MOLECULAR EVOLUTION OF THE EUKARYOTIC TRANSLATION ELONGATION FACTOR, EFL

by

GILLIAN HEATHER GILE
B.S. Humboldt State University, 2005

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES
(Botany)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

August 2009

ABSTRACT

The eukaryotic translation elongation factor EFL (for EF-Like) is a paralogue of the better-known elongation factor 1-alpha (EF-1α), which brings aminoacyl-tRNAs to the ribosome during translation. This essential protein was thought to be ubiquitous in eukaryotes until the recent discovery of EFL in a small number of diverse, mainly unicellular, eukaryotic organisms that were found to lack EF-1α. Because of the great evolutionary distances between EFL-encoding lineages and the near mutual exclusivity of the two proteins, the observed complex distribution of EFL was initially attributed entirely to multiple lateral gene transfers. In the enclosed chapters, the distribution of EFL was characterized in more detail in four distantly related eukaryotic lineages at both fine and broad taxonomic scales in order to better understand the effects that endosymbiotic gene transfer, differential loss, and lateral gene transfer have had on the molecular evolution of EFL. Endosymbiotic transfer of EFL was detected in the chlorarachniophytes, a group of algae whose secondary plastids retain a vestigial nucleus, known as a nucleomorph, in their reduced eukaryotic cytoplasm, known as the periplastid compartment (PPC). The endosymbiotically transferred EFL carries a bipartite targeting sequence similar to those of plastid-targeted proteins in this group and to plastid- and PPC-targeting sequences in cryptomonads to direct it to the PPC, suggesting similarities in the way these two lineages have solved their shared challenge of targeting to complex plastids with nucleomorphs. No clear phylogenetic evidence for lateral transfer of EFL has yet emerged; rather, differential loss of EFL and EF-1α from an ancestral state of co-occurrence was characterized in euglenozoans and detected in publicly available data from heterokonts and opisthokonts, unexpectedly revealing a significant role for this process in shaping the complex distribution of EFL and EF-1α. This finding serves as a cautionary reminder that adequate taxon sampling and a robust organismal phylogenetic hypothesis are crucial in order to correctly infer lateral gene transfer.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

My PhD work has been supported financially by scholarships and fellowships from a number of sources. I wish to thank the University of British Columbia for their support through an entrance scholarship and a University Graduate Fellowship, the Botany department for support through the Frances Chave memorial scholarship, and especially the Natural Sciences and Engineering Research Council of Canada for granting me a Postgraduate Doctoral Fellowship and a Postdoctoral Fellowship to allow me to continue my professional development.

My co-authors and I would like to thank the following people for their assistance with particular projects. Claudio Slamovits shared EST data from *Oxyrrhis marina* (chapter 2) and *Gymnochlora stellata* (chapter 5). EST sequencing and the taxonomically broad EST database (TBestDB) were supported by the Protist EST Program (PEP) of Genome Canada, Génome Québec, Genome Atlantic, and the Atlantic Canada Opportunities Agency (Atlantic Innovation Fund) (chapters 2, 4, and 5). Alastair Simpson and Susana Breglia provided genomic DNA from *Neobodo saliens* and *Entosiphon sulcatum*, respectively, and Sarah Jardeleza shared euglenozoan elongation factor primer sequences (chapter 4). Geoff Noble provided technical assistance (chapter 2). Their assistance is greatly appreciated.

I would especially like to thank the many people whose help and support allowed me to succeed in my graduate studies. These include all members of the Keeling, Fast, and Leander labs for collegiality and useful discussions, but more specifically Raheel Humayun for getting me started with PCR, Lena Burri for teaching me the basics of Western blots, Audrey de Koning and Matthew Rogers for introducing me to the Unix command line and phylogenetic analysis, Aleš Horak for advanced help in phylogenetics techniques, and Chitchai Chantangsi for many useful discussions and commiserations. I thank the members of my supervisory committee, Drs. Sean Graham, Jim Berger, and Brian Leander for their advice and guidance. I am especially indebted to my supervisor, Patrick Keeling, for all of the invaluable training, support, and guidance he has provided me over the course of my graduate studies.

CO-AUTHORSHIP STATEMENT

Chapter 2 is based on a published manuscript: Gile GH, Patron NJ, Keeling PJ. 2006. EFL GTPase in Cryptomonads and the Distribution of EFL and EF-1alpha in Chromalveolates. Protist. 157:435-444. The project was conceived of by PJ Keeling. EST data used in this study came from libraries generated by NJ Patron. I conducted all laboratory work and data analyses and PJ Keeling and I wrote the manuscript jointly.

Chapter 3 is based on a published manuscript: Gile GH, Novis PM, Cragg DS, Zuccarello GC, Keeling PJ. 2009. The Distribution of EF-1α, EFL, and a Non-canonical Genetic Code in the Ulvophyceae: Discrete Genetic Characters Support a Consistent Phylogenetic Framework. Journal of Eukaryotic Microbiology. 56:367-372. The project was suggested by PJ Keeling. DS Cragg, PM Novis, and I conducted laboratory work. I conducted data analyses with assistance from PM Novis and I wrote the manuscript. PJ Keeling provided significant insights into improving the manuscript.

Chapter 4 is based on a published manuscript: Gile GH, Faktorová D, Castlejohn CA, Burger G, Lang BF, Farmer MA, Lukeš J, Keeling PJ. 2009. Distribution and Phylogeny of EFL and EF-1α in Euglenozoa Suggest Ancestral Co-occurrence Followed by Differential Loss. PLoS ONE 4:e51162. PJ Keeling and J Lukeš initiated the project. I conducted most of the laboratory work with contributions from D Faktorová and CA Castlejohn. EST data used in this study came from libraries generated by G Burger and BF Lang. I conducted all data analyses and wrote the manuscript. All co-authors provided suggestions to improve the final manuscript.

Chapter 5 is based on a published manuscript: Gile GH and Keeling PJ. 2008. Nucleus-Encoded Periplastid-Targeted EFL in Chlorarachniophytes. Mol. Biol. Evol. 25:1967-1977. PJ Keeling conceived of the project. I conducted all laboratory work and data analyses and wrote the manuscript. PJ Keeling provided significant insights into improving the manuscript.

CHAPTER 1: INTRODUCTION

Literature Review

Structure and function of EF-1α

The eukaryotic translation elongation factor EFL (for Elongation Factor-Like) is a paralogue of the better-known Elongation Factor-1α (EF-1α), which plays an essential role in translation by depositing aminoacyl-tRNAs (aa-tRNAs) in the A-site of the ribosome. EF-1α is a moderate sized protein made up of an N-terminal GTPase domain and two β-barrel domains (Andersen et al. 2000). In its active, or GTP-bound conformation, EF-1α binds an aa-tRNA and approaches the ribosome where the tRNA enters the A-site (Andersen, Nissen, and Nyborg 2003; Negrutskii and El'skaya 1998). Correct codon-anticodon pairing allows EF-1α to align with an active site on the ribosome's surface, which stimulates EF-1α to cleave GTP. Upon GTP cleavage, EF-1α undergoes a conformational change in which the β-barrel domains swing away from the GTPase domain, releasing the aa-tRNA and causing EF-1α to spring away from the ribosome (Nilsson and Nissen 2005). EF-1α must bind to its guanine nucleotide exchange factor (GEF) EF-1β in order to reverse the conformational change and allow it to replace GDP with GTP for an additional cycle of peptide elongation (Andersen, Nissen, and Nyborg 2003).

EF-1α is one of the most abundant proteins in mammalian cells, second only to actin, making up 1-2% of total cellular protein, well in excess of ribosomes and other translational GTPases, despite its moderate size of ~ 50 kDa (Slobin 1980). In accordance with this observation, EF-1α is known to participate in other cellular functions besides translation. It can also bind and bundle actin filaments and microtubules (Edmonds et al. 1998; Gross and Kinzy 2005; Nakazawa et al. 1999; Yang et al. 1990). While bound to actin, EF-1α can simultaneously bind certain mRNAs, localizing them in the cell (Liu et al. 2002; Mickleburgh et al. 2006). EF-1α has also been implicated in ubiquitin-dependent protein degradation (Chuang et al. 2005; Gonen et al. 1994).

EF-1α in evolutionary biology

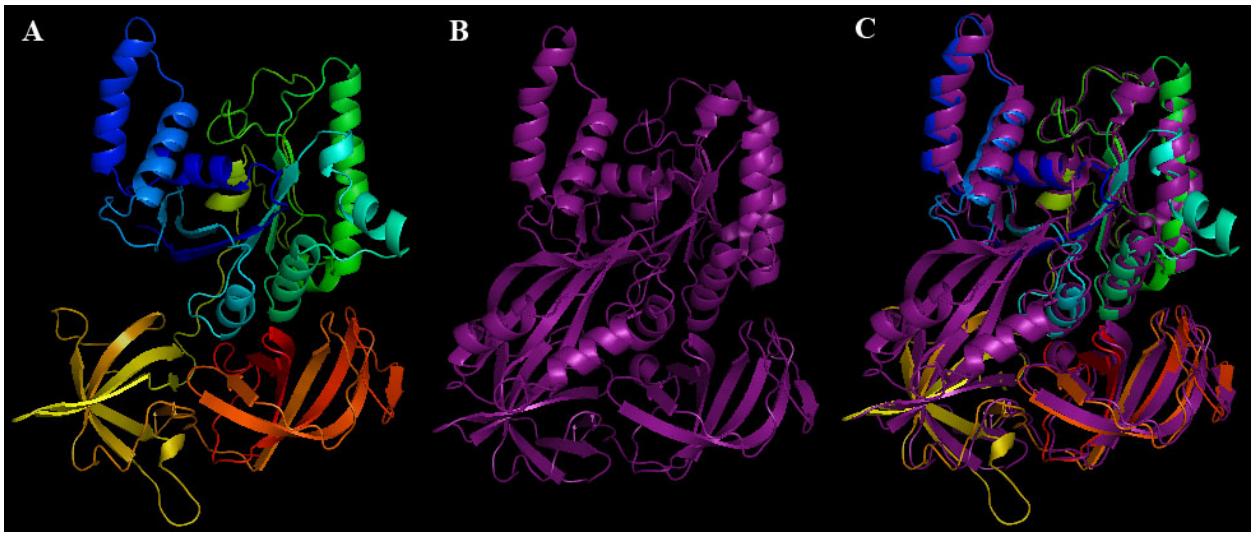
The ubiquity, overall conservation, and thorough characterization of EF-1α have proven to be useful characteristics for molecular evolutionary studies. EF-1α and its paralogue, EF-2 (called EF-Tu and EF-G in bacteria) are found in all three domains of life, so they were able to be used as outgroups to one another in order to root the universal tree of life (Baldauf, Palmer,

and Doolittle 1996; Iwabe et al. 1989). EF-1α sequences from diverse eukaryotes are well enough conserved that they can be aligned for deep-level phylogenetic reconstruction (Baldauf and Palmer 1993; Baldauf and Doolittle 1997; Hashimoto and Hasegawa 1996), though mutational saturation and lineage specific rate shifts limit EF-1α's ability to resolve the deepest eukaryotic relationships (Roger et al. 1999). These lineage specific rate shifts, also known as covarion behavior, in turn made EF-1α useful for developing models to extract phylogenetic signal from protein alignments displaying this characteristic (Lopez, Forterre, and Philippe 1999). Because EF-1α is functionally well characterized, its covarion behavior was further exploited to develop methods of predicting functional divergence from protein alignments (Gaucher, Miyamoto, and Benner 2001; Gaucher et al. 2002; Inagaki et al. 2003).

EFL can likely perform the same function as EF-1α

In all genomes so far found to lack EF-1α, EFL was identified as the most similar and thus the most likely protein to take over EF-1α's essential role in translation. Although EFL has not been characterized at the structural or functional levels, its high sequence similarity to EF-1α (typically 40-45% identity at the amino acid level) suggests a nearly identical structure (Figure 1.1). The sequence similarity is particularly pronounced at sites corresponding to EF-1α's binding sites for GTP/GDP, aa-tRNA, and EF-1β, which in turn do not significantly overlap with predicted functionally divergent regions, suggesting an ability to perform the same canonical function in translation elongation (Keeling and Inagaki 2004). Similar analyses comparing EF-1α to its other close paralogues, the eukaryotic release factor eRF3 and the heat shock suppressor HBS1, which are known to have different cellular functions, successfully predicted functional divergence at these sites, thereby helping to validate the method and support the inference that EFL is capable of EF-1α's core translation function (Inagaki et al. 2003). While EFL and EF-1α likely do not share a sister relationship in their family of related GTPases, this phylogenetic hypothesis cannot be rejected (Keeling and Inagaki 2004). Regardless of which protein is most closely related to EFL, altogether the evidence suggests that it can fill EF-1α's role in translation.

Figure 1.1. Comparison of tertiary structures of EFL and EF-1α.



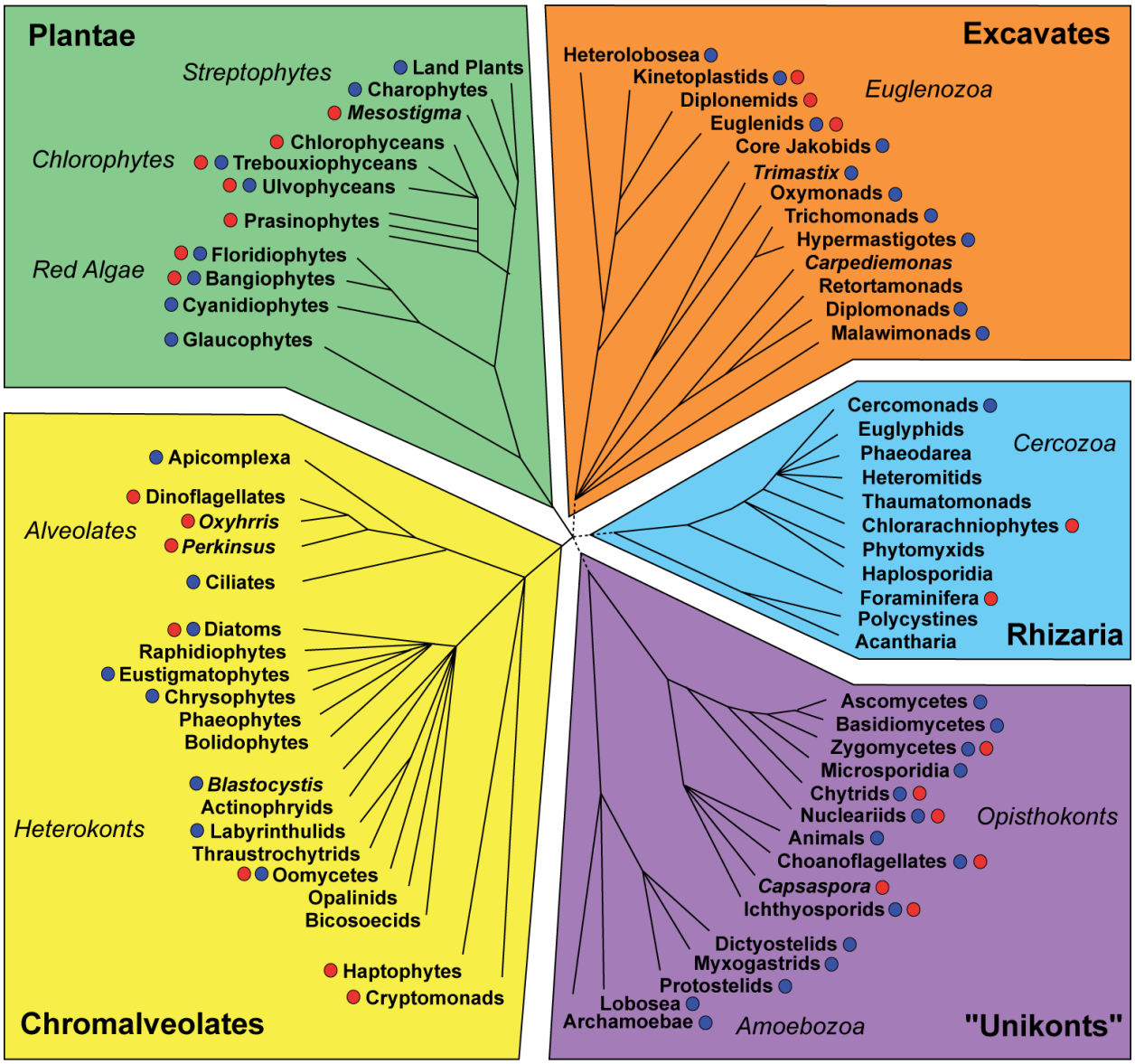
Tertiary structures of EF-1α and EFL. A) Homology model of *Bigelowiella natans* cytosolic EFL based on yeast EF-1α generated by SWISS-MODEL. B) Model of yeast EF-1α in complex with EF-1β, (IF60.pdb, Andersen et al. 2000) based on the solved crystal structure. C) Homology model of EFL from *B. natans* superimposed on yeast EF-1α. The model of EF-1β can be seen as a central domain in purple only with no equivalent in the model of EFL.

EFL and EF-1α have a complex distribution

EFL and EF-1α are each monophyletic in protein phylogenies, but the organisms in which they are found are not each other's closest relatives. EFL is scattered across the tree of eukaryotes in a complex pattern and its presence is nearly mutually exclusive with EF-1α. The two proteins can occur in closely related organisms, even within a genus, as in *Pythium*, or EFL can be found throughout a major group of diverse eukaryotes, as in the dinoflagellates (Figure 1.2). EFL was first reported from a handful of lineages including haptophytes and dinoflagellates, certain green algae, zygomycete fungi, a choanoflagellate, and a chlorarachniophyte (Keeling and Inagaki 2004). The extreme evolutionary distances between these lineages coupled with the mutually exclusive distribution made it unlikely that this complex pattern was due to differential loss from an ancient duplication. Instead, it was inferred that EFL had laterally transferred and functionally replaced EF-1α multiple times independently, an interpretation that ranked EFL among the most transfer-prone genes known (Keeling and Inagaki 2004). Follow-up studies in chromalveolate groups have supported this interpretation (Gile, Patron, and Keeling 2006; Kamikawa, Inagaki, and Sako 2008; Sakaguchi et al. 2009), though the discovery of both genes in the complete genome of the diatom *Thalassiosira pseudonana* (Kamikawa, Inagaki, and Sako 2008) and the PCR amplification of both genes from

the zygomycete fungus *Basidiobolus ranarum* (James et al. 2006) somewhat weakened the case against a long period of co-occurrence of these two proteins.

Figure 1.2. Global distribution of EF-1α and EFL.



Schematic view of the five-supergroup model of eukaryotic phylogeny adapted from Keeling et al. 2005 illustrating current knowledge of the distribution of EF-1α and EFL. Blue and red dots indicate lineages in which EF-1α and EFL are known to occur, respectively. For the diatom *Thalassiosira pseudonana* and the zygomycete fungus *Basidiobolus ranarum*, the presence of both a red and a blue dot indicates co-occurrence of the two proteins; in all other cases where both dots appear, the two proteins are encoded by different members of the indicated lineage. Absence of a coloured dot indicates lack of data from that lineage at the time of writing.

Lateral gene transfer is a major player in prokaryotic evolution

Lateral gene transfer (LGT, or HGT for horizontal gene transfer) is the transfer of genes between distinct taxa, as opposed to the vertical transfer of parental inheritance. It is known to play a significant role in prokaryotic evolution (Gogarten, Doolittle, and Lawrence 2002; Gogarten and Townsend 2005; Koonin, Makarova, and Aravind 2001; Koonin and Wolf 2008; Ochman, Lawrence, and Groisman 2000), and the mechanisms that enable transfer of genetic information, transduction, transformation, and conjugation, are well studied (Chen, Christie, and Dubnau 2005; Frost et al. 2005; Thomas and Nielsen 2005). There are many examples of prokaryotic LGT in the literature, but two particularly dramatic examples, from *Escherichia coli* and *Thermotoga maritima*, are often cited as evidence not only that LGT occurs, but that it can occur across vast evolutionary distances and contribute substantial proportions of a genome. A comparison of the genome sequences of three *E. coli* strains, one enterohaemorrhagic, one uropathogenic, and one benign lab strain, found that only 39% of the total number of genes encoded by these three organisms was shared among all three (Welch et al. 2002). The shared genes form a largely syntenic "backbone" in each genome that is punctuated, often at tRNA genes, by regions of anomalous codon usage and GC content where the unique genes are found. These details support the inference that the unique regions were recently inserted, rather than differentially lost from their recent common ancestor, which would have had to have an extraordinarily large genome. A complementary example, involving an equally remarkable frequency of LGT but between distantly related organisms sharing a habitat rather than closely related organisms in different habitats, is provided by the hyperthermophilic bacterium *T. maritima*. Using a combination of BLAST searches and phylogenetic analyses, the authors computed that 451, or 24%, of the total predicted genes are most similar to archaeal homologues, or that only have homologues among Archaea (Nelson et al. 1999). The non-random distribution of these genes across functional categories, predominantly in transporters, and the clustering of about a fifth of them into regions that are syntenic with Archaea and bear some archaeal repeat elements support the conclusion that these genes were laterally transferred. Furthermore, the inference of LGT is intuitively reasonable because these organisms share a difficult habitat; not only are they in close enough proximity to share genes, but adaptations to such an extreme habitat would seem at least as easily acquired by LGT as having evolved from scratch.

Methods of detecting LGT

Part of the strength of the *E. coli* and *T. maritima* examples is their presentation of

multiple lines of evidence to support the inference of LGT, thus mitigating the limitations of any single method of LGT detection. Many such methods have been employed, each with its own advantages and limitations (Ragan 2001a; Zhaxybayeva 2009). One phenomenon alluded to above and often interpreted as an indication of LGT is the natural tendency of certain prokaryotic genomes to develop characteristic patterns of GC and codon usage bias. Any section of the genome significantly deviating from the background compositional pattern can be considered of potentially foreign origin. Under this assumption, the amount of foreign DNA in the *E. coli* lab strain genome was estimated to be 17%, which is roughly in keeping with the estimates based on genes not shared with close relatives (Lawrence and Ochman 1997). A major drawback of detecting LGT with composition-based methods, however, is the tendency for xenologous DNA to acquire the same compositional biases as the host genome over time, a process called amelioration. As a result, these methods can only detect relatively recent transfers from donor genomes with noticeably different composition (Lawrence and Ochman 1997). Another common method for detecting LGT is the identification of sequences that are most similar to homologues in unrelated organisms via BLAST, as in the *T. maritima* example. One major weakness of this method is the potential for genes in the group of interest to be missing from public databases, because of either gene loss or insufficient sampling, as occurred during annotation of the human genome, resulting in a greatly inflated estimate of LGT from bacteria (Salzberg et al. 2001; Stanhope et al. 2001). Another problem, one that also contributed to erroneous identification of LGT in the human genome, is that the closest BLAST hit is not necessarily the nearest neighbour phylogenetically (Koski and Golding 2001). This weakness can be partially overcome by employing a modified BLAST search in which the pattern of hits from a given gene is compared to the average pattern of hits of all genes in the genome (Clarke et al. 2002).

The preferred method for detecting LGT is phylogenetic analysis (Doolittle et al. 2003; Ragan 2001b), which can not only provide an indication of support for the inferred relationships, but can identify the source lineage of a laterally acquired gene. Even this method has its drawbacks, however. Phylogenetic analyses are laborious and time consuming, making them impractical for attempts to quantify the percentage of foreign genes in a large genome, they are sensitive to taxon sampling, as are BLAST-based methods, and they require a robust organismal phylogeny before any gene phylogeny can be detected as incongruous (Doolittle et al. 2003; Gogarten and Townsend 2005). In the best possible phylogenetic scenario for inferring LGT, a gene sequence from the organism of interest is nested in a highly supported clade of foreign

sequences and away from a supported clade of its known relatives' sequences, thus clearly demonstrating both the occurrence of LGT and the direction of transfer. However, this best case is rarely achieved, and even this type of incongruous phylogeny could be explained by differential gene loss (Andersson 2005; Gogarten and Townsend 2005; Rogers et al. 2007). Thus the inference depends on assumptions as to the relative likelihood of LGT versus gene loss and the number of each type of event that must be invoked to explain a given distribution (Doolittle et al. 2003). These factors, in turn, depend on the fineness of the taxonomic scale under investigation, as a gene shared by many closely related organisms is more likely to have been vertically inherited than one that is only present in a few, distantly related organisms (Rogers et al. 2007).

LGT has a more moderate impact on eukaryotic genomes

Lateral transfer also occurs in eukaryotes, though not as frequently. Several attempts to quantify the contribution of LGT to eukaryotic genomes have indicated a generally moderate impact, with one possibly inflated exception. The greatest proportion of LGT yet detected in a eukaryotic nuclear data set was 20% of plastid targeted genes in the chlorarachniophyte *Bigelowiella natans* (Archibald et al. 2003). Eight out of the 13 total putative xenologues grouped with red algae or secondary algae with red algal plastids in phylogenetic analysis, and one bacterial gene was subsequently shown to be shared by chromists (Rogers et al. 2007). If this phylogenetic signal is considered to reflect common ancestry with chromalveolates, a supergroup united by an ancient endosymbiosis in which their ancestor acquired a red alga as a plastid, as recent phylogenomic analyses have suggested (Burki et al. 2007; Hackett et al. 2007; Minge et al. 2009), the percentage of laterally transferred genes drops to 6%. Furthermore, if the proportion of LGT is computed from the total number of genes in the dataset (78), rather than the subset of genes with resolved phylogenies (62), the proportion would be 17%, or as low as 5% if both possibilities are combined. Nevertheless, 5% of genes derived from lateral transfer is among the highest proportions yet determined for eukaryotes, and a higher rate of LGT among genes for plastid targeted proteins is intuitively reasonable if acquisition of targeting information for endosymbiotically transferred genes is considered difficult. Another protist inferred to carry a large proportion of xenologues is the diplomonad *Spironucleus salmonicida*. An EST and genome sequence survey of this fish parasite, not restricted to any particular class of genes, detected 84 laterally transferred genes, or 6% of the total predicted genes (Andersson et al. 2007). On a similar level, an EST survey of four rumen-dwelling ciliate taxa found 148 out of

3563 genes, or 4%, to be laterally acquired (Ricard et al. 2006). Other surveys to date have detected LGT proportions of less than 1% of the total number of predicted genes: 24 out of 5519 total genes in *Cryptosporidium parvum*, 47 out of 9068 in *Trypanosoma brucei*, and 96 out of 9938 total genes in *Entamoeba histolytica* were inferred to derive from LGT (Berriman et al. 2005; Huang et al. 2004; Loftus et al. 2005). In order to achieve a clearer picture of the extent of LGT and its impact on eukaryote evolution, and in particular the factors that drive its relative importance among various lineages, we will need a more comprehensive sample of eukaryotic genomes. In particular, more genome surveys of heterotrophic protists, currently under-represented among the predominantly parasitic and photoautotrophic protist genomes, would be welcome in order to test the intuitively appealing idea that LGT has played a bigger role in their evolution due to greater exposure to intracellular foreign DNA (Andersson 2005; Doolittle 1998).

Among the many examples of genes with complex histories involving lateral transfer, EFL is distinguished by its close relationship with a well-studied, core housekeeping gene. Most complex evolutionary histories have been reported for metabolic rather than housekeeping genes (Andersson et al. 2006; Rogers and Keeling 2004; Rogers et al. 2007; Sanchez-Perez et al. 2008), though lateral transfers have also been reported for aa-tRNA synthetases and a DNA polymerase in eukaryotes (Andersson 2005). This may be due in part to the greater likelihood of retaining a laterally acquired gene for a useful new function, as many of the laterally acquired genes in rumen-dwelling ciliates and *E. histolytica* (enabling anaerobic metabolism), and *Phytophthora* (virulence factors) attest (Belbahri et al. 2008; Loftus et al. 2005). However, examples of orthologous replacement, in which a gene already encoded in the recipient genome is lost while the newly acquired xenologue is kept, are also known. EFL and EF-1α can be considered part of this latter category, being capable of the same function, along with the recently discovered example of distinct paralogues of α-tubulin (Simpson, Perley, and Lara 2008). Although there are many examples in addition to these, overall the frequency of LGT in eukaryotes does not appear great enough to warrant concern that the true phylogeny may be blurred beyond recovery. In some cases LGT can even provide useful synapomorphies and therefore aid phylogenetic reconstruction (Harper and Keeling 2003; Huang and Gogarten 2006; Patron, Rogers, and Keeling 2004).

With the exception of endosymbiotic gene transfer

Genes transferred from endosymbiotic organelles represent a special case of LGT in eukaryotes that has certainly had a major impact on eukaryotic evolution. It is well known that endosymbiosis enabled photosynthesis in several lineages through multiple independent endosymbioses. The retention of an internalized foreign cell exposes the host nucleus to continual invasion by foreign DNA from the endosymbiont. Transfer and integration of organellar DNA into the nucleus is an ongoing process, and its impact can be clearly observed in eukaryotic genomes. For example, domestic cats and the model plant *Arabidopsis thaliana* have complete copies of their mitochondrial DNA integrated into their nuclear genomes, rice has a complete copy of its chloroplast genome on chromosome 10, and smaller stretches of organellar DNA have been detected in the nuclei of many other plants, animals, fungi and protists (Timmis et al. 2004). These recent transfers, in which sequence similarity between the nuclear and organellar copies can exceed 95%, seem to indicate that DNA transfer is a common occurrence. The rate of transfer from mitochondria to the nucleus in yeast has been experimentally estimated to be one in 20,000 cells per generation (Thorsness and Fox 1990), and from chloroplasts to the nucleus in tobacco leaves one in every 5,000,000 cells (Stegemann et al. 2003). Such estimates are unsurprisingly quite variable, with tobacco pollen showing a rate that is two orders of magnitude higher than the leaf cells, and yeast mitochondrial transfers varying several fold by strain and by temperature, but they nonetheless place endosymbiotic transfer rates on the order of mutation frequency (Timmis et al. 2004). Most such transferred DNA is eventually lost, but in a few cases genes may become functional and be retained in the nuclear genome. This can happen either because the transferred gene took over the function of a related host gene, which would then typically be lost, or because the transferred gene acquired targeting information so that its product can be targeted to the organelle.

Research Objectives

The overall goal of this research is to better understand how the evolutionary processes of endosymbiotic gene transfer, lateral gene transfer, and differential gene loss with EF-1α have contributed to the molecular evolution of EFL. To this end, the complex distribution of EFL and EF-1α was investigated in four groups of phylogenetically distant eukaryotes at both fine and broad taxonomic scales. Endosymbiotic gene transfer was inferred to have occurred in the chlorarachniophytes, the smallest taxonomic group investigated. Lateral gene transfer was inferred in the largest group investigated, the chromalveolates, but more recent work in the green

algae and particularly in the kinetoplastids reveals evidence for differential loss from an ancestral state of co-occurrence with EF-1α. This unexpected finding adds another dimension to our understanding of the distribution of these proteins, and serves as a cautionary reminder that inferences of lateral gene transfer from incongruous gene phylogenies can be greatly influenced by taxon sampling and uncertainty regarding the underlying organismal phylogeny.

Literature Cited

Andersen GR, Nissen P, Nyborg J. 2003. Elongation factors in protein biosynthesis. Trends Biochem. Sci. 28:434-441.

Andersen GR, Pedersen L, Valente L, Chatterjee I, Kinzy TG, Kjeldgaard M, Nyborg J. 2000. Structural basis for nucleotide exchange and competition with tRNA in the yeast elongation factor complex eEF1A:EEF1Balpha. Mol. Cell 6:1261-1266.

Andersson JO. 2005. Lateral gene transfer in eukaryotes. Cell Mol. Life Sci. 62:1182-1197.

Andersson JO, Hirt RP, Foster PG, Roger AJ. 2006. Evolution of four gene families with patchy phylogenetic distributions: Influx of genes into protist genomes. BMC Evol. Biol. 6:27.

Andersson JO, Sjögren AM, Horner DS, Murphy CA, Dyal PL, Svärd SG, Logsdon JM Jr, Ragan MA, Hirt RP, Roger AJ. 2007. A genomic survey of the fish parasite *Spironucleus salmonicida* indicates genomic plasticity among diplomonads and significant lateral gene transfer in eukaryote genome evolution. BMC Genomics 8:51.

Archibald JM, Rogers MB, Toop M, Ishida K, Keeling PJ. 2003. Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigelowiella natans*. Proc. Natl. Acad. Sci. USA 100:7678-7683.

Baldauf SL, Doolittle WF. 1997. Origin and evolution of the slime molds (Mycetozoa). Proc. Natl. Acad. Sci. USA 94:12007-12012.

Baldauf SL, Palmer JD. 1993. Animals and fungi are each other's closest relatives: Congruent evidence from multiple proteins. Proc. Natl. Acad. Sci. USA 90:11558-11562.

Baldauf SL, Palmer JD, Doolittle WF. 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc. Natl. Acad. Sci. USA 93:7749-7754.

Belbahri L, Calmin G, Mauch F, Andersson JO. 2008. Evolution of the cutinase gene family: Evidence for lateral gene transfer of a candidate *Phytophthora* virulence factor. Gene 408:1-8.

Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B et al. (102 co-authors). 2005. The genome of the African trypanosome *Trypanosoma brucei*. Science 309:416-422.

Burki F, Shalchian-Tabrizi K, Minge M, Skjæveland A, Nikolaev SI, Jakobsen KS, Pawlowski J. 2007. Phylogenomics reshuffles the eukaryotic supergroups. PLoS ONE 2:e790.

Chen I, Christie PJ, Dubnau D. 2005. The ins and outs of DNA transfer in bacteria. Science 310:1456-1460.

Chuang SM, Chen L, Lambertson D, Anand M, Kinzy TG, Madura K. 2005. Proteasome-mediated degradation of cotranslationally damaged proteins involves translation elongation factor 1A. Mol. Cell. Biol. 25:403-413.

Clarke GD, Beiko RG, Ragan MA, Charlebois RL. 2002. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. J. Bacteriol. 184:2072-2080.

Doolittle WF. 1998. You are what you eat: A gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet. 14:307-311.

Doolittle WF, Boucher Y, Nesbø CL, Douady CJ, Andersson JO, Roger AJ. 2003. How big is the iceberg of which organellar genes in nuclear genomes are but the tip? Philos. Trans. R. Soc. Lond. B. Biol. Sci. 358:39-57.

Edmonds BT, Bell A, Wyckoff J, Condeelis J, Leyh TS. 1998. The effect of F-actin on the binding and hydrolysis of guanine nucleotide by *Dictyostelium* elongation factor 1A. J. Biol. Chem. 273:10288-10295.

Frost LS, Leplae R, Summers AO, Toussaint A. 2005. Mobile genetic elements: The agents of open source evolution. Nat. Rev. Microbiol. 3:722-732.

Gaucher EA, Miyamoto MM, Benner SA. 2001. Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. Proc. Natl. Acad. Sci. USA 98:548-552.

Gaucher EA, Das UK, Miyamoto MM, Benner SA. 2002. The crystal structure of eEF1A refines the functional predictions of an evolutionary analysis of rate changes among elongation factors. Mol. Biol. Evol. 19:569-573.

Gile GH, Patron NJ, Keeling PJ. 2006. EFL GTPase in cryptomonads and the distribution of EFL and EF-1alpha in chromalveolates. Protist 157:435-444.

Gogarten JP, Townsend JP. 2005. Horizontal gene transfer, genome innovation and evolution. Nat. Rev. Microbiol. 3:679-687.

Gogarten JP, Doolittle WF, Lawrence JG. 2002. Prokaryotic evolution in light of gene transfer. Mol. Biol. Evol. 19:2226-2238.

Gonen H, Smith CE, Siegel NR, Kahana C, Merrick WC, Chakraburtty K, Schwartz AL, Ciechanover A. 1994. Protein synthesis elongation factor EF-1 alpha is essential for ubiquitin-dependent degradation of certain N alpha-acetylated proteins and may be substituted for by the bacterial elongation factor EF-Tu. Proc. Natl. Acad. Sci. USA 91:7648-7652.

Gross SR, Kinzy TG. 2005. Translation elongation factor 1A is essential for regulation of the actin cytoskeleton and cell morphology. Nat. Struct. Mol. Biol. 12:772-778.

Hackett JD, Yoon HS, Li S, Reyes-Prieto A, Rummele SE, Bhattacharya D. 2007. Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of rhizaria with chromalveolates. Mol. Biol. Evol. 24:1702-1713.

Harper JT, Keeling PJ. 2003. Nucleus-encoded, plastid-targeted glyceraldehyde-3-phosphate dehydrogenase (GAPDH) indicates a single origin for chromalveolate plastids. Mol. Biol. Evol. 20:1730-1735.

Hashimoto T, Hasegawa M. 1996. Origin and early evolution of eukaryotes inferred from the amino acid sequences of translation elongation factors 1alpha/Tu and 2/G. Adv. Biophys. 32:73-120.

Huang J, Gogarten JP. 2006. Ancient horizontal gene transfer can benefit phylogenetic reconstruction. Trends Genet. 22:361-366.

Huang J, Mullapudi N, Lancto CA, Scott M, Abrahamsen MS, Kissinger JC. 2004. Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*. Genome Biol. 5:R88.

Inagaki Y, Blouin C, Susko E, Roger AJ. 2003. Assessing functional divergence in EF-1alpha and its paralogs in eukaryotes and archaebacteria. Nucleic Acids Res. 31:4227-4237.

Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA 86:9355-9359.

James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J et al. (72 co-authors). 2006. Reconstructing the early evolution of fungi using a six-gene phylogeny. Nature 443:818-822.

Kamikawa R, Inagaki Y, Sako Y. 2008. Direct phylogenetic evidence for lateral transfer of elongation factor-like gene. Proc. Natl. Acad. Sci. USA 105:6965-6969.

Keeling PJ, Inagaki Y. 2004. A class of eukaryotic GTPase with a punctate distribution suggesting multiple functional replacements of translation elongation factor 1alpha. Proc. Natl. Acad. Sci. USA 101:15380-15385.

Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, Roger AJ, Gray MW. 2005. The tree of eukaryotes. Trends Ecol. Evol. 20:670-676.

Koonin EV, Wolf YI. 2008. Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 36:6688-6719.

Koonin EV, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes: Quantification and classification. Annu. Rev. Microbiol. 55:709-742.

Koski LB, Golding GB. 2001. The closest BLAST hit is often not the nearest neighbor. J. Mol. Evol. 52:540-542.

Lawrence JG, Ochman H. 1997. Amelioration of bacterial genomes: Rates of change and exchange. J. Mol. Evol. 44:383-397.

Liu G, Grant WM, Persky D, Latham VM,Jr, Singer RH, Condeelis J. 2002. Interactions of

elongation factor 1alpha with F-actin and beta-actin mRNA: Implications for anchoring mRNA in cell protrusions. Mol. Biol. Cell 13:579-592.

Loftus B, Anderson I, Davies R, Alsmark UC, Samuelson J, Amedeo P, Roncaglia P, Berriman M, Hirt RP, Mann BJ et al. (54 co-authors). 2005. The genome of the protist parasite *Entamoeba histolytica*. Nature 433:865-868.

Lopez P, Forterre P, Philippe H. 1999. The root of the tree of life in the light of the covarion model. J. Mol. Evol. 49:496-508.

Mickleburgh I, Chabanon H, Nury D, Fan K, Burtle B, Chrzanowska-Lightowlers Z, Hesketh J. 2006. Elongation factor 1alpha binds to the region of the metallothionein-1 mRNA implicated in perinuclear localization--importance of an internal stem-loop. RNA 12:1397-1407.

Minge MA, Silberman JD, Orr RJ, Cavalier-Smith T, Shalchian-Tabrizi K, Burki F, Skjæveland A, Jakobsen KS. 2009. Evolutionary position of breviate amoebae and the primary eukaryote divergence. Proc. Biol. Sci. 276:597-604.

Nakazawa M, Moreira D, Laurent J, Le Guyader H, Fukami Y, Ito K. 1999. Biochemical analysis of the interaction between elongation factor 1alpha and alpha/beta-tubulins from a ciliate, *Tetrahymena pyriformis*. FEBS Lett. 453:29-34.

Negrutskii BS, El'skaya AV. 1998. Eukaryotic translation elongation factor 1 alpha: Structure, expression, functions, and possible role in aminoacyl-tRNA channeling. Prog. Nucleic Acid Res. Mol. Biol. 60:47-78.

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA et al. (29 co-authors). 1999. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. Nature 399:323-329.

Nilsson J, Nissen P. 2005. Elongation factors on the ribosome. Curr. Opin. Struct. Biol. 15:349-354.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299-304.

Patron NJ, Rogers MB, Keeling PJ. 2004. Gene replacement of fructose-1,6-bisphosphate aldolase supports the hypothesis of a single photosynthetic ancestor of chromalveolates. Eukaryot. Cell. 3:1169-1175.

Ragan MA. 2001a. Detection of lateral gene transfer among microbial genomes. Curr. Opin. Genet. Dev. 11:620-626.

Ragan MA. 2001b. On surrogate methods for detecting lateral gene transfer. FEMS Microbiol. Lett. 201:187-191.

Ricard G, McEwan NR, Dutilh BE, Jouany JP, Macheboeuf D, Mitsumori M, McIntosh FM,

Michalowski T, Nagamine T, Nelson N et al. (17 co-authors). 2006. Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. BMC Genomics 7:22.

Roger AJ, Sandblom O, Doolittle WF, Philippe H. 1999. An evaluation of elongation factor 1 alpha as a phylogenetic marker for eukaryotes. Mol. Biol. Evol. 16:218-233.

Rogers MB, Keeling PJ. 2004. Lateral transfer and recompartmentalization of Calvin cycle enzymes of plants and algae. J. Mol. Evol. 58:367-375.

Rogers MB, Watkins RF, Harper JT, Durnford DG, Gray MW, Keeling PJ. 2007. A complex and punctate distribution of three eukaryotic genes derived by lateral gene transfer. BMC Evol. Biol. 7:89.

Sakaguchi M, Takishita K, Matsumoto T, Hashimoto T, Inagaki Y. 2009. Tracing back EFL gene evolution in the cryptomonads-haptophytes assemblage: Separate origins of EFL genes in haptophytes, photosynthetic cryptomonads, and goniomonads. Gene, in press.

Salzberg SL, White O, Peterson J, Eisen JA. 2001. Microbial genes in the human genome: Lateral transfer or gene loss? Science 292:1903-1906.

Sanchez-Perez GF, Hampl V, Simpson AG, Roger AJ. 2008. A new divergent type of eukaryotic methionine adenosyltransferase is present in multiple distantly related secondary algal lineages. J. Eukaryot. Microbiol. 55:374-381.

Simpson AG, Perley TA, Lara E. 2008. Lateral transfer of the gene for a widely used marker, alpha-tubulin, indicated by a multi-protein study of the phylogenetic position of *Andalucia* (Excavata). Mol. Phylogenet. Evol. 47:366-377.

Slobin LI. 1980. The role of eucaryotic factor tu in protein synthesis: The measurement of the elongation factor tu content of rabbit reticulocytes and other mammalian cells by a sensitive radioimmunoassay. Eur. J. Biochem. 110:555-563.

Stanhope MJ, Lupas A, Italia MJ, Koretke KK, Volker C, Brown JR. 2001. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. Nature 411:940-944.

Stegemann S, Hartmann S, Ruf S, Bock R. 2003. High-frequency gene transfer from the chloroplast genome to the nucleus. Proc. Natl. Acad. Sci. USA 100:8828-8833.

Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat. Rev. Microbiol. 3:711-721.

Thorsness PE, Fox TD. 1990. Escape of DNA from mitochondria to the nucleus in *Saccharomyces cerevisiae*. Nature 346:376-379.

Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet. 5:123-135.

Welch RA, Burland V, Plunkett G 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J et al. (19 co-authors). 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc. Natl. Acad. Sci. USA 99:17020-17024.

Yang F, Demma M, Warren V, Dharmawardhane S, Condeelis J. 1990. Identification of an actin-binding protein from *Dictyostelium* as elongation factor 1a. Nature 347:494-496.

Zhaxybayeva O. 2009. Detection and quantitative assessment of horizontal gene transfer. Methods Mol. Biol. 532:195-213.

CHAPTER 2: EFL GTPASE IN CRYPTOMONADS AND THE DISTRIBUTION OF EFL AND EF-1A IN CHROMALVEOLATES[1]

Introduction

Translation elongation factor-1α (EF-1α, called EF-Tu in bacteria) plays an integral role in cellular information flow by bringing charged tRNAs to the ribosome during peptide elongation. It is a highly conserved protein found across the three domains of life. Because of its core role in translation and many interactions with other proteins, it is considered essential and unlikely to be moved from genome to genome by lateral gene transfer, and it has been used in many analyses of phylogeny and molecular evolution (Baldauf and Palmer 1993; Baldauf and Doolittle 1997; Gaucher, Miyamoto, and Benner 2001; Inagaki et al. 2004). However, a recent investigation found that several eukaryotic genomes lack any evidence of an EF-1α gene, and instead encode a distantly related paralogue called EF-like, or EFL (Keeling and Inagaki 2004). EFL has been found in only a few lineages scattered across the tree of eukaryotes, and nearly all of these have close relatives that encode EF-1α but apparently not EFL. These relationships suggest that EFL has spread by eukaryote-to-eukaryote lateral gene transfer, functionally replacing EF-1α several times independently despite its crucial role in translation (Keeling and Inagaki 2004).

One lineage where EFL has been found is the hypothetical 'supergroup', chromalveolates. The chromalveolate hypothesis states that the chromists (cryptomonads, haptophytes, and heterokonts) and alveolates (ciliates, dinoflagellates and apicomplexans) share a common ancestor and that this ancestor acquired a secondary red algal plastid (Cavalier-Smith 1999). Although no single gene examined to date unites all chromalveolates at once, several host and endosymbiont-derived genes support this hypothesis (Fast et al. 2001; Yoon et al. 2002; Harper and Keeling 2003; Patron, Rogers, and Keeling 2004; Harper, Waanders, and Keeling 2005). Within the chromalveolates, two major lineages were found to contain EFL (dinoflagellates and haptophytes), while the other four were found to contain EF-1α (apicomplexans, ciliates, heterokonts and cryptomonads) (Keeling and Inagaki 2004). In the cases of dinoflagellates and haptophytes, this was based on multiple expressed sequence tag (EST) sequencing projects, from which no EF-1α was evident. Similarly, whole genomes and EST projects from apicomplexans, ciliates, and heterokonts bore no evidence of EFL. Only

[1] A version of this chapter has been published. Gile GH, Patron NJ, Keeling PJ. 2006. EFL GTPase in cryptomonads and the distribution of EFL and EF-1alpha in chromalveolates. Protist. 157:435-444.

cryptomonads lack data from genome-wide surveys, and here the evidence for EF-1α came from a single gene amplified by PCR (Harper, Waanders, and Keeling 2005).

We have used EST sequence data to clarify our understanding of EFL's distribution in chromalveolates. Previous sampling from haptophytes coarsely represents the entire range of known diversity (because it includes the earliest known lineage, the genus *Pavlova*). The distribution of EFL in dinoflagellates was less clear because it included no early-branching lineages, though later-branching dinoflagellates are known to encode EFL. We accordingly sought EFL and EF-1α in two of the most ancient groups in the dinoflagellate lineage, the parasite *Perkinsus marinus* and the predator *Oxyrrhis marina* (Goggin and Barker 1993; Reece et al. 1997; Saldarriaga et al. 2003; Leander and Keeling 2004). In addition, we have sampled a later-branching dinoflagellate that has a haptophyte endosymbiont, *Karlodinium micrum* (Tengs et al. 2000; Patron, Waller, and Keeling 2006), to see which EFL was retained from a partnership that involved two EFL-containing organisms. Most importantly, we used EST sequences from two cryptomonads, *Guillardia theta* and *Rhodomonas salina*, to reassess the presence and absence of EFL and EF-1α in this group. In both taxa, EFL was found but EF-1α was not present in our sampling. This refines our understanding of several aspects of the distribution of EFL in chromalveolates: in groups where EFL is found, it appears to be common to all members of that group; of the six major lineages of chromalveolates, half have EFL and half have EF-1α; and within this supergroup the lineages with EFL are not related to one another to the exclusion of those lineages with EF-1α. Phylogenetic analyses suggest that the ancestor of all chromalveolates had EF-1α, but the phylogeny of EFL is not consistent with a common origin of EFL in chromalveolates. At face value this suggests multiple origins of EFL within the supergroup.

Materials and Methods

Identification and characterization of cryptomonad and dinoflagellate EFL genes

Homologues of EFL were identified in expressed sequence tag (EST) projects from two cryptomonads, *Guillardia theta* (CCMP 327) and *Rhodomonas salina* (CCMP 1319), the dinoflagellate *Karlodinium micrum* (CCMP 415) and the non-photosynthetic sister to dinoflagellates, *Oxyrrhis marina* (CCMP 1788). Databases containing these EST sequencing projects (http://www.bch.umontreal.ca/pepdb/pep.html) were searched using tBLASTn for homologues of both EF-1α and EFL. In some cases multiple copies of the gene were found, but

in all such cases they were identical or nearly identical at the amino acid level, so one full-length EST was chosen to represent them and the clone was completely sequenced on both strands. *Perkinsus marinus* sequences were identified using tBLASTn searches from the genome-sequencing project (http://www.tigr.org/tdb/e2k1/pmg/). Three EFL sequences, named 1099751674524, 1099751674136, and 1099751675083 at the time of writing, were conceptually translated and added to the alignment for phylogenetic analyses. For *G. theta* and *R. salina,* EFL was also amplified from total RNA using a degenerate EFL primer designed for the 5' end of the gene (CTGTCGATCGTCATHTGYGGNCAYGTNGA) and a species-specific primer (CTTCTTAGCACCACCATCATCGCGAGCAAC for *G. theta* and CGCTTGTGGTGCATCTCCACGGTGAAGATC for *R. salina*) for the 3' end, using the Superscript III RT-PCR kit (Invitrogen). Products of the expected size were cloned using TOPO-TA cloning (Invitrogen) and several clones were sequenced on both strands. *Karlodinium micrum* ESTs were described in Patron et al. (2006), and *G. theta*, *R. salina*, and *O. marina* EST projects are currently ongoing. All new EFL sequences were deposited in GenBank under accession numbers DQ659242-DQ659245 and DQ666284.

Phylogenetic analyses

New sequences were added to an existing amino acid alignment of EFL and EF-1α (Keeling and Inagaki 2004) and phylogenetic trees were inferred using maximum likelihood (ML), distance, and Bayesian methods. Trees were inferred using both genes, which confirmed the new sequences were EFL (not shown). All other analyses were restricted to full-length or near full-length EFL sequences alone, from which 438 unambiguously aligned positions from 19 taxa were analyzed. ML trees were inferred using PhyML 2.4.4 (Guindon and Gascuel 2003) with input trees generated by BIONJ, the JTT model of amino acids substitution, the proportion of variable rates estimated from the data, and eight variable categories of substitution rates with a proportion of invariable sites also estimated. One thousand bootstrap trees were inferred with PhyML using the same parameters from the original tree. For distance analyses, gamma corrected distances were calculated by TREE-PUZZLE 5.2 (Strimmer and von Haeseler 1996) using the WAG substitution matrix with eight variable rate categories and invariable sites. Trees were inferred by weighted neighbour-joining using WEIGHBOR 1.0.1a (Bruno, Socci, and Halpern 2000). One thousand bootstrap re-sampling replicates were performed in batches of 250 using PUZZLEBOOT (shell script by A. Roger and M. Holder, http:/www.tree-puzzle.de) with rates and frequencies estimated using TREE-PUZZLE 5.2. MRBAYES 3.0 (Ronquist and

Huelsenbeck 2003) was used to perform Bayesian analysis using the JTT substitution model with rates assigned by four equally probable categories approximating a gamma distribution. One cold and three heated MCMC chains were run for one million generations, sampling one tree every thousand generations. After 4,000 generations, log likelihood values stabilized, and subsequent trees were used to compute the 50% majority-rule consensus tree which depicts estimated posterior probabilities for each clade (data not shown).

Approximately Unbiased (AU) tests (Shimodaira and Hasegawa 2001) were carried out to examine alternate positions of cryptomonads and *P. marinus* and the monophyly of chromalveolates. For cryptomonads, ML trees excluding *G. theta* and *R. salina* were optimized as described above, which gave the same topology of the remaining taxa as found in the trees with cryptomonads included. Cryptomonads were added to this optimized tree as sister to all major groups and at all other inter-group nodes, resulting in the ML tree and 10 alternatives. Site-likelihoods for these trees and 100 bootstrap trees were calculated by TREE-PUZZLE 5.1 using the –wsl option with the parameters used for the ML tree, and AU tests were performed using CONSEL 1.19 (Shimodaira 2002). The position of *P. marinus* and the monophyly of chromalveolates were tested using the same procedure. The entire analysis was repeated with the highly divergent sequence of *Bigelowiella natans* removed from the alignment.

Results and Discussion

An expressed gene for EFL in cryptomonads

Members of the translation factor GTPase family were sought from ongoing cryptomonad EST projects using known EFL and EF-1α sequences to search 14,080 *G. theta* sequences comprising 6,267 clusters and 2,848 *R. salina* sequences comprising 1,773 clusters. In both cases sequences corresponding to EFL were found, but EF-1α was not found in our sampling from either species. The *R. salina* EFL was represented by three non-overlapping clusters of ESTs: one with nine ESTs spanning the 3' end of the gene, one single EST at the 5' end, and one single EST in the middle of the gene. A single, truncated EST spanning the 3' end of the gene represented EFL from *G. theta*. The level of representation seen in *R. salina* is characteristic of EFL from other EST samples (Keeling and Inagaki 2004), but the single EST from 14,080 sequences in *G. theta* is unusual. Representation does not necessarily relate to expression levels, so we suspect the single sequence is most likely an indication of under-representation in the library. However, the sequence did not contain sites for the restriction enzyme used in library construction (*Not*I), so there is no obvious reason for its under-

representation. In neither case did the EST clusters cover the entire gene sequence, so a large fragment of the gene was amplified by RT-PCR, and 14 and 12 individual clones were sequenced from *R. salina* and *G. theta* respectively. In *G. theta,* only six synonymous variations were found among all sequences, and the RT-PCR fragments correspond exactly to the EST fragment in the region of overlap. In *R. salina,* two slightly different copies of the gene were found several times each in both RT-PCR and EST sequences (one copy was found in all three EST clusters and the second only in the 3' cluster). The sequences varied only at synonymous positions, but they shared no 3' UTR sequence similarity, confirming they are different loci. One full-length *R. salina* EFL had a 13 bp 5' UTR and 38 bp 3' UTR, while the second copy lacked sequence for the extreme 5' end and had a 53 bp 3' UTR. The *G. theta* sequence is slightly truncated at the 5' end (approximately the first eight codons are missing) and there is a 45 bp 3' UTR. We compared these sequences to an independent collection of *G. theta* ESTs recently released to public databases (Gould et al. 2006) and found two short, identical fragments (see accession AM183813).

Rhodomonas salina is the only cryptomonad previously reported to contain EF-1α (Harper, Waanders, and Keeling 2005). We specifically searched for this sequence in our databases of *R. salina* and *G. theta* ESTs, and found no evidence for its presence in either collection. The identity of this sequence is therefore open to question. Cryptomonads retain the genome of their red algal endosymbiont (Douglas et al. 2001), so it is possible this gene is derived from that endosymbiont, but since this gene was not demonstrably related to red algal EF-1α (Harper, Waanders, and Keeling 2005), this seems unlikely. If *R. salina* contains both cytosolic EFL and cytosolic EF-1α, it is of great importance, since the genes are nearly always mutually exclusive (one possible case in the fungus *Basidiobolus ranarum* has been reported but not yet confirmed: DQ282610 and DQ275340). However, the acquisition of *R. salina* EF-1α by RT-PCR has not been repeated, and it was part of a large survey that attempted to sequence EF-1α from many cryptomonads and failed. PCR from genomic DNA also failed to recover this gene (Harper, Waanders, and Keeling 2005). Altogether, we believe this sequence is most likely an artifact of contamination, but if further evidence confirms it does exist in *R. salina*, its origin is of great interest.

EFL from Karlodinium, Oxyrrhis, *and* Perkinsus

EFL has previously been reported from only a few dinoflagellates, namely a full-length mRNA from *Heterocapsa triquetra* and fragments from *Amphidinium carterae* and

Lingulodinium polyedrum (Bachvaroff et al. 2004; Hackett et al. 2004; Keeling and Inagaki 2004). To determine whether EFL originated within the lineage or predated the dinoflagellate radiation, we identified EFL in two deep-branching lineages that are sister-groups to true dinoflagellates, *Oxyrrhis marina* and *Perkinsus marinus*. Both are non-photosynthetic (*O. marina* is a predator and *P. marinus* is a parasite, Azevedo 1989; Droop 1953), and molecular phylogenetic data show that both diverged early in dinoflagellate evolution, with *P. marinus* branching prior to *O. marina* (Goggin and Barker 1993; Reece et al. 1997; Saldarriaga et al. 2003; Leander and Keeling 2004). The genome-sequencing project of *P. marinus* (http://www.tigr.org/tdb/e2k1/pmg/) and EST data from *O. marina* (40 individual ESTs from a total of 18,012) both contained multiple copies of EFL but no copy of EF-1α. This suggests that EFL originated before the radiation of extant dinoflagellate lineages.

In contrast to *P. marinus* and *O. marina*, *Karlodinium micrum* diverged relatively recently within dinoflagellates, but it is of interest because it, and the closely related genus *Karenia*, has lost its original dinoflagellate plastid and replaced it with an endosymbiotic haptophyte (Tengs et al. 2000). It is therefore a symbiotic partnership between two organisms, both of which are expected to encode EFL. The nucleus of the haptophyte has since been lost, but the *K. micrum* nuclear genome contains many genes for plastid-targeted proteins derived from this genome (Ishida and Green 2002; Yoon et al. 2005; Patron, Waller, and Keeling 2006). Several distinct EFL genes were identified from 47 ESTs from *K. micrum*, but all ESTs were extremely similar to one another and to homologues from other dinoflagellates (supported by phylogenetic evidence – see below). We found no evidence of an EFL gene of haptophyte origin in this genome.

Phylogenetic inference for EFL

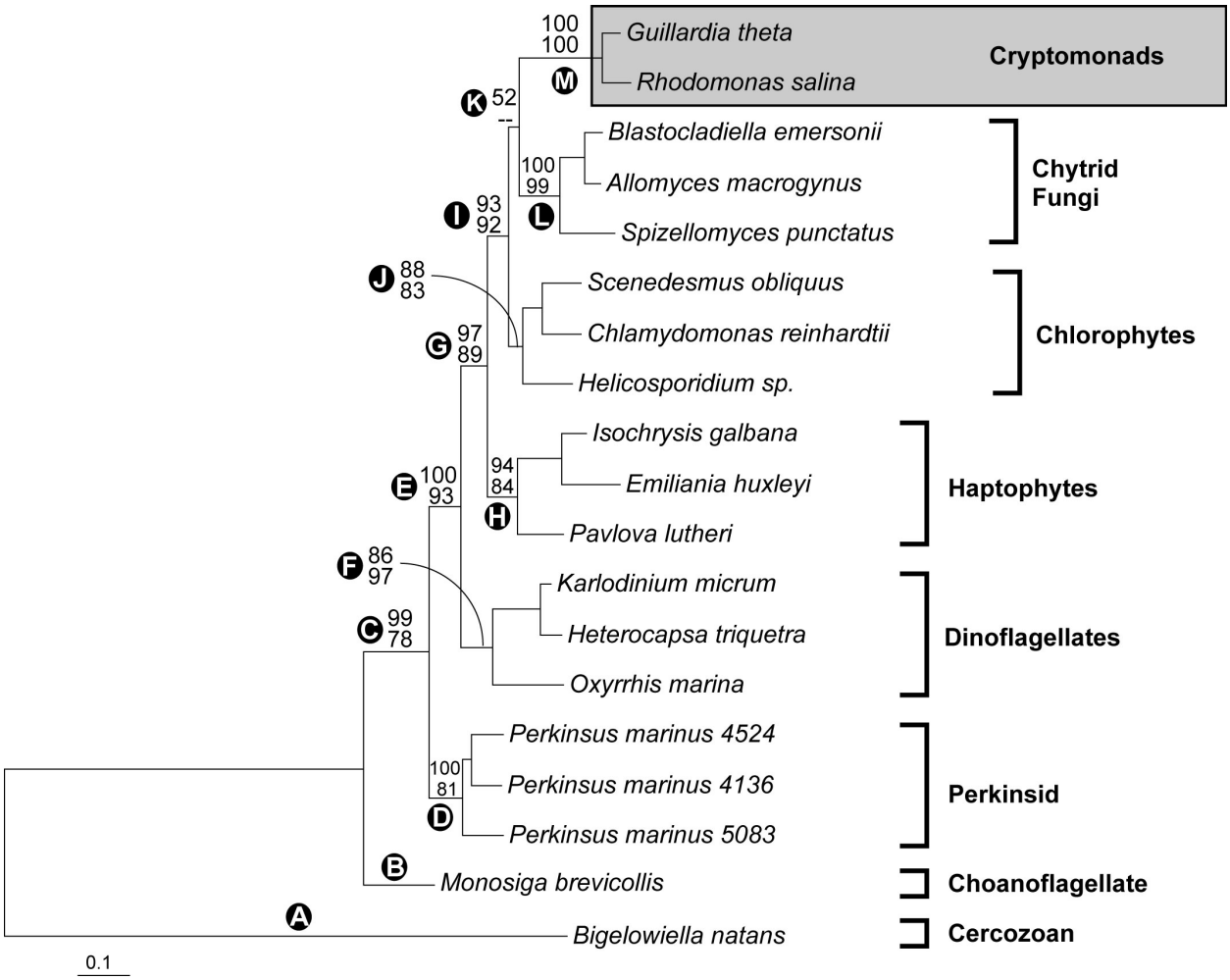
Phylogenetic trees of all full-length EFL sequences were inferred using a variety of methods, all of which yielded a topology similar to that shown in Figure 2.1. This unrooted maximum likelihood (ML) tree considers *Bigelowiella natans* as the outgroup because this sequence is always the earliest branch of EFL when analyzed with related GTPases (Keeling and Inagaki 2004). The *B. natans* sequence is highly divergent, however, and the root of the EFL tree must be taken with extreme caution, so all analyses were repeated excluding this sequence, which had no major effect on the tree or support levels (not shown). Most nodes are relatively strongly supported by ML and distance bootstrap methods, as well as by Bayesian posterior probabilities (which were all close to 1.0 except the node uniting chytrid fungi and cryptomonads which was 0.725, not shown). Most irrefutably supported lineages are recovered (i.e., green

algae, haptophytes, dinoflagellates, and cryptomonads), with the exception of the clade uniting dinoflagellates and *Perkinsus* (see below). *Oxyrrhis marina* is the sister group of the dinoflagellates *H. triquetra* and *K. micrum*, as expected given its position in phylogenies inferred from other proteins (Saldarriaga et al. 2003; Leander and Keeling 2004). The *K. micrum* EFL branches with dinoflagellates and not haptophytes, confirming its host origin.

Evolution of EFL in chromalveolates

Two connected features of EFL evolution specifically relating to chromalveolates stand out as unusual. First, why do so many chromalveolates contain EFL rather than EF-1α, and conversely, why are so many of the EFL-containing lineages chromalveolates? Second, why do the chromalveolate EFL genes not form a single clade? EFL is very rare in eukaryotes: it has been described in only seven lineages to date, and now nearly half of these are chromalveolates. On the other side of the same coin, half of the major lineages of chromalveolates contain EFL, meaning it is more abundant in this supergroup than in any other (Figure 2.2). Unlike many other protist groups, there is relatively deep sampling of molecular data from a broad diversity of chromalveolates. Noting that EFL has almost exclusively been found through genome-wide analyses (genome sequences or ESTs), it is possible that the high frequency of EFL in chromalveolates is simply due to the fact that this level of sampling is not widely available in protists. This suggests that improved sampling of protists as a whole may reveal many more lineages with EFL. Alternatively, EFL-containing chromalveolates may simply be more common than other eukaryotes, raising the question of whether the EFL-containing chromalveolate lineages acquired EFL several times independently, or whether it was present in their common ancestor. To distinguish between these possibilities we need to consider the known evolutionary relationships among the chromalveolate lineages and infer robust phylogenies of EFL and EF-1α.

Figure 2.1. Phylogeny of EFL



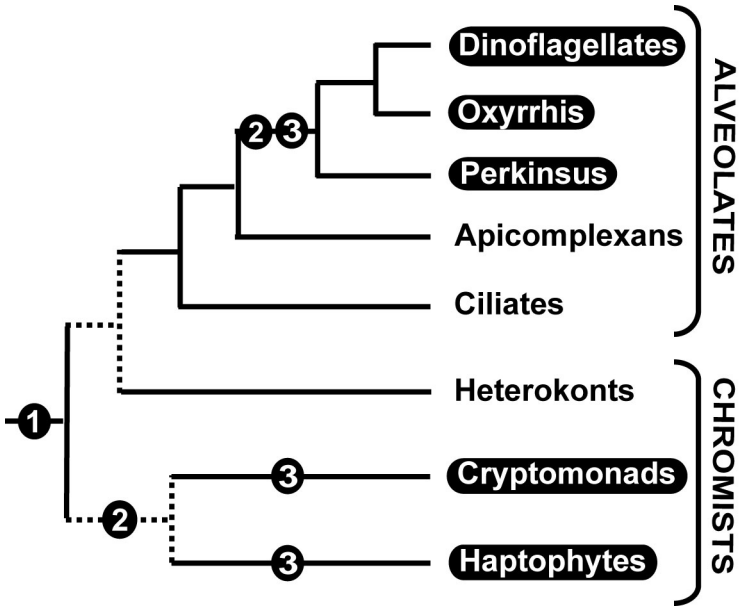
Protein maximum likelihood phylogeny of full-length EFL proteins. Major groups are bracketed and named to the right. New cryptomonad sequences are indicated by a box. Numbers at nodes correspond to bootstrap support from ML (top) and distance (bottom). Letters at nodes correspond to positions where alternative topologies were tested, the results of which are shown in Table 2.1. All analyses were repeated excluding the divergent *B. natans* sequence, but no major differences in either the tree topology or support were observed (data not shown).

In chromalveolate phylogeny (Figure 2.2), the monophyly of alveolates and branching order between them (ciliates first, then apicomplexans and dinoflagellates) are well established (Gajadhar et al. 1991; Wolters 1991; van de Peer, van der Auwera, and de Wachter 1996; Fast et al. 2002). The branching order among chromists, whether they are monophyletic or paraphyletic, and whether all three chromist groups are actually closely related to alveolates are all less clear, although many genes support a sister relationship between heterokonts and alveolates (van de Peer, van der Auwera, and de Wachter 1996; Baldauf et al. 2000; Harper, Waanders, and Keeling

2005). Regardless of those aspects of the chromalveolate phylogeny we do not yet know, there is no simple explanation for the distribution of EFL in chromalveolates. Dinoflagellates are certainly more closely related to other alveolates with EF-1α (ciliates and apicomplexans), and probably also more closely related to the EF-1α-containing heterokonts, than they are to EFL-containing haptophytes and cryptomonads. To arrive at the present distribution, therefore, EFL must have either been acquired by chromalveolates more than once, or co-existed with EF-1α for a long period of time, with different lineages subsequently losing one gene or the other. Even if EFL replaced the core translation role of EF-1α, EF-1α has several other functions in the cell and it is therefore likely that a complete loss of EF-1α would take more than just the appearance of EFL. Accordingly, we expect that the co-existence of both genes would be essential for some period of time, perhaps indefinitely under certain circumstances. It is therefore conceivable that EF-1α could 'recapture' its role in translation, making an early origin with subsequent lineage sorting a viable explanation (scheme 1 in Figure 2.2). On the other hand, EFL appears to have been acquired by several eukaryotic groups independently (Keeling and Inagaki 2004), so it may have originated in all three chromalveolate lineages independently (scheme 3 in Figure 2.2). Moreover, if we consider the possibility that cryptomonads and haptophytes are sister groups, then only two independent origins in chromalveolates would be needed to explain the distribution (scheme 2 in Figure 2.2).

If EFL originated once in chromalveolates, however, then chromalveolates should be monophyletic in phylogenies of both EF-1α and EFL. EF-1α phylogeny has been studied extensively (Baldauf 1999; Baldauf et al. 2000; Inagaki et al. 2002; Saldarriaga et al. 2003; Inagaki et al. 2004; Harper, Waanders, and Keeling 2005) and it has been shown that ciliate EF-1α genes are divergent and not monophyletic and are therefore difficult to interpret (Moreira et al. 2002), but the apicomplexan and heterokont homologues are related to one another with modest support in most analyses (Baldauf et al. 2000; Harper, Waanders, and Keeling 2005; Steenkamp, Wright, and Baldauf 2006). This suggests that EF-1α was present in the last common ancestor of at least heterokonts and alveolates.

Figure 2.2. Evolution of EFL in chromalveolates



Schematic of relationships between the six chromalveolate groups based on a variety of molecular and morphological data. Groups with names in black text possess EF-1α while groups with names in white text on black backgrounds possess EFL. Highly-supported relationships are shown as solid lines while hypothetical ones are shown as dotted lines. The alveolates (ciliates, dinoflagellates and apicomplexans) are strongly supported by virtually all known molecular and morphological data. There are also molecular data supporting a relationship between alveolates and heterokonts (see text for references), whereas the positions of haptophytes and cryptomonads are not well understood. Numbers indicate three possible scenarios to explain the current distribution of EFL and EF-1α. A single origin of EFL (at position labeled 1) with several losses of either EFL or EF-1α. Two independent origins of EFL are possible (at positions labeled 2) if haptophytes and cryptomonads are sister groups. Lastly, three origins of EFL (at positions labeled 3) are possible if all known lineages acquired EFL independently.

In EFL phylogeny, on the other hand, the EFL-encoding chromalveolates do not form a clade. In fact, despite their unusually frequent occurrence, no two EFL-encoding chromalveolate lineages cluster together: chlorophytes and chytrids are nested in the clade with *Perkinsus*, dinoflagellates, and haptophytes. This is not consistent with a single origin of EFL, and it is not simply due to a poorly resolved tree since most nodes are well supported. We specifically tested three of the more unusual aspects of this tree using Approximately Unbiased (AU) tests. First, the cryptomonads are never observed to branch with any other chromalveolate group, so we tested alternative trees where cryptomonads are moved to all internal branches (A through K in Figure 2.1). With the exception of node J, where cryptomonads are sister to green algae, and I, where cryptomonads are sister to a clade of chytrid fungi plus green algae (the topology found in

the distance tree), all these alternatives are rejected at the 5% level, including all positions with other chromalveolates (Table 2.1). It is noteworthy, however, that the tree placing cryptomonads with haptophytes is only rejected at 0.049, very close to the 5% level. Second, the position of *P. marinus* is unexpected because it is known to be a close relative of dinoflagellates (Goggin and Barker 1993; Reece et al. 1997; Saldarriaga et al. 2003; Leander and Keeling 2004), so we also tested all alternative positions of the three *P. marinus* sequences. In this case, all of the alternatives were rejected, including the expected position as sister to dinoflagellates (Table 2.1). Lastly, we forced all chromalveolates to be monophyletic. All four of these topologies were rejected regardless of the position of chromalveolates (Table 2.1). The same topologies were rejected at similar levels when *B. natans* was excluded from the analysis (data not shown).

Table 2.1. Approximately Unbiased (AU) test p-values.

Position	Cryptomonads	Perkinsus	Chromalveolates
A	0.001	0.045	0
B	0.001	0.044	0
C	0	**0.982**	0
D	0.003	NA	NA
E	0.022	NA	NA
F	0.004	0.002	NA
G	0.014	0.018	NA
H	0.049	0.001	NA
I	**0.154**	0.001	NA
J	**0.651**	0	0
K	**0.653**	0	NA
L	NA	0	0
M	NA	0	NA

Summary of AU tests comparing alternative phylogenetic positions of cryptomonad, *Perkinsus*, and chromalveolate EFL genes. Position corresponds to the label on Figure 2.1. At each position, the group being tested (cryptomonads column 1, *Perkinsus* column 2 and all chromalveolates constrained as a group in column 3) was grafted onto the position indicated and numbers are p-values from AU tests for that topology. NA indicates the position is identical to one of the other positions in that test.

These results appear to reject the conclusion that chromalveolate EFL genes are monophyletic, but it is important to note that this tree is unrooted: one could interpret it as a clade of chromalveolates with other EFL-encoding groups deriving from within (e.g., fungal and green algal genes coming from a cryptomonad or related source). In addition, we find the rejection of the monophyly of two closely related groups such as *P. marinus* and dinoflagellates highly suspicious, and therefore interpret the tree with caution regardless of the statistical

support. The phylogeny may indicate multiple origins of chromalveolate EFL genes, but independent transfers to closely related groups like *P. marinus* and dinoflagellates would require exceptionally strong evidence, and the paraphyletic relationship found here is not very compelling. Continued sampling of EFL diversity may well show that the tree is not as well supported as it appears with the current sampling.

Regardless of how many times the chromalveolates acquired EFL, its distribution in this group raises many interesting questions about its evolution. If it did arise more than once in chromalveolates or if it transferred from chromalveolates to other eukaryotes, this underscores the apparent mobility of this gene. If, on the other hand, it was acquired once in an ancestral chromalveolate, then several lineages must have subsequently lost it (at least ciliates and apicomplexans and probably also heterokonts), raising interesting questions about its functional relationship to EF-1α.

Literature Cited

Azevedo C. (1989). Fine structure of *Perkinsus atlanticus* n. sp. (Apicomplexa, Perkinsea) parasite of the clam *Ruditapes decussatus* from Portugal. J. Parasitol. 75: 627-635.

Bachvaroff TR, Concepcion GT, Rogers CR, Herman EM, Delwiche CF. (2004). Dinoflagellate expressed sequence tags data indicate massive transfer of chloroplast genes to the nuclear genome. Protist 155: 65-78.

Baldauf SL. (1999). A search for the origins of animals and chytrid fungi: comparing and combining molecular data. Am. Nat. 154: S178-S188.

Baldauf SL, Doolittle WF. (1997). Origin and evolution of the slime molds (Mycetozoa). Proc. Natl. Acad. Sci. USA 94: 12007-12012.

Baldauf SL, Palmer JD. (1993). Animals and chytrid fungi are each other's closest relatives: congruent evidence from multiple proteins. Proc. Natl. Acad. Sci. USA 90: 11558-11562.

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF. (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290: 972-977.

Bruno WJ, Socci ND, Halpern AL. (2000). Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. Mol. Biol. Evol. 17: 189-197.

Cavalier-Smith T. (1999). Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. J. Eukaryot. Microbiol. 46: 347-366.

Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu X, Reith M, Cavalier-Smith T, Maier UG. (2001). The highly reduced genome of an enslaved algal nucleus. Nature. 410: 1091-1016.

Droop MR. (1953). Phagotrophy in *Oxyrrhis marina* Dujardin. Nature 172: 250-251.

Fast NM, Kissinger JC, Roos DS, Keeling PJ. (2001). Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. Mol. Biol. Evol. 18: 418-426.

Fast NM, Xue L, Bingham S, Keeling PJ. (2002). Re-examining alveolate evolution using multiple protein molecular phylogenies. J. Eukaryot. Microbiol. 49: 30-37.

Gajadhar AA, Marquardt WC, Hall R, Gunderson J, Ariztia-Carmona EV, Sogin ML. (1991). Ribosomal RNA sequences of *Sarcocystis muris*, *Theileria annulata* and *Crypthecodinium cohnii* reveal evolutionary relationships among apicomplexans, dinoflagellates, and ciliates. Mol. Biochem. Parasitol. 45: 147-154.

Gaucher EA, Miyamoto MM, Benner SA. (2001). Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. Proc. Natl. Acad. Sci. USA 98: 548-552.

Goggin CL, Barker SC. (1993). Phylogenetic position of the genus *Perkinsus* (Protista, Apicomplexa) based on small subunit ribosomal RNA. Mol. Biochem. Parasitol. 60: 65-70.

Gould SB, Sommer MS, Hadfi K, Zauner S, Kroth PG, Maier U-G. (2006). Protein targeting into the complex plastid of cryptophytes. J. Mol. Evol. 62: 674-681.

Guindon S, Gascuel O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52: 696-704.

Hackett JD, Yoon HS, Soares MB, Bonaldo MF, Casavant TL, Scheetz TE, Nosenko T, Bhattacharya D. (2004). Migration of the plastid genome to the nucleus in a peridinin dinoflagellate. Curr. Biol. 14: 213-218.

Harper JT, Keeling PJ. (2003). Nucleus-encoded, plastid-targeted glyceraldehyde-3-phosphate dehydrogenase (GAPDH) indicates a single origin for chromalveolate plastids. Mol. Biol. Evol. 20: 1730-1735.

Harper JT, Waanders E, Keeling PJ. (2005). On the monophyly of the chromalveolates using a six-protein phylogeny of eukaryotes. Int. J. Syst. Evol. Microbiol. 55: 487-496.

Inagaki Y, Doolittle WF, Baldauf SL, Roger AJ. (2002). Lateral transfer of an EF-1alpha gene: origin and evolution of the large subunit of ATP sulfurylase in eubacteria. Curr. Biol. 12: 772-776.

Inagaki Y, Susko E, Fast NM, Roger AJ. (2004). Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1alpha phylogenies. Mol. Biol. Evol. 21: 1340-1349.

Ishida K, Green BR. (2002). Second- and third-hand chloroplasts in dinoflagellates: phylogeny of oxygen-evolving enhancer 1 (PsbO) protein reveals replacement of a nuclear-encoded plastid gene by that of a haptophyte tertiary endosymbiont. Proc. Natl. Acad. Sci. USA 99: 9294-9299.

Keeling PJ, Inagaki Y. (2004). A class of eukaryotic GTPase with a punctate distribution suggesting multiple functional replacements of translation elongation factor 1alpha. Proc. Natl. Acad. Sci. USA 101: 15380-15385.

Leander BS, Keeling PJ. (2004). Early evolutionary history of dinoflagellates and apicomplexans (Alveolata) as inferred from HSP90 and actin phylogenies. J. Phycol. 40: 341-350.

Moreira D, Kervestin S, Jean-Jean O, Philippe H. (2002). Evolution of eukaryotic translation elongation and termination factors: variations of evolutionary rate and genetic code deviations. Mol. Biol. Evol. 19: 189-200.

Patron NJ, Rogers MB, Keeling PJ. (2004). Gene replacement of fructose-1,6-bisphosphate aldolase (FBA) supports a single photosynthetic ancestor of chromalveolates. Eukaryot. Cell 3: 1169-1175.

Patron NJ, Waller RF, Keeling PJ. (2006). A tertiary plastid uses genes from two endosymbionts. J. Mol. Biol. 357: 1373-1382.

Reece KS, Siddall ME, Burreson EM, Graves JE. (1997). Phylogenetic analysis of *Perkinsus* based on actin gene sequences. J. Parasitol. 83: 417-423.

Ronquist F, Huelsenbeck JP. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572-1574.

Saldarriaga JF, McEwan ML, Fast NM, Taylor FJR, Keeling PJ. (2003). Multiple protein phylogenies show that *Oxyhrris marina* and *Perkinsus marinus* are early branches of the dinoflagellate lineage. Int. J. Sys. Evol. Microbiol. 53: 355-365.

Shimodaira H. (2002). An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51: 492-508.

Shimodaira H, Hasegawa M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17: 1246-1247.

Steenkamp ET, Wright J, Baldauf SL. (2006). The protistan origins of animals and fungi. Mol. Biol. Evol. 23: 93-106.

Strimmer K, von Haeseler A. (1996). Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13: 964-969.

Tengs T, Dahlberg OJ, Shalchian-Tabrizi K, Klaveness D, Rudi K, Delwiche CF, Jakobsen KS. (2000). Phylogenetic analyses indicate that the 19'hexanoyloxy-fucoxanthin-containing dinoflagellates have tertiary plastids of haptophyte origin. Mol. Biol. Evol. 17: 718-729.

Van de Peer Y, Van der Auwera G, De Wachter R. (1996). The evolution of stramenopiles and alveolates as derived by "substitution rate calibration" of small ribosomal subunit RNA. J Mol. Evol. 42: 201-210.

Wolters J. (1991). The troublesome parasites: molecular and morphological evidence that Apicomplexa belong to the dinoflagellate-ciliate clade. Biosystems 25: 75-84.

Yoon HS, Hackett JD, Pinto G, Bhattacharya D. (2002). A single, ancient origin of the plastid in the Chromista. Proc. Natl. Acad. Sci. USA. 99: 15507-15512.

Yoon HS, Hackett JD, Van Dolah FM, Nosenko T, Lidie KL, Bhattacharya D. (2005). Tertiary endosymbiosis driven genome evolution in dinoflagellate algae. Mol. Biol. Evol. 22: 1299-1308.

CHAPTER 3: THE DISTRIBUTION OF EF-1A, EFL, AND A NON-CANONICAL GENETIC CODE IN THE ULVOPHYCEAE: DISCRETE GENETIC CHARACTERS SUPPORT A CONSISTENT PHYLOGENETIC FRAMEWORK[2]

Introduction

Mattox and Stewart (1984) defined the green algal class Ulvophyceae mainly on the basis of the counter-clockwise offset of the cruciate flagellar basal apparatus and the mode of cytokinesis, which involves neither a phycoplast (precluding placement in the Chlorophyceae) nor a phragmoplast (a Charophycean feature). Based on this definition, O'Kelly and Floyd (1984) included five orders in the Ulvophyceae: Ulvales, Ulotrichales, Siphonocladales, Dasycladales, and Caulerpales; the Trentepohliales were omitted on the basis of anomalous features reminiscent of the Charophyceae, such as a multilayered structure in the flagellar root system and plasmodesmata between vegetative cells (O'Kelly and Floyd 1984). While some molecular phylogenetic analyses of 18S rRNA weakly recover ulvophycean monophyly (López-Bautista and Chapman 2003; Watanabe and Nakayama 2007), those with all orders represented suggest two distinct, non-sister lineages within a clade that also includes members of the Chlorophyceae and Trebouxiophyceae (Zechman et al. 1990; Watanabe, Kuroda, and Maiwa 2001). These lineages have been referred to as the Ulvophyceae I, which includes the orders Siphonocladales, Dasycladales, Caulerpales, and Trentepohliales, and the Ulvophyceae II, which includes the orders Ulvales and Ulotrichales (Watanabe, Kuroda, and Maiwa 2001). Overall, it seems likely that the Ulvophyceae is not monophyletic, unless more narrowly described to include just the Ulvophyceae II clade (Ulvophyceae *sensu* van den Hoek, Mann, and Jahns 1995; Watanabe, Kuroda, and Maiwa 2001), but a taxonomic revision awaits further evidence.

To further refine our understanding of relationships among ulvophyceans, we have examined taxa from the five orders identified by O'Kelly and Floyd (1984) plus *I. tetrasporus* which is not currently included in an order, for the presence of two discrete genetic characters: the presence of elongation factor 1α (EF-1α) versus elongation factor-like (EFL) proteins, and the presence of a non-canonical genetic code where TAA and TAG encode glutamine (see below). The eukaryotic elongation factor 1-alpha (EF-1α, also known as EF1A) plays an essential role in translation by bringing aminoacyl-tRNAs to the ribosome, and was thought to be ubiquitous. However, it was recently discovered that certain eukaryotic groups lack EF-1α

[2] A version of this chapter has been published. Gile GH, Novis PM, Cragg DS, Zuccarello GC, Keeling PJ. 2009. The distribution of elongation factor-1alpha (EF-1α), elongation factor-like (EFL), and a non-canonical genetic code in the Ulvophyceae: Discrete genetic characters support a consistent phylogenetic framework. Journal of Eukaryotic Microbiology. 56:367-372.

altogether and instead possess a distinct paralogue called EFL for Elongation Factor-Like (Keeling and Inagaki 2004). Within the Chlorophyta, all investigated species possess EFL, with one intriguing exception: the ulvophycean *Acetabularia acetabulum* (Dasycladales) possesses EF-1α. The relationships between green algal EFL genes are not well resolved, but the EF-1α gene from *A. acetabulum* is clearly related to EF-1α from charophytes and land plants, suggesting that at least EF-1α was present in the ancestor of Viridiplantae (Noble, Rogers, and Keeling 2007). The Ulvophyceae are therefore at the centre of the puzzle of EF-1α/EFL evolution in the green algae, but EFL and EF-1α have been characterized from only one ulvophycean order each: EFL in the Ulvales (Ulvophyceae II) from two *Ulva* species (Noble, Rogers, and Keeling 2007), and EF-1α in the Dasycladales (Ulvophyceae I) from *A. acetabulum* (Keeling and Inagaki 2004).

Another unusual molecular character in the Dasycladales is a non-canonical genetic code. This was first discovered in *A. acetabulum* (Schneider, Leible, and Yang 1989) and later in *Batophora oerstedii* (Schneider and de Groot 1991). In these genomes the canonical stop codons UAA and UAG specify glutamine. The same non-canonical code also occurs in the nuclear genomes of oxymonads (Keeling and Leander 2003; de Koning et al. 2007) and diplomonads (Keeling and Doolittle 1996, 1997), and it has likely arisen at least twice in ciliates (Baroin-Tourancheau et al. 1995; Lozupone, Knight, and Landweber 2001). In this study, we find that both characters are more broadly and informatively distributed than was previously recognized. EF-1α was found in Dasycladales, Siphonocladales, Caulerpales, and *I. tetrasporus,* whereas EFL was found in Ulvales and Ulotrichales. The non-canonical genetic code, in turn, was found in Dasycladales and Siphonocladales, but not in Caulerpales (or Ulvales or Ulotrichales). Together these characters support previous suggestions of a clade of Dasycladales, Siphonocladales, Caulerpales, and *I. tetrasporus* (Watanabe, Kuroda, and Maiwa 2001; Watanabe and Nakayama 2007) and a specific relationship between the Siphonocladales and Dasycladales within this clade, which is also consistent with ultrastructural characters (O'Kelly and Floyd 1984; Roberts, Stewart, and Mattox 1984; Sluiman 1989).

Materials and Methods

Culture sources

Seven strains of ulvophycean green algae from five orders were used in this study. Four were ordered from either the Provasoli-Guillard National Center for Culture of Marine Phytoplankton (CCMP) or the Culture Collection of Algae at the University of Texas at Austin

(UTEX): *Ochlochaete hystrix* Thwaites ex Harvey (CCMP 2319), *Urospora sp*. (CCMP 1082), *Ignatius tetrasporus* Bold et MacEntee (UTEX B 2012), and *Parvocaulis pusillus* (Howe) Berger, Fettweiss, Gleissberg, Liddle, Richter, Sawitsky & Zuccarello (UTEX LB 2710). *Caulerpa* cf. *racemosa* (Forsskål) Agardh was donated by the Vancouver Aquarium. *Chaetomorpha coliformis* (Montagne) Kuetzing was collected at Taylor's Mistake, New Zealand, 18 December 2007, and *Cladophora cf. crinalis* Harvey was collected at Wainui, Akaroa Harbour, New Zealand, 26 October 2007. Voucher specimens for the latter two collections were deposited at the Allan Herbarium, Lincoln, New Zealand, numbers CHR585485 and CHR585488.

RNA extraction, PCR, and sequencing methods

Total RNA was extracted from algal cell pellets of *Ochlochaete*, *Urospora*, *Ignatius*, *Parvocaulis*, and *Caulerpa* using Trizol reagent (Invitrogen, Carlsbad, CA) or the RNeasy Mini kit (Qiagen, Mississauga, ON, Canada) for *Chaetomorpha* and *Cladophora*. EFL and EF-1α cDNA sequences were obtained by 3' RACE using the following nested degenerate forward primers designed to be universal for any eukaryotic EFL and EF-1α: 5'-GTCGARATGCAYCAY-3' (outer) and 5'- CCGGGCGAYAAYGTNGG-3' (inner) using the FirstChoice RLM-RACE kit (Ambion, Austin, TX). Gene-specific reverse primers were designed from EFL and EF-1α 3' RACE sequences and used with the following nested degenerate forward primers using Superscript III One-Step RTPCR with Platinum Taq (Invitrogen): for EFL - 5'-CTGTCGATCGTCATHTGYGGN-3', 5'-CATGTCGACTCGGGCAAGTCNACNACNACNGG-3', and 5'-AACATCGTCGTGATHGGNCAYGTNGA-3'; for EF-1α - 5'-CATGTCGACTCGGGCAAGTCNACNACNACNGG-3' and 5'-TTCGAGAAGGAGGCNGCNGARATGAA-3'. PCR products of *Ochlochaete*, *Urospora*, *Ignatius*, *Parvocaulis* and *Caulerpa* were cloned using the TOPO-TA cloning kit (Invitrogen) and for *Chaetomorpha* and *Cladophora* using the pGEM-T Easy vector system (Promega, Madison, WI). New sequences obtained in both directions using BigDye Terminator v. 3.1 (Applied Biosystems, Foster City, CA) were deposited in GenBank under accession numbers FJ539138-FJ539144.

Phylogenetic analyses

New and previously published EFL and EF-1α sequences were translated and aligned by Multiple Alignment using Fast Fourier Transform (MAFFT, Katoh et al. 2002) and edited in

MacClade 4.08 (Maddison and Maddison 2003) to final matrix sizes of 32 taxa and 444 unambiguously aligned characters for EFL and 29 taxa and 426 characters for EF-1α. Phylogenetic trees were inferred using maximum likelihood (ML) and Bayesian methods. ProtTest 1.4 (Abascal, Zardoya, and Posada 2005) was used to determine the best amino acids replacement models and analysis parameters. ML trees were inferred using RAxML 7.0.3 (Stamatakis 2006) on the CIPRES portal (http://8ball.sdsc.edu:8889/cipres-web/Home.do) using the RtREV amino acid substitution matrix (Dimmic et al. 2002), four rate categories approximated by a Γ distribution with shape parameter α estimated from the data, amino acid frequencies calculated from the data, and in the case of EF-1α, the proportion of invariable sites also calculated from the data. The EFL alignment had a negligibly low proportion of invariable sites. Two hundred fifty bootstrap replicates were performed for both datasets, as computed to be sufficient by RAxML (Stamatakis, Hoover, and Rougemont 2008). MRBAYES 3.1.2 (Ronquist and Huelsenbeck 2003) was used to perform Bayesian analyses using the RtREV amino acids substitution matrix and the same parameters as the likelihood analyses. Five independent analyses for each of the EFL and EF-1α datasets were carried out in order to test for convergence, and all five analyses for each protein produced identical topologies and near-identical posterior probabilities. One cold and three heated chains were run for all analyses, sampling one tree per thousand generations, and 50% majority rule consensus trees were computed after observing that log likelihood (lnL) values stabilized at 5,000 generations and discarding the first five sampled trees. Consensus trees were also computed after discarding the first 100 sampled trees for the 5,000 generation runs and first 1,000 sampled trees for the 5,000,000 generations runs, with no effect on the consensus topology or posterior probabilities (data not shown). All five EFL analyses were run for 5,000,000 generations, and one consensus tree was arbitrarily chosen to indicate posterior probabilities on the ML topology. Four of the EF-1α analyses were run for one million generations and one for 5,000,000 generations; the longer run was chosen to represent the topology and posterior probabilities. For the EF-1α tree, ML branch lengths were computed using TREE-PUZZLE 5.1 (Schmidt et al. 2002), and displayed on the Bayesian topology.

Approximately Unbiased (AU) tests (Shimodaira 2002) were carried out using CONSEL 1.19 (Shimodaira and Hasegawa 2001) to evaluate the likelihood of alternate EF-1α and EFL topologies in which the Ulvophyceae is constrained as monophyletic. Site likelihoods were calculated by TREE-PUZZLE 5.1 (Schmidt et al. 2002) using the -wsl (with site likelihoods) option, the WAG amino acids substitution matrix (Whelan and Goldman 2001), and 4 Γ rate

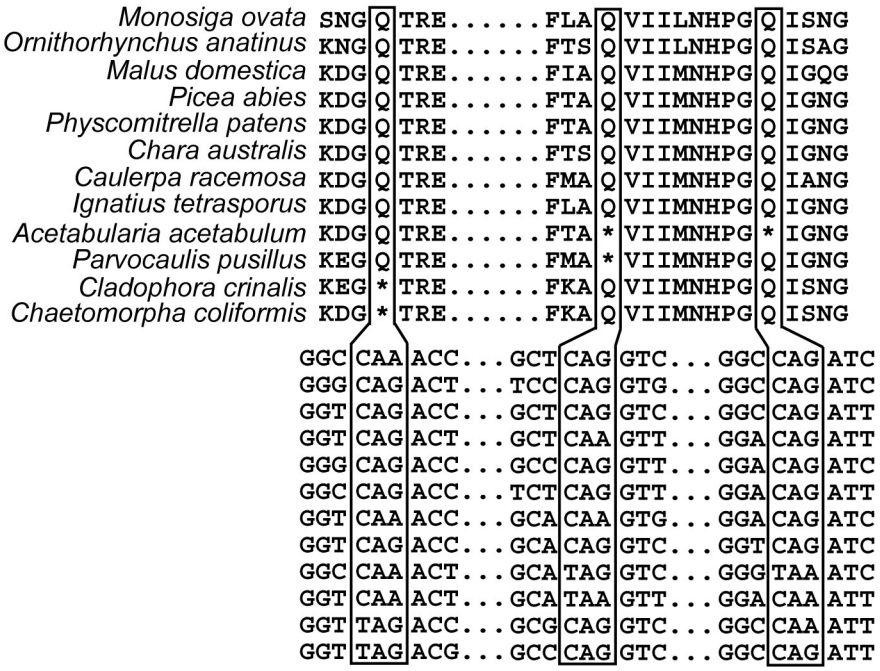
categories with parameter α, amino acid frequencies, and the proportion of invariable sites estimated from the data. Because the RtREV model is not available in this program, ML trees were also inferred with the WAG substitution model (Whelan and Goldman 2001), the second best model according to ProtTest. The resulting topologies were congruent with only minor, unsupported differences: *C. racemosa* and *I. tetrasporus* form a poorly supported clade (24%) in the EF-1α tree, and the EFL tree was identical to the Bayesian topology (see Results for details).

Results

Distribution of EFL and EF-1α and a non-canonical genetic code

In this study, seven species representing five ulvophycean orders were tested for the presence of EF-1α and EFL by RT-PCR. EFL sequences were determined from *Ochlochaete hystrix* (Ulvales) and *Urospora* sp. CCMP1082 (Ulotrichales), both members of the Ulvophyceae II clade. EF-1α sequences were determined from all other species tested, all of which belong to the Ulvophyceae I: *Caulerpa cf. racemosa* (Caulerpales), *Parvocaulis pusillus* (Dasycladales), *Cladophora cf. crinalis* and *Chaetomorpha coliformis* (Siphonocladales). None of the species investigated were found to express both genes. EF-1α sequences for *C. crinalis* and *C. coliformis* in the Siphonocladales and *P. pusillus* from the Dasycladales were found to use a non-canonical genetic code. In all three genes, a UAA or UAG codon was found at one or more positions that are otherwise highly conserved for the amino acid glutamine (Figure 3.1). The *C. racemosa* EF-1α sequence contained neither UAA nor UAG codons.

Figure 3.1. Non-canonical code in Ulvophyceae.

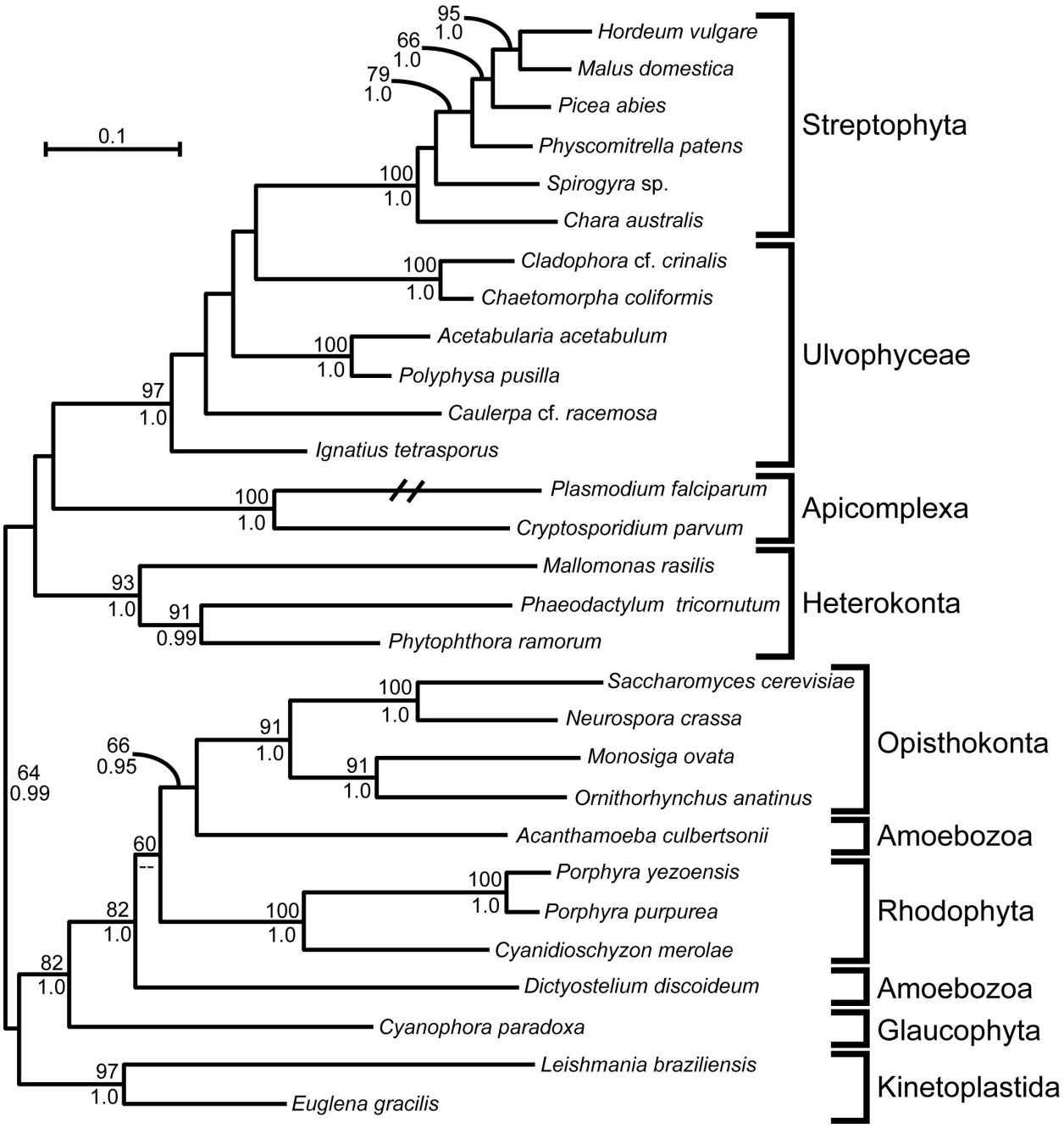


Selected regions of the elongation factor 1α (EF-1α) alignment indicating TAA and TAG codons at positions conserved for glutamine (Q) in the dasycladaleans *Parvocaulis pusillus* and *Acetabularia acetabulum* and the siphonocladaleans *Cladophora crinalis* and *Chaetomorpha coliformis*.

Phylogenetic analysis of EF-1α and EFL in the Ulvophyceae

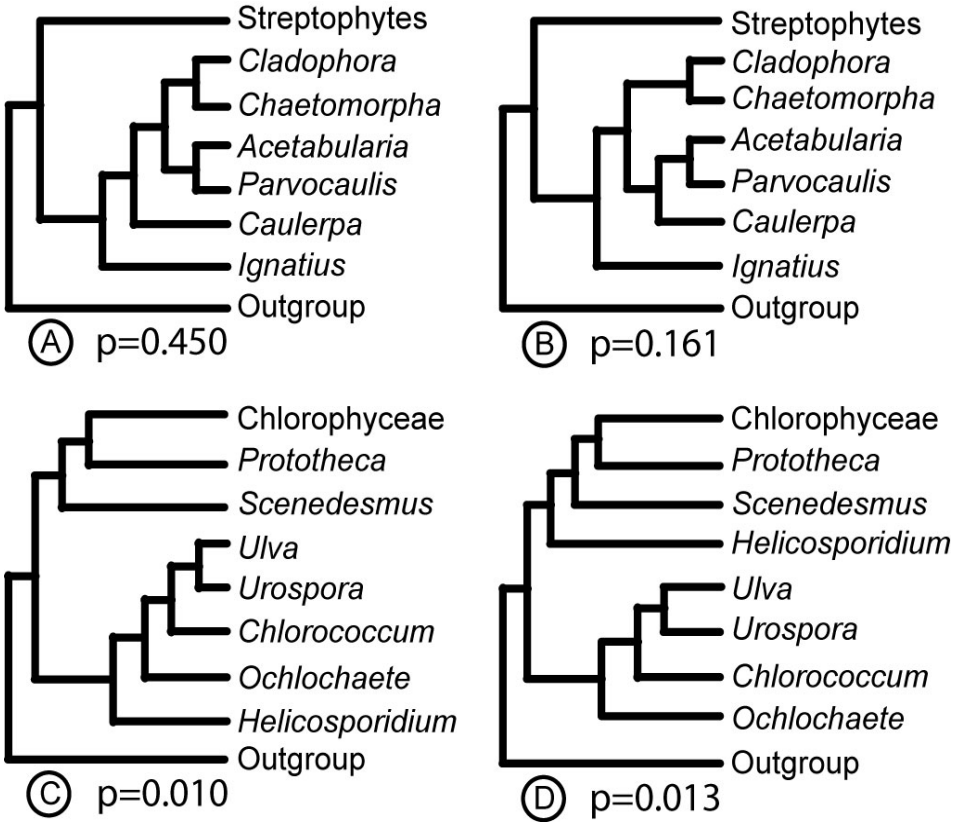
Phylogenetic analyses of EFL and EF-1α were carried out to gain insight into the evolutionary history of these proteins, but they should not be interpreted as reflective of ulvophycean relationships as neither of these proteins is well-suited for inferring organismal phylogenies (Keeling and Inagaki 2004; Roger et al. 1999). In the EF-1α phylogeny, the branch uniting streptophytes is strongly supported (100% bootstrap support, 1.0 posterior probability), as is the branch uniting streptophyte and ulvophycean EF-1α sequences (97% bootstrap support, 1.0 posterior probability, Figure 3.2). The ulvophycean sequences as a whole do not group together (only species from the same order form supported clades), but AU tests failed to reject the possibility of a monophyletic Ulvophyceae I with (p=0.45) or without (p=0.161) Siphonocladales and Dasycladales constrained as sister groups (Figure 3.3A, B). Therefore, while the best maximum likelihood topology of the EF-1α tree does not recover accepted relationships among members of the Ulvophyceae and Streptophyta, it is not inconsistent with their monophyly (Figure 3.2), suggesting that it was likely encoded in the common ancestor of these two groups.

Figure 3.2. Phylogeny of EF-1α.



Maximum likelihood (ML) phylogeny of elongation factor 1α (EF-1α) with major lineages bracketed to the right. Hash marks indicate branches whose lengths have been reduced by precisely one half, while ML bootstrap values of 50% or greater (above) and Bayesian posterior probabilities of 0.9 or greater (below) are indicated at nodes.

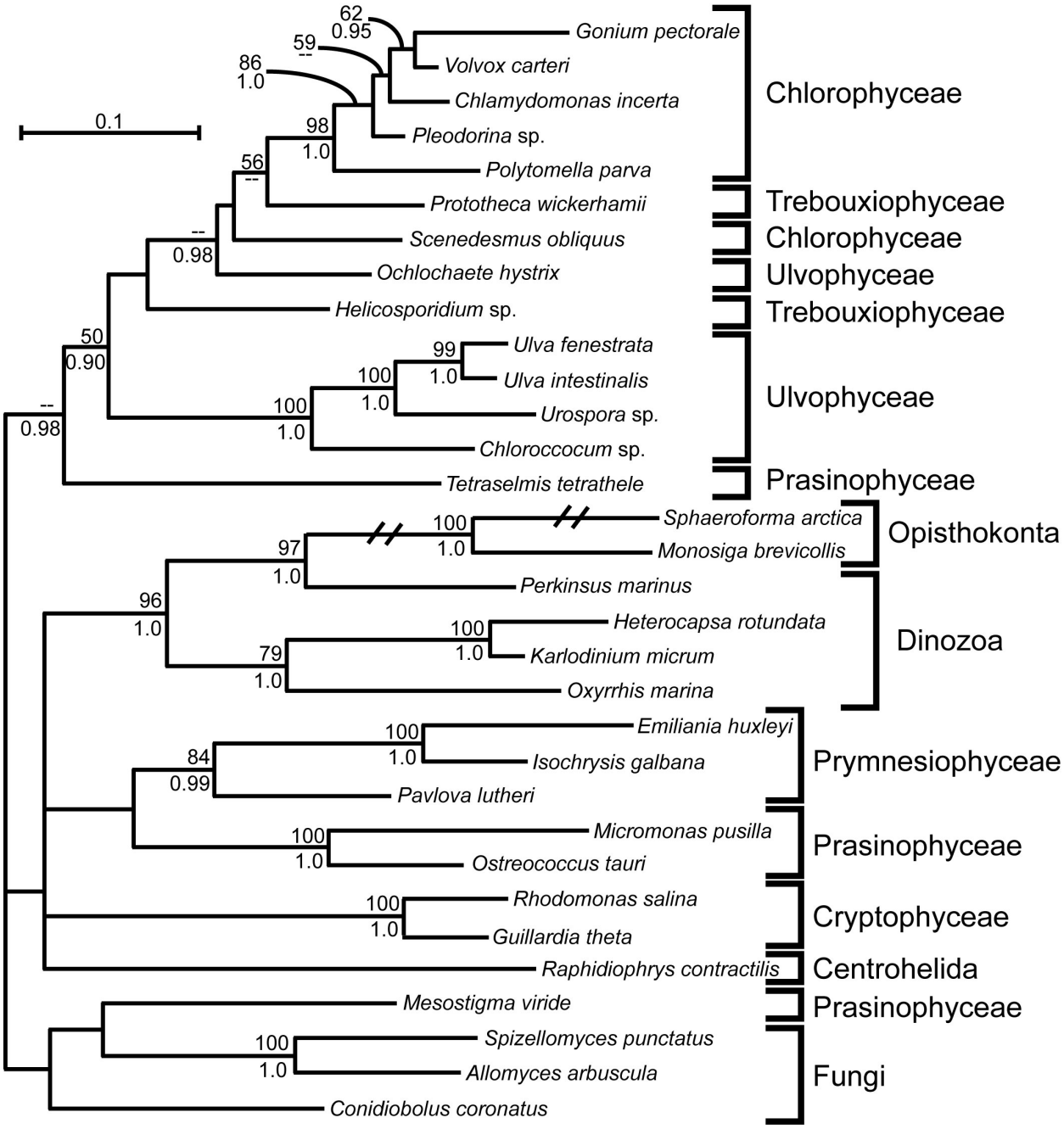
Figure 3.3. Topologies evaluated by approximately unbiased (AU) tests.



Alternative topologies of EF-1α (A, B) and EFL (C, D) tested to evaluate the probability that ulvophycean EF-1α and EFL sequences were vertically inherited. The p-value of each approximately unbiased (AU) test is indicated below its respective topology. A. Siphonocladales (*Cladophora* and *Chaetomorpha*) and Dasycladales (*Acetabularia* and *Parvocaulis*) are constrained as sisters within a monophyletic Ulvophyceae I, as suggested by the distribution of the non-canonical genetic code and ultrastructural data. B. Siphonocladales (*Cladophora* and *Chaetomorpha*) and Caulerpales (*Caulerpa*) are instead constrained as sisters as suggested by phylogenetic analyses of 18S rRNA sequences. C. Ulvophycean EFL is constrained as monophyletic with *Helicosporidium* (Trebouxiophyceae) placed according to its position in the ML analysis of EFL. D. Ulvophycean EFL is constrained as monophyletic but with *Helicosporidium* placed according to its position in the Bayesian analysis of EFL.

Our Bayesian EFL analysis (Figure 3.4) is congruent with previous analyses, but the new ulvophycean sequences are not monophyletic: *Urospora* and the *Ulva* species form a strongly supported clade, but *Ochlochaete hystrix*, which is more closely related to *Ulva* than either is to *Urospora* (according to phylogenetic analyses of 18S rRNA and chloroplast-encoded *tufA* gene sequences, O'Kelly, Wysor, and Bellows 2004), is excluded. The coccoid green alga identified as *Chlorococcum* sp. (NEPCC 478) also groups robustly with the *Ulva* clade, but as the support is high and many other coccoid green algae have been transferred out of their traditional morphology-based genera, even into new classes (Lewis & McCourt 2004), we suspect that this *Chlorococcum* strain is more likely a misidentified member of the Ulvales or Ulotrichales than a true *Chlorococcum* species. As a result of the unexpected placement of *O. hystrix*, the phylogeny of EFL is both consistent with a single origin of EFL in the Ulvophyceae, Trebouxiophyceae, and Chlorophyceae (collectively referred to as the "UTC clade") and inconsistent with a single origin of EFL in the Ulvophyceae. AU tests further supported the non-monophyly of ulvophycean EFL (Figure 3.3C, D). The ML topology (not shown) was essentially identical to the Bayesian topology, having only unsupported differences. The position of *O. hystrix* is the same but lacks support (44%), *Helicosporidium* falls as sister to the *Ulva* clade with poor support (6%), and *T. tetrathele* groups with *Raphidiophrys contractilis* with poor support (8%).

Figure 3.4. Phylogeny of EFL



Bayesian phylogeny of elongation factor-like (EFL) with maximum likelihood (ML) branch lengths. Major lineages are bracketed to the right. Hash marks indicate branches whose lengths have been reduced by precisely one half, while ML bootstrap values of 50% or greater (above) and Bayesian posterior probabilities of 0.9 or greater (below) are indicated at nodes.

Discussion

Previously, the distribution of EFL and EF-1α in the Ulvophyceae was only known from *Ulva fenestrata*, *Ulva intestinalis*, and *Acetabularia acetabulum*. We have used seven species from five orders of the Ulvophyceae to determine the presence of EFL and EF-1α by 3' RACE and RT-PCR, and their respective distributions were found to correspond to the two major groups within the class. EFL was found in the Ulvales from *O. hystrix* and previous *Ulva* sequences and in the Ulotrichales from *Urospora* sp. CCMP 1082, the two orders that make up the Ulvophyceae II group. EF-1α, on the other hand, was found in the Caulerpales from *Caulerpa* cf. *racemosa*, Dasycladales from *Parvocaulis pusillus* and previously from *A. acetabulum*, and Siphonocladales from *Cladophora* cf. *crinalis* and *Chaetomorpha coliformis*, all of which are members of the Ulvophyceae I group. *Ignatius tetrasporus* was also found to encode EF-1α, supporting its inclusion in Ulvophyceae I. None of the selected taxa were found to express both EFL and EF-1α. During the course of our research, other members of the Ulvophyceae I (i.e. *Ostreobium quekettii*, *Blastophysa rhizopus,* and *Codium*, *Derbesia*, and *Bryopsis* species) were independently found to encode EF-1α by examining gene fragments by PCR (Cocquyt, E. E., pers. comm.).

In the EF-1α sequences for *C. crinalis* and *C. coliformis* in the Siphonocladales and *P. pusillus* from the Dasycladales, the codons UAA or UAG were found at one or more positions that are otherwise highly conserved for the amino acid glutamine (Figure 3.1). We interpret this as being due to the use of a non-canonical code because the same code has been identified in *A. acetabulum* and confirmed by protein sequencing (Schneider et al. 1989) and also identified in the closely related dasycladalean *Batophora oerstedii* (Schneider and de Groot 1991). It is therefore no surprise to find this code in *P. pusillus*, which is even more closely related to *A. acetabulum* than *B. oerstedii* is (Zechman 2003), but its use by two members of the Siphonocladales is potentially informative. The EF-1α sequences from *C. racemosa* and *I. tetrasporus* use only canonical CAA and CAG codons for glutamine, and the coding sequences terminate with UGA codons, characteristics that are consistent with either genetic code. However, publicly available sequences from Caulerpales (i.e. *Bryopsis hypnoides* (lectin, Genbank EU410470), *Bryopsis maxima* (RNase Bm2, AB164318), *Bryopsis plumosa* (bryohealin, EU769118), and *Flabellia petiolata* (P-type ATPase, AJ972675), all use canonical glutamine codons, and coding sequences terminate with UGA, UAA or UAG codons, indicating that the Caulerpales as a whole use the universal code. Additional sequences from *Caulerpa* and other genera support this conclusion (unpubl. data and Cocquyt, E. E., pers. comm.).

The discovery of a non-canonical genetic code in the Siphonocladales expands the known distribution of this character within the Ulvophyceae I, and combined with ultrastructural evidence, supports a sister-group relationship between the Siphonocladales and Dasycladales. Genetic code changes in nuclear genomes are quite rare, having occurred in only a handful of eukaryotic lineages. Although the conversion of UAA and UAG codons from specifying stop to specifying glutamine has happened twice within the ciliates (Baroin-Tourancheau et al. 1995; Lozupone et al. 2001), this is unlikely to be the case in the Ulvophyceae for two main, interrelated reasons. First, Siphonocladales and Dasycladales undoubtedly share a recent common ancestor with only residual uncertainty about whether they are sisters or whether one of them is closer to the Caulerpales (Roberts et al. 1984; Sluiman 1989; Watanabe, Kuroda, and Maiwa 2001; Watanabe and Nakayama 2007; Zechman et al. 1990). The two lineages of ciliates known to share this code are far more distantly related to one another than the Siphonocladales are to the Dasycladales. Second, sisterhood of Siphonocladales and Dasycladales has been proposed previously on the basis of shared ultrastructural characters. These include a somewhat flattened cruciate arrangement of basal bodies and roots, a striated distal fiber connecting the two distal basal bodies, and a transverse septum in the flagellar transition zone (Roberts et al. 1984; Sluiman 1989). Molecular analyses neither support nor refute this hypothesis: they have either failed to resolve the relationships among these three orders (Zechman et al. 1990) or failed to include sequences from the Caulerpales (Watanabe, Kuroda, and Maiwa 2001; López-Bautista and Chapman 2003). Because the evidence from molecular phylogeny, morphology, and now the shared retention of EF-1α all support a monophyletic Ulvophyceae I, and Dasycladales and Siphonocladales share ultrastructural features, the most straightforward interpretation of the distribution of genetic codes is that Dasycladales and Siphonocladales share a common ancestor to the exclusion of Caulerpales.

The distributions of discrete genetic characters can be useful in inferring phylogenetic relationships, especially if they are consistent with other forms of data and when evidence from ultrastructural features and molecular phylogenies is in conflict or inconclusive. Previous molecular phylogenetic analyses provided conflicting placements of *I. tetrasporus*: a Bayesian analysis of 18S rRNA placed *I. tetrasporus* and its sister taxon *Pseudocharacium americanum* as early-diverging members of the Ulvophyceae II clade, while a distance analysis placed them at the base of the Ulvophyceae I, though neither placement was strongly supported (Watanabe and Nakayama 2007). The authors hypothesized that *I. tetrasporus* belongs with the Ulvophyceae I clade on the basis of their ultrastructural analysis, which is consistent with our findings. *Ignatius*

tetrasporus is only one of several putative ulvophyceans with uncertain affinities, however. Trentepohliales are hypothesized to be sisters to the Dasycladales and/or Siphonocladales, but their exact placement is uncertain (López-Bautista and Chapman 2003). If this hypothesis were correct, we would predict that this group also possesses EF-1α, and examination of their genetic code may be especially informative. The affinities of *Oltmannsiellopsis viridis* may also be clarified by determining which elongation factor it encodes. This taxon has been shown to branch at the base of the Ulvophyceae with strong support (Friedl and O'Kelly 2002), but because members of the Ulvophyceae I clade were omitted, its precise position remains unclear. Finally, certain trebouxiophytes show a weak affinity to the Ulvophyceae I clade in small subunit rRNA trees (Watanabe, Kuroda, and Maiwa 2001). It would be of interest to determine whether these taxa also possess EF-1α, and by extension, whether they might be better placed in the Ulvophyceae I.

Literature Cited

Abascal F, Zardoya R, Posada D. 2005. ProtTest: Selection of best-fit models of protein evolution. Bioinformatics 21:2104-2105.

Baroin-Tourancheau A, Tsao N, Klobutcher LA, Pearlman RE, Adoutte A. 1995. Genetic code deviations in the ciliates: evidence for multiple and independent events. EMBO J. 14:3262-3267.

Dimmic MW, Rest JS, Mindell DP, Goldstein RA. 2002. RtREV: An amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J. Mol. Evol. 55: 65-73.

Friedl T, O'Kelly CJ. 2002. Phylogenetic relationships of green algae assigned to the genus *Planophila* (Chlorophyta): evidence from 18S rDNA sequence data and ultrastructure. Eur. J. Phycol. 37:373-384.

van den Hoek C, Mann D, Jahns H. 1995. Algae: An introduction to phycology. Cambridge: Cambridge University Press. 623 p.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30: 3059-3066.

Keeling PJ, Doolittle WF. 1996. A non-canonical genetic code in an early diverging eukaryotic lineage. EMBO J. 15:2285-2290.

Keeling PJ, Doolittle WF. 1997. Widespread and ancient distribution of a non-canonical genetic code in diplomonads. Mol. Biol. Evol. 14:895-901.

Keeling PJ, Inagaki Y. 2004. A class of eukaryotic GTPase with a punctate distribution suggesting multiple functional replacements of translation elongation factor 1-alpha. Proc. Natl. Acad. Sci. USA. 101:15380-15385.

Keeling PJ, Leander BS. 2003. Characterization of a non-canonical genetic code in the oxymonad *Streblomastix strix*. J. Mol. Biol. 326:1337-1349.

de Koning AP, Noble GP. Heiss AA, Wong J, Keeling PJ. 2008. Environmental PCR survey to determine the distribution of a non-canonical genetic code in uncultivable oxymonads. Environ. Microbiol. 10:65-74.

Lewis LA, McCourt RM. 2004. Green algae and the origin of land plants. Am. J. Bot. 91:1535-1556.

López-Bautista JM, Chapman RL. 2003. Phylogenetic affinities of the Trentepohliales inferred from small-subunit rDNA. Int. J. Syst. Evol. Microbiol. 53:2099-2106.

Lozupone CA, Knight RD, Landweber LF. 2001. The molecular basis of nuclear genetic code change in ciliates. Curr. Biol. 11:65-74.

Maddison DR, Maddison WP. 2003. MacClade 4: Analysis of phylogeny and character evolution

[computer program]. Version 4.08. Sinauer Associates, Sunderland, Massachusetts.

Mattox KR, Stewart KD. 1984. Classification of the green algae: a concept based on comparative cytology. Irvine DEG, John DM editors. In: Systematics of the green algae. London: Academic Press p. 29-72.

Noble GP, Rogers MB, Keeling PJ. 2007. Complex distribution of EFL and EF-1alpha proteins in the green algal lineage. BMC Evol. Biol. 7:82.

O'Kelly CJ, Floyd GL. 1984. Correlations among patterns of sporangial structure and development, life histories, and ultrastructural features in the Ulvophyceae. Irvine DEG, John DM editors. In: Systematics of the green algae. London: Academic Press. p. 121-156.

O'Kelly CJ, Wysor B, Bellows WK. 2004. Gene sequence diversity and the phylogenetic position of algae assigned to the genera *Phaeophila* and *Ochlochaete* (Ulvophyceae, Chlorophyta). J. Phycol. 40:789-799.

Roberts KR, Stewart KD, Mattox KR. 1984. Structure and absolute configuration of the flagellar apparatus in the isogametes of *Batophora* (Dasycladales, Chlorophyta). J. Phycol. 20:183-191.

Roger AJ, Sandblom O, Doolittle WF, Philippe H. An evaluation of elongation factor 1alpha as a phylogenetic marker for eukaryotes. Mol. Biol. Evol. 16:218-233.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502-504.

Schneider SU, de Groot EJ. 1991. Sequences of two *rbcS* cDNA clones of *Batophora oerstedii*: Structural and evolutionary considerations. Curr. Genet. 20:173-175.

Schneider SU, Leible MB, Yang XP. 1989. Strong homology between the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase of two species of *Acetabularia* and the occurrence of unusual codon usage. Mol. Gen. Genet. 218:445-452.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492-508.

Shimodaira H, Hasegawa M. 2001. CONSEL: For assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246-1247.

Sluiman HJ. 1989. The green algal class Ulvophyceae: an ultrastructural survey and classification. Crypt. Bot. 1:83-94.

Stamatakis A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with

thousands of taxa and mixed models. Bioinformatics 22:2688-2690.

Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML web-servers. Syst. Biol. 57:758-771.

Watanabe S, Nakayama T. 2007. Ultrastructure and phylogenetic relationships of the unicellular green algae *Ignatius tetrasporus* and *Pseudocharacium americanum* (Chlorophyta). Phycol. Res. 55:1-16.

Watanabe S, Kuroda N, Maiwa F. 2001. Phylogenetic status of *Helicodictyon planctonicum* and *Desmochloris halophila* gen. et comb. nov. and the definition of the class Ulvophyceae (Chlorophyta). Int. J. Syst. Evol. Microbiol. 40:421-434.

Whelan S, 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.

Zechman FW. 2003. Phylogeny of the Dasycladales (Chlorophyta, Ulvophyceae) based on analyses of RUBISCO large subunit (*rbcL*) gene sequences. J. Phycol. 39:819-827.

Zechman FW, Theriot EC, Zimmer EA, Chapman RL. 1990. Phylogeny of the Ulvophyceae (Chlorophyta): Cladistic analysis of nuclear-encoded rRNA sequence data. J. Phycol. 26:700-710.

CHAPTER 4: DISTRIBUTION AND PHYLOGENY OF EFL AND EF-1A IN EUGLENOZOA SUGGEST ANCESTRAL CO-OCCURRENCE FOLLOWED BY DIFFERENTIAL LOSS[3]

Introduction

The essential eukaryotic translation elongation factor EF-1α and its distantly related paralogue EFL (for EF-Like) are GTPases with a complex, mutually exclusive distribution. While EF-1α is well known from plants, animals, and fungi, and has been characterized at the structural (Andersen et al. 2001) and functional (Negrutskii and El'skaya 1998) levels, EFL was discovered more recently in a small number of single-celled eukaryotes that were found to lack EF-1α (Keeling and Inagaki 2004). EFL is considered likely to perform the same canonical translation function as EF-1α due to their mutually exclusive distribution and the observation that EF-1α's binding sites for EF-1β, aminoacyl-tRNAs, and GTP are conserved in EFL (Keeling and Inagaki 2004), though no functional analyses of EFL have been carried out. Curiously, EFL-encoding lineages are scattered across the tree of eukaryotes, such that they are each more closely related to an EF-1α-encoding lineage than they are to one other. This complex pattern has persisted despite further studies of EFL in green algae (Noble, Rogers, and Keeling 2007), fungi (James et al. 2006), ichthyosporids (Ruiz-Trillo et al. 2006; Marshall et al. 2008), cryptophytes, haptophytes, red algae (Gile, Patron, and Keeling 2006; Sakaguchi et al. 2009), and diatoms (Kamikawa, Inagaki, and Sako 2008) that have greatly expanded its known distribution. In general, the phylogeny of EFL is incongruent with the phylogeny of the organisms in which it is found, which is not consistent with a single ancestral origin of eukaryotic EFL genes. As a result, multiple lateral gene transfers are often invoked to explain the complex distribution of EFL, despite the lack of compelling evidence for this interpretation. Only in one case did the phylogeny of EFL reveal a potential donor lineage for the putative lateral gene transfer (Kamikawa, Inagaki, and Sako 2008). In addition to lateral gene transfer, differential loss of EFL and EF-1α is a mechanism that can explain the unusual distribution of these two proteins. This possibility has not been explored as fully, although a close examination of the distribution of EFL in green algae pointed to this as a contributing factor in that lineage (Noble, Rogers, and Keeling 2007).

A clearer picture of the evolutionary history of EFL and EF-1α will depend on greater

[3] A version of this chapter has been published. Gile GH, Faktorová D, Castlejohn CA, Burger G, Lang BF, Farmer MA, Lukeš J, Keeling PJ. 2009. Distribution and phylogeny of EFL and EF-1α in Euglenozoa suggest ancestral co-occurrence followed by differential loss. PLoS ONE 4:e5162.

sampling, both on a broad scale to determine their distribution in eukaryotes as a whole and on a finer taxonomic scale in lineages where both proteins are found to gain insight into the processes behind this distribution. As part of an ongoing effort to address both levels of sampling, we have undertaken an EST- and PCR-based survey to determine the distribution of EFL and EF-1α in a previously under-sampled group, the Euglenozoa. The Euglenozoa are a phylum of protists with diverse habitats and lifestyles belonging to the somewhat contentious supergroup Excavata (Simpson 2003; Yoon et al. 2008) and comprising three major lineages: Euglenida, Kinetoplastea, and Diplonemida. There are approximately 1000 described species of euglenids, including the well-known *Euglena gracilis*, a photoautotrophic freshwater protist, and other non-photosynthetic bacteriovores, eukaryovores, and osmotrophs (Leander, Esson, and Breglia 2007). Kinetoplastids, which include human parasites of the genera *Trypanosoma* and *Leishmania*, are characterized by complex masses of DNA, known as kinetoplasts, found in their mitochondria (Riou and Delain 1969). There are only two described genera of diplonemids, although deep-sea environmental studies of small subunit ribosomal RNA (SSU rRNA) sequences have revealed considerable genetic diversity and two novel clades within this group (Lara et al. 2009). Within the Euglenozoa, the kinetoplastids and diplonemids are considered most likely to be sisters to the exclusion of euglenids (Maslov, Yasuhira, and Simpson 1999; Simpson and Roger 2004), although they are separated by a great evolutionary distance (Makiuchi et al. 2008).

Prior to this study, EF-1α sequences were known only from *E. gracilis* and a few of the medically important *Trypanosoma* and *Leishmania* species, and EFL was not known from any member of the Euglenozoa or even the excavate supergroup to which they belong. In the present study, we have examined 24 species spanning the phylogenetic diversity of Euglenozoa for the presence of EFL and EF-1α. EFL was found in six species scattered among all three euglenozoan lineages, whereas EF-1α was found in the remaining 18 species, but not from any diplonemid. None of the species examined was found to encode both proteins. The monophyly of euglenozoan EF-1α and close evolutionary similarity between EFL from *Neobodo saliens* and *Trypanoplasma borreli*, two kinetoplastids from distinct clades (Simpson and Roger 2004; von der Heyden et al. 2004; Simpson, Stevens, and Lukeš 2006) suggest that, at least in the kinetoplastids, this pattern is due to differential loss from an ancestral state of co-occurrence. Although we cannot rule out the unlikely possibility that lateral gene transfer produced this pattern, this is the clearest phylogenetic evidence from any group to date that differential loss has contributed to the complex distribution of EFL and EF-1α.

Materials and Methods

Culture sources and nucleic acids extraction

Three diplonemid species, five euglenid species, and 16 kinetoplastid species were tested for the presence of EFL and EF-1α by PCR, RT-PCR, or by searching EST libraries. Cell isolation and nucleic acids extraction methods were described previously for the diplonemids *Diplonema ambulator* ATCC 50223 and *Diplonema papillatum* ATCC 50162 (Marande, Lukeš, and Burger 2005), and *Rhynchopus euleiides* ATCC 50226 (Roy et al. 2007a; Roy et al. 2007b), the euglenids *Entosiphon sulcatum* (Breglia, Slamovits, and Leander 2007), *Peranema trichophorum* CCAP 1260/1 B and *Petalomonas cantuscygni* CCAP 1259/1 (Roy et al. 2007a), and the kinetoplastids *Blastocrithidia culicis* ATCC 30268, *Herpetomonas muscarum* ATCC 30260, *Herpetomonas pessoai* ATCC 30252 (Podlipaev et al. 2004), *Leishmania tarentolae* strain UC (Lukeš et al. 2006), *Leptomonas bifurcata* (Yurchenko et al. 2008), *Leptomonas costaricensis* (Yurchenko, Lukeš, Jirku et al. 2006), *Leptomonas podlipaevi* (Yurchenko, Lukeš, Xu, and Maslov 2006), *Neobodo saliens* (syn. *Bodo saliens*) ATCC 50358 (Atkins, Teske, and Anderson 2000), *Perkinsiella amoebae,* along with its host *Neoparamoeba branchiphila* strain AMOP1 (Dyková et al. 2003), *Trypanoplasma borreli* strain Tt-JH (Lukeš et al. 1994), *Trypanosoma avium* (Votypka et al. 2002), and *Trypanosoma brucei equiperdum* strain STIB818 (Lai et al. 2008). The remaining four kinetoplastid species were ordered from culture collections: *Rhynchobodo sp.* ATCC 50359, *Dimastigella trypaniformis* ATCC 50263, *Bodo saltans* CCAP 1907/2, and *Rhynchomonas nasuta* strain AZ-4 ATCC 50292. Total RNA was extracted from *Rhynchomonas nasuta* using the RNeasy Plant Mini Kit (Qiagen), and from *Trypanoplasma borreli* using Trizol reagent (Invitrogen). Genomic DNA was extracted from *Rhynchobodo sp*., *B. saltans*, and *D. trypaniformis* using the DNeasy Plant Mini Kit (Qiagen).

EST identification and assembly

EST libraries were generated as described (Rodríguez-Ezpeleta et al. 2009). EFL sequences from *D. ambulator*, *D. papillatum*, and *R. euleiides* and EF-1α sequences from three euglenids, *Astasia longa*, *Euglena gracilis,* and *P. trichophorum*, and seven non-euglenozoan excavates, *Histiona aroides*, *Jakoba bahamiensis*, *Jakoba libera*, *Malawimonas californiana*, *Reclinomonas americana*, *Seculamonas ecuadoriensis*, and *Stachyamoeba lipophora* were identified by tBLASTn search in the taxonomically broad EST database (TBestDB, http://amoebidia.bcm.umontreal.ca/pepdb/searches/login.php). Contigs of several ESTs were assembled using Sequencher 4.5 (GeneCodes) and examined for quality before export and

conceptual translation of consensus sequences.

Primer sets and sequencing

All non-EST sequences generated in this study were amplified from genomic DNA except for *R. nasuta*, which was amplified from cDNA. EF-1α sequences were amplified using nested degenerate primer pairs EF1a F1 and EF1a R1 followed by EF+ F2 and EF1a R2, except for sequences from *B. culicis*, *H. muscarum*, and *T. brucei equiperdum* which were amplified using EF1a F1 and EF1a Rc, and *B. saltans*, *D. trypaniformis*, *Rhynchobodo sp*., and *R. nasuta*, which were amplified using the degenerate primers EUG EF1a 1F and EUG EF1a 1R or 2R (Table 4.1). EFL from *N. saliens* was amplified using nested degenerate primer pairs EFL F1 and EFL R1 followed by EF+ F2 and EFL R2. EFL from *T. borreli* was amplified from genomic DNA with primers EFL F1 and EFL Rc, and subsequently confirmed by RT-PCR from total RNA using primers specific to the spliced leader RNA sequence and EFL sequence (data not shown). All templates were tested for both EFL and EF-1α, and none were found to encode both proteins. PCR products from *E. sulcatum, H. pessoai, L. tarentolae, P. amoebae, P. cantuscygni, R. nasuta*, and *T. avium* were TOPO-TA cloned into pCR 2.1 vector (Invitrogen) and sequenced on both strands. All other PCR products were sequenced directly on both strands. New sequences obtained in this study (Table 4.2) were deposited in GenBank under accession numbers FJ807237-FJ807268.

Phylogenetic analysis

New and previously published EFL and EF-1α sequences were translated and aligned using MAFFT (Katoh et al. 2002) and edited in MacClade 4.08 (Maddison and Maddison 2003) to final matrix sizes of 43 taxa and 478 characters for EFL and 51 taxa and 428 characters for EF-1α. In addition to these datasets, the EF-1α phylogeny was inferred with the anomalous, long-branch sequence from the heterolobosean *Acrasis rosea* (GenBank accession AAG48934) included. EFL phylogenies were also inferred from an alignment with the seven longest branches excluded: *Ditylum brightwellii, Thalassiosira pseudonana, Reticulomyxa filosa, Planoglabratella opercularis, Goniomonas amphinema,* and cytosolic sequences from *Bigelowiella natans* and *Gymnochlora stellata* (data not shown).

Phylogenetic trees were inferred using maximum likelihood (ML) and Bayesian methods. ProtTest 1.4 (Abascal, Zardoya, and Posada 2005) ranked RtREV the best amino acid substitution model for both proteins, but at present the LG model (Le and Gascuel 2008) is not

included among the models tested, so it was also used to infer trees as a point of comparison. ML trees were inferred with RAxML 7.0.4 (Stamatakis 2006) and PhyML 3.0 (Guindon and Gascuel 2003) using RtREV and LG amino acids substitution matrices, respectively (Dimmic et al. 2002; Le and Gascuel 2008), and using 4 rate categories approximated by a Γ distribution, with parameter α, amino acids frequencies, and proportion of invariable sites estimated from the data. Five hundred bootstrap replicates were performed in each program for each dataset. PhyloBayes 2.3 (Lartillot and Philippe 2006) was used to perform Bayesian analyses using 4 discrete Γ categories under the CAT mixture model which allows different different amino acid substitution models to be applied to different sites of the alignment and is therefore expected to better fit heterotachous datasets and better resist long branch attraction (Lartillot and Philippe 2004). For each dataset, two independent chains were run for 112,000 cycles, saving one tree in ten. The first 200 trees (representing 2000 cycles) were discarded as burn-in, and the remaining 11,000 trees from each chain in each dataset were used to test for convergence and compute the 50% majority rule consensus tree. Maxdiff values, which represent the largest discrepancy in frequency of any bipartition between the two chains, were 0.044 and 0.072 for EFL with long branches included and excluded, respectively, and 0.044 and 0.054 for EF-1α including and excluding the *A. rosea* sequence. When maxdiff values fall below 0.1, the two chains are considered likely to have converged on similar topologies.

Approximately Unbiased (AU) tests (Shimodaira 2002) were carried out to evaluate the likelihood of alternate EFL topologies in which euglenozoan sequences are constrained as monophyletic. Site-likelihoods for these trees were calculated by RAxML (Stamatakis 2006) using the RtREV amino acids substitution model (Dimmic et al. 2002) and four Γ rate categories with parameter α, amino acid frequencies, and the proportion of invariable sites estimated from the data. AU tests were performed using CONSEL 1.19 (Shimodaira and Hasegawa 2001).

Table 4.1. Names and sequences of primers used in this study

Name	Sequence, 5' to 3'
EFL F1	CTGTCGATCGTCATHTGYGGICAYGTHGA
EFL R1	GAACGCGATTCGGGATARNCCYTCRCA
EF+ F2	CATGTCGATGCAGGTAAGTCNACNACNACNGG
EFL R2	CTTCTTTCCTCCAGTYTCYTTNCC
EFL Rc	CTTGATRTTIAGICCIACRTTRTCNCC
EF1a F1	AACATCGTCGTGATHGGNCAYGTNGA
EF1a R1	ACGCCAACTGCTACNGTYTGNCKCAT
EF1a R2	CTGTCCAGGATGGTTCATDATDATNACYTG
EF1a Rc	CTTGATCACICCIACIGCNACNGT
EUG EF1a 1F	GGGIAARGAIAARGTICAYATNARYYT
EUG EF1a 1R	NCCNARIGGIGSRTARTCIKTRAA
EUG EF1a 2R	CCNACNGCIACITGYYGICGCATRTC

Table 4.2. New sequences obtained in this study

Species	EFL/EF-1α	Method
Diplonemids		
Diplonema ambulator ATCC 50223	EFL	ESTs
Diplonema papillatum ATCC 50162	EFL	ESTs
Rhynchopus euleiides ATCC 50226	EFL	ESTs
Kinetoplastids		
Blastocrithidia culicis ATCC 30268	EF-1α	PCR
Bodo saltans CCAP 1907/2	EF-1α	PCR
Dimastigella trypaniformis ATCC 50263	EF-1α	PCR
Herpetomonas muscarum ATCC 30260	EF-1α	PCR
Herpetomonas pessoai ATCC 30252	EF-1α	PCR
Leishmania tarentolae UC strain	EF-1α	PCR
Leptomonas bifurcata	EF-1α	PCR
Leptomonas costaricensis	EF-1α	PCR
Leptomonas podlipaevi	EF-1α	PCR
Neobodo saliens ATCC 50358	EFL	PCR
Perkinsiella amoebae	EF-1α	PCR
Rhynchobodo sp. ATCC 50359	EF-1α	PCR
Rhynchomonas nasuta strain AZ-4 ATCC 50292	EF-1α	RT-PCR
Trypanoplasma borreli strain Tt-JH	EFL	PCR
Trypanosoma avium	EF-1α	PCR
Trypanosoma brucei equiperdum strain STIB818	EF-1α	PCR
Euglenids		
Astasia longa	EF-1α	ESTs
Entosiphon sulcatum	EF-1α	PCR
Euglena gracilis	EF-1α	ESTs
Peranema trichophorum CCAP 1260/1 B	EF-1α	ESTs
Petalomonas cantuscygni CCAP 1259/1	EFL	PCR
Heterolobosean		
Stachyamoeba lipophora	EF-1α	ESTs
Jakobids		
Histiona aroides	EF-1α	ESTs
Jakoba bahamiensis	EF-1α	ESTs
Jakoba libera	EF-1α	ESTs
Reclinomonas americana	EF-1α	ESTs
Seculamonas ecuadoriensis	EF-1α	ESTs
Malawimonas		
Malawimonas californiana	EF-1α	ESTs
Amoebozoan		
Neoparamoeba branchiphila	EF-1α	PCR

Results

Distribution of EFL and EF-1α

Previously, only EF-1α sequences were known in the Euglenozoa, from *Trypanosoma* and *Leishmania* species and *E. gracilis*. We examined 24 species spanning the phylogenetic diversity of the Euglenozoa as well as seven non-euglenozoan excavate species for the presence of EFL and EF-1α by PCR or by searching EST libraries (Table 4.2). EFL was found in the diplonemids *D. ambulator*, *D. papillatum*, and *R. euleiides*, two deep-branching kinetoplastids *N. saliens* and *T. borreli*, and also in *P. cantuscygni*, a deep-branching euglenid (Breglia, Slamovits, and Leander 2007). All other species were found to encode EF-1α, including *N. branchiphila*, the amoebozoan host of *P. amoebae*, with which its DNA was co-purified. None of the species examined were found to encode both proteins, although this possibility cannot be ruled out. Where complete euglenozoan genomes exist, for the kinetoplastids *Trypanosoma brucei, Trypanosoma cruzi*, *Leishmania braziliensis*, *Leishmania infantum*, and *Leishmania major* (Berriman et al. 2005; El-Sayed et al. 2005; Ivens et al. 2005; Peacock et al. 2007) we can confirm that they each encode only EF-1α. To date there are only two documented cases of EFL and EF-1α co-occurrence: both genes were amplified by PCR in the zygomycete fungus *Basidiobolus ranarum* (James et al. 2006), and both are found in the complete genome of the diatom *Thalassiosira pseudonana*. While no expression data is available for the former, in the latter only EFL is expressed (Kamikawa, Inagaki, and Sako 2008).

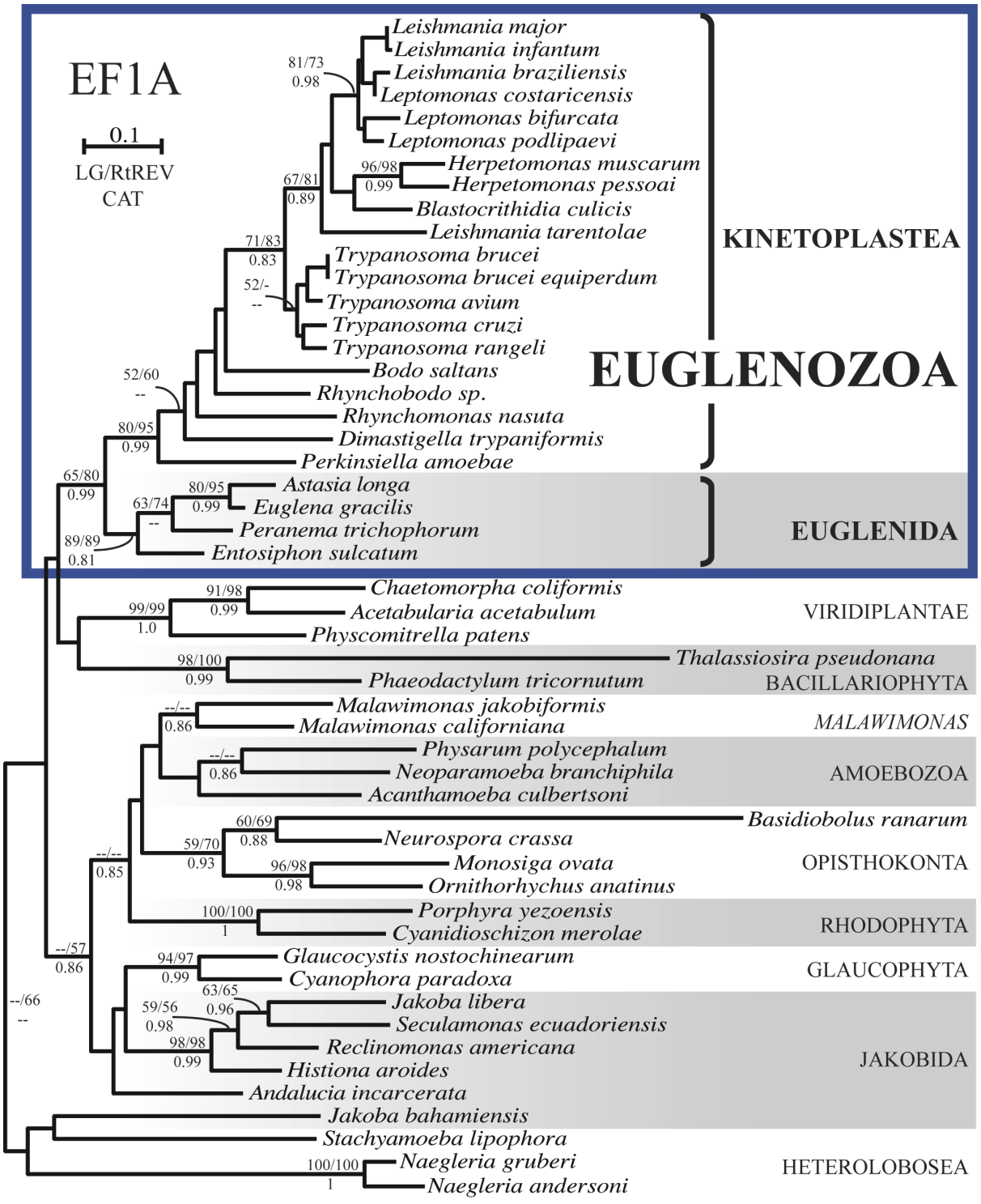
Phylogenetic analyses of EF-1α and EFL

The phylogeny of EF-1α is broadly concordant with accepted euglenozoan relationships. The monophyly of kinetoplastids, euglenids, and Euglenozoa as a whole are recovered with poor to good support depending on the method (Figure 4.1). Within the euglenids, the branching order of genera was poorly supported but consistent among methods and consistent with current hypotheses for the organismal phylogeny. The branching order within the kinetoplastids in ML trees roughly matches expectations (without much support) with the major exception that *R. nasuta* and *D. trypaniformis* did not form a clade, as they consistently group together in other published analyses (Simpson, Lukeš, and Roger 2002; Moreira, Lopez-Garcia, and Vickerman 2004; Simpson et al. 2004; von der Heyden et al. 2004; von der Heyden and Cavalier-Smith 2005). The overall prevalence of EF-1α in the Euglenozoa and its broad congruence with accepted organismal relationships suggest that EF-1α was present in the common ancestor of this group.

Preliminary EF-1α analyses were carried out with the EF-1α sequence from the heterolobosean *Acrasis rosea* (GenBank accession AAG48934) included. The position of this sequence was not resolved: rather than branching with other heteroloboseans, it formed a long branch within the *Herpetomonas* clade in ML analyses and at the base of kinetoplastids in the Bayesian analysis, and its inclusion reduced bootstrap support for trypanosomatid, kinetoplastid, and euglenozoan monophyly. Because of its uncertain placement, its disruptive effect on resolution throughout the kinetoplastid clade, and the fact that *A. rosea* is not a euglenozoan, this sequence was removed from the alignment for further analysis.

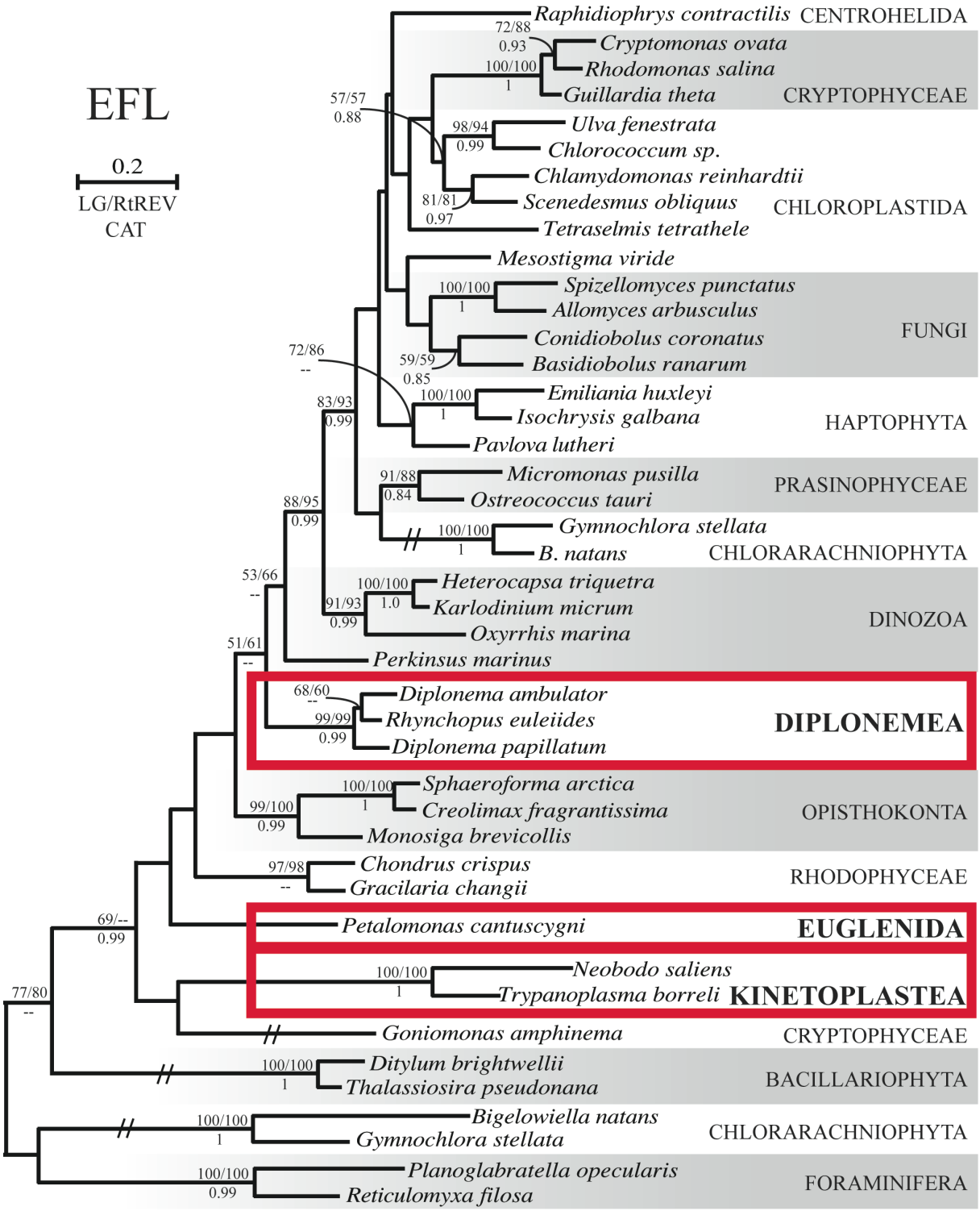
EFL phylogenies were inferred using the same models used for EF-1α. While much of the tree remains poorly resolved in all analyses, as is typical of EFL trees (Keeling and Inagaki 2004; Gile, Patron, and Keeling 2006; Noble, Rogers, and Keeling 2007; Kamikawa, Inagaki, and Sako 2008; Sakaguchi et al. 2009), three features emerge that are pertinent to the origin and evolution of EFL in the Euglenozoa (Figure 4.2). First, the three lineages of euglenozoan EFL, diplonemids, kinetoplastids, and *P. cantuscygni*, do not branch together. However, their positions are not clearly resolved, none of the nodes that separate them are supported, and the relative branching order of the three euglenozoan EFL lineages, *Goniomonas amphinema*, *Perkinsus marinus*, red algae, and a group of opisthokonts, varies greatly depending on the dataset analyzed and evolutionary model employed. Second, diplonemid EFL sequences robustly branch together in all analyses, suggesting that EFL is ancestral in this group. Third, and most importantly, the two kinetoplastid EFL sequences branch together with complete support in all analyses, providing strong evidence that EFL was present in their common ancestor as well. This is significant because *N. saliens* and *T. borreli* are members of two different subgroups in organismal phylogenies of kinetoplastids (Moreira, Le Guyader, Philippe 1999; Simpson, Lukeš, Roger 2002; Simpson and Roger 2004; Simpson et al. 2004; von der Heyden et al. 2004), which therefore places EFL at least as far back as the common ancestor of all kinetoplastids save the earliest-branching lineage that includes *P. amoebae* (Figure 4.3). Because the phylogeny of EF-1α suggests that this protein was also present in the ancestor of kinetoplastids, we infer that both genes must have co-existed through much of early kinetoplastid evolution, and it therefore appears that the complex distribution of EFL and EF-1α in the kinetoplastids is likely due to differential loss.

Figure 4.1. Phylogeny of EF-1α



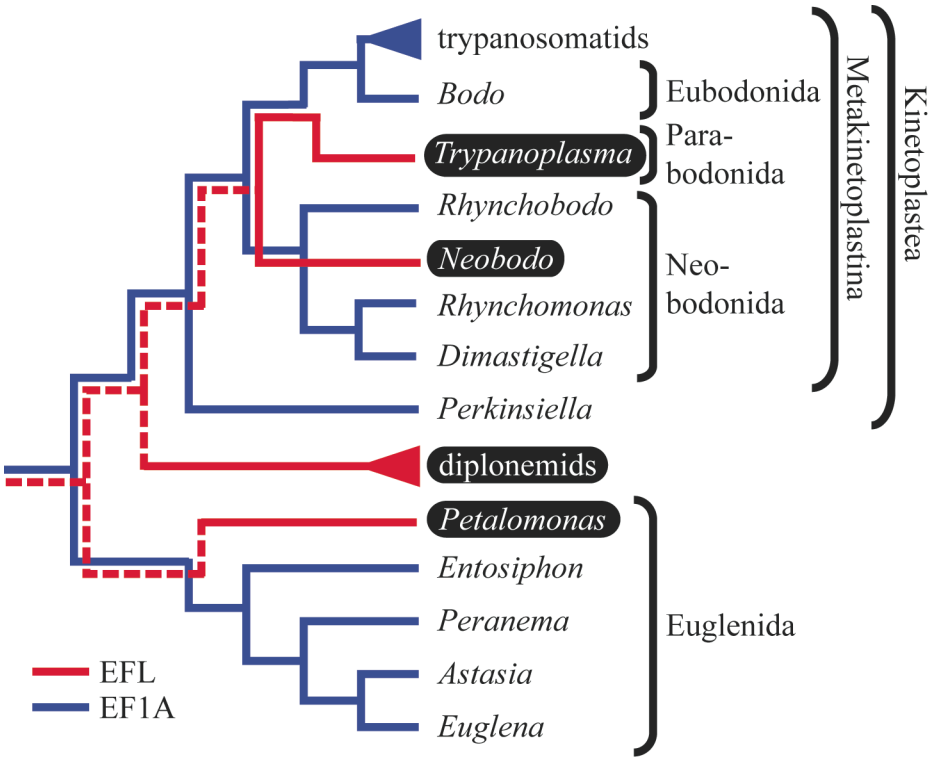
Maximum likelihood phylogeny of EF-1α including Bayesian posterior probabilities. The tree was inferred under LG, RtREV, and CAT amino acids substitution models using four Γ rate categories plus invariable sites; the LG topology, which has better support, is displayed. Bootstrap support greater than 50% and Bayesian posterior probabilities greater than 0.8 are displayed at nodes, with LG/RtREV ML bootstrap values above and CAT model posterior probability below. Euglenozoan taxa are boxed in blue.

Figure 4.2. Phylogeny of EFL



Maximum likelihood phylogeny of EFL including Bayesian posterior probabilities. The tree was inferred under LG, RtREV, and CAT amino acids substitution models using four Γ rate categories plus invariable sites; the LG topology is displayed. Bootstrap support greater than 50% and Bayesian posterior probabilities greater than 0.8 are displayed at nodes, with LG/RtREV ML bootstrap values above and CAT model posterior probability below. Branches with hatch marks are displayed at one half their actual length. Euglenozoan taxa are boxed in red.

Figure 4.3. Evolution of EFL and EF-1α in Euglenozoa



Schematic tree illustrating currently accepted phylogenetic relationships among euglenozoan taxa examined in this study. The presence of EFL (red) and EF-1α (blue) are traced along the organismal phylogeny to their origins with solid lines where there is phylogenetic evidence for their monophyly. Dotted lines hypothetically trace the presence of EFL back to the ancestor of Euglenozoa. Taxa shown in white text on black background encode EFL; all others encode EF-1α.

To test the possibility that EFL sequences from the three euglenozoan lineages are monophyletic, we carried out approximately unbiased (AU) tests to evaluate alternative topologies in which their monophyly was constrained. For each of four ML topologies, a euglenozoan clade in which kinetoplastids and diplonemids are sisters was grafted onto the positions where each of the three euglenozoan EFL lineages had individually branched in ML analyses. In tests including the entire dataset, euglenozoan EFL monophyly is not rejected at the 5% level when grafted to the diplonemid branch, but all other alternate topologies are rejected. Because significant rate heterogeneity is known in several EFL lineages, we also tested euglenozoan EFL monophyly using a second dataset where the seven longest-branching sequences were removed. A monophyletic Euglenozoa was once again grafted to the positions where the euglenid, diplonemid, and kinetoplastid lineages were placed in ML trees inferred from this dataset, and in this case AU tests fail to reject euglenozoan EFL monophyly in any position (Table 4.3). Overall, the phylogeny of EFL provides strong evidence for differential

loss of EFL and EF-1α in the kinetoplastid lineage, and the general failure of AU tests to reject euglenozoan EFL monophyly leaves open the possibility that differential loss after a single introduction of EFL may explain the entire distribution of EFL and EF-1α in Euglenozoa as a whole.

Table 4.3. Approximately Unbiased (AU) test p-values.

	Dataset	
Topology, position of Euglenozoa	EFL full	EFL short
LG, polyphyletic	0.454	0.444
LG, on kinetoplastids branch	0.001	0.164
LG, on *P. cantuscygni* branch	0.005	0.163
LG, on diplonemids branch	0.090	0.164
RtREV, polyphyletic	0.704	0.776
RtREV, on kinetoplastids branch	0.002	0.170
RtREV, on *P. cantuscygni* branch	0.000	0.170
RtREV, on diplonemids branch	0.039	0.167

Approximately Unbiased (AU) test p-values for topologies in which Euglenozoa are constrained as monophyletic with kinetoplastids and diplonemids as sister groups. Two datasets and two amino acid substitution models (LG and RtREV) were tested.

Discussion

Here we report the presence of EFL in the Euglenozoa, which occurs in a complex distribution that is not consistent with the known phylogenetic relationships of the organisms. Neither of these findings is unique to the Euglenozoa (Gile, Patron, and Keeling 2006; Noble, Rogers, and Keeling 2007; Sakaguchi et al. 2009); however, we also show that at least part of this complexity is consistent with differential loss of EFL and EF-1α from an ancestral state of co-occurrence rather than from multiple lateral transfer events. Three lines of evidence collectively support this interpretation. First, the monophyly of kinetoplastid EF-1α implies that this protein is ancestral in the kinetoplastids. Second, EFL sequences from *N. saliens* and *T. borreli* are closely related, implying that EFL was also present in their common ancestor. Third, analyses of other data consistently show that *T. borreli* and *N. saliens* are not sister taxa; rather, they belong to separate, consistently well-supported clades that have been named Parabodonida and Neobodonida, respectively (Moreira, Le Guyader, and Philippe 1999; Simpson, Lukeš, and Roger 2002; Simpson and Roger 2004; Simpson et al. 2004; von der Heyden et al. 2004). Therefore *N. saliens* is more closely related to other neobodonids such as *R. nasuta* and *D.*

trypaniformis, which, as we have demonstrated here, encode EF-1α. Although the branching order of kinetoplastid clades is somewhat variable, with notable differences in topology between SSU rRNA and heat shock protein phylogenies, neobodonids and parabodonids are always monophyletic groups, and are never sister to one another. The better-supported protein phylogenies favour a topology in which neo- and parabodonids branch as the deepest and next-deepest branches of the Metakinetoplastina (i.e., all kinetoplastids except the clade to which *P. amoebae* belongs), and their common ancestor is therefore also the ancestor of eubodonids and trypanosomatids (Figure 4.3). Taken together, these lines of evidence suggest that there was a period of co-occurrence of EFL and EF-1α in the stem lineage of modern kinetoplastids, and the complex distribution of these proteins is due to differential loss or continued co-existence, which we cannot rule out until complete genome sequences of these organisms are available. To explain this distribution through lateral gene transfer, one would need to invoke two independent transfers, coincidentally from closely related unidentified sources, or a transfer to either *N. saliens* or *T. borreli* followed by a transfer between the two, neither of which seems especially likely. Given the alternatives outlined above, we consider the scenario of co-occurrence followed by differential loss to be the most parsimonious.

If differential loss after a period of co-occurrence can explain the complex distribution of EFL and EF-1α within the Metakinetoplastina, how well can it explain the complex distribution in the Euglenozoa as a whole? Here, there is no strong evidence for either lateral gene transfer or differential loss. The distribution and phylogeny of EF-1α indicate that this protein is ancestral in the Euglenozoa, and the distribution of EFL in deep-branching members of all three euglenozoan lineages suggests that this protein may also be ancestral. The phylogeny of EFL, however, is too poorly supported to make strong conclusions in either direction. Taken at face value, three separate clades of euglenozoan EFL imply three independent acquisitions, but without a clear identification of donor lineages for any of these putative transfers, this does not constitute evidence for lateral gene transfer. Furthermore, the separation of these lineages is weak, and several of the EFL topologies with a monophyletic Euglenozoa cannot be rejected. Given the evidence for differential loss in the kinetoplastids and the occurrence of EFL in all three euglenozoan lineages, we surmise that EFL's complex distribution in the Euglenozoa as a whole may be due entirely to differential loss.

Where did the euglenozoan EFL ultimately originate? The closest relatives of Euglenozoa are the Heterolobosea and Jakobida, with Heterolobosea being the most likely sister group (Baldauf et al. 2000; Simpson 2003; Simpson, Inagaki, and Roger 2006; Rodríguez-

Ezpeleta et al. 2007). Only EF-1α sequences have been found in heterolobosean and jakobid taxa to date, including analyses of several EST projects described here, so at present there is no direct evidence for EFL in any excavate prior to the ancestor of Euglenozoa, although given the rapidity with which EFL has been discovered in diverse eukaryotes it would not be surprising if more excavate lineages are shown to possess it. Perhaps the anomalous EF-1α sequence of *A. rosea* is a hint that this species deserves further study. For both species in which EFL and EF-1α are currently known to co-occur, *T. pseudonana* and *B. ranarum*, EF-1α forms an unusually long branch (Figure 4.1), similar in length to the EF-1α sequence of *A. rosea* (not shown).

The Euglenozoa are very isolated in the tree of eukaryotes from other lineages currently known to encode EFL, and therefore the origin of EFL in the Euglenozoa is more simply explained by lateral gene transfer, but the demonstration here that differential loss plays a role in the distribution of EFL and EF-1α needs to be considered more carefully at all levels of the tree. There is evidence that this might have played a part in the distribution of EFL in green algae, where there is support for the retention of the ancestral EF-1α but no support for a common origin of EFL genes in distantly related lineages (Noble, Rogers, and Keeling 2007). Conversely, an analysis of EFL in diatoms has suggested a direct role for lateral transfer in that lineage (Kamikawa, Inagaki, and Sako 2008). The biggest question that remains is how lateral transfer and/or differential loss might have contributed to the distribution throughout eukaryotes as a whole. Without a robustly resolved phylogeny of EFL, which seems unlikely to emerge, we must remain open to the possibility that EFL's complex distribution is attributable to rampant lateral gene transfer; however, this study provides the strongest evidence to date that differential loss has also contributed to EFL's intriguing distribution.

Despite the considerable sequence divergence between EFL and EF-1α (typically 40-45% sequence identity), EFL is considered likely to perform the same canonical function as EF-1α, namely cleaving GTP to deposit aminoacyl-tRNAs in the A site of the ribosome. This inference is based on two main observations. First, the binding sites for aa-tRNAs, GTP, and the nucleotide exchange factor EF-1β are conserved between EF-1α and EFL: evolutionary rate shifts and divergence without rate shifts are confined primarily to non-binding sites. Second, EF-1α has an essential function in translation, and as the protein with the closest similarity to EF-1α in EF-1α-lacking genomes, EFL is the most likely candidate for executing this function (Keeling and Inagaki 2004). This leads to the question, why would one protein or the other be preferentially retained in different lineages? As yet there is very little data to address this

question, but part of the answer may lie among the many additional cellular processes in which EF-1α has been implicated, such as actin bundling (Gross and Kinzy 2005) and ubiquitin-dependent protein degradation (Gonen et al. 1994) for which EFL might not share EF-1α's binding sites. Minor functional differences may also help to explain our conclusion that these two proteins are better able to co-exist than their present distribution suggests. For the majority of duplicate gene pairs, from which we can draw a loose analogy to EFL and EF-1α, one copy tends to be lost quite rapidly unless it undergoes sub- or neofunctionalization (Lynch and Conery 2000). Much work is needed to determine whether such functional differences exist, and if so, whether there may be adaptive significance to the complex distribution of EFL and EF-1α.

Literature Cited

Abascal F, Zardoya R, Posada D. 2005. ProtTest: Selection of best-fit models of protein evolution. Bioinformatics 21:2104-2105.

Andersen GR, Valente L, Pedersen L, Kinzy TG, Nyborg J. 2001. Crystal structures of nucleotide exchange intermediates in the eEF1A-eEF1Balpha complex. Nat. Struct. Biol. 8:531-534.

Atkins MS, Teske AP, Anderson OR. 2000. A survey of flagellate diversity at four deep-sea hydrothermal vents in the eastern Pacific ocean using structural and molecular approaches. J. Eukaryot. Microbiol. 47:400-411.

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290:972-977.

Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B et al. (102 co-authors). 2005. The genome of the African trypanosome *Trypanosoma brucei*. Science 309:416-422.

Breglia SA, Slamovits CH, Leander BS. 2007. Phylogeny of phagotrophic euglenids (Euglenozoa) as inferred from hsp90 gene sequences. J. Eukaryot. Microbiol. 54:86-92.

Dimmic MW, Rest JS, Mindell DP, Goldstein RA. 2002. RtREV: An amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J. Mol. Evol. 55:65-73.

Dyková I, Fiala I, Lom J, Lukeš J. 2003. *Perkinsiella amoebae*-like endosymbionts of *Neoparamoeba* spp., relatives of the kinetoplastid *Ichthyobodo*. Eur. J. Protistol. 39:37-52.

El-Sayed NM, Myler PJ, Bartholomeu DC, et al. (82 co-authors). 2005. The genome sequence of *Trypanosoma cruzi*, etiologic agent of chagas disease. Science 309:409-415.

Gile GH, Patron NJ, Keeling PJ. 2006. EFL GTPase in cryptomonads and the distribution of EFL and EF-1alpha in chromalveolates. Protist 157:435-444.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696-704.

Ivens AC, Peacock CS, Worthey EA, et al. (101 co-authors). 2005. The genome of the kinetoplastid parasite, *Leishmania major*. Science 309:436-442.

James TY, Kauff F, Schoch CL, et al. (70 co-authors). 2006. Reconstructing the early evolution of fungi using a six-gene phylogeny. Nature 443:818-822.

Kamikawa R, Inagaki Y, Sako Y. 2008. Direct phylogenetic evidence for lateral transfer of elongation factor-like gene. Proc. Natl. Acad. Sci. USA 105:6965-6969.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: A novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Res. 30:3059-3066.

Keeling PJ, Inagaki Y. 2004. A class of eukaryotic GTPase with a punctate distribution suggesting multiple functional replacements of translation elongation factor 1alpha. Proc. Natl. Acad. Sci. USA 101:15380-15385.

Lai DH, Hashimi H, Lun ZR, Ayala FJ, Lukeš J. 2008. Adaptations of *Trypanosoma brucei* to gradual loss of kinetoplast DNA: *Trypanosoma equiperdum* and *Trypanosoma evansi* are petite mutants of *T. brucei*. Proc. Natl. Acad. Sci. USA 105:1999-2004.

Lara E, Moreira D, Vereshchaka A, Lopez-Garcia P. 2009. Pan-oceanic distribution of new highly diverse clades of deep-sea diplonemids. Environ. Microbiol. 11:47-55.

Lartillot N, Philippe H. 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55:195-207.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095-1109.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307-1320.

Leander BS, Esson HJ, Breglia SA. 2007. Macroevolution of complex cytoskeletal systems in euglenids. Bioessays 29:987-1000.

Lukeš J, Arts GJ, van den Burg J, de Haan A, Opperdoes F, Sloof P, Benne R. 1994. Novel pattern of editing regions in mitochondrial transcripts of the cryptobiid *Trypanoplasma borreli*. EMBO J. 13:5086-5098.

Lukeš J, Paris Z, Regmi S, Breitling R, Mureev S, Kushnir S, Pyatkov K, Jirků M, Alexandrov KA. 2006. Translational initiation in *Leishmania tarentolae* and *Phytomonas serpens* (Kinetoplastida) is strongly influenced by pre-ATG triplet and its 5` sequence context. Mol. Biochem. Parasitol. 148:125-132.

Maddison DR, Maddison WP. 2003. MacClade 4: Analysis of phylogeny and character evolution. 4.08. [computer program]

Makiuchi T, Annoura T, Hashimoto T, Murata E, Aoki T, Nara T. 2008. Evolutionary analysis of synteny and gene fusion for pyrimidine biosynthetic enzymes in Euglenozoa: An extraordinary gap between kinetoplastids and diplonemids. Protist 159:459-470.

Marande W, Lukeš J, Burger G. 2005. Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. Eukaryot. Cell. 4:1137-1146.

Marshall WL, Celio G, McLaughlin DJ, Berbee ML. 2008. Multiple isolations of a culturable, motile ichthyosporean (Mesomycetozoa, Opisthokonta), *Creolimax fragrantissima* n. gen., n. sp., from marine invertebrate digestive tracts. Protist 159:415-433.

Maslov DA, Yasuhira S, Simpson L. 1999. Phylogenetic affinities of *Diplonema* within the Euglenozoa as inferred from the SSU rRNA gene and partial COI protein sequences.

Protist 150:33-42.

Moreira D, Lopez-Garcia P, Vickerman K. 2004. An updated view of kinetoplastid phylogeny using environmental sequences and a closer outgroup: Proposal for a new classification of the class Kinetoplastea. Int. J. Syst. Evol. Microbiol. 54:1861-1875.

Moreira D, Le Guyader H, Philippe H. 1999. Unusually high evolutionary rate of the elongation factor 1 alpha genes from the Ciliophora and its impact on the phylogeny of eukaryotes. Mol. Biol. Evol. 16:234-245.

Negrutskii BS, El'skaya AV. 1998. Eukaryotic translation elongation factor 1 alpha: Structure, expression, functions, and possible role in aminoacyl-tRNA channeling. Prog. Nucleic Acid Res. Mol. Biol. 60:47-78.

Noble GP, Rogers MB, Keeling PJ. 2007. Complex distribution of EFL and EF-1alpha proteins in the green algal lineage. BMC Evol. Biol. 7:82.

Peacock CS, Seeger K, Harris D, et al. (41 co-authors). 2007. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. Nat. Genet. 39:839-847.

Podlipaev SA, Sturm NR, Fiala I, Fernandes O, Westenberger SJ, Dollet M, Campbell DA, Lukeš J. 2004. Diversity of insect trypanosomatids assessed from the spliced leader RNA and 5S rRNA genes and intergenic regions. J. Eukaryot. Microbiol. 51:283-290.

Riou G, Delain E. 1969. Electron microscopy of the circular kinetoplastic DNA from *Trypanosoma cruzi*: Occurrence of catenated forms. Proc. Natl. Acad. Sci. USA 62:210-217.

Rodríguez-Ezpeleta N, Teijeiro S, Forget L, Burger G, Lang BF. 2009. Construction of cDNA libraries: Focus on protists and fungi. In: John Parkinson, editor. Expressed Sequence Tags (ESTs): Generation and Analysis. Humana Press, Totowa, NJ, USA.

Rodríguez-Ezpeleta N, Brinkmann H, Burger G, Roger AJ, Gray MW, Philippe H, Lang BF. 2007. Toward resolving the eukaryotic tree: The phylogenetic positions of jakobids and cercozoans. Curr. Biol. 17:1420-1425.

Roy J, Faktorová D, Lukeš J, Burger G. 2007a. Unusual mitochondrial genome structures throughout the Euglenozoa. Protist 158:385-396.

Roy J, Faktorová D, Benada O, Lukeš J, Burger G. 2007b. Description of *Rhynchopus euleeides* n. sp. (Diplonemea), a free-living marine euglenozoan. J. Eukaryot. Microbiol. 54:137-145.

Ruiz-Trillo I, Lane CE, Archibald JM, Roger AJ. 2006. Insights into the evolutionary origin and genome architecture of the unicellular opisthokonts *Capsaspora owczarzaki* and *Sphaeroforma arctica*. J. Eukaryot. Microbiol. 53:379-384.

Sakaguchi M, Takishita K, Matsumoto T, Hashimoto T, Inagaki Y. 2009. Tracing back EFL gene evolution in the cryptomonads-haptophytes assemblage: Separate origins of EFL genes in

haptophytes, photosynthetic cryptomonads, and goniomonads. Gene, in press.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492-508.

Shimodaira H, Hasegawa M. 2001. CONSEL: For assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246-1247.

Simpson AG. 2003. Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota). Int. J. Syst. Evol. Microbiol. 53:1759-1777.

Simpson AG, Roger AJ. 2004. Protein phylogenies robustly resolve the deep-level relationships within Euglenozoa. Mol. Phylogenet. Evol. 30:201-212.

Simpson AG, Stevens JR, Lukeš J. 2006. The evolution and diversity of kinetoplastid flagellates. Trends Parasitol. 22:168-174.

Simpson AG, Inagaki Y, Roger AJ. 2006. Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of "primitive" eukaryotes. Mol. Biol. Evol. 23:615-625.

Simpson AG, Lukeš J, Roger AJ. 2002. The evolutionary history of kinetoplastids and their kinetoplasts. Mol. Biol. Evol. 19:2071-2083.

Simpson AG, Gill EE, Callahan HA, Litaker RW, Roger AJ. 2004. Early evolution within kinetoplastids (Euglenozoa), and the late emergence of trypanosomatids. Protist 155:407-422.

Stamatakis A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-2690.

von der Heyden S, Cavalier-Smith T. 2005. Culturing and environmental DNA sequencing uncover hidden kinetoplastid biodiversity and a major marine clade within ancestrally freshwater *Neobodo designis*. Int. J. Syst. Evol. Microbiol. 55:2605-2621.

von der Heyden S, Chao EE, Vickerman K, Cavalier-Smith T. 2004. Ribosomal RNA phylogeny of bodonid and diplonemid flagellates and the evolution of Euglenozoa. J. Eukaryot. Microbiol. 51:402-416.

Votypka J, Obornik M, Volf P, Svobodova M, Lukeš J. 2002. *Trypanosoma avium* of raptors (Falconiformes): Phylogeny and identification of vectors. Parasitology 125:253-263.

Yoon HS, Grant J, Tekle YI, Wu M, Chaon BC, Cole JC, Logsdon JM,Jr, Patterson DJ, Bhattacharya D, Katz LA. 2008. Broadly sampled multigene trees of eukaryotes. BMC Evol. Biol. 8:14.

Yurchenko VY, Lukeš J, Jirku M, Zeledon R, Maslov DA. 2006. *Leptomonas costaricensis* sp. n. (Kinetoplastea: Trypanosomatidae), a member of the novel phylogenetic group of insect trypanosomatids closely related to the genus *Leishmania*. Parasitology 133:537-546.

Yurchenko V, Lukeš J, Xu X, Maslov DA. 2006. An integrated morphological and molecular approach to a new species description in the Trypanosomatidae: The case of *Leptomonas podlipaevi* n. sp., a parasite of *Boisea rubrolineata* (Hemiptera: Rhopalidae). J. Eukaryot. Microbiol. 53:103-111.

Yurchenko VY, Lukeš J, Tesarova M, Jirku M, Maslov DA. 2008. Morphological discordance of the new trypanosomatid species phylogenetically associated with the genus *Crithidia*. Protist 159:99-114.

CHAPTER 5: NUCLEUS-ENCODED PERIPLASTID-TARGETED EFL IN CHLORARACHNIOPHYTES[4]

Introduction

One of the most important steps in the transition from endosymbiont to organelle is the establishment of a protein-targeting system. As endosymbionts integrate, many genes are transferred to the host nucleus, and those whose products are required in the plastid acquire targeting sequences that are recognized by a specific import apparatus. The targeting system of primary plastids such as those of green algae and plants has been relatively well studied, and most proteins are recognized via an amino-terminal extension known as a transit peptide. Transit peptides tend to share an overall positive charge due to a marked depletion in acidic residues and a modest enrichment in basic residues. Further generalizations can be made for specific subsets of photosynthetic eukaryotes; for example, transit peptides of land plants and to a lesser extent green algae are enriched in serine and threonine, but in general rules for one lineage may not apply to others.

Since the origin of plastids by primary endosymbiosis, plastids have subsequently moved between eukaryotic lineages by secondary and tertiary endosymbioses. Whereas primary plastids are bound by two membranes and located in the host cytosol, secondary and tertiary plastids are bounded by additional membranes and are located within the endomembrane system of the host. As a result, plastid-targeted proteins in secondary algae use a bipartite leader consisting of a signal peptide followed by a transit peptide (McFadden 1999; Patron and Waller 2007). The signal peptide allows proteins to cross the outermost membrane, which is part of the host endomembrane system (and in some taxa is detectably continuous with the endoplasmic reticulum), and the transit peptide is thought to mediate transfer across the two innermost membranes, which correspond to the two membranes of the primary plastid.

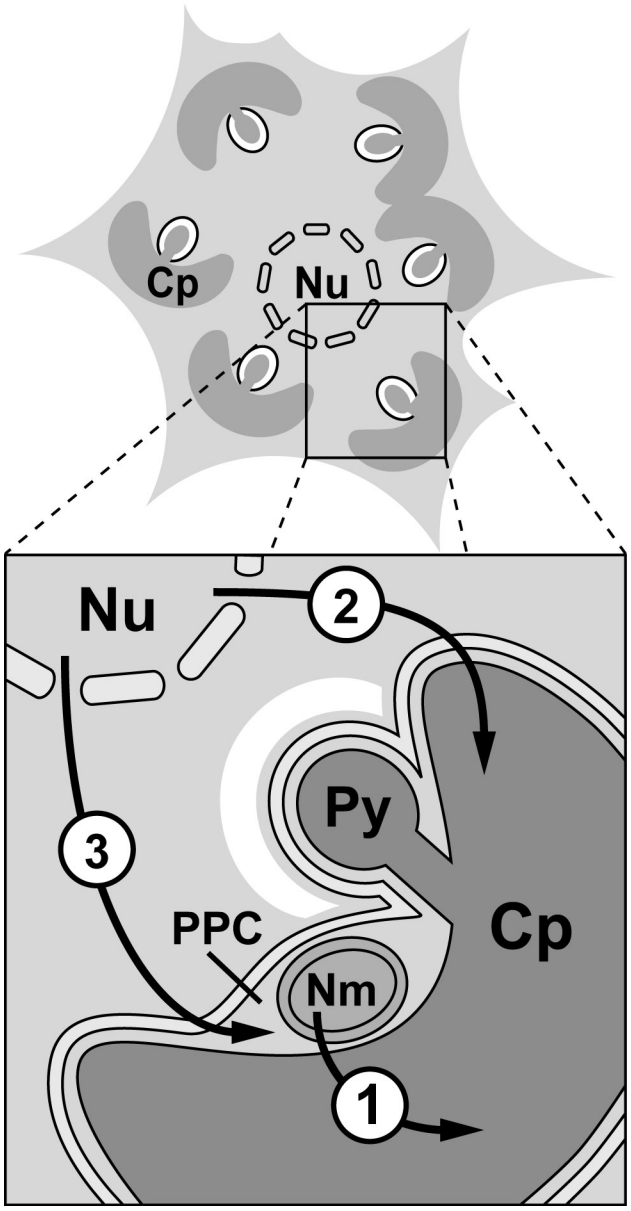
However, most secondary plastids (euglenids and dinoflagellates being the exceptions) have an additional membrane between the outer membrane and the two primary plastid membranes, which is thought to be derived from the plasma membrane of the endosymbiotic primary alga. How proteins cross this second membrane is the most poorly understood step of the system. This is partly because there is no obvious leader domain that mediates passage through this membrane, and partly because very few proteins are targeted across just the outer and second membranes. Such proteins might allow the requirements for each step of the process

[4] A version of this chapter has been published. Gile GH, Keeling PJ. 2008. Nucleus-encoded periplastid-targeted EFL in chlorarachniophytes. Mol. Biol. Evol. 25:1967-1977.

to be dissected, but in most algal lineages few proteins would be expected to function between the two pairs of plastid membranes. The two exceptions to this are chlorarachniophytes and cryptomonads. In all other secondary algae, the nucleus of the eukaryotic endosymbiont has been completely lost, but in these two lineages relict nuclei called nucleomorphs have been retained in the reduced eukaryotic cytoplasm known as the periplastid compartment (PPC) that lies between the inner and outer pairs of plastid membranes. The majority of nucleomorph genes in both lineages specify housekeeping functions, but in both cases many genes deemed to be essential for nucleomorph maintenance and expression are missing (Douglas et al. 2001; Gilson et al. 2006). These genes are believed to have been transferred to the host nucleus, from which their products would have to be targeted to the PPC. Thus, a system dominated in most secondary algae by a single plastid-targeting route (host nucleus to plastid) requires three different routes in chlorarachniophytes and cryptomonads: nucleus-encoded proteins are targeted to two distinct compartments (PPC and plastid) and nucleomorph-encoded proteins are targeted to the plastid (Figure 5.1). Each of these routes has to be specified by targeting information and the targeting information of PPC-targeted proteins has to be distinguishable from that of plastid-targeted proteins. By crossing only the first and second plastid membranes, the PPC-targeted proteins offer an opportunity to examine how proteins cross the second membrane that is rare or impossible in other lineages, although there is evidence that a few proteins are still targeted to this compartment in diatoms and apicomplexans (Sommer et al. 2007).

No PPC-targeted proteins have been identified in chlorarachniophytes, and it remains to be seen by what mechanism they cross the second membrane and by what mechanism the cells distinguish between plastid- and PPC-targeted proteins. This is significant because the outermost membrane is crossed using the endomembrane/secretion system of the host and the two innermost membranes are crossed using the plastid import system of the primary endosymbiont, so the crossing of the second membrane is the only part of the system that may have evolved completely independently in chlorarachniophytes and cryptomonads. Here we describe transit peptides from EFL, the first putatively PPC-targeted protein to be identified in chlorarachniophytes, and compare them to those of nucleus- and nucleomorph-encoded plastid-targeted proteins from two distantly related chlorarachniophytes.

Figure 5.1. Schematic view of plastid targeting in *Gymnochlora stellata*.



Arrows represent routes of protein targeting: 1) nucleomorph-encoded (Nm) proteins are targeted to the plastid (Cp); 2) nucleus-encoded (Nu) proteins are targeted to the plastid; and 3) nucleus-encoded proteins are targeted to the reduced eukaryotic cytosol of the plastid, the periplastid compartment (PPC). Py denotes pyrenoid.

EFL is a GTPase that is related to the translation elongation factor EF-1α and is thought to have taken over its essential role in translation in several eukaryotic lineages (Keeling and Inagaki 2004; Gile, Patron, and Keeling 2006; James et al. 2006; Ruiz-Trillo et al. 2006; Noble, Rogers, and Keeling 2007). The chlorarachniophyte *Bigelowiella natans* was previously shown to possess a nucleus-encoded EFL presumed to be of host ancestry (Keeling and Inagaki 2004), and a survey of EFL and EF-1α in the green algae showed that the ancestor of the

chlorarachniophytes' endosymbiont most likely encoded EFL as well (Noble, Rogers, and Keeling 2007). The *B. natans* nucleomorph genome encodes neither EFL nor EF-1α (Gilson et al. 2006), which suggests that an endosymbiont-derived EFL has transferred to the host genome and its product is PPC-targeted. Accordingly, we identified two evolutionarily distinct clades of EFL in chlorarachniophytes, and we here present evidence that one is a host protein and the other is targeted to the PPC. The putative PPC-targeted proteins include substantial amino-terminal bipartite leaders consisting of a signal peptide and a sequence with similarities to chlorarachniophyte transit peptides. Using these characteristics, we sought other potentially PPC-targeted proteins and identified a eukaryotic translation initiation factor with a similar leader that is missing from the *B. natans* nucleomorph genome. Western blotting of both types of EFL shows that the mature PPC-targeted protein is similar in size to the host-derived proteins, suggesting post-translational removal of its long leader. Immunolocalization of both proteins in *B. natans* confirmed that the leaderless EFL is cytosolic and showed a distinct localization pattern for the PPC-targeted protein. However, this pattern could not be distinguished from a plastid localization. Altogether, the evidence suggests that chlorarachniophytes have independently adopted the same overall strategy for PPC targeting as cryptomonads, namely the use of a bipartite targeting peptide similar to plastid-targeting peptides.

In order to characterize these PPC-targeting peptides, we compared them to nucleus- and nucleomorph-encoded plastid-targeted proteins from *B. natans* (Rogers et al. 2004; Gilson et al. 2006), and to corresponding classes of proteins we identified in an EST survey of the deep-branching chlorarachniophyte, *Gymnochlora stellata*, thereby including leader sequences from both eukaryotic genomes and across chlorarachniophyte diversity. In both species, plastid- and PPC-targeted proteins encoded in the nucleus share many characteristics while nucleomorph-encoded transit peptides differ, likely reflecting the high AT content of nucleomorph genomes.

Materials and Methods

Strains and culture conditions

Seven chlorarachniophyte species were examined in this study. *Chlorarachnion reptans* (strain NEPCC 449) and *Lotharella globosa* (strain NEPCC 811) were obtained from the Canadian Centre for the Culture of Microorganisms at UBC (CCCM). *Bigelowiella natans* (CCMP 621), *Lotharella amoeboformis* (CCMP 2058), *Lotharella vacuolata* (CCMP 240), *Gymnochlora stellata* (CCMP 2057), and unidentified chlorarachniophyte strain CCMP 1408 were obtained from the Provasoli-Guillard National Center for Culture of Marine Phytoplankton

(CCMP). All cultures were maintained in f/2 – Si or K medium at 22ºC on a 12/12-hour light/dark cycle.

DNA/RNA extraction, amplification, and sequencing

Total RNA was extracted from chlorarachniophyte cell pellets using Trizol reagent (Invitrogen, Carlsbad, CA). Genomic DNA was extracted from *B. natans*, *C. reptans*, *G. stellata*, and *L. vacuolata* using the DNeasy Plant Mini Kit (Qiagen, Mississauga, ON, Canada). Full length chlorarachniophyte EFL cDNA sequences were obtained by the following reactions: 1) 3' RACE using nested degenerate forward primers 5'-GTCGARATGCAYCAY-3' (outer) and 5'- CCGGGCGAYAAYGTNGG-3' (inner); 2) RT-PCR using gene-specific reverse primers designed from the 3' RACE products and different combinations of nested degenerate forward primers 5'- CTGTCGATCGTCATHTGYGGN-3', 5'-TCGTTCGCGTTCTTNTTYTWYATGGA-3', and 5'-GAGGAGCGCGAGCGNGGNGTNACNAT-3'; and 3) 5' RACE using specific reverse primers designed from the RT-PCR sequences. The incomplete sequence of *Reticulomyxa filosa* EFL was downloaded from the NCBI EST database and finished by PCR on genomic DNA using degenerate forward primers. *Thalassiosira weissflogii* EFL was amplified from genomic DNA using degenerate primers designed from *Thalassiosira pseudonana* EFL. The FirstChoice RLM-RACE kit (Ambion, Austin, TX) was used for all 3' RACE and 5' RACE reactions. Superscript III One-Step RTPCR with Platinum Taq (Invitrogen, Carlsbad, CA) was used for all RT-PCR reactions. PCR products were cloned using the TOPO-TA cloning kit (Invitrogen, Carlsbad, CA) and sequenced in both directions using BigDye Terminator v. 3.1 (Applied Biosystems, Foster City, CA).

Phylogenetic analyses

New and previously published EFL sequences were translated and aligned using MAFFT (Katoh et al. 2002) and edited in MacClade 4.08 (Maddison and Maddison 2003) to a final matrix of 50 taxa and 518 unambiguously aligned characters. Phylogenetic trees were inferred using maximum likelihood (ML) and Bayesian methods. ML trees were inferred using PhyML 2.4.4 (Guindon and Gascuel 2003) with input trees generated by BIONJ, the WAG model of amino acids substitution, and four rate categories approximating a gamma distribution plus a proportion of invariant sites. In all, 1,000 bootstrap replicates were performed with PhyML using the α parameter and proportion of invariant sites estimated from the original tree. MRBAYES 3.0 (Ronquist and Huelsenbeck 2003) was used to perform Bayesian analysis using

the WAG substitution model with rates assigned by 4 equally probable categories approximating a gamma distribution. One cold and three heated chains were run for two million generations, sampling one tree every hundred generations. The first 5000 sampled trees were discarded as burn-in, and subsequent trees were used to compute the 50% majority-rule consensus tree.

Sequence analysis of targeting leaders

Eight nucleomorph-encoded and 22 nucleus-encoded genes for plastid products were identified from an ongoing EST survey of *Gymnochlora stellata*. Genes were identified by similarity to the *B. natans* plastid-targeted proteins and by searching for EST clusters encoding full-length proteins with bipartite targeting sequences at the amino terminus. Several were truncated at the 5' end, and these were completed by 5' RACE as described above. Seventeen nucleomorph-encoded plastid-targeted genes from the complete *B. natans* nucleomorph genome (Gilson et al. 2006) and the 45 nucleus-encoded plastid-targeted genes from a previous EST survey (Rogers et al. 2004) were analyzed in parallel for comparison. Putative transit peptides were analyzed for amino acid content and hydrophobicity. Sliding window plots of acidic, basic, and hydroxylated amino acids and hydropathy profiles were generated as described previously (Rogers et al. 2004) using a window size of 5 residues. For nucleomorph-encoded transit peptides, residues 3-23 were analyzed. Nucleus-encoded transit peptides were aligned at the signal peptide cleavage point predicted by SignalP 3.0 (Bendtsen et al. 2004) and 15 residues upstream and 20 residues downstream were considered for analysis.

Amino acid frequencies of transit peptides were calculated and are included as a supplementary table. Mature plastid-targeted and cytosolic protein data sets were assembled and their amino acid frequencies were also computed in order to provide a point of comparison to the transit peptides. Nucleomorph-encoded mature protein data sets consisted of residues 100 to the end of the shortest of the plastid-targeted proteins, thereby including 83 residues from each of 8 proteins for *G. stellata* and 84 residues from 17 proteins for *B. natans*. The nucleus-encoded mature protein data sets included residues from 100 to the end of all plastid-targeted proteins. Twenty-eight nucleus-encoded cytosolic proteins were also identified in the *G. stellata* ESTs, and compared against the 38 nucleus-encoded cytosolic proteins from the previous *B. natans* EST survey (Rogers et al. 2004). In addition, 16 nucleomorph-encoded non-targeted proteins from *G. stellata* were compared to their homologues in the *B. natans* nucleomorph genome. New sequences from this study were deposited in Genbank under accession numbers EU810236-EU810337.

Immunoblotting and localization

Polyclonal antibodies were raised against a mixture of two synthetic peptide sequences unique to the putatively PPC-targeted EFL but conserved among diverse chlorarachniophytes, CDQAKYKEERYNEILK and KETGGKKVEDPKMLK (BioSynthesis Inc., Lewisville, TX) and against two synthetic peptides from the *B. natans* cytosolic EFL, CIVGVNKMDEKSVKYD and GKITDCKNNPVKTVS (AbCam, Cambridge, UK). The *B. natans* cystosolic EFL antibodies were affinity-purified before use. Cells were harvested from cultures of *G. stellata*, *B. natans*, and *L. vacuolata*, pelleted by centrifugation, resuspended in 0.5mL lysis buffer (50mM Tris pH 7.5, 200mM sorbitol, 1mM EDTA) with 5μL protease inhibitor cocktail (Sigma-Aldrich, St. Louis, MO) and 10μL PMSF in isopropanol (20mg/mL) and repeatedly shock-frozen in liquid N_2 to release the proteins. Cell lysates were added to sample buffer and boiled for 15 minutes before separation by SDS-PAGE electrophoresis on a 10% Tris-Glycine gel. Separated proteins were transferred to Hybond-P PVDF transfer membrane (Amersham, Buckinghamshire, UK) for 70 minutes at 100V. Membranes were blocked and incubated with primary antiserum at 1:1000 and then a peroxidase-conjugated goat-anti rabbit IgG antibody (BioRad, Hercules, CA) at 1:3000 dilution. Blots were visualized using the ECL detection system (Amersham, Buckinghamshire, UK).

For localization experiments, *B. natans* cells were fixed in cold 4% paraformaldehyde, settled on cover glass slides coated with Histogrip (Invitrogen, Carlsbad, CA), and blocked with 5% BSA at room temperature, shaking, for one hour. Primary antibodies were applied at a concentration of 1:200 in 5% BSA for one hour, shaking, at room temperature. Slides were washed, incubated with Alexa Fluor 488 goat anti-rabbit IgG (Molecular Probes, Invitrogen, Carlsbad, CA) at a concentration of 1:1000, washed again, mounted using ProLong Gold antifade reagent with DAPI (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions, and observed on an Axioplan 2 compound microscope (Zeiss) with an AttoArc 2 100W Mercury lamp (Zeiss) for fluorescence visualization.

Results

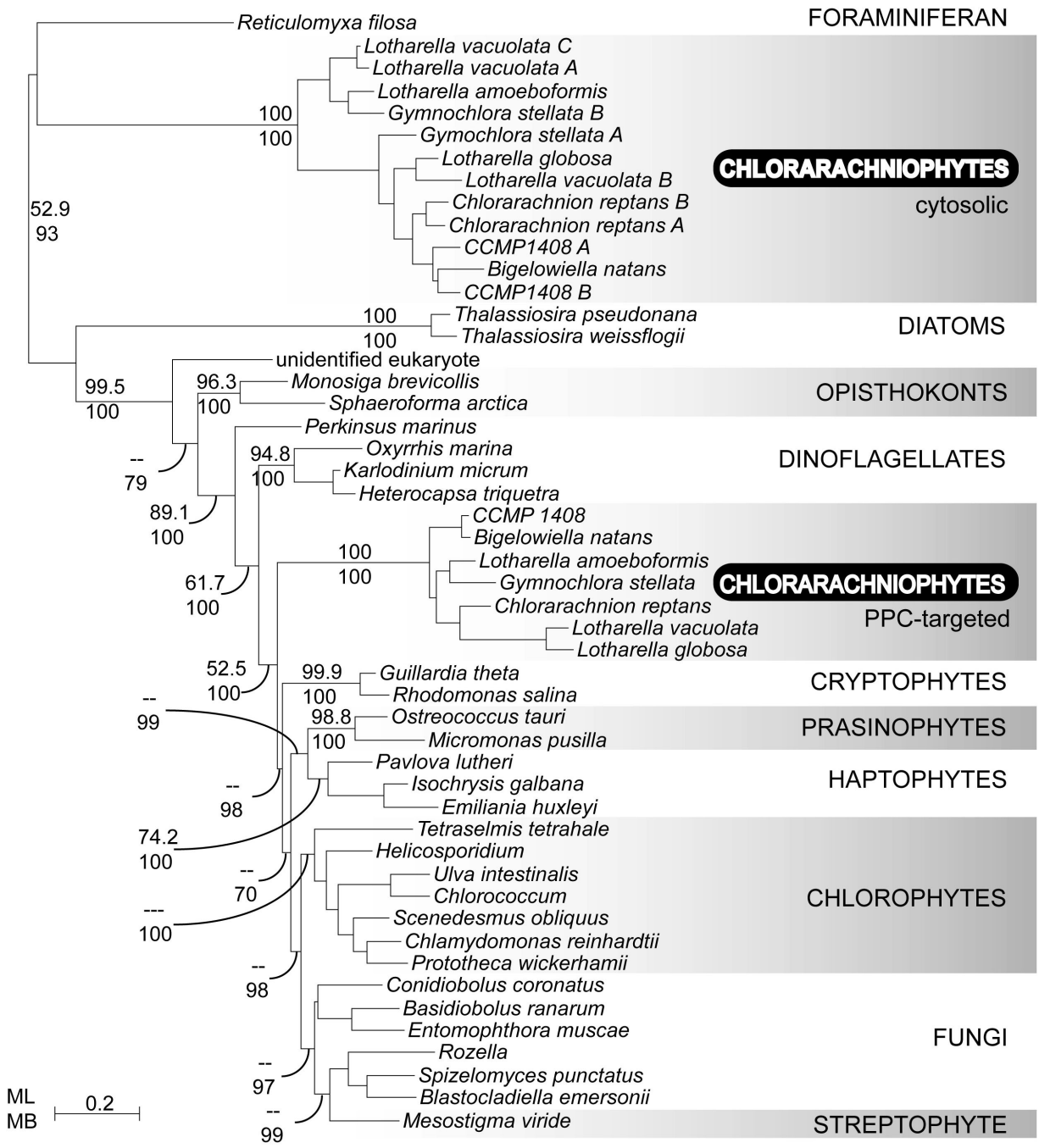
Chlorarachniophyte host nuclei encode two distinct clades of EFL

The 19 EFL sequences characterized from seven chlorarachniophyte strains group into two distantly related clades. One clade (referred to as "cytosolic" in Figure 5.2), is made up of relatively divergent sequences that share several distinguishing indels, and includes the EFL sequences found in both *B. natans* and *G. stellata* EST libraries. Some species were found to

encode two or three paralogues of this gene, some indicating recent duplications and others suggesting more ancient duplications. The clade is well supported overall (100%) and is related with weak support to the foraminiferan *Reticulomyxa filosa*, with which it shares two otherwise unique insertions. Genomic copies of cytosolic EFL were sequenced from *B. natans* and *G. stellata* and were found to completely lack introns. Complete 5' sequences were acquired by 5' RACE from *B. natans*, *L. globosa*, *Lotharella amoeboformis*, and the *L. vacuolata* sequences A and B, and none encode an extension longer than a few amino acids. Taken together, the evidence points to this gene being derived from the cercozoan host lineage and to its product functioning in the host cytoplasm.

The second clade (referred to as "PPC-targeted" in Figure 5.2) is also well supported (100%) and made up of slightly less divergent EFL sequences represented by a single sequence in each strain. This clade is not demonstrably related to green algal EFL, but the backbone of the tree is completely unresolved, and the green algae themselves do not group into a single clade. The possibility that the gene has a foreign origin cannot be ruled out, as approximately 20% of *B. natans* plastid-targeted proteins are thought to have been acquired by lateral gene transfer (Archibald et al. 2003). However, because EFL is relatively rare among eukaryotes compared to EF-1α, and because the chlorarachniophyte plastid is descended from a group where EFL is found in nearly all major subgroups (Noble, Rogers, and Keeling 2007), the simplest explanation for the origin of this second gene is that it is derived from the endosymbiont. None of the characterized copies of this gene appear to be encoded in the nucleomorph, however. This is supported by three lines of evidence. First, the complete nucleomorph genome of *B. natans* lacks EFL (Gilson et al. 2006). Second, the nucleomorph genome is >65% AT in the single copy regions (Gilson et al. 2006), but the PPC-targeted EFL sequences are 46.2% AT in *B. natans*, 49.8% in *G. stellata*, and 48.2% in *L. vacuolata*, consistent with expected nucleotide composition in the host nuclear genome. Third, and most compelling, nucleomorph introns are well studied and conspicuous: there are over 800 of them in the *B. natans* nucleomorph, and all are between 18 and 21bp in length (Gilson et al. 2006). Over 100 from *G. stellata* have been characterized and are similarly reduced (Slamovits, CH, unpublished data). We amplified and sequenced portions of the genomic copy of PPC-targeted EFL and found five introns in *L. vacuolata* and three introns each in *Chlorarachnion reptans* and *B. natans,* ranging from 47-176 bp in length.

Figure 5.2. Phylogeny of EFL



Protein maximum likelihood (ML) tree of 50 EFL sequences using 518 sites. Major groups are named to the right, and the inferred functional location of chlorarachniophyte proteins are indicated. Support for nodes greater than 50% is given from ML bootstrap values (above) and Bayesian posterior probability values (below).

Evidence for PPC targeting of EFL and characteristics of transit peptides

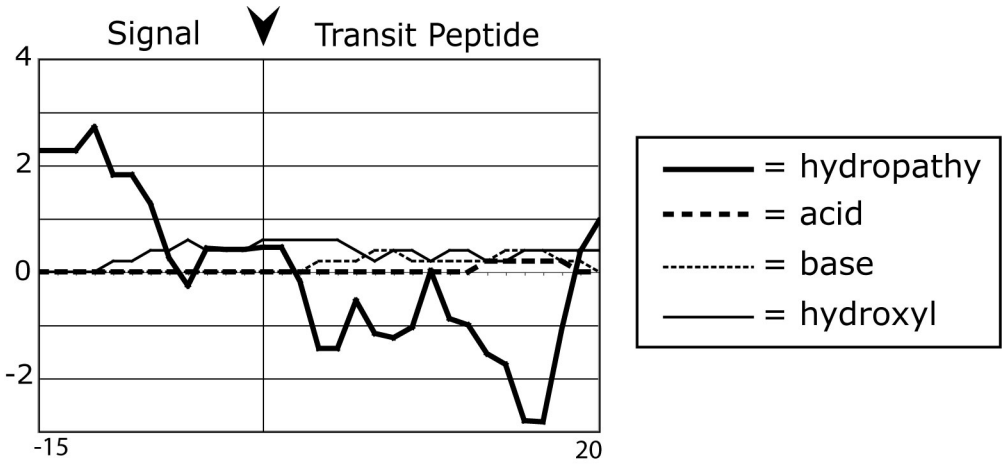
The putative PPC-targeted version of EFL is encoded in the host nucleus, so if it functions in the endosymbiont, it must be targeted. To examine these genes for evidence of

targeting peptides, we characterized the 5' end of PPC-targeted genes from *B. natans, G. stellata* and *L. vacuolata* by 5' RACE. In all three cases a bipartite targeting sequence was found, consisting of a signal peptide followed by a transit peptide-like sequence. Hydrophobic signal peptides of 39 amino acids in *B. natans* and 40 in *G. stellata* and *L. vacuolata* were predicted with high support (Figures 5.3 and 5.4 A-B), as is typical for chlorarachniophyte plastid-targeted proteins (Rogers et al. 2004). The predicted signal cleavage sites are ARQ for *B. natans*, ALA for *G. stellata,* and SFA for *L. vacuolata*, which conform to the expectations for signal cleavage sites of eukaryotic secreted proteins. Sliding window plots of the PPC-targeting sequences show extreme levels of hydrophobicity in the signal peptide relative to the plots of plastid-targeting sequences (Figure 5.4 A-B vs. C-D), but this is mainly due to the dampening effect that averaging has on the larger dataset of plastid-targeting peptides.

Between the predicted signal cleavage site and the start of sequence conservation with mature EFL proteins is between 59 (in *B. natans*) and 81 (in *L. vacuolata*) amino acids of sequence. With only one putative PPC-targeted protein from each species, characteristics of this class of targeting sequence within each species cannot be generalized, but comparing the *B. natans* and *G. stellata* PPC-targeted EFL targeting sequences to transit peptides from their plastid-targeted proteins is informative. A collection of nucleus-encoded plastid-targeted proteins from *B. natans* has been analyzed previously (Rogers et al. 2004), so we developed a comparable set of proteins from *G. stellata*. Twenty-four nucleus-encoded plastid-targeted proteins were identified from a *G. stellata* EST library and, where truncated, the complete sequence of their leaders acquired by 5' RACE. Twenty-eight *G. stellata* cytosolic proteins were also identified and completely sequenced to provide a baseline of amino acid composition of non-targeted sequences. Transit peptide lengths could not be determined unambiguously in all cases, so only the first 20 amino acids following the predicted signal cleavage site were analyzed for chemical characteristics. The transit peptides from *G. stellata* and *B. natans* plastid-targeted proteins have remarkably similar amino acid compositions, and most significantly they share characteristics with the transit peptides of the PPC-targeted EFL (Figure 5.5 A-D). In plastid- and PPC-targeted proteins in both species this region is enriched in serine (S) and arginine (R) relative to both the cytosolic and mature plastid-targeted proteins, although *G. stellata* is further enriched in alanine (A) and proline (P). They are both depleted in acidic residues (aspartic and glutamic acid, D and E) as is expected of transit peptides in general (Figure 5.4 C-D), and both are more severely depleted in lysine (K) than either aspartic or glutamic acid. These characteristics are reminiscent of plant transit peptides, which also tend to be enriched in serine

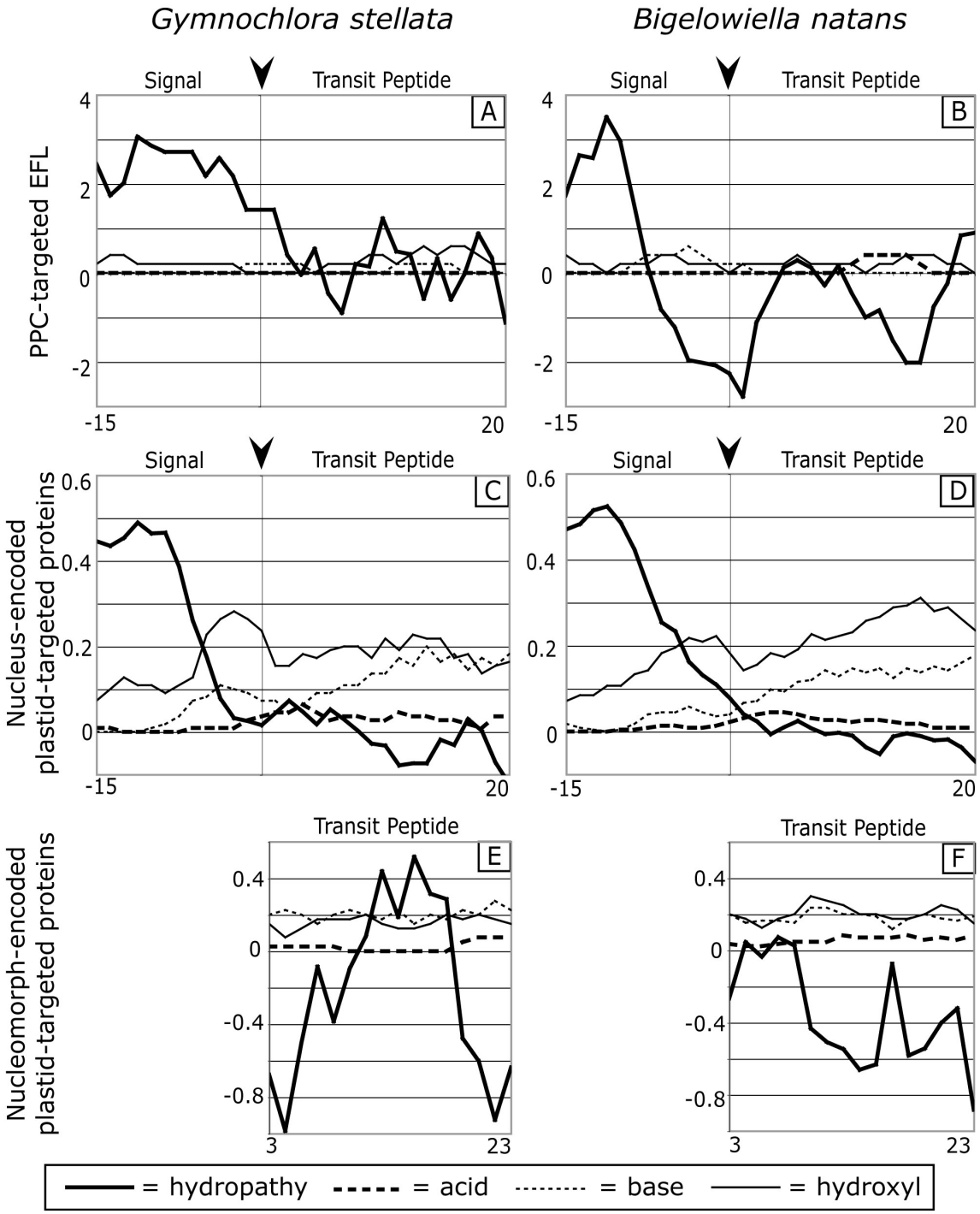
and arginine and depleted in the acidic residues. Overall, the N-terminal extensions of the PPC-targeted EFL proteins share all known characteristics of the chlorarachniophyte nucleus-encoded plastid-targeting leaders.

Figure 5.3. Signal and transit peptide characteristics of *Lotharella vacuolata* EFL



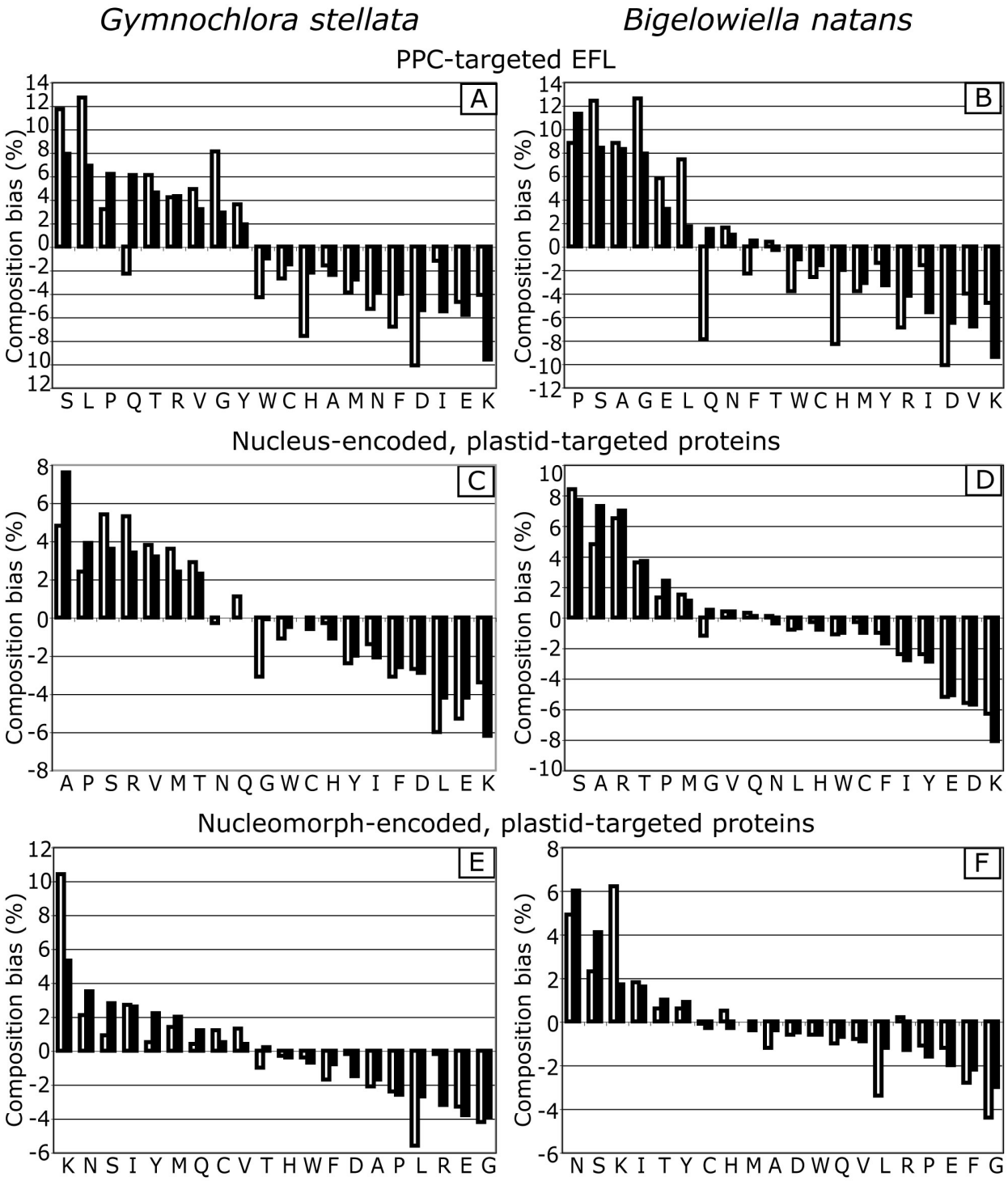
Sliding window plot of *L. vacuolata* targeted EFL signal and transit peptide characteristics. Hydropathy profiles were computed by dividing the total of the Kyte-Doolittle hydropathy scores of the residues in the window over the total window size (here 5 residues). Acid, base, and hydroxyl plots represent the number of residues with that property in the window divided by the size of the window (also 5 residues). Arrowhead indicates predicted signal peptide cleavage site.

Figure 5.4. Signal and transit peptide characteristics of two chlorarachniophytes



Sliding window plots of signal and transit peptide characteristics of PPC-targeted EFL (A and B), nucleus-encoded plastid-targeted proteins (C and D) and nucleomorph-encoded plastid-targeted proteins (E and F). Hydropathy profiles are computed by averaging the total of the Kyte-Doolittle hydropathy scores of the residues in the window and dividing by the total window size (here 5 residues). Acid, base, and hydroxyl plots represent the average number of residues with that property in the window divided by the size of the window (also 5 residues). Scores are averaged over all transit peptides in that class. Proteins from each class are aligned at their predicted signal cleavage sites (arrowheads).

Figure 5.5. Transit peptide amino acid frequencies

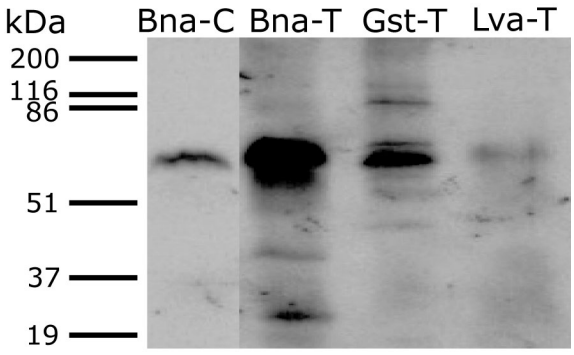


Relative amino acid composition of transit peptides of PPC-targeted EFL (A and B), nucleus-encoded plastid-targeted proteins (C and D) and nucleomorph-encoded plastid-targeted proteins (E and F). Bars indicate the difference between amino acid frequencies (%) in transit peptides versus mature targeted proteins (outlined bars) and transit peptides versus cytosolic proteins (solid bars).

Western blot results are consistent with post-import cleavage of PPC-targeting peptides

The predicted leader sequences in PPC-targeted EFL genes are substantial in size (over 100 amino acids), so we analyzed the size of mature PPC-targeted EFL by immunoblotting to determine whether there is any comparable reduction in the apparent size of the mature protein compared with the predicted size of the full-length gene product. Western blots of proteins from *B. natans*, *G. stellata*, and *L. vacuolata* with antibodies raised against peptide sequences specific to the PPC-targeted proteins were compared with blots of *B. natans* proteins reacted to antibodies raised against the *B. natans* cytosolic EFL. In all three species, the major PPC-targeted EFL band is significantly smaller than the predicted size of the full-length protein and similar in size to the *B. natans* cytosolic EFL (Figure 5.6). This is the size we would expect for the mature targeted EFL if the targeting sequence is cleaved. This method lacks the resolution to determine the exact length of the putatively cleaved sequence (assuming the major band is our protein of interest), but size estimates of PPC-targeted proteins suggest that cleavage would take place at or near the stretch of 2-3 glycine residues followed by a stretch of acidic and small neutral residues that immediately precedes the start of sequence homology to other EFL sequences. The lengths of the entire bipartite leader sequences would thus be approximately 90 amino acids in *B. natans*, 100 in *G. stellata*, and 110 in *L. vacuolata*.

Figure 5.6. Western blotting of EFL



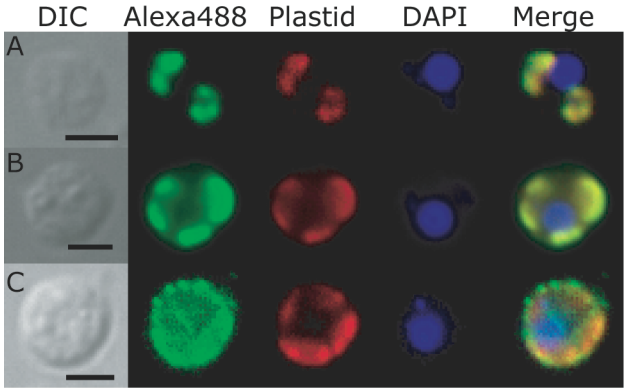
Western blots of proteins extracted from *B. natans* (Bna), *G. stellata* (Gst), and *L. vacuolata* (Lva) probed with antibodies raised to the *B. natans* cytosolic EFL (C) or epitopes common to the PPC-targeted clade of EFL (T).

Cytosolic and targeted EFL have distinct localization patterns

The cellular location of both EFL proteins was examined in *B. natans* using immunofluorescence and revealed two distinct localization patterns (Figure 5.7). The cytosolic

EFL protein appears to be distributed throughout the cytoplasm, as expected. The PPC-targeted EFL has a distinct localization pattern, co-localizing with plastid auto-fluorescence. From transmission electron micrographs, the PPC is known to be a thin layer that completely surrounds the plastid (Gilson and McFadden 1999, Moestrup and Sengco 2001). However, we were unable to detect a staining pattern that clearly surrounded the plastid without also occurring within it, even using confocal microscopy (data not shown). Moreover, the targeted EFL does not appear to co-localize with the nucleomorphs, which are faintly visible by DAPI staining as small satellites in Figure 5.7, and which would be expected to be surrounded by the largest visible volume of PPC. Assuming the antibody is specific for PPC-targeted EFL, and that the plastid signal is not due to some other artifact of fixation or binding, this observation can be interpreted in a number of ways. First, the 'PPC-targeted' EFL might actually function in the chloroplast. This would be quite remarkable, since EFL is so far strictly found in eukaryotic cytoplasm, and the *tufA* gene in the chloroplast genome is known to be transcriptionally active by its representation in the *B. natans* EST library. Alternatively, the targeting system may not be sufficiently differentiated to discriminate perfectly between plastid and PPC-targeted proteins, so that PPC proteins are present in both compartments, and the PPC signal is obscured by the plastid signal. Both of these scenarios are consistent with the overall resemblance of the targeted EFL leader to plastid-targeting peptides. The other major possibility is that the observed localization pattern does not reflect the actual location of this protein. This could be due to nonspecific binding of our polyclonal antibodies to chloroplast proteins (although BLAST similarity searches of our EFL epitopes against the nucleomorph and chloroplast genomes of *B. natan*s fail to return any matches), or it could be an artifact of the localization procedure such as degradation of the chloroplast membranes before fixation. A resolution between these possibilities should be available soon. The *B. natans* genome is currently being sequenced, and analysis of more putatively PPC-targeted proteins from that genome should clarify whether or not there is a distinct class of targeting peptides specific for PPC targeting and if the PPC-targeted EFL leader matches the characteristics of that class. In addition, a transformation system for *Lotharella amoeboformis* has been published recently (Hirakawa, Kofuji, and Ishida 2008), and when putatively PPC-targeted sequences are available from this species or when a transformation protocol is available for *B. natans* or *G. stellata*, GFP fusions can be used to determine the location of proteins much more clearly.

Figure 5.7. Immunolocalization of EFL

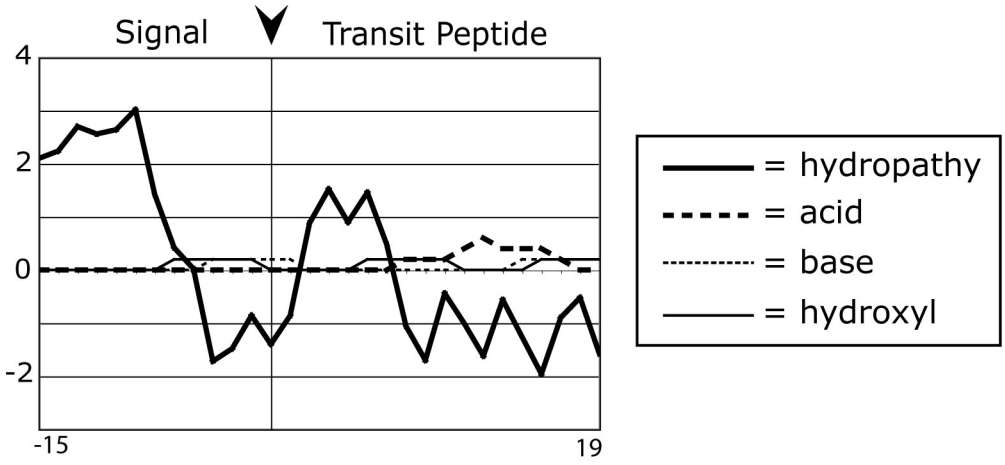


Immunolocalization of targeted (A and B) and cytosolic (C) EFL proteins in *B. natans*. Fixed cells of *B. natans* hybridized with antibodies specific to targeted EFL (A and B) and cytosolic EFL (C) are shown with differential interference contrast (DIC) optics in the leftmost column. Fluorescence was visualized under a mercury lamp in the four rightmost columns. From left to right, the fluorescence images show the Alexa 488 goat-anti-rabbit 2º antibody hybridized to EFL antibodies in green, plastid autofluorescence in red, DAPI fluorescence staining the prominent nucleus and adjacent tiny nucleomorphs in blue, and finally the three fluorescence images are merged, indicating co-localization of plastid autofluorescence and targeted EFL in A and B, and a cytosolic location for EFL in C. Scale bar represents 2μm.

PPC-targeted eIF1

If the leader on the PPC-targeted EFL does represent the characteristics of PPC-targeted peptides as a whole, then other putatively PPC-targeted proteins might be identifiable based on their possession of similar leaders. Accordingly, a eukaryotic translation initiation factor 1 (eIF1), was identified in the *G. stellata* EST library by virtue of a similar N-terminal extension to that found on the targeted clade of EFL. The leader consists of a 31-residue signal peptide and a short stretch of amino acids with similar characteristics to chlorarachniophyte transit peptides before the start of the eIF1 domain (Figure 5.8). Although the N-terminus is variable in eIF1 proteins from diverse eukaryotes, and some sequences have N-terminal extensions longer than that found on *G. stellata*, this is the only case in which SignalP strongly predicts a signal peptide. This protein is not likely encoded in the *G. stellata* nucleomorph because its AT content is only 57.7%, and because it is missing from the complete *B. natans* nucleomorph genome. Like the PPC-targeted EFL leaders, eIF1 has an acidic stretch near its C-terminus, a characteristic that is uncommon in transit peptides. While this may represent a recognizable trait by which the cell differentiates between PPC and plastid protein traffic, this possibility will need to be validated experimentally.

Figure 5.8. Signal and transit peptide characteristics of *Gymnochlora stellata* eIF1



Sliding window plot of *G. stellata* putatively targeted eIF1 signal and transit peptide generated as per Figure 5.3. Arrowhead indicates predicted signal peptide cleavage site.

Characteristics of G. stellata *and* B. natans *nucleomorph-encoded transit peptides*

The *B. natans* nucleomorph genome encodes 17 annotated genes for plastid-targeted proteins. We identified eight of these as full-length cDNAs in the *G. stellata* EST survey. Based on the start of sequence similarity between the nucleomorph proteins and their green algal homologues, transit peptides in both species ranged from 23 to 70 amino acids, but cleavage sites could not be unambiguously assigned. To characterize these leaders and determine how they differ from nucleus-encoded transit peptides, we assembled an additional set of 16 nucleomorph-encoded cytosolic proteins from the *G. stellata* ESTs and their counterparts from the *B. natans* nucleomorph genome.

In general, nucleomorph-encoded transit peptides are enriched in basic residues, especially lysine at 17%, and depleted in acidic residues, especially glutamic acid (Figure 5.4 E-F), relative to the mature proteins. Because the chlorarachniophyte plastid is descended from a green alga, we might expect to find enriched levels of serine and threonine, but they are only somewhat enriched relative to cytosolic proteins and not at all enriched relative to mature plastid-targeted proteins. The clearest trend in these transit peptides is the enrichment in asparagine (N) and lysine (K) (Figure 5.5 E-F and Table 5.1). This is reminiscent of apicomplexan and cryptomonad nucleomorph transit peptides, where the bias is driven by the high AT content of their genomes favoring amino acids with AT rich codons (Ralph et al. 2004) and contrasts sharply with the nucleus-encoded transit peptides in which lysine is the most severely depleted residue. Leucine (L) and glycine (G) are the most depleted in

chlorarachniophyte nucleomorph transit peptides, even more than the acidic residues.

Discussion

Origin and evolution of chlorarachniophyte EFL genes

We have shown that the host nuclear genomes of a diverse sample of chlorarachniophytes encode two phylogenetically distinct EFL genes. One is weakly related to foraminiferan relatives of chlorarachniophytes, encodes no targeting information, and is localized to the cytosol in immunolocalization experiments. We conclude that the product of this gene functions in the host cytoplasm. The other gene is not demonstrably related to green algal EFL, but we conclude that its product nevertheless most likely functions in the endosymbiont cytosol for a number of reasons. First, the translation function of EF-1α/EFL is essential and the endosymbiont likely encoded EFL when it was engulfed, but the relict nucleomorph genome no longer encodes either gene. Second, the PPC-targeted EFL encodes a leader with all the characteristics expected of a plastid-targeting leader in chlorarachniophytes, and immunoblotting indicates the leader is processed as expected for a targeting peptide. Immunofluorescence localization is consistent with a PPC and/or a plastid location of this protein, but since EFL is restricted to eukaryotes, and its presumed function in the plastid is fulfilled by the plastid-encoded *tufA* (Rogers et al. 2007), the logical compartment in which to assign the second EFL is the endosymbiont cytosol (although whether it also exists in the plastid in *B. natans* in a functional capacity needs to be clarified). The simplest explanation for this is that both host and endosymbiont used EFL at the time they were united, and the endosymbiont gene moved to the host nucleus from which its product is post-translationally targeted back to the compartment in which it has always functioned. This makes chlorarachniophytes an interesting case, as they are a union of two cells with the relatively rare EFL protein, unlike cryptomonads where the host uses EFL but the endosymbiont uses a nucleus-encoded, PPC-targeted EF-1α (Gould, Sommer, Kroth, et al. 2006).

Parallel evolution of PPC targeting in chlorarachniophytes and cryptomonads

The first putative nucleus-encoded PPC-targeted proteins have recently been described in cryptomonads, including EF-1α (Wastl and Maier 2000; Gould, Sommer, Kroth, et al. 2006; Sommer et al. 2007). Interestingly, a comparison of PPC- and plastid-targeting peptides came to the same conclusion as we reach here: the leaders are composed of signal peptides and transit peptide-like sequences (Gould, Sommer, Kroth, et al. 2006). Because cryptomonads and

chlorarachniophytes are very distantly related and their plastids were acquired by independent endosymbiotic events from two different primary algal groups, any similarity in their PPC-targeting systems must have evolved in parallel. This is significant because other major steps in the targeting pathway were assembled from existing machinery of the host (signal peptides and endomembrane targeting) or the endosymbiont (transit peptides and the TIC and TOC systems, although note that so far only one putative TOC component has been identified in chlorarachniophytes [Gilson et al. 2006] and none in cryptomonads). PPC targeting and the crossing of the second membranes are the only steps that could significantly differ between these independently evolved systems, and yet it now appears that both groups have arrived at the same solution, namely some modification of the transit peptide. This is more remarkable given the different plastid membrane topology in these two groups. The outermost membranes of chlorarachniophyte plastids are smooth and not continuous with the endoplasmic reticulum, whereas cryptomonad plastids reside within the rough endoplasmic reticulum. Thus the strategy for targeting between the entry to the endomembrane and the entry to the plastid itself could be quite different. It has been hypothesized that the second membrane houses a modified TOC complex that recognizes some transit peptides and not others (Cavalier-Smith 1999), and such a model is entirely consistent with our data for chlorarachniophytes. The only difference at present is that we have not determined the mechanism by which the chlorarachniophyte cell distinguishes between PPC- and plastid-targeting transit peptides. The phenylalanine identified as critical in cryptomonads (Gould, Sommer, Kroth, et al. 2006) was not used by the ancestor of green algae (Patron and Waller 2007), so if this is a key to PPC targeting in cryptomonads then chlorarachniophytes must have adopted a different key (and perhaps a less stringent key if the distinction between plastid and PPC targeting is as relaxed as our initial immunolocalization data suggest). Apicomplexan parasites may provide a useful comparison for understanding chlorarachniophyte plastid targeting because both groups have 4-membrane bound plastids with a smooth outer membrane that is not continuous with the endoplasmic reticulum. Although these groups are unrelated, similarities in plastid targeting between *Euglena gracilis* and dinoflagellates show that plastid membrane topology can influence targeting despite lack of relatedness (Patron et al. 2005, Durnford and Gray 2006). The EFL and eIF1 targeting sequences share an acidic stretch at their C-termini, but whether this is a general trend and whether it has any functional significance remains to be seen. Targeted chlorarachniophyte EFL sequences also have a hydrophilic sequence of alternating stretches of lysine and aspartic acid residues at their C-termini that is lacking in all other EFL proteins, including cytosolic chlorarachniophyte EFL.

If this extension is involved in targeting, it would represent a novel mechanism for targeting to a plastid, although peroxisomal proteins are targeted via an uncleaved C-terminal extension. However, this feature is lacking in the putatively PPC-targeted eIF1. The use of bipartite targeting sequences in both chlorarachniophytes and cryptomonads would suggest potential restrictions on the range of possible solutions to the problem of PPC targeting, and in both cases it was solved by modifying an existing system.

Three classes of transit peptides in chlorarachniophytes

While the nucleus-encoded plastid- and PPC-targeting peptides in chlorarachniophytes are unexpectedly similar to one another, the transit peptides of nucleomorph-encoded proteins are remarkably different, despite the fact that these proteins are destined for the same compartment and presumably recognized by the same import complexes. This is most likely due to differences in AT content between nuclear and nucleomorph genomes in chlorarachniophytes; cryptomonad nucleomorph-encoded transit peptides are similarly biased toward lysine and asparagine (Ralph et al. 2004), and the genomes are similarly AT-rich (Douglas et al. 2001; Gilson et al. 2006). The AT-rich genomes of *Plasmodium* species encode similarly asparagine and lysine-rich transit peptides (Ralph et al. 2004). If the same translocons are responsible for bringing nucleus- and nucleomorph-encoded proteins across the innermost pair of plastid membranes, the observed differences in amino acid usage further indicates a certain flexibility in transit peptide recognition.

Literature Cited

Archibald JM, Rogers MB, Toop M, Ishida K, Keeling PJ. 2003. Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigelowiella natans*. Proc. Natl. Acad. Sci. USA 100:7678-7683.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. 340:783-795.

Cavalier-Smith T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: Euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. J. Eukaryot. Microbiol. 46:347-366.

Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu X, Reith M, Cavalier-Smith T, Maier UG. 2001. The highly reduced genome of an enslaved algal nucleus. Nature 410:1091-1096.

Durnford DG, Gray MW. 2006. Analysis of *Euglena gracilis* plastid-targeted proteins reveals different classes of transit sequences. Eukaryot. Cell. 5:2079-2091.

Gile GH, Patron NJ, Keeling PJ. 2006. EFL GTPase in cryptomonads and the distribution of EFL and EF-1alpha in chromalveolates. Protist 157:435-444.

Gilson PR, McFadden GI. 1999. Molecular, morphological and phylogenetic characterization of six chlorarachniophyte strains. Phycol. Res. 47:7-19.

Gilson PR, Su V, Slamovits CH, Reith ME, Keeling PJ, McFadden GI. 2006. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: Nature's smallest nucleus. Proc. Natl. Acad. Sci. USA 103:9566-9571.

Gould SB, Sommer MS, Kroth PG, Gile GH, Keeling PJ, Maier UG. 2006. Nucleus-to-nucleus gene transfer and protein retargeting into a remnant cytoplasm of cryptophytes and diatoms. Mol. Biol. Evol. 23:2413-2422.

Gould SB, Sommer MS, Hadfi K, Zauner S, Kroth PG, Maier UG. 2006. Protein targeting into the complex plastid of cryptophytes. J. Mol. Evol. 62:674-681.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696-704.

Hirakawa Y, Kofuji R, Ishida K. 2008. Transient transformation of a chlorarachniophyte alga, *Lotharella amoebiformis* (Chlorarachniophyceae), with *uidA* and *egfp* reporter genes. J. Phycol. 44:814-820.

James TY, Kauff F, Schoch CL, et al. (70 co-authors). 2006. Reconstructing the early evolution of fungi using a six-gene phylogeny. Nature 443:818-822.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059-3066.

Keeling PJ, Inagaki Y. 2004. A class of eukaryotic GTPase with a punctate distribution suggesting multiple functional replacements of translation elongation factor 1alpha. Proc. Natl. Acad. Sci. USA 101:15380-15385.

Maddison DR, Maddison WP. 2003. MacClade 4: Analysis of phylogeny and character evolution. 4.08. [computer program]

McFadden GI. 1999. Plastids and protein targeting. J. Eukaryot. Microbiol. 46:339-346.

Moestrup Ø, Sengco M. 2001. Ultrastructural studies on *Bigelowiella natans*, gen. et sp. nov., a chlorarachniophyte flagellate. J. Phycol. 37:624-646.

Noble GP, Rogers MB, Keeling PJ. 2007. Complex distribution of EFL and EF-1alpha proteins in the green algal lineage. BMC Evol. Biol. 7:82.

Patron NJ, Waller RF. 2007. Transit peptide diversity and divergence: A global analysis of plastid targeting signals. Bioessays 29:1048-1058.

Patron NJ, Waller RF, Archibald JM, Keeling PJ. 2005. Complex protein targeting to dinoflagellate plastids. J. Mol. Biol. 348:1015-1024.

Ralph SA, Foth BJ, Hall N, McFadden GI. 2004. Evolutionary pressures on apicoplast transit peptides. Mol. Biol. Evol. 21:2183-2194.

Rogers MB, Gilson PR, Su V, McFadden GI, Keeling PJ. 2007. The complete chloroplast genome of the chlorarachniophyte *Bigelowiella natans*: Evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. Mol. Biol. Evol. 24:54-62.

Rogers MB, Archibald JM, Field MA, Li C, Striepen B, Keeling PJ. 2004. Plastid-targeting peptides from the chlorarachniophyte *Bigelowiella natans*. J. Eukaryot. Microbiol. 51:529-535.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Ruiz-Trillo I, Lane CE, Archibald JM, Roger AJ. 2006. Insights into the evolutionary origin and genome architecture of the unicellular opisthokonts *Capsaspora owczarzaki* and *Sphaeroforma arctica*. J. Eukaryot. Microbiol. 53:379-384.

Sommer MS, Gould SB, Lehmann P, Gruber A, Przyborski JM, Maier UG. 2007. Der1-mediated preprotein import into the periplastid compartment of chromalveolates? Mol. Biol. Evol. 24:918-928.

Steiner JM, Loffelhardt W. 2005. Protein translocation into and within cyanelles. Mol. Membr. Biol. 22:123-132.

Wastl J, Maier UG. 2000. Transport of proteins into cryptomonads complex plastids. J. Biol. Chem. 275:23194-23198.

Wunder T, Martin R, Loffelhardt W, Schleiff E, Steiner JM. 2007. The invariant phenylalanine of precursor proteins discloses the importance of Omp85 for protein translocation into cyanelles. BMC Evol. Biol. 7:236.

CHAPTER 6: CONCLUSION

EFL is more common than previously thought

At the outset of these projects, EFL was known from approximately 15 sequences from haptophytes, dinoflagellates, certain green algae, certain zygomycete fungi, a choanoflagellate, and a chlorarachniophyte (Keeling and Inagaki 2004). Over the course of this work, the number of known EFL gene sequences has more than quintupled, to a total of approximately 80 sequences annotated in GenBank, with many more present in public EST and metagenomic datasets, and many more again in unpublished EST libraries and genome projects in progress. EFL is now known to occur in all five supergroups of eukaryotes (Keeling et al. 2005, see Figure 1.2). New lineages found to encode EFL since its initial discovery include foraminifera, cryptomonads, certain red algae, diplonemids, certain kinetoplastids, a deep-branching euglenid, and *Raphidiophrys contractilis*, a centrohelid heliozoan, and three new clades of EFL that branch near heterokonts have been identified in environmental data but not yet associated with particular groups of organisms (pers. obs., not shown). Altogether, this indicates that EFL is far more common than initially suspected, and suggests that further EFL-encoding lineages are yet to be discovered. This dramatic increase in available data has provided a more nuanced understanding of the distribution of EFL while characterizing occurrences of endosymbiotic gene transfer and differential loss in diverse eukaryotic lineages.

Endosymbiotic gene transfer of EFL

On the finest taxonomic scale investigated, in chlorarachniophytes, EFL was inferred to have transferred from the nucleomorph, the green algal endosymbiont's reduced nucleus, to the nucleus of the host. At the outset of this project, it was known that the nucleomorph genome encodes neither EFL nor EF-1α (Gilson et al. 2006), so it was expected that some gene for one of these essential proteins is now encoded in the nucleus and targeted to the periplastid compartment (PPC) that houses the nucleomorph. The most likely candidate would be the endosymbiont-derived gene, which was most likely EFL as the plastids are descended from a chlorophyte green alga (Van de Peer et al. 1996; Ishida et al. 1997), but it was also possible that a duplicate host-derived copy or a xenologous copy could instead have acquired targeting information to reach the periplastid compartment: certain host-derived genes are targeted to the plastid in other lineages (Harper and Keeling 2003), and a significant proportion of characterized plastid targeted genes in *Bigelowiella natans* are thought to have been acquired by lateral transfer (Archibald et al. 2003). In addition, it was unknown at what point in the diversification of

chlorarachniophytes the EFL gene was lost from the nucleomorph, as *B. natans* is a late-branching chlorarachniophyte (Gilson and McFadden 1999; Silver et al. 2007) and no other nucleomorphs have been characterized.

EFL was characterized from representative chlorarachniophytes spanning the phylogenetic diversity of the group. Two distinct clades were found, one related to foraminiferan EFL that was inferred to be derived from the host nucleus, and one including targeting information that was inferred to be derived from the endosymbiont nucleus. These results further indicated that the loss of EFL from the nucleomorph took place early in the evolution of chlorarachniophytes, as none of the characterized EFL sequences showed nucleomorph characteristics, such as high AT content and abundant tiny introns, in agreement with the presence of targeting information. Although the targeted clade of EFL did not group with other green algae, there was no evidence for it being derived from another source (Figure 5.2). During its tenure within a chlorarachniophyte cell, the likely chlorophyte-derived copy of EFL would have experienced a period of accelerated evolution, as evidenced by the long branch in phylogenetic analyses (Figure 5.2), which may contribute to this clade's failure to group with other green algae. Considering that known green algal EFL sequences do not all group together and the closest free-living relative of chlorarachniophyte plastids may not have been sampled yet (or may be extinct), this clade is most likely derived from the endosymbiont and therefore EFL has undergone endosymbiotic gene transfer in chlorarachniophytes (Gile and Keeling 2008).

A similar endosymbiotic event took place in the cryptomonads, where a nucleus-encoded EF-1α with red algal affinities was characterized and shown to encode a similar targeting sequence (Gould et al. 2006). The bipartite targeting peptide found on cryptomonad EF-1α directed import to a "blob-like structure" on diatom plastids, which was interpreted as a periplastid compartment, despite the fact that diatom plastids do not have nucleomorphs. However, the plastids of cryptomonads and diatoms are considered to be homologous (Keeling 2009), and the detection of potentially PPC-targeted proteins along with a candidate import system in diatoms have bolstered this interpretation (Sommer et al. 2007). In chlorarachniophytes, the potential PPC-targeted EFL was localized by immunofluorescence, with inconclusive results; it is clearly not cytosolic, but the pattern could not be distinguished from plastid localization (Gile and Keeling 2008). A transient transformation protocol has since been developed for the chlorarachniophyte *Lotharella amoeboformis* (Hirakawa, Kofuji, and Ishida 2008), and the targeted EFL fused to GFP exhibits a clear PPC-localization signal, supporting the conclusion of PPC-targeting that was previously made on the basis of sequence analysis. As the

first PPC-targeted protein characterized in chlorarachniophytes, the endosymbiotically transferred EFL revealed the last unknown category of targeting information for complex plastids (Patron and Waller 2007).

Direct phylogenetic evidence for differential loss of EFL and EF-1α

The clearest evidence for differential loss of EFL comes from the Euglenozoa. EFL was found in each of the three major lineages of Euglenozoa: throughout the diplonemids, in one deep-branching euglenid, and in two deep-branching kinetoplastids, that crucially, are not sister taxa. Because the kinetoplastids group together in both EFL and EF-1α phylogenies, both proteins were inferred to have been inherited vertically from an ancestor that encoded both proteins. The presence of EFL in diplonemids and *Petalomonas cantuscygni* suggested that this inference could be extended to account for the distribution of EFL throughout the Euglenozoa. Although diplonemid, kinetoplastid, and euglenid EFL sequences branch separately, the phylogeny of EF-1α unites euglenozoans with strong support, indicating its presence in the Euglenozoan ancestor. Given the lack of support for any of the nodes separating the three euglenozoan EFL lineages and the strong evidence for differential loss in kinetoplastids, it would seem reasonable to conclude that differential loss could lead to the complex distribution in Euglenozoa as a whole (Gile, Faktorová et al. 2009).

Notably, this inference requires less weight to be given to the phylogeny of EFL and greater weight to be given to its distribution and to the unlikelihood of multiple lateral transfers to closely related organisms, a perspective that would influence the interpretation of lateral gene transfer if applied to other studies (see below). Although this study provided strong evidence that differential loss has occurred, the ultimate origin of EFL in the Euglenozoa remains unknown, and given the lack of evidence for EFL in any other excavate lineages, lateral transfer may still have played a major role by introducing EFL to the Euglenozoa.

Differential loss of EFL and EF-1α in the green lineage

Green algae are among the most thoroughly surveyed groups for EFL and EF-1α, and they provided the first suggestion that EFL and EF-1α might have been differentially lost from an ancestor in which they co-occurred, though the phylogeny of EFL is more consistent with multiple lateral transfers of EFL into this group. The initial description of EFL included sequences from chlorophycean and trebouxiophycean green algae and an environmental sequence that was later shown to belong to *Micromonas pusilla*, while land plants are known to

encode EF-1α (Keeling and Inagaki 2004). This distribution could be generalized as EFL in the Chlorophyta and EF-1α in the Streptophyta, which are the two main clades of the green lineage, except for the anomalous EF-1α sequence reported for *Acetabularia acetabulum*, which belongs to the Ulvophyceae, an order within the Chlorophyta. A follow up investigation further complicated the distribution by discovering EFL in other ulvophyceans and in a deep-branching streptophyte, *Mesostigma viride*, and recovering a close relationship between *A. acetabulum* EF-1α sequences and those of land plants. This relationship suggests that *A. acetabulum* EF-1α was inherited from its common ancestor with land plants, despite the EFL-encoding prasinophyte and chlorophyte taxa that branch between them. The phylogeny of EFL weakly recovers the monophyly of most chlorophyte green algae, likewise implying that EFL was acquired once in the ancestor of this lineage, though independent lateral transfers may have contributed to its presence in *M. viride* and the prasinophytes *Ostreococcus tauri* and *M. pusilla* (Noble, Rogers, and Keeling 2007).

Although a picture of predominantly vertical inheritance of EFL and EF-1α had emerged in the green algae, the anomalous *A. acetabulum* EF-1α sequence begged further investigation. To clarify this discrepancy, representatives of each order within the Ulvophyceae were examined for the presence of EFL and EF-1α. The Ulvophyceae consists of two main clades, sometimes referred to as the Ulvophyceae I, which includes the orders Caulerpales, Dasycladales, and Siphonocladales, and the Ulvophyceae II, which includes the Ulvales and Ulotrichales (Watanabe, Kuroda, and Maiwa 2001). In an unusual bout of consistency, all representatives of the Ulvophyceae I were found to encode EF-1α related to *A. acetabulum* and land plants, and all representatives of the Ulvophyceae II were found to encode EFL (Gile, Novis, et al. 2009). However, ulvophycean EFL monophyly was not recovered due to the exclusion of *Ochlochaete hystrix*, an ulvophycean that branches with the Ulvaceae in SSU rRNA analyses (O'Kelly, Wysor, and Bellows 2004). This exclusion may have been due in part to local disruption from the divergent EFL sequence of *Helicosporidium*, a parasitic trebouxiophyte. Approximately unbiased (AU) tests fail to reject the possibility of ulvophycean EFL monophyly with this sequence removed ($p=0.188$, not shown, see Figure 3.3 C or D topology), and another study examining a different set of ulvophycean taxa successfully recovered the monophyly of ulvophycean EFL (Cocquyt et al. 2009). In the case of the Ulvophyceae, a more detailed characterization of the distribution of EFL revealed vertical inheritance of both proteins, one in each of the two main ulvophycean lineages. This striking congruence with accepted ulvophycean relationships suggests the presence of EFL or EF-1α may be useful for inferring

relatedness in less well-resolved areas of the tree and in placing uncertain taxa (Gile, Novis, et al. 2009). The presence of EF-1α but not EFL in the draft genome sequence of "Chlorella vulgaris" C-169 (likely a Coccomyxa species) hints that an improved circumscription of the Trebouxiophyceae may also be aided by this character.

Overall, no strong evidence for lateral gene transfer was found in any of the three studies of EFL in the green algae, and differential loss seems likely to have played a major role in shaping the distribution of EFL and EF-1α. This conclusion is only slightly weakened by uncertainty in the organismal phylogeny and a lack of resolution in the phylogeny of EFL (Cocquyt et al. 2009; Gile, Novis, et al. 2009). Although a history of differential loss would appear to be in conflict with the mutually exclusive distribution of EFL and EF-1α, the general congruence of this distribution with major taxonomic groups suggests that the necessary period of co-occurrence began near the origin of the green lineage and may have ended shortly after the deepest divergence in the Ulvophyceae. Thus, the period of co-occurrence may have been brief compared to the subsequent period of lineage sorting that led to today's mutually exclusive distribution.

Evidence for lateral transfer of EFL in chromalveolates?

EFL is unusually common among chromalveolates. Since its initial description, EFL's known distribution has expanded from haptophytes and dinoflagellates to include cryptomonads (Gile, Patron, and Keeling 2006; Sakaguchi et al. 2009) and diatoms (Kamikawa, Inagaki, and Sako 2008), while Apicomplexa and ciliates are known to encode EF-1α (Figure 2.2). Interestingly, of the two complete diatom genomes, *Thalassiosira pseudonana* encodes both proteins but expresses only EFL, while *Phaeodactylum tricornutum* encodes and expresses only EF-1α. In all other diatoms tested, only EFL could be detected from both genomic DNA and cDNA (Kamikawa, Inagaki, and Sako 2008). Thus EFL is known to be present in four out of six major lineages of chromalveolates.

In all three studies of EFL in this supergroup, its complex distribution has been attributed to lateral gene transfer. In two of the three studies, the inference of lateral gene transfer was based on the failure of chromalveolate EFL sequences to group together (Gile, Patron, and Keeling 2006; Sakaguchi et al. 2009). Cryptomonad, haptophyte, perkinsid, and dinoflagellate EFL sequences branch separately, despite the now accepted sister relationship between cryptomonads and haptophytes and the close relationship of *Perkinsus marinus* to dinoflagellates (Saldarriaga et al. 2003; Rice and Palmer 2006; Patron, Inagaki, and Keeling 2007). EFL

subsequently characterized from the non-photosynthetic cryptomonad *Goniomonas amphinema* failed to group even with other cryptomonads (Sakaguchi et al. 2009), though the long branch associated with this sequence calls its placement into question. Both studies concluded that the phylogeny of EFL is inconsistent with a single origin of EFL in chromalveolates, and therefore that the complex distribution was likely due to multiple lateral gene transfers, though the number of independent transfers and their donor lineages could not be determined.

In the third study, certain diatoms were found to encode EFL. The phylogenetic analysis of EFL recovered a supported sister relationship between diatoms and the foraminiferan *Planoglabratella opercularis*, with the chlorarachniophyte *B. natans* branching at the base. This topology was interpreted as direct evidence for lateral gene transfer of EFL from foraminifera to diatoms (Kamikawa, Inagaki, and Sako 2008). However, there are three main difficulties with this interpretation. First of all, other phylogenetic analyses of EFL instead group chlorarachniophytes with foraminifera (Gile and Keeling 2008; Gile, Faktorová, et al. 2009), or root the EFL tree on the branch between chlorarachniophytes and diatoms (Cocquyt et al. 2009), weakening the support for a specific relationship between diatom and foraminiferan EFL. While these conflicting phylogenetic analyses do not remove the possibility that lateral transfer occurred, they remove some of the certainty with which the donor lineage can claim to have been identified. Another issue is inadequate taxon sampling. The discovery of EFL and EF-1α sequences in ESTs from oomycete species (pers. obs.) points to a major role for differential loss in this group. In both EFL and EF-1α phylogenies, oomycete and diatom taxa group together, indicating that both were vertically inherited from an ancestor that encoded both (see Figure 3.2 for EF-1α, EFL not shown). The common ancestor of oomycetes and diatoms is also the ancestor of all other photosynthetic heterokonts according to a current phylogenetic hypothesis (Cavalier-Smith and Chao 2006; Riisberg et al. 2009). While this does not preclude EFL's introduction by lateral transfer, it requires that EFL was transferred not to diatoms, but to the common ancestor of diatoms and oomycetes, thereby providing stronger evidence for differential loss. The third issue, one that is not consistent with lateral gene transfer, is the possibility Rhizaria are related to heterokonts and alveolates, as recent phylogenomic analyses have proposed (Burki et al. 2007; Hackett et al. 2007; Burki, Shalchian-Tabrizi, and Pawlowski 2008; Hampl et al. 2009; Minge et al. 2009). If these phylogenies are accurate, the close relationship between diatom and foraminiferan EFL is due to vertical inheritance rather than lateral gene transfer.

Altogether, evidence for lateral gene transfer in chromalveolates has been weak. In the

case of diatoms, the inferred lateral transfer may actually be a case of vertical inheritance. In the other two surveys, lateral transfer was inferred because the phylogeny of EFL failed to unite chromalveolates and because the mutually exclusive distribution of EFL and EF-1α made a long period of co-occurrence seem unlikely (the presence of both genes in *T. pseudonana* was unknown at the time). However, new evidence for co-occurrence and differential loss of the two genes weakens the case against differential loss. Furthermore, no single gene has yet recovered chromalveolate monophyly, so the null hypothesis that vertical inheritance would be reflected in the phylogeny of EFL is dubious. Although the contribution of lateral gene transfer cannot be ruled out, perhaps the prevalence of EFL in this supergroup, especially its presence throughout the cryptomonads and haptophytes, which are thought to comprise the deepest branching lineage, would be better interpreted as an indication that the ancestor of chromalveolates encoded EFL.

Prospects for future work

Despite these advances in our understanding of the distribution of EFL and EF-1α, many groups remain to be investigated. For example, EFL is known to occur in a complex distribution in opisthokonts, where it is present in certain ichthyosporids, choanoflagellates, and chytrid and zygomycete fungi (Keeling and Inagaki 2004; James et al. 2006; Ruiz-Trillo et al. 2006; Marshall et al. 2008). The fungal EFL sequences were discovered as part of the fungal tree of life initiative to create a taxon-rich multi-gene phylogeny of fungi (James et al. 2006). EFL-encoding fungal lineages are intermingled with EF-1α-encoding lineages at the base of the fungal tree, such that their common ancestor is the ancestor of all fungi. Fungi are monophyletic in both EFL and EF-1α phylogenies, so the authors concluded that EFL and EF-1α were both present in this ancestor and differentially lost in each lineage, except for one zygomycete, *Basidiobolus ranarum*, that retains both genes (James et al. 2006). The choanoflagellates and ichthyosporids have been less well sampled and never formally investigated, but EFL and EF-1α sequences are known from different species in both groups (Ragan, Murphy, and Rand 2003; Keeling and Inagaki 2004; Ruiz-Trillo et al. 2006; King et al. 2008; Marshall et al. 2008), and phylogenies of both proteins unite these groups with strong support (see Figure 3.4 for EFL, EF-1α not shown). As with the fungi, ichthyosporids and choanoflagellates branch sequentially at the base of the animal lineage, and thus their ancestor would also be the ancestor of all animals (Carr et al. 2008; Shalchian-Tabrizi et al. 2008). Differential loss thus appears to be the major contributor to EFL's punctate distribution at the base of both of the two main opisthokonts lineages, which raises the possibility that both genes are ancestral in the entire supergroup, but

more detailed phylogenetic analyses will be needed to confirm this hypothesis. Similar to the case in chromalveolates and Euglenozoa, however, the fungal-lineage and animal-lineage EFL groups branch separately, so there is no phylogenetic support for this last possibility. Certainly the distribution of EFL and EF-1α in the opisthokonts deserves further study.

Several EFL sequences are also present in public marine metagenomic data. A few of these sequences branch with previously characterized EFL sequences, such as those of the prasinophytes *Ostreococcus* and *Micromonas*, but others form distinct clades that have yet to be identified. These sequences likely derive from picoplanktonic eukaryotes, as the metagenomic data comes from surface seawater filtered to 2μm (Venter et al. 2004). Marine picoeukaryotes are quite diverse, and therefore these sequences could belong to any of a number of potential lineages, though the supported relationship with diatoms and oomycetes suggests that they might be one of several known picoplanktonic lineages of heterokonts (Not et al. 2007). Relatives or smaller life stages of foraminifera may also be present. If the grouping of heterokont, foraminiferan, and chlorarachniophyte EFL is, in fact, indicative of vertical inheritance, then it represents one of the deepest relationships recovered by the phylogeny of EFL. It would therefore be particularly informative to determine the source of the environmental EFL lineages. If they belong to other heterokonts and rhizarians, this would further support both the possibility that EFL was vertically inherited and the possibility that Rhizaria belong in the chromalveolates. However, these lineages may belong to unrelated organisms, which could instead provide the first strong evidence for lateral gene transfer of EFL. In either case, the environmental EFL sequences represent a promising avenue of future research.

Conclusions and broader significance

A surprising outcome of this work has been the emergence of evidence that differential loss of EFL and EF-1α has contributed to the complex distribution of these proteins. The strongest published evidence for differential loss has come from a relatively young group, the kinetoplastids. In general, the phylogeny of EFL is unable to provide strong evidence for either lateral gene transfers or differential loss at the broadest taxonomic scales, and therefore the evolutionary history of EFL at the supergroup level is likely to remain a matter of speculation. Nonetheless, with the detection of likely cases of differential loss, a predominantly vertical inheritance of EFL from an ancient paralogy in the ancestor of eukaryotes has become a possibility worth considering. This scenario is approached by the potential co-occurrence of EFL and EF-1α in both the ancestor of green algae and the ancestor of opisthokonts, representing

two of earliest splits in eukaryote evolution (Berney and Pawlowski 2006). Of course, at this deepest level, it would be equally parsimonious to posit a single lateral transfer from an ancient opisthokont ancestor to an early green alga (or vice versa), a view that has some support from the phylogeny of EFL.

Given all this uncertainty, the most tangible lessons from the molecular evolution of EFL are the potential pitfalls in inferring lateral gene transfer. The first of these is the crucial importance of taxon sampling. With many more eukaryotic groups now known to encode EFL, the evolutionary distance between them has shrunk and so has the length of time these proteins would have to co-occur in order to produce today's distribution. An even better example comes from the heterokonts, where the discovery of EFL in *Pythium* species turned the conclusion of lateral transfer to diatoms into strong evidence for differential loss in the heterokonts. The second lesson is that inferences of lateral gene transfer are extremely sensitive to the assumed organismal phylogeny. While this should be obvious, given that lateral transfer can only be detected through incongruous gene and species phylogenies, the story of EFL in diatoms provides a particularly striking example. It also serves as a timely reminder that the goal of determining the precise origin and series of evolutionary events leading to the distribution of EFL, as with understanding any evolutionary transition, will be difficult until we have a robustly resolved tree of life.

Literature Cited

Archibald JM, Rogers MB, Toop M, Ishida K, Keeling PJ. 2003. Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigelowiella natans*. Proc. Natl. Acad. Sci. USA 100:7678-7683.

Berney C, Pawlowski J. 2006. A molecular time-scale for eukaryote evolution recalibrated with the continuous microfossil record. Proc. Biol. Sci. 273:1867-1872.

Burki F, Shalchian-Tabrizi K, Pawlowski J. 2008. Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes. Biol. Lett. 4:366-369.

Burki F, Shalchian-Tabrizi K, Minge M, Skjæveland A, Nikolaev SI, Jakobsen KS, Pawlowski J. 2007. Phylogenomics reshuffles the eukaryotic supergroups. PLoS ONE 2:e790.

Carr M, Leadbeater BS, Hassan R, Nelson M, Baldauf SL. 2008. Molecular phylogeny of choanoflagellates, the sister group to Metazoa. Proc. Natl. Acad. Sci. USA 105:16641-16646.

Cavalier-Smith T, Chao EE. 2006. Phylogeny and megasystematics of phagotrophic heterokonts (kingdom Chromista). J. Mol. Evol. 62:388-420.

Cocquyt E, Verbruggen H, Leliaert F, Zechman FW, Sabbe K, De Clerck O. 2009. Gain and loss of elongation factor genes in green algae. BMC Evol. Biol. 9:39.

Gile GH, Keeling PJ. 2008. Nucleus-encoded periplastid-targeted EFL in chlorarachniophytes. Mol. Biol. Evol. 25:1967-1977.

Gile GH, Patron NJ, Keeling PJ. 2006. EFL GTPase in cryptomonads and the distribution of EFL and EF-1alpha in chromalveolates. Protist 157:435-444.

Gile GH, Faktorová D, Castlejohn CA, Burger G, Lang BF, Farmer MA, Lukeš J, Keeling PJ. 2009. Distribution and phylogeny of EFL and EF-1α in Euglenozoa suggest ancestral co-occurrence followed by differential loss. PLoS ONE 4:e5162.

Gile GH, Novis PM, Cragg DS, Zuccarello GC, Keeling PJ. 2009. The distribution of elongation factor-1 alpha (EF-1a), elongation factor-like (EFL), and a non-canonical genetic code in the ulvophyceae: Discrete genetic characters support a consistent phylogenetic framework. J Eukaryot Microbiol, in press.

Gilson PR, Su V, Slamovits CH, Reith ME, Keeling PJ, McFadden GI. 2006. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: Nature's smallest nucleus. Proc. Natl. Acad. Sci. USA 103:9566-9571.

Gilson PR, McFadden GI. 1999. Molecular, morphological and phylogenetic characterization of six chlorarachniophyte strains. Phycol. Res. 47:7-19.

Gould SB, Sommer MS, Kroth PG, Gile GH, Keeling PJ, Maier UG. 2006. Nucleus-to-nucleus gene transfer and protein retargeting into a remnant cytoplasm of cryptophytes and

diatoms. Mol. Biol. Evol. 23:2413-2422.

Hackett JD, Yoon HS, Li S, Reyes-Prieto A, Rummele SE, Bhattacharya D. 2007. Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of Rhizaria with chromalveolates. Mol. Biol. Evol. 24:1702-1713.

Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AG, Roger AJ. 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". Proc. Natl. Acad. Sci. USA 106:3859-3864.

Harper JT, Keeling PJ. 2003. Nucleus-encoded, plastid-targeted glyceraldehyde-3-phosphate dehydrogenase (GAPDH) indicates a single origin for chromalveolate plastids. Mol. Biol. Evol. 20:1730-1735.

Hirakawa Y, Kofuji R, Ishida K. 2008. Transient transformation of a chlorarachniophyte alga, *Lotharella amoebiformis* (Chlorarachniophyceae), with *uidA* and *egfp* reporter genes. J. Phycol. 44:814-820.

Ishida K, Cao Y, Hasegawa M, Okada N, Hara Y. 1997. The origin of chlorarachniophyte plastids, as inferred from phylogenetic comparisons of amino acid sequences of EF-Tu. J. Mol. Evol. 45:682-687.

James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J et al. (72 co-authors). 2006. Reconstructing the early evolution of fungi using a six-gene phylogeny. Nature 443:818-822.

Kamikawa R, Inagaki Y, Sako Y. 2008. Direct phylogenetic evidence for lateral transfer of elongation factor-like gene. Proc. Natl. Acad. Sci. USA 105:6965-6969.

Keeling PJ. 2009. Chromalveolates and the evolution of plastids by secondary endosymbiosis. J. Eukaryot. Microbiol. 56:1-8.

Keeling PJ, Inagaki Y. 2004. A class of eukaryotic GTPase with a punctate distribution suggesting multiple functional replacements of translation elongation factor 1alpha. Proc. Natl. Acad. Sci. USA 101:15380-15385.

Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, Roger AJ, Gray MW. 2005. The tree of eukaryotes. Trends Ecol. Evol. 20:670-676.

King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I et al. (36 co-authors). 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature 451:783-788.

Marshall WL, Celio G, McLaughlin DJ, Berbee ML. 2008. Multiple isolations of a culturable, motile ichthyosporean (Mesomycetozoa, Opisthokonta), *Creolimax fragrantissima* n. gen., n. sp., from marine invertebrate digestive tracts. Protist 159:415-433.

Minge MA, Silberman JD, Orr RJ, Cavalier-Smith T, Shalchian-Tabrizi K, Burki F, Skjæveland A, Jakobsen KS. 2009. Evolutionary position of breviate amoebae and the primary

eukaryote divergence. Proc. Biol. Sci. 276:597-604.

Noble GP, Rogers MB, Keeling PJ. 2007. Complex distribution of EFL and EF-1alpha proteins in the green algal lineage. BMC Evol. Biol. 7:82.

Not F, Gausling R, Azam F, Heidelberg JF, Worden AZ. 2007. Vertical distribution of picoeukaryotic diversity in the Sargasso Sea. Environ. Microbiol. 9:1233-1252.

O'Kelly CJ, Wysor B, Bellows WK. 2004. Gene sequence diversity and the phylogenetic position of algae assigned to the genera *Phaeophila* and *Ochlochaete* (Ulvophyceae, Chlorophyta). J. Phycol. 40:789-799.

Patron NJ, Inagaki Y, Keeling PJ. 2007. Multiple gene phylogenies support the monophyly of cryptomonad and haptophyte host lineages. Curr. Biol. 17:887-891.

Patron NJ, Waller RF. 2007. Transit peptide diversity and divergence: A global analysis of plastid targeting signals. Bioessays 29:1048-1058.

Ragan MA, Murphy CA, Rand TG. 2003. Are Ichthyosporea animals or fungi? Bayesian phylogenetic analysis of elongation factor 1alpha of *Ichthyophonus irregularis*. Mol. Phylogenet. Evol. 29:550-562.

Rice DW, Palmer JD. 2006. An exceptional horizontal gene transfer in plastids: Gene replacement by a distant bacterial paralog and evidence that haptophyte and cryptophyte plastids are sisters. BMC Biol. 4:31.

Riisberg I, Orr RJ, Kluge R, Shalchian-Tabrizi K, Bowers HA, Patil V, Edvardsen B, Jakobsen KS. 2009. Seven gene phylogeny of heterokonts. Protist 160:191-204.

Ruiz-Trillo I, Lane CE, Archibald JM, Roger AJ. 2006. Insights into the evolutionary origin and genome architecture of the unicellular opisthokonts *Capsaspora owczarzaki* and *Sphaeroforma arctica*. J. Eukaryot. Microbiol. 53:379-384.

Sakaguchi M, Takishita K, Matsumoto T, Hashimoto T, Inagaki Y. 2009. Tracing back EFL gene evolution in the cryptomonads-haptophytes assemblage: Separate origins of EFL genes in haptophytes, photosynthetic cryptomonads, and goniomonads. Gene, in press.

Saldarriaga JF, McEwan ML, Fast NM, Taylor FJ, Keeling PJ. 2003. Multiple protein phylogenies show that *Oxyrrhis marina* and *Perkinsus marinus* are early branches of the dinoflagellate lineage. Int. J. Syst. Evol. Microbiol. 53:355-365.

Shalchian-Tabrizi K, Minge MA, Espelund M, Orr R, Ruden T, Jakobsen KS, Cavalier-Smith T. 2008. Multigene phylogeny of Choanozoa and the origin of animals. PLoS ONE 3:e2098.

Silver TD, Koike S, Yabuki A, Kofuji R, Archibald JM, Ishida K. 2007. Phylogeny and nucleomorph karyotype diversity of chlorarachniophyte algae. J. Eukaryot. Microbiol. 54:403-410.

Sommer MS, Gould SB, Lehmann P, Gruber A, Przyborski JM, Maier UG. 2007. Der1-mediated

preprotein import into the periplastid compartment of chromalveolates? Mol. Biol. Evol. 24:918-928.

Van de Peer Y, Rensing SA, Maier UG, De Wachter R. 1996. Substitution rate calibration of small subunit ribosomal RNA identifies chlorarachniophyte endosymbionts as remnants of green algae. Proc. Natl. Acad. Sci. USA 93:7732-7736.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W et al. (x co-authors. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66-74.

Watanabe S, Kuroda N, Maiwa F. 2001. Phylogenetic status of *Helicodictyon planctonicum* and *Desmochloris halophila* gen. et comb, nov. and the definition of the class Ulvophyceae (Chlorophyta). Phycologia 40:421.