Assessing the Monophyly of Red Algae and Green Plants Via Conserved Core Informational Genes

Justin Perry

Department of Biology

June 2015

Director of Thesis: Dr. John W. Stiller

Major: Molecular Biology & Biotechnology

For well over a century the existence of a monophyletic relationship between red algae and green plants has been debated. Many scholars have sought to address this issue, however, a consistent solution to the problem has not been found. Addressing a monophyletic relationship of red algae and green plants is important for understanding early eukaryotic evolution. Elucidating this relationship will allow for a more in depth evaluation of the origin and spread of photosynthesis in eukaryotes, and will further develop an understanding of the evolution of primary producers, which are of paramount importance in supporting the earth's ecosystems.

The goal of this project is to apply a method that provides an accurate and consistent way to classify more ancient phylogenetic relationships. Although a great deal of work has been done in the past on this question, the need for a more consistent method that is minimally affected by phylogenetic artifacts has never been greater. This is because of the rapid increase in the amount of available sequence data, as well as the number of new taxa that are being sequenced. By providing a more accurate methodology for investigating broad scale relationships we hope to ameliorate some of the issues seen previously in evaluations of deep phylogenetic relationships.

The first goal of this project was to develop a set of core conserved genes related to information processing in cells that span the broad range of eukaryotic life to circumvent known

issues from previous studies where selection of markers was problematic. These genes perform highly conserved functions in the cell and, therefore, are less likely to be negatively influenced by problems that create phylogenetic discontinuities. For example, all living organisms must transcribe and translate their genes into proteins. As such, the transcriptional and transitional machinery required to accomplish this task is highly conserved across all forms of life. Although they are responsible for functioning of the central dogma of molecular biology, this research shows that universal conservation of many of these genes across the broad range of eukaryotic life is uncertain. Thorough analyses of 47 conserved genes indicated that the most reliable markers for ancient phylogenetic inferences are core subunits of DNA-dependent RNA polymerases.

Genes encoding the two largest subunits of each of three eukaryotic RNA polymerases were recovered from a list of organisms that span eukaryotic diversity via BLAST searches of two major bioinformatics databases National Center for Biotechnology Information (NCBI) and the Department of Energy's Joint Genomics Institute (JGI). The sequences were aligned using multiple sequence alignment software packages, edited by hand, and then used as input into phylogenetic analysis programs. The resulting alignments recovered a polyphyletic relationship among red algae and green plants. Statistical analyses were applied to each tree, allowing for a clear determination that polyphyly was strongly supported by these data. The further hope is that this project will provide a method that is useful, not only for addressing red/green monophyletic issues, but also for future problematic phylogenies.

Assessing the Monophyly of Red Algae and Green Plants Via Core Conserved Informational Genes

A Thesis

Presented To the Faculty of the Department of Biology

East Carolina University

In Partial Fulfillment of the Requirements for the Degree of Master of Science, Molecular Biology and Biotechnology

By

Justin Perry

June, 2015

ASSESSING THE MONOPHYLY OF RED ALGAE AND GREEN PLANTS VIA CONSERVED CORE INFORMATIONAL GENES

by

Justin Perry

APPROVED BY:

DIRECTOR OF THESIS:_____
John W. Stiller, PhD

COMMITTEE MEMBER:_____
Michael S. Brewer II, PhD

COMMITTEE MEMBER:_____
Timothy W. Christensen, PhD

COMMITTEE MEMBER:_____
Adam M. Reitzel, PhD

CHAIR OF THE DEPARTMENT OF Biology:

_____
Jeffrey S. McKinnon , PhD

DEAN OF THE GRADUATE SCHOOL:

_____
Paul J. Gemperline, PhD

Acknowledgements

I would like to thank my wife and my family for their support and encouragement. Without them none of this would have been possible. I would also like to thank Dr. John W. Stiller for the critical role he has played in shaping my education and development as a scientist. Dr. Stiller has consistently gone above and beyond what anyone would expect from a thesis advisor and words cannot express the gratitude I have for all he has done.

Table of Contents

Chapter 1

Introduction

For well over a century the place of red algae on the tree of life has been highly contested (see Ragan and Gutell, 2005 for review). Red algal relationships with green plants, as well as other organisms, are at the heart of many questions related to early eukaryotic evolution. As the cost of genomic sequencing has decreased and the technology for obtaining those sequences continues to improve, the amount of sequence data available for comparison has exploded. Yet, as exciting as it is to be in an era of so much rapid growth and expansion, extreme caution must be taken with massive data sets to ensure that conclusions drawn are accurate. With more than 3,500 sequences added to the National Center for Biotechnology Information's (NCBI) sequence database each day (Benson et al., 2015), and over 1 trillion total bases of sequence data publically available (NCBI, 2015), it has become commonplace to think that more data are the answer to more robust analyses. However, merely adding more sequence data is not enough to resolve the inconsistencies seen among phylogenetic trees (Philippe et al., 2011). Consequently, a better understanding of how to effectively assess organismal relationships throughout the tree of life has never been more paramount.

Problematic phylogenetic relationships permeate the tree of life (Baldauf, 2003; Dunn et al., 2008; Halanych, 2004; Moreira et al., 2000; Philippe and Laurent, 1998; Philippe et al., 2011; Ragan and Gutell, 1995; Stiller and Hall, 1997; Williams et al., 2012) and, therefore, require increased attention to ensure that accurate and precise conclusions can be drawn from phylogenetic trees that are generated. Because the relationship is heavily debated, a consistent and accurate way to assess the monophyly of red algae and green plants will be a valuable step in

evaluating many other problematic phylogenetic relationships. Moreover, the approach presented in this thesis has the potential to provide more definitive answers to other highly contested questions about early eukaryotic evolution. The potential impact outside of the field of phylogenomics may not be initially obvious; however, a better understanding of the issues at the heart of early eukaryotic evolution can have profound and unknown impacts on future research in a multitude of disciplines.

History

Historically a list of scholars as far back as Pliny the Elder (approx. date) have found evidence placing red algae in the Kingdom Plantae (Ragan and Gutell, 1995). Even hundreds of years ago, however, there were those who did not agree that a close relationship between red algae and green plants was clearly defined. Wilson and Cassin said: "… there are numerous expressions in the works of naturalists of all times, which show a suspicion that organisms exist which are not to be regarded as either animal or vegetable in their structure or nature" (Ragan and Gutell, 1995). This example serves to illustrate the point that even very early in the development of taxonomic ideas, there was often a minority of scholars who did not agree with the popular opinion. Although placing red algae within the kingdom Plantae appeased most nineteenth and twentieth century scholars, it did not satisfy H.F. Copeland. In 1938 Copeland classified red algae into Kingdom Protista. While ribosomal RNA (rRNA) sequence data favored Copeland's reclassification (Ragan and Gutell, 1995), this support did not the end the debate. Initially, morphological and cytological characteristics were used to classify red algae and other organisms (Lipscomb, 1985), as well as in phylogenetic analyses to determine their relationships. Cytological characters, interestingly, placed red algae at the base of the eukaryotic tree (Lipscomb, 1985) adding further question about where red algae actually belong on the tree-of-

life. Although this type of characterization seemed to work well for quite some time, as the use of molecular characters began to increase, many of the ideas previously supported by morphological and cytological data began to break down.

There are a plethora of scholarly articles relating to the history and development of organismal classification, and each has its own interpretation of the most accurate and appropriate methods for classifying organisms on the tree of life. Of the various papers in the last three decades, the work of Carl Woese and colleagues stands out as perhaps having the largest impact on the field of systematics. This work illustrates how new ideas in systematics can have profound implications for the field, and provides several key ideas that had much broader implications for biology in general (Woese et al., 1990).

Early Use of Molecular Data

Among the arguments made for the restructuring systematic classification are several important ideas that should be mentioned in relation to the debate surrounding red/green monophyly. Most important among these is the idea that molecular sequences, rather than phenotypic characteristics, typically provide more reliable information when dealing with evolutionary relationships (Woese, Kandler, and Wheelis, 1990). Although accepting this as true seems almost second nature in 21st century biology, this was not always the case. The failings of morphological classification to properly evaluate organismal relationships called for the issue to be readdressed. Furthermore, prior to the 1970s classifications were largely limited to metazoan and metaphyta (Woese, Kandler, and Wheelis, 1990). This problem appeared to be resolved for the most part with the revolution in sequencing technology. Advancements in sequencing have

made it possible to trace the evolutionary history back to the most recent common ancestor of all cells (Woese, Kandler, and Wheelis, 1990).

These two ideas, the importance of using molecular data and the impact it can have on evaluating evolutionary relationships, influenced all subsequent investigations of broad scale eukaryotic systematics. They also provide the framework for this thesis project. Additionally, it is important to follow up the earlier reference to Copeland's contributions to the reclassification of red algae, specifically, that they are a "highly evolved group of unknown origin" (Copeland, 1938). While this statement is interesting in itself, it is important to think of it in light of Woese's argument regarding the new possibilities of addressing such issues through the use of molecular data. Red algae provide an interesting test of how well modern molecular approaches answer phylogenetic problems that previously seemed intractable.

Although the previous discussion in no way does justice to the complex issues involved in the history of red algal systematics, it does provide a framework for understanding the goals of this research. It is with this background we can move forward to discussing issues that have arisen in red/green relationships using molecular characters. Just as in early systematic debates, the use of various molecular data to build phylogenetic trees has had a turbulent past. The dawn of the sequencing age, however, brought a great change in the way phylogenies were constructed. In the late 1980's the first red algal phylogenies were built using 5s ribosomal RNA; however, in relatively short order 5s rRNA trees were found to be unreliable because of the lack of informative sites and the highly constrained nature of the molecules (Halanych, 1991; Steele et al., 1991). 5s rRNA phylogenies gave way to the use of nuclear-encoded small subunit rRNA genes (ssu rDNA), which are highly conserved among organisms as an essential component of protein synthesis. By the end of the 1980s, eukaryotic ssu rDNA trees were well defined and,

while some issues remain with respect to the resolution of eukaryotic relationships (Ragan and Gutell, 1995), a new era of tree building was born. ssu rDNA proved very useful in early tree building because it was larger in size, had more information-rich characters, and was easy to amplify via the polymerase chain reaction (Sogin, 1990). This is not to say that ssu rDNA data were necessarily the best characters for building phylogenetic trees, but their use represented a major paradigm shift from the previous systematic approaches. Using ssu rDNA appeared to be a reliable way of building trees; however, as sequencing technology and molecular biology have developed, so too have the methods used for evaluating phylogenetic relationships.

The impact that initial ssu rRNA trees had cannot be overstated. They provided a launching point for new and updated approaches for dealing with organismal relationships. Research on the origin of red algae has been shaped by several influences throughout its history. More recently, however, the predominate influences have come from evidence from nuclear, mitochondrial and plastid genes (Ragan and Gutell, 1995). Using these sources of evidence provides an excellent place to segue into how these issues relate specifically to this project.

At the time Ragan and Gutell wrote *Are Red Algae Plants?* (1995) these molecular data sources were new and largely untapped. In fact when mentioning each of them, several points were made that show just how new these ideas were, even in 1995. Perhaps one of the best examples is the following: "The non-rRNA molecular biology of red algae is in its infancy, and that of red algal nuclear genes even more so" (Ragan and Gutell, 1995). The importance of this statement is twofold. Primarily it illustrates that a shift away from the sole use of ssu rRNA to build trees. It also provides an example, albeit unintentional, of a changing mindset. This shift in mindset, away from the old way of thinking about classifying organisms based solely on phenotypic traits, to the new incorporation of molecular data, was critical in developing this

field. Even more important was the idea that all of the answers did not lie within ssu rRNA trees. Fortunately, many scholars in the field at this time saw the need to evaluate more data and use additional molecular evidence to build phylogenies. Although algal genomics has come a long way since the early 90's, and many more molecular markers are in use, even as recently as 2007 some suggested that the field is still in its infancy (Grossman, 2007). While this was certainly a time of excitement and discovery in the field of comparative genomics and even systematics, it did not come without its own set of problems.

As is often the case with the development of new procedures in any field, a lot of new data were generated very quickly. These new data also brought an abundance of conflicting new hypotheses. An understanding of these conflicts requires an evaluation of the total evidence from each category of sequence data (nuclear, plastid, mitochondria); as a whole they provide a much more complete picture of the field and its issues than assessing each category individually. To completely understand and interpret data from each of these areas requires at least some input from the others. For example, elucidating the origin of plastids cannot be accomplished without also understanding the impact of subsequent loss and/or reduction of nuclear genes from both host and endosymbiont (Stiller et al., 2003). Similarly, evaluating whether or not there have been one or multiple primary endosymbiotic events requires an evaluation of nuclear genes as well as subsequent life history strategies. This interconnectedness, while critical to evaluating red/green monophyly, also makes it increasingly difficult to uncover a reliable solution. These issues will be addressed below and are critical to understanding the multi-faceted approach required to answer the question of red/green monophyly.

Evidence from Plastid and Mitochondrial Data

There are an abundance of issues that continue to confound the red algae and green plant monophyly debate. Although there have been many papers written to evaluate the issue over the past three decades only a handful will be evaluated here. While these papers represent only a brief overview of the work that has been done, they serve as a representative sample of the research that has contributed to the current knowledge of the field.

It is essential to study the origins of algae because, as primary producers, they are among the most important organisms on the planet (Bhattacharya and Medlin, 1998). Critical to the understanding of these relationships is plastid evolution, more specifically gene reduction and gene transfer as they relate to endosymbiotic events (Bhattacharya and Medlin, 1998). Inferences of genome reduction and gene transfers have played a critical role in evaluating red algae and green plant monophyly. Additionally, it has become essential to ask how deeper evolutionary phylogenetic relationships can be addressed by using the "framework" provided by modern molecular methods (Bhattacharya and Medlin, 1998).

Bhattacharya and Medlin (1998) proposed a novel phylogeny based on 16s plastid ribosomal RNA analysis of simple-plastid containing algae (a host cell), which placed glaucocystophytes, rhodophytes and chlorophytes in a monophyletic group. Despite the strong support they found for this monophyletic relationship, analyses of nuclear small subunit rDNA (Sogin, 1989), and other nuclear genes at the time had produced data that did not support a monophyletic Plantae (Stiller and Hall, 1997), thus highlighting the idea that more needed to be done to accurately resolve the phylogenetic relationship of these clades.

The impact early plastid studies had on how red/green monophyly issues have been evaluated cannot be overstated. Effectively, an understanding of how plastids were acquired and

their subsequent relationships helped to guide the theories that shaped how red/green relations are determined (Bhattacharya and Medlin, 1995, 1998; Delwiche, 1999; Delwiche and Palmer, 1997). Some studies have focused on complete nuclear genome analysis of plastid related proteins to suggest a monophyletic origin of plastids (McFadden and van Dooren, 2004). Others have used plastid genome data to draw conclusions about phylogenetic relationships of the host cells (Rodriguez-Ezpeleta et al., 2005). No matter what type of data are used, or what results obtained, there is one central question that lies at the heart of plastid research: Do plastids and their host cells share the same phylogenetic history?

In a paper evaluating the origin of plastids and the effects of convergent evolution on genome content, Stiller, Reel and Johnson (2003) presented several key ideas that had an impact on the field. They provided an interesting example of a problem that has yet to be addressed in this discussion, convergent evolution of plastids. Stiller and colleagues discussed the varying hypotheses surrounding the origin of primary plastids, and the issues that arise when dealing with how the "primary" plastid lineages (green plants, red algae, and Glaucocystophytes) obtained their plastids (see Delwiche and Palmer, 1997 and Delwiche 1999 for reviews).

It has now been commonly accepted that plastids were obtained via a single endosymbiotic event (Cavalier-Smith, 2000; Palmer 2000), however the idea is not without challenges. For example, one of the major lines of evidence for establishing this theory originally was the high similarity in genome content shared among all plastids. Through the course of evolution approximately 90% of the original cyanobacterial genome has been lost either through transport to the nucleus, or simply because the gene wasn't required to maintain the endosymbiont once it was engulfed (Martin et al., 1998). The remaining conserved regions of the plastid genome has been shown through multiple analyses to have strong similarity across other

photosynthetic species (Kowallik, 1994). While this evidence was initially used to support a single plastid origin (Palmer 1993, Kowallik 1994) it was found that not all genes are easily transferred to the host nucleus (Race et al., 1999). This led to an increased focus on the impacts of convergent evolution on plastid genome content, as well as how to assess the difference between selective genome reduction versus random gene loss. It became clear that many of the genes that remain after the reduction of endosymbiont's genomes are related to the core function of that organelle, which clearly would be under strong selection for retention (Stiller et al., 2003).

Based on this observation, two groups of genes (tRNA and ribosomal proteins) were analyzed in all three primary plastids as well as in the relatively unreduced mitochondrion of *Reclinomonas*. The tRNA and ribosomal protein genes were chosen because neither is related to the defining biochemical roles of its organelle, photosynthesis and cellular respiration respectively. The results of ribosomal protein and tRNA analyses provided no evidence for a single plastid origin (Stiller et al., 2003), and suggested that the conclusion made about a single or multiple endosymbiotic events depends largely on the how evidence is interpreted. Furthermore, while both single and multiple plastid origins are in need of further investigation, the manner in which one evaluates plastid origins should also be reconsidered (Howe et al., 2008; Stiller et al., 2003).

The typical depiction of plastid origins, in which a single endosymbiotic event with a cyanobacteria diverged into the three primary plastid lineages (red, green, and glaucocystophyte) can be juxtaposed with the idea of "multiple plastid origins", in which multiple independent events resulted in the primary plastid lineages seen today (Stiller et al., 2003). One would see the same results in looking at the intermediate forms of each of the three primary lineages; however,

each of these forms is extinct and thus no longer able to be characterized. Therefore it has been proposed that there are two equally valid explanations of plastid origins based on available data. The traditional hypothesis for a single establishment of primary plastids, or the other equally acceptable explanation of incorporating multiple independent events and subsequent extinctions, both ultimately are consistent with most results (Stiller et al., 2003).

The mindset that a single plastid origin means a monophyletic relationship between red algae and green plants continues to be an issue in current phylogenetic interpretations. Although this introspective approach was not the key focus of the research presented by Stiller's group, their work became influential by introducing a new way of viewing phylogenetic data, and for its contributions to the origin of plastids debate.

In the same issue of the *Journal of Phycology* that presented the paper above (Stiller et al., 2003), Jeffrey Palmer presented a review article seeking to address the evidence concerning how many endosymbioses had occurred, for both primary and secondary plastids (Palmer, 2003). A variety of research (Delwiche and Palmer, 1997, Bhattacharya and Medlin, 1998, Martin et al., 1998, 2002, McFadden, 2001) investigating these issues came do different conclusions than the study by Stiller, Reel, and Johnson (2003). Yet Palmer concluded that, despite the evidence presented by cumulative research, albeit slightly skewed toward primary plastid monophyly, no definitive answer had been produced regarding the red/green monophyly debate (Palmer, 2003).

Taking the two papers mentioned above as representative of the field, it became clear that the red/green monophyly debate remained unresolved, and more recent publications on a wide range of topics dealing with red algae indicate the question remains open (Chan et al., 2011; Qui et. al., 2015). In an attempt to remedy this situation it was proposed that the genome sequence of

the red alga *Cyanidioschyzon merolae* would help to provide answers about the origin of plastids and even suggested that algal genomes could be the Rosetta stone for understanding protein targeting in secondary plastids (McFadden and van Dooren, 2004). Given that some scholars postulate independent plastid origins based on polyphyly of reds and greens, while others support plastid monophyly based on their observations that endosymbiont and host genes typically unite the primary plastid lineages, there seems to be no clear answer to the question (McFadden and van Dooren, 2004). It also has been suggested that this grouping could be considered a phylogenetic artifact (see Stiller, 2007).

When assessing whether the relationships among plastids are support for relationships among their host cell lineages, several factors have to be taken into consideration; some organisms preferentially adopt organisms as symbionts and modern cyanobacteria are not like their ancestors of more than a billion years ago (Stiller and Hall, 1997). Therefore, interpretations of a single or multiple endosymbiotic origin of plastids often rely more on the assumptions made about relationships than on the actual data (Stiller et al., 2003). Although most recent evolutionary, genetic, and biochemical investigations have been done under the assumption of a single plastid origin, to explain certain results obtained under this conceptual framework several complicated assumptions have to be made (Stiller, 2014). This can result in misleading interpretations. For example, one of the most widely cited early pieces of evidence for a red/green relationship was the recovery of red algae and green plants as a monophyletic group based on mitochondrial sequences. Interestingly, these sequences grouped together only if certain genes and organisms were excluded from analyses (Burger et al., 1999; Stiller et al., 2003). Different sets of mitochondrial genes that should be equally reliable result in polyphyly of

reds and greens, relationships that are more easily explained via multiple plastid origins (Stiller et al., 2003).

After evaluation of mitochondrial genomes proved unreliable in determining plastid origins, researchers looked elsewhere to explain relationships among plastids and photosynthetic organisms. To better address the question of plastid origins, outside the impact of mitochondrial genomes or the other issues previously discussed, a data set of proteins involved in plastid protein translocation machinery were evaluated (McFadden and van Dooren, 2004). The predominant idea behind this study was an understanding of how proteins move across the multiple membranes of secondary plastids would make it easier to understand their origins. Of the translocation machinery investigated, a particularly interesting protein, Tic110, a central component of the inner membrane apparatus, can be identified in green algae, plants, red algae, cryptomonads and diatoms, but not in cyanobacteria (McFadden and van Dooren, 2004). This evidence was used to support the idea of a common origin of plastids because Tic110 is present in both primary red and green plastids but also subsequent secondary plastids in cryptomonads and diatoms.

Although some foundational knowledge about early movement of plastids between taxa has been established (see Stiller, 2014 for review), little more than the most basic assumptions about plastid origins can be made for certain. As was the case with early molecular characters and mitochondrial genes, a consistent solution to the problem of red/green monophyly cannot yet be answered by plastid data alone. Having now established that host plastid phylogenies do not necessarily share the same relationships as their host cells a different approach must be taken to address the issues of red/green monophyly. The predominant choice in recent years has been to tackle these problems by building phylogenies with nuclear genes.

Evidence from Nuclear Genes

In the early 2000s new molecular markers (amino acid sequences) were being placed alongside rDNA data to build phylogenies to test evolutionary relationships with increasing success. Using the 20 character states of amino acid sequences over the 4 states of nucleotide sequences has proven to be more phylogenetically informative and reduces the likelihood of parallel changes occurring by chance. Such was the case when evidence was found for a sister relationship between red algae and green plants via the use of a variety of protein sequences, most notably elongation factor 2 (EF-2), and gene fusion analysis (Moreira et al., 2000). EF-2 was considered a more accurate and reliable molecular marker for elucidating a relationship between red algae and green plants because of its conserved function (Moreira et al, 2000). Previous phylogenies that had been inferred by other research groups, from both mitochondrial and nuclear encoded genes, were argued to have a variety of problems and conflicts, leading to the conclusion that past phylogenies had been inconclusive and prone to contradictory results (Moreira et al, 2000).

While most of the data presented by Moreira and colleagues supported the red/green sister grouping they proposed, they also discussed several issues that became increasingly relevant to subsequent analyses. Most important are the effects of long-branch attraction (LBA) artifacts on the construction of phylogenetic trees. In brief, LBA is a tree-building artifact caused by more rapid evolution in some sequences, which can lead to artificial clustering of the longest branches of a tree, regardless of their actual phylogenetic relationships. In other words, the longest branches of a tree are more likely to group together because of convergent changes, providing inaccurate results. A more thorough treatment of how such artifacts influence phylogenetic tree building will be discussed later.

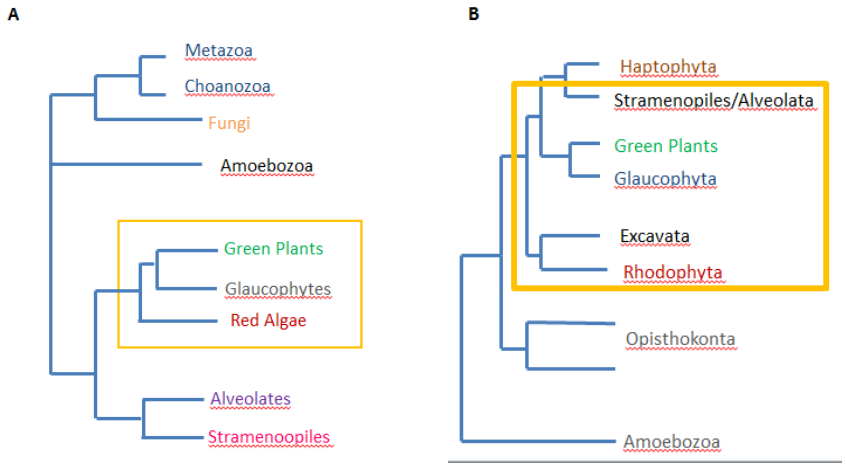
Another point that was highlighted in early protein studies was the importance of ensuring that the data evaluated are of appropriate length and are obtained from a representative sampling of taxa (Moreira et, al., 2000). While not the first to make these points, Moreira and colleagues highlighted the importance of appropriate sequences and taxon sampling when addressing the origin of rhodophytes and chlorophytes. Both of these provisos have played large roles in the development of the methodology for this project.

As promising as phylogenetic analyses of nuclear genes have been, they too have yielded contradictory results. Relationships based on slowly evolving genes would appear to show the most promise for providing solid evidence in one direction or another about red/green monophyly (Nozaki et al., 2007). Nozaki and colleagues took this approach, focusing on reevaluating analyses by Rodriguez-Ezpeleta et al. (2005) of 143 nuclear genes that supported a monophyletic red/green relationship. In addition to these 143 nuclear genes, Rodriguez-Ezpeleta and colleagues sampled taxa primarily from stramenopiles and alveolates, which warranted the addition of more taxa outside these two groups to better illustrate global phylogenetic relationships (Nozaki et al., 2007). Through the use of only 19 of the most slowly evolving genes, and the addition of taxa to balance out sampling, an entirely different result was obtained, indicating polyphyletic origins of red algae and green plants (Nozaki et al., 2007).

The work of Nozaki and colleagues was significant because it used virtually the same data as in a previous study (Rodriguez-Ezpeleta et al., 2005), minus rapidly evolving sequences and with minor but significant adjustments, and provided a completely contradictory result. Among the issues Nozaki and colleagues raised is the potential for rapidly evolving genes commonly used in phylogenetic reconstruction to be subject to long-branch attraction. Figure 1 illustrates the phylogenetic differences resulting from the two comparative approaches. Figure

1A depicts a monophyletic Plantae as obtained by multiple analyses (Baldauf et al. 2000; Moreira et al., 2000; Rodriguez-Ezpeleta et al., 2005; Chan et al., 2011), while Figure 1B presents a polyphyletic red/green association recovered using other data sets and approaches (Stiller and Hall, 1997; Stiller et al., 2001; Nozaki et al. 2003; Stiller and Harrell, 2005; Nozaki et al., 2007). These trees highlight the point that entirely different results can be obtained using similar sets of data, and illustrates that there has been no consistent answer to the red/green monophyly question from nuclear gene analyses.

Figure 1: Competing hypothesis of red algal and green plant origins. (A) depicts a monophyletic red/green relationship, (B) depicts a polyphyletic relationship. Both A and B have been recovered from multiple groups; tree A by (Baldauf et al. 2000; Moreira, Le Guyader, Philippe, 2000; Rodriguez-Ezpeleta et al., 2005; Chan et al., 2011), and tree B by (Stiller and Hall, 1997; Stiller, Riley, and Hall, 2001; Nozaki et al. 2003; Stiller and Harrell, 2005; Nozaki et al., 2007) using a variety of methods and data sets.



The varied data sets discussed thus far have yielded contradictory results. Phylogenies based on plastid rDNA established a monophyletic red/green tree, (Bhattacharya and Medlin, 1995), which was later supported by analyses of plastid protein sequences (see Palmer, 2003 for review). Moreira et al. (2000) found a monophyletic red/green association based on nuclear genes, and highlighted the need to use an appropriate sequence length and number of taxa to avoid issues of LBA. Subsequent studies of large numbers of nuclear-encoded genes provided

further support for red/green monophyly (Baldauf et al. 2000; Moreira, Le Guyader, Philippe, 2000; Rodriguez-Ezpeleta et al., 2005; Chan et al., 2011); however, depending on the data set and methodology a paraphyletic red/green relationship also was recovered, for example from slowly evolving genes (Nozaki et al., 2005, 2007). With this historical context in mind, a discussion of current issues facing the field can begin.

Current ideas in the field

At its inception, molecular phylogenetic research focused primarily on a handful of markers that were thought to be capable of answering tough questions. As time progressed more and more genes were lumped together to create larger and larger data sets. Papers from the late 1990's and early 2000's typically used fewer than 20 genes for analysis (Ragan and Gutell, 1995; Morier, Le Guyader and Philippe, 2000; Stiller, Riley and Hall, 2001), whereas papers since that time have significantly increased the number of genes evaluated (Rodriguez-Ezpeleta et al., 2005, Chan et all, 2011, Williams et al., 2012, Cavalier-Smith, 2014). While it generally is assumed that more genes taken together provide a more correct answer, sometimes more sequences can actually aggravate problems with the data (Phillippe et al., 2011). This begs the question, if the use of more molecular markers doesn't resolve the issue, what approach should be taken?

Appropriate Gene and Taxon Selection

Despite evidence that they can be unreliable (Phillippe et al., 2011), large concatenated data sets continue to be used to address broad scale phylogenetic questions. While the majority of phylogenomic studies continue to use such data, some researchers have explored the use of

fewer, more reliable (resistant to phylogenetic artifacts) genes to provide clearer answers to the tough questions. A reevaluation of the traditionally accepted "three-domain" tree of life (Williams et al, 2012) provided an example of an approach to gene selection that could be most appropriate for addressing the origin of red algae. While that paper dealt largely with the evaluation of the eocyte hypothesis (eukaryotes evolved from the prokaryotic Crenarchaeota a phylum within the Archaea (Lake, 1988)) not red/green monophyly, the criteria Williams and colleagues used for selecting their proteins were the basis of the initial methodology projected for this study. Additionally, they highlighted the value of careful taxon sampling, the impact of phylogenetic artifacts, and the need for careful evaluation of the relationships resulting from any given study because of the possibility of ancient and possibly extinct lineages being absent from the data.

Careful taxon sampling plays a significant role in developing a thorough methodology for evaluating difficult phylogenomic questions. Poor taxon sampling can often lead to tree imbalances, which result in misleading and inaccurate conclusions (Heath et al., 2008). While advances in modern sequencing technology have helped to provide more data from a broader array of species, there are still certain important groups that lack sequence data from key species. This lack of sequence data makes it difficult to address questions of global relatedness in sparsely sampled genera. While little can be done immediately to remedy this issue, as time passes more organisms will be added to genomics databases and help fill in the critical missing data in current studies. In the meantime it is important to choose the most reliable molecular markers and balanced taxon sampling for phylogenetic analysis.

Among the most potentially reliable genes for tree-of-life investigations are single copy orthologous genes involved in "informational" processes, like transcription and translation, as

well as ribosomal RNA (Williams et al., 2012). Of the numerous possibilities, 29 proteins conserved across all three domains, and 64 that are conserved between eukaryotes and Archaea, were identified as meeting these criteria (Williams et al, 2012). This highlights an important point; the number of genes selected or taxa sampled could be secondary to the types of molecular markers chosen. As discussed previously, the use of conserved molecular markers like rDNA, tRNA, ribosomal proteins, cytochrome C oxidase, and DNA-dependent RNA polymerase II subunits (RPB1) have generated a great deal of data, but with conflicting results. The use of 'informational genes" illustrates a shift in thinking about how phylogenetically informative data sets are constructed, and the importance of critically evaluating the markers one selects.

Because they are critical to the processes that sustain an organism's viability, it seems intuitive that genes like EF2 (a translation elongation factor), DNA-dependent RNA polymerases or ribosomal proteins would be conserved across all domains of life. This idea has been supported in numerous papers (Woese, 1987; Pühler et al., 1989; Lecompte et al., 2002; Vannini and Cramer, 2012) and continues to influence modern molecular phylogenetics. The use of "informational genes" has gained support for several reasons. Many are present and homologous across all domains of life and are associated with processes (replication, transcription, and translation) that are more resistant to phylogenetic artifacts and horizontal gene transfer (HGT) or endosymbiotic gene transfer (EGT) (Jain et al., 1999; Brochier et al., 2000; Abby et al., 2012).

Along with their being more resistant to tree-building artifacts, using "informational genes" avoids sequences related to certain biochemical functions that are not common to all organisms. For example, using proteins involved in photosynthesis to construct a broad scale eukaryotic phylogeny should be avoided because they are specific to photoautotrophs. Similarly, avoiding genes associated with cellular respiration and metabolic processes also help build a

stronger data set by removing lineage specific subtleties that could lead to inaccurate assumptions of relatedness. Therefore, in evaluating what should be considered evolutionarily conserved it is important to determine the most appropriate data for answering the question at hand. If the goal is to address deep evolutionary roots and relationships, then genes should be chosen that have been relatively resistant to large changes over evolutionary time and are resistant to phylogenetic artifacts. These issues can best be addressed by using conserved informational processing genes.

Phylogenetic Artifacts

It has been suggested that many "informational genes" are more resistant to horizontal gene transfer (HGT), an argument known as the "complexity hypothesis" (Jain et al., 1999). This hypothesis suggests that genes involved in transcription, translation and replication are more resistant to lateral movement because of their complexity in structure and interactions with other proteins within the cell (Jain et al., 1999). Furthermore it has been demonstrated empirically that operational (housekeeping) genes involved in metabolic pathways are much more likely to be transferred horizontally (Jain et al., 1999). Interestingly the probability of a gene being subject to HGT is considered inversely proportional to the number of interactions it has with other proteins (Jain et al., 1999).

Horizontal Gene Transfer is best described as a mechanism by which genetic material is transferred from one organism to another in a non-genealogical manner (Goldenfeld and Woese, 2007). While HGT was initially described in microbes, the lateral movement of genes has played a significant role in eukaryotic evolution as well, although predominantly in unicellular eukaryotes (Andersson, 2005). HGT was first described, although not fully understood, as early

as the 1950s by Victor Freeman (Freeman, 1951), but the significance of the process wasn't understood until the mid-1980s (Syvanen, 1985). Starting in the early 1990s, research on HGT really began to take off and continues to impact the reconstruction of organismal relationships (Boto, 2010). HGT plays an important role in determining prokaryotic relationships between closely related species, partially because of the high rate of gene transfer that occurs between these organisms (Than et al., 2006). Although useful in these cases, HGT also leads to inaccurate gene and species phylogenies (Than et al., 2006). Endosymbiotic gene transfer, or EGT, while similar to HGT, describes the transfer of genetic information from an endosymbiont to its host; or, in other words the correlated transfer of numerous genes from a once free-living organism that has been engulfed by another organism. EGT has played a significant role in shaping all eukaryotic chromosomes, from both the ancestors of mitochondria and plastids (Timmis et al., 2004). Additionally, EGT plays a significant role in confounding phylogenetic relationships of host cells across the tree of life, as transfer of genetic material between endosymbiont and host makes it difficult to accurately determine ancient relationships (Lane and Archibald, 2008).

Thus far a great deal of attention has been given to the impacts of HGT and EGT but there are other methodological issues that have arisen in broad scale phylogenomic analyses that need to be addressed to fully understand the current conflicts in the field. As more taxa and new molecular sequence data become available, there also is an increased need to be aware of potential phylogenetic artifacts; historically, the most important has been long-branch attraction (LBA).

Since its discovery by Joseph Felsenstein (1978), long-branch attraction has been a recurrent and critical problem when phylogenies are developed. The primary idea of long-branch attraction, as mentioned above, is that more rapidly evolving species group together on

phylogenetic trees regardless of their actual phylogenetic relationships (Felsenstein, 1978). There are often times when LBA impacts phylogenies to a degree that trees with long branches included show entirely different relationships from ones with long branches excluded (see Stiller and Hall, 1999, Stiller, Riley, and Hall, 2001, Dacks, et al., 2002).

While it is important to be cognizant of LBA, there are other pressing issues facing modern phylogenomics; these include compositional bias, the problem of covarions, and the need to address "psychological" attitudes towards change (Philippe and Laurent, 1998). Although compositional biases are believed to be handled relatively well computationally, covarions and attitudes are harder to address. The covarion model states that the constraints on the evolution of any given site in a molecular sequence alignment can vary in some parts of the tree while remaining invariable in others, or can vary differently across evolutionary lineages and through time (Philippe and Laurent, 1998) (Fitch, 1971). The presence of covarions suggests that there should be little useful phylogenetic signal among protein sequences given 400 million years of divergence (Penny et al., 2001) and even more problematic among reds and greens, which likely diverged over 1.2 billion years ago (Butterfield, 2000). Penny et al. showed that, once enough time has passed (400 MY under their model), the accumulation of covarions is expected to overwhelm useful historical signal in a sequence alignment, and that no phylogenetic branches should be recovered accurately. Although recognized as an issue for years, covarions remain exceedingly difficult to model into tree-building algorithms (Galtier, 2001; Lopez et al., 1999). Thus, it is important to apply a scenario that minimizes sequence covariation, particularly when the species evaluated include rapidly evolving taxa with higher degrees of site variability (Philippe and Laurent, 1998). It is also critical to understand that "psychological" changes are needed with respect to how uncertainties in phylogenomic data are handled. Often, newer

methods are not well received even if the older "tried and true" approaches aren't as accurate or, in some cases are just plain incorrect (Philippe and Laurent, 1998).

Summary

The exhausting amount of data and number of analyses to date, combined with the lack of resolution of the origin of red algae and green plants, begs the question: Is there a consistent way to resolve whether red algae and green plants share a monophyletic relationship? In looking at the diverse methods used in the examples above, the answer would appear to be no. While there have certainly been marvelous developments in the way relationships are evaluated, for almost three decades broad scale eukaryotic phylogenomics has not found a consistent way to accurately resolve the backbone of the eukaryotic tree. That is not to say that the data that have been generated in the past 30 years are all for naught, but rather that approaches are required that best take in to account all of the complicating issues that have been discussed. Having established the historical context of research on red algal and green plant relationships, a discussion of the methodology to further investigate these questions can now be addressed. This project has been centered around two main goals; (1) develop an approach using carefully selected genes and taxa that are least affected by phylogenetic artifacts and is large enough to compare highly conserved regions of core informational genes across a broad range of the Eukarya, (2) use this data set to provide strong evidence for the resolution of the longstanding controversy over the origins of red algae and green plants. Although this project does not provide answers to all of the issues discussed previously, it certainly takes a step in the right direction towards refining how phylogenetic relationships are evaluated. By examining a broad set of informational genes from previous studies (Williams et al., 2012) and others that have been hand selected to be most

refractory to the phylogenetic artifacts that have been reviewed, potential answers to the question of red algal origins can begin to be addressed.

Chapter 2 - Selection of Molecular Markers

Introduction

At the time this project began very little effort had been put into creating large molecular data sets based on *a priori* considerations of genes that were thought to be immune to demonstrating phylogenetic artifacts and other confounding factors known to be an issue in phylogenomic data sets. To address the goals of this project a data set was constructed building on what previous researchers have indicated are conserved "informational genes" that help define the three domains of life (Ciccarelli et al., 2006; Harris et al., 2003; Jain et al., 1999; Williams et al., 2012; Woese et al., 1990). While this project was underway, a related data set was published (Williams et al., 2012) to address different evolutionary questions. It showed that a data set of informational genes could be used to address difficult phylogenetic questions outside the red/green debate, thus adding to the credibility of my proposed methodology. The genes that were ultimately selected for this project adhered to strict set of criteria that enabled thorough testing of the phylogenetic questions addressed, while minimizing, as much as possible, the potential for phylogenetic artifacts that were discussed previously.

The "complexity hypothesis" (Jain et al., 1999) also was taken into account when narrowing down appropriate genes for this study. With these ideas in mind an initial set of approximately 45 genes was selected from 47 species for analysis. Table 1 shows the original genes selected and Table 2 the initial taxa. This data set would have provided more than 2,000 total sequences for comparison. Although not as large as some from recent studies (Cavalier-Smith et al., 2014; Williams et al., 2012) this set of markers contained what were thought to be the informational genes most highly conserved across all eukaryotes. The genes in Table 1 were

selected and subsequently divided into 3 core groups: transcription, translation, and DNA replication. The initial concept was to compare phylogenetic results from each of the groups individually, and as a combined data set, to determine whether the phylogenetic signals from each of the major groups of information processing genes were consistent with each other. However, this approach proved a bit more challenging than originally thought.

Table 1: A list of the genes selected initially selected for analysis. Each gene is grouped into a larger classification that corresponds to their known functional categories (Transcription, Translation, DNA Replication).

Gene	Description	
Transcription		
ELP3	Catalytic histone subunit of RNAPol II elongator complex- transcriptional elongation.	related 18 genes to transcription
RPA1	Replication protein- DNA binding	
RPA2	Replication protein	
RPB1	DNA dependent RNA Polymerase Subunit	
RPB2	DNA dependent RNA Polymerase Subunit	
RPB3	DNA dependent RNA Polymerase Subunit	
RPC1	RNA Polymerase III largest subunit	
RPC2	RNA polymerase III second largest subunit	
RPC3		
Spt5	Transcription initiation factor	
TBP	TATA box binding protein	
TFIIH	Transcription, cell cycle control and DNA repair. Phosphorylates the CTD	
TFIIB	RNAP II preinitiation complex	
TFIIE	Recruits TFIIH and stimulates RNAPII CTD	
TFIIA	Joins TFIID and TBP in binding to promoter	
DNA Replication		
DNA Polymerase subunits	DNA replication	14 genes related to DNA replication
Mcm	Mini chromosomal maintenance	
(Mcm 10 , Mcm 2-7)	Replication initiation	
Topoisomerase		
Translation		
Ribosomal Proteins	Site of translation	13 genes related to translation
Telomerase	Maintains telomere ends- adds TTAGGG	
		45 total genes x 47 organisms = 2,115 sequences

Table 2: A list of the taxa selected for use based on availability of complete genome sequences in a major bioinformatics database (NCBI or JGI). Taxa highlighted in red were removed from the list because they were missing one or more genes from Table 1.

Organism	Abreviation for analyses	Classification
Acanthamoeba	Acan	Amoebozoan
Amphimedon	Amph	Animal
Arababidopsis	Arab	Viridiplantae
Aspergillus	Aspe	Fungi
Aureococcus	Aure	Stramenopile
Batrachochytrium	Batr	Fungi
Bigellowiella	Bige	Rhizaria
Blastocystis	Blas	Stramenopile
Brachypodium	Brac	Viridiplantae
Caenorhabditis	Caen	Animal
Capsaspora	Caps	Opisthokont
Chondrus	Chon	Rhodophyte
Chlamydomonas	Chla	Viridiplantae
Coprinopsis	Copr	Opisthokont
Cryptosporidium	Cryp	Apicomplexan
Cyanidioschyzon	Cyan	Rhodophyte
Cyanophora	Cyanop	Glaucosystophyte
Dictyostelium	Dict	Amoebozoan
Drosophila	Dros	Animal
Ectocarpus	Ecto	Stramenopile
Emiliania	Emil	Haptophyte
Entamoeba	Enta	Amoebozoan
Galdieria	Gald	Rhodophyte
Giardia	Giar	Excavate
Guillardia	Guil	Cryptomonad
Homo	Homo	Animal
Laccaria	Lacc	Fungi
Leishmania	Leis	Euglenoid
Magnaporthe	Magn	Fungi
Monosiga	Mono	Choanozoa
Micromonas	Micr	Viridiplantae
Naegleria	Naeg	Excavate
Ostreoccus	Ostr	Viridiplantae
Paramecium	Para	Cilliate
Perkinsus	Perk	Apicomplexan
Phaeodactylum	Phae	Stramenopile
Physcomitrella	Phys	Viridiplantae
Phytopthora	Phyt	Stramenopile
Plasmodium	Plas	Apicomplexan
Salpingoeca	Salp	Opisthokont

Schizosaccharomyces	Schi	Fungi
Selaginella	Sela	Viridiplantae
Tetrahymena	Tetr	Cilliate
Thalassiosira	Thal	Stramenopile
Trichomonas	Tric	Excavate
Trypanosoma	Tryp	Euglenoid
Volvox	Volv	Viridiplantae
	Total	47 initially, 39 final

A great deal of research went into compiling a set of molecular markers that are both highly conserved and perform a specific function in information processing in the cell. These genes presumably had three major advantages that would help to overcome some of the previous problems seen in broad scale phylogenetic analyses: 1) information processing genes have been shown to be more refractory to HGT than other functional classes (Rivera et al., 1998), 2) they are generally carried as single copy genes rather than gene families, reducing issues of lineage sorting of paralogous sequences, and 3) they have highly conserved functions across all eukaryotes, where studied, and have large core domains that are more resistant to covarions and other biases that lead to phylogenetic artifacts.

Meeting Basic Biological Thinking Criteria

Aside from meeting these criteria, the genes selected were presumed to make sense biologically as carrying out essential functions that should be present across all eukaryotes. For example, ELP3 is involved in transcriptional elongation and is a key part of the RNA polymerase II holoenzyme complex (Wittschieben et al., 1999). Therefore if an organism transcribes protein-encoding genes, it was presumed that elongation occurs as described in model systems and, thus, requires the ELP3 gene. This same thought process was applied as the initial list was created;

however, as each gene was evaluated carefully across all species, it became clear that in some cases simply making sense biologically didn't translate into a reality of universal conservation. In several cases it was discovered that a putatively conserved gene, based on literature review, could not be recovered from a number of species, or even major taxa, in this study. This will be discussed below in more detail through evaluation of single gene data set and trees. This discovery required further careful evaluation of each gene to ensure both its universal conservation and consistency with the criteria behind using informational processing genes in the first place. Because most of the previous work done on informational gene conservation has been evaluated in light of relationships across the three domains of life (Ciccarelli et al., 2006; Foster et al., 2009; Harris et al., 2003; Jain et al., 1999; Rivera et al., 1998; Williams et al., 2012; Woese et al., 1990) it was not essential to assess whether any given gene was present within the eukaryotic domain.

What others have done

The markers selected based on other studies did contain genes present across the tree-of-life (Williams et al., 2012); however, upon closer inspection, these large data sets contained genes already shown to be problematic with respect to the evolutionary questions to be addressed in this study. For example, Elongation Factor (EF2) was previously shown to have issues with possible horizontal gene recombination and/or paralogy (Stiller et al., 2001), but was included in the larger data set used by Williams et al (2012). It should be noted that, although Williams et al. (2012) did provide a great starting point for building list of broadly conserved genes, the questions they sought to address were quite different. Thus, problems related to HGT within eukaryotes were not an issue with respect to relationships among the Bacteria, Archaea and Eukarya. In addition, all informational genes conserved across eukaryotes are not necessarily

present in bacteria or archaeans, simply because of the vast differences in evolutionary history. Nevertheless, it is well supported that informational genes place the Archaea and Eukarya as more closely related to the exclusion of bacteria (Foster et al., 2009; Williams et al., 2012). The similarity in features between eukaryotes and archaeans extends beyond physiological to include a broader resemblance in their informational systems (Forterre, 2013). Thus many potentially important informational genes are not found in the Williams et al. (2012) study because it compared all three domains of life. This absence of informational genes is understandable in large-scale comparisons of the three domains of life and is, in part, what drove the necessity of expanding the data set for this project. A logical reason for the absence of informational genes, however, is not obvious with respect to some other more recent large-scale phylogenomic investigations.

A recent large-scale phylogenomic study examined a range of 73-122 species and between 173-192 genes, all of which were carefully selected and evaluated to avoid issues of paralogy (Cavalier-Smith et al., 2014). However, aside from including the problematic EF2 gene, the most striking characteristic about this large data set was the absence of any large DNA or RNA polymerase subunits. There is no doubt that the genes selected for this large study are highly conserved across eukaryotes. Certainly ribosomal proteins, heat shock proteins, GTP-binding proteins, histones, catalytic subunits for ATP synthase, and others are all critical to cellular function. Key issues arise, however, when considering the underlying biology of the genes being considered. It is interesting that some recent studies (Cavalier-Smith et al., 2014) have not included RNA polymerase genes despite the fact that they should be among the most reliable markers because they are part of large multi-protein complexes that have been shown (Jain et al., 1999) to be most resistant to HGT. Ensuring that the genes one selects satisfy the

"complexity hypothesis" has become increasingly important as HGT has become one of the most compelling problems currently facing phylogenomics. The following paragraphs will look at the genes selected and discuss the initial results from the data set that lead to either inclusion or removal of each gene from the final data set.

Methodology for Selecting Genes

To ensure that the most reliable data set possible was recovered for tracing the evolutionary history of the eukaryotic nucleus, each of the genes in Table 1 underwent a strict selection screening based on the following criteria: 1) a gene had to be present in all of the taxa selected, 2) sequences from each species could not be missing significant portions of core sequence domains, 3) they had to be present as a single copy in all genomes or, if paralogous copies were present, they had to be closely related group-specific paralogs, and 4) there must be no obvious evidence of HGT as determined by single gene phylogenies, meaning that single gene trees had to recover major well-defined, higher-order taxa as monophyletic. This stringent methodology surprisingly eliminated a large portion of the presumed data set; of the 45 genes selected initially, only 6 were retained after final evaluation. A few examples of the selection process will be discussed in subsequent paragraphs to better explain how the final 6 genes were chosen.

Given the support in the literature for the utility of the genes in Table 1, it was surprising to discover that a large number of them could not be recovered from the broad range of taxa selected. Perhaps the two most interesting examples of missing information from this data set are from the DNA polymerase subunits and the mini-chromosome maintenance 10 (MCM10) gene. Canonical DNA polymerase subunits (α, β, γ, δ, ε) have been described for some time as

conserved across eukaryotes, and as sharing a high degree of similarity among all three domains of life (see Hübscher et al., 2002 for review). Classical biological thinking would suggest that all living organisms replicate their DNA in the same manner and, as such, require the same homologous core subunits of DNA polymerase. Upon further investigation, however, several subunits were not found in all species. Although subunits δ and ε were recovered from all taxa, they were eventually removed. Their removal was based on the idea that if the whole of DNA polymerase subunits could not satisfy the "complexity hypothesis" then they should be removed. In other words, if the complex protein environment thought to be shared by all eukaryotes is not conserved across the domain, it is likely that changes in the biochemical context of the protein domains' interactions must have occurred as well, thereby resulting in covarions. This ultimately led to the exclusion of the DNA polymerase subunits all together.

The role of MCM proteins is well studied and suggests that yeast and animal cells require the formation of a MCM2-7 hexamer to carry out DNA replication (Liu et al., 2009). Although MCM10 is thought to be required for DNA replication in eukaryotes the exact role is still not well understood (Thu and Bielinsky, 2013). Initially all of the MCM 2-7 subunits as well as MCM10 were included in the dataset but after further investigation it was discovered that various MCM 2-7 subunits either could not be recovered from multiple taxa, or were missing a significant portion of core domain sequences in different taxa. Therefore the MCM2-7 complex also was excluded from the DNA replication gene data, set leaving only MCM 10. Interestingly careful analysis of conservation of MCM10 found it also to be unrecoverable from numerous taxa. Additionally, this analysis showed, for the first time, that the C terminal domain of MCM10 is not found in higher land plants or fungi, although present in green algae and metazoa. Such variation in domain architecture is a recipe for covarions in phylogenetic analyses. Although

more work is still required to determine the implications of the evolutionary distribution of MCM10 and its C-terminal domain, the survey rendered MCM10 unacceptable for this analysis, along with proteins with which it interacts. Table 3 shows a list of originally selected genes and the criteria used to either keep or eliminate them from the data set.

Table 3: Table depicting the criteria for genes to be included in the study. An "X" in a particular column denotes that gene does not meet that criterion. Genes highlighted in green without any markings were used for further analyses.

Gene	Present	Complete Sequence	Single Copy	Recover well defined groups
DNA Polymerase subunits	X	X		X
ELP3	X	X		
MCM 2-7			X	
MCM10	X	X		
Ribosomal proteins			X	
RPA1				
RPA2				
RPB1				
RPB2				
RPB3				X
RPC1				
RPC2				
RPC3				
Spt5		X		X
TBP				X
Telomerase	X	X		
TFIIA	X	X		
TFIIB	X	X		
TFIIE	X	X		
TFIIH	X	X		X
Topoisomerase	X	X		

Symbol Key
Missing significant data
Does not recover well defined groups
Paralogy detected
Could not recover gene or all components of a complex

Table 3 shows that of all of the information genes related to transcription, translation and DNA replication, only the two largest subunits of the DNA-dependent RNA polymerases passed all required tests for inclusion in phylogenetic analysis. A great deal of previous work, both molecular and phylogenetic, has supported the conservation of function of transcriptional subunits, particularly the conserved domains of the largest subunits of RNA polymerases (Butler and Kadonaga, 2002; Lane and Darst, 2010; Murakami et al., 2013, 2002; Proshkina, 2006; Pühler et al., 1989; Vannini and Cramer, 2012; Wittschieben et al., 1999). Furthermore the core subunits of RNA polymerase II have been even more thoroughly investigated at both molecular and phylogenetic levels (Butler and Kadonaga, 2002; Lane and Darst, 2010; Murakami et al., 2013, 2002; Proshkina, 2006; Pühler et al., 1989; Stiller and Hall, 1998, 2002; Stiller and Harrell, 2005; Vannini and Cramer, 2012; Wittschieben et al., 1999). These prior results, combined with the thorough evaluation these genes underwent in this study, helped to solidify the decision to use only RNA polymerase subunits to address the question of red algae and green plant monophyly. Moreover, extensive and careful investigations have demonstrated no evidence that a core eukaryotic RNA polymerase subunit ever has been transferred horizontally, or carried for long evolutionary periods as duplicated paralogs across species (Iyer et al., 2004; Lane and Darst, 2010). Although the initial data set contained eight polymerase subunits, once each was evaluated via the strict criteria discussed above, only the 6 largest subunits (A1-2, B1-2, C1-2) were retained. The potential reliability of these data comes not only from the theoretical considerations, but also from empirical evidence (shown in the trees in the following chapter); specifically, single gene trees of each subunit show no evidence of HGT across taxa, or even the possibility that subunits can be swapped between the RNA polymerases within the same organism.

As enigmatic as it is to find that genes previously considered to be conserved molecular markers are missing from annotated genomes of certain taxa, it should be noted that their absence from this data set does not mean they definitely are absent in those taxa. A great deal of effort went into carefully evaluating each gene to ensure that a reciprocal BLAST (Basic Local Alignment Search Tool) search into NCBI or JGI returned the actual gene queried; however, problems of misannotation and poor quality sequence data can lead to inaccurate conclusions and misinformation. Although only "complete genomes" were queried, until each gene from each organism was carefully evaluated, it cannot be determined unequivocally that certain genes are missing. Additionally, while problematic genes were removed via the criteria above, there were several cases that led to the removal of what could ultimately prove to be acceptable genes. In these cases it was decided that if nearly all other genes of the category (transcription, translation, DNA replication) failed our criteria, the one or two genes remaining would not accurately represent the category, and missing genes could reflect dramatic functional differences that could lead to covarions in phylogenetic reconstruction. In total, only the two largest subunits of each of the three eukaryotic RNA polymerases were selected as the most reliable data set to address the issue of red/green monophyly.

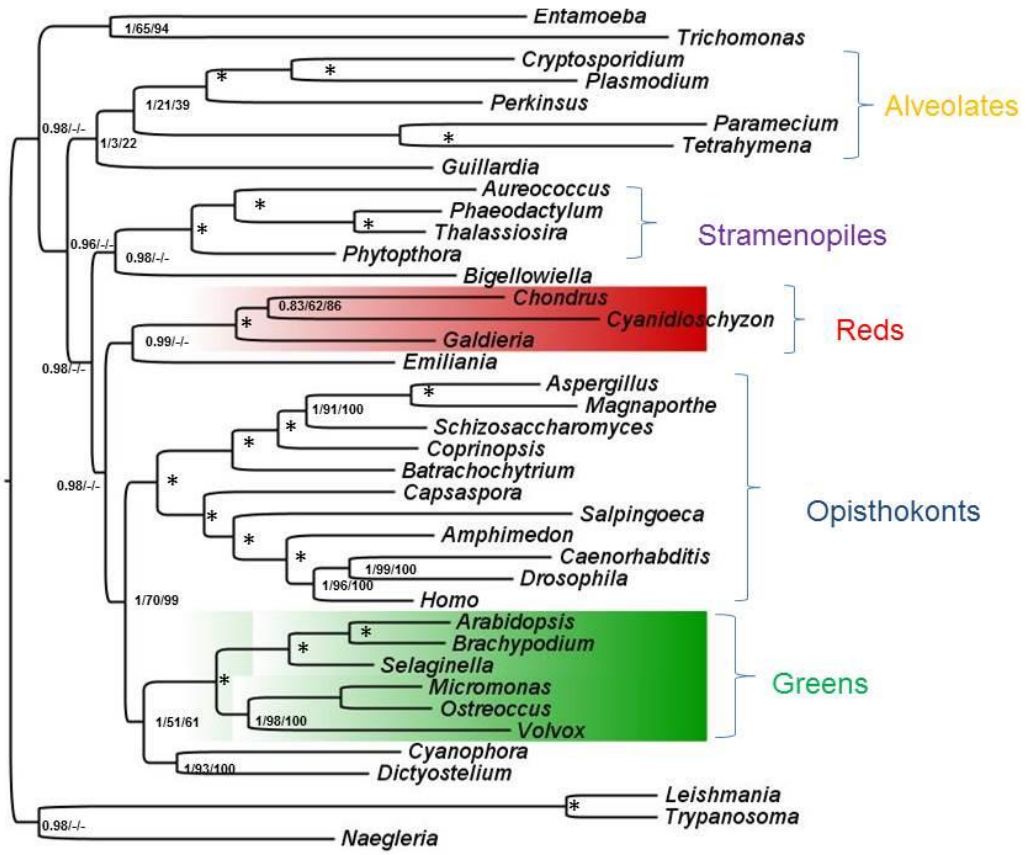
Chapter 3 – Phylogenetic analyses of RNAP subunits

The question of red/green monophyly was addressed through multigene analyses of 39 species, using the two largest subunits of three RNA polymerases (4,313 positions). This chapter presents the best tree topologies obtained from maximum-likelihood (ML) via PhyML version 3.0 (Guindon et al., 2010) and Bayesian inference in MrBayes v 3.2.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003), statistical tests of alternative evolutionary hypotheses related to red and green plants, as well as evidence that the results are not dictated by endosymbiotic gene transfer (EGT).

Upon collection and verification of sequences by the methods discussed previously, a multiple sequence alignment (MSA) was created for the data set using MEGA 6.06 (Tamura et al., 2013). This MSA was analyzed using Gblocks (Castresana, 2000; Talavera and Castresana, 2007), allowing for gaps in the final blocks, to identify the conserved domains in the alignment in an unbiased manner. The conserved regions were then subjected to ML and Bayesian phylogenetic analyses.

In the initial tests of this data set, both Bayesian and ML trees supported a polyphyletic relationship between red algae and green plants with moderate to strong support. While both analyses recovered similar topologies supporting polyphyly of reds and greens, the Bayesian tree was found to be the most likely tree after thorough evaluation with FastTree (Price et al., 2009, 2010) and Consel (Shimodaira and Hasegawa, 2001) using Gamma20 likelihood and a WAG substitution model. The best (Bayesian) topology is illustrated in **Figure 2.**

Figure 2: Best tree topology from Bayesian inference on 39 taxa and 4313 positions, with support values from Bayesian, ML, and jackknifed analyses. Bayesian posterior probabilities/ML bootstrap values/jackknifed ML bootstrap values.



* Values = 1/100/100

As noted above, maximum-likelihood analyses and Bayesian inference recovered trees with slightly different topologies. While all major taxa were recovered as monophyletic, several taxa represented by individual species changed positions between the two trees. The inconsistencies between Bayesian and ML trees can largely be attributed to low support values along the backbone of the tree in ML analysis; however, the goal of this project focused on the specific question of a putative relationship between red algae and green plants, not on global

relationships of the eukaryotic tree. The ML and Bayesian trees were not significantly different based on tests in Consel; however, Bayesian analyses recovered the topology with the highest likelihood in all analyses of both the larger and core data sets and, therefore, the Bayesian tree is displayed in all tree figures shown (e.g. see **Figures 2** & **3**)

While the sequences used should be relatively unaffected by phylogenetic artifacts, and most representative of reliable and phylogenetically informative positions, the number of total positions used is small compared to other recent studies (Cavalier-Smith et al., 2014; Williams et al., 2012). To estimate the strength of the phylogenetic signal in polymerase subunits, if the same tree-building signal was found across data sets of comparable size to those in larger phylogenomics studies, a modified power analysis was performed on this alignment. To accomplish this analysis the data set was resampled by jackknifing to produce alignments of ~50,000 positions. The jackknifed data set was then subjected to ML bootstrap analyses. Interestingly jackknifed ML analysis recovered a similar topology to the ML tree from the smaller (original) data set; however, bootstrap values were significantly improved on the jackknifed tree. These results indicate that, given a larger data set with comparable signal to the six genes analyzed here, monophyly of red algae and green plants would be rejected even more strongly.

To determine whether a monophyletic Archaeplastida can be rejected based on these data, two additional hypothesis tests were run in Fasttree and Consel to confirm the results. The best Bayesian topology for this data set was rearranged to 1) force reds and greens together in monophyletic relationship, and 2) create a monophyletic clade containing reds, greens and the glaucophyte *Cyanophora* (the so-called Archaeplastida). Interestingly, when *Cyanophora*, previously argued to group monophyletically with reds and greens based on a shared

photosynthetic history (Moreira et al., 2000; Rodríguez-Ezpeleta et al., 2005), was added to this larger data set, red/green polyphyly was still shown to be supported over the monophyly of either reds/greens or reds/greens and glaucophytes (**Figure 4**). Although the monophyly of reds and greens is rejected at just short of significance in the non-Jackknifed data set, once the data were jackknifed the values become highly significant (**Figure 4**).

After confirming polyphyly of reds and greens using the largest available data set, the analyses were taken a step further to evaluate the signal obtained from only the most stringently selected sequence data. In this "core" alignment we removed all regions of the alignment with missing data, as well as potentially misaligned sequences because of presence of inferred indels. The resulting alignment contained only the most functionally conserved RNA polymerase domains, which significantly limits the potential impact of covarions.

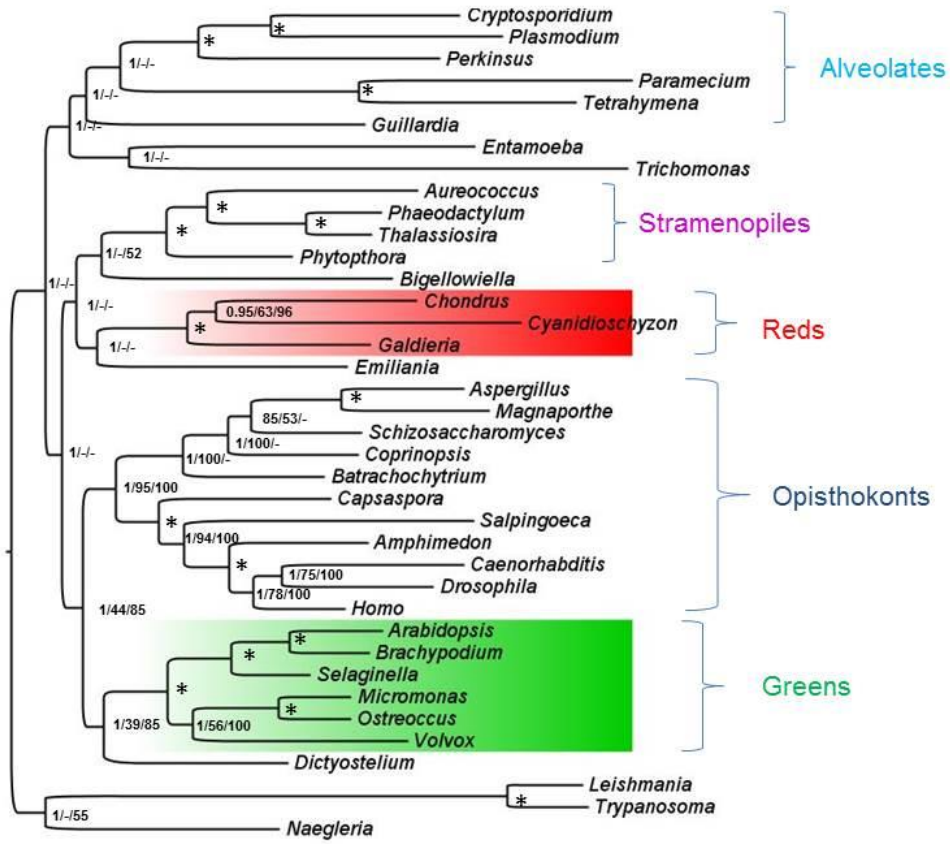
It was particularly important to consider the impact of missing data with *Cyanophora* due to its putatively shared photosynthetic history with reds and greens. That is, it was possible that missing data from *Cyanophora* artificially created red/green polyphyly; therefore, it was critical to remove taxa with large amounts of sequences missing from the alignment to ensure the red/green polyphyly recovered wasn't an artifact of missing data.

Additionally, any potentially misaligned sequences were removed through rigorous evaluation of the data set. This evaluation was performed on the alignment previously evaluated by Gblocks, which already had objectively identified conserved domains. These domains were further trimmed, by hand, to regions anchored on both ends by invariable sites and without any insertions or deletions present. In limiting the data to only these regions, we significantly reduced

the likelihood of covarions arising, by including functional and physical structures shared among polymerases across the three domains of life (see Vannini and Cramer, 2012 for review).

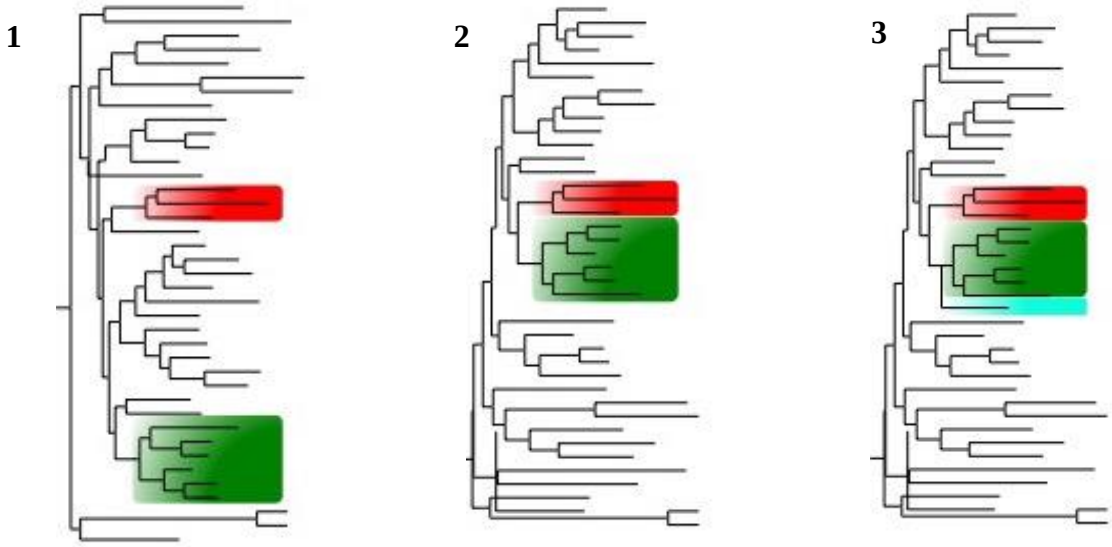
This careful removal of missing or potentially incorrectly aligned sequences produced a smaller alignment of 2,941 positions from 38 taxa. These sequences were termed the "final core" data and were subjected to the same treatments described above, which again confirmed a polyphyletic red/green association (**Figure 3**). To further assess whether the lack of a relationship between reds and greens was statistically significant, a Consel analysis was run on both jackknifed and the original core data sets, in which Consel was given the topology of the tree inferred through Bayesian inference and a tree in which reds and greens were forced into a monophyletic clade (**Figure 4**). Using the jackknifed data set, all statistical evaluations confirmed that a monophyletic red/green relationship is rejected (**Figure 4**). In contrast, although analyses of the original core data set also yielded polyphyletic red/green trees, topology tests did not reject a monophyletic red green relationship at a statistically significant level (**Figure 4**). Despite this lack of significance, the smaller core data set still supports a polyphyletic relationship among red algae and green plants in both initial and jackknifed analyses, indicating this inference was not an artifact of missing or poorly aligned sequences in the original, larger alignment (**Figure 4**).

Figure 3- Best Bayesian tree topology from analyses of the final core data set of 38 taxa and 2941 positions. ML and Jackknifed ML trees were similar, but with lower likelihood scores and are not shown. Node values are as follows: Bayesian posterior probabilities/ML bootstrap/ Jackknifed ML bootstrap.



* Values = 1/100/100

Figure 4: Alternative tree topologies tested. **1)** Represents the optimal tree recovered from Bayesian inference **2)** is the same tree with Reds and Greens are forced into a monophyletic relationship **3)** is the same tree with Reds, Greens and *Cyanophora* as a monophyletic group. Statistical values for evaluating the best tree from the larger (containing *Cyanophora*) and final core data sets, as well as their respective jackknifed are presented in the associated table. The trees depicted represent the topology for the larger data set but illustrate the same hypothesis test in the final core data set, with the exception of the absence of *Cyanophora*. **Key to the table: rank:** the trees depicted, **item:** the label for the tree, **au:** the p-value of the approximately unbiased test calculated from the multiscale bootstrap, **kh:** Kishino-Hasegawa test, **sh:** Shimodaira-Hasegawa test, **wkh:** weighted Kishino-Hasegawa test, **wsh:** weighted Shimodaira-Hasegawa test.



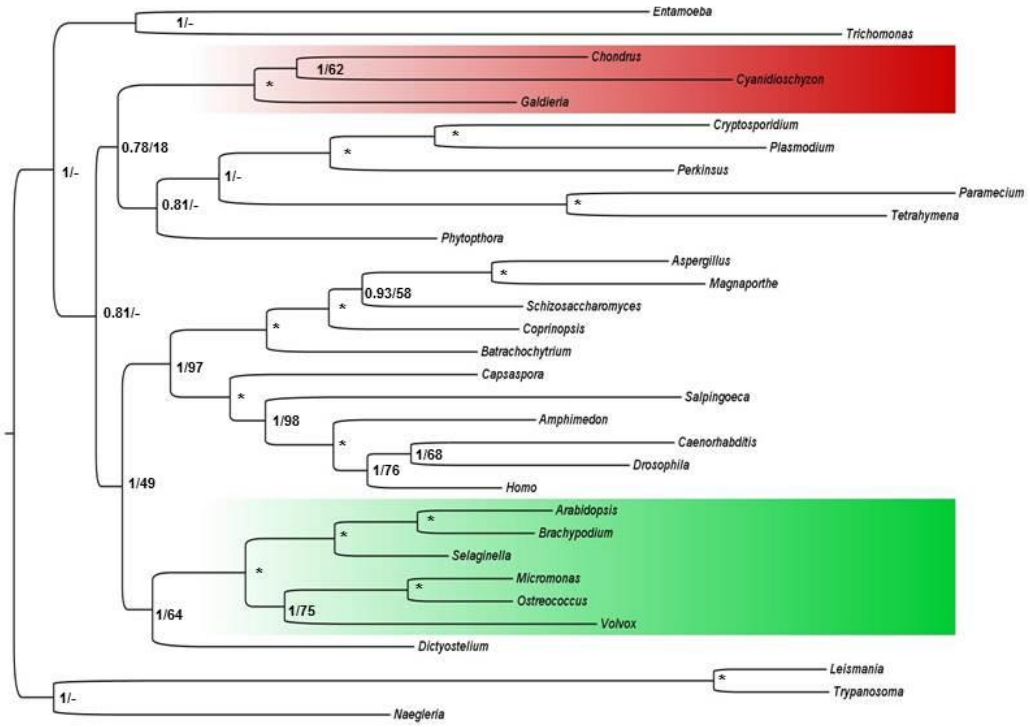
* Data set without *Cyanophora*

Larger Non-Jackknifed	**Rank**	**Item**	**Au**	**kh**	**sh**	**Wkh**	**wsh**
Best Tree	**1**	**1**	**0.939**	**0.934**	**0.964**	**0.934**	**0.971**
Red/Green monophyly	2	2	0.069	0.066	0.087	0.066	0.123
Monophyletic Plantae	3	3	3e-04	0.002	0.002	0.002	0.003
Larger jackknifed							
Best Tree	**1**	**1**	**0.989**	**0.987**	**0.992**	**0.987**	**0.994**
Red/Green monophyly	2	2	0.011	0.013	0.013	0.013	0.022
Monophyletic plantae	3	3	2e-09	0	0	0	0
Final* Non-Jackknifed							
Best Tree	**1**	**1**	**0.834**	**0.827**	**0.827**	**0.827**	**0.827**
Red/Green monophyly	2	2	0.166	0.173	0.173	0.173	0.173
Final* Jackknifed							
Best Tree	**1**	**1**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Red/Green monophyly	2	2	2e-05	3e-04	3e-04	3e-04	3e-04

Although the sequences used in this study should be among the least impacted by horizontal gene transfer, we further investigated the potential influences of EGT on phylogenetic analyses by excluding algal taxa containing secondary plastids derived from red and green algae. If nuclear RNA polymerase genes were transferred to these organisms along with plastids, they would tend to pull reds and greens away from each other, and toward the secondary recipients of their plastids. The removal of these taxa does not change overall tree topology substantially, clearly showing that reds and/or greens are not artificially recovered as polyphyletic because of EGT (**Figure 5**).

Taking the results of these analyses in totality several things become clear. As expected, the largest subunits of eukaryotic RNA polymerases appear to be resistant to HGT, and recover a polyphyletic relationship between red algae and green plants, with very strong statistical support in power analyses that apply comparably sized data sets to those used in previous phylogenomic studies that recovered a monophyletic Archaeplastida.

Figure 5. Tree topology of Final Core best tree with non-Red/Green photosynthetic organisms removed. Node values are as follows: Bayesian posterior probabilities/ML bootstrap/ Jackknifed ML bootstrap.



***** Values = 1/100/100

Chapter 4 -Discussion

Using the most reliable molecular markers, a core data set of the two largest subunits of three RNA polymerases, which are among the genes least likely to be affected by phylogenetic artifacts, red algae and green plants have been shown to be polyphyletic. This result was demonstrated through recovery of this relationship in both Bayesian and ML phylogenetic analyses, and by statistical tests that refute a monophyletic red/green association based on these data. A reliable demonstration a polyphyletic relationship between red algae and green plants helps shed light on several of the most problematic issues facing algal phylogenomics.

While these results cannot specifically address where and when plastids originated they can aid in future evaluations of plastid origin questions. For quite some time there has been a need to reevaluate how the evolution of plastids has been interpreted in the past three decades (Stiller, 2014). Empirical evidence can be interpreted as being consistent with either a single or multiple plastid origins, and the debate continues because the phylogenetic relationships of Archeaplastida have been largely ambiguous. The results presented here help obtain a more firm understanding the origin of plastids, and provide a solid footing from which to test more direct hypotheses regarding plastid evolution.

The polyphyletic red/green association recovered by this study adds to the small yet increasing body of evidence that rejects a monophyletic Plantae. Thus, a particularly interesting case can be made for multiple origins of primary plastids given the results we have presented.

While previous studies have shown support for monophyly of primary plastid containing lineages (Cavalier-Smith, 2000; McFadden, 2001; Moreira and Philippe, 2001; Palmer, 2000, 2003) our demonstration of polyphyly supports alternative hypotheses of independent or multiple

plastid origins (Howe et al., 2008; Stiller, 2014, 2014; Stiller and Harrell, 2005). One strong piece of evidence that is consistent with our result is that no red algal plastid has ever been shown to replace its phycobilisomes with chlorophyll b or c and no green plastid has ever reverted to the use of phycobilisomes. Therefore, assuming a single origin of plastids in the common ancestor of red and green algae invokes evolutionary processes that are never observed in nature (see Stiller, 2014 for review). Similarly, evaluating chlorophyll antennae complexes in plastids has yet to support evidence of the inheritance of photosynthetic pigment complexes that is proposed under either the Archaeplastida or chromalveolate hypothesis (Stiller, 2014). Taking in conjunction with our strong rejection of a monophyletic Archeaplastida, these observations suggest that it is highly improbable for plastids to have evolved via direct decent from a single common ancestor. The fact that we can confidently reject Archaeaplastid monophyly with this core set of molecular markers is an exciting result, particularly when the totality of evidence from other studies is taken into consideration.

Although polyphyletic relationship between reds and greens helps to solidify support for multiple plastid origins, one cannot neglect the potential role intervening and intermediate taxa have played in plastid evolution. It is very likely that extinct or unstudied intervening lineages have influenced extant plastid containing taxa (Stiller et al., 2003). These taxa have likely played significant roles in shaping the current landscape of algal and plant diversity, yet are rarely mentioned when considering plastid origins. While it is impossible to know how much influence extinct and missing taxa have on what we currently observe, it remains highly probable that they played a much larger role than they receive credit for. Acknowledging these factors when evaluating plastid origins is becoming increasingly important and has been incorporated into a new "empirical framework" for interpreting plastid evolution (Stiller, 2014). By better

understanding the role and evolution of plastids we can more confidently assess current relationships see among extant members of the Archaeplastida.

As critical as it is to better understand the origins of plastids, this project was primarily focused on evaluating the monophyly of red algae and green plants via nuclear encoded genes. The methodology we have employed in recovering this polyphyletic red/green association could be of potential use in evaluating other problematic phylogenies outside this debate. Problematic relationships exist throughout the tree of life (Baldauf, 2003; Dunn et al., 2008; Halanych, 2004; Moreira et al., 2000; Philippe and Laurent, 1998; Philippe et al., 2011; Ragan and Gutell, 1995; Stiller and Hall, 1997; Williams et al., 2012) yet, while the methods employed across studies seem to provide accurate answers to *a priori* considerations of respective relationships, all still produce results that is subject to speculation.

We believe that the data we used here to elucidate relationships provide a solution to other mismatched collections of molecular markers. The justification for choosing the genes we have selected has already been discussed at great length. It should be reiterated, however, that this data set is the most resistant to phylogenetic artifacts, issues of paralogy and covarions, and has been shown to contain only the most functionally conserved regions of what are already considered universally conserved genes. We have collected a set of molecular markers that is probably the most free of the issues that plague typical large scale phylogenomic data sets. Incorporating broader taxon sampling with this methodology (as sequences become available) has potential to answer previously intractable or difficult phylogenetic questions.

As exciting and successful as this study has been, it does not go unrecognized that there are still several issues that need to be addressed. As has been the case since the dawn of

molecular systematics an increase in taxon sampling is critical to more accurately evaluating problematic phylogenies. While it is certainly important to consider points made earlier about more taxa not being the only answer (Philippe et al., 2011), perhaps it is better to restate this idea as more "accurately" sampled taxa. Throughout the collection of genes used in this study there were numerous instances in which putatively conserved sequences had not been found in various complete genomes. Much of this represented misannotation of genomic sequences. There was no particular pattern to these discoveries and our experience suggests that misannotation of gene sequences pervades phylogenomic databases. As the rate of sequencing increases rapidly, the need for more thorough screening and care in genome annotation is at an all-time high. Finding so many poorly annotated sequences is one of the reasons we chose not to perform this study using the popular automated pipelines employed in many large-scale studies. Based on our results, such pipeline could allow inappropriately annotated data to be included in a data set and, as a result, potentially influence the final interpretation of relationships.

Additionally, automated screening of BLAST results, for example, can exclude misannotated sequences from the larger alignment and, if left unchecked, could lead to the conclusion that a particular taxon was missing the gene of interest. Each of our 45 genes in the original data set was reciprocally BLASTed to ensure it was appropriately annotated. In some cases this led to removal of a gene from the list; however, in other cases it allowed us to include sequence data that otherwise would have been excluded had we not evaluated each gene by hand. While this methodology is certainly more time consuming, it ensures that sequences recovered for each phylogenetic analysis have been checked carefully to confirm their proper identities and orthologies. Although perhaps overly optimistic, the hope is that an increase in scrutiny of

sequences added to large genomics databases will help to remedy some of the issues that have plagued eukaryotic phylogenomics.

Having shown that, in both our larger and final core data sets, the two largest subunits of three RNA polymerases do not significantly reject red/green monophyly in topology tests (without jackknifed power analyses) we are left with several possibilities. The small number of positions in our original alignments seems to be the most likely factor in preventing stronger statistical support for red/green polyphyly. Additionally, we have demonstrated that in the broad sampling of the Eukarya, genes from major gene families are missing or turn out to be paralogous. Therefore, we could not generate a large enough sample of reliable sequences by "natural" means. Thus, it is worth posing the question, what does this say about the direction the field of eukaryotic phylogenomics has taken? We have searched "informational genes" included in transcription, translation, and DNA replication that have been accepted for many years to be conserved across all domains of life yet we can only recover only six genes that are not susceptible to factors known to result in phylogenetic artifacts. Thus, it remains to be seen whether it is possible to "naturally" assemble a data set with enough genes resistant to artifacts to provide statistically significant results without inferences from power analyses. These questions will continue to go largely unanswered until more taxa and better sequence data for problematic taxa becomes available.

The demonstration of a polyphyletic relationship between red algae and green plants, with reasonably strong statistical support compared to comparably sized data sets, shows great promise for changing how problematic phylogenies are assessed. The recovery of this relationship provides support for multiple origins of primary plastids, rejects a monophyletic Archaeplastida recovered from earlier studies involving molecular markers of lesser quality, and

outlines a methodology for addressing future problematic phylogenetic relationships. Broader taxon sampling and future sequencing of additional taxa that are not currently represented in genomics databases, will improve the power of this methodology and could help to bring about major changes in the way current phylogenomic evaluations are performed. The number of new phylogenetic studies published continues to grow rapidly. As more research points to other previously unrecognized phylogenetic relationships, it becomes increasingly important to ensure that the results obtained are accurate and not the result of sequence-based artifacts. In his 1874 publication *The descent of man, and selection in relation to sex,* Charles Darwin suggested "False facts are highly injurious to the progress of science, or they often long endure; but false views, if supported by some evidence, do little harm, as everyone takes a salutary pleasure in proving their falseness; and when this is done, one path towards error is closed and the road to truth is often at the same time opened." Guided by the words of Darwin and the promise of the methodology discussed above for dealing with problematic phylogenies, the convoluted nature of algal phylogenomics could become substantially easier to manage.

References

Abby, S.S., Tannier, E., Gouy, M., and Daubin, V. (2012). Lateral gene transfer as a support for the tree of life. Proceedings of the National Academy of Sciences *109*, 4962–4967.

Amit Roy, S.R. (2014). Molecular Markers in Phylogenetic Studies-A Review. Journal of Phylogenetics & Evolutionary Biology *02*.

Andersson, J.O. (2005). Lateral gene transfer in eukaryotes. CMLS, Cell. Mol. Life Sci. *62*, 1182–1197.

Ané, C., and Sanderson, M. (2005). Missing the Forest for the Trees: Phylogenetic Compression and Its Implications for Inferring Complex Evolutionary Histories. Systematic Biology *54*, 146–157.

Archibald, J.M. (2007). Nucleomorph genomes: structure, function, origin and evolution. BioEssays *29*, 392–402.

Baldauf, S.L. (2003). The Deep Roots of Eukaryotes. Science *300*, 1703–1706.

Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., and Doolittle, W.F. (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data. Science *290*, 972–977.

Barbash, Z.S., Weissman, J.D., Mu, J., and Singer, D.S. (2013). Core promoter elements are not essential for transcription in mammals. Epigenetics & Chromatin *6*, P91.

Beiko, R., Doolittle, W.F., and Charlebois, R. (2008). The Impact of Reticulate Evolution on Genome Phylogeny. Systematic Biology *57*, 844–856.

Benson, D.A., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2015). GenBank. Nucleic Acids Research *43*, D30–D35.

Bhattacharya, D., and Medlin, L. (1995). The Phylogeny of Plastids: A Review Based on Comparisons of Small-Subunit Ribosomal RNA Coding Regions. Journal of Phycology *31*, 489–498.

Bhattacharya, D., and Medlin, L. (1998). Algal phylogeny and the origin of land plants. Plant Physiology *116*, 9–15.

Bhattacharya, D., Yoon, H.S., and Hackett, J.D. (2004). Photosynthetic eukaryotes unite: endosymbiosis connects the dots. BioEssays *26*, 50–60.

Body\l, A., Mackiewicz, P., and Stiller, J.W. (2007). The intracellular cyanobacteria of Paulinella chromatophora: endosymbionts or organelles? Trends in Microbiology *15*, 295–296.

Bodył, A., Mackiewicz, P., and Stiller, J.W. (2009). Early steps in plastid evolution: current ideas and controversies. BioEssays *31*, 1219–1232.

Body\l, A., Stiller, J.W., and Mackiewicz, P. (2009). Chromalveolate plastids: direct descent or multiple endosymbioses? Trends in Ecology & Evolution *24*, 119–121.

Boto, L. (2010). Horizontal gene transfer in evolution: facts and challenges. Proceedings of the Royal Society of London B: Biological Sciences *277*, 819–827.

Brochier, C., Philippe, H., and Moreira, D. (2000). The evolutionary history of ribosomal protein RpS14: Trends in Genetics *16*, 529–533.

Burger, G., Saint-Louis, D., Gray, M.W., and Lang, B.F. (1999). Complete sequence of the mitochondrial DNA of the red alga Porphyra purpurea: cyanobacterial introns and shared ancestry of red and green algae. The Plant Cell Online *11*, 1675–1694.

Butler, J.E., and Kadonaga, J.T. (2002). The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes & Development *16*, 2583–2592.

Butterfield, N.J. (2000). Bangiomorpha pubescens n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. Paleobiology *26*, 386–404.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. *17*, 540–552.

Cavalier-Smith, T. (2000). Membrane heredity and early chloroplast evolution. Trends Plant Sci. *5*, 174–182.

Cavalier-Smith, T., Chao, E.E., Snell, E.A., Berney, C., Fiore-Donno, A.M., and Lewis, R. (2014). Multigene eukaryote phylogeny reveals the likely protozoan ancestors of opisthokonts (animals, fungi, choanozoans) and Amoebozoa. Molecular Phylogenetics and Evolution *81*, 71–85.

Chan, C.X., Yang, E.C., Banerjee, T., Yoon, H.S., Martone, P.T., Estevez, J.M., and Bhattacharya, D. (2011). Red and green algal monophyly and extensive gene sharing found in a rich repertoire of red algal genes. Curr. Biol. *21*, 328–333.

Ciccarelli, F.D., Doerks, T., Mering, C. von, Creevey, C.J., Snel, B., and Bork, P. (2006). Toward Automatic Reconstruction of a Highly Resolved Tree of Life. Science *311*, 1283–1287.

Copeland, H.F. (1938). The Kingdoms of Organisms. The Quarterly Review of Biology *13*, 383–420.

Cramer, P. (2001). Structural Basis of Transcription: RNA Polymerase II at 2.8 Angstrom Resolution. Science *292*, 1863–1876.

Dacks, J.B., Marinets, A., Doolittle, W.F., Cavalier-Smith, T., and Logsdon, J.M. (2002). Analyses of RNA Polymerase II Genes from Free-Living Protists: Phylogeny, Long Branch Attraction, and the Eukaryotic Big Bang. Mol Biol Evol *19*, 830–840.

Darwin, C. R. (1874). *The descent of man, and selection in relation to sex*. London: John Murray. 2nd ed., Scanned copy. Retrieved from http://darwin-online.org.uk/converted/published/1874_Descent_F944/1874_Descent_F944.html

Delwiche, C.F. (1999). Tracing the Thread of Plastid Diversity through the Tapestry of Life. The American Naturalist *154*, S164–S177.

Delwiche, C.F., and Palmer, J.D. (1997). The origin of plastids and their spread via secondary symbiosis. In Origins of Algae and Their Plastids, D.D. Bhattacharya, ed. (Springer Vienna), pp. 53–86.

Dorrell, R.G., and Smith, A.G. (2011). Do Red and Green Make Brown?: Perspectives on Plastid Acquisitions within Chromalveolates. Eukaryotic Cell *10*, 856–868.

Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., et al. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. Nature *452*, 745–749.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Felsenstein, J. (1978). Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. Systematic Zoology *27*, 401.

Fitch, W.M. (1971). The nonidentity of invariable positions in the cytochromes c of different species. Biochemical Genetics *5*, 231–241.

Forterre, P. (2013). The Common Ancestor of Archaea and Eukarya Was Not an Archaeon. Archaea *2013*, e372396.

Foster, P.G., Cox, C.J., and Embley, T.M. (2009). The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. Philosophical Transactions of the Royal Society of London B: Biological Sciences *364*, 2197–2207.

Freeman, V.J. (1951). Studies on the virulence of bacteriophage-infected strains of Corynebacterium diphtheriae. J. Bacteriol. *61*, 675–688.

Galtier, N. (2001). Maximum-Likelihood Phylogenetic Analysis Under a Covarion-like Model. Mol Biol Evol *18*, 866–873.

G Burger, D.S.-L. (1999). Complete sequence of the mitochondrial DNA of the red alga Porphyra purpurea. Cyanobacterial introns and shared ancestry of red and green algae. The Plant Cell *11*, 1675–1694.

Geyer, C.J. (1991). Markov chain Monte Carlo maximum likelihood.

Goldenfeld, N., and Woese, C. (2007). Biology's next revolution. Nature *445*, 369–369.

Grossman, A.R. (2007). In the grip of algal genomics. Adv. Exp. Med. Biol. *616*, 54–76.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321.

Halanych, K.M. (1991). 5S ribosomal RNA sequences inappropriate for phylogenetic reconstruction. Molecular Biology and Evolution *8*, 249–253.

Halanych, K.M. (2004). The New View of Animal Phylogeny. Annual Review of Ecology, Evolution, and Systematics *35*, 229–256.

Han, L., Masani, S., Hsieh, C., and Yu, K. (2014). DNA Ligase I Is Not Essential for Mammalian Cell Viability. Cell Reports *7*, 316–320.

Harris, J.K., Kelley, S.T., Spiegelman, G.B., and Pace, N.R. (2003). The Genetic Core of the Universal Ancestor. Genome Res. *13*, 407–412.

Hastings, W.K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika *57*, 97–109.

Heath, T.A., Zwickl, D.J., Kim, J., and Hillis, D.M. (2008). Taxon Sampling Affects Inferences of Macroevolutionary Processes from Phylogenetic Trees. Systematic Biology *57*, 160–166.

Hillis, D.M., and Bull, J.J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic Biology *42*, 182–192.

Holder, M., and Lewis, P.O. (2003). Phylogeny estimation: traditional and Bayesian approaches. Nature Reviews Genetics *4*, 275–284.

Howe, C.J., Barbrook, A.C., Nisbet, R.E.R., Lockhart, P.J., and Larkum, A.W.D. (2008). The origin of plastids. Philos. Trans. R. Soc. Lond., B, Biol. Sci. *363*, 2675–2685.

Hübscher, U., Maga, G., and Spadari, S. (2002). Eukaryotic Dna Polymerases. Annual Review of Biochemistry *71*, 133–163.

Huelsenbeck, J.P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics *17*, 754–755.

Inagaki, Y., Nakajima, Y., Sato, M., Sakaguchi, M., and Hashimoto, T. (2009). Gene Sampling Can Bias Multi-Gene Phylogenetic Inferences: The Relationship between Red Algae and Green Plants as a Case Study. Molecular Biology and Evolution *26*, 1171–1178.

Iyer, L.M., Koonin, E.V., and Aravind, L. (2004). Evolution of bacterial RNA polymerase: implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer. Gene *335*, 73–88.

Jain, R., Rivera, M.C., and Lake, J.A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. Proceedings of the National Academy of Sciences *96*, 3801–3806.

Kowallik, K.V. (1994). FROM ENDOSYMBIONTS TO CHLOROPLASTS: EVIDENCE FOR A SINGLE PROKARYOTIC/EUKARYOTIC ENDOCYTOBIOSIS. Endocytobiosis & Cell Res. *10*, 137–149.

Lake, J.A. (1988). Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. Nature *331*, 184–186.

Lane, C.E., and Archibald, J.M. (2008). The eukaryotic tree of life: endosymbiosis takes its TOL. Trends Ecol. Evol. (Amst.) *23*, 268–275.

Lane, W.J., and Darst, S.A. (2010). Molecular Evolution of Multisubunit RNA Polymerases: Sequence Analysis. Journal of Molecular Biology *395*, 671–685.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947–2948.

Larkum, A.W.D., Lockhart, P.J., and Howe, C.J. (2007). Shopping for plastids. Trends in Plant Science *12*, 189–195.

Lartillot, N., and Philippe, H. (2004). A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. Mol Biol Evol *21*, 1095–1109.

Lartillot, N., and Philippe, H. (2006). Computing Bayes Factors Using Thermodynamic Integration. Syst Biol *55*, 195–207.

Lartillot, N., Brinkmann, H., and Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol *7*, S4.

Lartillot, N., Lepage, T., and Blanquart, S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics *25*, 2286–2288.

Lecompte, O., Ripp, R., Thierry, J.-C., Moras, D., and Poch, O. (2002). Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. Nucl. Acids Res. *30*, 5382–5390.

Leigh, J.W., Susko, E., Baumgartner, M., and Roger, A.J. (2008). Testing Congruence in Phylogenomic Analysis. Systematic Biology *57*, 104–115.

León-Bañares, R., González-Ballester, D., Galván, A., and Fernández, E. (2004). Transgenic microalgae as green cell-factories. Trends in Biotechnology *22*, 45–52.

Lewis, P.O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. Systematic Biology *50*, 913–925.

Lipscomb, D.L. (1985). The Eukaryotic Kingdoms. Cladistics *1*, 127–140.

Liu, Y., Richards, T.A., and Aves, S.J. (2009). Ancient diversification of eukaryotic MCM DNA replication proteins. BMC Evolutionary Biology *9*, 60.

Lopez, P., Forterre, P., and Philippe, H. (1999). The root of the tree of life in the light of the covarion model. J. Mol. Evol. *49*, 496–508.

Martin, W., Stoebe, B., Goremykin, V., Hapsmann, S., Hasegawa, M., and Kowallik, K.V. (1998). Gene transfer to the nucleus and the evolution of chloroplasts. Nature *393*, 162–165.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., and Penny, D. (2002). Evolutionary analysis of Arabidopsis, cyanobacterial, and

chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc. Natl. Acad. Sci. U.S.A. *99*, 12246–12251.

Mau, B., Newton, M.A., and Larget, B. (1999). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics *55*, 1–12.

McFadden, G.I. (2001). Primary and Secondary Endosymbiosis and the Origin of Plastids. Journal of Phycology *37*, 951–959.

McFadden, G.I., and van Dooren, G.G. (2004). Evolution: Red Algal Genome Affirms a Common Origin of All Plastids. Current Biology *14*, R514–R516.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics *21*, 1087.

Moreira, D., and Philippe, H. (2001). Sure facts and open questions about the origin and evolution of photosynthetic plastids. Res. Microbiol. *152*, 771–780.

Moreira, D., Le Guyader, H., and Philippe, H. (2000). The origin of red algae and the evolution of chloroplasts. Nature *405*, 69–72.

Murakami, K., Elmlund, H., Kalisman, N., Bushnell, D.A., Adams, C.M., Azubel, M., Elmlund, D., Levi-Kalisman, Y., Liu, X., Gibbons, B.J., et al. (2013). Architecture of an RNA Polymerase II Transcription Pre-Initiation Complex. Science *342*, 1238724–1238724.

Murakami, K.S., Masuda, S., Campbell, E.A., Muzzin, O., and Darst, S.A. (2002). Structural Basis of Transcription Initiation: An RNA Polymerase Holoenzyme-DNA Complex. Science *296*, 1285–1290.

Newton, M.A., Mau, B., and Larget, B. (1999). Markov chain Monte Carlo for the Bayesian analysis of evolutionary trees from aligned molecular sequences. Lecture Notes-Monograph Series 143–162.

Nozaki, H. (2005). A new scenario of plastid evolution: plastid primary endosymbiosis before the divergence of the "Plantae," emended. Journal of Plant Research *118*, 247–255.

Nozaki, H., Matsuzaki, M., Takahara, M., Misumi, O., Kuroiwa, H., Hasegawa, M., Shin-i, T., Kohara, Y., Ogasawara, N., and Kuroiwa, T. (2003). The phylogenetic position of red algae revealed by multiple nuclear genes from mitochondria-containing eukaryotes and an alternative hypothesis on the origin of plastids. J. Mol. Evol. *56*, 485–497.

Nozaki, H., Iseki, M., Hasegawa, M., Misawa, K., Nakada, T., Sasaki, N., and Watanabe, M. (2007). Phylogeny of primary photosynthetic eukaryotes as deduced from slowly evolving nuclear genes. Molecular Biology and Evolution *24*, 1592–1595.

Pace, N.R. (2006). Time for a change. Nature *441*, 289–289.

Palmer, J.D. (1993). A genetic rainbow of plastids. Nature *364*, 762–763.

Palmer, J.D. (2000). Molecular evolution: A single birth of all plastids? Nature *405*, 32–33.

Palmer, J.D. (2003). The symbiotic birth and spread of plastids: how many times and whodunit? Journal of Phycology *39*, 4–12.

Penny, D., McComish, B.J., Charleston, M.A., and Hendy, M.D. (2001). Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J. Mol. Evol. *53*, 711–723.

Philippe, H., and Laurent, J. (1998). How good are deep phylogenetic trees? Curr. Opin. Genet. Dev. *8*, 616–623.

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T.J., Manuel, M., Wörheide, G., and Baurain, D. (2011). Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLoS Biol *9*, e1000602.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. *26*, 1641–1650.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS ONE *5*, e9490.

Proshkina, G.M. (2006). Ancient origin, functional conservation and fast evolution of DNA-dependent RNA polymerase III. Nucleic Acids Research *34*, 3615–3624.

Pühler, G., Leffers, H., Gropp, F., Palm, P., Klenk, H.P., Lottspeich, F., Garrett, R.A., and Zillig, W. (1989). Archaebacterial DNA-dependent RNA polymerases testify to the evolution of the eukaryotic nuclear genome. Proc. Natl. Acad. Sci. U.S.A. *86*, 4569–4573.

Qiu, H., Price, D.C., Yang, E.C., Yoon, H.S., and Bhattacharya, D. (2015). Evidence of ancient genome reduction in red algae (Rhodophyta). Journal of Phycology.

Race, H.L., Herrmann, R.G., and Martin, W. (1999). Why have organelles retained genomes? Trends in Genetics *15*, 364–370.

Ragan, M.A., and Gutell, R.R. (1995). Are red algae plants? Botanical Journal of the Linnean Society *118*, 81–105.

Rannala, B., and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. Journal of Molecular Evolution *43*, 304–311.

Reyes-Prieto, A., Moustafa, A., and Bhattacharya, D. (2008). Multiple Genes of Apparent Algal Origin Suggest Ciliates May Once Have Been Photosynthetic. Current Biology *18*, 956–962.

Rivera, M.C., Jain, R., Moore, J.E., and Lake, J.A. (1998). Genomic evidence for two functionally distinct gene classes. PNAS *95*, 6239–6244.

Rodríguez-Ezpeleta, N., Brinkmann, H., Burey, S.C., Roure, B., Burger, G., Löffelhardt, W., Bohnert, H.J., Philippe, H., and Lang, B.F. (2005). Monophyly of Primary Photosynthetic Eukaryotes: Green Plants, Red Algae, and Glaucophytes. Current Biology *15*, 1325–1330.

Ronquist, F., and Deans, A.R. (2010). Bayesian Phylogenetics and Its Influence on Insect Systematics. Annual Review of Entomology *55*, 189–206.

Ronquist, F., and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics *19*, 1572–1574.

Roure, B., Baurain, D., and Philippe, H. (2013). Impact of Missing Data on Phylogenies Inferred from Empirical Phylogenomic Data Sets. Mol Biol Evol *30*, 197–214.

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

Sogin, M.L. (1990). Amplification of ribosomal RNA genes for molecular evolution studies. In PCR Protocols: A Guide to Methods and Applications, pp. 307–314.

Stamatakis, A. (2014). RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics btu033.

Steele, K.P., Holsinger, K.E., Jansen, R.K., and Taylor, D.W. (1991). Assessing the Reliability of 5S rRNA Sequence Data for Phylogenetic Analysis in Green Plants. Mol Biol Evol *8*, 240.

Stiller, J.W. (2003). WEIGHING THE EVIDENCE FOR A SINGLE ORIGIN OF PLASTIDS1. Journal of Phycology *39*, 1283–1285.

Stiller, J.W. (2007). Plastid endosymbiosis, genome evolution and the origin of green plants. Trends in Plant Science *12*, 391–396.

Stiller, J.W. (2011). Experimental design and statistical rigor in phylogenomics of horizontal and endosymbiotic gene transfer. BMC Evolutionary Biology *11*, 259.

Stiller, J.W. (2014). Toward an empirical framework for interpreting plastid evolution. J. Phycol. *50*, 462–471.

Stiller, J.W., and Hall, B.D. (1997). The origin of red algae: implications for plastid evolution. Proceedings of the National Academy of Sciences *94*, 4520–4525.

Stiller, J.W., and Hall, B.D. (1998). Sequences of the largest subunit of RNA polymerase II from two red algae and their implications for rhodophyte evolution. Journal of Phycology *34*, 857–864.

Stiller, J.W., and Hall, B.D. (1999). Long-branch attraction and the rDNA model of early eukaryotic evolution. Molecular Biology and Evolution *16*, 1270–1279.

Stiller, J.W., and Hall, B.D. (2002). Evolution of the RNA polymerase II C-terminal domain. Proceedings of the National Academy of Sciences *99*, 6091–6096.

Stiller, J.W., and Harrell, L. (2005). The largest subunit of RNA polymerase II from the Glaucocystophyta: functional constraint and short-branch exclusion in deep eukaryotic phylogeny. BMC Evolutionary Biology *5*, 71.

Stiller, J.W., and Waaland, J.R. (1993). MOLECULAR ANALYSIS REVEALS CRYPTIC DIVERSITY IN PORPHYRA (RHODOPHYTA) 1. Journal of Phycology *29*, 506–517.

Stiller, J.W., Riley, J., and Hall, B.D. (2001). Are Red Algae Plants? A Critical Evaluation of Three Key Molecular Data Sets. Journal of Molecular Evolution *52*, 527–539.

Stiller, J.W., Reel, D.C., and Johnson, J.C. (2003). A SINGLE ORIGIN OF PLASTIDS REVISITED: CONVERGENT EVOLUTION IN ORGANELLAR GENOME CONTENT1. Journal of Phycology *39*, 95–105.

Stiller, J.W., Huang, J., Ding, Q., Tian, J., and Goodwillie, C. (2009). Are algal genes in nonphotosynthetic protists evidence of historical plastid endosymbioses? BMC Genomics *10*, 484.

Stiller, W.J., Reel, D.C., and Johnson, J.C. (2002). The Case for a Single-Plastid Origin Revisited: Convergent Evolution in Organellar Gene Content. Journal of Phycology *38*, 34–34.

Sogin ML (1989) Evolution of eukaryotic microorganisms and their small subunit RNAs. Am Zool 29:487–499

Syvanen, M. (1985). Cross-species gene transfer; implications for a new theory of evolution. Journal of Theoretical Biology *112*, 333–343.

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. *56*, 564–577.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol Biol Evol *30*, 2725–2729.

Than, C., Ruths, D., Innan, H., and Nakhleh, L. (2006). Identifiability issues in phylogeny-based detection of horizontal gene transfer. In Comparative Genomics, (Springer), pp. 215–229.

Thu, Y.M., and Bielinsky, A.-K. (2013). Enigmatic roles of Mcm10 in DNA replication. Trends Biochem Sci *38*, 184–194.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y., and Martin, W. (2004). Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet *5*, 123–135.

Vannini, A., and Cramer, P. (2012). Conservation between the RNA Polymerase I, II, and III Transcription Initiation Machineries. Molecular Cell *45*, 439–446.

Williams, T.A., Foster, P.G., Nye, T.M.W., Cox, C.J., and Embley, T.M. (2012). A congruent phylogenomic signal places eukaryotes within the Archaea. Proceedings of the Royal Society of London B: Biological Sciences rspb20121795.

Wittschieben, B.Ø., Otero, G., de Bizemont, T., Fellows, J., Erdjument-Bromage, H., Ohba, R., Li, Y., Allis, C.D., Tempst, P., and Svejstrup, J.Q. (1999). A Novel Histone Acetyltransferase Is an Integral Subunit of Elongating RNA Polymerase II Holoenzyme. Molecular Cell *4*, 123–128.

Woese, C.R. (1987). Bacterial evolution. Microbiol Rev *51*, 221–271.

Woese, C.R. (2002). On the evolution of cells. Proceedings of the National Academy of Sciences *99*, 8742–8747.

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. U.S.A. *87*, 4576–4579.

Yoon, H.S., Muller, K.M., Sheath, R.G., Ott, F.D., and Bhattacharya, D. (2006). DEFINING THE MAJOR LINEAGES OF RED ALGAE (RHODOPHYTA)1. Journal of Phycology *42*, 482–492.