis to many branches of the eukaryotic tree; however, the history of photosynthetic
genome (plastid) is still controversial. Fortuitously, endosymbiosis may leave a genomic footprint through the
transfer of endosymbiont genes to the “host” genome (endosymbiotic gene transfer, EGT). EGT can be detected
through comparison of host genomes to uncover the history of past plastid acquisitions. Here we focus on a lineage
of chlorophyll c–containing algae and protists (“chromalveolates”) that are postulated to share a common red algal sec-
condary endosymbiont. This plastid is originally of cyanobacterial origin through primary endosymbiosis and is closely
related among the Plante (i.e., red, green, and glaucophyte algae). To test these ideas, an automated phylogenomics
pipeline was used with a novel unigene data set of 5,081 expressed sequence tags (ESTs) from the haptophyte alga Emi-
liania huxleyi and genome or EST data from other chromalveolates, red algae, plants, animals, fungi, and bacteria. We
focused on nuclear-encoded proteins that are targeted to the plastid to express their function because this group of genes is
expected to have phylogenies that are relatively easy to interpret. A total of 708 genes were identified in E. huxleyi that had a
significant Blast hit to at least one other taxon in our data set. Forty-six of the alignments that were derived from the 708
genes contained at least one other plastidome (i.e., besides E. huxleyi), red and/or green algae (or land plants), and one or
two cyanobacteria, whereas 15 alignments contained E. huxleyi, one or more other plastidomes, and only cyanobacteria. Detailed phylogenetic analyses of these data sets turned up 19 cases of EGT that did not contain significant paralogy and had strong bootstrap support at the internal nodes, allowing us to confidently identify the source of the plastid-
targeted gene in E. huxleyi. A total of 17 genes originated from the red algal lineage, whereas 2 genes were of green algal
origin. Our data demonstrate the existence of multiple red algal genes that are shared among different chromalveolates,
suggesting that at least a subset of this group may share a common origin.

Introduction

An important challenge in evolutionary biology is to clarify the origin and spread of the photosynthetic organelle
(plastid) in algae and plants. It is believed that plastids ultimately trace their origin to a single primary endosymbiosis (fig. 1A); that is, the acquisition and retention of a cyanobacterium by a heterotrophic eukaryote (e.g., Bhattacharya and Medlin 1995; Delwiche 1999). Subsequently, the ∼2.5- to 5-million base pair genome of the pro-
karyote was reduced through outright gene loss and gene transfer (endosymbiotic gene transfer, EGT) to the “host”
nucleus (Martin and Hermann 1998), resulting in a typical plastid genome that encodes between 100 and 200 genes.
The products of the transferred genes are involved in photosynthesis, plastid (i.e., plastid-targeted proteins) with the evolution of an N-terminal extension that allowed passage of the protein through the two plastid membranes (McFadden 1999; McFadden and van Dooren 2004). Under this well-established scenario, in phylogenetic trees (fig. 1B), plastid-encoded and most
nuclear-encoded plastid-targeted proteins are sister to their cyanobacterial counterparts and are relatively more dis-
tantly related to eukaryotic homologs (Martin et al. 2002; Yoon et al. 2002).

The primordial alga that resulted from the primary endosymbiosis diverged into three lineages (fig. 1B), the
Rhodophyta (red algae), the Glaucophyta, and the Viridiplantae (green algae plus land plants), that together are referred to as the Plantae (Cavalier-Smith 2004). There is growing evi-
dence that the Plantae members share a close evolutionary relationship (e.g., Moreira, Le Guyader, and Philippe 2000;
Sanderson, and D. Bhattacharya, unpublished data). The remaining algae obtained their plastid through secondary or
tertiary endosymbiosis. In the first case, a nonphotosyn-
thetic protist engulfed a red or a green alga (Gibbs 1993) resulting in a “secondary” plastid (e.g., fig. 1C), whereas in the second case, an alga containing a secondary plastid was engulfed (e.g., Yoon et al. 2005). Until now, tertiary endosymbiosis is only known from the dinoflagellates (Tengs et al. 2000; Ishida and Green 2002; Bhattacharya, Yoon, and Hackett 2004; Yoon et al. 2005).

Following secondary endosymbiosis, the process of EGT also occurred from both the nuclear genome of the
engulfed alga that contained the genes encoding plastid-targeted proteins required for plastid function and from the
secondary plastid genome (McFadden 1999). The transferred genes evolved bipartite protein targeting signals for travers-
ing the 3–4 membranes of the secondary plastid (McFadden 1999). In phylogenies, these nuclear-encoded (now secondary)
plastid-targeted proteins are predicted to be nested within their donor taxa (red or green algae) that are sister to cyano-
bacterial homologs, as described above (see fig. 1D).

Here we used phylogenomics (e.g., Eisen and Fraser 2003; Huang et al. 2004) and a novel unigene data set of 5,081 expressed sequence tags (ESTs) from the haptophyte alga Emiliania huxleyi to study nuclear-encoded proteins that are targeted to the plastid in this species and in other related “chromalveolate” protists. The chromalveolates in-
clude the alveolates (dinoflagellates, apicomplexans, and ciliates) and chromists (cryptophytes, haptophytes, and

1 These authors have contributed equally to the manuscript.
2 Present address: Biology Department, Woods Hole Oceanographic Institution.

Key words: chromalveolates, Emiliania huxleyi, endosymbiosis, gene transfer, phylogenomics.

E-mail: debashi-bhattacharya@uiowa.edu.

doi:10.1093/molbev/msj075
Advance Access publication December 15, 2005

© The Author 2005. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.
For permissions, please e-mail: journals.permissions@oxfordjournals.org
stramenopiles) and are postulated to share a single red algal secondary endosymbiosis in their common ancestor (Cavalier-Smith 1999). Our analysis combined available genomic sequences (both complete genome and EST data) from plants, animals, fungi, bacteria, red algae, and chromalveolates. The preliminary trees identified with our phylogenomic pipeline were used as starting points for extensive database searches from which we prepared phylogenies that included the available sequence data. Our analyses resulted in 19 protein maximum likelihood trees that did not contain significant paralogy (allowing relative ease of interpretation) and had significant bootstrap support at the internal nodes to allow us to robustly identify the source of the EGT. A total of 17 genes were of the expected red algal origin, consistent with the chromalveolate hypothesis, whereas two genes had a green algal ancestry. Our data provide evidence for significant red algal EGT in chromalveolates and suggest that some of these taxa may share a monophyletic origin.

**Materials and Methods**

**Phylogenomic Pipeline**

We generated a set of 5,081 unique complementary DNAs (cDNAs) (J. D. Hackett, D. Bhattacharya, and M. B., Soares unpublished data) from the haptophyte alga *E. huxleyi* CCMP 1280 (as in Hackett et al. 2004). Briefly, total RNA from a culture of *E. huxleyi* was extracted using Trizol (Invitrogen, Carlsbad, Calif.) and the mRNA purified was extracted using the Oligotex mRNA Midi Kit (Qiagen, Santa Clarita, Calif.). Start and normalized cDNA libraries were constructed according to Bonaldo, Lennon, and Soares (1996). The cDNA clones were sequenced from the 3’ end, and plastid genes were identified through amino acid (aa) sequence similarity searches. Clustering of the 8,299 clones into 5,081 nonredundant sets was performed using the program Uiclust (Trivedi et al. 2002). The *E. huxleyi* ESTs have been released as one submission to the GenBank dbEST database (http://www.ncbi.nlm.nih.gov/dbEST/index.html, CX771261–CX779560). These data were used as input for the phylogenomics approach using the PhyloGenie package of computer programs (Lupas and Frickey 2004). PhyloGenie provides “high-throughput” phylogenetic reconstruction and serves as an automated pipeline in which the following analyses can be implemented: Blast search, extraction of homologous sequences from the Blast results, generation of alignments, phylogenetic tree reconstruction, and calculation of bootstrap support values for individual phylogenies. We created a local protein database for the Blast search by retrieving completed genome sequences from the National Center

The 5,081 DNA sequences were translated into the six possible open reading frames using the Transeq program in the Emboss package (http://emboss.sourceforge.net/). The final fasta file that included all of the data was formatted using the formatdb program in the Blast package (http://www.ncbi.nlm.nih.gov/BLAST/) and comprised the database for the PhyloGenie Blast search. We initially included a minimal set of 10 species in the local database comprising the dinoflagellates Alexandrium tamarense (Hackett et al. 2005) and Karenia brevis (Lidie et al. 2005), the diatom (stramenopile) Thalassiosira pseudonana (Armbust et al. 2004), the apicomplexan Plasmodium falciparum, the green alga Chlamydomonas reinhardtii, the red alga C. merolae (Matsuzaki et al. 2004), Drosophila melanogaster, Saccharomyces cerevisiae, the cyanobacterium Nostoc sp. PCC 7120, and Escherichia coli. We set the minimum Expect “e value” for the Blast search of these data at 10. To run PhyloGenie, it was necessary to give the java virtual machine the right to use up to 1,000 megabytes of memory via the “java -jar blammer.jar” command. Otherwise, the program did not run and returned the “java.lang.outOfMemoryError” message. All hits with an e value better than 0.01 were then taken to build the hidden Markov model (hmm) alignments. All other parameters were kept as default. The program TreeView (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html) was used to visualize the resulting trees.

To simplify the search for plastid-targeted proteins in chromalveolates, we retained trees in which only one gene copy or a small number of closely related copies were found for each species in the alignment. Two classes of trees were retained for more extensive analysis. The first class included E. huxleyi and at least one other chromalveolate (usually the diatom T. pseudonana), a red or green alga (i.e., C. merolae and/or C. reinhardtii), and Nostoc. All trees containing D. melanogaster and/or S. cerevisiae or lacking the cyanobacteria were excluded from this group. The second class addressed the potential issue that plastid-targeted proteins of cyanobacterial origin may be too highly divergent in the red and/or green algae to be easily identified with our pipeline and therefore included trees that contained E. huxleyi and at least one other chromalveolate and Nostoc. To verify these results, the candidate genes from the initial run were used as input for a second run under PhyloGenie with the addition in the local database of the predicted proteins from the following eight genome data sets and the EST data set from Toxoplasma gondii (available from NCBI): eukaryotes—Arabidopsis thaliana, Giardia intestinalis, Guillardia theta (nucleomorph genome), Trypanosoma brucei and prokaryotes—Halobacterium sp. NRC-1, Sulfolobus tokodaii, Synechococcus elongatus PCC 7942, Trichodesmium erythraeum. Trees that had complex patterns of gene family evolution such as deep paralogy with duplicated genes distributed across different eukaryotic lineages or had low bootstrap support at nodes (e.g., due to small protein size) were again discarded from subsequent analyses. This conservative approach most certainly resulted in an underestimation of the number of nuclear-encoded plastid-targeted genes in E. huxleyi and other chromalveolates but provided a manageable set of candidate trees for building the final in-depth alignments.

Building the Final Alignments

All potential homologs of the candidate E. huxleyi sequences that were used to build the final alignments were identified using Blast searches (e value \( \leq 10^{-10} \)) against the GenBank nonredundant (nr) and Expressed Sequence Tag (dbEST) and other databases. In particular, we focused on red algal and chromalveolate data including Galdieria sulphuraria (Weber et al. 2004; Michigan State University Galdieria Database http://genomics.msu.edu/galdieria/sequence_data.html), Porphyra yezoensis (Asamizu et al. 2003; http://www.kazusa.or.jp/en/plant/porphyra/EST), and Phaeodactylum tricornutum (Scala et al. 2002; http://avesthagen.szbowlver.com/). Overlapping aa sequences from each taxon were aligned using ClustalW and adjusted manually under BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html).

The intracellular destination of the proteins was inferred from the similarity to known plastid-targeted proteins (primarily in the annotated genomes of C. merolae and A. thaliana) and from analysis of the N-terminal extensions using the following transit peptide prediction programs: TargetP (plant version) (http://www.cbs.dtu.dk/services/TargetP), PlasmoAP (http://www.plasmodb.org/cgi-bin/plasmoap.cgi), and Prediction of Apicoplast-Targeted Sequences (PATS, http://gecco.org.chemie.uni-frankfurt.de/pats/pats-index.php). The length of the N-terminal extension predicted by TargetP was verified using the protein alignments. The annotated genome data from C. merolae were used to identify gene function in E. huxleyi and in other chromalveolates.

Phylogenetic Analysis

For each data set, a phylogeny was reconstructed under maximum likelihood (ML) using the PHYML V2.4.3 computer program (Guindon and Gascuel 2003) with the WAG + I + G evolutionary model and tree optimization. The alpha value for the gamma distribution was calculated using eight rate categories. To assess the stability of monophyletic groups in the ML trees, we calculated PHYML bootstrap (100 replicates) support values (Felsenstein 1985). In addition, we calculated bootstrap values (100 replications) using the neighbor joining (NJ) method with JTT + G distance matrices (PHYLIP V3.63, http://evolution.genetics.washington.edu/phylip.html). The NJ analysis was done with randomized taxon addition. Finally, we generated Bayesian posterior probabilities for nodes in the ML tree using MrBayes V3.0b4 (Huelsenbeck and Ronquist 2001) and the WAG + G model with Metropolis-coupled Markov chain Monte Carlo from a random starting tree. The Bayesian analyses were run for 1,000,000 generations with trees sampled each 1,000 cycles. Four chains were run
simultaneously of which three were heated and one was cold, with the initial 500,000 cycles (500 trees) being discarded as the “burn in.” A consensus tree was made with the remaining 500 phylogenies to determine the posterior probabilities at the different nodes.

Results and Discussion

The goal of our research was to identify the sources of nuclear-encoded plastid-targeted proteins in the haptophyte alga *E. huxleyi* and in the limited genome data available from other chromalveolates. To this end, we analyzed all genes of cyanobacterial origin in the nuclear genome of *E. huxleyi* that were found either only in chromalveolates or in chromalveolates and members of the Plantae. The expectation under the chromalveolate hypothesis, described later in detail, is that the majority of plastid-targeted genes of cyanobacterial origin in this group should be monophyletic and specifically related to the red algae (i.e., as sister to the Cyanidales; Yoon et al. 2002, 2005). Furthermore, the branching pattern within the chromalveolate clade should, in spite of the uncertainty associated with estimating ancient splits with single-protein trees (e.g., Graybeal 1998; Hillis et al. 2003; Rodriguez-Ezpeleta et al. 2005), be generally congruent between the different genes that originated through EGT, or at least they should not disagree with significant bootstrap support. It should also be noted that previous work has identified genes encoding plastid-targeted (and mitochondrial-targeted; Funes et al. 2002) proteins of green algal origin in chromalveolates (e.g., *hemB*; Hackett et al. 2004); therefore this is an alternative significant source of nuclear genes in our study group.

Chromalveolate Hypothesis

Chromalveolate monophyly would unify a broad assemblage of protists but remains strongly in question because of incomplete phylogenetic data. Analyses of plastid genes (e.g., Yoon et al. 2002; Hagopian et al. 2004; Bachvaroff, Sanchez Puerta, and Delwiche 2005) provide evidence that chromist plastids are closely related to each other and likely monophyletic, consistent with a single origin of the organelle in this group. The topology of plastid gene trees shows an early divergence of the cryptophytes with the haptophytes and stramenopiles forming a sister group. The highly divergent alveolate sequences are more difficult to place in plastid gene trees, but a recent analysis by Yoon et al. (2005) reveals that dinoflagellate secondary plastids have a weakly supported sister group relationship to the stramenopiles. Analysis of 10 plastid-encoded proteins from the dinoflagellate *Amphidinium operculatum* also places this species within (not sister to) the chromists (Bachvaroff, Sanchez Puerta, and Delwiche 2005). Plastid monophyly does not, however, prove the chromalveolate hypothesis because these organelles could potentially have resulted from multiple independent secondary endosymbioses involving closely related red algae or tertiary endosymbioses involving existing chromalveolates (e.g., a stramenopile origin of the dinoflagellate peridinin plastid).

Phylogenies of the host nuclear genes are equivocal with respect to chromalveolate monophyly. Trees inferred from concatenated nuclear data sets strongly support a sister group relationship between the stramenopiles and alveolates (e.g., Baldauf et al. 2000; Harper, Waanders, and Keeling 2005). However, the position of the cryptophytes and haptophytes remains uncertain with a recent analysis of a six-protein data set providing weak support for their monophyly. This group was however distantly related to the stramenopile + alveolate clade in the trees (fig. 1E; Harper, Waanders, and Keeling 2005).

An alternative approach to assess chromalveolate monophyly that was taken here is to study the phylogenies of nuclear genes encoding plastid-targeted proteins. Because the red algae (as members of the Plantae) are distantly related to chromalveolates (fig. 1E; e.g., Baldauf et al. 2000; Rodriguez-Ezpeleta et al. 2005), nuclear genes shared among chromalveolates that have a well-supported sister group relationship to the red algae would most likely ( barring multiple red algal horizontal transfers) have originated through EGT via the secondary endosymbiosis. These trees would not prove chromalveolate monophyly but rather test the prediction that genes of red algal origin are shared by the different members of this lineage. Such a finding would be most easily explained by a single origin of the genes in the chromalveolate common ancestor through EGT from a red algal endosymbiont. Nuclear-encoded proteins of red algal origin have been reported for several species of chromalveolates; e.g., *fisZ* in chromalveolates and cryptophytes (Miyagishima et al. 2004), *apI* and *apII* in dinoflagellates (Hackett et al. 2004), and genes of red algal origin involved in the amylopectin pathway have been found in apicomplexans (Coppin et al. 2005). Analyses of the plastid-targeted glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and fructose-1,6-bisphosphate aldolase (FBA) also support chromalveolate monophyly (Fast et al. 2001; Patron, Rogers, and Keeling 2004). These compelling data are, however, based on unusual evolutionary histories. In the first case, the gene encoding plastid-targeted GAPDH has resulted from the duplication of the cytosolic gene of the secondary plastid host and the retargeting of one of the copies to the plastid, whereas in the second case, the plastid-targeted FBA arose from the retargeting of a class II FBA gene of uncertain origin.

Results of the Phylogenomic Analysis

The initial phylogenomic analysis returned a total of 708 genes (i.e., the inferred protein sequences) in the *E. huxleyi* ESTs that had a significant Blast hit to at least one other taxon in our local database. Of these alignments, completion of the second round of analysis showed that 46 contained at least one other chromalveolate (i.e., besides *E. huxleyi*), a Plantae member, and one or more cyanobacteria, whereas 15 alignments contained *E. huxleyi*, other chromalveolates, and only cyanobacteria. This list of 61 genes is shown in table S1 (Supplementary Material online). Detailed phylogenetic analyses of these data sets resulted in 19 trees (*E. huxleyi* shown in large boldface text in each phylogeny) that contained plastid-targeted proteins that did not contain significant paralogy and had strong ML bootstrap support at the internal nodes, allowing us to confidently identify their sources in *E. huxleyi* (see table 1). We note that many genes that are of cyanobacterial origin and
encode plastid-targeted proteins (e.g., chlorophyll-binding proteins) gave rise to trees that were simply too complex for us to infer the history of each gene copy among the taxa and these were not considered further. In addition, by disposing of all trees that included non-Plantae taxa, we likely excluded genes or gene family members that encode plastid-targeted proteins, but the presence of taxa such as the ophistokonts suggests an ancient gene origin (or potentially ancient lateral transfer) that predates the Plantae. In our hands, the laborious process of gene selection was largely guided by the effect of taxon addition. The 19 genes considered here for in-depth analysis became more resolved and easier to interpret with the addition of taxa (e.g., the second round of phylogenomics), suggesting that they were useful phylogenetic markers with the available data. These genes were primarily annotated using the C. merolae genome data (see table 1) and a total of 17 originated from the red algal lineage, whereas 2 genes were of green algal origin. Although we do not have taxon sampling from each taxonomic group of chromalveolates in these trees, there are at least two different lineages represented in each phylogeny. In the following section, we describe in detail the putative function of several outstanding examples from this gene set and their inferred evolutionary histories (other trees [except geranyl geranyl diphosphate synthase] are found in figs. S2, S3, and S4 [Supplementary Material online]). We did not present the results of the analysis of PSBO and FTSZ because these have been previously reported (e.g., Ishida and Green 2002; Hackett et al. 2004; Miyagishima et al. 2004).

Thylakoid Lumen and Pentapeptide Proteins

Our analysis of the E. huxleyi ESTs revealed two conserved proteins that are members of the pentapeptide repeat protein family (PEP). From our Blast search, when using the haptophyte sequences as the query, significant hits were found to homologs in cyanobacteria, red algae, green algae, and plants (i.e., e value $10^{-10}$). One of these sequences showed a high similarity to the thylakoid lumen protein (TLP) that is annotated as plastid targeted in the dinoflagellate Heterocapsa triquetra (Patron et al. 2005) and in plants (Kieselbach et al. 1998).

### Table 1

Phylogenetic Affinities of Plastid-Targeted Proteins of Cyanobacterial (CB) Origin in *Emiliania huxleyi* (Ehux) and in Other Chromalveolates

<table>
<thead>
<tr>
<th>Ehux clone number</th>
<th>Affinity in Phylogenomic Pipeline</th>
<th>Cyanidioschyzon merolae Best Hit</th>
<th>Putative Plastid Function and Protein Name</th>
<th>Length in Cyanidioschyzon merolae</th>
</tr>
</thead>
<tbody>
<tr>
<td>UI-EH-HG1-aac-a-07-0-Uls1</td>
<td>CB + red + chromalveolate</td>
<td>CMQ241C</td>
<td>Plastid-specific 30S ribosomal protein (PSRP-1)</td>
<td>324 aa</td>
</tr>
<tr>
<td>UI-EH-HG2-aar-m-02-0-Uls1</td>
<td>—</td>
<td>CMK150C</td>
<td>Translation elongation factor G (EF-G)</td>
<td>770 aa</td>
</tr>
<tr>
<td>UI-EH-HG2-abf-h-21-0-Uls1</td>
<td>—</td>
<td>CMO201C</td>
<td>Hypothetical conserved pentapeptide protein</td>
<td>321 aa</td>
</tr>
<tr>
<td>UI-EH-HG2-aaz-g-11-0-Uls1</td>
<td>—</td>
<td>CMJ151C</td>
<td>Thylakoid luminal 17.4-kDa protein (P17.4)</td>
<td>288 aa</td>
</tr>
<tr>
<td>UI-EH-HG2-abc-h-12-0-Uls1</td>
<td>—</td>
<td>CMJ054C</td>
<td>Glutamyl-transfer RNA reductase</td>
<td>532 aa</td>
</tr>
<tr>
<td>UI-EH-HG2-aaar-j-22-0-Uls1</td>
<td>—</td>
<td>CMF290C</td>
<td>Manganese-stabilizing protein precursor (PSBO)</td>
<td>329 aa</td>
</tr>
<tr>
<td>UI-EH-HG2-aao-d-08-0-Uls1</td>
<td>—</td>
<td>CMS004C</td>
<td>Plastid division protein (FTSZ)</td>
<td>503 aa</td>
</tr>
<tr>
<td>UI-EH-HG2-aee-o-17-0-Uls1</td>
<td>—</td>
<td>CMQ234C</td>
<td>Dihydrolipoamide dehydrogenase (LpDH)</td>
<td>547 aa</td>
</tr>
<tr>
<td>UI-EH-HG2-aan-r-13-0-Uls1</td>
<td>—</td>
<td>CMH148C</td>
<td>FKBP-type peptidyl-prolyl cis-trans isomerase (FKBP_C)</td>
<td>540 aa</td>
</tr>
<tr>
<td>UI-EH-HG0-aah-I-09-0-Uls1</td>
<td>—</td>
<td>CMM270C</td>
<td>Magnesium chelatase subunit D (CHLD)</td>
<td>738 aa</td>
</tr>
<tr>
<td>UI-EH-HG2-aas-a-19-0-Uls1</td>
<td>—</td>
<td>CMX011C</td>
<td>Plastid ribosomal protein L10 (RPLJ)</td>
<td>240 aa</td>
</tr>
<tr>
<td>UI-EH-HG2-aaz-a-06-0-Uls1</td>
<td>—</td>
<td>CMK058C</td>
<td>Geranylgeranyl diphasphate synthase (GGPS)</td>
<td>388 aa</td>
</tr>
<tr>
<td>UI-EH-HG2-abc-I-24-0-Uls1</td>
<td>—</td>
<td>CMR476C</td>
<td>Glycerol-3-phosphate dehydrogenase</td>
<td>425 aa</td>
</tr>
<tr>
<td>UI-EH-HG0-aag-l-11-0-Uls1</td>
<td>—</td>
<td>CMJ035C</td>
<td>Phosphoglycerate kinase</td>
<td>477 aa</td>
</tr>
<tr>
<td>UI-EH-HG2-aau-i-17-0-Uls1</td>
<td>—</td>
<td>CMQ124C</td>
<td>Hypothetical conserved (seed maturation protein)</td>
<td>406 aa</td>
</tr>
<tr>
<td>UI-EH-HG2-aap-j-11-0-Uls1</td>
<td>—</td>
<td>CML119C</td>
<td>GDP cyclohydrolase</td>
<td>551 aa</td>
</tr>
<tr>
<td>UI-EH-HG2-abg-p-17-0-Uls1</td>
<td>—</td>
<td>CMQ210C</td>
<td>II/3,4-dihydroxy-2-butanone 4-phosphate synthase (DHBP synthase)</td>
<td>414 aa</td>
</tr>
<tr>
<td>UI-EH-HG2-aau-e-01-0-Uls1</td>
<td>—</td>
<td>CMH047C</td>
<td>mRNA-binding proteins (e.g., CSP41)</td>
<td>429 aa</td>
</tr>
<tr>
<td>UI-EH-HG0-aak-a-24-0-Uls1</td>
<td>CB + green + chromalveolate</td>
<td>CMF117C</td>
<td>Phosphoribokinase</td>
<td>444 aa</td>
</tr>
<tr>
<td>UI-EH-HG2-aau-e-05-0-Uls1</td>
<td>—</td>
<td>CMT220C</td>
<td>Chlorophyll a synthase (CHLG)</td>
<td>403 aa</td>
</tr>
</tbody>
</table>

**Note.**—The boxed proteins are gene family members.
plastid targeting (probability, \(p = 0.727\); predicted transit peptide cleavage site, TPlen = 23, \(p = 0.941\); TPlen = 34, respectively). Analysis of the N-terminal extension in \(C. \) merolae TLP and PEP (and \(C. \) reinhardtii TLP) with TargetP did not provide any support for plastid or mitochondrial targeting of these proteins.

The TLP and PEP protein sequences were included in a single alignment for the phylogenetic analysis with the branch connecting these paralogs (see filled circle in fig. 2A) used to root each subtree (134 aa, fig. 2A). Although all the nodes within subtrees are not fully resolved, likely due to the small size of the data set, the ML tree is most easily
interpreted as supporting a cyanobacterial origin of the TLP and PEP genes in plants and algae. The gene duplication that gave rise to these genes occurred in the cyanobacteria prior to their transfer into the Plantae nuclear genome following primary endosymbiosis. These genes entered the nucleus of chromalveolates (i.e., alveolates, haptophytes, and stramenopiles for TLP) via secondary EGT from a red alga. A similar evolutionary history is found for the duplicated mRNA-binding proteins (see fig. S3, Supplementary Material online). These topologies are predicted by the chromalveolate hypothesis (fig. 1D), although the data do not allow us to resolve the branching order within the chromalveolate clade. To test these findings, we concatenated the TLP and PEP protein sequences into a single alignment (269 aa) to gain more phylogenetic resolution. The A. tamarense TLP and H. triquetra PEP proteins were combined to create a “Dinoflagellate” sequence in this data set. The ML and NJ bootstrap support values supporting the cyanobacterial origin of TLP and PEP were 100% in the concatenated protein tree (fig. S1, Supplementary Material online), and the red algal origin of these genes in chromalveolates and the monophyly of the chromalveolate clade was supported by the Bayesian inference (prob > 0.95, 1.0) and by the ML (83%, 70%) and NJ (62%, 64%) bootstrap analyses, respectively.

FKBP_C Protein

FKBP_C, or trigger factor-like protein, is a member of the FKBP family of immunophilins (He, Li, and Luan 2004) and is composed of a central FKBP domain that has a peptidyl-prolyl cis-trans isomerase activity and N-terminal and C-terminal ribosome-associated chaperone domains (Zarnt et al. 1997). The gene encoding FKBP_C has been found in all groups of eubacteria and in the plastids of algae and plants (Matsuzaki et al. 2004; Romano et al. 2005) but not in other eukaryotes, suggesting a cyanobacterial origin of the algal/plant gene. This hypothesis is robustly supported by the FKBP_C ML tree (321 aa, fig. 2B), in which the algal/plant FKBP_C sequences are sister to homologs in cyanobacteria (ML/NJ bootstrap values = 100%). In plants, the FKBP_C gene is nuclear encoded, synthesized in the cytosol, and targeted to the plastid stroma (He, Li, and Luan 2004). FKBP_C in the red alga C. merolae also contains an N-terminal extension; however, TargetP was unable to detect a significant targeting signal in this protein. Although the function of eukaryotic FKBP_C remains unknown, its structural similarity to the bacterial homolog suggests that this protein may be involved in protein synthesis and targeting processes in plant plastids (Romano et al. 2005).

Our analysis shows that this gene is nuclear encoded in three chromalveolates: E. huxleyi, and the two diatoms P. tricornutum, and T. pseudonana. The ML tree provides moderate bootstrap support for the monophyly of chromalveolate FKBP_C (ML = 89%) and their close evolutionary relationship to homologs in the red algae (ML = 79%, NJ = 71%; as in fig. 1D). Within eukaryotes there is, as would be expected, a close relationship between the diatom sequences (i.e., P. tricornutum and T. pseudonana; ML = 100%, NJ = 100%) and a sister group relationship between chlorophyte (Ulva linza) and land plant FKBP_C (ML = 95%, NJ = 93%).

Hypothetical Plastid-Targeted Protein aau170

Analysis of the E. huxleyi EST library with PhyloGe-nie revealed two closely related genes of a hypothetical nuclear-encoded protein that is clearly of cyanobacterial origin. We named these proteins aau170, corresponding to a shortened version of the identification number of the E. huxleyi clone encoding one of these copies (U1-EH-HG2-aaui1-17-0-ULs1, see table 1). The Blast searches against GenBank (BlastP against the nr database and TBlastN against the est_others database) and against our local database revealed homologs in plants (>30 species, many of these were annotated as seed maturation-like protein), photosynthetic protists (green algae, red algae, stramenopiles, haptophytes), and apicomplexans. According to TargetP, the plant (e.g., Asparagus officinalis [prob = 0.939; TPlen = 82], A. thaliana [prob = 0.869; TPlen = 58]) homologs contain N-terminal extensions for plastid targeting, whereas the sequence in C. merolae was predicted to have a similar targeting potential for either the mitochondrion (prob = 0.525; TPlen = 76) or the plastid (prob = 0.544; Tplen = 76). The two aau170 homologs from T. pseudonana contain typical stramenopile bipartite plastid-targeting sequences that consist of a 33 aa–long N-terminal signal peptide followed by a 29 aa–long transit peptide. According to PlasmoAP, the significant N-terminal extensions in P. falciparum and Plasmodium yoelii aau170 did not encode an apicoplast-targeting signal (3/5 tests returned positive). However, analysis of these sequences with PATS suggested the existence of full-length apicoplast-targeting signals in these taxa (P. falciparum, prob = 1.00; P. yoelii, prob = 0.996).

Phylogenetic analysis of this data set (192 aa) provides strong bootstrap (ML = 89%, NJ = 87%) and Bayesian support for a single origin of the aau170 genes in chromalveolates from a red algal source (ML = 94%, NJ = 91%, fig. 2C). The sister group relationship between the red algal chromalveolate and the chlorophyte (C. reinhardtii) + land plant clades (ML = 99%, NJ = 97%) and the close phylogenetic relationship of these eukaryotic sequences to homologs in cyanobacteria fit in well with the scenario shown in figure 1. These data argue strongly for the existence of a red algal endosymbiont in apicomplexans that is shared with the stramenopiles and haptophytes, consistent with the findings of Coppin et al. (2005).

Plastid-Specific 30S and L10 Ribosomal Proteins

Plastid-specific 30S ribosomal protein (PSRP-1) is a member a family of proteins that are believed to bind to ribosomes and regulate protein translation in the plastid (Yamaguchi and Subramanian 2000). A likely homolog of PSRP-1 in E. coli (protein Y) has also been implicated in the regulation of translation during cold shock (for a review, see Wilson and Nir-Haras 2004). Homologs of PSRP-1 are widespread in cyanobacteria, other eubacteria (known as Sigma 54 modulation protein or S30EA ribosomal protein in this group), and in sequenced land plant genomes. Analysis of the N-terminal extensions in plants such as Spinacia
oleracea (prob = 0.975; TPlen = 64) and Lycopersicon esculentum (prob = 0.929; TPlen = 75) suggest strongly that these proteins are plastid targeted, whereas TargetP is unable to find a targeting signal for the C. merolae PSRP-1 homolog. Phylogenetic analyses confirm the suspected cyanobacterial origin of the nuclear gene encoding PSRP-1 (Johnson, Kraft, and Subramanian 1990) with the ML tree providing strong bootstrap support for the origin of haptophyte and stramenopile PSRP-1 from a red algal source (ML = 90%, NJ = 89%). PSRP-1 in the distantly related chlorarachniophyte amoeba Bigelowiella natans is monophyletic with the chromalveolate clade, but this likely represents an independent lateral transfer of a chromalveolate gene into this organism (see Archibald et al. 2003).

The ML tree inferred from the plastid-targeted L10 ribosomal protein also conforms to the expectations of the chromalveolate hypothesis (fig. 2E). This protein contains a strong signal for plastid targeting in plants (e.g., A. thaliana [prob = 0.887; TPlen = 40], O. sativa [prob = 0.763; TPlen = 50]) and in C. merolae (prob = 0.769; TPlen = 36). L10 in the cryptophyte G. theta has been annotated as being plastid targeted (GenBank CAH25357). The chromalveolates form a monophyletic group in the L10 tree with weak ML (69%) bootstrap support and with Bayesian (P = 1.0) support. The interrelationships of the chromalveolate taxa conform to the expectation from plastid gene trees with cryptophytes as sister to a clade defined by the haptophytes and stramenopiles (e.g., Yoon et al. 2004, 2005). However, this is the case only when the highly divergent C. merolae sequence is removed from the analysis. This red alga branches, without bootstrap support, as sister to E. huxleyi (see filled square in fig. 2E) when retained in the data set.

Magnesium-Chelatase Subunits CHLI and CHLD

Magnesium-chelatase is involved in chlorophyll synthesis; that is, $\text{Mg}^{2+}$ is inserted into protoporphyrin IX to form Mg-protoporphyrin IX (Romano et al. 2005). This function is usually carried out by an association of three protein subunits: CHLD, CHLH, and CHLI (Jensen et al. 1999; Gibson et al. 1995). The genes for plastid-encoded CHLI and nuclear-encoded CHLD are related through a gene duplication and fusion event and share about 40% aa identity (Jensen et al. 1996). We identified one chlD sequence that has homologs in cyanobacteria, red algae, land plants, and chromalveolates. CHLD in all the studied land plants contains a plastid-targeting signal (e.g., Pisum sativum, prob = 0.831; TPlen = 51, A. thaliana, prob = 0.822; Tplen = 49), and this protein appears to be plastid targeted in C. merolae (prob = 0.641; TPlen = 56). CHlD is found in the plastid of plants and all algae (Jensen et al. 1996) except peridinin-containing dinoflagellates (Hackett et al. 2004; Bachvaroff et al. 2004). We analyzed CHLD and CHLI separately to compare the topology of these trees.

The CHLI (324 aa) tree (fig. 3A) is typical (i.e., fig. 1B) for plastid-encoded proteins (Yoon et al. 2005). The monophyly of the green and red + chromalveolate clades, their distant phylogenetic relationship to the glaucophyte Cyanophora paradoxa, as well as the origin of this gene from a cyanobacterial primary endosymbiont are supported (the latter only weakly) in the ML tree. The CHLD phylogeny (564 aa) provides essentially the same topology with moderate bootstrap support (ML = 92%, NJ = 64%) for the sister group relationship of chromalveolates and red algae (fig. 3B). Again, this result is consistent with the scenario shown in figure 1D and implies that the chlD gene had a cyanobacterial origin in the Plantae and that the chromalveolates most likely obtained this sequence through red algal EGT.

Genes of Green Algal Origin

Of the 20 protein ML trees that were analyzed in detail after the phylogenomics approach, two (chlorophyll a synthase, phosphorubulokinase) suggested a green algal rather
than a rhodophyte ancestry for nuclear-encoded plastid-targeted genes in chromalveolates. A green algal contribution to alveolate nuclear genomes or the existence of a plastid of green algal origin in these taxa has been previously suggested and hotly debated (e.g., Funes et al. 2002) although these data are inconclusive. It is unclear whether the best known examples, apicoplast-encoded elongation factor *tufA* (Köhler et al. 1997), nuclear-encoded *hemB* (Hackett et al. 2004), and the nuclear-encoded mitochondrial-targeted *cox2a* and *cox2b* subunits (Funes et al. 2002), reflect independent lateral transfer events or result from EGT from a green alga that was once resident in the cell (for details, see Hackett et al. 2004). Our data do not conclusively resolve this issue but certainly suggest that the red algal contribution was substantial and appears to be a magnitude higher than that of green algae.

**Chlorophyll a Synthase**

Chlorophyll a synthase catalyzes the final step in chlorophyll biosynthesis, the introduction of the tetrapteryl side chain, and is also implicated in the regulation of photosynthesis (Schmid et al. 2001). This well-studied enzyme (included in the Pfam UbiA prenyltransferase family) is widespread in cyanobacteria and other eubacteria and in plant and algal nuclear genomes. In our Blast searches, the eukaryotic genes were more closely related to cyanobacterial orthologs (< 10^{-100}) than to orthologs in other eubacteria (approximately 10^{-50} to 10^{-20}) supporting their origin in algae/plants through primary endosymbiosis. The plant proteins in our alignment all contain a plastid-targeting signal (e.g., *A. thaliana*, prob = 0.849; Tplen = 57, *Avena sativa*, prob = 0.947; Tplen = 45) as does this protein from *C. merolae* (prob = 0.820; Tplen = 50). The protein ML tree of chlorophyll a synthase (fig. 4A) provides strong support for the monophyly of the plastid proteins (ML = 100%, NJ = 98%) relative to the cyanobacteria, and the NJ (74%) and Bayesian (prob = 1.0) methods suggest a green algal origin of the gene in chromalveolates and in *B. natans*.

**Phosphoribulokinase**

Class II phosphoribulokinase (PRK) is found in most photosynthetic organisms including cyanobacteria, photosynthetic algae, and land plants. Along with ribulose bisphosphate carboxylase/oxygenase, GAPDH, fructose-1,6-bisphosphate, and sedoheptulose-1,7-bisphosphatase (SBPase), PRK is considered to be a key enzyme of the Calvin cycle (Graciet, Lebreton, and Gontero 2004) and is involved only in Calvin cycle–specific functions. In eukaryotes, PRK is located in the plastid stroma where it catalyzes the conversion of ribulose-5-bisphosphate and adenosine triphosphate to ribulose-1,5-bisphosphate and adenosine 5′-diphosphate at the final step of ribulose bisphosphate regeneration (Michels, Wedel, and Kroth 2005; Porter et al. 1986). Cytosolic isoforms of PRK have not been reported thus far. As would be expected, PRK from plants have a strong plastid-targeting signal (e.g., *O. sativa*, prob = 0.947; TPlen = 51, *A. thaliana*, prob = 0.739; Tplen = 54), and a targeting signal is also detected by TargetP for *C. merolae* PRK (prob = 0.638; TPlen = 44).

Phylogenetic analysis of PRK confirms its cyanobacterial origin in algae and plants (fig. 4B). The eukaryotic clade is divided into the red (NJ = 86%), chromalveolate (ML = 66%, NJ = 83%), and green lineages (ML = 96%, NJ = 94%). The Bayesian and bootstrap (ML = 100%, NJ = 93%) support for the monophyly of chromalveolate and “green” PRKs supports a green algal origin of this gene in at least the dinoflagellates (*A. tamarense*, *Amphidinium carteri*, *Heterocapsa triquetra* [see arrow in fig. 4B]), haptophytes, and stramenopiles. Within the chromalveolates, there is significant Bayesian (prob = 1.0) and bootstrap

---

**FIG. 4.—Examples of green algal lateral gene transfer in the chromalveolates.** (A) ML tree of chlorophyll a synthase. (B) ML tree of phosphoribulokinase. These phylogenies are rooted with the cyanobacteria, and the numbers above and below the branches are the results of ML and NJ bootstrap analyses, respectively. The thick branches indicate ≥0.95 posterior probability from Bayesian inference. Only bootstrap values ≥60% are shown. Branch lengths are proportional to the number of substitutions per site (see scale bars). Lineage designations are as in figure 2. The position of the highly divergent alveolate taxa (i.e., three dinoflagellates) is marked with the arrow. For details, see figure S4 (Supplementary Material online).
(ML = 100%) support for a close relationship between the haptophytes and dinoflagellates (see fig. S4, Supplementary Material online). This result may, however, reflect the highly variable PRK divergence rates among the different eukaryotic clades combined with a shared rate elevation in haptophytes and dinoflagellates could that lead to artifactual long-branch attraction.

Summary

Understanding the role of EGT in shaping algal and plant genomes remains a significant challenge in comparative genomics (e.g., Martin et al. 2002; Lopez-Juez and Pyke 2005). Here we have used phylogenomics to address the origin of nuclear-encoded plastid-targeted proteins in chromalveolates with the goal of testing the monophyly of this group and assessing whether these protein trees show a topology relative to the cyanobacteria that is expected for the Plantae (see fig. 1). Our analyses identified 19 ML trees that are straightforward to interpret. Seventeen of the trees support strongly a red algal origin of chromalveolate plastid-targeted proteins and two support a green algal origin. These data provide evidence for extensive red algal EGT in chromalveolates and add to the growing support for a close evolutionary relationship of members of this group (e.g., Yoon et al. 2002; Fast et al. 2001; Patron, Rogers, and Keeling 2004; Harper, Waanders, and Keeling 2005).

Although we suggest that the chromalveolate hypothesis is the most parsimonious explanation for our data, there are several caveats to our analysis. First, the cryptophytes are missing from all of our nuclear data sets except for the L10 ribosomal protein (fig. 2E) and glutamyl–transfer RNA reductase (fig. S2, Supplementary Material online) trees. In the latter case, the cryptophyte G. theta does not group with the other chromalveolates (with moderate bootstrap support). Whether this result reflects the poor resolution associated with a single-protein analysis or a potential independent origin of this plastid or the plastid-targeted gene remains to be determined. Clearly, cryptophytes need to be included in a larger number of nuclear gene trees to determine their position within the chromalveolates. Second, the relative positions of chromalveolate members are in conflict between the different trees. The existing nuclear gene trees (e.g., Baldauf et al. 2000; Harper, Waanders, and Keeling 2005) suggest that stramenopiles and alveolates should form a monophyletic group, whereas plastid gene/genomes trees (e.g., Hagopian et al. 2004; Bachvaroff, Sanchez Puerta, and Delwiche 2005; Yoon et al. 2005) suggest a sister group relationship between haptophytes and stramenopiles. Both these results (as well as other topologies) are found in our trees (e.g., stramenopiles + alveolates; aaui170 [fig. 2C], dihydrolipoamide dehydrogenase [fig. 2S, Supplementary Material online]), presently making it impossible to ascertain the true interrelationships of chromalveolates. Taxon sampling from additional chromalveolates may address this issue, although any single tree may not unambiguously support a given topology due to a deficit of phylogenetic signal (see Yoon et al. 2005).

In conclusion, our results lead to three major insights into endosymbiosis; (1) uncovering sufficient and convincing examples of EGT will likely require a concerted comparative genomic approach (e.g., Martin et al. 2002) rather than reliance on anecdotal findings; (2) the present results suggest that only a small subset of candidate genes will likely be of sufficient length, conservation, and free of extensive paralogy to address ancient gene transfer events; and (3) although our data are consistent with the chromalveolate hypothesis, the often unresolved interrelationships of chromalveolates in our trees leaves open the possibility that the transferred genes of red algal origin may have originated through independent horizontal transfers in different lineages or through a series of endosymbioses involving red algae or algae containing red algal endosymbionts. A resolved host tree of chromalveolates combined with a more detailed understanding of endosymbiotic events in this lineage will ultimately prove or modify this fascinating model of eukaryotic evolution. And finally, our analysis identified a novel putatively plastid-targeted protein (aaui170, according to PATS) that is conserved across photosynthetic eukaryotes and in apicomplexans. Phylogenomics offers, therefore, the opportunity to identify novel proteome components in algal/plant plastids and in the apicoplast. These proteins could be of potential importance in understanding organelle function.

Supplementary Material

Supplementary table S1, figures S1–S4, and figure 2S and are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by grants from the National Science Foundation awarded to D.B. (MCB 02-36631, EF 04-31117), S.L. and T.N. were partially supported by an Avis E. Cone Research Fellowship from the University of Iowa. J.D.H. was supported by an Institutional NRSA grant (T 32 GM98629) from the National Institutes of Health. We are grateful to T. Frickey for assistance with the use of PhylOGenie.

Literature Cited


Phylogenomic Analysis of Chromalveolates 673


Martin Embley, Associate Editor

Accepted December 12, 2005