SYMBIOSIS AND ITS IMPACT ON EUKARYOTE EVOLUTION

by

Shannon J. Sibbald

Submitted in partial fulfilment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
August 2017

© Copyright by Shannon J. Sibbald, 2017

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Endosymbiosis has had a significant impact on eukaryotic evolution, from various coevolving partnerships to the origin of mitochondria and plastids. Cryptophytes are a lineage of unicellular algae that harbor a red-algal plastid derived from secondary endosymbiosis and belong to a phylum (Cryptista) thought (by some) to be ancestrally non-photosynthetic. Furthermore, Cryptista has traditionally been difficult to place in the eukaryotic tree of life. To investigate Cryptista's relationship to other eukaryotes and the evolution of red-algal complex plastids, I searched for an algal footprint in genomic data from *Goniomonas avonlea*, a close heterotrophic relative of cryptophytes. Overall, a close association of Cryptista to Archaeplastida was revealed – specifically to green/glaucophyte algae – and little evidence supporting red-algal plastid ancestry in *G. avonlea* was found. Additionally, I investigated a novel, recently established, eukaryote-eukaryote endosymbiosis not involving photosynthesis. Characterization of novel isolates and comprehensive phylogenetic analyses on *Paramoeba-Perkinsela* revealed a strong signal for co-evolution.

LIST OF ABBREVIATIONS USED

AA	Amino Acids
AGD	Amoebic Gill Disease
ASW	Artificial Seawater
AU	Approximately Unbiased
BLAST	Basic Local Alignment Search Tool
BMGE	Block Mapping and Gathering with Entropy
CAT	Cross-species Alignment Tool
DIAMOND	Double Index Alignment of Next-generation Sequencing Data
DIC	Differential Interference Contrast
dNTP	Deoxy-nucleoside Triphosphate
dsDNA	Double Stranded DNA
EGR	Endosymbiotic Gene Replacement
EGT	Endosymbiotic Gene Transfer
EST	Expressed Sequence Tag
GTR	General Time Reversible
H_0	Null Hypothesis
H_1	Alternative Hypothesis
IPTG	Isopropyl β-D-1-thiogalactopyranoside
LB	Luria-Bertani Medium
LGT	Lateral Gene Transfer
MAFFT	Multiple Alignment using Fast Fourier Transform
MCMC	Markov Chain Monte Carlo
MCOA	Multiple Co-inertia Analysis
miRNA	Micro-RNA

ML	Maximum Likelihood
MMETSP	Marine Microbial Eukaryote Transcriptome Sequencing Project
MRO	Mitochondria Related Organelle
NR	NCBI Non-redundant Database
OTU	Operational Taxonomic Unit
PCR	Polymerase Chain Reaction
Pi	Nucleotide Diversity
PLO	*Perkinsiella*-like-organism
PMSF	Posterior Mean Site Frequency
PPC	Periplastidal Compartment
rDNA	Ribosomal DNA
SAR	Stramenopile-Alveolata-Rhizaria
SH-aLRT	SH(Shimodaira–Hasegawa)-like approximate likelihood ratio tests
TIC	Translocase of the Inner Chloroplast Membrane
TIM	Translocase of the Inner Membrane
TOC	Translocon on the Outer Chloroplast Membrane
TOM	Translocase of the Outer Membrane
UFboot	Ultra-fast Bootstrap Approximation

ACKNOWLEDGEMENTS

I would like to start out by first and foremost thanking my supervisor John Archibald who took me into his lab almost four years ago when I was just an undergraduate and has mentored me ever since. I cannot express the gratitude I have towards him for all the opportunities he has and continues to provide me with. When I first began, I did not even know what a protist is!

I would also like to thank my supervisory committee, Andrew Roger and Alastair Simpson, who provided ample insight and guidance into how to further my research. Special thanks goes out to Laura Eme (a former post-doc of Andrew Roger) who spent numerous hours leading me through the phylogenetic process and answering every question I could come up with. She also provided many of the scripts that were used or edited for the purpose of these analyses (which I am extremely grateful for).

To all the members of the Archibald lab that have been here over the last few years with me – thank you for helping me out in some way or another! In particular, this work could not have been completed without Ugo Cenci, who spearheaded the *Goniomonas avonlea* genome analyses, Eunsoo Kim (American Museum of Natural History) who generated the genomic/transcriptomic data used throughout my analyses, and Bruce Curtis who provided ample bioinformatic insight. Morgan Colp also deserves special mention for helping me to generate sequence data for a part of the *Paramoeba* project. The *Paramoeba* project would not have been possible without the initial isolation and culturing of novel strains by Charles O'Kelly (Friday Harbor Laboratories) and Yana Eglit, who also aided in microscopy. And finally, thanks goes out to all of the CGEB group for providing feedback on my work in lab meetings and exposing me to a vast variety of research and ideas.

CHAPTER 1 INTRODUCTION

Symbiosis has undoubtedly had a significant impact on the evolution of eukaryotes. Many examples of symbioses exist in nature, varying in kind and combination of prokaryotic and eukaryotic partnerships. Symbiotic relationships are key in promoting evolutionary processes and co-evolution of the partners involved (Lopéz-Garcia et al. 2017). One striking (and well known) example of symbiosis involves coral reefs and the dinoflagellate genus *Symbiodinium*. This relationship is essential to the well-being and survival of coral reefs; the dissociation of the coral with the dinoflagellate symbiont results in a phenomenon known as coral bleaching (i.e., loss of pigmentation) and eventually death of the coral from starvation (Roth 2014). With the genome sequencing of *Symbiodinium kawagutti* by Lin et al. (2015), it was revealed that these two partners are biochemically intertwined. *S. kawagutti* expresses proteins involved in translocating products of photosynthesis out of the cell that are capable of being imported into the coral cells via their corresponding translocators. Other than metabolic exchange, Lin et al. (2015) found evidence for *Symbiodinium* being able to control specific gene expression in the coral using a microRNA (miRNA) based regulatory system. This particular relationship is an example of an endosymbiosis – a specific type of symbiosis that occurs when the symbiont lives inside the host cell. This example, however, shows only a glimpse into the impacts endosymbiosis has had on eukaryotic evolution.

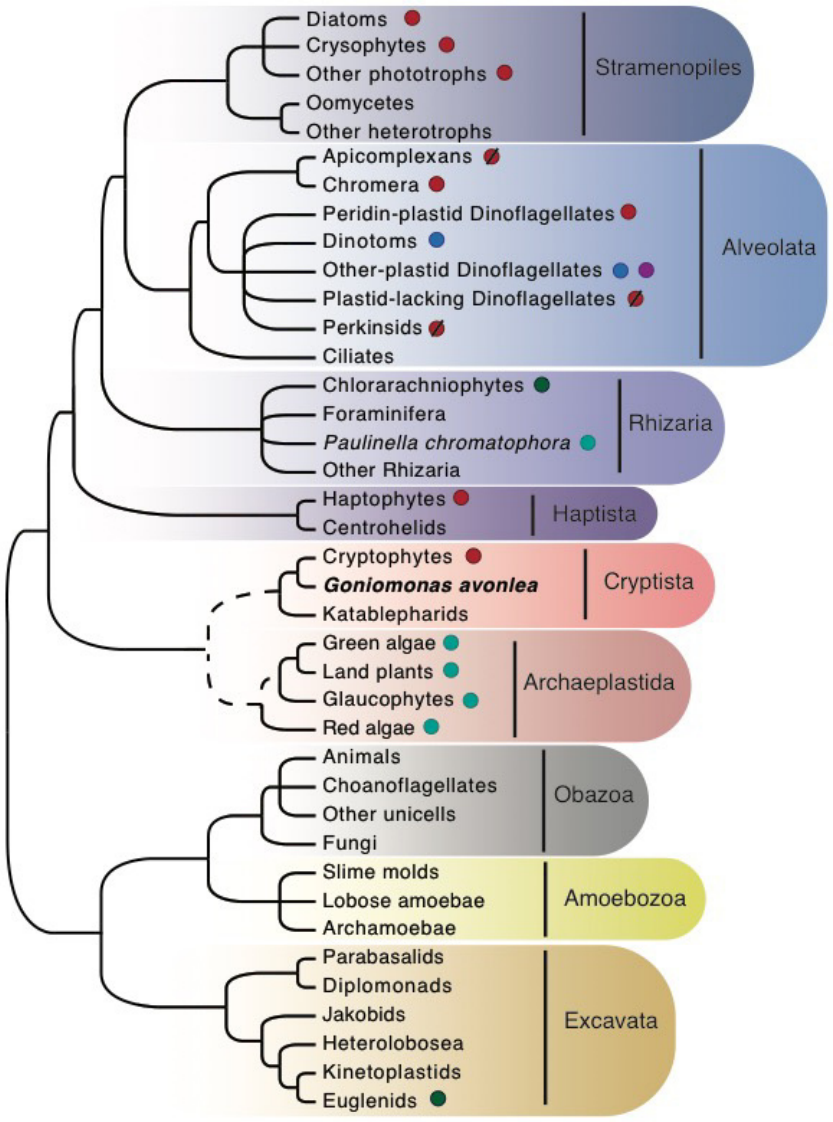
Endosymbiosis sits at the heart of the origin of two well known cellular organelles – mitochondria and chloroplasts (plastids) – and is even thought to be central in the origin of the eukaryotic cell itself (e.g. see Lopéz-Garcia et al. 2017). Mitochondria or mitochondria-related organelles (MROs) are found in all extant eukaryotes with the exception of a single excavate, *Monocercomonoides* sp., where they were secondarily lost (Karnkowska et al. 2016). Mitochondria are involved in a variety of key metabolic tasks within the cell. Although it is most prominently known for its role in energy generation, mitochondria are also involved in fatty acid and amino acid metabolism as well as iron-sulfur cluster assembly. Phylogenomics strongly supports a single origin of mitochondria in a proto-eukaryote host cell from an α-proteobacterium specifically related to the order Rickettsiales (Wang and Wu 2015). Other genetic factors such as a distinct organization

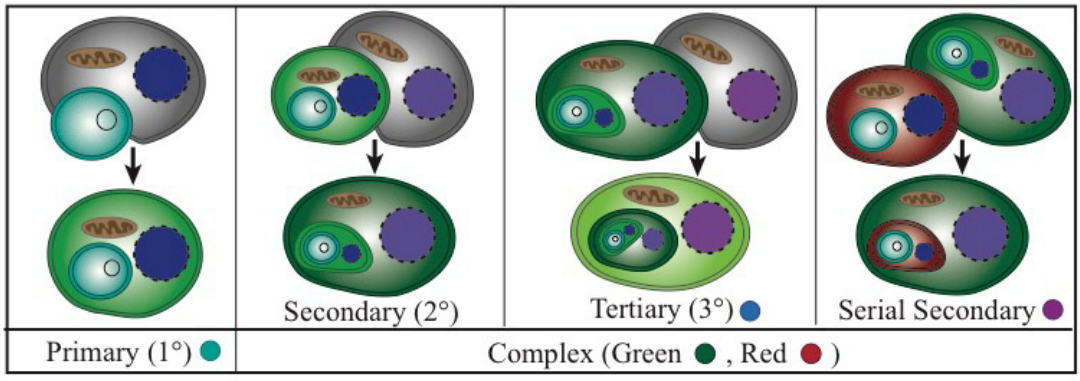
and expression of mitochondrial genes (compared to bacterial genomes; Gray 2015) and similarly reduced genetic repertoires that are a subset of the most gene-rich genomes of Rickettsiales provide further evidence pointing towards a single common origin of this organelle (Gray 2012). While it is widely accepted to be an organelle of endosymbiotic origin, the exact details with regards to when and how mitochondria emerged remain a challenge to determine (Gray 2015).

Unlike mitochondria where a single endosymbiotic event occurred, the origin of photosynthesis in eukaryotes and its spread to a diverse set of taxa widely distributed throughout the eukaryotic tree of life is a result of multiple distinct endosymbiotic events (Figure 1.1). Primary plastids found in Archaeplastida (Glaucophyta, Rhodophyta and Viridiplantae, which includes both green algae and land plants) are thought to have originated via a single primary endosymbiosis approximately 900-1,300 million years ago involving a heterotrophic common ancestor of Archaeplastida and a cyanobacterium (Eme et al. 2014). This has been consistently supported through various forms of phylogenetic and molecular evidence including the monophyly of Archaeplastida in a few nuclear gene trees (Rodriguez-Ezpeleta et al. 2005), monophyly of primary plastids within cyanobacteria in plastid gene trees (Rodriguez-Ezpeleta et al. 2005) and the presence of shared derived features such as an evolved protein targeting and import system (translocase of the inner chloroplast membrane (TIC)/translocase of the outer chloroplast membrane (TOC) complex (Shi and Theg, 2013)) and the presence of unique inverted repeats in plastid genomes that are not found in extant cyanobacteria (McFadden 2001). An additional primary endosymbiotic event estimated to have occurred much more recently (approximately 60 million years ago; Nowack, Melkonian and Glöckner, 2008) involving a different cyanobacterial lineage and a rhizarian, *Paulinella chromatophora*, is the only known exception to the singularity of primary plastids.

Plastids found across the rest of the tree of life are a consequence of higher order endosymbioses involving heterotrophic eukaryotes and eukaryotes with established plastids (Figure 1.1). Secondary plastids found in the distantly related euglenids and chlorarachniophytes originated from two independent endosymbioses involving green

Figure 1.1: Schematic of the eukaryotic tree of life based on Burki et al. (2016) and the distribution of plastids throughout. Type of plastid (primary or complex) is indicated next to a given lineage. While complex plastids of green algal origin are known to have occurred via independent secondary events, the endosymbiotic history of red algal derived plastids is uncertain. Where known, specific complex events of tertiary endosymbioses and serial secondary endosymbioses are shown. Known plastid losses or loss of photosynthesis but presence of a non-photosynthetic version of the organelle are indicated with a line through their plastid circle. A dashed line in the backbone of the tree represents uncertainty in phylogenetic placement of the corresponding lineages.


Diatoms
Crysophytes
Other phototrophs
Oomycetes
Other heterotrophs
Stramenopiles
Apicomplexans
Chromera
Peridin-plastid Dinoflagellates
Dinotoms
Other-plastid Dinoflagellates
Plastid-lacking Dinoflagellates
Perkinsids
Ciliates
Alveolata
Chlorarachniophytes
Foraminifera
Paulinella chromatophora
Other Rhizaria
Rhizaria
Haptophytes
Centrohelids
Haptista
Cryptophytes
Goniomonas avonlea
Katablepharids
Cryptista
Green algae
Land plants
Glaucophytes
Red algae
Archaeplastida
Animals
Choanoflagellates
Other unicells
Fungi
Obazoa
Slime molds
Lobose amoebae
Archamoebae
Amoebozoa
Parabasalids
Diplomonads
Jakobids
Heterolobosea
Kinetoplastids
Euglenids
Excavata


Secondary (2°)
Tertiary (3°)
Serial Secondary
Primary (1°)
Complex (Green , Red)

algae (Archibald 2015). Complex plastids (including those acquired from secondary, tertiary or higher endosymbiotic events) derived from a red alga are found in a wider variety of lineages including the haptophytes, cryptophytes, stramenopiles (including diatoms and brown/golden algae) and alveolates (in both photosynthetic and non-photosynthetic forms) (Archibald 2015) and dominate primary production occurring in the open ocean (Falkowski et al. 2004). Many uncertainties surround red algal complex plastids to this day – where and how many times they were established and how they were horizontally spread is unclear.

Why has the evolutionary history of red-algal complex plastids been so difficult to discern? In part, this is due to conflicting phylogenetic signals created by complicated events in plastid evolutionary history such as plastid replacement and cryptic plastid loss. The dinoflagellates are a phylum that exemplify a great deal of conceivable plastid possibilities (as can be seen in Figure 1.1). While most photosynthetic dinoflagellate lineages harbor a complex plastid of red-alga origin (also known as peridinin-plastids), the dinotoms have undergone a tertiary endosymbiosis, permanently replacing their photosynthetic organelle with a diatom endosymbiont and repurposing their red-alga plastid as an eyespot (Hehenberger, Imanian and Keeling 2014). Other lineages, the Kareniaceae and *Dinophysis*, have replaced their original plastid with a tertiary one of haptophyte or cryptophyte origin, respectively (Tengs et al. 2000; Hackett et al. 2003). There are even instances where the red-algal derived plastid has been replaced by a secondary plastid of green algal origin (*Lepidodinium*; Matsumoto et al. 2011) and where there is evidence for plastid loss entirely (*Hematodinium* sp.; Gornik et al. 2015) or the presence of a cryptic plastid (*Crypthecodinium cohnii*; Sanchez-Puerta et al. 2007). Non-photosynthetic taxa like *C. cohnii* create additional complications as it is difficult to distinguish whether plastid loss should be inferred or if the lineage was ancestrally non-photosynthetic (Keeling 2010).

Across the eukaryotic tree of life, photosynthetic lineages are frequently intertwined with non-photosynthetic taxa (as can be seen in Figure 1.1). With plastid loss being difficult to prove, cryptic non-photosynthetic plastids being found in lineages such as the apicomplexans (Waller and McFadden 2005), and examples of plastid replacements

(both serial secondary endosymbioses and tertiary/quaternary endosymbiotic events), it is not surprising that, in the case of red-algal complex plastids, uncertainties prevail. As a result, many hypotheses have been proposed to explain the origin and spread of these complex plastids exist including the chromalveolate hypothesis (Cavalier-Smith 1999) and a variety that involve multiple serial endosymbiotic events (e.g. Sanchez-Puerta and Delwiche 2008; Stiller et al. 2014; Petersen et al. 2014; Boydł 2017; see Chapter 3 for details). How do we know the source of an endosymbiont-turned-organelle to begin with? Simply put, it is 'in the DNA'. What is clear is that endosymbiosis results in substantial change to the genomic landscape of both the host and their endosymbiont as the two become integrated and intertwined; a lot can be learned from the genetic footprint that remains within a modern plastid's genome and the genes transferred to the host's nucleus during endosymbiont incorporation (e.g. Curtis et al. 2012).

As an endosymbiont transitions into an organelle, it becomes reduced at both the cellular and genomic levels and grows increasingly integrated with and dependent upon its host (Keeling 2013). This typically includes loss of the endosymbiont's organelles, the transfer of genes from the endosymbiont to the host nucleus (endosymbiotic gene transfer (EGT)) and loss of redundant genes (e.g., those whose protein products can be produced by the host and utilized by the endosymbiont). Hand in hand with this is the development of regulatory elements and some kind of targeting and import system to allow the protein products of EGTs to be targeted back to the endosymbiont, cross its surrounding membranes and function appropriately (Keeling 2013). In the case of mitochondria, this is accomplished using a mitochondrial targeting peptide and the TIM/TOM complex (translocase of the inner and outer membrane; Dolezal et al. 2006). Primary plastids utilize a similar mechanism involving a different targeting signal (transit peptide) and TIC/TOC complex (e.g., Archibald 2015). In the case of higher order plastid endosymbioses, additional membranes that belong to the host's endomembrane system need to be crossed. In these cases, an additional targeting signal, a signal peptide, is required to cross the outermost membrane(s) into the periplastidal compartment (PPC) and a transit peptide for the protein product to enter the plastid (e.g., Archibald 2015).

An endosymbiont's genome is typically characterized as highly reduced, fast

evolving and rich in adenosine and thymidine nucleotides (there are, of course, exceptions; Smith and Keeling 2015). Due to gene loss and gene transfer, present day mitochondrial genomes encode only a fraction of the genes that would have been present in its free-living α-proteobacterial ancestor. In humans and most metazoa, only 37 genes are encoded by a single circular mitochondrial DNA – in protists the size and structure of mitochondrial genomes vary (Gray et al. 1998), however, having nuclear mitochondrial DNA is a common feature in all eukaryotes (Huang et al. 2004). Plastid genomes experienced a similar fate. The cyanobacterium-derived endosymbiont present in photosynthetic eukaryotes has a highly reduced genome containing ~70-200 (at most) of the 2000-3000 typically found in its free-living cyanobacterial counterparts (McCutcheon and Moran 2011). Complex plastids have undergone an additional layer (or layers) of endosymbiosis and subsequent reduction and integration with its host. In most instances, this has involved additional gene transfers from the primary algal nucleus to the point of completion and loss of this nucleus. In the cryptophytes and chlorarachniophytes, however, a highly reduced remnant of the primary alga nucleus remains (the nucleomorph; Curtis et al. 2012) that depends upon a substantial number of nucleus encoded proteins to function.

Why do organelles lose genetic material in the first place? Selosse et al. (2001) proposed that it could be because of the high cost of maintaining an organelle genome from a biochemical standpoint, making out-sourcing the maintenance to the host beneficial. Another hypothesis put forth by Allen and Raven (1996) suggests that the nucleus is a less toxic environment for DNA as free radicals produced during the normal function of the mitochondria and plastids can have a negative impact on organelle genomes. Whether EGT is a cost-cutting mechanism, a safe-guard for the organelle genome, or occurs for other reasons is not yet obvious (Daley and Whelan 2005). It is generally thought that the initial wave of EGT occurred relatively quickly (Martin 2003) and that varying amounts of DNA from mitochondria and plastids are continually being transferred to the nucleus even today (Richly and Leister 2004). Furthermore, the process of EGT is believed to be random in that transfers can include partial coding or non-coding regions up to entire genes (and even whole chromosomes in *Arabidopsis* (Lin et al. 1999)) and genetic material is not deliberately selected for transfer to the host nucleus (Leister 2005).

EGT that occurs during integration of the endosymbiont with the host cell offers vast evolutionary possibilities to create innovations from the introduction of new genes and duplicate genes into the host nucleus. Not only are these genes used to maintain the endosymbiont/organelle and perform their function there, they can also acquire targeting signals to other locations in the cell or function in the host cytosol, promoting an environment of mix-and-match biochemistry (Archibald 2015). As such, EGTs can acquire new roles in host cell biochemical pathways either by replacing a host gene with a redundant function or by gene duplication that results in a recently diverged paralog with a new function (Archibald 2015). Organisms with complex plastids derived from higher order endosymbiotic events have massive amounts of gene transfer (EGT) resulting in a chimera of genes present in the host nucleus from a variety of sources with very different evolutionary origins and histories. As a result, EGT can make phylogenies of different genes difficult to interpret and potentially discordant with one another. The higher the order of the endosymbiotic event involved and the more closely related the partners, the more complicated it becomes to detect and disentangle the source of the gene transfer (Archibald 2015).

In Chapter 2, I discuss a novel eukaryote-eukaryote endosymbiosis that does not involve a photosynthetic endosymbiont – the only known example of this kind. *Paramoeba* species (a lobose amoebae) exist in a stable, obligate endosymbiotic relationship with the distantly related excavate, *Perkinsela* sp. (Tanifuji et al. 2011). Genomic evidence in *Perkinsela* sp. such as significant genome reduction suggests that this system is an example of an endosymbiont in the transition to becoming a full-fledged organelle (Tanifuji et al. 2011). While this makes it a potentially useful model to study the process of how an endosymbiont becomes an organelle, many uncertainties remain regarding the nature of this relationship. By characterizing numerous novel isolates and investigating the probable coevolution occurring between *Paramoeba* spp. and *Perkinsela* sp. (e.g., Young et al. 2014) I aimed to shed some evolutionary light on the genus and endosymbiotic association.

In Chapter 3, I examine the evolution of complex plastids of red-algal origin and the super-group Cryptista. By searching for an algal footprint in genomic data obtained from *Goniomonas avonlea*, a member of the closest heterotrophic relatives to the

photosynthetic cryptophytes (the goniomonads), I attempt to pinpoint the acquisition of secondary plastids within Cryptista. Additionally, evidence of plastid loss or lack thereof in the goniomonads will hopefully allow for a more accurate annotation of EGT in cryptophytes such as *Guillardia theta* (see Curtis et al. 2012) and provide further insight into how plastid endosymbiosis alters the genome and biochemistry of its host. Here I also present a phylogenomic investigation of the global eukaryotic position of Cryptista and begin to look for multiple signals emerging from a marker-gene dataset. Finally, in Chapter 4 I briefly summarize the conclusions of Chapters 2 and 3 and show how symbiosis, literally meaning 'living together', has and continues to shape eukaryote evolution.

CHAPTER 2 DIVERSITY AND EVOLUTION OF *PARAMOEBA*

This chapter includes work published in Sibbald, S. J., Cenci, U., Colp, M., Eglit, Y., O'Kelly, C. J. and Archibald, J. M. (2017), Diversity and Evolution of *Paramoeba* spp. and their Kinetoplastid Endosymbionts. *J. Euk. Microbiol.* doi:10.1111/jeu.12394.

2.1 INTRODUCTION TO *PARAMOEBA*

Paramoeba/*Neoparamoeba* species are small lobose amoebae with dactylopodiate pseudopodia, best known as pathogens of various fish and invertebrates in marine and estuary environments around the world. They are of significant economic and ecological interest because of their association with Amoebic Gill Disease (AGD) in various fishes, paramoebiasis in blue crabs and lobster, and wasting disease in green sea urchins (Caraguel et al. 2007; Dykov a et al. 2005; Feehan et al. 2013; Fiala and Dykova 2003; Lee et al. 2006; Mouton et al. 2014; Sprague et al. 1969; Tanifuji et al. 2011; Young et al. 2007, 2008). Despite the ongoing interest in the study of *Paramoeba*/*Neoparamoeba* spp., the pathobiology of the organism is still poorly understood. This is in part owing to difficulties in isolating and culturing *Paramoeba*/*Neoparamoeba* spp. (Lee et al. 2006).

Historically, the presence or absence of micro-scales on the amoeba cell surface has been used to separate species into either the genus *Paramoeba* or *Neoparamoeba* (Page 1987). Phylogenetic analysis combined with microscopic observations has suggested, however, that micro-scale presence versus absence is not a distinguishing feature between these genera, leading to debate as to whether *Paramoeba* and *Neoparamoeba* are in fact distinct taxa (Feehan et al. 2013; Young et al. 2014). For convenience, I will use *Paramoeba* herein, consistent with the nomenclature used by Feehan et al. (2013) (but nevertheless retain the original names associated with previous GenBank sequence submissions).

Paramoeba cells are typically characterized by the presence of a membrane-bound, nucleus-associated compartment known as the “parasome” (Dykova et al. 2003).

Originally thought to be an organelle, it was not until the 1970s that it gradually became clear that the parasome was actually a eukaryotic cell of endosymbiotic origin (Dykova et al. 2003; Perkins and Castagna 1971). The similarities between the parasome and *Perkinsiella amoeba*, an endosymbiont of *Janickina* amoebae (Hollande 1980), resulted in it being referred to as a *Perkinsiella*-like-organism (PLO), and later, *Perkinsela* sp. (Young et al. 2014). Based on 18S ribosomal DNA (rDNA) phylogenies, *Perkinsela* sp. was shown to be a member of the Kinetoplastea, and particularly closely related to *Ichthyobodo necator*, an ectoparasite of fish (Dykova et al. 2003; Tanifuji et al. 2011). Although all currently known *Paramoeba* spp. possess one or more *Perkinsela* sp. within them, its presence alone cannot be considered as a diagnostic feature of the genus, as endosymbionts have been described in a few other genera and families of amoebae (Dykova et al. 2000).

Perkinsela sp. is an aflagellate body that is surrounded by a single membrane (i.e. it is not surrounded by any host-derived membranes) (Dykova et al. 2003). It is often binuclear with the nuclei localized to opposite poles, and the vast majority of the cell volume is occupied by a single large mitochondrion structurally similar to that of other kinetoplastids (Dykova et al. 2003; Young et al. 2014). *Perkinsela* sp. possesses many other interesting features of kinetoplastids including mitochondrial RNA editing and spliced leader (SL) trans splicing (Tanifuji et al. 2011). No ultrastructural differences between the *Perkinsela* sp. within different amoeba strains or species are obvious under light or transmission electron microscopy (Dykova et al. 2003, 2005, 2008). Likewise, no readily discernable differences between *Paramoeba* strains and species have been observed (Dykova et al. 2000). This has resulted in an increased reliance on molecular methods for species characterization.

The association between *Paramoeba* spp. (at least, those species previously assigned to *Neoparamoeba*) and *Perkinsela* sp. is unique in that it involves two eukaryotes existing in a seemingly stable and obligatory endosymbiosis, a relationship in which photosynthesis does not play a role. This is unlike the eukaryote-eukaryote endosymbiotic events that led to the evolution of "complex" plastids (chloroplasts) in algae such as diatoms and haptophytes (Archibald 2009). Why this particular host-endosymbiont relationship was established is still a mystery. While endosymbionts are known to undergo

significant molecular and cell biological changes as a result of adaptation to intracellular life (e.g. EGT to the host nucleus, massive gene loss because of gene redundancy, and a reduction in genomic G + C content (Timmis et al. 2004)), preliminary investigation of the *Perkinsela* sp. genome relative to free-living kinetoplastids does not suggest extensive genetic integration between *Perkinsela* sp. and its host (Tanifuji et al. 2011). This is indicative of a relatively recent adaptation to intracellularity, the study of which has the potential to shed light on the early stages of reductive evolution and the transition from eukaryotic endosymbiont to organelle. Furthermore, *Paramoeba* spp. and *Perkinsela* sp. have not successfully been cultured separately from one another, and *Perkinsela* sp. is invariably found in close association with the amoebae nucleus (Dykova et al. 2003; Tanifuji et al. 2011).

All together, these facts point towards a stable, obligatory relationship between these two eukaryotes, albeit one in which the benefits to one or both them are unknown. Previous molecular investigations of *Paramoeba* spp. and their endosymbionts have revealed a strong signal for co-evolution suggesting a single endosymbiosis followed by vertical inheritance (e.g. Caraguel et al. 2007; Dykova et al. 2008; Young et al. 2014). To establish a more robust framework for inferring and investigating the evolutionary history and relationships between *Paramoeba* species and *Perkinsela* sp., I expanded the available molecular dataset, particularly for *Perkinsela* sp., by characterizing novel isolates using 18S rDNA. In this chapter, I present 18S characterization of 33 new sequences from *Paramoeba* spp. and 16 sequences from their associated *Perkinsela* sp. from new isolates and an acquired strain of *P. invadens.* Furthermore, I present evidence for coevolution occurring between these two eukaryotes and discuss the observed microheterogeneity in the 18S gene of *Paramoeba* spp..

2.2 METHODS

2.2.1 Cell culturing and DNA isolation

Novel isolates of Amoebozoa were isolated, by Charles J. O'Kelly (Friday Harbor Laboratories), from marine seawater, sediment, and seaweed samples, plated onto natural seawater solidified with 1.5% agar and inspected by microscopy for colonies of amoebae. Three of these isolates (O5, 5G5, KPF3) were obtained from samples collected at Keahole Point on the Big Island of Hawai'i, one novel isolate (FHL) from a tidal mud flat on San Juan Island, Washington, and the final strain (*P. invadens* SMB60) was obtained from a private culture collection (Feehan et al. 2013). Isolates were shipped to Dalhousie University, where I maintained them at room temperature on 1.5% solid Bacto-agar medium prepared with 40% artificial seawater (ASW; 24.72 g/liter NaCl, 0.67 g/liter KCl, 1.364 g/liter CaCl2-2H2O, 4.66 g/liter MgCl2-6H2O, 6.29 g/liter MgSO4-7H2O, and 0.18 g/liter NaHCO3) with the bacterium *Halomonas* sp. (strains O5, 5G5, KPF3, FHL) or *Escherichia coli* (*P. invadens* SMB60) provided as a food source. *Halomonas* sp. was cultured on a similar solid agar using MY100 media (1 g/liter tryptone, 1 g/liter yeast extract and 1 g/liter glucose in 40% ASW) for 24 h at 37 °C and then maintained at room temperature, while *E. coli* was grown on LB media (10 g tryptone, 5 g yeast extract and 5 g NaCl per liter, adjusted to pH 7.0) at 37 °C. Time between subculturing of each isolate varied from 1 to 4 weeks depending on the observed cell density using light microscopy.

Cells were isolated from plates via the addition of ASW to the agar surface and mildly agitating the cultures with light shaking for 30 min. Afterwards, the surface of the agar was scraped and cells suspended in the liquid overlay were collected. Plates were rinsed and subjected to additional shaking until minimal amoebae were observed remaining on the agar surface. For DNA extraction, 10–15 densely covered plates (10 cm diameter) for each isolate were used. Cells were collected by centrifugation at 5,020 g at 8 °C for 10 min. Total genomic DNA was extracted using a standard phenol-chloroform method similar to Lane et al. (2006). The concentration of each DNA sample was determined using a Qubit dsDNA broad-range assay.

2.2.2 Amplification and sequencing of 18S rDNA

Primers were designed to specifically amplify a section of the 18S rDNA gene of *Paramoeba* spp. or *Perkinsela* sp. (PAR-F: 5'- GTAGTATAGAGGACTACCATGGTG-3'; PAR-R: 5'-CACAGACCTGTTATTGCCTCAAA-3'; PLO-F: 5'-CCAAC GAGTATCAATTGGAGGACA-3'; PLO-R: 5'-GGACCTGCTGTTGCCCAAAATGC-3'). PCR reactions were carried out using TaKaRa Ex Taq (Takara Bio USA, Inc., Mountain View, CA) standard protocol (50 µl PCR reactions using 5 µl 10x *Ex Taq* buffer, 4 µl dNTP mixture containing 2.5 mM of each dNTP, 100–200 ng of DNA, 1.0 µM final concentration each of the forward and reverse primer, 0.5 µl of TaKaRa Ex Taq (5 units/µl)) under the following conditions: initial denaturation at 98 °C for 5 min; 35 cycles of denaturation at 98 °C for 10 s, annealing at 57 °C for 30 s and extension at 72 °C for 97 s; and a final extension at 72 °C for 10 min. Universal 18S eukaryotic primers (EukA and EukB (Medlin et al. 1988)) were also used in some instances as they provided longer amplification products. PCR products were analyzed using gel electrophoresis (1% agarose gel) to ensure specificity of amplification. PCR products were then prepared for cloning using a Macherey-Nagel NucleoSpin Gel and PCR Clean-Up Kit as per the manufacturer's protocol; DNA concentrations were determined as above.

Cleaned PCR products were ligated into pGEM-T Vectors (Promega, Madison, WI) following the manufacturer's protocol and incubated overnight at 4 °C. Resulting ligation products were transformed into JM109 High Efficiency Competent Cells (Promega) as per standard protocol. Either 50 or 250 µl of the transformation mixture was plated onto LB plates supplemented with ampicillin (100 µg/ml), IPTG (0.5 mM) and X-gal (80 µg/ml). Plates were left for 24 h at 37 °C before being transferred to 4 °C for 2 h to facilitate color development. Multiple independent white clones per ligation reaction were selected from the LB plates, transferred to 5 ml LB + ampicillin (100 µg/ml) liquid medium and incubated at 37 °C overnight with moderate shaking. The insert-containing plasmids were extracted and purified using a PureYield Plasmid Miniprep Kit (Promega) as per manufacturer's protocol. Incorporation of the expected insert was confirmed using EcoR1 digestion and subsequent agarose gel electrophoresis. Plasmids containing the expected insert were

Sanger sequenced (GeneWiz, South Plainfield, NJ). Resulting 18S rDNA regions were assembled for individual clones and submitted to GenBank under accession numbers shown in Table 2.1.

2.2.3 Phylogenetic analysis

For phylogenetic analysis, 89 *Paramoeba/Neoparamoeba* spp. 18S rDNA sequences were retrieved from GenBank (Table 2.2) along with out-group sequences from the closely related amoeba species *Korotnevella* spp., *Vexillifera armata* and *Pseudoparamoeba pagei*. Alignments including the 33 newly obtained 18S rDNA *Paramoeba* spp. sequences were produced using MAFFT-linsi (multiple alignment using fast Fourier transform; version 7.205; Katoh and Standley 2013) and manually refined. Ambiguously aligned regions were removed using BMGE (block mapping and gathering with entropy; version 1.1; Criscuolo and Gribaldo 2010) with default parameters. The resulting alignment was used to create phylogenies based on maximum-likelihood (ML) methods using IQ-TREE (Version 1.4.3) (Nguyen et al. 2015) with 500 bootstrap replicates under the substitution model GTR+I+G and Bayesian inference using PhyloBayes (version 4.1) (Lartillot et al. 2009). For Bayesian inference, four simultaneous Markov chain Monte Carlo (MCMC) chains were run under the CAT + GTR model with a sample frequency of 0.1 until a pair of chains were determined to converge (maxdif < 0.1). The first 500 sampled generations were discarded as burn-in and posterior probabilities were calculated using majority consensus rule. Phylogenetic analysis of *Perkinsela* sp. was carried out in the same manner, with 44 18S rDNA sequences taken from GenBank (Table 2.2) along with out-group sequences from various *Ichthyobodo* spp. combined with 16 newly obtained 18S rDNA sequences.

To analyze co-evolution between the host amoebae and their *Perkinsela* sp., phylogenetic trees were inferred using the methods described above using a subset of available host 18S sequences for which corresponding *Perkinsela* sp. sequences were available (Table 2.2). A test for co-evolution between host and endosymbiont was performed using ParaFit (Legendre et al. 2002) over 9,999 random permutations with the

Table 2.1. Newly sequenced isolates of *Paramoeba* and their associated *Perkinsela* sp. with corresponding GenBank accession numbers. Multiple clones of each strain were sequenced resulting in 33 new *Paramoeba* (host) and 16 new *Perkinsela* sp. (PLO) 18S rDNA sequences. Clone identification correlates with the primers used to amplify the 18S gene (H-PAR_F/R, U-EukA/B and P-PLO_F/R), where 'H' indicates *Paramoeba* specific, 'P' indicates *Perkinsela* specific, and 'U' indicates universal eukaryotic 18S primers.

Species	Strain	Host Clone	Host Accession	PLO Clone	PLO Accession
P. branchiphila	5G5	H1	KY465840	P1	KY465859
		H2	KY465841		
		H3	KY465842	P2	KY465861
		U1	KY465845		
		U3	KY465843	P3	KY465858
		U4	KY465844		
		U5	KY465846	P4	KY465862
		U6	KY465847		
	KPF3	H1	KY465831	P1	KY465856
		H2	KY465832	P2	KY465857
	O5	H1	KY465836	P1	KY465860
		H2	KY465837		
		H3	KY465838	P2	KY465864
		H4	KY465839		
		U1	KY465833	P3	KY465863
		U2	KY465834		
		U3	KY465835		
P. invadens	SMB60	H1	KY465824	P1	KY465867
		H2	KY465821		
		H3	KY465826	P2	KY465866
		H4	KY465823		
		H5	KY465825	P3	KY465868
		H6	KY465820		
		H7	KY465828	P4	KY465865
		H8	KY465822		
		H9	KY465827		
		H10	KY465830		
		H11	KY465829		
P. pemaquidensis	FHL	H2	KY465848	P1	KY465853
		H3	KY465849	P2	KY465854
		U1	KY465851	P3	KY465855
		U2	KY465852		
		U3	KY465850		

Table 2.2. GenBank accession numbers and strain identification for the 18S rDNA gene used in phylogenetic analysis from *Paramoeba (Neoparamoeba)* spp. and their corresponding *Perkinsela* sp. endosymbiont (when available).

Species	Host Accession	PLO Accession	Strain
P.(N.) branchiphila	AY193724	AY163355	AFSM3
	AY193725	EU331011	SM68
	AY193726	EU331002	SM53
	AY714365	EU331002	ST4N
	AY714366	EU331016	SEDMH1
	AY714367	EU331004	NRSS
	EF675599	EU331029	SU4
	EF675600	EU331028	AMOPI
	EF675601	EU331027	TG1162
	EF675602	EU331026	TG1267
	EF675603	EU331025	RP
	HQ132923	-	DE1A
	HQ132924	-	DE2A
	HQ132925	-	DE3A
	HQ132926	HQ132931	DE4A
	HQ132927	-	DE5B
	HQ132928	-	DE6D
	HQ132929	HQ132932	DE11D
	HQ132930	-	DE5A
P. invadens	KC790384	KC790389	SMB/A12
	KC790385	KC790388	SP/S5
	KC790386	KC790389	SMB/A11
	KC790387	KC790388	SP/S9
P.(*N.) pemaquidensis*	AF371967	EU331005	PA027
	AF371968	-	AVG8194
	AF371969	-	CCAP/1560_4
	AF371970	-	CCAP/1560_5
	AF371971	-	ATCC/50172
	AF371972	-	ATCC/30735
	AY183887	-	ATCC/30735
	AY183889	-	ATCC/50172
	AY193722	EU331031	AFSM2V
	AY193723	EU331032	AFSM11
	AY686577	-	-
	AY686578	-	-

Species	Host Accession	PLO Accession	Strain
P.(*N.) pemaquidensis*	AY714350	EU331010	NETH2T3
	AY714351	EU331014	NP251002
	AY714352	EU331020	GILLNOR1
	AY714353	EU331006	SEDC1
	AY714353	-	SEDC1
	AY714354	EU331007	GILLNOR2
	AY714355	EU331015	ST8V
	AY714356	EU331003	FRS
	AY714357	EU331008	SEDCB1
	AY714358	-	PA027
	AY714359	EU331033	SEDST1
	AY714360	EU331017	SED5A
	AY714361	EU331012	WTUTS
	AY714362	EU331018	SEDCT1
	AY714363	EU331019	NETC1
	AY714364	EU331013	NETC2
	EF675604	EU331024	NET12AFL
	EF675605	EU331023	WT2708
	EF675606	EU331022	GILLRICH3
	EF675607	EU331021	TUN1
	EU331036	EU331034	LITHON
	EU884493	EU884495	PAL2
	EU884494	EU884496	ASL1
P.(N.) perurans	EF216899	EU884498	GD/D1/2
	EF216900	EU884499	GD/D1/3
	EF216902	EU884497	GD/D1/1/1
	EF213903	-	GD/D1/1/2
	EF216901	-	GD/D1/4
	EF216904	-	GD/HAC/2/1
	EF216905	-	GD/HAC/2/2
	EF474477	-	461L1
	EF474478	-	461L4
	EF474479	-	591L1
	EF474480	-	591L3
	GQ407108	-	Chile
	KF146711	-	Isolate_1_4
	KF146712	-	Isolate_5
	KF146713	-	Isolate_6
	KF179520	-	LB200313

Species	Host Accession	PLO Accession	Strain
P. atlantica	JN202436	JN202437	CCAP/1560_9
P. eilhardi	AY686575	-	CCAP/1560_2
	JN202438	-	CCAP/1560_2_10802
	JN202439	-	CCAP/1560_2_10803
	JN202440	-	CCAP/1560_2_10807
	JN202441	-	CCAP/1560_2_10808
P.(N.) aestuarina	AF371973	-	CCAP/1560_7
	AY121848	-	ATCC/50744
	AY121851	-	ATCC/50805
	AY121852	-	ATCC/50806
	AY686574	-	CCAP/1560_7
	DQ229957	-	W4/3
	DQ229958	-	S131/2
	DQ229959	-	SL200
	EU331035	EU331030	SU03

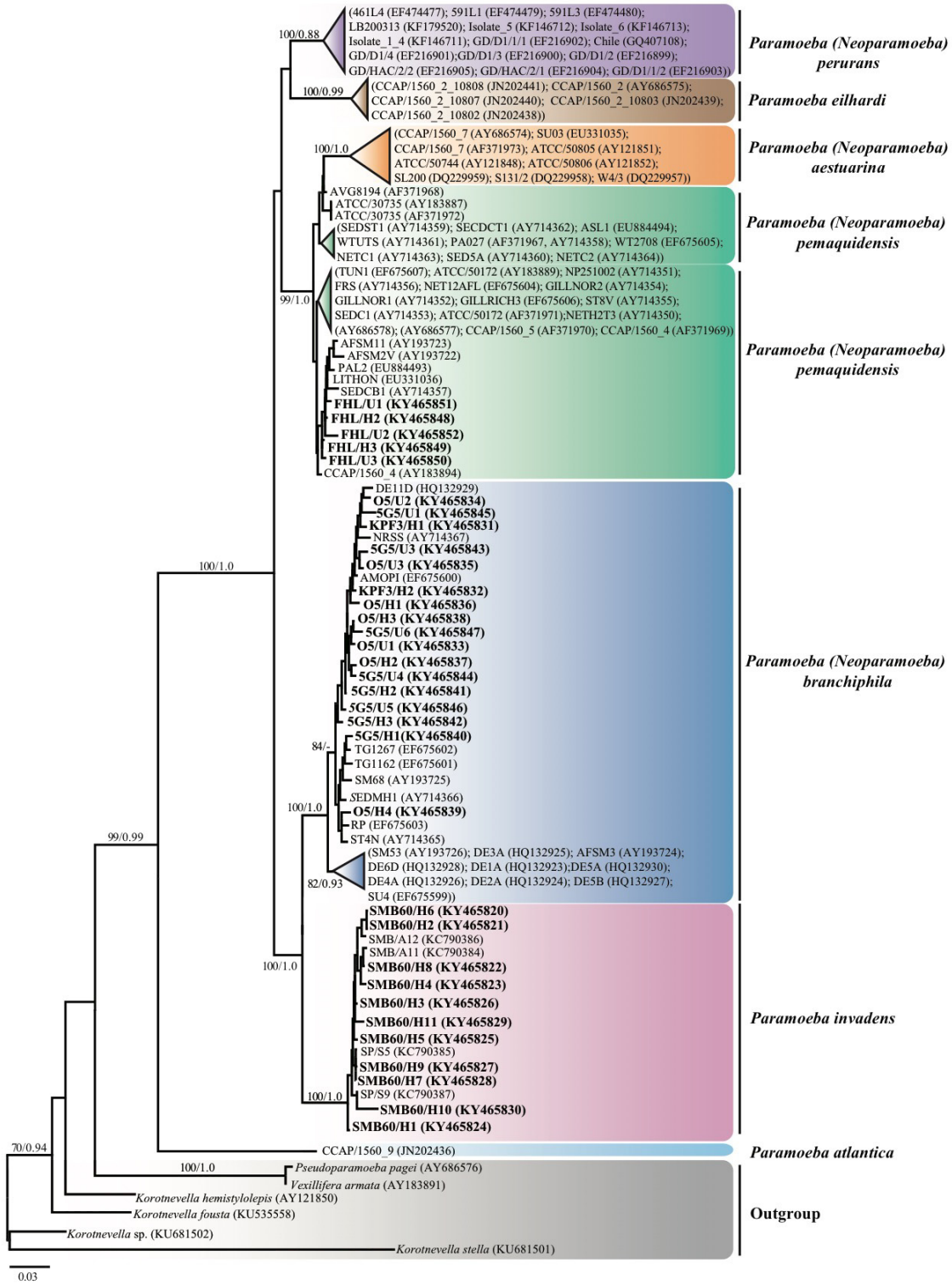
null hypothesis that each *Perkinsela* sp. endosymbiont is randomly associated with a *Paramoeba* spp. host. Alternate tree topologies were tested using the Approximately Unbiased (AU) test (Shimodaira 2002) implemented in Consel (version 1.20) (Shimodaira and Hasegawa 2001). To examine microheterogeneity in the 18S rDNA of isolates, estimations of the nucleotide diversity (Pi) between clones of an isolate (intra-isolate), between isolates (inter-isolate), and within the total data-set were calculated according to equation 10.5 in Nei (1987).

2.3 RESULTS

2.3.1 Sequencing and strain characterization

An overview of 18S rDNA sequences obtained from various host isolates and their *Perkinsela* sp. is shown in Table 2.1. A total of 33 and 16 new 18S rDNA sequences were obtained for five different isolates of *Paramoeba* spp. and their *Perkinsela* sp. endosymbionts respectively. Novel Hawaiian isolates O5, 5G5 and KPF3 were determined to be strains of *P. branchiphila* based on phylogenetic positioning within this species with maximum bootstrap support and posterior probability in ML and Bayesian phylogenies respectively (Figure 2.1). Sequences from all three of the new *P. branchiphila* isolates were found to branch within one subgroup of this species with moderate support (84% bootstrap in ML). While multiple clones of each isolate were sequenced (seven of strain O5, eight of 5G5 and two of KPF3), clone-specific sequences derived from each isolate did not group together in the phylogeny to the exclusion of the other isolates and existing sequences within the *P. branchiphila* sub-group. The other novel isolate, strain FHL from Washington, was determined to be part of the *P. pemaquidensis* clade based on highly supported phylogenetic positioning in the ML and Bayesian inferred phylogenies (Figure 2.1). Five 18S rDNA clones of FHL were sequenced and found to branch close to one another, but not exclusively together. Finally, 11 distinct 18S rDNA clones were sequenced for *P. invadens* strain SMB60. These sequences branched within, and were

Figure 2.1: Maximum likelihood (ML) phylogeny of 18S rDNA sequences of *Paramoeba* spp**.** The phylogeny is based on an alignment of 89 existing and 33 new (highlighted in bold) 18S rDNA sequences and 1667 unambiguously aligned sites. The ML tree is shown with 500 bootstrap replicates (GTR + I + G model) rooted in mid-point with posterior probabilities based on Bayesian inference are mapped onto the ML tree. Multiple clones of the novel isolates KPF3, O5, 5G5, and FHL as well as the *P. invadens* strain SMB60 are shown (highlighted in bold). Out-group sequences are highlighted in gray. Only bootstrap support values > 70% and posterior probabilities > 0.80 are shown. The scale bar shows an inferred 0.06 substitutions per site.


(461L4 (EF474477); 591L1 (EF474479); 591L3 (EF474480);
LB200313 (KF179520); Isolate_5 (KF146712); Isolate_6 (KF146713);
Isolate_1_4 (KF146711); GD/D1/1/1 (EF216902); Chile (GQ407108);
GD/D1/4 (EF216901);GD/D1/3 (EF216900); GD/D1/2 (EF216899);
GD/HAC/2/2 (EF216905); GD/HAC/2/1 (EF216904); GD/D1/1/2 (EF216903))
Paramoeba (Neoparamoeba) perurans
(CCAP/1560_2_10808 (JN202441); CCAP/1560_2 (AY686575);
CCAP/1560_2_10807 (JN202440); CCAP/1560_2_10803 (JN202439);
CCAP/1560_2_10802 (JN202438))
Paramoeba eilhardi
(CCAP/1560_7 (AY686574); SU03 (EU331035);
CCAP/1560_7 (AF371973); ATCC/50805 (AY121851);
ATCC/50744 (AY121848); ATCC/50806 (AY121852);
SL200 (DQ229959); S131/2 (DQ229958); W4/3 (DQ229957))
Paramoeba (Neoparamoeba) aestuarina
AVG8194 (AF371968)
ATCC/30735 (AY183887)
ATCC/30735 (AF371972)
(SEDST1 (AY714359); SECDCT1 (AY714362); ASL1 (EU884494);
WTUTS (AY714361); PA027 (AF371967, AY714358); WT2708 (EF675605);
NETC1 (AY714363); SED5A (AY714360); NETC2 (AY714364))
Paramoeba (Neoparamoeba) pemaquidensis
(TUN1 (EF675607); ATCC/50172 (AY183889); NP251002 (AY714351);
FRS (AY714356); NET12AFL (EF675604); GILLNOR2 (AY714354);
GILLNOR1 (AY714352); GILLRICH3 (EF675606); ST8V (AY714355);
SEDC1 (AY714353); ATCC/50172 (AF371971);NETH2T3 (AY714350);
(AY686578); (AY686577); CCAP/1560_5 (AF371970); CCAP/1560_4 (AF371969)
AFSM11 (AY193723)
AFSM2V (AY193722)
PAL2 (EU884493)
LITHON (EU331036)
SEDCB1 (AY714357)
FHL/U1 (KY465851)
FHL/H2 (KY465848)
FHL/U2 (KY465852)
FHL/H3 (KY465849)
FHL/U3 (KY465850)
CCAP/1560_4 (AY183894)
Paramoeba (Neoparamoeba) pemaquidensis
DE11D (HQ132929)
O5/U2 (KY465834)
5G5/U1 (KY465845)
KPF3/H1 (KY465831)
NRSS (AY714367)
5G5/U3 (KY465843)
O5/U3 (KY465835)
AMOPI (EF675600)
KPF3/H2 (KY465832)
O5/H1 (KY465836)
O5/H3 (KY465838)
5G5/U6 (KY465847)
O5/U1 (KY465833)
O5/H2 (KY465837)
5G5/U4 (KY465844)
5G5/H2 (KY465841)
5G5/U5 (KY465846)
5G5/H3 (KY465842)
5G5/H1(KY465840)
TG1267 (EF675602)
TG1162 (EF675601)
SM68 (AY193725)
SEDMH1 (AY714366)
O5/H4 (KY465839)
RP (EF675603)
ST4N (AY714365)
(SM53 (AY193726); DE3A (HQ132925); AFSM3 (AY193724);
DE6D (HQ132928); DE1A (HQ132923);DE5A (HQ132930);
DE4A (HQ132926); DE2A (HQ132924); DE5B (HQ132927);
SU4 (EF675599))
Paramoeba (Neoparamoeba) branchiphila
SMB60/H6 (KY465820)
SMB60/H2 (KY465821)
SMB/A12 (KC790386)
SMB/A11 (KC790384)
SMB60/H8 (KY465822)
SMB60/H4 (KY465823)
SMB60/H3 (KY465826)
SMB60/H11 (KY465829)
SMB60/H5 (KY465825)
SP/S5 (KC790385)
SMB60/H9 (KY465827)
SMB60/H7 (KY465828)
SP/S9 (KC790387)
SMB60/H10 (KY465830)
SMB60/H1 (KY465824)
Paramoeba invadens
CCAP/1560_9 (JN202436)
Paramoeba atlantica
Pseudoparamoeba pagei (AY686576)
Vexillifera armata (AY183891)
Korotnevella hemistylolepis (AY121850)
Korotnevella fousta (KU535558)
Korotnevella sp. (KU681502)
Korotnevella stella (KU681501)
Outgroup
100/0.88
100/0.99
100/1.0
99/1.0
100/1.0
84/-
100/1.0
82/0.93
100/1.0
99/0.99
100/1.0
70/0.94
100/1.0
0.03

interspersed among, *P. invadens* sequences determined previously (Feehan et al. 2013).

Microscopic characterization using both brightfield and differential interference contrast (DIC) light microscopy confirmed the presence of one or two *Perkinsela* sp. endosymbionts within each of the novel isolates, a subset of which are shown in Figure 2.2. These novel isolates are similar in size, with no obvious features distinguishing between them under the light microscope. Using 18S rDNA sequences obtained from *Perkinsela* sp., ML and Bayesian inferred phylogenies placed *Perkinsela sp.* endosymbionts of strains O5, 5G5 and KPF3 within *Perkinsela* sp. isolated from strains of *P. branchiphila* with maximum support and posterior probability (Fig. 2.3, phylogeny on the right). These sequences fall within the same sub-group of *P. branchiphila* as the nuclear 18S sequences with moderate support, and multiple clones of each isolate (three O5, four 5G5 and two KPF3) generally branch close together. Three clones from *Perkinsela* sp. of isolate FHL were sequenced and found to branch together within other existing sequences from *Perkinsela* sp. of *P. pemaquidensis* strains with high bootstrap support (93%) and moderate posterior probability (0.83). Finally, sequences from four independent clones of *Perkinsela* sp. associated with the *P. invadens* strain SMB60 branched among other *Perkinsela* sp. sequences from *P. invadens* with maximum support and posterior probability.

2.3.2 Co-evolution and alternate tree topologies

Comparison of host and endosymbiont 18S rDNA phylogenies made from a subset of *Paramoeba* spp. strains for which corresponding *Perkinsela* sp. sequences are available largely shows parallel positioning of inter-species relationships that are moderately to highly supported (Figure 2.3). A "host-parasite" coevolution test was performed using ParaFit (Legendre et al. 2002), based on the genetic distances obtained in the ML phylogenies under GTR+I+G with the null hypothesis (H_0) that the host amoebae and endosymbionts are randomly associated and the alternative hypothesis (H_1) that the host and endosymbionts are associated in a fixed, non-random manner. Based on specifying known host-endosymbiont associations, the ParaFit test indicated that there is a non-

Figure 2.2: Microscopic observations using Differential Interference Contrast (DIC) for a sub-set of novel *Paramoeba* spp. isolates. Observations were made using the X100 objective. One to two Perkinsela sp. (caret) as well as the nucleus (solid triangle) are shown. Scale bars represent 5 μm. **(A)** Strain O5, *Paramoeba branchiphila*. **(B)** Strain KPF3, *Paramoeba branchiphila*. **(C)** Strain FHL, *Paramoeba pemaquidensis*.

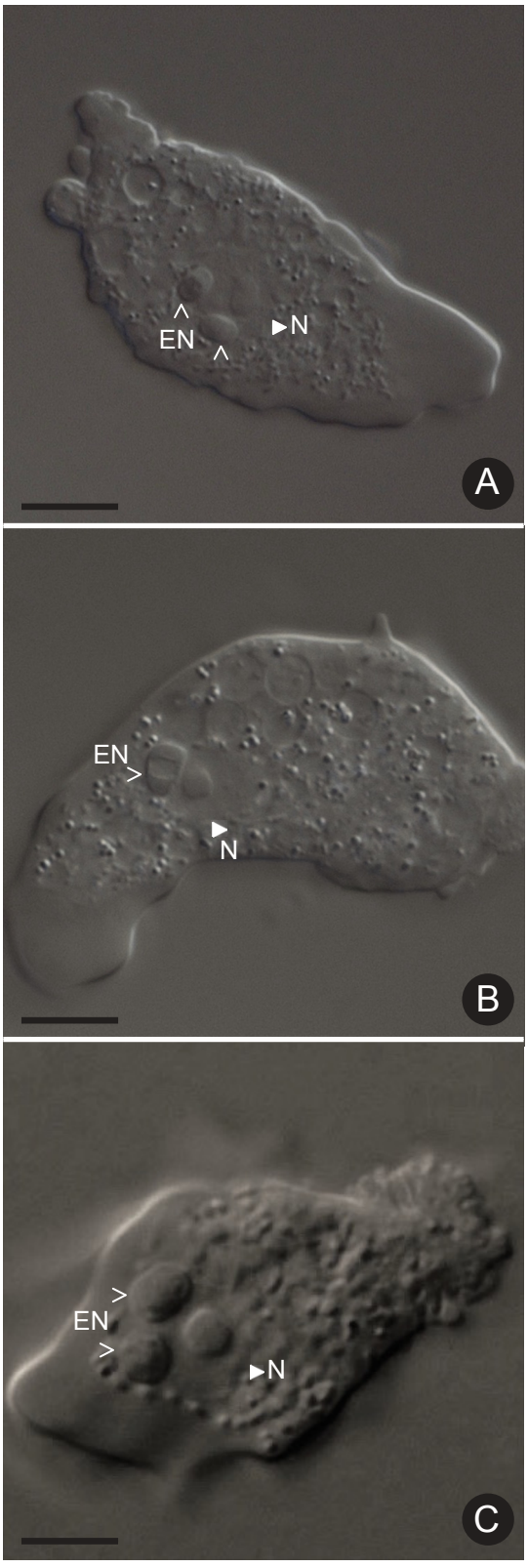
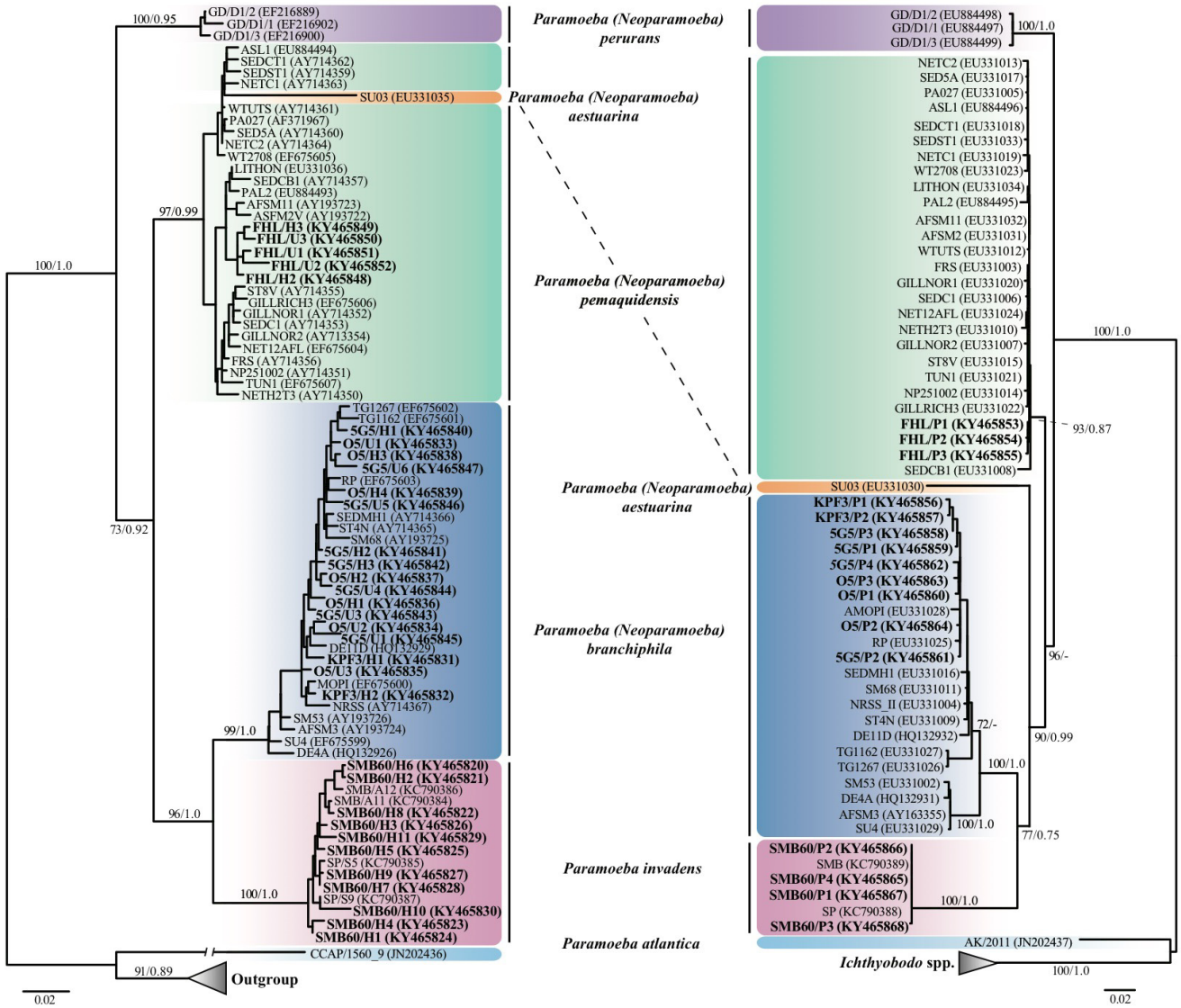

EN
N
A
EN
N
B
EN
N
C

Figure 2.3: Co-evolution analysis using phylogenies of *Paramoeba* spp. strains (left) and their corresponding *Perkinsela* sp. (right) based on 18S rDNA alignments. Strains and novel isolate sequences for which host and *Perkinsela* sp. sequences are available were selected for analysis, resulting in alignments of 79 nuclear 18S sequences (1,275 unambiguous sites) and 60 *Perkinsela* sp. 18S sequences (1,107 unambiguous sites). Maximum-likelihood (ML) trees are shown with 500 bootstrap replicates (GTR + I + G model) rooted at the mid-point with posterior probabilities based on Bayesian inference values mapped onto the ML tree. Multiple clones of the newly cultivated isolates KPF3, O5, 5G5 and FHL, as well as new sequences from *P. invadens* strain SMB60 and their *Perkinsela* sp., are shown in bold. Congruence between these two trees is largely seen, with the exception of the phylogenetic positioning of *(Neo)Paramoeba aestuarina* (shown by a dashed line). Only bootstrap support values > 70% and posterior probabilities > 0.80 are shown. The scale bar shows an inferred 0.02 substitutions per site. The *P. atlantica* branch highlighted by hash marks has been reduced by 50%.


GD/D1/2 (EF216889)
GD/D1/1 (EF216902)
GD/D1/3 (EF216900)
ASL1 (EU884494)
SEDCT1 (AY714362)
SEDST1 (AY714359)
NETC1 (AY714363)
SU03 (EU331035)
WTUTS (AY714361)
PA027 (AF371967)
SED5A (AY714360)
NETC2 (AY714364)
WT2708 (EF675605)
LITHON (EU331036)
SEDCB1 (AY714357)
PAL2 (EU884493)
AFSM11 (AY193723)
ASFM2V (AY193722)
FHL/H3 (KY465849)
FHL/U3 (KY465850)
FHL/U1 (KY465851)
FHL/U2 (KY465852)
FHL/H2 (KY465848)
ST8V (AY714355)
GILLRICH3 (EF675606)
GILLNOR1 (AY714352)
SEDC1 (AY714353)
GILLNOR2 (AY713354)
NET12AFL (EF675604)
FRS (AY714356)
NP251002 (AY714351)
TUN1 (EF675607)
NETH2T3 (AY714350)
TG1267 (EF675602)
TG1162 (EF675601)
5G5/H1 (KY465840)
O5/U1 (KY465833)
O5/H3 (KY465838)
5G5/U6 (KY465847)
RP (EF675603)
O5/H4 (KY465839)
5G5/U5 (KY465846)
SEDMH1 (AY714366)
ST4N (AY714365)
SM68 (AY193725)
5G5/H2 (KY465841)
5G5/H3 (KY465842)
O5/H2 (KY465837)
5G5/U4 (KY465844)
O5/H1 (KY465836)
5G5/U3 (KY465843)
O5/U2 (KY465834)
5G5/U1 (KY465845)
DE11D (HQ132929)
KPF3/H1 (KY465831)
O5/U3 (KY465835)
MOPI (EF675600)
KPF3/H2 (KY465832)
NRSS (AY714367)
SM53 (AY193726)
AFSM3 (AY193724)
SU4 (EF675599)
DE4A (HQ132926)
SMB60/H6 (KY465820)
SMB60/H2 (KY465821)
SMB/A12 (KC790386)
SMB/A11 (KC790384)
SMB60/H8 (KY465822)
SMB60/H3 (KY465826)
SMB60/H11 (KY465829)
SMB60/H5 (KY465825)
SP/S5 (KC790385)
SMB60/H9 (KY465827)
SMB60/H7 (KY465828)
SP/S9 (KC790387)
SMB60/H10 (KY465830)
SMB60/H4 (KY465823)
SMB60/H1 (KY465824)
CCAP/1560_9 (JN202436)
Outgroup
0.02
Paramoeba (Neoparamoeba) perurans
Paramoeba (Neoparamoeba) aestuarina
Paramoeba (Neoparamoeba) pemaquidensis
Paramoeba (Neoparamoeba) aestuarina
Paramoeba (Neoparamoeba) branchiphila
Paramoeba invadens
Paramoeba atlantica
GD/D1/2 (EU884498)
GD/D1/1 (EU884497)
GD/D1/3 (EU884499)
NETC2 (EU331013)
SED5A (EU331017)
PA027 (EU331005)
ASL1 (EU884496)
SEDCT1 (EU331018)
SEDST1 (EU331033)
NETC1 (EU331019)
WT2708 (EU331023)
LITHON (EU331034)
PAL2 (EU884495)
AFSM11 (EU331032)
AFSM2 (EU331031)
WTUTS (EU331012)
FRS (EU331003)
GILLNOR1 (EU331020)
SEDC1 (EU331006)
NET12AFL (EU331024)
NETH2T3 (EU331010)
GILLNOR2 (EU331007)
ST8V (EU331015)
TUN1 (EU331021)
NP251002 (EU331014)
GILLRICH3 (EU331022)
FHL/P1 (KY465853)
FHL/P2 (KY465854)
FHL/P3 (KY465855)
SEDCB1 (EU331008)
SU03 (EU331030)
KPF3/P1 (KY465856)
KPF3/P2 (KY465857)
5G5/P3 (KY465858)
5G5/P1 (KY465859)
5G5/P4 (KY465862)
O5/P3 (KY465863)
O5/P1 (KY465860)
AMOPI (EU331028)
O5/P2 (KY465864)
RP (EU331025)
5G5/P2 (KY465861)
SEDMH1 (EU331016)
SM68 (EU331011)
NRSS_II (EU331004)
ST4N (EU331009)
DE11D (HQ132932)
TG1162 (EU331027)
TG1267 (EU331026)
SM53 (EU331002)
DE4A (HQ132931)
AFSM3 (AY163355)
SU4 (EU331029)
SMB60/P2 (KY465866)
SMB (KC790389)
SMB60/P4 (KY465865)
SMB60/P1 (KY465867)
SP (KC790388)
SMB60/P3 (KY465868)
AK/2011 (JN202437)
Ichthyobodo spp.
0.02
100/0.95
100/1.0
97/0.99
73/0.92
99/1.0
96/1.0
100/1.0
91/0.89
100/1.0
93/0.87
96/-
72/-
90/0.99
100/1.0
100/1.0
77/0.75
100/1.0
100/1.0

random association between *Paramoeba* spp. and *Perkinsela* sp., and that the two phylogenies are generally congruent (P-value < 0.0001).

Tests for individual amoeba-endosymbiont links indicate significant coevolution for the majority of established associations (P-value < 0.05), with the exception of the single *P.(N.) aestuarina* strain (P-value = 0.860), three *P.(N.) pemaquidensis* strains (Lithon, PAL2, and AFSM11) and two *P.(N.) branchiphila* strains (RP and SU4) where the null hypothesis was not rejected (P-value > 0.05). The only incongruence inferred at the inter-species level is the relative position of the single *P.(N.) aestuarina* strain for which both host and endosymbiont sequences were available. However, evaluating alternate tree topologies using the AU test, I could not reject alternative placements of either *P.(N.) aestuarina* or its corresponding *Perkinsela* sp. within the phylogenies. Among the topologies that could not be rejected were the hypotheses that *P.(N.) aestuarina* in fact branches sister to *P.(N.) pemaquidensis* rather than within it and that the *Perkinsela* sp. of *P.(N.) aestuarina* branches within the *Perkinsela* sp. of *P.(N.) pemaquidensis* rather than sister to the *P.(N.) branchiphila* and *P. invadens* clade (P-value > 0.05).

2.3.3 Microheterogeneity in 18S rDNA

Analysis of 18S rDNA from multiple clones of each isolate revealed unexpected levels of variability between clones. For example, pairwise comparisons between 11 clones from *P. invadens* strain SMB60 showed nucleotide differences ranging from 1 to 30 across the ~1,200 bases sequenced, with an average of 15.45 differences per sequence pair. In contrast, nucleotide differences between sequences obtained from clones of the corresponding *Perkinsela* sp. showed minimal variability (0–1 nucleotide differences). To estimate the level of intra-isolate variability in both the *Paramoeba* and *Perkinsela* sp. 18S genes, the average number of nucleotide differences per site was determined between populations using equation 10.5 from Nei (1987) (Table 2.3). The nucleotide diversity (Pi) observed at the intra-isolate level in the novel isolates was much higher within the *Paramoeba* 18S gene compared with the corresponding *Perkinsela* sp. 18 rDNA. This was

Table 2.3. Nucleotide diversity (Pi) within clones of novel isolates (intra-isolate) of *Paramoeba* spp. and *Perkinsela* sp. and within all existing strains. Pi values were determined using equation 10.5 from Nei (1987) and represent the average number of nucleotide differences per site between the population at the isolate level or the total dataset. *P(N). branchiphila* is split into two sub-groups (SG), where SG-1 contains all *P. branchiphila* novel isolates from this study. The number of sequences used to estimate the nucleotide diversity is shown in brackets next to the Pi value. An asterisk (*) signifies that only a single sequence exists in the database.

Species	Strains	Pi *Paramoeba*	Pi *Perkinsela* sp.
P. invadens	SMB60	0.0114 (11)	0.000417 (4)
	SMB	0.0104 (2)	– *
	SP	0.0127 (2)	– *
	All	0.0109 (15)	0.000278 (6)
P (N). branchiphila	O5	0.0214 (7)	0.00666 (3)
	5G5	0.0277 (8)	0.00593 (4)
	KPF3	0.0147 (2)	0.00388 (2)
	Existing	0.0281 (19)	0.0344 (13)
	SG-1 existing	0.0205 (9)	0.0282 (9)
	SG-1 all	0.0230 (26)	0.0241 (18)
	SG-2	0.0230 (10)	0.00976 (4)
	All	0.0281 (36)	0.0344 (22)
P (N). pemaquidensis	FHL	0.0202 (5)	0.00167 (3)
	Existing	0.0221 (35)	0.0118 (24)
	All	0.0229 (40)	0.0110 (27)
All and new and existing *Paramoeba* spp.		0.0744 (122)	0.0732 (60)
All existing *Paramoeba* spp.		0.0462 (89)	0.0436 (44)

also generally observed among all existing strains within each species (i.e. at the intra-species level), while the total dataset showed similar Pi values for both *Paramoeba* spp. and their *Perkinsela* sp. endosymbionts.

2.4 DISCUSSION

The *Paramoeba-Perkinsela* sp. relationship is an endosymbiosis in which the endosymbiont is not photosynthetic, setting it apart from all other known eukaryote-eukaryote endosymbioses (Archibald 2009). In nature, *Paramoeba* spp. have been observed with multiple *Perkinsela* sp. per cell – however, in culture conditions the number of *Perkinsela* sp. tends to decrease over time until only one remains (Dykova et al. 2003). Nevertheless, with the possible exception of the scale-bearing species *P. eilhardi* (see Hollande 1940, Anderson 1977 and Smirnov 1997), *Paramoeba* spp. have always been observed with at least one *Perkinsela* sp. endosymbiont and the two eukaryotes have not successfully been cultured separately. The reason(s) underlying the establishment and persistence of this apparently obligate relationship remains unclear. Characterization of novel isolates has the potential to resolve evolutionary relationships within *Paramoeba* spp. and elucidate the exact nature of this intriguing association. To this end, I generated sequence data from four novel isolates and one existing strain, resulting in 33 and 16 new 18S rDNA sequences for *Paramoeba spp*. and their *Perkinsela* sp., respectively.

2.4.1 Evolutionary relationships of *Paramoeba* spp.

Phylogenies based on 122 18S rDNA sequences from *Paramoeba* spp. (Figure 2.1) show P.*(N.) pemaquidensis* and *P.(N.) aestuarina* forming a distinct clade together that is highly supported under both ML (99% bootstrap support) and Bayesian (1.0 posterior probability) methods. Within this tree*, P.(N.) aestuarina* forms a monophyletic group with maximum

support whose branching location relative to *P.(N.) pemaquidensis* sequences is unclear. Whether *P.(N.) pemaquidensis* is monophyletic or paraphyletic is uncertain. Although strains of this species form multiple, poorly supported clades with *P.(N.) aestuarina* emerging from among them, the AU test could not reject the monophyly of P.*(N.) pemaquidensis* (P-value = 0.470), an alternative topology observed in the phylogenetic analyses of Dykova et al. (2008). In my analysis, *P.(N.) branchiphila* and *P. invadens* form a maximally supported clade (100% bootstrap support, 1.0 posterior probability), with both species forming robust monophyletic groups sister to one another. Furthermore, *P.(N.) branchiphila* appears to form two distinct sub-clades with moderate support, suggesting that perhaps *P.(N.) branchiphila* consists of two different sub-species (as seen in Feehan et al. 2013 and Young et al. 2014).

Paramoeba eilhardi and *P.(N.) perurans* each form separate monophyletic groups with maximum bootstrap support and high posterior probability. While they appear together in the phylogeny as a poorly supported clade sister to *P.(N.) pemaquidensis* and *P.(N.) aestuarina*, a previous study by Feehan et al. (2013) showed these two groups branching sister to the *P. invadens* and *P.(N.) branchiphila* clade with low support. This alternative relationship observed by Feehan et al. (2013) could not be rejected using our dataset with an AU test (P- value > 0.05). It is also possible that *P. eilhardi* and *P.(N.) perurans* do not form a clade, but rather branch separately in the tree, as the placement of either species in various alternative positions could not be rejected by the AU test either. As seen in previous studies such as those of Feehan et al. (2013) and Kudryavtsev et al. (2011), I found *P. atlantica* to branch basal to all other *Paramoeba* spp. with high support. Overall, the evolutionary relationships between these various species remain uncertain. The addition of new 18S sequences to the dataset, particularly in less represented species such as *P. eilhardi*, as well as sequencing multiple 18S copies per strain should aid in interpreting inter-species relationships with greater confidence.

2.4.2 Coevolution of *Paramoeba* spp. and *Perkinsela* sp.

It is generally thought that *Perkinsela* sp. is vertically inherited within *Paramoeba* spp. (e.g., Dykova et al. 2003). Vertical inheritance and the noticeably close association of the amoeba nucleus with *Perkinsela* sp., such as seen in Fig. 2.2, suggests that the host and endosymbiont are indeed evolving in a highly coordinated fashion (Dykova et al. 2003, 2008). Phylogenetic analyses of *Perkinsela* sp. 18S rDNA sequences were found to give similar results to that of the host *Paramoeba* spp. sequences (Figure 2.3). However, complete congruence between the host and endosymbiont 18S rDNA phylogeny is not seen in both our study and in previous studies (Caraguel et al. 2007; Dykova et al. 2008), particularly at the strain level. Complete congruence is perhaps not to be expected, as the resolution between strains of a given species is quite low and the overall relationship between strains of *Paramoeba* spp. remains uncertain. The topology of the host and endosymbiont 18S rDNA phylogenies differ in the placement of one particular species, *P.(N.) aestuarina*; analogous to Dykova et al. (2008), the host phylogeny generated herein shows this species emerging from within *P.(N.) pemaquidensis*, which is not observed in the *Perkinsela* sp. phylogeny where the corresponding endosymbiont of *P.(N.) aestuarina* branches sister to the *P.(N.) branchiphila* and *P. invadens* clade.

With the exception of the positioning of the single *P.(N.) aestuarina* strain, overall inter-species relationships are very similar between the two phylogenies, shown by the parallel positioning of host and endosymbiont 18S rDNA sequences (Figure 2.3). The lack of congruence in this instance does not refute the idea of coevolution between *Paramoeba* and *Perkinsela* sp. The ParaFit test shows that the association of *Perkinsela* sp. and *Paramoeba* spp. is specific rather than random (P-value < 0.0001), supporting the hypothesis that co-evolution is occurring between the host and endosymbiont. This suggests that there was one endosymbiotic event in the common ancestor of all *Paramoeba* spp. and that the endosymbiont has subsequently been vertically inherited. These results are similar to those observed in Caraguel et al. (2007), Dykova et al. (2008), Tanifuji et al. (2011) and Young et al. (2014).

Besides incongruence in the positioning of *P.(N.) aestuarina* and its *Perkinsela* sp.

in Figure 2.3, there are slight differences in topology between this phylogeny and the expanded host phylogeny in Figure 2.1, particularly with respect to the placement of *P.(N.) perurans*. This is most likely because of the inclusion of only a subset of *Paramoeba* spp. strains for which corresponding *Perkinsela* sp. 18S rDNA sequences are available, resulting in no representation of *P. eilhardi* and a reduction of *P.(N.) aestuarina* diversity to a single sequence. Previous studies have noted difficulties in obtaining *Perkinsela sp*. 18S rDNA sequences (e.g., Dykova et al. 2003), resulting in a reduced set of endosymbiont-derived sequences for numerous *Paramoeba* strains. However, to accurately investigate the evolutionary origin of the eukaryotic endosymbiont, characterization of *Perkinsela* sp. from multiple strains across all *Paramoeba* spp. is essential. Data from *P. eilhardi* may be especially important, as there is some evidence that the *Perkinsela* endosymbiosis is not obligate in this species. Anderson (1977) and Smirnov (1997) described amoebae with the scale structure of *P. eilhardi* but without parasomes; the latter assigned the name *Korotnevella nivo* to this entity. Hollande (1940) described what he thought were dispersal stages of the *P. eilhardi* parasome that are consistent with the morphology of small kinetoplastid flagellates including *Ichthyobodo*. Further investigation is needed, particularly in obtaining a larger dataset of *Perkinsela* sp. sequences from *P. eilhardi* and *P.(N.) aestuarina*. This will allow for a more comprehensive comparison of host and endosymbiont phylogenies and biology.

2.4.3 Intra-genomic variability in the 18S rDNA of *Paramoeba* spp.

Sequencing of multiple 18S rDNA clones from individual *Paramoeba* spp. isolates showed high levels of variability (Table 2.3). These results are consistent with those obtained by Feehan et al. (2013). 18S rDNA-based phylogenetic analyses often assume that intra-genomic variability is minimal, but as pointed out by Caraguel et al. (2007), and as can be seen both here and in the study by Dykova et al. (2005), this assumption may not be valid for *Paramoeba* spp. I observed many nucleotide differences between copies of the 18S rDNA amplified from a single isolate, leading to high levels of microheterogeneity. On the other hand, similar levels of variability within an isolate were not detected in 18S rDNA

sequences of *Perkinsela* sp. Within *P. invadens* strain SMB60, the 11 clones of the host 18S rDNA showed 27-fold greater nucleotide diversity compared with the corresponding *Perkinsela* sp. clones. Likewise, nucleotide diversity within the various novel isolates of *P. branchiphila* (strains O5, 5G5 and KPF3) was found to be 3–5 fold greater in the host versus *Perkinsel*a sp. 18S rDNA, and 12-fold greater in the newly obtained *P. pemaquidensis* FHL isolate. Nucleotide diversity between clones of a strain and between strains of a species seems to be similar. For example, isolate FHL (Pi = 0.0202) has a similar nucleotide diversity compared with the entire set of 18S rDNA sequences from the species to which it belongs, i.e. *P. pemaquidensis* (Pi = 0.0229).

Limitations exist in the ability to analyze the extent of microheterogeneity within *Paramoeba* spp. because of the nature of the data that currently exist. The microheterogeneity observed in clones within strains of *P. invadens*, *P. branchiphila* and *P. pemaquidensis* here and in other studies (e.g., Caraguel et al. 2007; Dykova et al. 2005) suggests that strain-related differences in the 18S rDNA sequences existing in the database may not be accurately represented, depending on whether a single sequence or multiple sequences from clones of an isolate were examined. With the high levels of microheterogeneity observed herein, analyzing a single 18S rDNA sequence from a particular isolate may lead to an over or under estimation of nucleotide differences compared with other strains within the database. Furthermore, the presence of single sequences that are unrepresentative of the observed microheterogeneity may impact phylogenetic resolution at the inter-strain level.

Considering these results, the presence of microheterogeneity in the 18S rDNA gene of *Paramoeba* spp. means that while it may be a good marker for identification at the species level, it is of limited utility for strain identification. Perhaps the use of 18S rDNA from *Perkinsela* sp. of *Paramoeba* spp. should be considered for molecular characterization, as it does not appear to exhibit the same levels of microheterogeneity as the host nuclear gene and appears to have co-evolved with the amoebae. This would require the use of specific *Perkinsela* sp. 18S rDNA primers, such as those used in this study, as universal eukaryotic 18S primers have been found to preferentially amplify the *Paramoeba* spp. 18S rDNA, at least in part owing to the significantly lower DNA content of the endosymbiont nucleus relative to the host (Dykova et al. 2003; Tanifuji et al. 2011).

Differences between the 18S rDNA of closely related *Perkinsela* sp. strains, however, may not be great enough to resolve inter-strain relationships. Nevertheless, nucleotide differences at the inter-species level appear sufficiently high to be able to conclusively identify an isolate whether the nuclear or *Perkinsela* sp. gene is used. All things considered, more sequence data from both host and endosymbiont genomes, including protein genes, will be needed to better resolve the evolution of the various candidate species within *Paramoeba* and gain a better understanding of the nature of this unique relationship.

CHAPTER 3 PHYLOGENOMICS OF *GONIOMONAS AVONLEA*

3.1 INTRODUCTION TO *G. AVONLEA*

Goniomonas avonlea, a marine, bacterivorous flagellate isolated from a sandy beach in Prince Edward Island, is a non-photosynthetic, plastid-lacking goniomonad (Kim and Archibald 2013). Goniomonads like *G. avonlea* occupy a key phylogenetic position within the phylum Cryptista as the closest heterotrophic relatives to the photosynthetic cryptophytes (Okamoto et al. 2009) – a position that is crucial to pinpointing the acquisition of complex red algal plastids and understanding their evolution. As mentioned in Chapter 1, little is known for certain with regard to the origin and horizontal spread of complex plastids derived from red algae. Complex red algal plastids are clearly monophyletic (even if no consensus can be reached on the relationships between their host lineages), which points towards red-algal derived complex plastids having a single secondary endosymbiotic origin (Muñoz-Gómez et al. 2017). When they were acquired and how they spread horizontally throughout diverse eukaryotic lineages (see Figure 1.1 for their distribution throughout the tree) is still debated (e.g., Archibald 2015).

One of the most prominent hypotheses of the past two decades, the chromalveolate hypothesis (Cavalier-Smith 1999), has seen a gradual diminishment in support over the last few years. The chromalveolate hypothesis suggests that all red-algal complex plastids have a single secondary origin in a common ancestor of all chromalveolate taxa (stramenopiles, alveolates, haptophytes and cryptophytes), and is based upon the idea that the number of inferred plastid establishments should be minimized due to the difficulties associated with evolving an organelle. Under our current understanding of the eukaryotic tree of life (e.g., Burki et al. 2016a), this requires extensive plastid loss as non-photosynthetic lineages frequently interrupt those that are photosynthetic (i.e., the chromalveolate taxa are not monophyletic). Additionally, due to the branching pattern of Cryptista with Archaeplastida observed in the most recent multi-gene phylogenies (e.g., Burki et al. 2016a), under the chromalveolate hypothesis, secondary plastids would have had to originate before red algal plastids even existed. While the plastid genomes of complex red-algal bearing taxa are monophyletic, the conflicting evolutionary histories of the plastid and nucleus have

resulted in alternative hypotheses centered around scenarios involving a single core secondary endosymbiosis followed by additional higher order endosymbiotic events (e.g., Sanchez-Puerta and Delwiche 2008; Stiller et al. 2014; Petersen et al. 2014; Bodyl 2017).

Stiller et al. (2014) proposed one such scenario, specifying the partners and order of horizontal transfers involved. Here, the initial secondary endosymbiosis is suggested to have occurred in a heterotrophic ancestor of cryptophytes. The secondary plastid in cryptophytes is then proposed to have spread to ochrophytes (photosynthetic stramenopiles) via a tertiary endosymbiosis and then passed to haptophytes by a quaternary endosymbiotic event. This 'cryptophyte first model' places the initial secondary event in the ancestor of cryptophytes on the basis of plastid gene phylogenies, statistical analyses of EGTs in ochrophytes, haptophytes and cryptophytes, and the preservation of a relic of the primary red algal nucleus (i.e., the nucleomorph; Stiller et al. 2014). Whether this particular hypothesis is correct or not, the analyses of Stiller et al. (2014) highlights the value of being able to accurately detect EGTs and endosymbiotic gene replacements (EGRs) in cryptophytes and other complex red algal plastid lineages, which has a significant effect on our ability to correctly reconstruct the evolutionary history of complex red-algal plastids and the eukaryotic tree of life as a whole (Lane and Archibald, 2008).

Reconstructing the tree of life is challenging enough on its own due to the long evolutionary histories that have occurred since the origin of eukaryotes – EGTs (and LGTs) provide an additional difficulty if left undetected as they can have vastly different evolutionary histories from those genes of true vertical decent. Haptophyta and Cryptista, phyla comprised of both photosynthetic and non-photosynthetic species, have been particularly challenging to place in eukaryotic tree of life (Burki et al. 2016a). Previously it was thought that Haptophyta and Cryptista formed a monophyletic group branching at the base of the Stramenopile-Alveolata-Rhizaria (SAR) clade (e.g., Burki et al. 2009) or sister to Archaeplastida (e.g., Katz and Grant 2015). Recent large scale phylogenomic studies, however, place Haptophyta and Centrohelida together (a clade referred to as Haptista) as sister to SAR, while Cryptista branches completely separate in a close, highly supported relationship with Archaeplastida (e.g. Burki et al. 2016a). As Burki et al. (2016a) pointed out, this branching pattern altogether rules out the chromalveolate hypothesis. It is

possible that undetected EGTs in the cryptophytes remain in the multi-gene datasets used to infer these phylogenies, causing conflicting signals and artificial attraction to Archaeplastida (Burki et al. 2016a).

Cryptista is now generally thought to be an ancestrally non-photosynthetic clade due to a lack of molecular evidence for a cryptic plastid or plastid derived genes in the katablepharids, one of the early diverging plastid-lacking clades within this phylum (Burki et al. 2012b). Being the closest heterotrophic lineage to the cryptophytes, the goniomonads are important for understanding the impact of endosymbiosis on the cryptophyte genome, and can potentially be invaluable in more accurately annotating cryptophyte genomes and detecting EGTs and EGRs. However, while *G. avonlea* is aplastidic today (Kim and Archibald 2013), its photosynthetic evolutionary history is uncertain. Two scenarios exist where goniomonads either (i) evolved from a plastid bearing ancestor and lost the organelle secondarily or (ii) were primitively plastid lacking and diverged prior to the uptake of a plastid in the common ancestor of cryptophytes. To this end, in this chapter I present an in-depth phylogenomic analysis of *G. avonlea* in search of evidence for a significant red algal footprint in its genome.

As the algal endosymbiont becomes integrated into the host cell, genes are transferred from the endosymbiont to the host nucleus (as discussed in Chapter 1). Movement of DNA (or RNA; see Timmis 2012) from the algal endosymbiont to the host nucleus is thought to occur not only during the transition from an endosymbiont to a fully-fledged organelle, but also before and after the organelle has evolved. EGT has had a notable role in the evolution of algae and their nuclear genomes (e.g. Timmis et al. 2004, Curtis et al. 2012). For example, a thorough genomic investigation into the cryptophyte *Guillardia theta* detected 508 genes of putative algal origin in its nucleus (of varying algal source; Curtis et al. 2012). While gene transfer is thought to be continuously occurring from the plastid to the nucleus (to varying extents in different lineages; Richly and Leister 2004), the limited transfer window hypothesis posits that the pulse of EGT occurred mostly during a window prior to permanent establishment of the endosymbiont, particularly in algae with single plastids as lysis of their plastid would be lethal (Barbrook et al. 2006). In higher plants, however, this window does not appear to be closed; studies measuring the

rate of EGT in the lab have shown these such transfers continuously occur at surprisingly high rates (Huang, Ayliffe and Timmis 2003; Stegemann et al. 2003; Timmis et al. 2004).

As a result, organisms that once had a plastid are expected to retain a footprint of algal endosymbiosis in the form of EGTs – EGTs they should have in common with their photosynthetic relatives (Curtis et al. 2012). Although plastid loss is difficult to prove, putative algal genes in plastid-lacking protists have been cited as evidence of gene transfer from a photosynthetic endosymbiont that was subsequently completely lost or lost photosynthetic abilities. One such example of this is the heterotrophic dinoflagellate *Crypthecodinium cohnii* (Sanchez-Puerta et al. 2007). In an analysis of sequence data from *C. cohnii*, Sanchez-Puerta et al. (2007) found fully intact N-terminal bipartite sequences indicative of plastid targeting as well as numerous genes showing cyanobacterial or algal origin providing significant evidence for plastid bearing ancestry and the presence of a reduced, non-photosynthetic plastid. Similarly, in the heterotrophic ciliates *Tetrahymena thermophila* and *Paramecium tetraurelia*, 16 proteins were identified of possible algal origin, 14 of which had homologs in other chromalveolate taxa (Reyes-Prieto et al. 2008). But, as the authors point out, it is impossible to tell if these are examples of EGT or if they are LGTs and a product of ciliates phagotrophic nature, or a phylogenetic artifact. Interestingly, one ciliate species, *Myrionecta rubra*, exhibits kleptoplasty and 'steals' chloroplasts from its prey, the cryptophyte *Gemingera cryophila* (Johnson et al. 2007). This scenario provides a concrete example that, by the same ideas behind the 'you are what you eat' hypothesis (Doolittle 1998), suggests algal genes in ciliates may be a result of having engulfed algal prey. A more obvious example of loss of photosynthesis involves the apicomplexans that retain a remnant of their original plastid in the form of a non-photosynthetic, essential organelle called the apicoplast which functions in fatty acid, isoprenoid and heme synthesis (Waller and McFadden 2005) and one apicomplexan species, *Cryptosporidium parvum*, who shows evidence of complete plastid loss (Zhu, Marchewka and Keithyl 2000; Huang, J. et al. 2004).

In the absence of cytological evidence of an active or vestigial plastid in *G. avonlea* (Kim and Archibald 2013), I searched for common algal EGTs in *G. avonlea* and cryptophytes (i.e., a shared red-algal footprint). The extent of EGT into photosynthetic

eukaryotes is unclear (Moreira and Deschamps 2014), but it has had a large enough impact to affect our ability to determine host phylogenies (Archibald 2015). Using heterotrophic lineages that are closely related to plastid bearing ones, such as *G. avonlea* to *G. theta* and other cryptophytes, one should be able to more accurately identify EGTs and better understand the genome mosaicism by looking for genes present in the genome where the gene tree and species tree do not agree (although these could still be the result of LGT or phylogenetic artifacts; Bodył, Stiller and Mackiewicz 2009). However, this becomes difficult to do if Cryptista truly branches sister to or within Archaeplastida (Burki et al. 2016) as this greatly increases the uncertainty in assigning gene origins to the host or the primary algal endosymbiont. Here I present a phylogenomic analysis from a targeted approach using previously predicted algal EGTs in *G. theta* (Curtis et al. 2012) as well as a more comprehensive investigation involving all predicted proteins in *G. avonlea*. Additionally, I present concatenated marker-gene phylogenies inferred in hope of gaining a better understanding of the relationship of Cryptista to other eukaryote phyla (in particular Archaeplastida) and an analysis of alternative signals emerging within the dataset.

3.2 METHODS

3.2.1 Creating the *G. avonlea* dataset

Generation of genome and transcriptome data for *G. avonlea*, along with gene model and protein predictions, was completed by Eunsoo Kim and Bruce Curtis. In order to investigate the most complete set of protein coding genes within *G. avonlea*, I combined all predicted proteins from the transcriptome (13,506) and a subset of predicted gene models from the genome (25,266) in a complementary approach. To ensure only non-redundant gene models were added to the dataset, genome-predicted gene models were subjected to a homology search against the *G. avonlea* transcriptome using double index alignment of next-generation sequencing data (DIAMOND; Buchfink et al. 2015). Predicted gene models that had a homolog to a predicted protein with higher than 90%

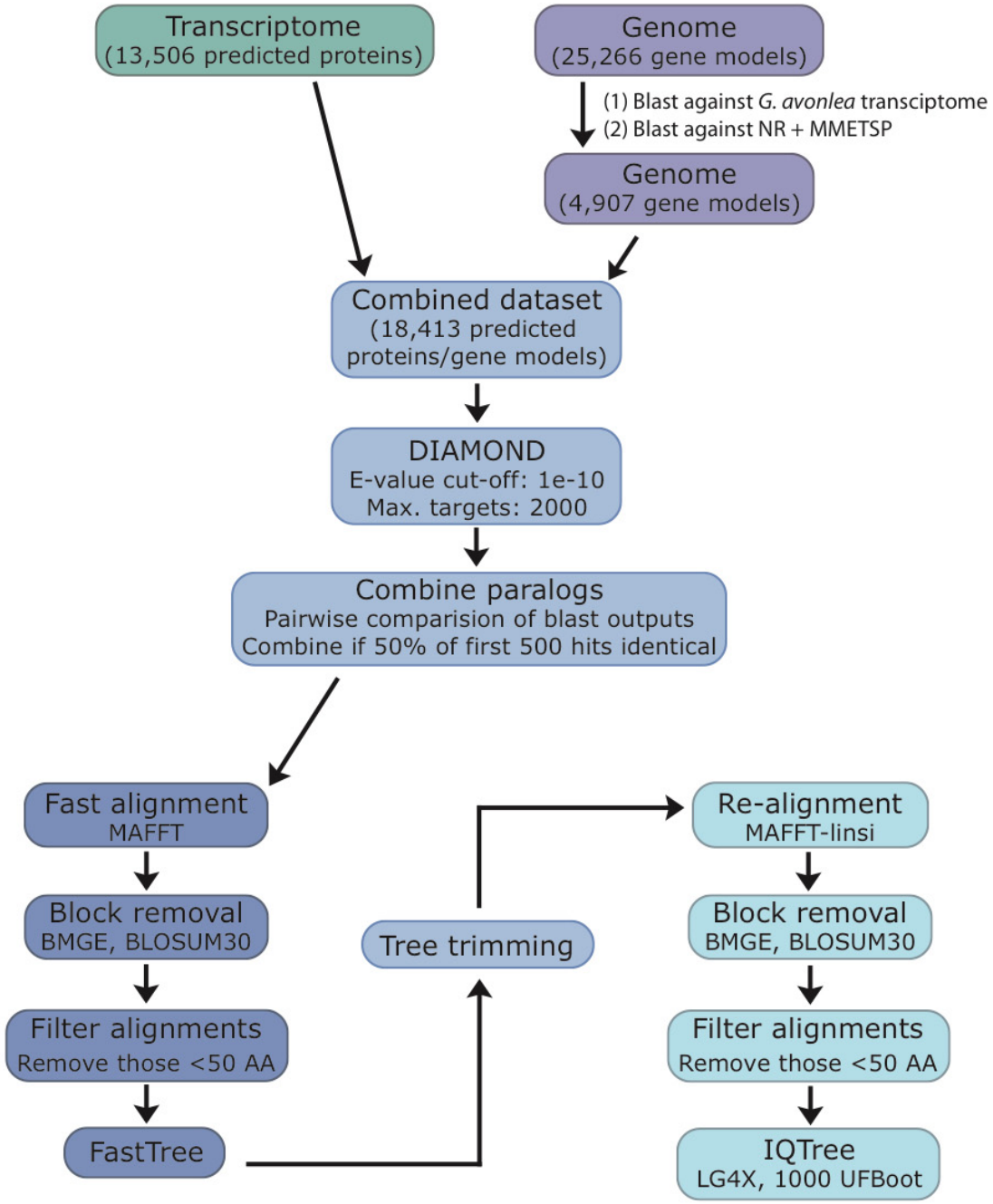
similarity were considered redundant and not considered further. Homologs to the remaining gene models were searched for using DIAMOND against a custom database of protein sequences from NR (NCBI non-redundant database) and the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP; Keeling et al. 2014; retrieved from http://imicrobe.us/project/view/104). Only predicted gene models with at least four homologs below an e-value of 1e-05 were added to the transcriptome dataset (i.e., gene models for which a phylogenetic tree could be inferred). This resulted in the addition of 4,923 predicted gene models to the transcriptome dataset, creating a combined dataset of 18,429 predicted proteins and gene models to be used in subsequent analyses.

3.2.2 From gene predictions to single gene trees

Homologs of each of the 18,429 predicted proteins from the *G. avonlea* transcriptome and genome dataset were retrieved by searching against a custom reference database consisting of NR and MMETSP protein sequences (including *Goniomonas* sp. and *G. pacifica* sequence data, and nucleomorph genomes) using DIAMOND (Buchfink et al. 2015) with the '--more sensitive' option to ensure recovery of all hits similar to the query. For each query, up to 2,000 hits with an e-value cut off 1e-10 were retrieved. Protein coding genes from *G. theta* (24,822; Curtis et al. 2012) were subjected to an identical homology search as above, but with the addition of the *G. avonlea* combined dataset to the custom database. Recent paralogs in the *G. avonlea* dataset were identified by pairwise comparison of DIAMOND outputs; if two queries had 50% or greater identical hits in their first 500 they were deemed paralogous and were combined in a non-redundant fashion.

Single gene trees were created for each of the remaining queries using a phylogenomic pipeline as outlined in Figure 3.1. Sequences were initially aligned using MAFFT (version 7.205; Katoh and Standley 2013) with default parameters. Ambiguously aligned regions were then removed using BMGE (version 1.1; Criscuolo and Gribaldo 2010) under default parameters except for the scoring matrix; for this analysis, BLOSUM30 was used to trim distantly related sequences in a more relaxed fashion. Any

Figure 3.1: Outline of the procedure used to generate single gene trees for each of the predicted proteins/gene models in *G. avonlea*.


Transcriptome
(13,506 predicted proteins)
Genome
(25,266 gene models)
(1) Blast against G. avonlea transciptome
(2) Blast against NR + MMETSP
Genome
(4,907 gene models)
Combined dataset
(18,413 predicted proteins/gene models)
DIAMOND
E-value cut-off: 1e-10
Max. targets: 2000
Combine paralogs
Pairwise comparision of blast outputs
Combine if 50% of first 500 hits identical
Fast alignment
MAFFT
Block removal
BMGE, BLOSUM30
Filter alignments
Remove those <50 AA
FastTree
Tree trimming
Re-alignment
MAFFT-linsi
Block removal
BMGE, BLOSUM30
Filter alignments
Remove those <50 AA
IQTree
LG4X, 1000 UFBoot

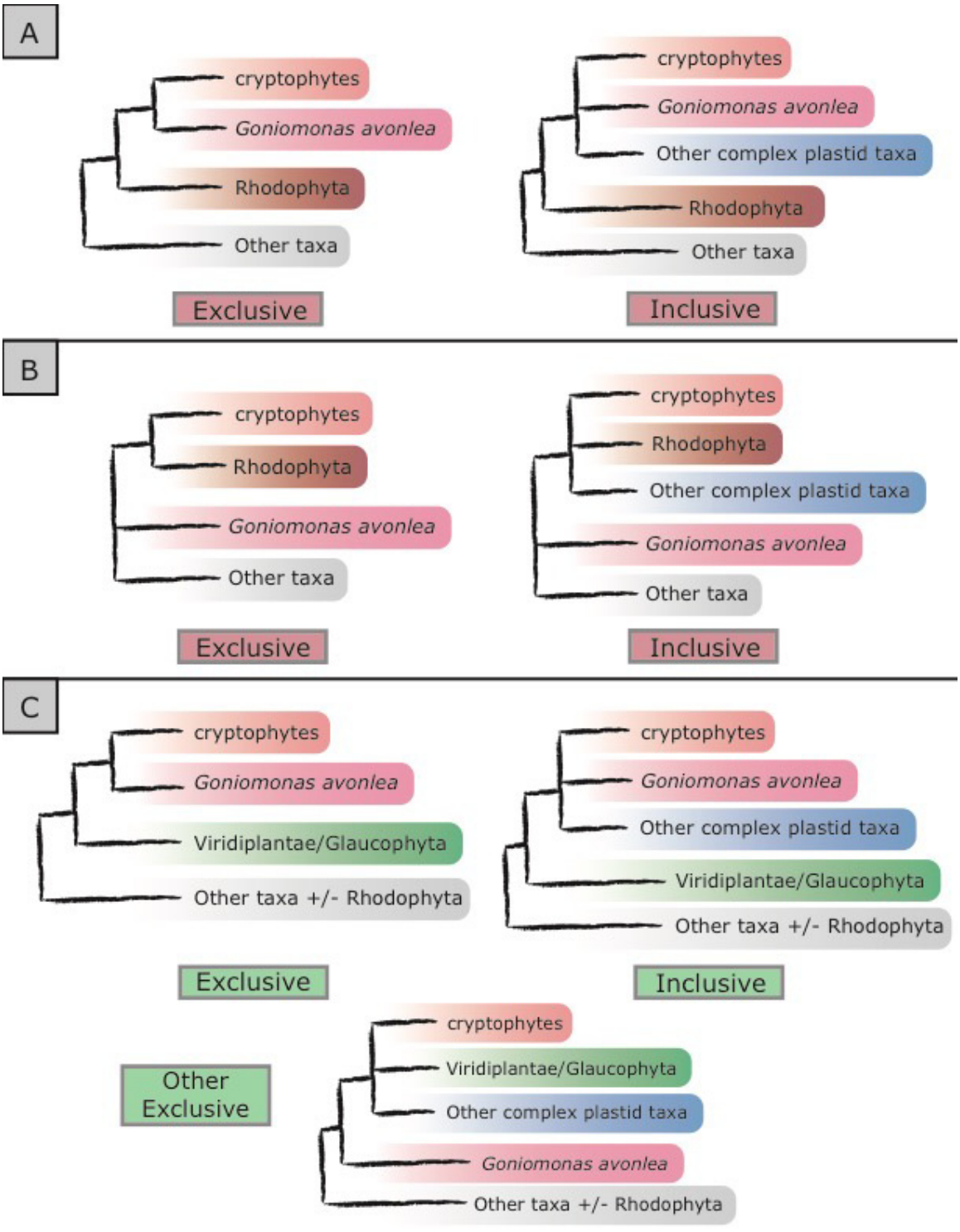
resulting trimmed alignment shorter than 50 AA was removed from further analyses. Alignments that met the minimum length requirement were used to create phylogenies based on approximately-maximum-likelihood methods using FastTree (Price et al 2009). Thereafter, the resulting tree was used in combination with the initial alignment to reduce taxonomic redundancy (ensuring that sequences belonging to the phyla Cryptista, Glaucophyta and Rhodophyta were retained) using an in-house tree-trimming script (written by Laura Eme). The reduced sequence sets were then re-aligned using MAFFT-linsi (version 7.205; Katoh and Standley 2013) with default parameters. Removal of ambiguously aligned regions was carried out as above and alignments shorter than 50 AA were once again discarded. Remaining alignments were used to infer phylogenies based on ML methods in IQ-TREE (Version 1.4.3; Nguyen et al. 2015) under the LG4X model with 1000 ultra-fast bootstrap approximations (UFboot) (Minh et al 2013).

3.2.3 Identifying genes of algal origin in *G. avonlea*

Homologs in the *G. avonlea* transcriptome and genome dataset to predicted algal EGTs in cryptophytes were identified using BLAST (basic local alignment search tool) in a protein-protein search against a custom database consisting of the 508 *G. theta* predicted algal EGTs in *G. theta* (Curtis et al. 2012). Any *G. avonlea* sequence that had a hit with an e-value less than 1e-10 was considered a potential shared algal EGT and its corresponding ML tree (as generated in section 3.2.2) was manually evaluated. Trees were sorted based on the topology of *G. avonlea* and *G. theta* in relation to each other and various combinations of Archaeplastida lineages and secondarily photosynthetic taxa as shown in Figure 3.2. Potential algal EGT genes in *G. avonlea* were annotated using InterPro (Finn et al. 2016) and their subcellular localization was predicted using TargetP (Emanuelsson et al. 2000) under both plant and non-plant modes due to the uncertain evolutionary history of *G. avonlea*.

In order to determine the phylogenetic position of *G. avonlea* and its surroundings in all single gene trees generated, an in-house pattern detecting script (a modified version of a script written by Laura Eme) was used. Trees were initially sorted based on the nearest

Figure 3.2: Schematic showing topologies of interest in identifying genes of putative algal ancestry in plastid bearing (cryptophytes) and plastid lacking (*Goniomonas avonlea*) Cryptista. **(A)** Genes showing shared red algal (Rhodophyta) ancestry amongst Cryptista. Cryptista branches either exclusively with Rhodophyta (left) or inclusively with other complex plastid bearing taxa present (right). **(B)** Genes showing red algal ancestry amongst Cryptophyta only. Cryptomonads branch either exclusively with Rhodophyta (left) or inclusively with other complex plastid bearing taxa present (right) to the exclusion of *G. avonlea*. **(C)** Genes showing ancestry with green (Viridiplantae) and/or glaucophyte (Glaucophyta) algae amongst Cryptista lineages. Topologies of genes with shared green/glaucophyte algal ancestry amongst Cryptista (top) show Cryptista branching either exclusively with them (left) or inclusively with other complex plastid bearing taxa present (right). Below this is another topology where either an exclusive or inclusive signal to green and/or glaucophyte algae is present in cryptophytes to the exclusion of *G. avonlea*.


A
cryptophytes
Goniomonas avonlea
Rhodophyta
Other taxa
Exclusive
cryptophytes
Goniomonas avonlea
Other complex plastid taxa
Rhodophyta
Other taxa
Inclusive
B
cryptophytes
Rhodophyta
Goniomonas avonlea
Other taxa
Exclusive
cryptophytes
Rhodophyta
Other complex plastid taxa
Goniomonas avonlea
Other taxa
Inclusive
C
cryptophytes
Goniomonas avonlea
Viridiplantae/Glaucophyta
Other taxa +/- Rhodophyta
Exclusive
cryptophytes
Goniomonas avonlea
Other complex plastid taxa
Viridiplantae/Glaucophyta
Other taxa +/- Rhodophyta
Inclusive
Other Exclusive
cryptophytes
Viridiplantae/Glaucophyta
Other complex plastid taxa
Goniomonas avonlea
Other taxa +/- Rhodophyta

neighboring eukaryotic super-group to *G. avonlea* and underwent additional pattern detection to progressively determine further phylogenetic context until either (i) no additional phylogenetic information was present in the tree, (ii) no clear taxonomic identity could be assigned to the next neighboring clade, or (iii) the most recent iteration of pattern detection showed strong affinity to Obazoa, Amoebozoa, Excavata or non-cyanobacterial prokaryotes. Any tree with cyanobacteria present or that showed a potential red-algal signal was manually evaluated and the corresponding protein-coding gene was annotated and subcellular localization predicted as above.

3.2.4 Multi-gene trees and assessing the phylogenetic position of Cryptista

To investigate the phylogenetic position of *G. avonlea* and Cryptista in the eukaryotic tree of life, a 250 marker gene, 150 operational taxonomic units (OTUs) un-aligned dataset from Burki et al. (2016a) was obtained from the Dryad Digital Repository (Burki et al. 2016b). The number of OTUs was systematically reduced to 98 in the interest of decreasing the complexity of phylogenetic analyses while maintaining taxonomic diversity (see Supplementary Table A1 for OTUs retained versus removed from the complete marker gene dataset and Supplementary Table A2 for gene names and abbreviations). Generally, closely related OTUs were evaluated for percent gene coverage across the marker gene dataset (see Supplementary Table A1) and the OTU with the least amount of missing data was included for phylogenetic analyses. *Telonema subtilis* and *Picomonas* sp. were amongst the OTUs removed – while they represent orphan lineages and do not belong to any particular eukaryotic super-group, they are problematic for resolution of the tree due to their poor representation in genomic databases. To decrease the amount of missing data for goniomonads in the Burki et al. (2016b) dataset, I replaced the only goniomonad data present with transcriptome data from *G. avonlea* to increase the percent gene coverage from 77.60% to 91.60%.

Predicted proteins from the *G. avonlea* transcriptome were added to the marker gene dataset by performing a homology search using BLASTp for each of the marker genes

against the transcriptome data using any Cryptista sequence as the query (if available) or the first sequence in the marker gene set (if no Cryptista sequence data was available). The most statistically significant hit from *G. avonlea* to each marker gene was added to the dataset if the e-value was below 1e-10. Each marker gene was then aligned as before using MAFFT-linsi (version 7.205; Katoh and Standley 2013) and ambiguous sites were removed using BMGE and default parameters (version 1.1; Criscuolo and Gribaldo 2010). Each marker gene trimmed alignment was used to infer a single gene tree using ML methods in IQTREE under the LG4X model (Version 1.4.3; Nguyen et al. 2015) and 1000 UFboot (Minh et al. 2013). The resulting phylogenies were manually inspected for any obvious problems (e.g., long branch attractions). The individual marker gene alignments were concatenated and the resulting alignment was used to infer a ML phylogeny in IQTREE (Version 1.4.3; Nguyen et al. 2015) using the model LG+C60+F+PMSF (posterior mean site frequency; Wang et al. 2017) with 100 standard bootstrap iterations.

Towards further investigating the phylogenetic position of Cryptista, the process of creating a multi-gene tree was repeated on modified versions of the above dataset where (i) photosynthetic Cryptista were removed from the dataset (i.e., cryptophytes), and (ii) potential 'problem' genes that produce discordant topologies for a given OTU and thus evolved in a different way or represent noise in the data were detected using an outlier detecting program (PhyloMCOA; De Vienne et al. 2012) and removed from the dataset (see Supplementary Table A3 for a list of outlier genes detected for each OTU). PhyloMCOA (De Vienne et al. 2012) considers the position of each OTU in each single gene tree and uses nodal distances between OTUs to look at similarities and differences in an OTUs phylogenetic position using multiple co-inertia analysis (MCOA). Additionally, *G. avonlea* predicted proteins were added to a second marker gene dataset containing 351 marker genes and 383 OTUs (derived from Kang et al. 2017 and Brown et al. (unpublished)) as above. As with the Burki et al. 2016a marker gene dataset, the number of OTUs was systematically reduced to 101 (see Supplementary Table A4 for OTUs retained versus removed from the Kang et al. (2017) complete marker gene dataset and Supplementary Table A5 for gene names and abbreviations). Individual marker genes were aligned, concatenated and used to infer a ML tree as above.

To explore alternative signals emerging from Cryptista in the marker gene dataset based on Burki et al. (2016a), 183 of the 250 genes (those that contained a homolog in *G. avonlea* and at least one other cryptophyte) were randomly partitioned into four equally sized bins (+/- one gene). Each subset of 45 or 46 aligned and trimmed marker genes were concatenated and used to infer a phylogeny based on ML methods in IQTREE (Version 1.4.3; Nguyen et al. 2015) using the model LG+C20+F. This process was repeated 25 times, resulting in 100 randomly generated marker gene subset trees. The topology of Cryptista in each of these random permutation trees was manually evaluated.

3.3 RESULTS

3.3.1 Common algal EGTs in *G. avonlea* and *G. theta*

Based on sequence homology searches, only 144 of the 508 predicted algal EGTs in *G. theta* (Curtis et al. 2012) were found to have obvious homologs in *G. avonlea*. As shown in Figure 3.3, manual evaluation of phylogenies for each of the 144 potential common EGTs with regards to algal signal (for patterns as shown in Figure 3.2) resulted in only nine being assigned as showing any red algal signal in both cryptophytes and *G. avonlea* (e.g., Figure 3.4). On the other hand, *G. avonlea* appeared sister to an amoebozoan 16 times in these trees, a topology not expected to be observed frequently. Potential EGRs, where cryptophytes show a significant red algal signal to the exclusion of *G.* avonlea, were observed in 16/144 of these phylogenies (e.g., Figure 3.5). A large proportion of these trees (51/144) showed an unambiguously green or glaucophyte algal signal, while 50 of them did not show any significant algal signal (in contrast to Curtis et al. 2012) and thus could not be assigned as unambiguously algal. Functional annotation and subcellular localization predictions on the resulting nine potential red algal EGTs common in both Cryptophyta and *G. avonlea* are shown in Table 3.1).

Figure 3.3: Distribution of topologies observed in *Goniomonas avonlea* homologs to 144 of the predicted algal EGTs in *Guillardia theta* (Curtis et al. 2012). Patterns observed are as explained in Figure 3.2. The number of trees in which a pattern is observed as well as the percent this comprises of the 144 phylogenies is indicated. Only 9/144 homologs in *G. avonlea* showed any affinity to red algae.

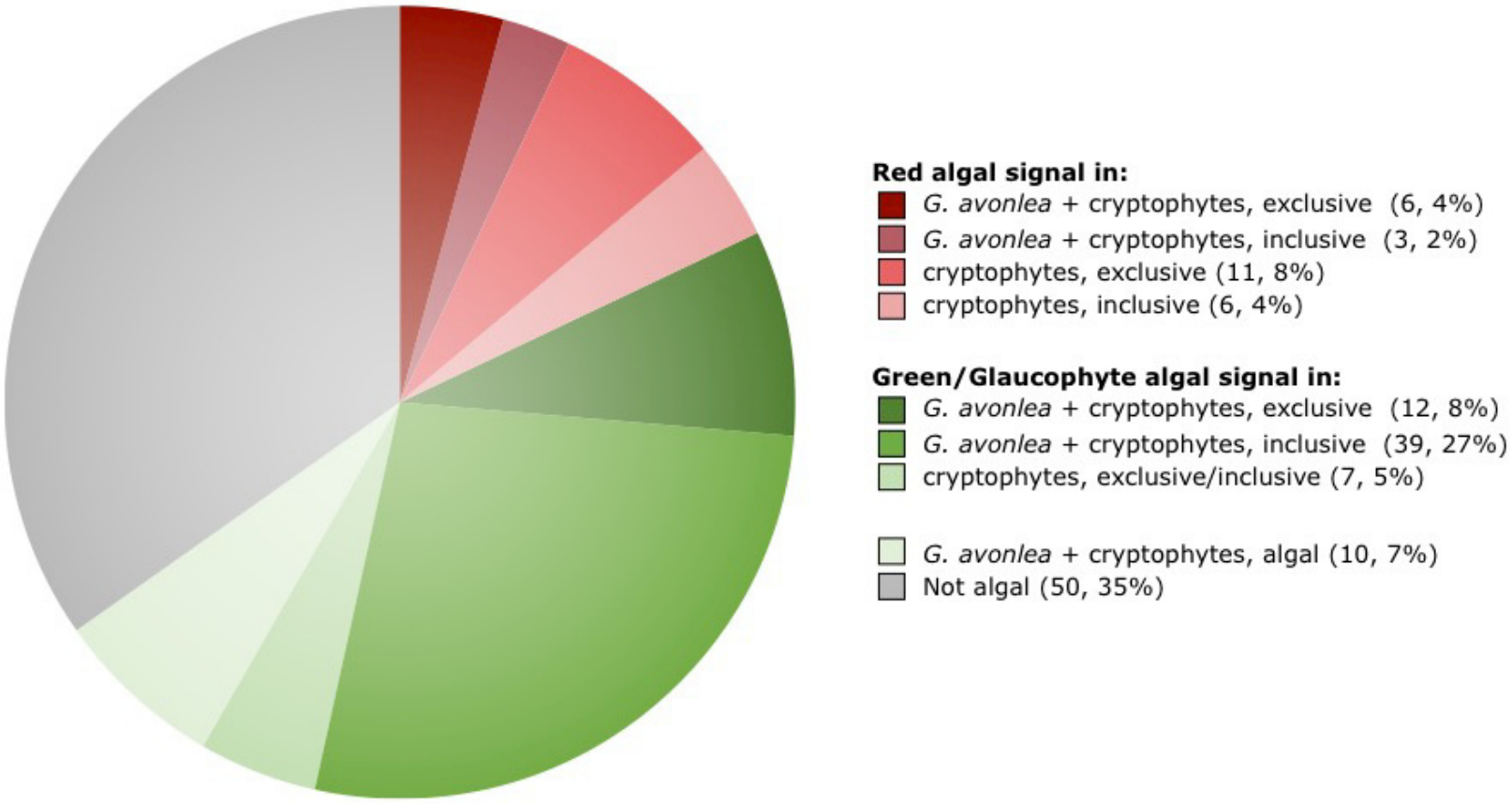

Red algal signal in:
G. avonlea + cryptophytes, exclusive (6, 4%)
G. avonlea + cryptophytes, inclusive (3, 2%)
cryptophytes, exclusive (11, 8%)
cryptophytes, inclusive (6, 4%)
Green/Glaucophyte algal signal in:
G. avonlea + cryptophytes, exclusive (12, 8%)
G. avonlea + cryptophytes, inclusive (39, 27%)
cryptophytes, exclusive/inclusive (7, 5%)
G. avonlea + cryptophytes, algal (10, 7%)
Not algal (50, 35%)

Figure 3.4: Maximum likelihood (ML) phylogeny of a single gene in *G. avonlea* (comp57164_c0, ubiquitin activating enzyme E1) and its homologs inferred under the model LG4X with 1000 UFboot replicates (682 unambiguously aligned sites). Based on the topology of the tree, this gene represents a candidate EGT of red algal origin found in both *G. avonlea* and cryptophytes. OTUs are colored according to their eukaryotic super group with sequences from *G. avonlea* highlighted in bright red. Clades containing 10 or more species from the same super group were collapsed for simplicity. The tree shown has been rooted in midpoint. Black dots indicate maximal support for a particular node. When not maximal, only bootstrap support values > 70% are shown. The scale bar shows an inferred 0.2 substitutions per site.

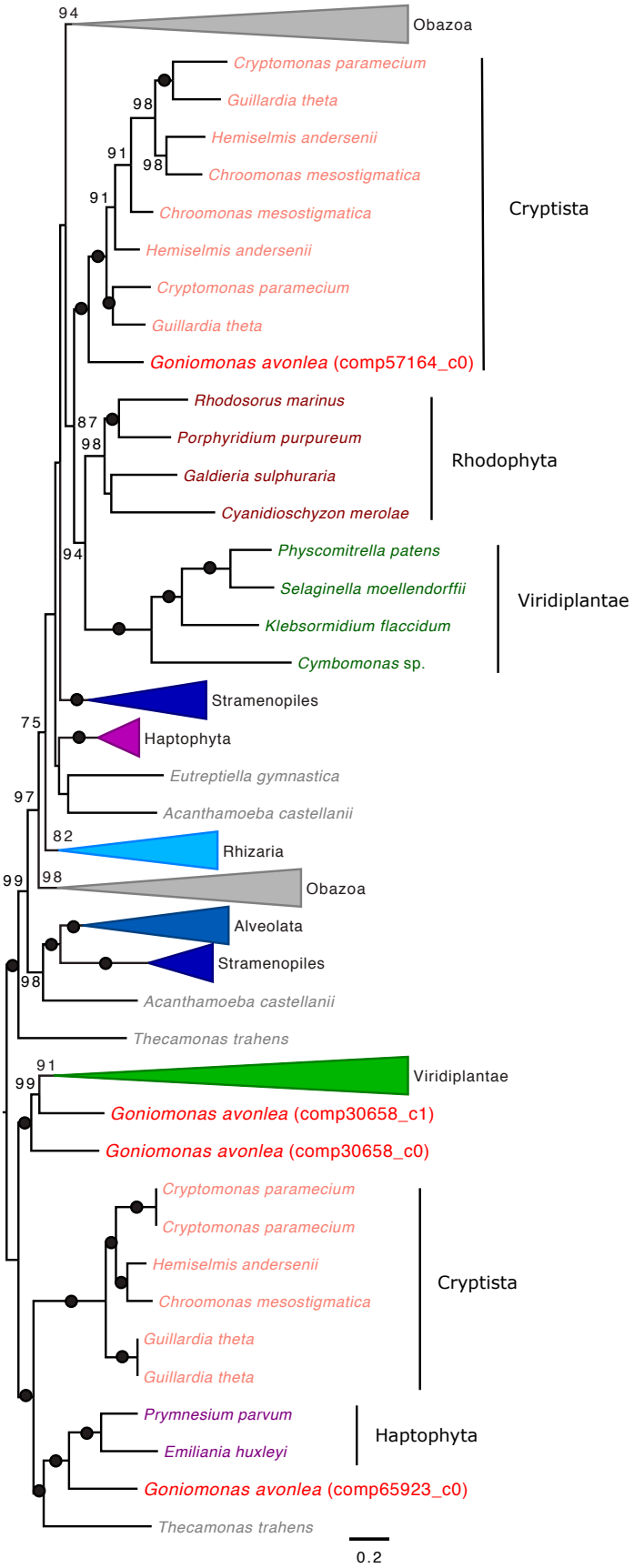

94
Obazoa
Cryptomonas paramecium
Guillardia theta
98
Hemiselmis andersenii
91
98
Chroomonas mesostigmatica
91
Chroomonas mesostigmatica
Hemiselmis andersenii
Cryptomonas paramecium
Guillardia theta
Goniomonas avonlea (comp57164_c0)
Cryptista
Rhodosorus marinus
87
Porphyridium purpureum
98
Galdieria sulphuraria
Cyanidioschyzon merolae
Rhodophyta
94
Physcomitrella patens
Selaginella moellendorffii
Klebsormidium flaccidum
Cymbomonas sp.
Viridiplantae
Stramenopiles
75
Haptophyta
Eutreptiella gymnastica
97
Acanthamoeba castellanii
82
Rhizaria
99
98
Obazoa
Alveolata
Stramenopiles
98
Acanthamoeba castellanii
Thecamonas trahens
91
99
Viridiplantae
Goniomonas avonlea (comp30658_c1)
Goniomonas avonlea (comp30658_c0)
Cryptomonas paramecium
Cryptomonas paramecium
Hemiselmis andersenii
Chroomonas mesostigmatica
Guillardia theta
Guillardia theta
Cryptista
Prymnesium parvum
Emiliania huxleyi
Haptophyta
Goniomonas avonlea (comp65923_c0)
Thecamonas trahens
0.2

Figure 3.5: Maximum likelihood (ML) phylogeny of a single gene in *G. avonlea* (comp56792_c0, 4-alpha-glucanotransferase, mitochondria targeted) and its homologs showing a candidate red-algal EGT in cryptophytes (plastid targeted) that is not present in *G. avonlea*. This phylogeny was inferred under the model LG4X with 1000 UFboot replicates (using 326 unambiguously aligned sites). OTUs are colored according to their eukaryotic super group with sequences from *G. avonlea* highlighted in bright red. Clades containing 10 or more species from the same super group were collapsed for simplicity. The tree shown has been rooted in midpoint. Black dots indicate maximal support for a particular node. When not maximal, only bootstrap support values > 70% are shown. The scale bar shows an inferred 0.4 substitutions per site.

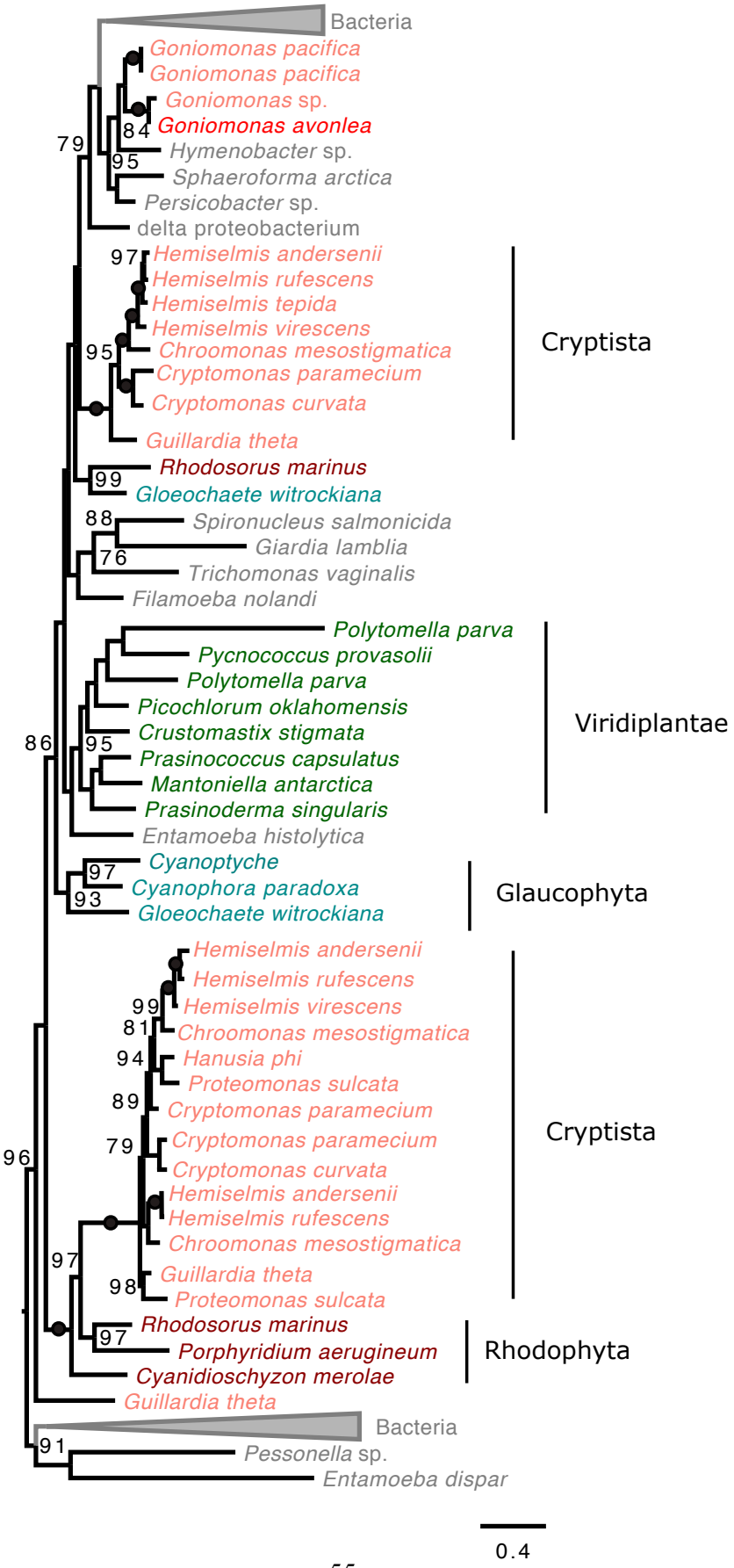

Bacteria
Goniomonas pacifica
Goniomonas pacifica
Goniomonas sp.
84
Goniomonas avonlea
79
Hymenobacter sp.
95
Sphaeroforma arctica
Persicobacter sp.
delta proteobacterium
97
Hemiselmis andersenii
Hemiselmis rufescens
Hemiselmis tepida
Hemiselmis virescens
95
Chroomonas mesostigmatica
Cryptomonas paramecium
Cryptomonas curvata
Cryptista
Guillardia theta
Rhodosorus marinus
99
Gloeochaete witrockiana
88
Spironucleus salmonicida
Giardia lamblia
76
Trichomonas vaginalis
Filamoeba nolandi
Polytomella parva
Pycnococcus provasolii
Polytomella parva
Picochlorum oklahomensis
Crustomastix stigmata
86
95
Prasinococcus capsulatus
Mantoniella antarctica
Prasinoderma singularis
Viridiplantae
Entamoeba histolytica
Cyanoptyche
97
Cyanophora paradoxa
93
Gloeochaete witrockiana
Glaucophyta
Hemiselmis andersenii
Hemiselmis rufescens
99
Hemiselmis virescens
81
Chroomonas mesostigmatica
94
Hanusia phi
Proteomonas sulcata
89
Cryptomonas paramecium
79
Cryptomonas paramecium
96
Cryptomonas curvata
Hemiselmis andersenii
Hemiselmis rufescens
Chroomonas mesostigmatica
97
Cryptista
Guillardia theta
98
Proteomonas sulcata
Rhodosorus marinus
97
Porphyridium aerugineum
Cyanidioschyzon merolae
Rhodophyta
Guillardia theta
Bacteria
91
Pessonella sp.
Entamoeba dispar
0.4

Table 3.1 Functional annotation and subcellular localization predictions for nine potential common EGTs of red algal origin in cryptophytes and *G. avonlea*. LC = low confidence prediction (reliability class 3, 4 or 5); 'Other' indicates the sequence was not predicted to contain a signal peptide, mitochondrial targeting signal or plastid transit peptide.

G. avonlea gene	Corresponding *G. theta* homolog	Targeting signal (plant/non-plant)	Protein function
comp89552_c0	Gtheta_algalgenes_118616	Signal peptide	Na+/H+ exchanger
comp52815_c0	Gtheta_algalgenes_159018	Mitochondrial	ATP binding
comp54781_c1	Gtheta_algalgenes_133922	Other	Transcription factor DP
comp57164_c0	Gtheta_algalgenes_64494	Other	Ubiquitin/SUMO-activating enzyme E1
comp57471_c1	Gtheta_algalgenes_94821	Other	Translation initiation factor
comp61465_c0	Gtheta_algalgenes_54608, Gtheta_algalgenes_109969	Other	ATP-dependent RNA helicase Ski2
g21336.t1	Gtheta_algalgenes_114557	Signal peptide (LC)	Na+/H+ exchanger
g26967.t1	Gtheta_algalgenes_95701	Other	Heat shock protein (Hsp90)
g32900.t1	Gtheta_algalgenes_118616	Signal peptide	Na+/H+ exchanger

3.3.2 Phylogenetic distribution of predicted proteins and gene models

Considering the top blast hit to each predicted protein and gene model in the *G. avonlea* dataset and its broad taxonomic classification, an expected affinity to other Cryptista was observed most frequently (approximately 27% of the time; Figure 3.6). Surprisingly, the second most common top-hit (15%) was from Obazoa, with Viridiplantae and Alveolata each appearing as the most significant hit in 11% of sequence homology searches. Notably, the number of instances where an amoebozoan was the most similar sequence (1124, 6.5%) was considerably greater than those where a red alga was most similar (130, 0.7%). Considering the taxonomic distribution of top blast hits to the *G. theta* nuclear genome, *G. theta* shows a 3.6 times greater enrichment in red algal signal than *G. avonlea* (Table 3.2).

Analysis of the phylogenetic affinity for all predicted proteins and gene models in the *G. avonlea* combined dataset showed the nearest neighbor most commonly (and expectedly) as Cryptista (Figure 3.7). Approximately half of the single gene trees produced could not be assigned a definitive nearest neighbor. As in the top-blast hit results, a surprising number of single gene trees showed *G. avonlea* branching sister to sequences affiliated with Obazoa (858/11,955) and the number of instances where Rhodophyta was sister to *G. avonlea* specifically (93/11,955) was much less than those where *G. avonlea* was sister to Amoebozoa (248/11,955). Targeting signal prediction and functional annotation of the single gene trees where *G. avonlea* branched with Cyanobacteria specifically (10/11,955) did not identify any obvious plastid targeted or functioning proteins.

Results of additional phylogenetic affiliations (i.e., the next sister groups) for those predicted gene models and proteins in *G. avonlea* showing an immediate sister relationship to Cryptista (2,689/11,955) is shown in Figure 3.8. Approximately half of these 2,689 genes showed either no additional taxa in the phylogeny or were poorly resolved and could not be confidently assigned an additional phylogenetic pattern. Those that could be assigned a phylogenetic pattern most commonly showed Cryptista and *G. avonlea* sister to Obazoa (300/1,372). The number of times a sister relationship was observed between the Cryptista and *G. avonlea* clade and Rhodophyta (68/1,372) was slightly less than those found sister to Amoebozoa (76/1,372). Results of additional phylogenetic affiliations (i.e.,

Figure 3.6: The taxonomic distribution of the top blast hit to each predicted protein and gene model in the *G. avonlea* dataset. The top blast hit was defined as the most significant homolog to *G. avonlea* (i.e., lowest e-value) excluding any other *Goniomonas* sequence. Super groups shown are abbreviated as follows: Cry = Cryptista, Rho. = Rhodophyta, Vir = Viridiplantae, Gla = Glaucophyta, Hap = Haptista, Str = Stramenopiles, Alv = Alveolata, Rhi = Rhizaria, Oba = Obazoa, Amo = Amoebozoa, Cya = Cyanobacteria, Bac = Bacteria (non-cyanobacteria), and Arc = Archaea.

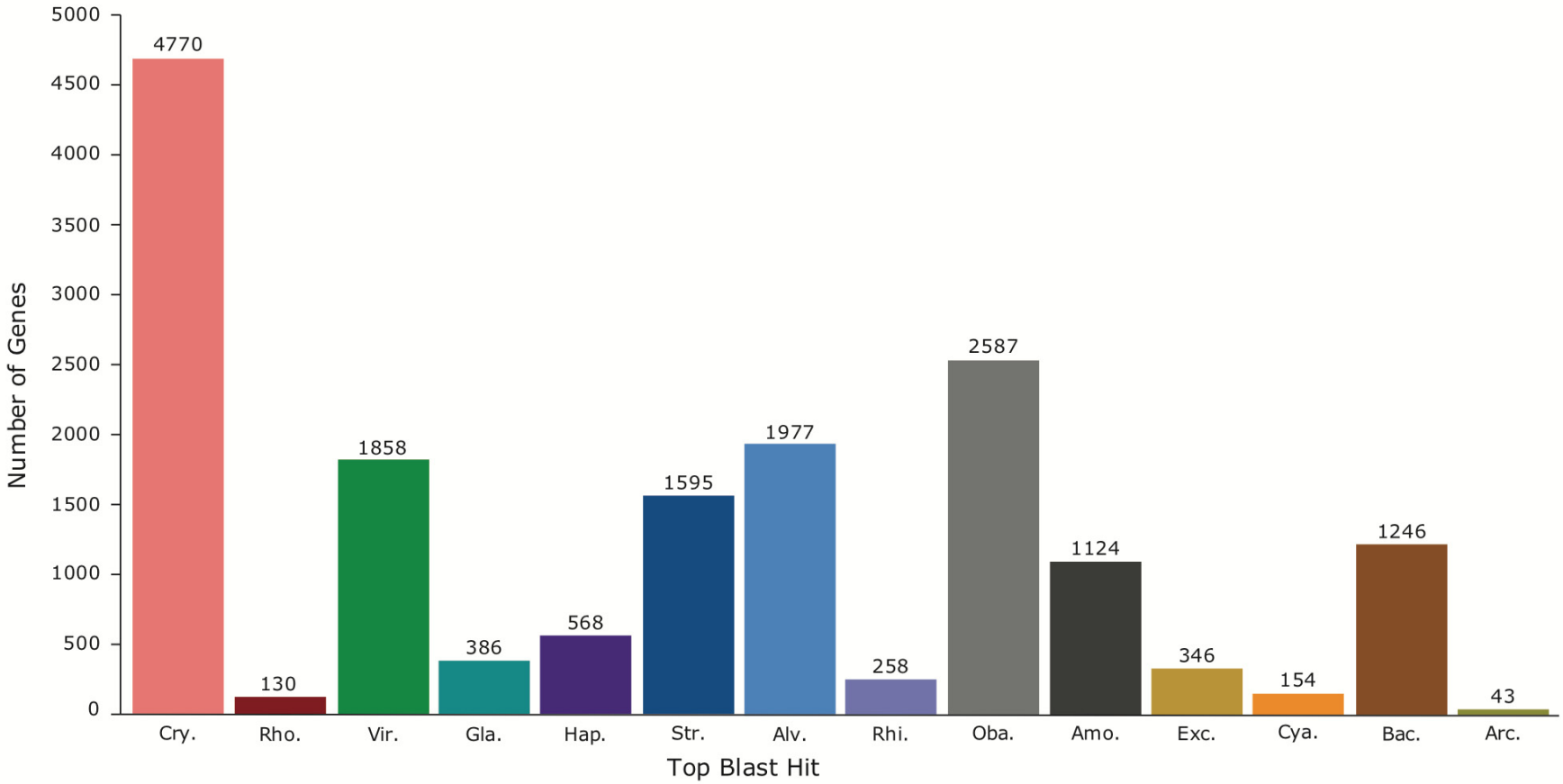

Number of Genes
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
4770
130
1858
386
568
1595
1977
258
2587
1124
346
154
1246
43
Cry.
Rho.
Vir.
Gla.
Hap.
Str.
Alv.
Rhi.
Oba.
Amo.
Exc.
Cya.
Bac.
Arc.
Top Blast Hit

Table 3.2. Contribution of Rhodophyta and Amoebozoa affiliated genes to the genomes of the photosynthetic cryptophyte *G. theta* and non-photosynthetic goniomonad *G. avonlea.* Top hits to Amoebozoa were considered as a control taxa to determine if red algal signal is above a baseline expected outcome due to non-phylogenetic signal. The ratio of Rhodophyta to Amoebozoa signal was corrected for abundance of sequences from each group within the database queried.

Genome	Total Top-Blast Hits		Rhodophyta : Amoebozoa (corrected)
	Rhodophyta	Amoebozoa	
G. theta	108	267	0.86
G. avonlea	130	1124	0.24

Figure 3.7: The phylogenetic position of *G. avonlea* across all 11,955 single gene trees generated from the combined predicted proteins and gene models dataset. Phylogenetic position was determined as the super-group of the majority of OTUS in the closest clade to *G. avonlea* (i.e. nearest neighbor) with bootstrap support ≥ 70%. Not shown are 5,327 trees that did could not be assigned a clear nearest neighbor. Super groups shown are abbreviated as follows: Cry = Cryptista, Rho. = Rhodophyta, Vir = Viridiplantae, Gla = Glaucophyta, Hap = Haptista, Str = Stramenopiles, Alv = Alveolata, Rhi = Rhizaria, Oba = Obazoa, Amo = Amoebozoa, Cya = Cyanobacteria, Bac = Bacteria (non-cyanobacteria), and Arc = Archaea.

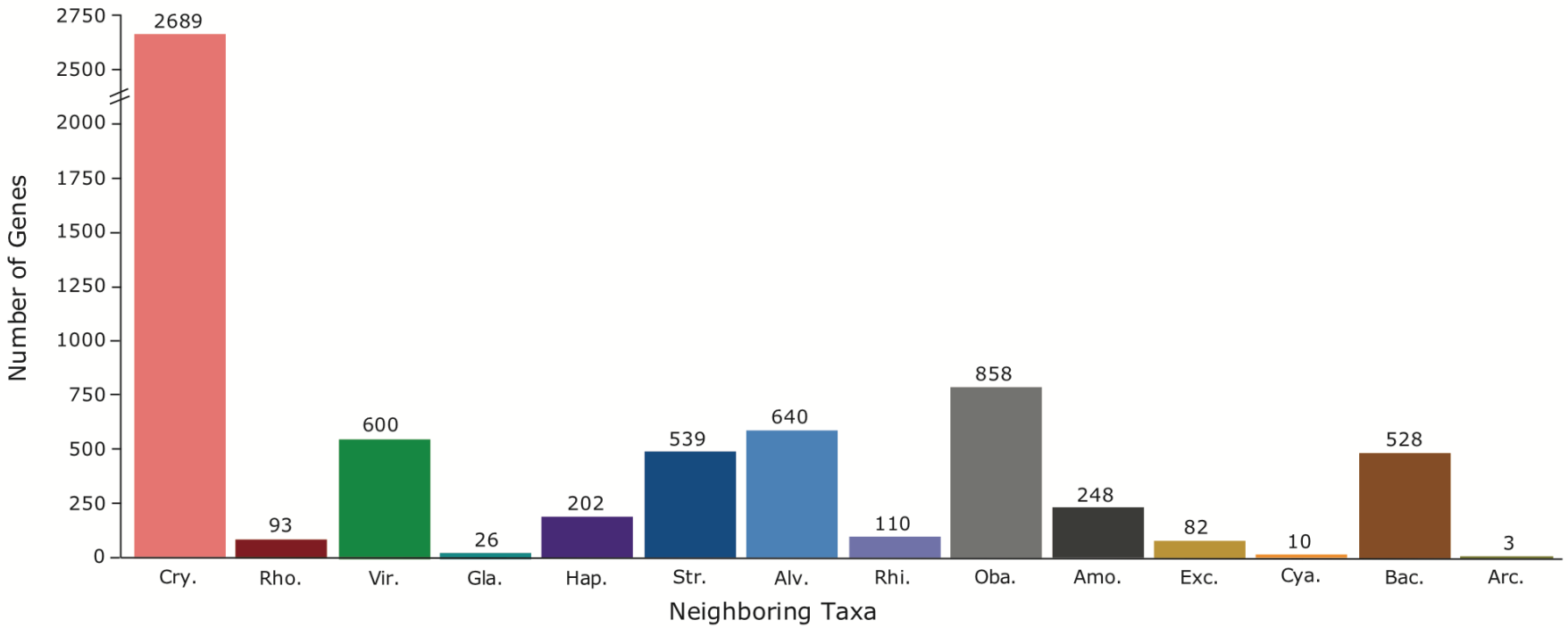
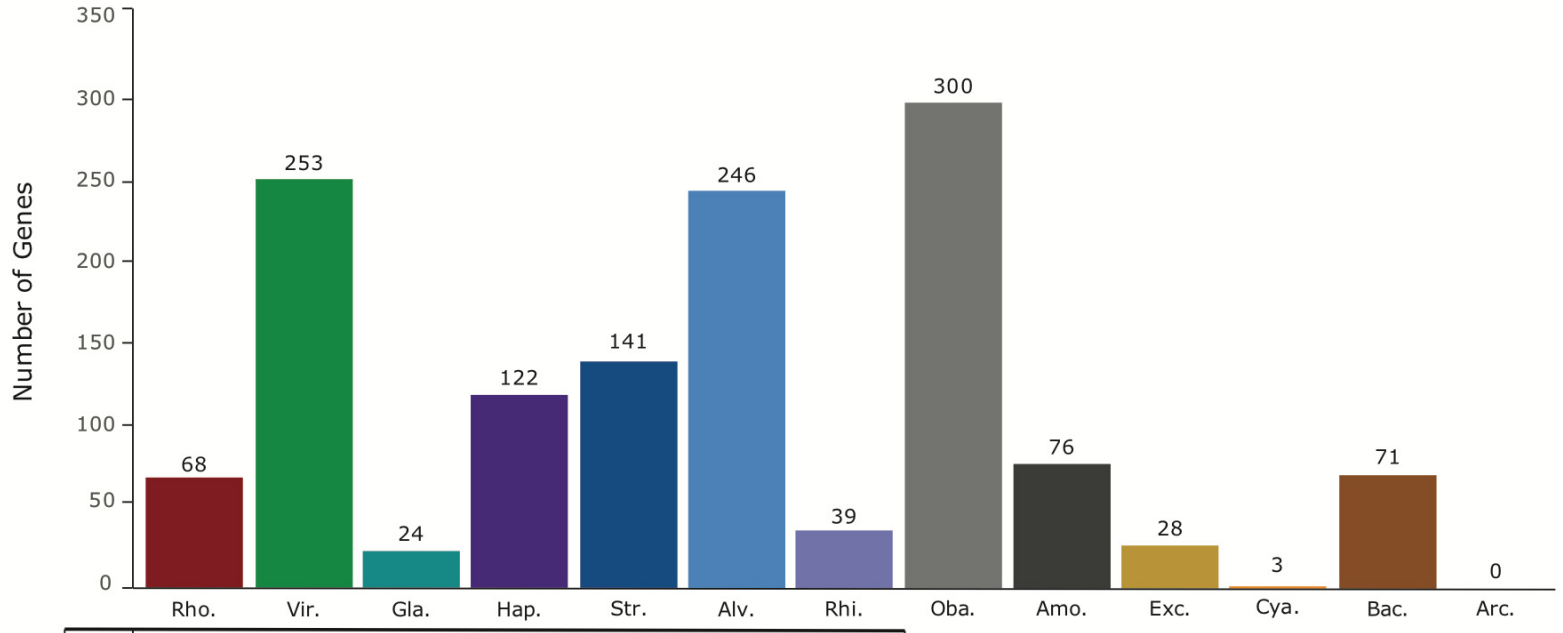

Number of Genes
2750
2500
2000
1750
1500
1250
1000
750
500
250
0
2689
93
600
26
202
539
640
110
858
248
82
10
528
3
Cry.
Rho.
Vir.
Gla.
Hap.
Str.
Alv.
Rhi.
Oba.
Amo.
Exc.
Cya.
Bac.
Arc.
Neighboring Taxa

Figure 3.8: The phylogenetic position of Cryptista across all single gene trees generated where *G. avonlea* branches sister to or within Cryptista. Phylogenetic position was determined as the super-group of the majority of OTUS in the closest clade to *G. avonlea* and Cryptista (i.e., nearest neighbor) with bootstrap support ≥ 70%. An additional round of topology detection was used to determine the next nearest neighbor to any clade showing a relationship between *G. avonlea*, Cryptista and an additional photosynthetic eukaryotic group (shown in a table below the histogram). Super groups shown are abbreviated as follows: Cry = Cryptista, Rho. = Rhodophyta, Vir = Viridiplantae, Gla = Glaucophyta, Hap = Haptista, Str = Stramenopiles, Alv = Alveolata, Rhi = Rhizaria, Oba = Obazoa, Amo = Amoebozoa, Cya = Cyanobacteria, Bac = Bacteria (non-cyanobacteria), and Arc = Archaea. A dash ('-') indicates no tree showed the corresponding branching pattern.



Next Neighboring Taxa

Rho.	-	8	1	5	2	6	1
Vir.	3	-	6	15	20	33	3
Gla.	1	3	-	-	3	2	2
Hap.	-	23	1	-	14	16	3
Str.	9	21	-	5	-	29	1
Alv.	3	21	2	11	22	-	-
Rhi.	3	3	1	5	2	11	-
Oba.	9	44	1	19	17	27	3
Amo.	1	9	-	1	2	4	1
Exc.	1	4	-	-	1	5	-
Cya.	2	1	-	1	-	-	-
Bac.	7	3	-	8	6	10	2
Arc.	-	-	-	2	1	-	-

the next sister groups) for those predicted gene models and proteins in *G. avonlea* showing an immediate sister relationship to super groups containing primary or red-algal secondarily photosynthetic taxa other than Rhodophyta or Cryptista are shown in Supplementary Figures B1-B7. Manual evaluation of 98 single gene trees showing red-algal signal in both *G. avonlea* and Cryptista (exclusively (68) or inclusively (30) with other complex red-algal derived plastid bearing taxa) resulted in reduction of candidate red-algal EGTs in *G. avonlea* and photosynthetic Cryptista from 98 to 10 (such as the phylogeny shown in Figure 3.9); subcellular localization and functional predictions for these genes (along with topology details) is shown in Table 3.3.

3.3.3 Phylogenetic position of Cryptista in the eukaryotic tree of life

Using a dataset based on Burki et al. (2016a) consisting of 98 OTUs and 250 marker genes (see Supplementary Table A1 and A2) with *G. avonlea* added to it, a multi-gene phylogeny was inferred as shown in Figure 3.10 under the model LG + C60 + F+ PMSF with 100 bootstrap replicates. With the exception of Archaeplastida, the monophyly of eukaryotic supergroups as well as that of the SAR clade was recovered with maximum support. Here, Haptista was found to branch sister to the SAR super-group with nearly maximal support (99% bootstrap support). The monophyly of Archaeplastida was broken-up by the positioning of Cryptista, which was found to branch with Archaeplastida (with maximum support) and, more specifically, sister to a clade of Viridiplantae and Glaucophyta (99% bootstrap support) to the exclusion of Rhodophyta with 82% bootstrap support.

When OTUs affiliated with Cryptista were removed from the dataset and a similar phylogeny was inferred (Figure 3.11), the monophyly of Archaeplastida was recovered with maximum support with Viridiplantae and Glaucophyta forming a maximally supported clade to the exclusion of Rhodophyta. No other topological changes were observed. Upon reintroducing Cryptista OTUs not known to have ever harbored a plastid (i.e., exclusion of the photosynthetic cryptophytes), the monophyly of Archaeplastida was once again broken by the branching of non-photosynthetic Cryptista within it. Here, this subset of Cryptista branch sister to a clade of Viridiplantae and Glaucophyta (92% standard

Figure 3.9: Maximum likelihood (ML) phylogeny of a single gene in *G. avonlea* (comp62470_c4, nucleic acid binding domains) and its homologs (inferred under the model LG4X with 1000 UFboot replicates across 214 unambiguously aligned sites) that shows a candidate shared EGT of red algal origin in *G. avonlea* and photosynthetic cryptophytes. OTUs are colored according to their eukaryotic super group with sequences from *G. avonlea* highlighted in bright red. The tree shown has been rooted in midpoint. Black dots indicate maximal support for a particular node. When not maximal, only bootstrap support values > 70% are shown. The scale bar shows an inferred 0.2 substitutions per site.

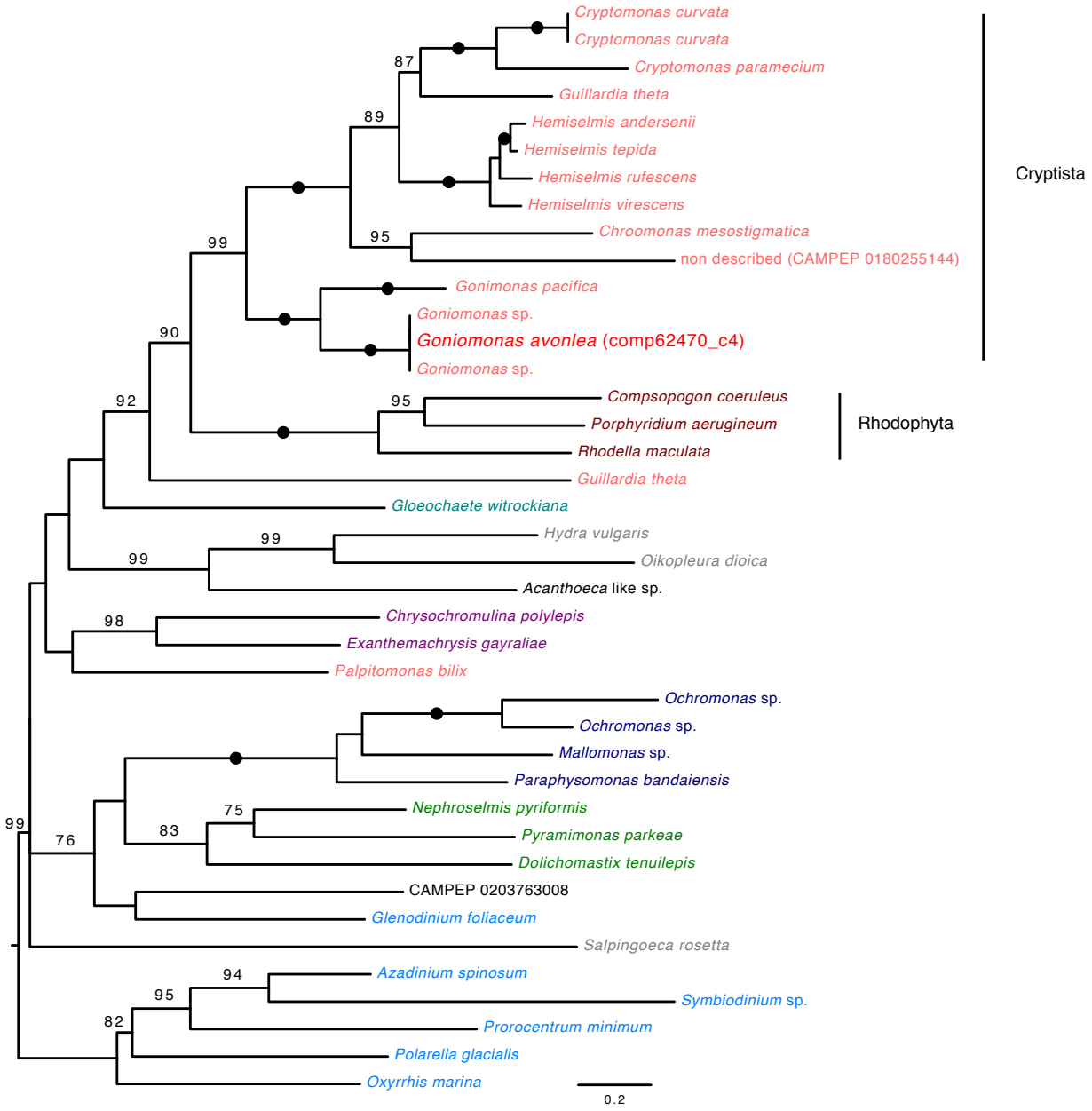

Cryptomonas curvata
Cryptomonas curvata
Cryptomonas paramecium
Guillardia theta
Hemiselmis andersenii
Hemiselmis tepida
Hemiselmis rufescens
Hemiselmis virescens
Chroomonas mesostigmatica
non described (CAMPEP 0180255144)
Gonimonas pacifica
Goniomonas sp.
Goniomonas avonlea (comp62470_c4)
Goniomonas sp.
Cryptista
Compsopogon coeruleus
Porphyridium aerugineum
Rhodella maculata
Rhodophyta
Guillardia theta
Gloeochaete witrockiana
Hydra vulgaris
Oikopleura dioica
Acanthoeca like sp.
Chrysochromulina polylepis
Exanthemachrysis gayraliae
Palpitomonas bilix
Ochromonas sp.
Ochromonas sp.
Mallomonas sp.
Paraphysomonas bandaiensis
Nephroselmis pyriformis
Pyramimonas parkeae
Dolichomastix tenuilepis
CAMPEP 0203763008
Glenodinium foliaceum
Salpingoeca rosetta
Azadinium spinosum
Symbiodinium sp.
Prorocentrum minimum
Polarella glacialis
Oxyrrhis marina
87
89
95
99
90
92
95
99
99
98
99
75
83
76
99
94
95
82
0.2

Table 3.3. Functional annotation and subcellular localization predictions for 10 potential common EGTs of red algal origin in cryptophytes and *G. avonlea* identified based on tree pattern detection that were not predicted in Curtis et al. (2012). Whether a homolog in *G. theta* is present in the phylogeny is indicated. LC = low confidence prediction (reliability class 3, 4 or 5); 'Other' indicates the sequence was not predicted to contain a signal peptide, mitochondrial targeting signal or plastid transit peptide.

G. avonlea gene	Homolog in *G. theta*?	Targeting signal (plant/non-plant)	Protein function
comp118753_c0	Yes	Other	Protein kinase-like
comp51629_c0	No	Other	None predicted
comp62470_c4	Yes	Signal peptide/Other	Domains involved in nucleic acid binding
g10210.t1	No	Mitochondrial	FAD/NAD(P)-binding
g13890.t1	No	Other	Dihydroprymidine dehydrogenase (pyrimidine degradation)
g20552.t1	Yes	Chloroplast (LC)/Other	Histone H2A
g34362.t1	Yes	Other	WD40-repeat (protein binding)
g34578.t1	Yes	Mitochondrial	Proteasome beta 3 subunit
g6101.t1	Yes	Signal peptide	Phosphate transporter
g7117.t1	Yes	Other	Nop domain

Figure 3.10: Maximum likelihood (ML) phylogeny of a 250 marker gene set as in Burki et al. (2016a) that includes new transcriptome data from *Goniomonas avonlea*. The phylogeny is based on a concatenated marker gene alignment of 71,151 unambiguously aligned sites across 98 OTUs. The ML tree shown was generated under the model LG + C60 + F + PMSF with 100 standard bootstrap replicates and has been rooted in mid-point. Black dots indicate maximal support for a particular node. When not maximal, only bootstrap support values > 70% are shown. The scale bar shows an inferred 0.2 substitutions per site. For consistency, taxon names are provided as in Burki et al. (2016a).

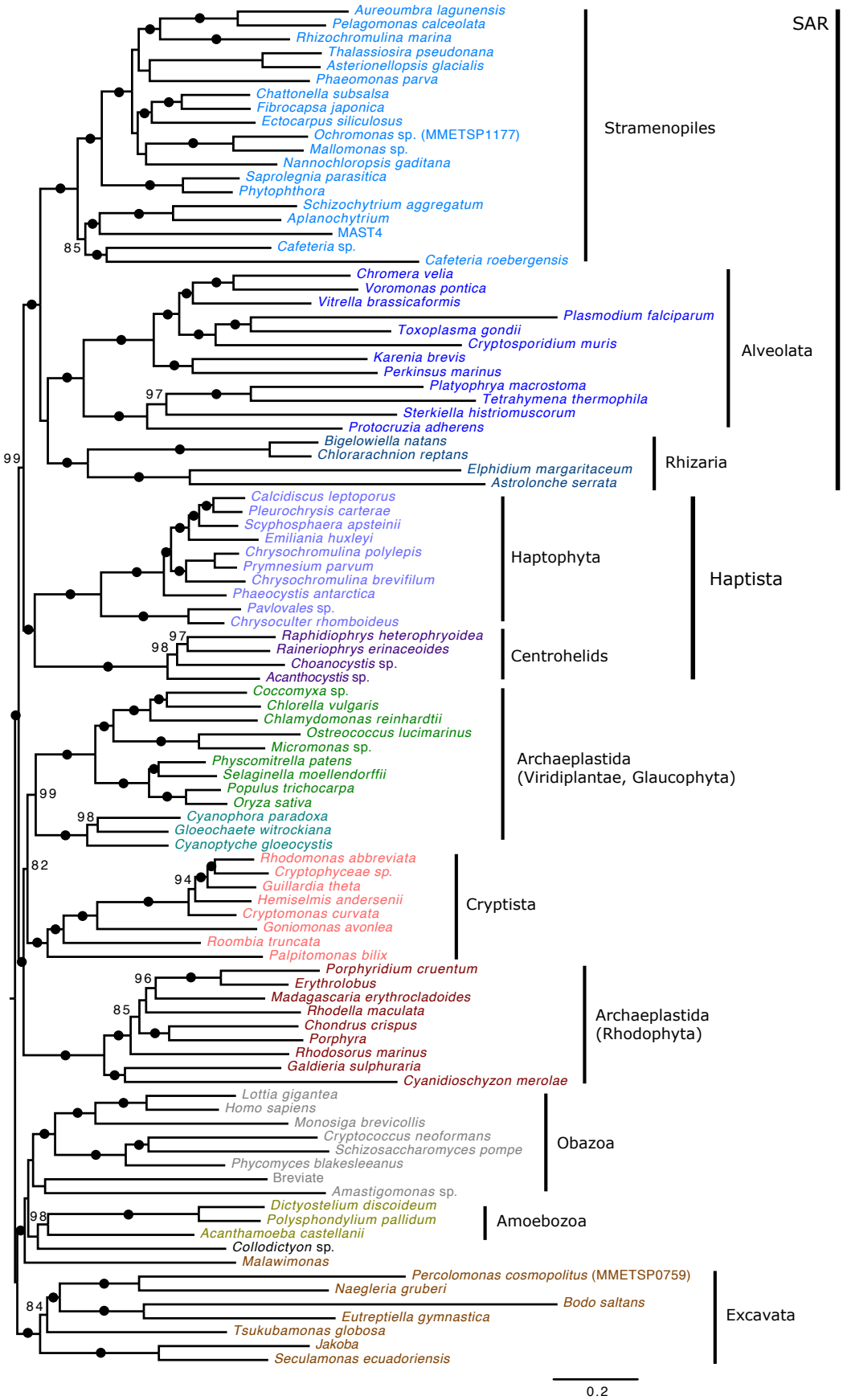

Aureoumbra lagunensis
Pelagomonas calceolata
Rhizochromulina marina
Thalassiosira pseudonana
Asterionellopsis glacialis
Phaeomonas parva
Chattonella subsalsa
Fibrocapsa japonica
Ectocarpus siliculosus
Ochromonas sp. (MMETSP1177)
Mallomonas sp.
Nannochloropsis gaditana
Saprolegnia parasitica
Phytophthora
Schizochytrium aggregatum
Aplanochytrium
MAST4
Cafeteria sp.
Cafeteria roebergensis
Stramenopiles
SAR
Chromera velia
Voromonas pontica
Vitrella brassicaformis
Plasmodium falciparum
Toxoplasma gondii
Cryptosporidium muris
Karenia brevis
Perkinsus marinus
Platyophrya macrostoma
Tetrahymena thermophila
Sterkiella histriomuscorum
Protocruzia adherens
Alveolata
Bigelowiella natans
Chlorarachnion reptans
Elphidium margaritaceum
Astrolonche serrata
Rhizaria
Calcidiscus leptoporus
Pleurochrysis carterae
Scyphosphaera apsteinii
Emiliania huxleyi
Chrysochromulina polylepis
Prymnesium parvum
Chrysochromulina brevifilum
Phaeocystis antarctica
Pavlovales sp.
Chrysoculter rhomboideus
Haptophyta
Raphidiophrys heterophryoidea
Raineriophrys erinaceoides
Choanocystis sp.
Acanthocystis sp.
Centrohelids
Haptista
Coccomyxa sp.
Chlorella vulgaris
Chlamydomonas reinhardtii
Ostreococcus lucimarinus
Micromonas sp.
Physcomitrella patens
Selaginella moellendorffii
Populus trichocarpa
Oryza sativa
Archaeplastida
(Viridiplantae, Glaucophyta)
Cyanophora paradoxa
Gloeochaete witrockiana
Cyanoptyche gloeocystis
Rhodomonas abbreviata
Cryptophyceae sp.
Guillardia theta
Hemiselmis andersenii
Cryptomonas curvata
Goniomonas avonlea
Roombia truncata
Palpitomonas bilix
Cryptista
Porphyridium cruentum
Erythrolobus
Madagascaria erythrocladoides
Rhodella maculata
Chondrus crispus
Porphyra
Rhodosorus marinus
Galdieria sulphuraria
Cyanidioschyzon merolae
Archaeplastida
(Rhodophyta)
Lottia gigantea
Homo sapiens
Monosiga brevicollis
Cryptococcus neoformans
Schizosaccharomyces pompe
Phycomyces blakesleeanus
Breviate
Amastigomonas sp.
Obazoa
Dictyostelium discoideum
Polysphondylium pallidum
Acanthamoeba castellanii
Amoebozoa
Collodictyon sp.
Malawimonas
Percolomonas cosmopolitus (MMETSP0759)
Naegleria gruberi
Bodo saltans
Eutreptiella gymnastica
Tsukubamonas globosa
Jakoba
Seculamonas ecuadoriensis
Excavata
85
97
99
97
98
99
98
82
94
96
85
98
84
0.2

Figure 3.11: Maximum likelihood (ML) phylogeny of a 250 marker gene set as in Burki et al. (2016a) with Cryptista removed from the dataset. The phylogeny is based on a concatenated marker gene alignment of 71,477 unambiguously aligned sites across 90 OTUs. The ML tree shown was generated under the model LG + C60 + F + PMSF with 100 standard bootstrap replicates and has been rooted in mid-point. Black dots indicate maximal support for a particular node. When not maximal, only bootstrap support values > 70% are shown. The scale bar shows an inferred 0.2 substitutions per site. For consistency, taxon names are provided as in Burki et al. (2016a).

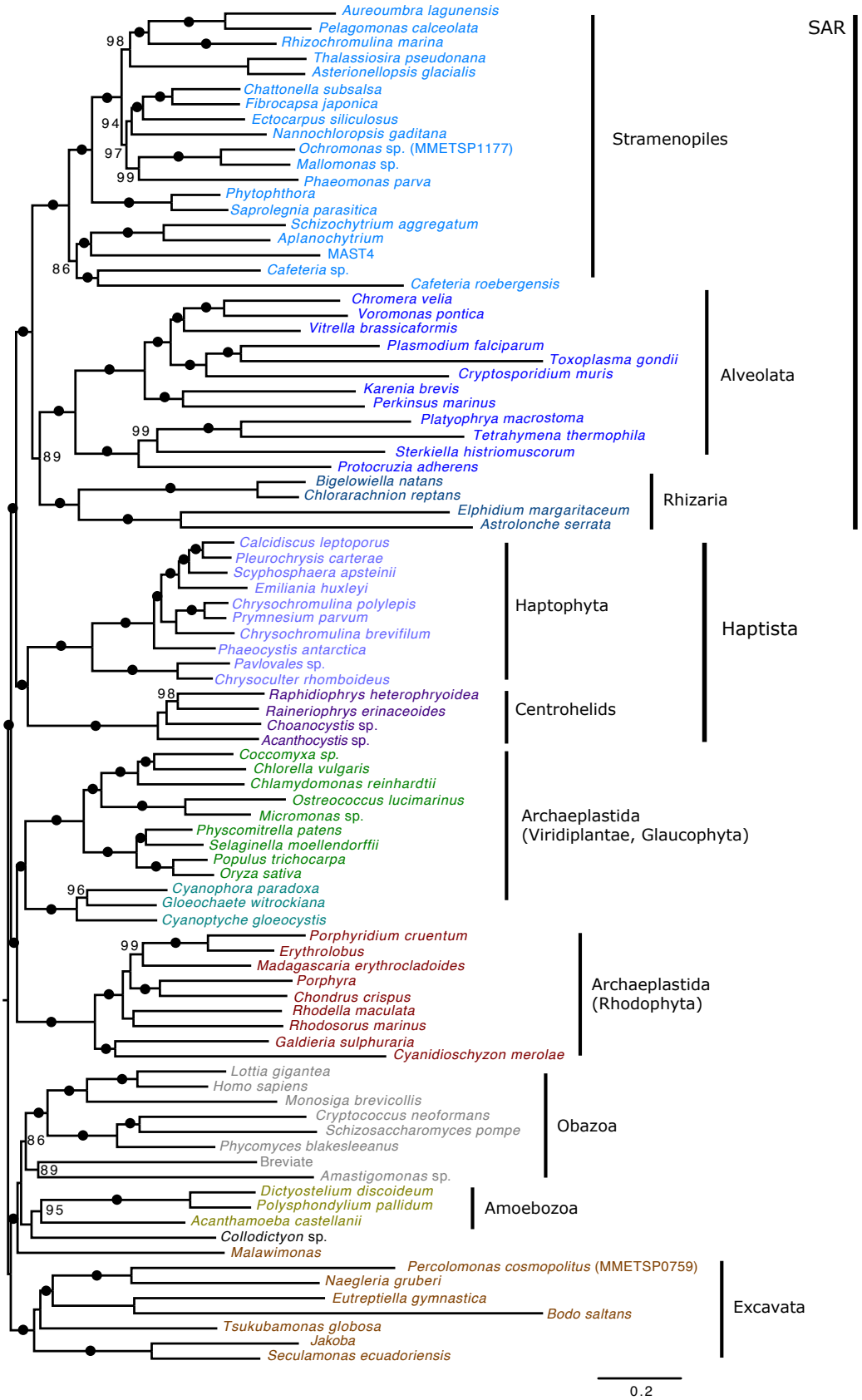
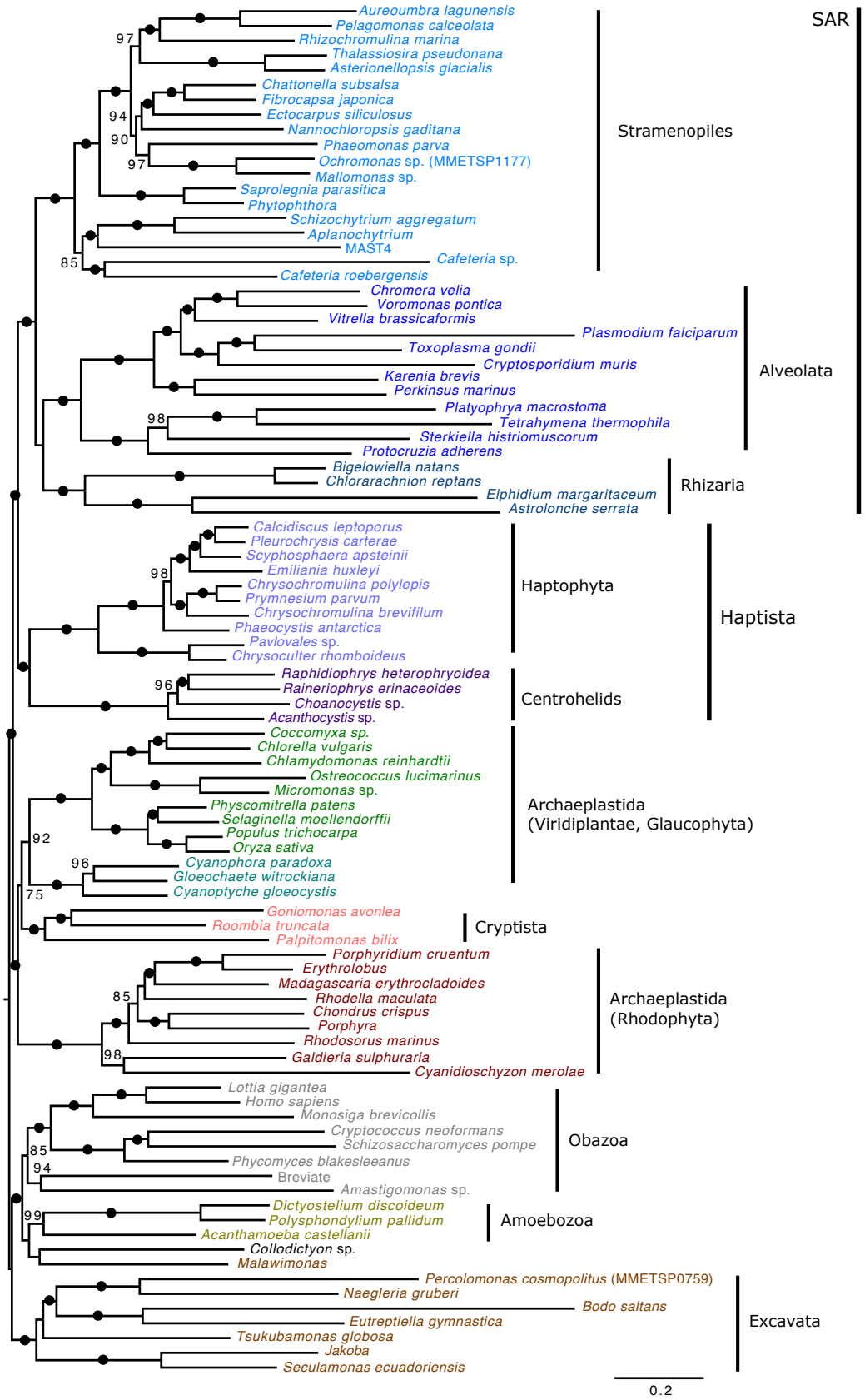

Aureoumbra lagunensis
Pelagomonas calceolata
Rhizochromulina marina
Thalassiosira pseudonana
Asterionellopsis glacialis
Chattonella subsalsa
Fibrocapsa japonica
Ectocarpus siliculosus
Nannochloropsis gaditana
Ochromonas sp. (MMETSP1177)
Mallomonas sp.
Phaeomonas parva
Phytophthora
Saprolegnia parasitica
Schizochytrium aggregatum
Aplanochytrium
MAST4
Cafeteria sp.
Cafeteria roebergensis
Stramenopiles
SAR
Chromera velia
Voromonas pontica
Vitrella brassicaformis
Plasmodium falciparum
Toxoplasma gondii
Cryptosporidium muris
Karenia brevis
Perkinsus marinus
Platyophrya macrostoma
Tetrahymena thermophila
Sterkiella histriomuscorum
Protocruzia adherens
Alveolata
Bigelowiella natans
Chlorarachnion reptans
Elphidium margaritaceum
Astrolonche serrata
Rhizaria
Calcidiscus leptoporus
Pleurochrysis carterae
Scyphosphaera apsteinii
Emiliania huxleyi
Chrysochromulina polylepis
Prymnesium parvum
Chrysochromulina brevifilum
Phaeocystis antarctica
Pavlovales sp.
Chrysoculter rhomboideus
Haptophyta
Raphidiophrys heterophryoidea
Raineriophrys erinaceoides
Choanocystis sp.
Acanthocystis sp.
Centrohelids
Haptista
Coccomyxa sp.
Chlorella vulgaris
Chlamydomonas reinhardtii
Ostreococcus lucimarinus
Micromonas sp.
Physcomitrella patens
Selaginella moellendorffii
Populus trichocarpa
Oryza sativa
Cyanophora paradoxa
Gloeochaete witrockiana
Cyanoptyche gloeocystis
Archaeplastida
(Viridiplantae, Glaucophyta)
Porphyridium cruentum
Erythrolobus
Madagascaria erythrocladoides
Porphyra
Chondrus crispus
Rhodella maculata
Rhodosorus marinus
Galdieria sulphuraria
Cyanidioschyzon merolae
Archaeplastida
(Rhodophyta)
Lottia gigantea
Homo sapiens
Monosiga brevicollis
Cryptococcus neoformans
Schizosaccharomyces pompe
Phycomyces blakesleeanus
Breviate
Amastigomonas sp.
Obazoa
Dictyostelium discoideum
Polysphondylium pallidum
Acanthamoeba castellanii
Collodictyon sp.
Malawimonas
Amoebozoa
Percolomonas cosmopolitus (MMETSP0759)
Naegleria gruberi
Eutreptiella gymnastica
Bodo saltans
Tsukubamonas globosa
Jakoba
Seculamonas ecuadoriensis
Excavata
98
94
97
99
86
99
89
98
99
96
99
86
89
95
0.2

Figure 3.12: Maximum likelihood (ML) phylogeny of a 250 marker gene set as in Burki et al. (2016a) with plastid-bearing Cryptista (cryptophytes) removed from the dataset. The phylogeny is based on a concatenated marker gene alignment of 71,277 unambiguously aligned sites across 93 OTUs that includes new transcriptome data from *Goniomonas avonlea*. The ML tree shown was generated under the model LG + C60 + F + PMSF with 100 standard bootstrap replicates and has been rooted in mid-point. Black dots indicate maximal support for a particular node. When not maximal, only bootstrap support values > 70% are shown. The scale bar shows an inferred 0.2 substitutions per site. For consistency, taxon names are provided as in Burki et al. (2016a).


Aureoumbra lagunensis
Pelagomonas calceolata
Rhizochromulina marina
Thalassiosira pseudonana
Asterionellopsis glacialis
Chattonella subsalsa
Fibrocapsa japonica
Ectocarpus siliculosus
Nannochloropsis gaditana
Phaeomonas parva
Ochromonas sp. (MMETSP1177)
Mallomonas sp.
Saprolegnia parasitica
Phytophthora
Schizochytrium aggregatum
Aplanochytrium
MAST4
Cafeteria sp.
Cafeteria roebergensis
Stramenopiles
SAR
Chromera velia
Voromonas pontica
Vitrella brassicaformis
Plasmodium falciparum
Toxoplasma gondii
Cryptosporidium muris
Karenia brevis
Perkinsus marinus
Platyophrya macrostoma
Tetrahymena thermophila
Sterkiella histriomuscorum
Protocruzia adherens
Alveolata
Bigelowiella natans
Chlorarachnion reptans
Elphidium margaritaceum
Astrolonche serrata
Rhizaria
Calcidiscus leptoporus
Pleurochrysis carterae
Scyphosphaera apsteinii
Emiliania huxleyi
Chrysochromulina polylepis
Prymnesium parvum
Chrysochromulina brevifilum
Phaeocystis antarctica
Pavlovales sp.
Chrysoculter rhomboideus
Haptophyta
Haptista
Raphidiophrys heterophryoidea
Raineriophrys erinaceoides
Choanocystis sp.
Acanthocystis sp.
Centrohelids
Coccomyxa sp.
Chlorella vulgaris
Chlamydomonas reinhardtii
Ostreococcus lucimarinus
Micromonas sp.
Physcomitrella patens
Selaginella moellendorffii
Populus trichocarpa
Oryza sativa
Archaeplastida
(Viridiplantae, Glaucophyta)
Cyanophora paradoxa
Gloeochaete witrockiana
Cyanoptyche gloeocystis
Goniomonas avonlea
Roombia truncata
Palpitomonas bilix
Cryptista
Porphyridium cruentum
Erythrolobus
Madagascaria erythrocladoides
Rhodella maculata
Chondrus crispus
Porphyra
Rhodosorus marinus
Galdieria sulphuraria
Cyanidioschyzon merolae
Archaeplastida
(Rhodophyta)
Lottia gigantea
Homo sapiens
Monosiga brevicollis
Cryptococcus neoformans
Schizosaccharomyces pompe
Phycomyces blakesleeanus
Breviate
Amastigomonas sp.
Obazoa
Dictyostelium discoideum
Polysphondylium pallidum
Acanthamoeba castellanii
Collodictyon sp.
Malawimonas
Amoebozoa
Percolomonas cosmopolitus (MMETSP0759)
Naegleria gruberi
Bodo saltans
Eutreptiella gymnastica
Tsukubamonas globosa
Jakoba
Seculamonas ecuadoriensis
Excavata
97
94
90
97
85
98
96
92
96
75
85
94
99
98
0.2

bootstrap support) to the exclusion of Rhodophyta with 75% bootstrap support. Using standard error of bootstrap value with a 95% confidence interval, it was determined that for 100 replicates this support value (75%) is not significantly different from the support value for this positioning when all Cryptista are included (82%). However, inferring a phylogeny using a second dataset based on Kang et al. (2017) consisting of 105 OTUs and 351 marker genes (see Supplementary Tables A4 and A5) with *G. avonlea* added to it resulted in a different topology (Figure 3.13). Here, the monophyly of all eukaryotic super groups was recovered with maximum or near maximum support, including a monophyletic Archaeplastida (97%) branching sister to Cryptista (100%).

Assessing the topology of all single marker-gene trees in the dataset based on Burki et al. (2016a) for specific genes in specific OTUs that display discordant topologies (as determined by considering pairwise nodal distances between OTUs and comparing this to a consensus topology) using PhyloMCOA (De Vienne et al. 2012) resulted in the identification and removal of 223 outlier sequences from various OTUs (see Supplementary Table A3). The phylogeny inferred based on this modified dataset resolved identical topologies to those previously observed (Figure 3.14). Here, the support for the position of Cryptista internal to Archaeplastida and sister to the Viridiplantae and Glaucophyta clade increased to 90% (significantly different from a bootstrap of 82% as determined using standard error of a bootstrap value at a 95% confidence interval). Inferring phylogenies based on random subsets of the marker-gene set (46 or 47 of the 250 marker-genes) and evaluating the phylogenetic position of Cryptista within them resulted in observing a consistent relationship with one or more Archaeplastida groups in 93 of 100 randomly generated gene sets (Figure 3.15). While Cryptista was most frequently observed sister to the Viridiplantae and Glaucophyta clade (30%), 24% of trees showed an exclusive relationship with Glaucophyta, 13% exclusively with Rhodophyta and 20% as sister to a monophyletic Archaeplastida. Notably, a sister relationship between Cryptista and Haptophyta was never observed.

Figure 3.13: Maximum likelihood (ML) phylogeny of a 351 marker gene set as in Kang et al. (2017) that includes new transcriptome data from *Goniomonas avonlea*. The phylogeny is based on a concatenated marker gene alignment of 99,322 unambiguously aligned sites across 105 OTUs. The ML tree shown was generated under the model LG + C60 + F + PMSF with 100 standard bootstrap replicates and has been rooted in mid-point. Black dots indicate maximal support for a particular node. When not maximal, only bootstrap support values > 70% are shown. The scale bar shows an inferred 0.2 substitutions per site.

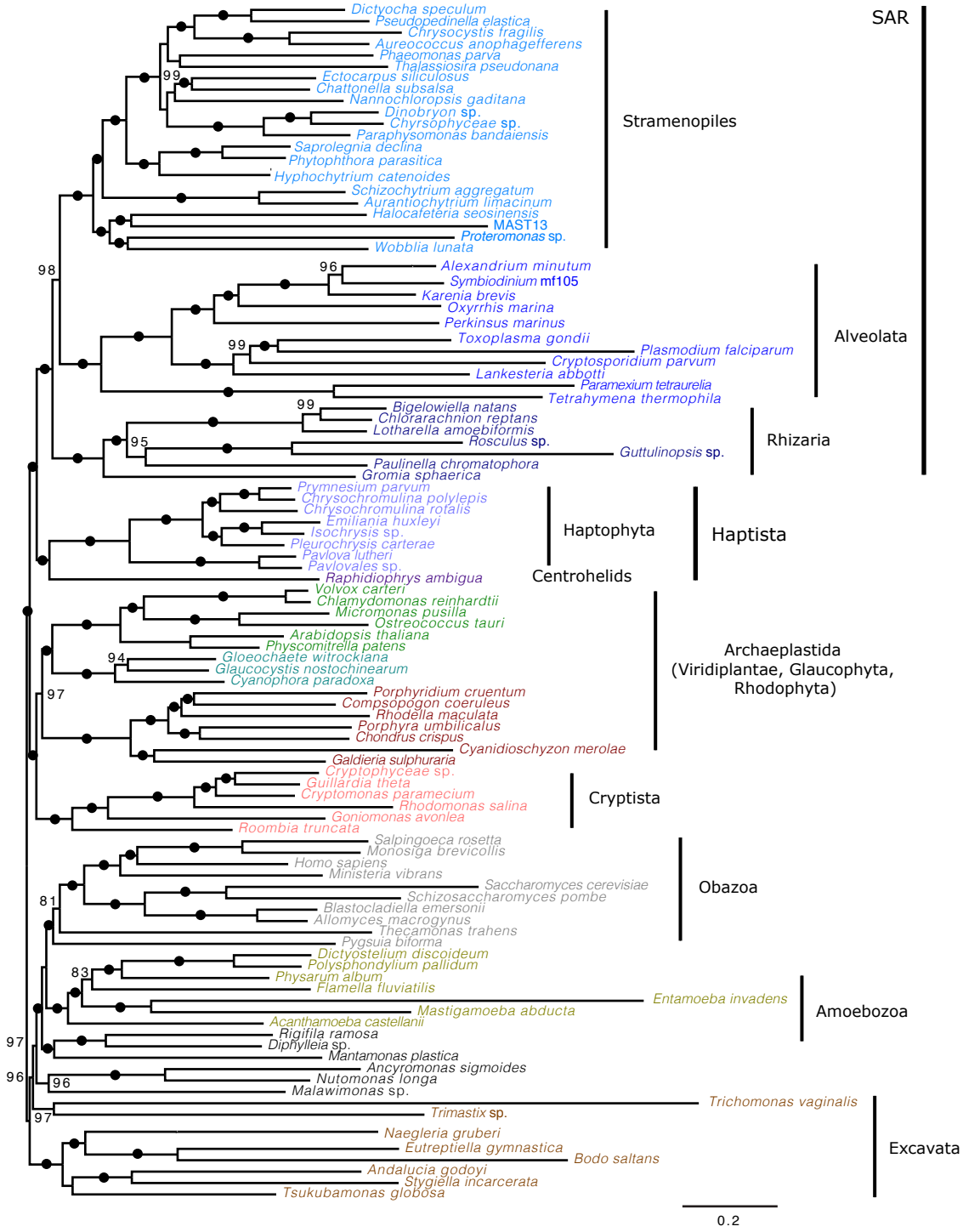

Dictyocha speculum
Pseudopedinella elastica
Chrysocystis fragilis
Aureococcus anophagefferens
Phaeomonas parva
Thalassiosira pseudonana
Ectocarpus siliculosus
Chattonella subsalsa
Nannochloropsis gaditana
Dinobryon sp.
Chyrsophyceae sp.
Paraphysomonas bandaiensis
Saprolegnia declina
Phytophthora parasitica
Hyphochytrium catenoides
Schizochytrium aggregatum
Aurantiochytrium limacinum
Halocafeteria seosinensis
MAST13
Proteromonas sp.
Wobblia lunata
Stramenopiles
SAR
Alexandrium minutum
Symbiodinium mf105
Karenia brevis
Oxyrrhis marina
Perkinsus marinus
Toxoplasma gondii
Plasmodium falciparum
Cryptosporidium parvum
Lankesteria abbotti
Paramexium tetraurelia
Tetrahymena thermophila
Alveolata
Bigelowiella natans
Chlorarachnion reptans
Lotharella amoebiformis
Rosculus sp.
Guttulinopsis sp.
Paulinella chromatophora
Gromia sphaerica
Rhizaria
Prymnesium parvum
Chrysochromulina polylepis
Chrysochromulina rotalis
Emiliania huxleyi
Isochrysis sp.
Pleurochrysis carterae
Pavlova lutheri
Pavlovales sp.
Haptophyta
Haptista
Raphidiophrys ambigua
Centrohelids
Volvox carteri
Chlamydomonas reinhardtii
Micromonas pusilla
Ostreococcus tauri
Arabidopsis thaliana
Physcomitrella patens
Gloeochaete witrockiana
Glaucocystis nostochinearum
Cyanophora paradoxa
Porphyridium cruentum
Compsopogon coeruleus
Rhodella maculata
Porphyra umbilicalus
Chondrus crispus
Cyanidioschyzon merolae
Galdieria sulphuraria
Archaeplastida
(Viridiplantae, Glaucophyta,
Rhodophyta)
Cryptophyceae sp.
Guillardia theta
Cryptomonas paramecium
Rhodomonas salina
Goniomonas avonlea
Roombia truncata
Cryptista
Salpingoeca rosetta
Monosiga brevicollis
Homo sapiens
Ministeria vibrans
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Blastocladiella emersonii
Allomyces macrogynus
Thecamonas trahens
Pygsuia biforma
Obazoa
Dictyostelium discoideum
Polysphondylium pallidum
Physarum album
Flamella fluviatilis
Entamoeba invadens
Mastigamoeba abducta
Acanthamoeba castellanii
Rigifila ramosa
Diphylleia sp.
Mantamonas plastica
Ancyromonas sigmoides
Nutomonas longa
Malawimonas sp.
Amoebozoa
Trichomonas vaginalis
Trimastix sp.
Naegleria gruberi
Eutreptiella gymnastica
Bodo saltans
Andalucia godoyi
Stygiella incarcerata
Tsukubamonas globosa
Excavata
98
96
99
99
95
99
94
97
81
83
97
96
96
97
0.2

Figure 3.14: Maximum likelihood (ML) phylogeny of a 250 marker gene set as in Burki et al. (2016a) generated after removal of specific genes in individual taxa that were determined to produce a discordant signal via analysis using PhyloMCOA. The phylogeny is based on a concatenated marker gene alignment of 71,425 unambiguously aligned sites across 98 OTUs that includes new transcriptome data from *Goniomonas avonlea*. The ML tree shown was generated under the model LG + C60 + F + PMSF with 100 standard bootstrap replicates and has been rooted in mid-point. Black dots indicate maximal support for a particular node. When not maximal, only bootstrap support values > 70% are shown. The scale bar shows an inferred 0.2 substitutions per site.

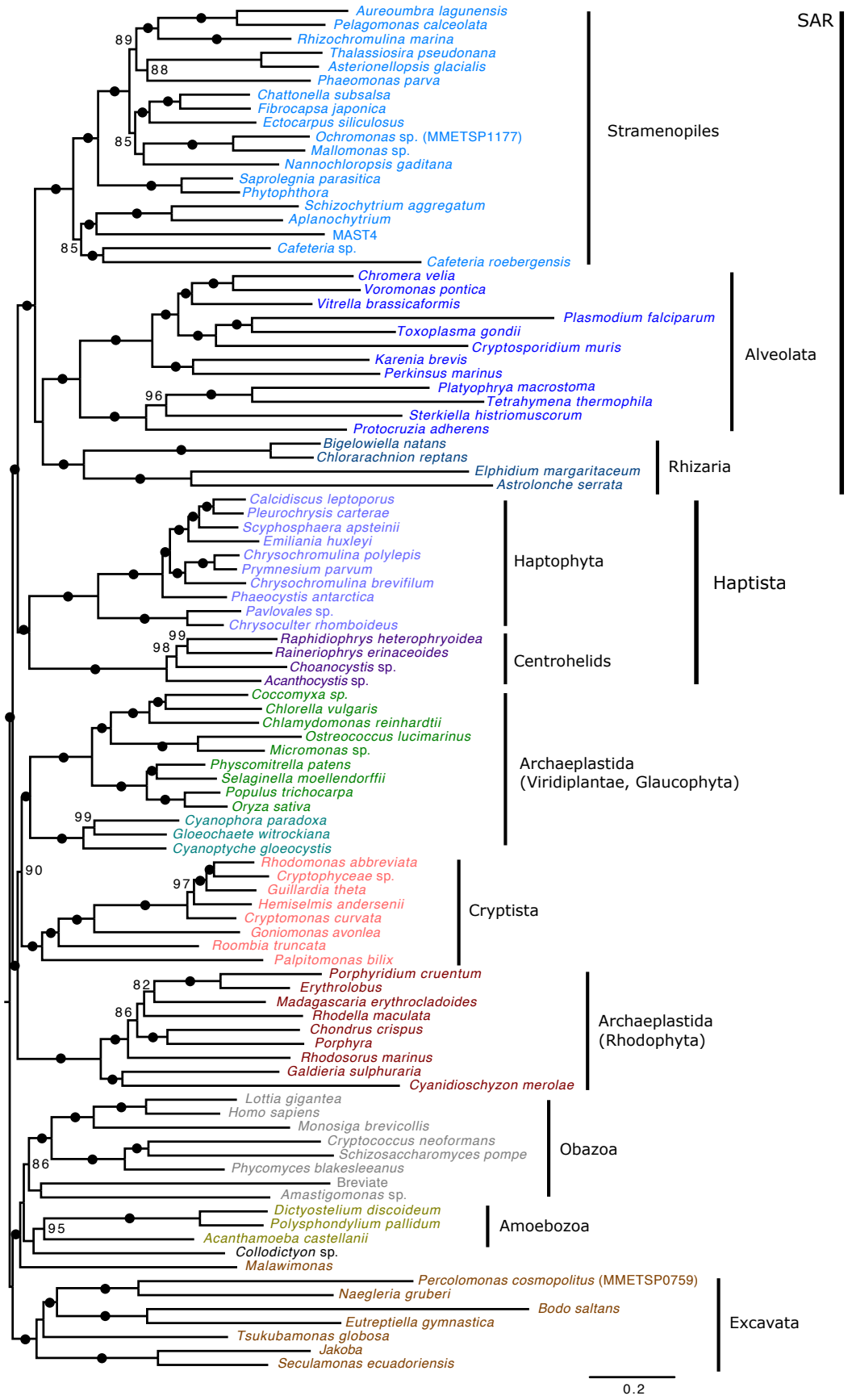
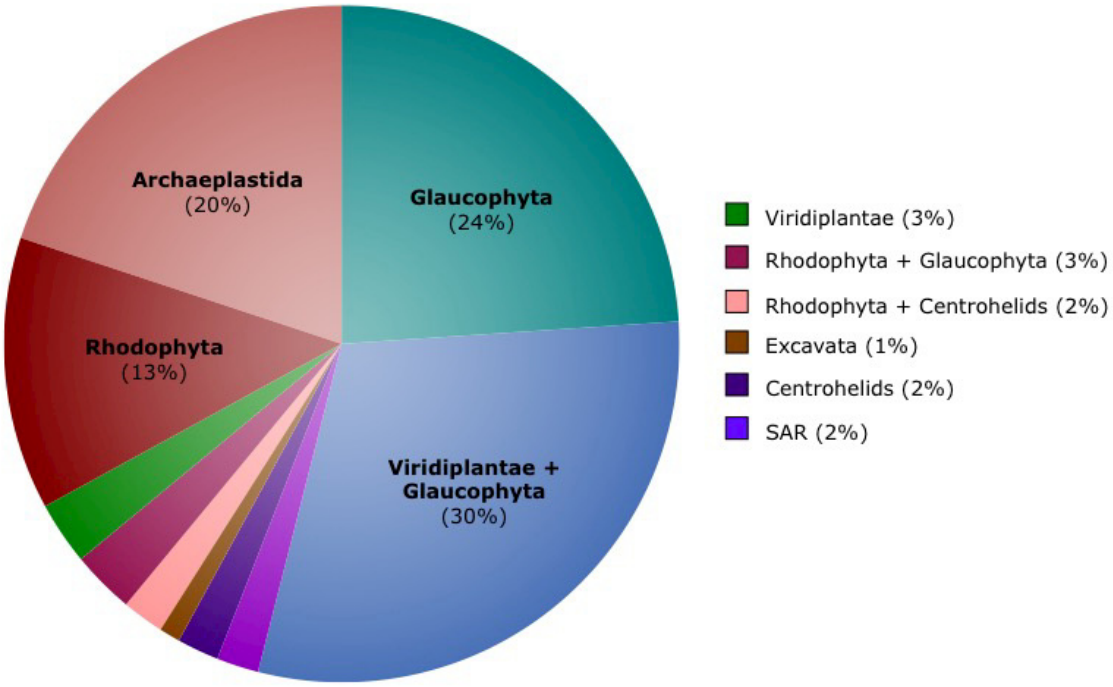

Aureoumbra lagunensis
Pelagomonas calceolata
Rhizochromulina marina
Thalassiosira pseudonana
Asterionellopsis glacialis
Phaeomonas parva
Chattonella subsalsa
Fibrocapsa japonica
Ectocarpus siliculosus
Ochromonas sp. (MMETSP1177)
Mallomonas sp.
Nannochloropsis gaditana
Saprolegnia parasitica
Phytophthora
Schizochytrium aggregatum
Aplanochytrium
MAST4
Cafeteria sp.
Cafeteria roebergensis
Stramenopiles
SAR
Chromera velia
Voromonas pontica
Vitrella brassicaformis
Plasmodium falciparum
Toxoplasma gondii
Cryptosporidium muris
Karenia brevis
Perkinsus marinus
Platyophrya macrostoma
Tetrahymena thermophila
Sterkiella histriomuscorum
Protocruzia adherens
Alveolata
Bigelowiella natans
Chlorarachnion reptans
Elphidium margaritaceum
Astrolonche serrata
Rhizaria
Calcidiscus leptoporus
Pleurochrysis carterae
Scyphosphaera apsteinii
Emiliania huxleyi
Chrysochromulina polylepis
Prymnesium parvum
Chrysochromulina brevifilum
Phaeocystis antarctica
Pavlovales sp.
Chrysoculter rhomboideus
Haptophyta
Raphidiophrys heterophryoidea
Raineriophrys erinaceoides
Choanocystis sp.
Acanthocystis sp.
Centrohelids
Haptista
Coccomyxa sp.
Chlorella vulgaris
Chlamydomonas reinhardtii
Ostreococcus lucimarinus
Micromonas sp.
Physcomitrella patens
Selaginella moellendorffii
Populus trichocarpa
Oryza sativa
Archaeplastida
(Viridiplantae, Glaucophyta)
Cyanophora paradoxa
Gloeochaete witrockiana
Cyanoptyche gloeocystis
Rhodomonas abbreviata
Cryptophyceae sp.
Guillardia theta
Hemiselmis andersenii
Cryptomonas curvata
Goniomonas avonlea
Roombia truncata
Palpitomonas bilix
Cryptista
Porphyridium cruentum
Erythrolobus
Madagascaria erythrocladoides
Rhodella maculata
Chondrus crispus
Porphyra
Rhodosorus marinus
Galdieria sulphuraria
Cyanidioschyzon merolae
Archaeplastida
(Rhodophyta)
Lottia gigantea
Homo sapiens
Monosiga brevicollis
Cryptococcus neoformans
Schizosaccharomyces pombe
Phycomyces blakesleeanus
Breviate
Amastigomonas sp.
Obazoa
Dictyostelium discoideum
Polysphondylium pallidum
Acanthamoeba castellanii
Collodictyon sp.
Malawimonas
Amoebozoa
Percolomonas cosmopolitus (MMETSP0759)
Naegleria gruberi
Bodo saltans
Eutreptiella gymnastica
Tsukubamonas globosa
Jakoba
Seculamonas ecuadoriensis
Excavata
89
88
85
85
85
96
99
98
99
90
97
82
86
86
95
0.2

Figure 3.15: The phylogenetic position of Cryptista within each ML tree inferred under the model LG+C20+F using randomly generated subsets of marker genes (46 or 47 of the 250 marker-genes) from the Burki et al. (2016) dataset. Only marker genes for which a homolog was present in *Goniomonas avonlea* and at least one additional Cryptista were included. The distribution shown is based on a total of 100 randomly generated marker gene subset trees.


Archaeplastida
(20%)
Glaucophyta
(24%)
Rhodophyta
(13%)
Viridiplantae +
Glaucophyta
(30%)
Viridiplantae (3%)
Rhodophyta + Glaucophyta (3%)
Rhodophyta + Centrohelids (2%)
Excavata (1%)
Centrohelids (2%)
SAR (2%)

3.4 DISCUSSION

By all ultrastructural evidence, goniomonads like *G. avonlea* appear to lack an active or vestigial plastid (Hill 1991; Kim and Archibald 2013). Although no sequence data exist from freshwater goniomonads (outside of 18S sequences, and a few select genes), sequence data does exist for two other marine *Goniomonas* species and were retained in the database used in analysis of *G. avonlea* single gene trees. Data from *Goniomonas* sp. stems from an expressed sequence tag (EST) project, which are typically used for the purpose of gene discovery and are incomplete in terms of gene coverage (Parkinson and Blaxter 2009; Philippe et al. 2011). On the other hand, sequences from *G. pacifica* generated from the MMETSP dataset seems to contain a high level of contamination and should be interpreted with caution (a universal low level of contamination in sequenced transcriptomes was identified in the MMETSP dataset by Keeling et al. (2014)). The genome data generated for *G. avonlea* was found to be highly fragmented and incomplete while the transcriptome appeared to be missing some key metabolic enzymes and partially incomplete. Thus, for this study a hybrid dataset of transcriptome-predicted proteins supplemented with non-redundant predicted gene models from the genome was used to ensure that the most complete set of protein coding genes was considered. Until this analysis, no genome-wide study had been performed in search of a significant red-algal footprint in goniomonads to try to determine if the lineage experienced a secondary loss of photosynthesis or was ancestrally non-photosynthetic.

3.4.1 *G. avonlea* and *G. theta* share few potential red-algal EGTs

In a genome-wide analysis of the cryptophyte *G. theta* by Curtis et al. (2012), 508 genes were predicted to be algal EGTs on the basis of phylogenetics. A homolog to each of the 508 predicted algal EGTs in *G. theta* (Curtis et al. 2012) was not expected in *G. avonlea* as a proportion of these are involved in maintenance and function of the PPC and plastid. Many, however, are predicted to have been repurposed and function in the host cytosol and host compartments. If *G. avonlea* secondarily lost the red-algal derived plastid found in

plastid-bearing cryptophytes, it is expected that a significant amount of common EGTs showing red-algal ancestry would be present in its genome (Stiller et al. 2009; Curtis et al. 2012). Of the 508 predicted algal EGTs in *G. theta*, 144 were found to have a homolog in *G. avonlea*. Only nine of these showed a topology consistent with common red-algal ancestry (see Table 3.1) while 51 showed a common origin of clear green/glaucophyte algal ancestry, 14 of which did not contain a single red algal homolog anywhere in the tree, and 10 showed a common ambiguous algal ancestry.

Is this red-algal footprint in *G. avonlea* significant? Stiller et al. (2009) suggested that when searching for significant EGT in heterotrophic species the relative signal to red algae and a distantly related control taxa known to have an entirely non-photosynthetic evolutionary history should be compared. If the proposed algal signal is due to EGT it should be comparatively stronger than the signal to the chosen control, whose observed frequency of phylogenetic affinity represents the expected outcome of background phylogenetic noise. Here, any taxa from the super-group Amoebozoa could be considered a negative control as it meets the criteria outlined above. In the targeted approach of searching for previously predicted algal EGTs in *G. theta*, *G. avonlea* branched with Amoebozoa in 16 phylogenies – seven more phylogenies than those with a topology suggesting a shared red-algal footprint. This suggests that the observed red-algal footprint in *G. avonlea* is not greater than what is expected due to background phylogenetic noise and is, therefore, not significant. This is in contrast to the clear red-algal footprint observed in *G. theta* (with or without *G. avonlea*) in 26/144 phylogenies inferred here, greater than the number of occurrences in which it branches with the amoebozoan control (1/144). Furthermore, when taking into consideration predicted subcellular localization of the nine-potential shared red-algal EGTs in *G. avonlea*, none appear to contain both a signal and transit peptide. Three of these are predicted to contain signal peptides only and function as sodium proton exchangers; these, however, cannot be definitively assigned as plastid functioning as they could be targeted anywhere in the host endomembrane system and are commonly present in lipid bilayers across all domains of life (Orlowski and Grinstein 2004).

While there does not appear to be a significant shared red-algal footprint between *G. avonlea* and *G. theta*, a substantial proportion of phylogenies show a common affinity

to green and/or glaucophyte algae (51/144 trees, 14/51 with no red-algal homolog present). This mirrors the analysis performed by Curtis et al. (2012) where approximately half of the predicted algal EGTs showed an unambiguous green/glaucophyte ancestry (and over half of these had no red-algal homolog). At the time, these were considered as likely stemming from an algal endosymbiont; the unresolved phylogenetic positioning of Cryptista, however, meant that there were uncertainties in assigning these as predicted algal EGTs and in interpreting the green-red mosaicism observed (Curtis et al. 2012). Now, whether are not these can be conclusively associated with plastid-ancestry is even more unclear because of the highly supported relationship of Cryptista and Archaeplastida (perhaps even branching specifically with Viridiplantae and Glaucophyta) in nuclear gene trees observed both here (see Section 3.4.4 and Figures 3.10 to 3.14), in Burki et al. (2016a), and, more uncertainly, previous phylogenomic analyses too (e.g. Burki et al. 2012b, Brown et al. 2013). As a result, without an in-tact plastid targeting signal or specific plastid function, these 'green genes' could in fact be attributed to vertical inheritance. In the case of *G. avonlea*, none of the predicted green/glaucophyte algal genes show evidence of explicit plastid targeting or function (data not shown).

It is possible that some of the phylogenies showing an ambiguous algal signal or green/glaucophyte algal signal (particularly when at least one red algal homolog is also present) are truly of red algal origin and exist in the genome as a consequence of plastid ancestry and EGT. There is a significantly greater representation of green algae/land plant transcriptomes and genomes available in databases compared to that of red or glaucophyte algae (Sibbald and Archibald 2017) resulting in a taxonomic sampling bias when building phylogenies. With the sequencing of underrepresented primary algal lineages, the number of predicted red-algal EGTs in cryptophytes and *G. avonlea* (as well as other complex red-alga plastid bearing taxa) may increase (Curtis et al. 2012). Other factors such as phylogenetic artifacts and poor branching resolution add to the complications of determining whether an algal gene is of red or green ancestry. As mentioned above, untangling the sources of genes in the nucleus of Cryptista is complicated, particularly in light of their nuclear archaeplastidal relationship. Detecting EGT relies upon the ability to differentiate nuclear evolutionary history from that of secondary plastid evolutionary history and becomes much more difficult when these two sources are closely related (and,

as an added complication, presumably makes LGT easier; Archibald 2015). Overall, there is currently little evidence for the heterotrophic *G. avonlea* having a significant red-algal footprint in common with *G. theta,* or for the presence of a red-algal secondary plastid in their common ancestor.

3.4.2 A BLAST-based analysis of algal signal in *G. avonlea*

The top BLAST hit (i.e., most significant homolog) does not always reflect the nearest neighbor in a phylogenetic analysis (Koski and Golding 2001). Homologous hits to a query sequence using BLAST is heavily influenced by the composition of the database making it less likely to retrieve hits to taxa with less database representation (Stiller et al. 2009). Furthermore, sequence homology searches like BLAST can be misled by evolutionary rate variation and multi-domain proteins (Eisen 2000). Even though a top BLAST hit approach is error prone, it provides attractive advantages in terms of speed and automation (Eisen 2000) while still allowing for one to examine trends in relationships of a species to broad eukaryotic lineages (Stiller et al. 2009). Additionally, this approach allows one to consider predicted proteins and/or gene models in a dataset that have less than four significant homologs and for which a phylogenetic tree cannot be inferred (potentially representing a species specific LGT; Stiller et al. 2009).

In an analysis of the top BLAST hit to each *G. avonlea* predicted protein/gene model included in the dataset (Figure 3.6), 27% of sequences had a highest scoring homolog to another member of Cryptista. This was expected as they are members of the same phylum. The next eukaryotic group *G. avonlea* had the most best matches to (14% of queries), however, was the unrelated Obazoa (opisthokonts, breviates and apusomonads; Brown et al. 2013). A similar result was found in an analysis of *G. theta* by Curtis (2012) where 18% of top matches were found to be opisthokonts. Why there is such a strong signal of both *G. avonlea* and *G. theta* to Obazoa in these BLAST based analyses is uncertain; it may be reflective of the database composition as there is an over-representation of Obazoa in terms of genome/transcriptome availability (Sibbald and Archibald 2017) and should be interpreted with caution. Nevertheless, a significant proportion of these datasets also produced phylogenies supporting this affiliation (discussed in Section 3.4.3).

Of note is the abundance of top hits to Viridiplantae and Glaucophyta (~13%) in the *G. avonlea* dataset. While this suggests a fairly strong green/glaucophyte algal signal in *G. avonlea*, it may be showing vertical rather than endosymbiotic ancestry due to their relationship with Cryptista in most recent multi-gene phylogenies (Burki et al. 2016a; discussed in Section 3.4.4). When it comes to red algae, there is no indication of significant signal. As can be seen in Figure 3.6, there were considerably more top hits to an amoebozoan control taxa (~6%) compared to rhodophytes (~1%). This suggests that the observed red-algal affinity is most likely due to background phylogenetic noise and likely not EGT. When considering the contribution of red-algal and amoebozoan top hits in both *G. avonlea* and *G. theta* (taking into account database composition), *G. theta* showed a 3.6-fold greater red-algal signal (Table 3.2). A comparative signal to red algae in the nucleus of *G. avonlea*, as was found in *G. theta*, was not expected due to the loss of plastid-functioning/related genes that are no longer essential (Stiller et al. 2009) – however, there is no indication of excess signal in *G. avonlea* on the basis of top BLAST hits. This is similar to the outcome of a study by Stiller et al. (2009) on the red-algal contribution to the genomes of heterotrophic oomycetes in comparison to their photosynthetic diatom relatives where they concluded that there was no unusual red-algal contribution in the oomycete genomes examined.

3.4.3 Searching for an algal signal using single gene trees

As expected, many of *G. avonlea* sequences for which meaningful trees could be inferred branched within or sister to Cryptista in their corresponding single gene tree (Figure 3.7). Similar to the BLAST-based approach discussed above, the second most frequently observed affinity was to Obazoa. Additionally, during the second round of pattern detection, the next-nearest neighbor of *G. avonlea* and Cryptista was most frequently Obazoa (Figure 3.8). This unexpected relationship could be a result of BLAST-based artifacts due to database composition; however, observing this close relationship in such a large proportion of phylogenies suggests that there may be an alternative explanation. Another possibility is that there were similar, separate gene family expansions in both

Cryptista and Obazoa leading to a disproportionate number of genes per gene family in these lineages and artefactual phylogenetic attraction. This idea receives some support in the presence of large numbers of paralogs in the *G. avonlea* and *G. theta* genomes (Curtis 2012). Furthermore, in instances where the neighboring Obazoa consists of one or a few unicellular taxa it could be inferred that LGT has occurred (in one direction or the other). A large obazoan signal to in both *G. avonlea* and Cryptista was not investigated in detail here and further work should be done to determine its source and to assess the nature and prevalence of paralogs in *G. avonlea* and cryptophytes.

Many of the phylogenies that underwent automated topology detection resulted in an uncertain/ambiguous assignment of phylogenetic history to the *G. avonlea* genes and were ultimately deemed uninformative. Single gene trees are often poorly resolved due do the complexity of signals present in the data, making it difficult to glean any meaningful conclusions on the evolutionary history of a particular gene in any given species out of background phylogenetic noise (Archibald 2015). Many of the single gene trees contained multiple paralogs from *G. avonlea* and other Cryptista species, adding difficulties to automated and manual curation. Additionally, many predicted protein-coding genes showed lineage-specific proteins with limited taxa sampling resulting in an inability to confidently determine neighboring relationships. As the first step in constructing single gene trees relies on detecting homologs using BLAST, problematic sequences that are below e-value cutoffs can enter the dataset and disrupt phylogenetic signal; this includes paralogs and proteins that share a high degree of similarity to a single domain rather than across the entire sequence (Eisen 2000). LGT is also problematic here as it can disrupt monophyletic sub-trees (Eisen 2000). Due to the unresolved, error-prone nature of single gene trees, any tree of interest should be manually inspected.

Potential EGTs in *G. avonlea* of red algal ancestry that are unrelated to plastid function (i.e., functioning in the host biochemical processes) would likely be shared between *G. avonlea* and its photosynthetic relatives like *G. theta* (Stiller et al. 2009). Thus, genes showing an affinity of *G. avonlea* to Rhodophyta in the absence of any other cryptophyte are not strong candidates for endosymbiotic ancestry, while those showing a common *G. avonlea*-cryptophyte-Rhodophyta relationship are potential red algal EGTs.

Using the automated pattern detecting script followed by manual curation to filter out poorly resolved trees or those showing strong affiliations to non-photosynthetic lineages, 10 protein coding genes of potential common red-algal origin were detected (Table 3.3) that were not previously examined using the targeted approach discussed in Section 3.4.1. Seven of these contained a homolog in *G. theta* that was not identified in Curtis et al. (2012). Most of these were predicted to function in the host cytosol or contain a mitochondrial targeting signal and not exclusively suggestive of endosymbiotic ancestry. The fact that *G. avonlea* was more frequently found sister to Amoebozoa than to Rhodophyta, both when it branches with (76 Amoebozoa: 68 Rhodophyta) and without (248 Amoebozoa: 93 Rhodophyta) other Cryptista, suggests a minimal red-algal footprint in its genome (see Section 3.4.2 for a discussion on the use of Amoebozoa as a null hypothesis). Further investigation using the automated topology detection method should be performed to find new examples of EGT/EGR where *G. avonlea* clearly does not show a red-algal origin while cryptophytes do.

The few phylogenies investigated here showing a relationship of *G. avonlea* and Cryptista with cyanobacteria (with or without Rhodophyta present) or with *G. avonlea* sister to cyanobacteria directly do not show a topology supporting plastid ancestry. Cyanobacterial genes in *G. avonlea* should only be considered as evidence of a plastid if they show greater affinity to red algal genomes than to bacterial groups (Stiller et al. 2009), as there are alternate explanations for the presence of these genes. It is possible that the goniomonads obtained foreign algal genes by LGT instead of EGT due to their phagotrophic nature. This has been observed elsewhere, such as in the genome of choanoflagellates where it was reasoned that algal genes were a result of LGT and not EGT as a plastid ancestry in this lineage or its ancestors has never been inferred (Sun et al. 2010). Common genes appearing of cyanobacterial or algal ancestry in the cryptophytes and goniomonads may have alternative explanations such as a prey bias in their phagotrophic common ancestors (Stiller et al. 2009). Differentiating between LGT and EGT is difficult, as is distinguishing between EGT and vertical ancestry, particularly when the evolutionary history of a species is uncertain.

In contrast to a lack of red algal signal, there appears to be a strong affiliation of *G. avonlea* and Cryptista to Viridiplantae (either branching exclusively or inclusively with

other complex plastid-bearing taxa), consistent with the BLAST analysis and targeted EGT search discussed above. A red-green mosaicism has been frequently observed in taxa harboring a complex red-algal derived plastid (Archibald et al. 2003; Mousafa et al. 2009; Woehle et al. 2011; Burki et al. 2012a; Curtis et al 2012; Deschamps and Moreira 2012; Dorrell et al. 2017). The 'green' signal originally observed in the diatoms was found to be so substantial that it led to the suggestion of a cryptic plastid replacement of an original plastid of green algal ancestry with the complex red algal derived plastid found in extant lineages today (>70% of predicted EGTs showed green algal ancestry; Moustafa et al. 2009). In a re-analysis of this data by Deschamps and Moreira (2012) with an increased taxonomic sampling of red-algae, the assignment of many of the 'green' genes by Moustafa et al. (2009) were determined to be unresolved, of vertical descent or ambiguous algal ancestry suggesting that while there are genes of putative green-algal origin in the diatoms genome, they are not in such abundance as to suggest a cryptic green-algal plastid. On the other hand, a cryptic green algal signal observed in chromerids (Alveolata) was attributed to poor taxon sampling of red algae and artefacts associated with overly simplistic automated tree-sorting (Woehle et al. 2011).

Where do the 'green genes' found so prevalently in complex red-algal derived plastid bearing taxa come from? They could be a consequence of the phagotrophic lifestyles of the ancestors of these lineages and a result of repeated LGT (Archibald 2015). It is also possible that they stem from an ancient green algal endosymbiont in the ancestor of a lineage harboring a red-algal derived plastid today (Archibald 2015). In an attempt to explain the red-green mosaicism found in extant 'chromalveolates', Dorrell and Smith (2011) proposed that there was an ancient green algal derived plastid in the common ancestor of chromalveolates that was lost and followed by a secondary uptake of a red-algal plastid in haptophytes/cryptophytes that was spread via serial endosymbiosis. With the current phylogenetic placement of Cryptista (discussed further in the following sections), it is possible that many of these 'green genes' found within these 'chromalveolate' lineages (such as the diatoms) are a product of EGT in tertiary or higher endosymbiotic events involving a cryptophyte, whose nuclear genes have a close affinity with Archaeplastida, rather than a result of past green algal endosymbiotic ancestry.

3.4.4 A Cryptista-Archaeplastida relationship

As single gene trees are sensitive to the presence of LGTs and EGTs, paralogy, and reconstruction artefacts due to the limited phylogenetic signal they typically contain, they are generally not sufficient to infer deep eukaryotic relationships and multi-gene analysis should be used (Philippe et al. 2005; Leigh et al. 2008). Inferring phylogenies using datasets consisting of multiple genes helps to increase the phylogenetic signal and resolution of branches (Parfrey et al. 2010). As mentioned in Section 3.1, Cryptista (along with Haptista) has, in the past, been challenging to place in the eukaryotic tree of life (Burki et al. 2016a). Past analyses have placed Cryptista in a highly supported clade with Haptophyta either sister to SAR (e.g., Burki et al. 2009) or sister to Archaeplastida (e.g., Katz and Grant 2015), albeit with low support. A recent study by Burki et al. (2016a) using a large multigene dataset of 250 marker-genes and 150 taxa recovered a highly supported, alternative relationship: Haptista branching sister to SAR (highly supported, 98% UFboot) and Cryptista branching with Archaeplastida (maximally supported, 100% UFboot). Specifically, they found Cryptista as sister to a clade of Viridiplantae and Glaucophyta to the exclusion of Rhodophyta.

To investigate this relationship further, I inferred a phylogeny using a taxonomically reduced version of their dataset that included sequences from *G. avonlea* (see Supplementary Table A1 for taxa retained versus removed and Supplementary Table A2 for a list of marker genes used). Increasing taxon sampling is generally thought to be one of the key ways to alleviate systematic error in phylogenetic inference as it breaks up long branches (Hillis 1998; Wiens 2005) allowing for more highly resolved phylogenies to be obtained. Missing data among incompletely sequenced lineages, however, creates a vulnerability to systematic error and can cause poorly supported and/or incorrect branching due to a decrease in the ratio of phylogenetic to non-phylogenetic signal (Rodriguez-ezpeleta et al. 2007; Leigh et al. 2008; Parfrey et al. 2010; Philippe et al. 2017). At the same time, the computational power and time required to infer a phylogeny is directly influenced by the number of taxa and genes involved in the analysis and is amplified when using highly sophisticated models of evolution (Philippe et al. 2017). In order to decrease the amount of time and power required to infer multi-gene trees in this analysis, I reduced

the number of taxa represented while maintaining taxonomic diversity and minimizing the amount of missing data.

Other than the use of a reduced taxon sampling dataset covering the same taxonomic diversity and addition of a more complete goniomonad, a key difference between the Burki et al. (2016a) analysis and the analysis presented here is the assessment of branch support using standard bootstrap under the PMSF model (Wang et al. 2016). Due to the computational burden of assessing ML phylogenies under highly sophisticated models of evolution prior to development of the PMSF model by Wang et al. (2016), Burki et al. (2016a) assessed branch support using approximation methods such as UFboot (Minh et al. 2013) and SH-like approximate likelihood ratio tests (SH-aLRT; Guindon et al. 2010). While these are extremely fast methods of approximating branch support in large phylogenomic datasets, SH-aLRT derived support values can be difficult to interpret (Minh et al. 2013) and have a tendency to be inflated at moderate to severe model violations (Guindon et al. 2010; Minh et al. 2013). UFboot derived support values also tend to be inflated under severe model violations (Minh et al. 2013). Standard bootstrapping, on the other hand, has been shown to be conservative and underestimate the true probabilities of observing a particular branching order (Anisimova et al. 2011).

Phylogenies inferred here based on a modified Burki et al. (2016b) dataset recovered identical relationships to those inferred by Burki et al. (2016a) (Figures 3.10 to 3.14). While the position of Cryptista as sister to Viridiplantae and Glaucophyta had variable support in ML analyses in Burki et al. (2016a), the same relationship was observed here with 82% standard bootstrap support (Figure 3.10). The branching of Cryptista with a paraphyletic Archaeplastida was similarly maximally supported, while Haptista was recovered as sister to SAR with a slightly higher support value (99% standard bootstrap). To further investigate the branching position of Cryptista within Archaeplastida, additional phylogenies were inferred using modified datasets. When all Cryptista affiliated OTUs were removed, the monophyly of Archaeplastida was recovered (Figure 3.11) and the branching order of the rest of the phyla remained the same. When plastid-bearing Cryptista were removed from the dataset, non-photosynthetic, plastid-lacking Cryptista remained as sister to Viridiplantae and Glaucophyta with lower bootstrap support ((Figure 3.12) non-significant as determined using standard error of a bootstrap at a 95% confidence interval),

suggesting that the Archaeplastida-Cryptista relationship is not solely due to plastid ancestry (however, there is an uncertainty surrounding the endosymbiotic history of plastid-lacking taxa in Cryptista). Furthermore, this relationship does not appear to be caused by a few genes within Archaeplastida or Cryptista as removal of specific genes in OTUs identified as producing discordant topologies (and thus contributing to non-phylogenetic signal; Philippe et al. 2011; Philippe et al. 2017) significantly increased the support of this branch to 90% (as determined using standard error of a bootstrap at a 95% confidence interval) (Figure 3.14).

The seemingly stable position of Cryptista internal to Archaeplastida suggests that the position is not entirely caused by undetected EGTs in the dataset. If this were the case, one would expect Cryptista to be artificially attracted to Rhodophyta rather than Viridiplantae and Glaucophyta (as this was the source of their secondary plastid; Burki et al. 2016a) and for the support of the relationship when only non-photosynthetic Cryptista are included in the phylogeny to significantly decrease. However, it is possible that instances of EGT have gone undetected within Cryptista due to the close evolutionary relationship of their nuclear genes (either Viridiplantae and Glaucophyta specifically or Archaeplastida as a whole) and the source of their plastid (Rhodophyta) make it extremely difficult to disentangle the sources of genes in the nucleus. It should be noted that with a different marker-gene dataset based on Kang et. al (2017), which encompasses the Burki et al. (2016a) 250 marker gene set, an additional 100 marker genes and different OTU selection, Cryptista branches sister to Archaeplastida with maximum support and the monophyly or Archaeplastida is recovered with 97% standard bootstrap support (Figure 3.13). Further research needs to be done comparing the composition of these two datasets to determine what is causing the difference in branching position.

Attempts to search for alternative signals in the dataset by inferring phylogenies based on random subsets of marker genes from the modified Burki et al. (2016b) dataset showed a consistent relationship of Cryptista branching sister to at least one Archaeplastida lineage (95/100 iterations; Figure 3.15). While 30 of these iterations showed Cryptista as sister to Viridiplantae and Glaucophyta, 20 showed a sister relationship to a monophyletic Archaeplastida, further contributing to the uncertainty in Cryptista's exact position. Interestingly, 24/100 iterations showed Cryptista branching specifically with Glaucophyta

while only three showed an exclusive relationship with Viridiplantae, perhaps suggesting that Cryptista shares a closer ancestry to Glaucophyta. In support of this, a study by Burki et al. (2012) found that when cryptophytes were removed from their marker-gene dataset, the katablepharids (a non-photosynthetic Cryptista lineage) branched exclusively with Glaucophytes rather than sister to a monophyletic Archaeplastida. However, they also suggested that this relationship may be due to compositional heterogeneity or similar slow rates of evolution causing 'short branch exclusion'. Further work should be done in search of common genes causing the specific Cryptista-Glaucophyta affinity to see if compositional heterogeneity or slow evolutionary rates are suspected among them.

Whether the true placement of Cryptista is somewhere within or sister to Archaeplastida was not resolved by this study. Relationships that are observed consistently in phylogenies inferred from all or select genes under varying taxa are thought to be accurately reconstructed (Parfrey et al. 2010; Philippe et al. 2005). While exact placement is uncertain, a relationship with Archaeplastida was highly supported and consistent across all analyses and the alternative relationship of Cryptista with Haptista was never observed. In order to resolve this uncertainty, further analyses should be done to assess the monophyly of Archaeplastida including a more in-depth search for multiple signals emerging from the dataset as well as manual curation of individual marker-gene phylogenies with a larger taxon representation. Additionally, genes common in the randomly generated subsets that resulted in an affinity of Cryptista to Rhodophyta specifically (13/100 iterations) should be further explored, as they may indicate examples of EGT which are problematic in multi-gene phylogenies as their evolutionary history differs greatly from that of truly ancestral nuclear genes.

3.4.5 Implications on plastid evolution

As mentioned in Section 3.1, various hypotheses exist for the evolution of complex red algal plastids. The branching pattern observed in multi-gene phylogenies in this study and recent work by Burki et al. (2016a) strongly argues against the chromalveolate hypothesis (Cavalier-Smith 1999). Under the host phylogeny observed here where Cryptista and

Archaeplastida are in a highly supported sister relationship, the singularity of chromalveolate plastids would require the secondary plastid of red algal origin to be acquired prior to establishment and divergence of the primary plastids from which it evolved. Thus, the more probable scenario is a unique secondary endosymbiosis of a red alga followed by horizontal spread of the plastid via one or more of higher order endosymbiotic events (e.g., Sanchez-Puerta and Delwiche 2008; Stiller et al. 2014; Petersen et al. 2014; Burki et al. 2016a; Bodyl 2017). The precise order of horizontal spread, and the lineage responsible for the secondary event, remains unclear. If the cryptophyte-first model of serial endosymbiotic plastid evolution proposed by Stiller et al. (2014) turns out to be the true evolutionary scenario, the lack of evidence for a red-algal footprint in katablepharids and palpitomonas (Burki et al. 2012b, Yabuki et al. 2010) suggests that secondary plastid acquisition occurred somewhere after their divergence from the common ancestor of goniomonads and cryptophytes. An overall lack of significant signal to red algae in *G. avonlea* suggests that the red-alga derived secondary plastid was acquired after the divergence of the goniomonads from the cryptophytes. The specific relationship between Cryptista and Archaeplastida may provide an explanation for the abundance of green-algal genes in photosynthetic lineages throughout SAR (Archibald et al. 2003; Mousafa et al. 2009; Woehle et al. 2011; Burki et al. 2012a; Curtis et al 2012; Deschamps and Moreira 2012; Dorrell et al. 2017) if a cryptophyte was the source of their complex plastids as suggested by Stiller et al. (2014).

Whether Cryptista is sister to a monophyletic Archaeplastida or branching internally sister to Viridiplantae and Glaucophyta is uncertain. However, if Cryptista is branching in such a way as to break the monophyly of Archaeplastida there are further implications for the evolution of primary plastids. One scenario consistent with this topology is the presence of an ancestral primary plastid in Cryptista that was secondarily lost some point after diverging from Viridiplantae and Glaucophyta (followed by the cryptophytes re-acquiring a plastid via secondary endosymbiosis with a red alga). Another scenario questions the singularity of primary plastids, which, while generally accepted, has been questioned by some (e.g., Burki 2017; Kim and Maruyama 2014; Stiller 2014; Howe et al. 2008; Stiller and Hall 1997) and finds plausibility in the separately originating primary photosynthetic organelle of *P. chromatophora (*Nowack, Melkonian and Glöckner

2008). Stiller and Hall (1997) argued that it is possible the primary plastids of Archaeplastida lineages originated from independent endosymbioses of related cyanobacteria that may have been preferential endosymbionts. There is also the possibility that the primary plastids of Viridiplantae and Rhodophyta are not really primary and actually stem from a secondary endosymbiosis with a glaucophyte (who consequently contain the only true primary plastid; Stiller and Hall 1997).

While it is tempting to speculate on alternative evolutionary scenarios, the phylogenetic position of Cryptista is still uncertain. This is evidently seen in the highly supported, yet differing positions of Cryptista obtained under the different marker-gene datasets used in this study. As mentioned above, additional research should be done to assess the phylogenetic position of Cryptista and monophyletic nature of Archaeplastida and, consequently, the origin of primary plastids. It is possible that primary plastid evolution is different from what is generally accepted today and that complex plastid evolution is more convoluted or significantly simpler than current models propose (Archibald 2015). There are many underrepresented lineages in regard to genome sequence availability across protists (Sibbald and Archibald 2017) including those that are key to studying plastid evolution such as Glaucophyta, who only have a single genome sequenced (Price et al. 2012), and Rhodophyta. All it may take is the discovery and/or sequencing of one species to change the way we view organelle evolution and the evolutionary relationships of eukaryotes as a whole.

CHAPTER 4 CONCLUSION

Endosymbiosis involves a close interaction between two cells where one lives inside the other. The intimate nature of this relationship fosters co-evolutionary processes and involves integration of the endosymbiont and the host at both metabolic and genetic levels (via EGT and EGR), leading to a highly reduced endosymbiont and, in the case of plastids and mitochondria, the evolution of organelles. Here I presented phylogenetic analyses of two very different systems where endosymbiosis plays a significant role. The first involved an obligate endosymbiosis involving an amoebozoan host (*Paramoeba* spp.) and kinetoplastid endosymbiont (*Perkinsela* sp.). Through molecular characterization of novel isolates of *Paramoeba* spp. and their associated endosymbiont, I found strong evidence for coevolution occurring. While the exact nature of their relationship is uncertain, this system may prove useful in studying the transition from endosymbiont to organelle. The second system involves a close heterotrophic relative to the secondarily photosynthetic cryptophytes, *G. avonlea*. Taking advantage of the fact that genes are transferred from an endosymbiont to the host nucleus during plastid integration, I probed genomic data from *G. avonlea* for a red algal endosymbiotic footprint. Overall, phylogenetic analyses revealed few genes indicative of endosymbiotic ancestry in *G. avonlea*, suggesting that the goniomonads, along with Cryptista as a whole, were ancestrally non-photosynthetic. It should be noted that while evidence of red algal ancestry would strongly support a plastid-bearing past in *G. avonlea*, an absence of a significant signal does not guarantee a plastid was never there. While analysis of all single gene trees did reveal a strong relationship of *G. avonlea* and green/glaucophyte algae, these genes could not be conclusively assigned as EGTs due to the significant sister relationship of Cryptista and Archaeplastida in phylogenomic analyses with varying marker-gene datasets. There is difficulty in assigning genes as putative EGTs in Cryptista, particularly with their relationship to Archaeplastida in phylogenomic analyses, and it remains to be determined if these 'green genes' are of endosymbiotic origin.

APPENDIX A – SUPPLEMENTARY TABLES FOR CHAPTER 3

Supplementary Table A1. Operational taxonomic units (OTUs) retained versus removed from a 250 gene, 150 OTU marker gene dataset (Burki et al. 2016) and their corresponding percent gene coverage. The number of OTUs was systematically reduced from 150 to 98 to reduce the complexity of phylogenetic analyses while maintaining taxonomic diversity. Transcriptomic data was added from *Goniomonas. avonlea* (highlighted in bold) to increase phylogenetic signal from a goniomonad (91.60% gene coverage compared to the previously used *Goniomonas* sp. at 77.60%).

Super-group	Taxa	Retained	Percent Gene Coverage
Alveolata	*Amphidinium carterae*	N	90.40%
	Babesia bovis	N	82.40%
	Chromera velia	Y	87.60%
	Colpodella	N	74.40%
	Cryptosporidium muris	Y	83.60%
	Euplotes	N	82.40%
	Hematodinium sp.	N	68.40%
	Karenia brevis	Y	94.80%
	Litonotus pictus	N	58.00%
	Noctiluca scintillans	N	84.40%
	Paramecium tetraurelia	N	83.20%
	Perkinsus marinus	Y	84.00%
	Plasmodium falciparum	Y	83.20%
	Platyophrya macrostoma	Y	85.20%
	Protocruzia adherens	Y	70.00%
	Sterkiella histriomuscorum	Y	82.00%
	Strombidium inclinatum	N	71.60%
	Tetrahymena thermophila	Y	90.80%
	Toxoplasma gondii	Y	84.40%
	Vitrella brassicaformis	Y	69.20%
	Voromonas pontica	Y	59.20%
Stramenopiles	*Aplanochytrium*	Y	93.20%
	Asterionellopsis glacialis	Y	93.60%
	Aurantiochytrium limacinum	N	88.40%
	Aureococcus anophageferrens	N	88.40%
	Aureoumbra lagunensis	Y	73.20%
	Blastocystis hominis	N	83.60%
	Bolidomonas pacifica	N	85.20%
	Cafeteria sp.	Y	88.40%
	Chattonella subsalsa	Y	90.00%
	Dictyocha speculum	N	91.20%
	Ectocarpus siliculosus	Y	95.60%

Super-group	Taxa	Retained	Percent Gene Coverage
	Fibrocapsa japonica	Y	76.40%
	Heterosigma akashiwo	N	81.20%
	Mallomonas sp.	Y	81.20%
Stramenopiles	MAST4	Y	43.20%
	Nannochloropsis gaditana	Y	75.60%
	Ochromonas sp. (MMETSP1177)	Y	70.80%
	Odontella aurita	N	92.00%
	Paraphysomonas imperforata	N	85.20%
	Pelagomonas calceolata	Y	93.20%
	Phaeodactylum tricornutum	N	90.00%
	Phaeomonas parva	Y	72.80%
	Phytophthora	Y	98.80%
	Pinguiococcus pyrenoidosus	N	60.00%
	Pseudopedinella elastica	N	92.80%
	Rhizochromulina marina	Y	87.20%
	Saprolegnia parasitica	Y	92.40%
	Schizochytrium aggregatum	Y	88.40%
	Spumella elongata	N	93.20%
	Thalassiosira pseudonana	Y	91.20%
	Thraustochytrium sp.	N	90.40%
	Vaucheria litorea	N	70.40%
Rhizaria	Astrolonche serrata	Y	47.60%
	Bigelowiella natans	Y	89.20%
	Chlorarachnion reptans	Y	86.00%
	Elphidium margaritaceum	Y	86.80%
	Gromia sphaerica	N	51.60%
	Reticulomyxa filosa	N	92.40%
Haptista	*Acanthocystis* sp.	Y	91.20%
	Calcidiscus leptoporus	Y	85.20%
	Choanocystis sp.	Y	92.00%
	Chrysochromulina brevifilum	Y	86.00%
	Chrysochromulina polylepis	Y	92.00%
	Chrysoculter rhomboideus	Y	74.80%
	Emiliania huxleyi	Y	96.00%
	Isochrysis galbana	N	90.40%
	Pavlovales sp.	Y	89.60%
	Phaeocystis antarctica	Y	93.20%
	Phaeocystis sp.	N	75.60%
	Pleurochrysis carterae	Y	94.00%
	Prymnesium parvum	Y	93.20%
	Raineriophrys erinaceoides	Y	93.60%
	Raphidiophrys ambigua	N	44.40%
	Raphidiophrys heterophryoidea	Y	94.40%

Super-group	Taxa	Retained	Percent Gene Coverage
	Scyphosphaera apsteinii	Y	86.00%
Cryptista	*Cryptomonas curvata*	Y	82.00%
	Cryptophyceae sp.	Y	93.60%
Cryptista	***Goniomonas avonlea***	**NEW**	**91.60%**
	Goniomonas pacifica	N	54.40%
	Goniomonas sp.	N	77.60%
	Guillardia theta	Y	96.00%
	Hemiselmis andersenii	Y	91.60%
	Hemiselmis rufescens	N	90.80%
	Palpitomonas bilix	Y	78.40%
	Rhodomonas abbreviata	Y	84.00%
	Rhodomonas sp.	N	76.80%
	Roombia truncata	Y	68.80%
Viridiplantae	*Arabidopsis*	N	95.20%
	Brachypodium distachyon	N	94.00%
	Chlamydomonas reinhardtii	Y	90.40%
	Chlorella vulgaris	Y	84.00%
	Coccomyxa sp.	Y	93.20%
	Micromonas sp.	Y	93.20%
	Mimulus guttatus	N	92.80%
	Oryza sativa	Y	95.60%
	Ostreococcus lucimarinus	Y	82.40%
	Physcomitrella patens	Y	94.80%
	Populus trichocarpa	Y	94.40%
	Selaginella moellendorffii	Y	93.60%
	Volvox carteri	N	91.60%
Glaucophyta	*Cyanophora paradoxa*	Y	86.00%
	Cyanoptyche gloeocystis	Y	67.60%
	Gloeochaete witrockiana	Y	82.80%
Rhodophyta	*Chondrus crispus*	Y	83.60%
	Compsopogon coeruleus	N	59.60%
	Cyanidioschyzon merolae	Y	81.20%
	Erythrolobus	Y	72.80%
	Galdieria sulphuraria	Y	86.80%
	Madagascaria erythrocladoides	Y	62.40%
	Porphyra	Y	84.80%
	Porphyridium aerugineum	N	62.40%
	Porphyridium cruentum	Y	83.60%
	Rhodella maculata	N	61.20%
	Rhodosorus marinus	Y	72.80%
Excavata	*Bodo saltans*	Y	84.80%
	Eutreptiella gymnastica	Y	86.00%

Super-group	Taxa	Retained	Percent Gene Coverage
Excavata	*Jakoba*	Y	40.40%
	Malawimonas	Y	53.60%
	Naegleria gruberi	Y	82.00%
	Neobodo designis	N	88.80%
	Percolomonas cosmopolitus (MMETSP0758)	N	68.00%
	Percolomonas cosmopolitus (MMETSP0759)	Y	80.40%
	Reclinomonas americana	N	52.40%
	Sawyeria marylandensis	N	40.40%
	Seculamonas ecuadoriensis	Y	36.80%
	Tsukubamonas globosa	Y	68.40%
Obazoa	*Amastigomonas* sp.	Y	85.60%
	Batrachochytrium dendrobatidis	N	96.40%
	Branchiostoma floridae	N	95.20%
	Breviate	Y	91.20%
	Cryptococcus neoformans	Y	90.80%
	Danio rerio	N	94.40%
	Daphnia pulex	N	91.60%
	Homo sapiens	Y	99.60%
	Lottia gigantea	Y	92.00%
	Monosiga brevicollis	Y	90.00%
	Nematostella vectensis	N	96.00%
	Neurospora crassa	N	89.20%
	Phycomyces blakesleeanus	Y	92.80%
	Schizosaccharomyces pompe	Y	86.40%
	Thecamonas trahens	N	81.20%
	Ustilago maydis	N	86.00%
Amoebozoa	*Acanthamoeba castellanii*	Y	60.80%
	Dictyostelium discoideum	Y	91.60%
	Dictyostelium purpureum	N	88.80%
	Polysphondylium pallidum	Y	92.40%
Orphan Lineages	*Collodictyon* sp.	Y	38.40%
	Picobiliphyte MS584 11	N	19.20%
	Telonema	N	52.00%

Supplementary Table A2 – Gene abbreviations and corresponding full gene names for the 250 marker genes used in the Burki et al. (2016) dataset.

Abbreviation	Full Gene Name
abce1	ATP-binding cassette sub-family E member 1
abt1	Activator of basal transcription 1
agx	UDP-N-acteylglucosamine pyrophosphorylase 1
alg11	asparagine-linked glycosylation protein 11
ap1m1	AP-1 complex subunit mu-1
ap1s2	adaptor-related protein complex 1, sigma 2 subunit
ap2m1	AP-2 complex subunit mu-1
ap3m1	AP-3 complex subunit mu-1
ap3s1	adaptor-related protein complex 3, sigma 1 subunit
ap4s1	adaptor-related protein complex 4, sigma 1 subunit
arp2	actin-related protein 2
arpc3	Actin-related protein 2/3 complex subunit 3
arpc4	Actin-related protein 2/3 complex subunit 4
asf1a	Histone chaperone ASF1A
atad1	ATPase family AAA domain-containing protein 1
atp6v1a	ATPase, H+ transporting, lysosomal V0 subunit a1
atp6v1b	V-type proton ATPase catalytic subunit A
atp6v1c	V-type proton ATPase subunit B
atp6v1d	V-type proton ATPase subunit D
atp6v1e	V-type proton ATPase subunit E
bat1	Spliceosome RNA helicase BAT1
bms1	Ribosome biogenesis protein BMS1 homolog
brf1	Transcription factor IIIB 90 kDa subunit
bysl	Bystin
C16orf80	chromosome 16 open reading frame 80
calm	Calmodulin
capza1	F-actin-capping protein subunit alpha
capzb	F-actin-capping protein subunit beta
ccdc65	coiled-coil domain containing 65
fntb	Protein farnesyltransferase subunit beta
clgn	Calmegin
cop-beta	coatomer protein complex, subunit beta 2
cope	coatomer protein complex, subunit epsilon
copg2	Coatomer subunit gamma-2
cops6	COP9 constitutive photomorphogenic homolog subunit 6
coq4	coenzyme Q4 homolog
coro1c	coronin, actin binding protein, 1C
crfg	Nucleolar GTP-binding protein 1
dcaf13	DDB1 and CUL4 associated factor 13
dimt1l	DIM1 dimethyladenosine transferase 1-like
dkc1	Dyskerin
dnai2	dynein, axonemal, intermediate chain 2

Abbreviation	Full Gene Name
dnal1	dynein, axonemal, light chain 1
dpagt1	UDP-N-acetylglucosamine--dolichyl-phosphate N-acetylglucosaminephosphotransferase
dph1	Diphthamide biosynthesis protein 1
drg2	developmentally regulated GTP binding protein 2
eftud1	Elongation factor Tu GTP binding domain-containing protein 1
eftud2	Elongation factor Tu GTP-binding domain-containing protein 2
eif1a	Eukaryotic translation initiation factor 1A
eif1b	Eukaryotic translation initiation factor 1B
eif2a	Eukaryotic translation initiation factor 2 subunit 1
eif2b	Eukaryotic translation initiation factor 2 subunit 2
eif2g	Eukaryotic translation initiation factor 2 subunit 3
eif3i	eukaryotic translation initiation factor 3 subunit I
eif5A	Eukaryotic translation initiation factor 5A-1
eif5b	Eukaryotic translation initiation factor 5B
eif6	Eukaryotic translation initiation factor 6
emg1	Probable ribosome biogenesis protein NEP1
gspt2	Eukaryotic peptide chain release factor GTP-binding subunit
etf1	Eukaryotic peptide chain release factor subunit 1
fam96b	family with sequence similarity 96, member B
fbl	rRNA 2'-O-methyltransferase fibrillarin
ftsj1	Putative tRNA
gas8	growth arrest-specific 8
gdi2	Rab GDP dissociation inhibitor beta
gnb2L1	Guanine nucleotide-binding protein subunit beta-2-like 1
gnb3	Transducin beta chain 3
gnl2	Nucleolar GTP-binding protein 2
gpn1	GPN-loop GTPase 1
gpn2	GPN-loop GTPase 2
gpn3	GPN-loop GTPase 3
grwd1	glutamate-rich WD repeat containing 1
hsp90	Heat shock protein HSP 90
hsp75	Heat shock protein 75 kDa, mitochondrial
hyou1	hypoxia up-regulated 1
ift46	intraflagellar transport 46 homolog
ift57	intraflagellar transport 57 homolog
ift88	intraflagellar transport 88 homolog
imp4	IMP4, U3 small nucleolar ribonucleoprotein
ino1	Inositol-3-phosphate synthase 1
kars	lysyl-tRNA synthetase
kpnb1	Importin subunit beta-1
krr1	KRR1 small subunit processome component homolog
lsm4	U6 snRNA-associated Sm-like protein LSm4
mak16	Protein MAK16 homolog
mat1a	methionine adenosyltransferase I, alpha

Abbreviation	Full Gene Name
mcm2	DNA replication licensing factor MCM2
mcm3	DNA replication licensing factor MCM3
mcm4	DNA replication licensing factor MCM4
mcm5	DNA replication licensing factor MCM5
mcm6	DNA replication licensing factor MCM6
mcm7	DNA replication licensing factor MCM7
mcm9	DNA replication licensing factor MCM9
metap2	Methionine aminopeptidase 2
mettl1	tRNA (guanine-N(7)-)-methyltransferase isoform a
naa15	N(alpha)-acetyltransferase 15, NatA auxiliary subunit
nae1	NEDD8 activating enzyme E1 subunit 1
nat10	N-acetyltransferase 10
ncbp2	Nuclear cap-binding protein subunit 2
ndufv1	NADH dehydrogenase (ubiquinone) flavoprotein 1
ndufv2	NADH dehydrogenase (ubiquinone) flavoprotein 2, mitochondrial
nhp2	H/ACA ribonucleoprotein complex subunit 2
nhp2L1	NHP2-like protein 1
nip7	60S ribosome subunit biogenesis protein NIP7 homolog
nmt2	Glycylpeptide N-tetradecanoyltransferase 2
nop2	Probable 28S rRNA (cytosine(4447)-C(5))-methyltransferase
nop56	Nucleolar protein 56
nop58	Nucleolar protein 58
nsa2	NSA2 ribosome biogenesis homolog
nsf	Vesicle-fusing ATPase
oplah	5-oxoprolinase
osgep	Probable O-sialoglycoprotein endopeptidase
pcna	Proliferating cell nuclear antigen
pls3	Plastin-3
pno1	RNA-binding protein PNO1
polr1a	DNA-directed RNA polymerase I subunit RPA1
polr1b	DNA-directed RNA polymerase I subunit RPA2
polr1c	DNA-directed RNA polymerases I and III subunit RPAC1
polr1d	DNA-directed RNA polymerases I and III subunit RPAC2
polr2a	DNA-directed RNA polymerase II subunit RPB1
polr2b	DNA-directed RNA polymerase III subunit RPC2
polr2f	DNA-directed RNA polymerases I, II, and III subunit RPABC2
polr2h	DNA-directed RNA polymerases I, II, and III subunit RPABC3
polr3b	DNA-directed RNA polymerase II subunit RPB2
ppp2r3	protein phosphatase 2, regulatory subunit B, alpha
prpf8	Pre-mRNA-processing-splicing factor 8
psma1	Proteasome subunit alpha type-1
psma2	Proteasome subunit alpha type-2
psma3	Proteasome subunit alpha type-3
psma4	Proteasome subunit alpha type-4

Abbreviation	Full Gene Name
psma5	Proteasome subunit alpha type-5
psma6	Proteasome subunit alpha type-6
psma7	Proteasome subunit alpha type-7
psmb1	Proteasome subunit beta type-1
psmb2	Proteasome subunit beta type-2
psmb3	Proteasome subunit beta type-3
psmb4	Proteasome subunit beta type-4
psmb5	Proteasome subunit beta type-5
psmb6	Proteasome subunit beta type-6
psmb7	Proteasome subunit beta type-7
psmc1	26S protease regulatory subunit 4
psmc2	26S protease regulatory subunit 7
psmc3	26S protease regulatory subunit 6A
psmc4	26S protease regulatory subunit 6B
psmc5	26S protease regulatory subunit 8
psmc6	26S protease regulatory subunit S10B
psmd1	26S proteasome non-ATPase regulatory subunit 1
psmd12	Proteasome 26S subunit, non-ATPase, 12
psmd14	26S proteasome non-ATPase regulatory subunit 14
rad51	DNA repair protein RAD51 homolog 1
ran	GTP-binding nuclear protein Ran
rbm19	Probable RNA-binding protein 19
rcl1	RNA 3'-terminal phosphate cyclase-like protein
rfc2	Replication factor C subunit 2
rfc4	Replication factor C subunit 4
rfc5	Replication factor C subunit 5
rpf1	ribosome production factor 1
rpl10	60S ribosomal protein L10
rpl10a	60S ribosomal protein L10a
rpl11	60S ribosomal protein L11
rpl12	60S ribosomal protein L12
rpl13	60S ribosomal protein L13
rpl13a	60S ribosomal protein L13a
rpl14	60S ribosomal protein L14
rpl15	60S ribosomal protein L15
rpl17	60S ribosomal protein L17
rpl18	60S ribosomal protein L18
rpl18a	60S ribosomal protein L18a
rpl19	60S ribosomal protein L19
rpl21	60S ribosomal protein L21
rpl23	60S ribosomal protein L23a
rpl24	60S ribosomal protein L24
rpl26	60S ribosomal protein L26
rpl27a	60S ribosomal protein L27a
rpl3	60S ribosomal protein L3

Abbreviation	Full Gene Name
rpl30	60S ribosomal protein L30
rpl31	60S ribosomal protein L31
rpl32	60S ribosomal protein L32
rpl34	60S ribosomal protein L34
rpl35	60S ribosomal protein L35
rpl35a	60S ribosomal protein L35a
rpl36a	60S ribosomal protein L36a
rpl37a	60S ribosomal protein L37a
rpl4	60S ribosomal protein L4
rpl5	60S ribosomal protein L5
rpl6	60S ribosomal protein L6
rpl7	60S ribosomal protein L7
rpl7a	60S ribosomal protein L7a
rpl8	60S ribosomal protein L8
rpl9	60S ribosomal protein L9
rplp0	60S acidic ribosomal protein P0
rps10	40S ribosomal protein S10
rps11	40S ribosomal protein S11
rps12	40S ribosomal protein S12
rps13	40S ribosomal protein S13
rps14	40S ribosomal protein S14
rps15	40S ribosomal protein S15
rps15a	40S ribosomal protein S15a
rps16	40S ribosomal protein S16
rps17	40S ribosomal protein S17
rps18	40S ribosomal protein S18
rps19	40S ribosomal protein S19
rps2	40S ribosomal protein S2
rps20	40S ribosomal protein S20
rps23	40S ribosomal protein S23
rps24	40S ribosomal protein S24
rps25	40S ribosomal protein S25
rps26	40S ribosomal protein S26
rps27	40S ribosomal protein S27
rps3	40S ribosomal protein S3
rps3a	40S ribosomal protein S3a
rps4y1	40S ribosomal protein S4
rps5	40S ribosomal protein S5
rps6	40S ribosomal protein S6
rps8	40S ribosomal protein S8
rps9	40S ribosomal protein S9
rpsaT	40S ribosomal protein SA
ruvbl1	RuvB-like 1
sars	Seryl-tRNA synthetase, cytoplasmic
sbds	Ribosome maturation protein SBDS

Abbreviation	Full Gene Name
sco1	Protein SCO1 homolog, mitochondrial
sec61	protein transport protein Sec61 subunit alpha isoform 2 isoform a
snd1	staphylococcal nuclease and tudor domain containing 1
srp54	Signal recognition particle 54 kDa protein
srpr	Signal recognition particle receptor subunit alpha
stxbp1	syntaxin binding protein 1
suclg1	Succinyl-CoA ligase (GDP-forming) subunit alpha
tbp	TATA-box-binding protein
tcp1-alpha	T-complex protein 1 subunit alpha
tcp1-beta	T-complex protein 1 subunit beta
tcp1-delta	T-complex protein 1 subunit delta
tcp1-epsilon	T-complex protein 1 subunit epsilon
tcp1-eta	T-complex protein 1 subunit eta
tcp1-gamma	T-complex protein 1 subunit gamma
tcp1-theta	T-complex protein 1 subunit theta
tcp1-zeta	T-complex protein 1 subunit zeta
tm9sf1	transmembrane 9 superfamily member 1 isoform a
tubb	Tubulin beta
tubg	Tubulin gamma
uba3	ubiquitin-like modifier activating enzyme 3
vbp1	von Hippel-Lindau binding protein 1
vpc	Transitional endoplasmic reticulum ATPase
vps18	vacuolar protein sorting 18
vps26b	vacuolar protein sorting 26
vps4	Vacuolar protein sorting-associated protein 4A
wbscr22	Williams Beuren syndrome chromosome region 22
xpb	TFIIH basal transcription factor complex helicase XPB subunit
xpo1	Exportin-1
ykt6	YKT6 v-SNARE homolog

Supplementary Table A3. Individual genes in specific species that were determined to be outliers based on analysis using PhyloMCOA (De Vienne et al. 2012). These discordant genes were removed from their corresponding OTU in the Burki et al. (2016) 250 marker gene dataset prior to generating the phylogeny shown in Figure 3.14. Full gene names can be found in Supplementary Table A2.

OTU	Gene
Amastigomonas sp.	POLR2B
Aplanochytrium	ABT1
Asterionellopsis glacialis	atp6v1e
	osgep
	rpl11
	rpsa
	AP1S2
Astrolonche serrata	capzb
	EFTUD1
	tubb
	VPS18
Aureoumbra lagunensis	psma2
	rpl21
	rps12
	xpb
	erf3b
	rpl3
Bodo saltans	STXBP1
	arpc4
	eif2g
Cafeteria roebergensis	PRPF8
	arpc4
	ASF1
	CCDC65
Cafeteria sp.	rps13
	BRF1
	KARS
Calcidiscus leptoporus	rps5
	mcm7
	POLR2F
	ran
	FTSJ1
	IFT57

OTU	Gene
Chattonella subsalsa	mito
	POLR2H
	rps20
Chlamydomonas reinhardtii	TM9SF1
	COPS6
	rps11
	EIF1B
Chlorella vulgaris	rpl15
	psmb5
	rpl10
Chondrus crispus	NAE1
	rpl10
Chromera velia	rpl30
	RPL34
	rps15
	BMS1
	mcm4
	NAT10
Chrysochromulina brevifilum	polr1a
	EFTUD1
	VPS18
	ino1
Chrysochromulina polylepis	POLR2B
	POLR1D
	rps14
	emg1
Chrysoculter rhomboideus	GRWD1
	POLR2H
	psma7
	psmc3
Choanocystis sp.	RPS25
Coccomyxa sp.	emg1
	KARS
	psmc4
	RFC5
Cryptococcus neoformans	RUVBL1
	rpl37a
	ALG11
	ap3m1

OTU	Gene
Cryptococcus neoformans	COPE
Cryptomonas curvata	COPS6
	EFTUD1
Cryptophyceae sp.	EIF1B
	KARS
	NMT
	POLR1D
Cryptosporidium muris	POLR2F
	RFC2
Cyanidioschyzon merolae	RFC4
	rpl10
	rps20
	tcp1-epsilon
	atp6v1d
	NAE1
Cyanophora paradoxa	rps5
	KARS
	ndufv1
	psmc5
	rpl19
Cyanoptyche gloeocystis	atad1
Ectocarpus siliculosus	FTSJ1
Emiliania huxleyi	ino1
	mcm2
Erythrolobus	mcm4
Eutreptiella gymnastica	mcm6
	psmb3
	psmc6
	RFC5
Fibrocapsa japonica	rpl27a
Galdieria sulphuraria	gnb2L1
	gpn2
	mito
	rpl13a
	rpl30
	rpl32
	rps3
	VPS18

OTU	Gene
Gloeochaete witrockiana	BMS1
Goniomonas avonlea	BMS1
	COP-beta
	mito
	pno1
	polr1a
	IFT46
	IFT88
	PLS3
	POLR2H
	rpl27a
Hemiselmis andersenii	rpl6
Homo sapiens	DNAI2
Jakoba	nsf
Karenia brevis	rps4y1
	rps8
	BMS1
	FAM96B
	mcm9
	ndufv1
	mito
	psma5
	psmb4
	rpl26
	ap1m1
	HYOU1
	oplah
	psma2
Lottia gigantea	ap3m1
	EFTUD1
	eif6
Madagascaria erythrocladoides	psmb7
	ABT1
	hsp90
Malawimonas	ABT1
Mallomonas sp.	rpl15
MAST4	rps4y1
	tcp1-gamma

OTU	Gene
MAST4	nsf
Micromonas sp.	sars
	tcp1-gamma
Monosiga brevicollis	IFT88
	psmb3
	PPP2R3
	ran
	rpl3
Naegleria gruberi	rps13
Nannochloropsis gaditana	rps3
Ochromonas sp. (MMETSP1177)	rps5
Oryza sativa	PSMD12
	RPL34
	ALG11
Ostreococcus lucimarinus	DPH1
	FTSJ1
	NAE1
	ran
Palpitomonas bilix	rpl11
Pavlovales sp.	ASF1
	eif6
	gpn3
	GRWD1
	rpl18
	VPS18
Pelagomonas calceolata	EIF3I
	nsf
	atad1
	NAE1
	rpl11
	psma7
Percolomonas cosmopolitus (MMETSP0759)	psmc6
Perkinsus marinus	eif6
	ap2m1
	FTSJ1
	gnb2L1
	ALG11
	AP4S1

OTU	Gene
Perkinsus marinus	C16orf80
	psmc3
Phaeomonas parva	XPO1
	ap1m1
Phycomyces blakesleeanus	BMS1
	eftud2
	FAM96B
	gdi2
Plasmodium falciparum	sbds
	IFT46
	dpagt1
	eif2b
	oplah
	POLR2H
	rpl11
	RPL34
	sbds
	tubg
	DRG2
	nhp2
	RFC4
	tubb
Platyophrya macrostoma	capzb
	COPG2
Pleurochrysis carterae	DNAL1
	ino1
	mcm9
	NOP2
Polysphondylium pallidum	psmb1
Porphyra	psmb6
	RCL1
	rpl24
	rpl31
	RPL34
Porphyridium cruentum	rps23
	rps4y1
	IMP4
	DIMT1L
	mito

OTU	Gene
Prymnesium parvum	NCBP2
	rpl14
	rpl26
	atp6v1c
Raphidiophrys heterophryoidea	MAK16
	POLR1D
Rhizochromulina marina	rps20
Rhodomonas abbreviata	SND1
Rhodosorus marinus	mcm7
	RPS19
	eftud2
	gdi2
	NSA2
	ASF1
Saprolegnia parasitica	BRF1
Schizochytrium aggregatum	COPE
	PCNA
Schizosaccharomyces pombe	rpl4
	rps15a
	xpb
	DNAL1
	NMT
Scyphosphaera apsteinii	etf1
	METTL1
	psmb5
	rpl12
	rpl31
Seculamonas ecuadoriensis	capza1
	rpl7a
	tcp1-beta
Selaginella moellendorffii	IFT46
	psmc1
	rpl27a
Sterkiella histriomuscorum	BYSL
	rpl9
	rps23

OTU	Gene
Tetrahymena thermophila	rps4y1
Thalassiosira pseudonana	C16orf80
	DNAI2
	LSM4
	rpl27a
Toxoplasma gondii	ABT1
	AP1S2
	C16orf80
	gpn3
	rpl13a
	rps12
Vitrella brassicaformis	psma2
	rps9
	tubb
Voromonas pontica	NAA15

Supplementary Table A4. Operational taxonomic units (OTUs) retained from a 351 gene, 383 OTU marker gene dataset (Kang et al. 2017) and their corresponding percent gene coverage. The number of OTUs was systematically reduced from 383 to 101 to reduce the complexity of phylogenetic analyses while maintaining taxonomic diversity. Transcriptomic data was added from *Goniomonas avonlea* (highlighted in bold).

Super-group	Taxa	Percent Gene Coverage
Alveolata	*Alexandrium minutum*	41.03%
	Cryptosporidium parvum	72.93%
	Karenia brevis	81.77%
	Lankesteria abbotti	72.36%
	Oxyrrhis marina	82.34%
	Paramecium tetraurelia	90.31%
	Perkinsus marinus	84.05%
	Plasmodium falciparum	65.81%
	Symbiodinium mf105	65.53%
	Tetrahymena thermophila	87.46%
	Toxoplasma gondii	60.97%
Stramenopiles	*Aurantiochytrium limacinum*	94.02%
	Aureococcus anophagefferens	84.05%
	Chattonella subsalsa	84.90%
	Chrysocystis fragilis	73.79%
	Chyrsophyceae sp.	88.03%
	Dictyocha speculum	87.18%
	Dinobryon sp.	80.63%
	Ectocarpus siliculosus	94.02%
	Halocafeteria seosinensis	94.59%
	Hyphochytrium catenoides	92.59%
	MAST13	86.89%
	Nannochloropsis gaditana	83.48%
	Paraphysomonas bandaiensis	90.03%
	Phaeomonas parva	74.64%
	Phytophthora parasitica	95.16%
	Proteromonas sp.	82.62%
	Pseudopedinella elastica	87.18%
	Saprolegnia declina	95.16%
	Schizochytrium aggregatum	92.02%
	Thalassiosira pseudonana	86.89%

Super-group	Taxa	Percent Gene Coverage
	Wobblia lunata	96.58%
Rhizaria	*Bigelowiella natans*	91.17%
	Chlorarachnion reptans	87.18%
	Gromia sphaerica	50.43%
	Guttulinopsis sp.	82.05%
	Lotharella amoebiformis	86.04%
	Paulinella chromatophora	72.36%
	Rosculus sp.	88.60%
Haptista	*Chrysochromulina rotalis*	84.33%
	Chrysochromulinapolylepis	69.52%
	Emiliania huxleyi	78.06%
	Isochrysis sp.	74.36%
	Pavlova lutheri	43.02%
	Pavlovales sp.	80.91%
	Pleurochrysis carterae	85.19%
	Prymnesium parvum	85.75%
	Raphidiophrys ambigua	56.41%
Cyptista	*Cryptomonas paramecium*	87.75%
	Cryptophyceae sp.	80.34%
	Guillardia theta	92.88%
	Goniomonas avonlea	**90.31%**
	Rhodomonas salina	19.66%
	Roombia truncata	72.65%
Viridiplantae	*Arabidopsis thaliana*	90.60%
	Chlamydomonas reinhardtii	88.60%
	Micromonas pusilla	84.05%
	Ostreococcus tauri	80.91%
	Physcomitrella patens	92.88%
	Volvox carteri	89.74%
Glaucophyta	*Cyanophora paradoxa*	87.46%
	Glaucocystis nostochinearum	39.89%
	Gloeochaete witrockiana	90.88%
Rhodophyta	*Chondrus crispus*	79.77%
	Compsopogon coeruleus	60.11%
	Cyanidioschyzon merolae	70.66%
	Galdieria sulphuraria	80.91%
	Porphyra umbilicalus	76.07%

Super-group	Taxa	Percent Gene Coverage
	Porphyridium cruentum	74.64%
	Rhodella maculata	64.10%
Excavata	*Andalucia godoyi*	90.88%
	Andalucia incarcerata	41.31%
	Bodo saltans	82.91%
	Eutreptiella gymnastica	69.80%
	Naegleria gruberi	89.17%
	Trichomonas vaginalis	74.64%
	Trimastix sp.	76.07%
	Tsukubamonas globosa	69.80%
Obazoa	*Thecamonas trahens*	88.32%
	Pygsuia biforma	88.60%
	Allomyces macrogynus	94.02%
	Blastocystis hominis	76.64%
	Homo sapiens	98.58%
	Ministeria vibrans	87.18%
	Monosiga brevicollis	92.59%
	Saccharomyces cerevisiae	85.75%
	Salpingoeca rosetta	90.31%
	Schizosaccharomyces pombe	85.75%
Amoebozoa	*Acanthamoeba castellanii*	80.91%
	Dictyostelium discoideum	92.02%
	Entamoeba invadens	74.93%
	Flamella fluviatilis	75.78%
	Mastigamoeba abducta	83.19%
	Physarum album	83.76%
	Polysphondylium pallidum	90.03%
Orphans	*Diphylleia* sp.	95.44%
	Ancyromonas sigmoides	74.64%
	Malawimonas sp.	86.04%
	Mantamonas plastica	91.45%
	Nutomonas longa	84.05%
	Rigifila ramosa	90.60%

Supplementary Table A5 – Gene abbreviations and corresponding full gene names for a 351 marker gene dataset (Brown, unpublished) based on marker genes used in Brown et al. (2013) (159 genes; highlighted in blue), Burki et al. (2012) (94 genes, highlighted in green) and Kang et al. (2017) (99 genes, highlighted in orange). Highlighted in bold are genes that are also included in the Burki et al. (2016) marker gene dataset (181/250).

Abbreviation	Full Gene Name
AAP	Amino acid permease
ABHD13	abhydrolase domain containing 13
Actin	Actin
ADK2	Adenosine kinase 2
AGB1	GTP binding protein beta 1
AGX	UDP-N-acteylglucosamine pyrophosphorylase 1
AKT	RAC-alpha serine/threonine-protein kinase
AKTIP	AKT-interacting protein
ALAT1	Alanine aminotransferase 1
ALDR	Aldose reductase
ALG11	asparagine-linked glycosylation protein 11
ALIS1	ALA-interacting subunit 1
AMP2B	Antimicrobial peptide 2
AOAH	Acyloxyacyl hydrolase
AP1S2	adaptor-related protein complex 1, sigma 2 subunit
AP3M1	AP-3 complex subunit mu-1
AP3S1	adaptor-related protein complex 3, sigma 1 subunit
AP4M	AP-4 complex subunit mu-1
AP4S1	adaptor-related protein complex 4, sigma 1 subunit
APBLC	Beta-adaptin-like protein C
ar21	Actin-related protein 2/3 complex subunit 3
arf3	ADP-ribosylation factor 1
ARL6	ADP-ribosylation factor-like 6
ARP2	actin-related protein 2
ARP3	actin-related protein 3
arpc1	Clathrin assembly protein complex 1 medium chain
ARPC4	Actin-related protein 2/3 complex subunit 4
ATEHD2	EH domain-containing protein 2
ATG2	Autophagy-related protein 2
atp6	V-type proton ATPase 16 kDa proteolipid subunit c2
ATP6V0A1	ATPase, H+ transporting, lysosomal V0 subunit a1
ATP6V0D1	ATPase, H+ transporting, lysosomal V0 subunit d1
ATPDIL14	Protein disulfide isomerase-like 1-4
ATSAR2	Putative GTP-binding protein, SAR2B
Atub	Tubulin alpha chain
BAT1	Spliceosome RNA helicase BAT1
Btub	Tubulin beta chain
C16orf80	chromosome 16 open reading frame 80
C22orf28	chromosome 22 open reading frame 28

Abbreviation	Full Gene Name
C3H4	C3H4 type zinc finger protein
calr	Calreticulin
capz	F-actin-capping protein subunit beta
CATB	Catalase isozyme B
CC1	Cytochrome c
CCDC113	coiled-coil domain containing 113
CCDC37	coiled-coil domain containing 37
CCDC40	coiled-coil domain containing 40
CCDC65	coiled-coil domain containing 65
cct-A	T-complex protein 1 subunit alpha
cct-B	T-complex protein 1 subunit beta
cct-D	T-complex protein 1 subunit delta
cct-E	T-complex protein 1 subunit epsilon
cct-G	T-complex protein 1 subunit gamma
cct-N	T-complex protein 1 subunit eta
cct-T	T-complex protein 1 subunit theta
cct-Z	T-complex protein 1 subunit zeta
CDK5	CDK5 regulatory subunit associated protein 1-like 1
CLAT	Choline O-acetyltransferase
COP-beta	coatomer protein complex, subunit beta 2
COPE	coatomer protein complex, subunit epsilon
COPG2	coatomer subunit gamma-2
COPS2	COP9 constitutive photomorphogenic homolog subunit 2
COPS6	COP9 constitutive photomorphogenic homolog subunit 6
COQ4-mito	coenzyme Q4 homolog
CORO1C	coronin, actin binding protein, 1C
cpn60	Chaperonin CPN60-like 2, mitochondrial
crfg	Nucleolar GTP-binding protein 1
CRNL1	Crooked neck-like protein 1
CS	citrate synthase
CTP	Dynein light chain 1, cytoplasmic
D2HGDH-mito	D-2-hydroxyglutarate dehydrogenase
DCAF13	DDB1 and CUL4 associated factor 13
DHSA1	Succinate dehydrogenase [ubiquinone] flavoprotein subunit 1, mitochondrial
DHSB3	Succinate dehydrogenase [ubiquinone] iron-sulfur subunit 3, mitochondrial
DHYS	Deoxyhypusine synthase
DIMT1L	DIM1 dimethyladenosine transferase 1-like
DNAI2	dynein, axonemal, intermediate chain 2
DNAJ	Chaperone protein DnaJ
DNAL1	dynein, axonemal, light chain 1
DNM	Dynamin-1-like protein
DPH5	Diphthine methyl ester synthase

Abbreviation	Full Gene Name
DPP3	dipeptidyl-peptidase 3
DRG2	developmentally regulated GTP binding protein 2
ECHM	Enoyl-CoA hydratase, mitochondrial
ef1alpha	Elongation factor 1-alpha
EF2	Elongation factor 2
EFG-mito	G elongation factor
EFTUD1	Elongation factor Tu GTP binding domain-containing protein 1
EIF3B	eukaryotic translation initiation factor 3 subunit B
EIF3C	eukaryotic translation initiation factor 3 subunit C
EIF3I	eukaryotic translation initiation factor 3 subunit I
EIF4A3	eukaryotic translation initiation factor 4A3
EIF4E	eukaryotic translation initiation factor 4E
ERLIN1	ER lipid raft associated 1
ETFA	Electron transfer flavoprotein subunit alpha, mitochondrial
FA2H	fatty acid 2-hydroxylase
FAH	Fumarylacetoacetase
FAM18B	family with sequence similarity 18, member B2
FAM96B	family with sequence similarity 96, member B
FAM	family with sequence similarity 49, member B
fh	fumarase hydratase
fibri	rRNA 2'-O-methyltransferase fibrillarin 2
FOLD	Bifunctional protein FolD
fpps	Farnesyl pyrophosphate synthase 2
FTSJ1	Putative tRNA
G6PD6	Glucose-6-phosphate 1-dehydrogenase, cytoplasmic
GAS8	growth arrest-specific 8
GCST	Aminomethyltransferase, mitochondrial
gdi2	Rab GDP dissociation inhibitor beta
GDI	Rab GDP dissociation inhibitor alpha
glcn	UDP-N-acetylglucosamine--dolichyl-phosphate N-acetylglucosamine phosphotransferase
GLGB2	1,4-alpha-glucan branching enzyme GlgB 2
GMPP3	mannose-1-phosphate guanylyltransferase 3
gnb2l	Guanine nucleotide-binding protein subunit beta-like protein A
gnbpa	Guanine nucleotide-binding protein alpha-1 subunit
GNL2	Nucleolar GTP-binding protein 2
GPD1L	glycerol-3-phosphate dehydrogenase 1-like
grc5	60S ribosomal protein L10-2
GRWD1	glutamate-rich WD repeat containing 1
GSS	glutathione synthetase
Gtub	Tubulin gamma-2 chain
H2A	Histone 2A
H2B	Histone 2B
h3	Histone H3

Abbreviation	Full Gene Name
h4	Histone H4
HDDC2	HD domain-containing protein 2
HGO	Homogentisate 1,2-dioxygenase
HM13	Minor histocompatibility antigen H13
hmt1	Arginine methyltransferase pam1
HSP70C	Heat shock cognate 70 kDa protein 1
hsp70mt	Heat shock 70 kDa protein, mitochondrial
HSP90	Heat shock protein 90
HYOU1	hypoxia up-regulated 1
if2b	Eukaryotic translation initiation factor 2 subunit beta
if2g	Eukaryotic translation initiation factor 2 subunit gamma
if2p	Eukaryotic translation initiation factor 5B
if6	Eukaryotic translation initiation factor 6-1
IFT46	intraflagellar transport 46 homolog
IFT57	intraflagellar transport 57 homolog
IFT88	intraflagellar transport 88 homolog
IMB1	Importin subunit beta-1
IMP4	IMP4, U3 small nucleolar ribonucleoprotein
ino1	Inositol-3-phosphate synthase 1
IP5PD	Type I inositol polyphosphate 5-phosphatase 13
IPO4	importin-4
IPO5	importin-5
ITIH4	Inter-alpha-trypsin inhibitor heavy chain H4
KARS	lysyl-tRNA synthetase
KDELR2	KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 2
l10a	60S ribosomal protein L10a-1
l12e-D	60S ribosomal protein L7a
LRRC48	leucine rich repeat containing 48
LTA4H	Leukotriene A-4 hydrolase
mat	S-adenosylmethionine synthase 1
mcm-A	DNA replication licensing factor MCM5
mcm-B	DNA replication licensing factor MCM2
mcm-C	DNA replication licensing factor MCM3 homolog
mcm-D	DNA replication licensing factor MCM7
mcm-E	DNA replication licensing factor MCM4
metap2	Methionine aminopeptidase 2
METTL1	tRNA (guanine-N(7)-)-methyltransferase isoform a
MLST8	Target of rapamycin complex subunit LST8
MMAA-mito	methylmalonic aciduria
MOCS3	Adenylyltransferase and sulfurtransferase MOCS3
mra1	Multicopy suppressor of ras1
MTHFR	methylenetetrahydrofolate reductase
MTLPD2	Dihydrolipoyl dehydrogenase 2, mitochondrial

Abbreviation	Full Gene Name
MYG1	UPF0160 protein MYG1, mitochondrial
NAA15	N(alpha)-acetyltransferase 15, NatA auxiliary subunit
NAE1	NEDD8 activating enzyme E1 subunit 1
NAPA	Alpha-soluble NSF attachment protein
ndf1	NADH dehydrogenase [ubiquinone] flavoprotein 1
NDPK2	Nucleoside diphosphate kinase 2
NDUFV2-mito	NADH dehydrogenase (ubiquinone) flavoprotein 2
NFS1-mito	NFS1 nitrogen fixation 1
NLN-mito	neurolysin
NMD3	Nonsense-mediated mRNA decay protein 3
NMT1	Glycylpeptide N-tetradecanoyltransferase 1
NOP5A	nucleolar protein 5-1
NSA2	NSA2 ribosome biogenesis homolog
nsf1-C	Vacuolar protein sorting-associated protein 4
nsf1-E	Mitochondrial inner membrane i-AAA protease supercomplex subunit YME1
nsf1-G	26S protease regulatory subunit 8 homolog A
nsf1-H	ATPase family AAA domain-containing protein 1
nsf1-I	26S protease regulatory subunit 7
nsf1-J	26S protease regulatory subunit 10B
nsf1-K	26S protease regulatory subunit 6A
nsf1-L	26S protease regulatory subunit 6B
nsf1-M	26S proteasome regulatory subunit 4
nsf2-A	Cell division control protein 48 homolog E
nsf2-F	Vesicle-fusing ATPase 2: NSF (N-ethylmaleimide sensitive factor)
ODB2	Organellar DNA-binding protein 2
ODBA	2-oxoisovalerate dehydrogenase subunit alpha
ODBB	2-oxoisovalerate dehydrogenase subunit beta
ODO2A	Dihydrolipoamide succinyltransferase component of 2-oxoglutarate dehydrogenase complex 1
ODPA2	Pyruvate dehydrogenase E1 component subunit alpha-2
ODPB	Pyruvate dehydrogenase E1 component subunit beta
oplah	5-oxoprolinase
orf2	RNA-binding protein pno1
osgep	O-sialoglycoprotein endopeptidase
PABPC4	poly(A) binding protein, cytoplasmic 4
pace2-A	GPN-loop GTPase 1 homolog
pace2B	GPN-loop GTPase 2
Pace2C	GPN-loop GTPase 3
pace5	Ribosome maturation protein SBDS
PACRG	PARK2 co-regulated
PCY2	Ethanolamine-phosphate cytidylyltransferase
PELO	Protein pelota homolog
PGM2	Phosphoglucomutase-2

Abbreviation	Full Gene Name
PGMP	Phosphoglucomutase
PIK3C3	phosphatidylinositol 3-kinase catalytic subunit type 3
PLS3	Plastin-3
PMM2	phosphomannomutase 2
PMPCB	Mitochondrial-processing peptidase subunit beta
pp2A-b	Serine/threonine-protein phosphatase PP2A-2 catalytic subunit
PP2BC	Serine/threonine-protein phosphatase PP2B catalytic subunit
PPP2R3	protein phosphatase 2, regulatory subunit B, alpha
PPP2R5C	protein phosphatase 2, regulatory subunit B, gamma
PPX2	Serine/threonine-protein phosphatase PP-X isozyme 2
PR19A	Pre-mRNA-processing factor 19 homolog 1
PROSC	proline synthetase co-transcribed
PSD11	26S proteasome non-ATPase regulatory subunit 11
PSD7	26S proteasome non-ATPase regulatory subunit 7
psma-A	Proteasome subunit alpha type-5-A
psma-B	Proteasome subunit alpha type-7-B
psma-C	Proteasome subunit alpha type-4
psma-E	Proteasome subunit alpha type-1-A
psma-F	Proteasome subunit alpha type-3
psma-G	Proteasome subunit alpha type-6-A
psma-H	Proteasome subunit beta type-2-A
psma-J	Proteasome subunit beta type-1
psmb-K	Proteasome subunit beta type-7-B
psmb-L	Proteasome subunit beta type-6
psmb-M	Proteasome subunit beta type-5-B
psmb-N	Proteasome subunit beta type-4
PSMD12	Proteasome 26S subunit, non-ATPase, 12
PSMD6	26S proteasome non-ATPase regulatory subunit 6
psmd	26S proteasome non-ATPase regulatory subunit 14
PTPL	Protein tyrosine phosphatase
PURA	Adenylosuccinate synthetase
PYGB	phosphorylase, glycogen
rac	Rac-like GTP-binding protein RAC1
rad23	Probable DNA repair protein RAD23
Rad51A	DNA repair protein RAD51 homolog 1
ran	GTP-binding nuclear protein Ran
RBX1	ring-box 1, E3 ubiquitin protein ligase
rf1	Eukaryotic peptide chain release factor subunit 1-2
RHEB	GTP-binding protein Rheb
RICTOR	Rapamycin-insensitive companion of mTOR
rla2a	60S acidic ribosomal protein P2-1
rla2b	60S acidic ribosomal protein P1-2
RPAC1	DNA-directed RNA polymerases I and III subunit RPAC1
RPF1	ribosome production factor 1

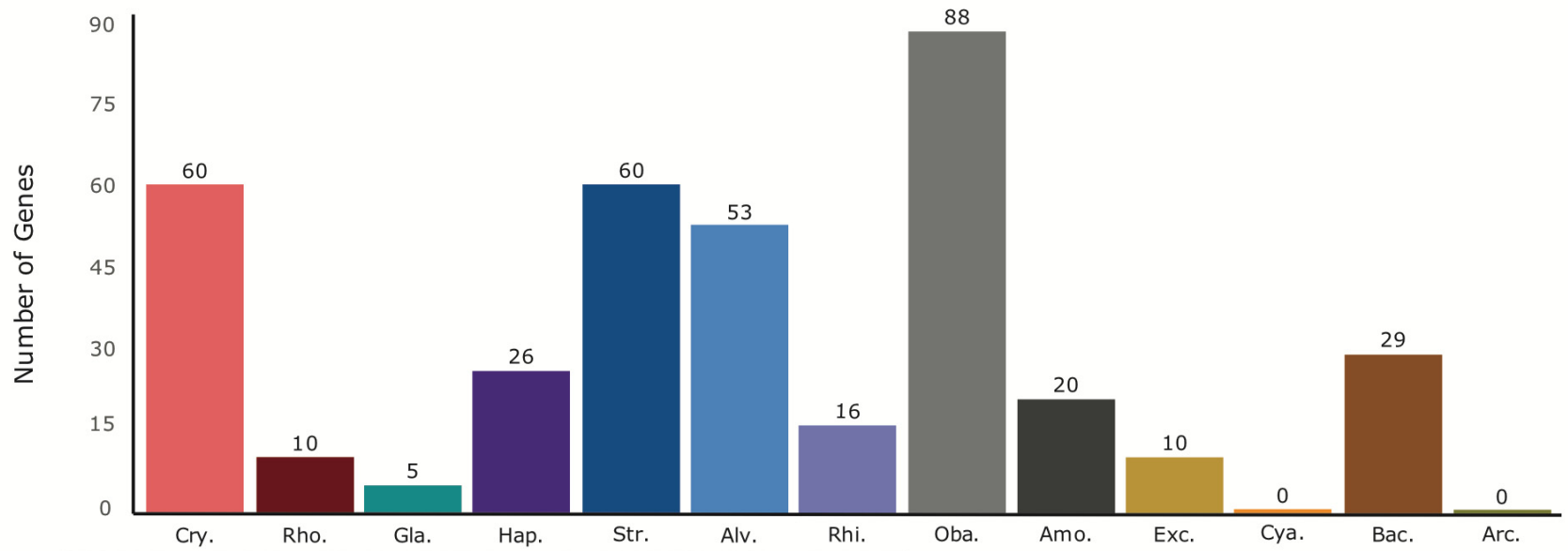
Abbreviation	Full Gene Name
rpl11	60S ribosomal protein L11-2
rpl12	60S ribosomal protein L12-1
Rpl13A	60S ribosomal protein L13a-1
Rpl13e	60S ribosomal protein L13-1
Rpl14e	60S ribosomal protein L14-1
Rpl15	60S ribosomal protein L15-2
rpl17	60S ribosomal protein L17-2
Rpl18	60S ribosomal protein L18-3
rpl19	60S ribosomal protein L19-2
rpl20	60S ribosomal protein L18a-2
rpl21	60S ribosomal protein L21-1
Rpl24A	60S ribosomal protein L24-1
rpl26	60S ribosomal protein L26-2
rpl27	60S ribosomal protein L27a-3
Rpl2	60S ribosomal protein L8-2
rpl30	60S ribosomal protein L30-1
rpl31	60S ribosomal protein L31-1
rpl32	60S ribosomal protein L32-2
rpl33	60S ribosomal protein L35a-1
rpl35	60S ribosomal protein L35-3
Rpl3	60S ribosomal protein L3-2
rpl43	60S ribosomal protein L37a-2
rpl44	60S ribosomal protein L36a
Rpl4b	60S ribosomal protein L4-2
Rpl5	60S ribosomal protein L5-2
rpl6	60S ribosomal protein L6-1
Rpl7a	60S ribosomal protein L7-3
rpl9	60S ribosomal protein L9-1
RPN1B	26S proteasome non-ATPase regulatory subunit 2 homolog B
rpo-A	DNA-directed RNA polymerase I subunit rpa1
rpo-B	DNA-directed RNA polymerase II subunit RPB1
rpo-C	DNA-directed RNA polymerase III subunit RPC1
RPPK	Ribose-phosphate pyrophosphokinase
rppO	60S acidic ribosomal protein P0-3
rps10	40S ribosomal protein S10-1
rps11	40S ribosomal protein S11-1
rps12	40S ribosomal protein S12
rps14	40S ribosomal protein S14-1
rps15	40S ribosomal protein S15-1
rps16	40S ribosomal protein S16-1
rps17	40S ribosomal protein S17-1
rps18	40S ribosomal protein S18
rps20	40S ribosomal protein S20-1
rps23	40S ribosomal protein S23-1
rps26	40S ribosomal protein S26-1

Abbreviation	Full Gene Name
rps27	40S ribosomal protein S27-1
rps2	40S ribosomal protein S2-1
rps3	40S ribosomal protein S3-1
rps4	40S ribosomal protein S4-1
rps5	40S ribosomal protein S5-1
rps6	40S ribosomal protein S6-1
rps8	40S ribosomal protein S8-1
RPTOR	associated protein of mTOR
RRAGD	Ras-related GTP-binding protein D
RRM1	ribonucleotide reductase M1
s15a	40S ribosomal protein S15a
s15p	40S ribosomal protein S13
sap40	40S ribosomal protein Sa-1
SCO1-mito	Protein SCO1 homolog, mitochondrial
SCSB	Succinate--CoA ligase [ADP-forming] subunit beta
SEC23	Protein transport protein SEC23
SF3B2	Splicing factor 3B subunit 2
SND1	staphylococcal nuclease and tudor domain containing 1
SPTC2	Serine palmitoyltransferase 2
SPTLC1	serine palmitoyltransferase 1 isoform a
sra	Signal recognition particle receptor subunit alpha
srp54	Signal recognition particle 54 kDa protein
STXBP1	syntaxin binding protein 1
suca	Succinyl-CoA ligase [ADP-forming] subunit alpha-2
SYGM1	Glycine--tRNA ligase
SYNJ	Synaptojanin
TAL	Transaldolase
tfiid	TATA-box-binding protein 1
TM9SF1	transmembrane 9 superfamily member 1 isoform a
TMS	TMS
topo1	DNA topoisomerase 1
trs	Threonyl-tRNA synthetase
UBA3	ubiquitin-like modifier activating enzyme 3
ubc	Ubiquitin-conjugating enzyme E2 9
UBE12	Ubiquitin-activating enzyme E1 2
UBE2J2	ubiquitin-conjugating enzyme E2, J2
Ubq	Ubiquitin
VAPA	Vesicle-associated membrane protein-associated protein A
VARS	valyl-tRNA synthetase
vata	V-type proton ATPase catalytic subunit A
vatb	V-type proton ATPase subunit B2
vatc	V-type proton ATPase subunit C
vate	V-type proton ATPase subunit E
VBP1	von Hippel-Lindau binding protein 1

Abbreviation	Full Gene Name
VPS18	vacuolar protein sorting 18
VPS26B	vacuolar protein sorting 26
WBSCR22	Williams Beuren syndrome chromosome region 22
WD66	66 kDa stress protein
wd	WD repeat domain phosphoinositide-interacting protein 3
wrs	tRNA synthetase class I (W and Y) family protein
xpb	DNA repair helicase XPB1
XRP2	retinitis pigmentosa 2
YKT6	YKT6 v-SNARE homolog

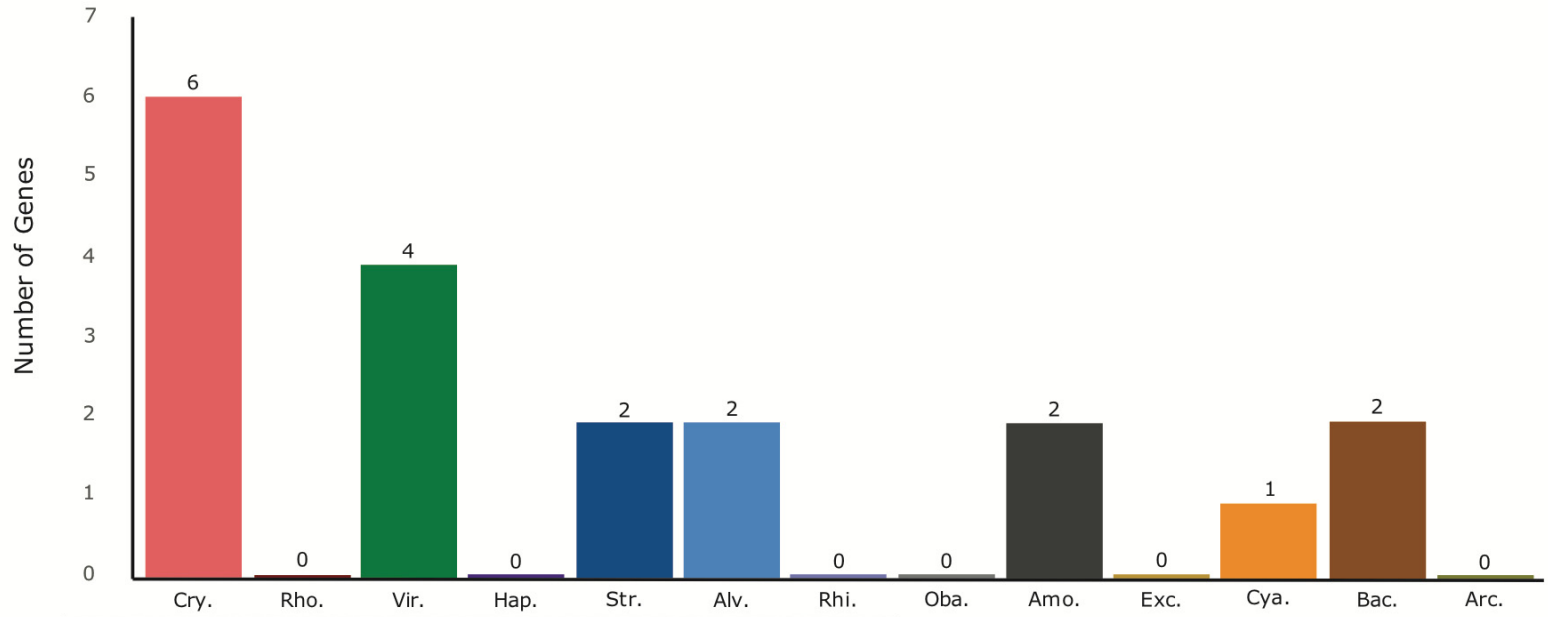
APPENDIX B – SUPPLEMENTARY FIGURES FOR CHAPTER 3

Figure B1: The phylogenetic position of *Goniomonas avonlea* across all single gene trees generated from the combined predicted proteins and gene models' dataset where *G. avonlea* branches sister to Viridiplantae. Phylogenetic position was determined as the super-group of the majority of OTUS in the closest clade to *G. avonlea* and Viridiplantae (i.e. nearest neighbor) with bootstrap support ≥ 70% (shown in the histogram). An additional round of topology detection was used to determine the next nearest neighbor to any clade showing a relationship between *G. avonlea*, Viridiplantae and an additional photosynthetic eukaryotic group (shown in a table below the histogram). Super groups shown are abbreviated as follows: Cry = Cryptista, Rho. = Rhodophyta, Vir = Viridiplantae, Gla = Glaucophyta, Hap = Haptista, Str = Stramenopiles, Alv = Alveolata, Rhi = Rhizaria, Oba = Obazoa, Amo = Amoebozoa, Cya = Cyanobacteria, Bac = Bacteria (non-cyanobacteria), and Arc = Archaea. A dash ('-') indicates no tree showed the corresponding branching pattern.



Next Neighboring Taxa							
Cry.	-	1	1	2	4	6	1
Rho.	3	-	-	1	1	-	-
Gla.	1	-	-	1	-	-	-
Hap.	6	1	-	-	3	7	-
Str.	6	-	1	2	-	10	1
Alv.	4	-	-	1	11	-	-
Rhi.	7	-	-	-	5	1	-
Oba.	11	2	-	2	7	7	3
Amo.	-	-	1	1	-	2	1
Exc.	-	-	-	-	3	-	-
Cya.	-	-	-	-	-	-	-
Bac.	1	-	-	1	2	1	-
Arc.	-	-	-	1	-	1	-

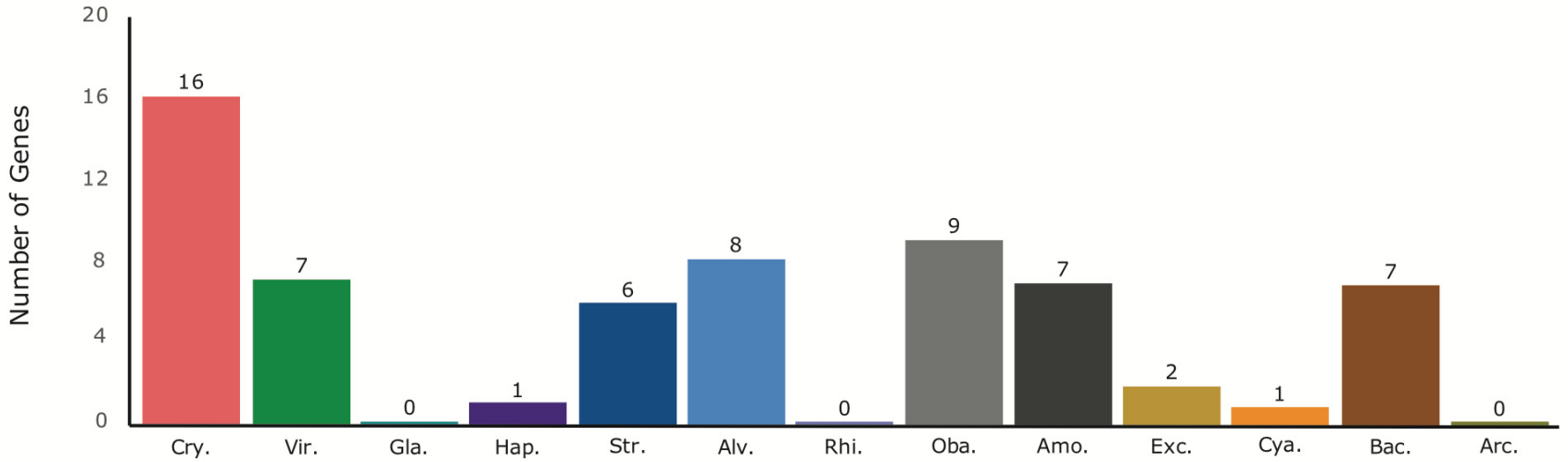
Figure B2: The phylogenetic position of *Goniomonas avonlea* across all single gene trees generated from the combined predicted proteins and gene models' dataset where *G. avonlea* branches with Glaucophyta. Phylogenetic position was determined as the super-group of the majority of OTUS in the closest clade to *G. avonlea* and Glaucophyta (i.e. nearest neighbor) with bootstrap support ≥ 70% (shown in the histogram). An additional round of topology detection was used to determine the next nearest neighbor to any clade showing a relationship between *G. avonlea*, Glaucophyta and an additional photosynthetic eukaryotic group (shown in a table below the histogram). Super groups shown are abbreviated as follows: Cry = Cryptista, Rho. = Rhodophyta, Vir = Viridiplantae, Gla = Glaucophyta, Hap = Haptista, Str = Stramenopiles, Alv = Alveolata, Rhi = Rhizaria, Oba = Obazoa, Amo = Amoebozoa, Cya = Cyanobacteria, Bac = Bacteria (non-cyanobacteria), and Arc = Archaea. A dash ('-') indicates no tree showed the corresponding branching pattern.



Next Neighboring Taxa

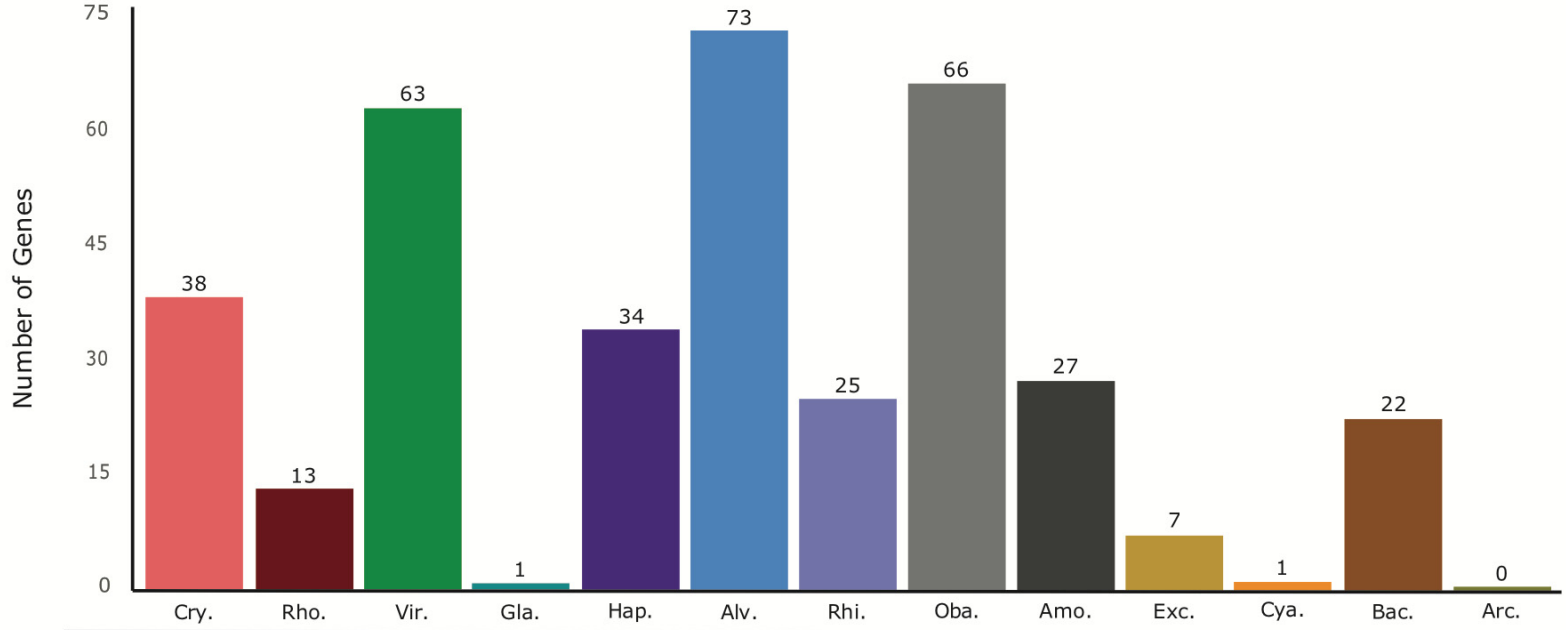
	Cry.	Rho.	Vir.	Hap.	Str.	Alv.	Rhi.
Cry.	-	-	1	-	-	-	-
Rho.	-	-	-	-	-	-	-
Vir.	2	-	-	-	-	-	-
Hap.	-	-	-	-	-	-	-
Str.	-	-	-	-	-	1	-
Alv.	-	-	-	-	-	-	-
Rhi.	-	-	-	-	-	-	-
Oba.	-	-	1	-	-	-	-
Amo.	-	-	-	-	-	-	-
Exc.	-	-	-	-	-	-	-
Cya.	-	-	-	-	-	-	-
Bac.	-	-	-	-	-	-	-
Arc.	-	-	-	-	-	-	-

Figure B3: The phylogenetic position of *Goniomonas avonlea* across all single gene trees generated from the combined predicted proteins and gene models' dataset where *G. avonlea* branches sister to Rhodophyta. Phylogenetic position was determined as the super-group of the majority of OTUS in the closest clade to *G. avonlea* and Rhodophyta (i.e. nearest neighbor) with bootstrap support ≥ 70% (shown in the histogram). An additional round of topology detection was used to determine the next nearest neighbor to any clade showing a relationship between *G. avonlea*, Rhodophyta and an additional photosynthetic eukaryotic group (shown in a table below the histogram). Super groups shown are abbreviated as follows: Cry = Cryptista, Rho. = Rhodophyta, Vir = Viridiplantae, Gla = Glaucophyta, Hap = Haptista, Str = Stramenopiles, Alv = Alveolata, Rhi = Rhizaria, Oba = Obazoa, Amo = Amoebozoa, Cya = Cyanobacteria, Bac = Bacteria (non-cyanobacteria), and Arc = Archaea. A dash ('-') indicates no tree showed the corresponding branching pattern.



Next Neighboring Taxa	Cry.	Vir.	Gla.	Hap.	Str.	Alv.	Rhi.
Cry.	-	-	-	-	-	2	-
Vir.	1	-	-	-	1	1	-
Gla.	-	-	-	-	-	-	-
Hap.	2	1	-	-	1	-	-
Str.	2	-	-	-	-	-	-
Alv.	2	-	-	-	-	-	-
Rhi.	1	-	-	-	-	-	-
Oba.	-	3	-	-	-	1	-
Amo.	1	-	-	-	-	-	-
Exc.	-	-	-	-	-	-	-
Cya.	-	-	-	1	-	-	-
Bac.	1	1	-	-	3	1	-
Arc.	-	-	-	-	-	-	-

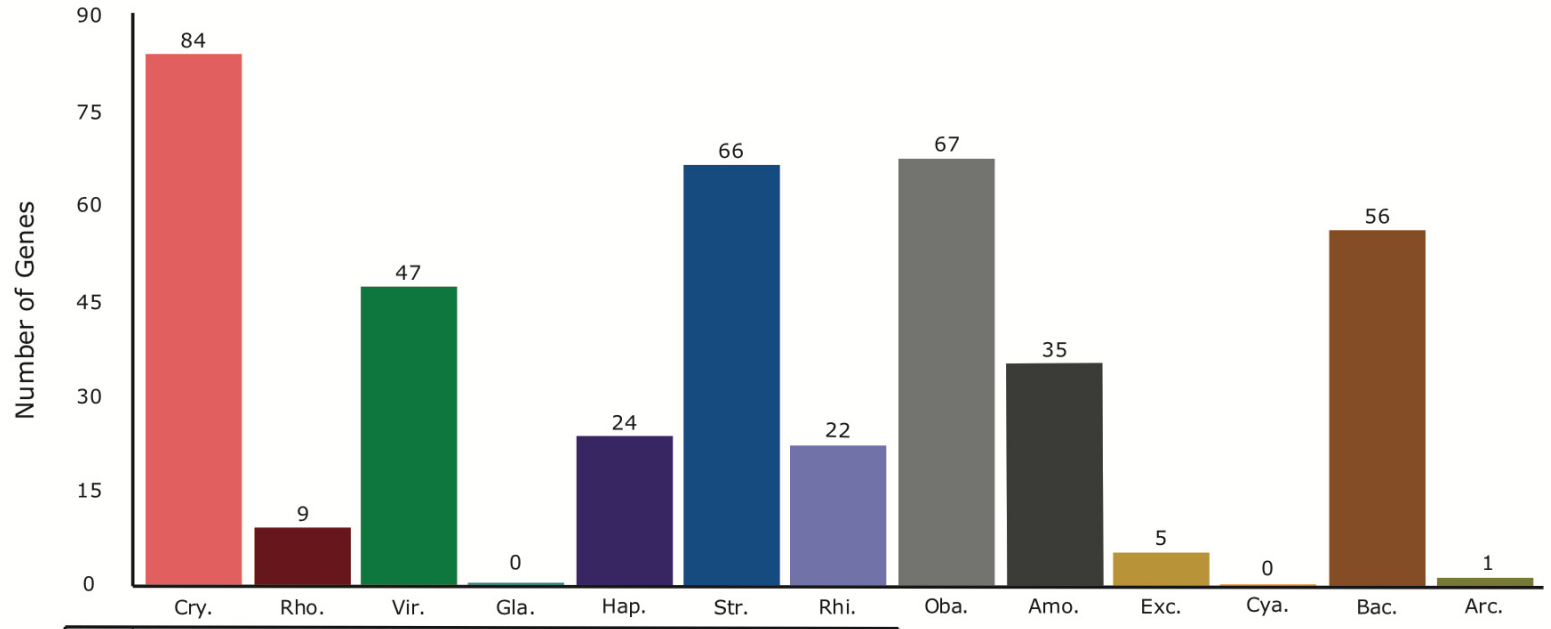
Figure B4: The phylogenetic position of *Goniomonas avonlea* across all single gene trees generated from the combined predicted proteins and gene models' dataset where *G. avonlea* branches sister to Stramenopiles. Phylogenetic position was determined as the super-group of the majority of OTUS in the closest clade to *G. avonlea* and Stramenopiles (i.e. nearest neighbor) with bootstrap support ≥ 70% (shown in the histogram). An additional round of topology detection was used to determine the next nearest neighbor to any clade showing a relationship between *G. avonlea*, Stramenopiles and an additional photosynthetic eukaryotic group (shown in a table below the histogram). Super groups shown are abbreviated as follows: Cry = Cryptista, Rho. = Rhodophyta, Vir = Viridiplantae, Gla = Glaucophyta, Hap = Haptista, Str = Stramenopiles, Alv = Alveolata, Rhi = Rhizaria, Oba = Obazoa, Amo = Amoebozoa, Cya = Cyanobacteria, Bac = Bacteria (non-cyanobacteria), and Arc = Archaea. A dash ('-') indicates no tree showed the corresponding branching pattern.



Next Neighboring Taxa

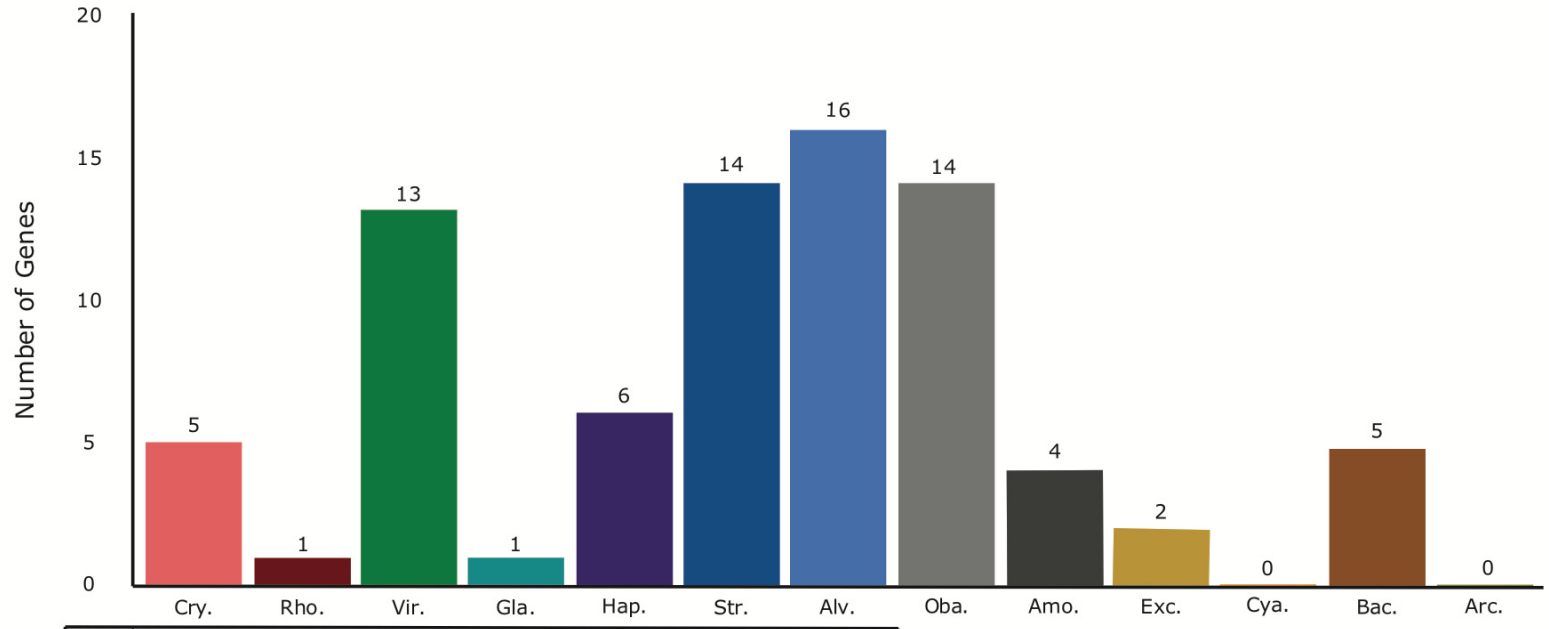
Cry.	-	1	6	-	4	6	-
Rho.	-	-	-	-	-	-	-
Vir.	1	-	-	-	5	10	4
Gla.	-	-	-	-	-	-	-
Hap.	5	-	8	-	-	6	-
Alv.	7	1	5	-	5	-	2
Rhi.	2	-	2	-	2	2	-
Oba.	2	2	6	1	3	13	5
Amo.	-	-	1	-	2	4	1
Exc.	1	-	-	-	1	2	-
Cya.	-	-	-	-	-	-	-
Bac.	2	2	4	-	6	9	2
Arc.	-	2	-	-	-	-	-

Figure B5: The phylogenetic position of *Goniomonas avonlea* across all single gene trees generated from the combined predicted proteins and gene models' dataset where *G. avonlea* branches sister to Alveolata. Phylogenetic position was determined as the super-group of the majority of OTUS in the closest clade to *G. avonlea* and Alveolata (i.e. nearest neighbor) with bootstrap support ≥ 70% (shown in the histogram). An additional round of topology detection was used to determine the next nearest neighbor to any clade showing a relationship between *G. avonlea*, Alveolata and an additional photosynthetic eukaryotic group (shown in a table below the histogram). Super groups shown are abbreviated as follows: Cry = Cryptista, Rho. = Rhodophyta, Vir = Viridiplantae, Gla = Glaucophyta, Hap = Haptista, Str = Stramenopiles, Alv = Alveolata, Rhi = Rhizaria, Oba = Obazoa, Amo = Amoebozoa, Cya = Cyanobacteria, Bac = Bacteria (non-cyanobacteria), and Arc = Archaea. A dash ('-') indicates no tree showed the corresponding branching pattern.



Next Neighboring Taxa	Cry.	Rho.	Vir.	Gla.	Hap.	Str.	Rhi.
Cry.	-	-	3	-	5	6	2
Rho.	1	-	-	-	-	1	1
Vir.	9	1	-	-	4	9	-
Gla.	-	-	1	-	-	-	-
Hap.	4	-	3	-	-	7	5
Str.	8	2	9	-	4	-	1
Rhi.	-	-	4	-	-	6	-
Oba.	9	-	7	-	2	6	2
Amo.	1	1	2	-	1	8	2
Exc.	1	-	1	-	1	3	1
Cya.	1	-	-	-	-	1	-
Bac.	3	-	3	-	1	5	4
Arc.	-	-	-	-	-	-	-

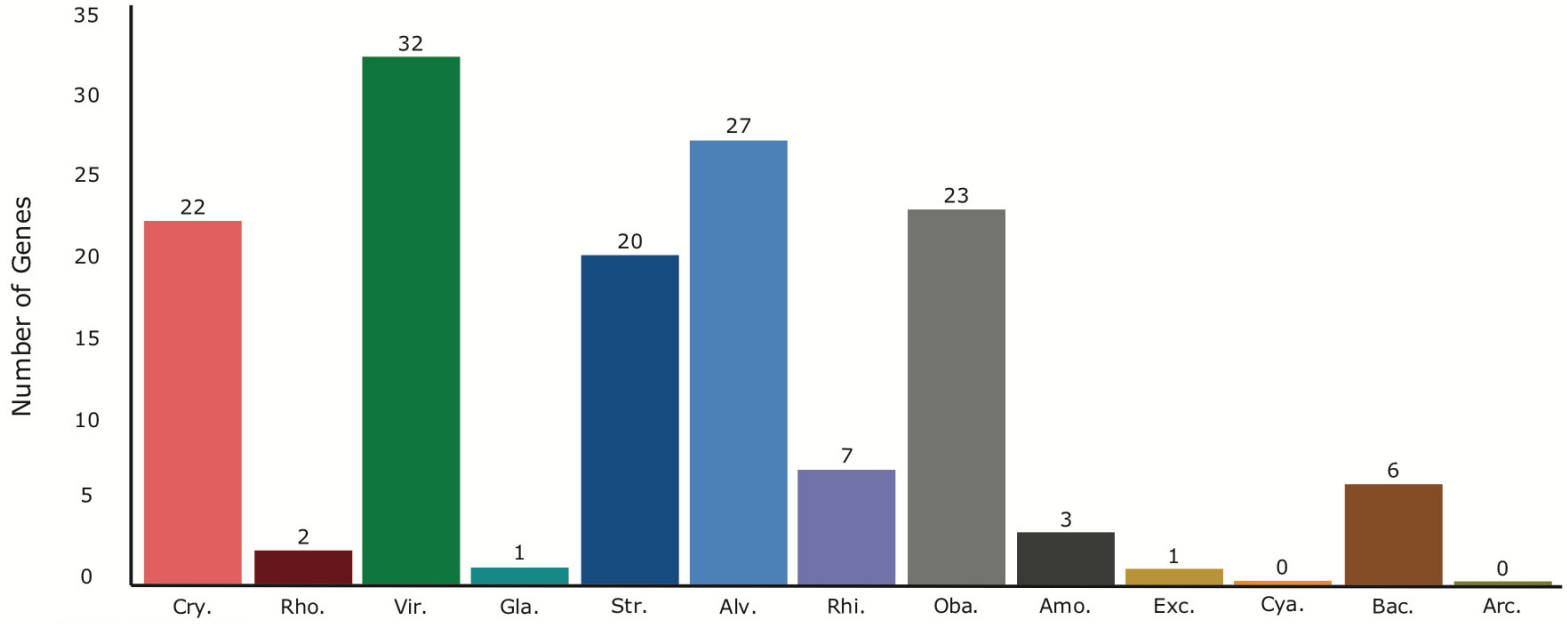
Figure B6: The phylogenetic position of *Goniomonas avonlea* across all single gene trees generated from the combined predicted proteins and gene models' dataset where *G. avonlea* branches sister to Rhizaria. Phylogenetic position was determined as the super-group of the majority of OTUS in the closest clade to *G. avonlea* and Rhizaria (i.e. nearest neighbor) with bootstrap support ≥ 70% (shown in the histogram). An additional round of topology detection was used to determine the next nearest neighbor to any clade showing a relationship between *G. avonlea*, Rhizaria and an additional photosynthetic eukaryotic group (shown in a table below the histogram). Super groups shown are abbreviated as follows: Cry = Cryptista, Rho. = Rhodophyta, Vir = Viridiplantae, Gla = Glaucophyta, Hap = Haptista, Str = Stramenopiles, Alv = Alveolata, Rhi = Rhizaria, Oba = Obazoa, Amo = Amoebozoa, Cya = Cyanobacteria, Bac = Bacteria (non-cyanobacteria), and Arc = Archaea. A dash ('-') indicates no tree showed the corresponding branching pattern.



Next Neighboring Taxa

Cry.	-	-	-	-	-	1	-
Rho.	--	-	-	-	-	1	-
Vir.	-	-	-	-	1	-	1
Gla.	-	-	1	-	-	-	-
Hap.	-	-	2	-	-	-	2
Str.	-	1	3	-	1	-	4
Alv.	-	-	-	-	-	5	-
Oba.	2	-	-	-	-	1	1
Amo.	-	-	-	-	-	-	1
Exc.	-	-	-	-	-	-	1
Cya.	-	-	-	-	-	-	-
Bac.	-	-	2	1	2	-	-
Arc.	-	-	-	-	-	-	-

Figure B7: The phylogenetic position of *Goniomonas avonlea* across all single gene trees generated from the combined predicted proteins and gene models' dataset where *G. avonlea* branches sister to Haptista. Phylogenetic position was determined as the super-group of the majority of OTUS in the closest clade to *G. avonlea* and Haptista (i.e. nearest neighbor) with bootstrap support ≥ 70% (shown in the histogram). An additional round of topology detection was used to determine the next nearest neighbor to any clade showing a relationship between *G. avonlea*, Haptista and an additional photosynthetic eukaryotic group (shown in a table below the histogram). Super groups shown are abbreviated as follows: Cry = Cryptista, Rho. = Rhodophyta, Vir = Viridiplantae, Gla = Glaucophyta, Hap = Haptista, Str = Stramenopiles, Alv = Alveolata, Rhi = Rhizaria, Oba = Obazoa, Amo = Amoebozoa, Cya = Cyanobacteria, Bac = Bacteria (non-cyanobacteria), and Arc = Archaea. A dash ('-') indicates no tree showed the corresponding branching pattern.



Next Neighboring Taxa

Cry.	-	-	3	-	3	1	-
Rho.	-	-	-	-	1	-	-
Vir.	2	-	-	-	1	3	1
Gla.	-	-	-	-	-	-	-
Str.	1	-	4	-	-	4	-
Alv.	2	-	5	-	4	-	1
Rhi.	1	-	2	-	2	1	-
Oba.	6	-	6	-	-	-	3
Amo.	-	-	1	-	1	3	-
Exc.	-	-	-	-	1	2	-
Cya.	-	1	-	-	-	-	-
Bac.	-	-	2	-	1	2	-
Arc.	-	-	-	-	-	-	-

REFERENCES

Allen, J., Raven, J., Allen, J., & Raven, J. 1996. Free-radical-induced mutation vs redox regulation: Costs and benefits of genes in organelles. J. Mol. Evol., 5:482-492. doi:10.1007/BF02352278

Anderson, O. R. 1977. Fine structure of a marine ameba [sic] associated with a blue-green alga in the Sargasso Sea. J. Protozool., 24:370-376. doi: 10.1111/j.1550-7408.1977.tb04753.x

Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C., & Gascuel, O. 2011. Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. Syst. Biol., 60: 685–699. doi:10.1093/sysbio/syr041

Archibald, J.M., Rogers, M. B., Toop, M., Ishida, K., & Keeling, P. J. 2003. Lateral gene transfer and the evlution of plastid-targeted proteins in the secondary plastid-containing alga *Bigelowiella natans*. Proc. Natl. Acad. Sci. 100:7678-7683. doi:10.1073/pnas.1230951100

Archibald, J. M. 2009. The puzzle of plastid evolution. Curr. Biol., 19:R81-R88. doi:10.1016/j.cub.2008.11.067

Archibald, J. M. 2015. Genomic perspectives on the birth and spread of plastids. Proc. R. Soc. B, 33:10147–10153. doi:10.1073/pnas.1421374112

Barbrook, A. C., Howe, C. J., & Purton, S. 2006. Why are plastid genomes retained in non-photosynthetic organisms? Trends Plant Sci., 2:101–108. doi:10.1016/j.tplants.2005.12.004

Bodył, A. 2017. Did some red alga-derived plastids evolve via kleptoplastidy? A hypothesis. Biol. Rev. Camb. Philos. Soc., doi:10.1111/brv.12340

Bodył, A., Stiller, J., & Mackiewicz, P. 2009. Chromalveolate plastids: direct descent or multiple endosymbioses? Trends Ecol. Evolut., 3:119–121. doi:10.1016/j.tree.2008.11.003

Brown, M., Sharpe, S., Silberman, J., Heiss, A., Lang, F., Simpson, A., & Roger, A. 2013. Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. Proc. R. Soc. Lond. B: Biol., 280:20131755. doi:10.1098/rspb.2013.1755

Buchfink, B., Xie, C., & Huson, D. 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods, 12:59–60. doi:10.1038/nmeth.3176

Burki, F, Inagaki, Y, Bråte, J, & Archibald, J. M. 2009. Large-scale phylogenomic analyses reveal that two enigmatic protist lineages, Telonemia and Centroheliozoa, are related to photosynthetic chromalveolates. Genome Biol. Evol., 1:231–238. doi:10.1093/gbe/evp022

Burki, F., Flegontov, P., Oborník, M., Cihlář, J., Pain, A., Lukeš, J., & Keeling, P. 2012a. Re-evaluating the Green versus Red Signal in Eukaryotes with Secondary Plastid of Red Algal Origin. Genome Biol. Evol., 4: 626–635. doi:10.1093/gbe/evs049

Burki, F., Okamoto, N., Pombert, J.-F., & Keeling, P. 2012b. The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. Proc. R. Soc. Lond. Biol., 1736:2246–2254. doi:10.1098/rspb.2011.2301

Burki, F., Kaplan, M., Tikhonenkov, D., Zlatogursky, V., Minh, B., Radaykina, L., … Keeling, P. 2016a. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. Proc. R. Soc. B, 283:20152802. doi:10.1098/rspb.2015.2802

Burki F, Kaplan M, Tikhonekov DV, Zlatogursky V, Minh BQ, Radaykina LV, Smirnov A, Mylnikov AP, Keeling PJ. 2016b. Data from: Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta, and Cryptista. Dryad Digital Repository. doi:org/10.5061/dryad.rj87v

Burki, F. 2017. The Convoluted Evolution of Eukaryotes With Complex Plastids. Adv. Bot. Res. *In Press.* doi:10.1016/bs.abr.2017.06.001

Caraguel, C. G. B., O'Kelly, C. J., Legendre, P., Frasca Jr., S., Gast, R. J., Després, B. M., Cawthorn, R.J., & Greenwood, S. J. 2007. Microheterogeneity and coevolution: An examination of rDNA sequence characteristics in Neoparamoeba pemaquidensis and its prokinetoplastid endosymbiont. J. Euk. Microbiol., 54:418-426. doi:10.1111/j.1550-7408.2007.00281.x

Cavalier-Smith, T. 1999. Principles of Protein and Lipid Targeting in Secondary Symbiogenesis: Euglenoid, Dinoflagellate, and Sporozoan Plastid Origins and the Eukaryote Family Tree 1, 2. J. Euk. Microbiol., 4:347–366. doi:10.1111/j.1550-7408.1999.tb04614.x

Criscuolo, A. & Gribaldo, S. 2010. BMGE (block mapping and gathering with entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol., 10:210-231. doi:10.1186/1471-2148-10-210

Curtis, B. A., 2012. Endosymbiotic gene transfer in the nucleomorph containing organisms *Bigelowiella natans* and *Guillardia theta*. Halifax, NS: Dalhousie University.

Curtis, B., Tanifuji, G., Burki, F., Gruber, A., Irimia, M., Maruyama, S., … Archibald, J. 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature, 492:59–65. doi:10.1038/nature11681

Daley, D., & Whelan, J. 2005. Why genes persist in organelle genomes. Genome Biol., 110. doi: 10.1186/gb-2005-6-5-110

Deschamps, P. & Moreira, D. 2012. Reevaluating the green contribution to diatom genomes. Genome. Bio. Evol. 4:683-688. doi:10.1093/gbe/evs/053

De Vienne, D., Ollier, S., & Aguileta, G. 2012. Phylo-MCOA: a fast and efficient method to detect outlier genes and species in phylogenomics using multiple co-inertia analysis. Mol. Biol. Evol., 29:1587–98. doi:10.1093/molbev/msr317

Dolezal, P., Likic, V., Tachezy, J., & Lithgow, T. 2006. Evolution of the molecular machines for protein import into mitochondria. Science, 5785:314–318. doi:10.1126/science.1127895

Doolittle, F. 1998. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet., 14:307–311. doi:10.1016/S0168-9525(98)01494-2

Dorrell, R., Gile, G., McCallum, G., Méheust, R., Bapteste, E., Klinger, C., … Bowler, C. 2017. Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. eLife, 6:e23717. doi:10.7554/eLife.23717

Dorrell, R.G., & Smith, A.G.. 2011. Do red and green make brown?: perspectives on plastid acquisitions within chromalveolates. Euk. Cell. 10:856-868. doi:10.1128/EC.00326-10

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. doi:10.1078/0932-4739-00901

Dyková, I., Fiala, I. & Pecková, H. 2008. Neoparamoeba spp. and their eukaryotic endosymbionts similar to Perkinsela amoebae (Hollande, 1980): Coevolution demonstrated by SSU rRNA gene phylogenies. Eur. J. Protistol., 44:269-277. doi:10.1016/j.ejop.2008.01.004

Dyková, I., Figueras, A. & Peric, Z. 2000. Neoparamoeba page, 1987: Light and electron microscopic observations on six strains of different origin. Dis. Aquat. Org., 43:217-223. doi:10.3354/dao043217

Dyková, I., Nowak, B. F., Crosbie, P. B. B., Fiala, I., Pecková, H., Adams, M. B., Macháckovà, B. & Dvoráková, H. 2005. Neoparamoeba branchiphila n. sp., and related species of the genus Neoparamoeba page, 1987: Morphological and molecular characterization of selected strains. J. Fish Dis., 28:49-64. doi:10.1111/j.1365-2761.2004.00600.x

Eisen, J. 2000. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. Curr. Opin. Genet. Dev., 10:606–611. doi:10.1016/S0959-437X(00)00143-X

Emanuelsson, O., Nielsen, H., Brunak, S., & von Heijne, G. 2000. Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. J. Mol. Biol., 300:1005–1016. doi:10.1006/jmbi.2000.3903

Eme, L., Sharpe, S., Brown, M., Roger, A., Eme, L., Sharpe, S., … Roger, A. 2014. On the Age of Eukaryotes: Evaluating Evidence from Fossils and Molecular Clocks. Cold Spring Harb. Perspect. Biol., doi:10.1101/cshperspect.a016139

Falkowski, P., Katz, M., Knoll, A., Quigg, A., Raven, J., Schofield, O., & Taylor, F. J. R. 2004. The Evolution of Modern Eukaryotic Phytoplankton. Science 305:354–360. doi:10.1126/science.1095964

Feehan, C. J., Johnson-Mackinnon, J., Scheibling, R. E., Lauzon-Guay, J. & Simpson, A. G. B. 2013. Validating the identity of Paramoeba invadens, the causative agent of recurrent mass mortality of sea urchins in Nova Scotia, Canada. Dis. Aquat. Org., 103:209-227. doi:10.3354/dao02577

Fiala, I. & Dykova, I. 2003. Molecular characterization of Neoparamoeba strains isolated from gills of Scophthalmus maximus. Dis. Aquat. Org., 55:11-16. doi:10.3354/dao055011

Finn, R., Attwood, T., Babbitt, P., Bateman, A., Bork, P., Bridge, A., … Mitchell, A. 2017. InterPro in 2017— beyond protein family and domain annotations. Nucleic Acids Res., 45:D190–D199. doi:10.1093/nar/gkw1107

Gornik, S., Febrimarsa, Cassin, A., MacRae, J., Ramaprasad, A., Rchiad, Z., … Waller, R. 2015. Endosymbiosis undone by stepwise elimination of the plastid in a parasitic dinoflagellate. Proc. Nat. Sci., 112:5767–5772. doi:10.1073/pnas.1423400112

Gray, M., Lang, F., Cedergren, R., Golding, B., Lemieux, C., Sankoff, D., … Burger, G. 1998. Genome structure and gene content in protist mitochondrial DNAs. Nucleic Acids Res., 4:865–878. doi:10.1093/nar/26.4.865

Gray, M. 2012. Mitochondrial Evolution. Cold Spring Harb. Perspect. Biol., 9:a011403. doi:10.1101/cshperspect.a011403

Gray, M. 2015. Mosaic nature of the mitochondrial proteome: Implications for the origin and evolution of mitochondria. Proc. R. Soc. B, 33:10133–10138. doi:10.1073/pnas.1421379112

Grzebyk, D., Schofield, O., Vetriani, C., Falkowski, P., Grzebyk, D., Schofield, O., … Falkowski, P. 2003. The Mesozoic radiation of eukaryotic algae: the portable plastid hypothesis. J. Phycol., 39:259-267. doi:10.1046/j.1529-8817.2003.02082.x

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst. Biol. 59: 307–321. doi:10.1093/sysbio/syq010

Hackett, J., Maranda, L., Yoon, H., & Bhattacharya, D. 2003. Phylogenetic evidence for the cryptophyte origin of the plastid of dinophysis (Dinophysiales, Dinophyceae). J. Phycol., doi:10.1046/j.1529-8817.2003.02100.x

Hehenberger, E., Imanian, B., Burki, F., & Keeling, P. 2014. Evidence for the Retention of Two Evolutionary Distinct Plastids in Dinoflagellates with Diatom Endosymbionts. Genome Biol. Evol., 9:2321–2334. doi:10.1093/gbe/evu182

Hill, D.R.A. 1991. Diversity of heterotrophic cryptomonads. In the Biology of Free-Living Heterotrophic Flagellates. pp. 235-240. Clarendon Press, Oxford.

Hollande, A. 1940. Le "Nebenkörper" de certains Cryptomonas et la critique du cycle de Paramoeba eilhardi selon Schaudinn. Bull. Soc. Zool. Fr., 65:211-216.

Hollande, A. 1980. Identification du parasome (nebenkern) de Janickina pigmentifera à un symbionte (Perkinsella amoebae nov. gen. – nov. sp.) apparenté aux flagellés kinetoplastidiés. Protistologica, 16:613-625.

Howe, C., Barbrook, A., Nisbet, R. E., Lockhart, P, Larkum, A. W., Howe, C., … Larkum, A. W. 2008. The origin of plastids. Philos Trans R Soc Lond B Biol Sci., 363:2675–2685. doi:10.1098/rstb.2008.0050

Huang, C., Ayliffe, M., & Timmis, J. 2003. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature, 422:72–76. doi:10.1038/nature01435

Huang, C., Ayliffe, M., Timmis, J., Huang, C., Ayliffe, M., & Timmis, J. 2004. Simple and complex nuclear loci created by newly transferred chloroplast DNA in tobacco. Proc. Natl. Acad. Sci. U.S.A., doi:10.1073/pnas.0400853101

Huang, J., Mullapudi, N., Lancto, C. A., Scott, M., Abrahamsen, M.S., & Kissinger, J.C. Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*. Genome. Biol. 5:R88. doi:10.1186/gb-2004-5-11-r88

Johnson, M., Oldach, D., Delwiche, C., & Stoecker, D. 2007. Retention of transcriptionally active cryptophyte nuclei by the ciliate *Myrionecta rubra*. Nature, 7126:426–428. doi:10.1038/nature05496

Kang, S., Tice, A. K., Spiegel, F. W., Silberman, J. D., Pánek, T., Čepička I., . . . Brown, M. W. 2017. Between a pod and a hard test: the deep evolution of amoebae. Mol. Biol. Evol., doi:10.1093/molbev/msx162

Karnkowska, A., Vacek, V., Zubáčová, Z., Treitli, S., Petrželková, R., Eme, L., … Hampl, V. 2016. A Eukaryote without a Mitochondrial Organelle. Curr. Biol., 10:1274–1284. doi:10.1016/j.cub.2016.03.053

Katoh, K. & Standley, D. M. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol., 30:772-780. doi:10.1093/molbev/mst010

Katz, L.A., & Grant, J.R. 2014. Taxon-rich phylogenomic analyses resolve the eukaryotic tree of life and reveal the power of subsampling by sites. Syst. Biol., 3:406-415. doi: 10.1093/sysbio/syu126

Keeling, P.J. 2010. The endosymbiotic origin, diversification and fate of plastids. Phil. Trans. R. Soc. B, 1541:729-748. doi: 10.1098/rstb.2009.0103

Keeling, P., Burki, F., Wilcox, H., Allam, B., Allen, E., Amaral-Zettler, L., … Worden, A. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. PLoS Biology., 12:e1001889. doi:10.1371/journal.pbio.1001889

Keeling, P. 2013. The Number, Speed, and Impact of Plastid Endosymbioses in Eukaryotic Evolution. Plant Biol., 64:583–607. doi:10.1146/annurev-arplant-050312-120144

Kim, E., & Archibald, J.M.. 2013. Ultrastructure and Molecular Phylogeny of the Cryptomonad *Goniomonas avonlea* sp. nov. Protist, 2:160–182. doi:10.1016/j.protis.2012.10.002

Kim, E., & Maruyama, S. 2014. A contemplation on the secondary origin of green algal and plant plastids. Acta. Soc.Bot. Pol., 83:331–336. doi:10.5586/asbp.2014.040

Koski, L., & Golding, B. 2001. The Closest BLAST Hit Is Often Not the Nearest Neighbor. J. Mol. Evol., 52:540–542. doi:10.1007/s002390010184

Kudryavtsev, A., Pawlowski, J. & Hausmann, K. 2011. Description of Paramoeba atlantica n. sp. (amoebozoa, dactylopodida) - a marine amoeba from the eastern atlantic, with emendation of the dactylopodid families. Acta Protozool., 50:239-253. doi:10.4467/16890027AP.11.023.0023

Lane, C. E., Khan, H., MacKinnon, M., Fong, A., Theophilou, S. & Archibald, J. M. 2006. Insight into the diversity and evolution of the cryptomonad nucleomorph genome. Mol. Biol. Evol., 23:856-865. doi:10.1093/molbev/msj066

Lane, C. E., & Archibald, J. M. 2008. The eukaryotic tree of life: endosymbiosis takes its TOL. Trends Ecol. Evolut., 5:268–75. doi:10.1016/j.tree.2008.02.004

Lartillot, N., Lepage, T. & Blanquart, S. 2009. PhyloBayes 3: A bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics, 25:2286-2288. doi:10.1093/bioinformatics/btp368

Lee, L. E. J., Van Es, S. J., Walsh, S. K., Rainnie, D. J., Donay, N., Summerfield, R. & Cawthorn, R. J. 2006. High yield and rapid growth of Neoparamoeba pemaquidensis in co-culture with a rainbow trout gill-derived cell line RTgill-W1. J. Fish Dis., 29:467-480. doi:10.1111/j.1365-2761.2006.00740.x

Legendre, P., Desdevises Y. & Bazin, E. 2002. A Statistical Test for Host-Parasite Coevolution. Syst. Biol., 51:216-234. doi:10.1080/10635150252899734

Leigh, J., Susko, E., Baumgartner, M., & Roger, A. 2008. Testing Congruence in Phylogenomic Analysis. Syst. Biol., 57:104–115. doi:10.1080/10635150801910436

Leister, D. 2005. Origin, evolution and genetic effects of nuclear insertions of organelle DNA. Trends Genet., 12:655–663. doi:10.1016/j.tig.2005.09.004

Lin, S., Cheng, S., Song, B., Zhong, X., Lin, X., Li, W., … Morse, D. 2015. The Symbiodinium kawagutii genome illuminates dinoflagellate gene expression and coral symbiosis. Science, 6261:691-694. doi:10.1126/science.aad0408

Lin, X., Kaul, S., Rounsley, S., Shea, T., Benito, M.-I., Town, C., … Venter, C. 1999. Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana. Nature, 402:761–768. doi:10.1038/45471

López-García, P., Eme, L., & Moreira, D. 2017. Symbiosis in eukaryotic evolution. J. Theoret. Biol., 17:30095-30104. doi:10.1016/j.jtbi.2017.02.031

Martin, W. 2003. Gene transfer from organelles to the nucleus: Frequent and in big chunks. Proc. Natl. Acad. Sci., 15:8612–8614. doi:10.1073/pnas.1633606100

Matsumoto, T., Shinozaki, F., Chikuni, T., Yabuki, A., Takishita, K., Kawachi, M., … Inagaki, Y. 2011. Green-colored plastids in the dinoflagellate genus *Lepidodinium* are of core chlorophyte origin. Protist, 2:268–76. doi:10.1016/j.protis.2010.07.001

McCutcheon, J., & Moran, N. 2011. Extreme genome reduction in symbiotic bacteria. Nature Rev. Microbiol., 10:13–26. doi:10.1038/nrmicro2670

McFadden, G. 2001. Primary and secondary endosymbiosis and the origin of plastids. J. Phycol., 6:951-959. doi:10.1046/j.1529-8817.2001.01126.x

Medlin L., Elwood H. J., Stickel S. & Sogin M. L. 1988. The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. Gene, 71:491–499. doi:10.1016/0378-1119(88)90066-2

Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. 2013. Ultrafast Approximation for Phylogenetic Bootstrap. Mol. Biol. Evol., 30:1188-1195. doi: 10.1093/molbev/mst024

Moreira, D., & Deschamps, P. 2014. What Was the Real Contribution of Endosymbionts to the Eukaryotic Nucleus? Insights from Photosynthetic Eukaryotes. Cold Spring Harb. Perspect. Biol., 7:a016014. doi:10.1101/cshperspect.a016014

Moustafa, A., Beszteri, B., Maier, U., Bowler, C., Valentin, K., & Bhattacharya, D. 2009. Genomic Footprints of a Cryptic Plastid Endosymbiosis in Diatoms. Science, 324:1724–1726. doi:10.1126/science.1172983

Mouton, A., Crosbie, P., Cadoret, K. & Nowak, B. 2014. First record of amoebic gill disease caused by Neoparamoeba perurans in South Africa. J. Fish Dis., 37:407-409. doi:10.1111/jfd.12133

Muñoz-Gómez, S., Mejía-Franco, F., Durnin, K., Colp, M., Grisdale, C., Archibald, J., … Slamovits, C. 2017. The New Red Algal Subphylum Proteorhodophytina Comprises the Largest and Most Divergent Plastid Genomes Known. Curr. Biol., 27:1677-1684. doi:10.1016/j.cub.2017.04.054

Nei M. 1987. Molecular Evolutionary Genetics. Columbia University Press, New York. p. 254- 266.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Mol. Biol. Evol., 32:268-274. doi:10.1093/molbev/msu300

Nowack, E., Melkonian, M., & Glöckner, G. 2008. Chromatophore Genome Sequence of Paulinella Sheds Light on Acquisition of Photosynthesis by Eukaryotes. Curr. Biol., 6:410–418. doi:10.1016/j.cub.2008.02.051

Okamoto, N., & Inouye, I. 2006. *Hatena arenicola* gen. et sp. nov., a Katablepharid Undergoing Probable Plastid Acquisition. Protist, 4:401–419. doi:10.1016/j.protis.2006.05.011

Orlowski J., & Grinstein, S. 2004. Diversity of the mammalian sodium/proton exchanger SLC9 gene family. Eur. J. Physiol. 447:549-565. doi:10.1007/s00424-003-1110-3

Page, F. C. 1987. The classification of "naked" amoebae (Phylum Rhizopoda). Archiv für Protistenkunde, 133:199-217. doi:10.1016/S0003-9365(87)80053-2

Parfrey, L., Grant, J., Tekle, Y., Lasek-Nesselquist, E., Morrison, H., Sogin, M., … Katz, L. 2010. Broadly Sampled Multigene Analyses Yield a Well-Resolved Eukaryotic Tree of Life. Syst. Biol., 59:518–533. doi:10.1093/sysbio/syq037

Parkinson, J., & Blaxter, M. 2009. Methods in Molecular Biology. Methods in molecular biology. Springer (Clifton, N.J.) (533:1–12). doi:10.1007/978-1-60327-136-3_1

Perkins, F. O. & Castagna, M. 1971. Ultrastructure of the Nebenköper or "secondary" nucleus of the parasitic amoeba Paramoeba perniciosa (Amoebida, Parmoebidae). J. Invertebr. Pathol., 17-186-193.

Petersen, J., Ludewig, A.-K., Michael, V., Bunk, B., Jarek, M., Baurain, D., & Brinkmann, H. 2014. Chromera velia, endosymbioses and the rhodoplex hypothesis – plastid evolution in cryptophytes, alveolates, stramenopiles, and haptophytes (CASH lineages). Genome Biol. Evol., 3:666–684. doi: 10.1093/gbe/evu043

Philippe, H., de Vienne, D., Ranwez, V., Roure, B., Baurain, D., & Delsuc, F. 2017. Pitfalls in supermatrix phylogenomics. Eur. J. Taxon., 283:1-25. doi:10.5852/ejt.2017.283

Philippe, H., Brinkmann, H., Lavrov, D., Littlewood, T., Manuel, M., Wörheide, G., & Baurain, D. 2011. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLoS Biol., 9:e1000602. doi:10.1371/journal.pbio.1000602

Philippe, H., Delsuc, F., Brinkmann, H., & Lartillot, N. 2005. PHYLOGENOMICS. Annu. Rev. Ecol. Evol. Syst. 36:541–562. doi:10.1146/annurev.ecolsys.35.112202.130205

Price, D., Chan, C., Yoon, H., Yang, E., Qiu, H., Weber, A., … Bhattacharya, D. 2012. *Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. Science. 335:843-847. doi:10.1126/science.1213561

Price , M. N., Dehal , P. S. & Arkin, A. P. 2009. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. Mol. Biol. Evol., 26:1641-1650. doi:10.1093/molbev/msp077

Reyes-Prieto, A., Moustafa, A., & Bhattacharya, D. 2008. Multiple Genes of Apparent Algal Origin Suggest Ciliates May Once Have Been Photosynthetic. Curr. Biol., 13:956–962. doi:10.1016/j.cub.2008.05.042

Richly, E., Leister, D., Richly, E., & Leister, D. 2004. NUMTs in Sequenced Eukaryotic Genomes. Mol. Biol. Evol., doi:10.1093/molbev/msh110

Rodríguez-Ezpeleta, N., Brinkmann, H., Burey, S., Roure, B., Burger, G., Löffelhardt, W., … Lang, F. 2005. Monophyly of Primary Photosynthetic Eukaryotes: Green Plants, Red Algae, and Glaucophytes. Curr. Biol., 15:1325–1330. doi:10.1016/j.cub.2005.06.040

Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, F., & Philippe, H. 2007. Detecting and Overcoming Systematic Errors in Genome-Scale Phylogenies. Syst. Biol., 56:389–399. doi:10.1080/10635150701397643

Roth, M. 2014. The engine of the reef: photobiology of the coral–algal symbiosis. Front. Microbiol., 5:1-22. doi:10.3389/fmicb.2014.00422

Sanchez-Puerta, V., & Delwiche, C. 2008. A hypothesis for plastid evolution in chromalveolates. J. Phycol., 5:1097-1107. doi:10.1111/j.1529-8817.2008.00559.x

Sanchez-Puerta, V., Lippmeier, C., Apt, K., & Delwiche, C. 2007. Plastid Genes in a Non-Photosynthetic Dinoflagellate. Protist, 1:105–117. doi:10.1016/j.protis.2006.09.004

Selosse, M.-A., Albert, B., Godelle, B., Selosse, M.-A., Albert, B., & Godelle, B. 2001. Reducing the genome size of organelles favours gene transfer to the nucleus. Trends Ecol. Evolut., doi:10.1016/S0169-5347(00)02084-X

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol., 51:492-508. doi:10.1080/10635150290069913

Shi, L., & Theg, S. M. 2013. The chloroplast protein import system: From algae to trees. Biochim. Biophys. Acta, 1833:312-331. doi:10.1016/j.bbamcr.2012.10.002

Shimodaira, H., &, M., Hasegawa. 2001. CONSEL: For assessing the confidence of phylogenetic tree selection. Bioinformatics, 17:1246-1247. doi:10.1093/bioinformatics/17.12.1246

Sibbald, S., & Archibald, J. 2017. More protist genomes needed. Nat. Ecol. Evol., 1:0145. doi:10.1038/s41559-017-0145

Sibbald, S. J., Cenci, U., Colp, M., Eglit, Y., O'Kelly, C. J. and Archibald, J. M. 2017. Diversity and Evolution of *Paramoeba* spp. and their Kinetoplastid Endosymbionts. J. Euk. Microbiol., doi:10.1111/jeu.12394

Smirnov, A. V. 1997. Two new species of marine amoebae: Hartmannella lobifera n. sp. and Korotnevella nivo n. sp. (Lobosea, Gymnamoebida). Archiv für Protistenkunde, 147:283-292. doi: 10.1016/S0003-9365(97)80055-3

Smith, D., & Keeling, P. 2015. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. Proc. Nat. Acad. Sci., 112:10177–10184. doi:10.1073/pnas.1422049112

Sprague, V., Beckett, R. L., & Sawyer, T. K. 1969. A new species of Paramoeba (Amoebida, Paramoebidae) parastitic in the crab Callinectes sapidus. J. Invertebr. Pathol., 14:167-174. doi:10.1016/0022-2011(69)90103-7

Stegemann, S., Hartmann, S., Ruf, S., & Bock, R. 2003. High-frequency gene transfer from the chloroplast genome to the nucleus. Proc. Nat. Acad. Sci., 100:8828–8833. doi:10.1073/pnas.1430924100

Stiller, J., Huang, J., Ding, Q., Tian, J., & Goodwillie, C. 2009. Are algal genes in nonphotosynthetic protists evidence of historical plastid endosymbioses? BMC Genom., 10:1–16. doi:10.1186/1471-2164-10-484

Stiller, J. W. 2014. Toward an empirical framework for interpreting plastid evolution. J. Phycol., 50:462–471. doi:10.1111/jpy.12178

Stiller, J. W., Schreiber, J., Yue, J., Guo, H., Ding, Q., & Huang, J. 2014. The evolution of photosynthesis in chromist algae through serial endosymbioses. Nat. Commun., 5:5764. doi:10.1038/ncomms6764

Sun, G., Yang, Z., Ishwar, A., & Huang, J. 2010. Algal Genes in the Closest Relatives of Animals. Mol. Biol. Evol., 27:2879–2889. doi:10.1093/molbev/msq175

Tanifuji, G., Kim, E., Onodera, N. T., Gibeault, R., Dlutek, M., Cawthorn, R. J., Fiala, I., Lukeš, J., Greenwood, S. J. & Archibald, J. M. 2011. Genomic characterization of *Neoparamoeba pemaquidensis* (Amoebozoa) and its kinetoplastid endosymbiont. Euk. Cell, 10:1143-1146. doi:10.1128/EC.05027-11

Tengs, T., Dahlberg, O., Shalchian-Tabrizi, K., Klaveness, D., Rudi, K., Delwiche, C., … Jakobsen, K. 2000. Phylogenetic Analyses Indicate that the 19′Hexanoyloxy-fucoxanthin-Containing Dinoflagellates Have Tertiary Plastids of Haptophyte Origin. Mol. Biol. Evol., doi:10.1093/oxfordjournals.molbev.a026350

Timmis, J. N., Ayliffe, M. A., Huang, C. Y. & Martin, W. 2004. Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet., 5:123-135. doi:10.1038/nrg1271

Timmis, J. 2012. Endosymbiotic Evolution: RNA Intermediates in Endosymbiotic Gene Transfer. Curr. Biol., 22:R296–R298. doi:10.1016/j.cub.2012.03.043

Waller, R.F. & McFadden, G.I. 2005. The apicoplast: a review of the derived plastid of apicomplexan parasites. Curr. Issues Mol. Biol, 1:56-79.

Wang, H.C., Susko, S, Minh B.Q & Roger A.J. 2017. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *In Press.*

Wang, Z., & Wu, M. 2015. An integrated phylogenomic approach toward pinpointing the origin of mitochondria. Scientific Reports, *1*:7949. doi:10.1038/srep07949

Woehle, C., Dagan, T., Martin, W., & Gould, S. 2011. Red and Problematic Green Phylogenetic Signals among Thousands of Nuclear Genes from the Photosynthetic and Apicomplexa-Related *Chromera velia*. Gen. Biol. Evol., 3:1220–1230. doi:10.1093/gbe/evr100

Yabuki, A., Inagaki, Y., Ishida, K., Yabuki, A., Inagaki, Y., & Ishida, K. 2010. *Palpitomonas bilix* gen. et sp. nov.: A Novel Deep-branching Heterotroph Possibly Related to Archaeplastida or Hacrobia. Protist. 161:523-538. doi:10.1016/j.protis.2010.03.001

Young, N. D., Crosbie, P. B., Adams, M. B., Nowak, B. F. & Morrison, R. N. 2007. Neoparamoeba perurans n. sp., an agent of amoebic gill disease of Atlantic salmon (Salmo salar). Int. J. Parasitol., 37:1469–1481. doi:10.1016/j.ijpara.2007.04.018

Young, N. D., Dyková, I., Crosbie, P. B., Wolf, M., Morrison, R. N., Bridle, A. R. & Nowak, B. F. 2014. Support for the coevolution of Neoparamoeba and their endosymbionts, Perkinsela amoebae-like organisms. Eur. J. Protistol., 50:509-523. doi:10.1016/j.ejop.2014.07.004

Young, N. D., Dyková, I., Snekvik, K., Nowak, B. F. & Morrison, R. N. 2008. Neoparamoeba perurans is a cosmopolitan aetiological agent of amoebic gill disease. Dis. Aquat. Org., 78:217-223. doi:10.3354/dao01869

Zhu, G., Marchewka, M., & Keithly, J. 2000. *Cryptosporidium parvum* appears to lack a plastid genome. Microbiology, 146:315–321. doi:10.1099/00221287-146-2-315