Title
Letters from the past: Molecular time travel to the origins of photosynthesis, RuBisCO, and the cyanobacterial phylum

Author
Shih, Patrick M.

Publication Date
2013

Letters from the past: Molecular time travel to the origins of photosynthesis, RuBisCO, and the cyanobacterial phylum

By

Patrick M Shih

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Plant Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Krishna K. Niyogi, Co-Chair
Adjunct Associate Professor Cheryl A. Kerfeld, Co-Chair
Associate Professor Chelsea D. Specht
Professor Kenneth H. Sauer

Fall 2013

Abstract

Letters from the past: Molecular time travel to the origins of photosynthesis, RuBisCO, and the cyanobacterial phylum

by

Patrick M Shih

Doctor of Philosophy in Plant Biology

University of California, Berkeley

Professor Krishna K. Niyogi, Co-Chair
Adjunct Associate Professor Cheryl A. Kerfeld, Co-Chair

The invention of oxygenic photosynthesis has forever changed the face of the Earth by producing the oxygen present in our atmosphere. It is thought that this unique metabolism arose from the bacterial phylum, Cyanobacteria. Moreover, this group has significantly contributed to eukaryotic diversity via endosymbiosis, as a cyanobacterium is considered to be the progenitor of the original plastid organelle, more commonly recognized as chloroplasts in plants. The antiquity of these events, dearth of convincing methodologies, and lack of conclusive evidence all contribute to the considerable challenges we face in understanding the timing of these major evolutionary and geological transitions. In order to address these problems, I have employed various techniques focusing on improving our understanding of the role of cyanobacteria and photosynthesis in shaping the world we have today.

Evolutionary relationships are difficult to reconstruct due to phylogenetic noise (e.g. horizontal gene transfer and homoplasy), resulting in uncertainty in our ability to build accurate phylogenetic trees. In order to address this issue, fifty-four strains of cyanobacteria were chosen for genome sequencing based on improving the phylogenetic coverage of the phylum. Not only does the diversity-driven and phylum-level approach identify many novel genes, but it also clarifies the phylogenetic placement of various cyanobacterial subclades, protein families, and endosymbiosis events.

The timing of ancient events, such as primary endosymbiosis events, has primarily been dependent on the fossil record. This is problematic as microfossils are difficult to interpret and assign to extant lineages. Conversely, molecular clock methods have been just as widely varying. We devised a new approach to increase the amount of dating information incorporated into molecular clock analyses, improving the accuracy of the predicted dates. We date the plastid and mitochondrial endosymbiosis events to approximately 900 and 1200 million years ago, respectively.

Finally, I focus on the protein evolution of a specific protein crucial to photosynthesis: RuBisCO. Here, I use ancestral sequence reconstruction methods to predict, synthesize, and characterize ancestral versions of RuBisCO. We show that ancestral RuBisCOs have lower rates of carboxylation, reflective of the high CO_2 and low O_2 Precambrian atmosphere.

Table of Contents

Chapter 1

Interpretations and implications of systematics concerning early life, photosynthesis, and cyanobacteria

Abbreviations:

TOL	Tree of Life
Gya	Billion years ago
GOE	Great Oxidation Event
GEBA	Genomic Encyclopedia of Bacteria and Archaea

Introduction

An attribute pervasive to human culture is our universal interest in understanding where we come from and the origins of life. Because scientists are no different, the scientific community has not been discouraged from speculation on the topic, much of which is just as debated as many philosophical questions. However, in spite of our inability to produce strongly persuasive evidence for many of the hypotheses concerning the origins of life, many of these ideas are commonly accepted as fact and can be routinely found in textbooks. Inherently, the nature of these ancient events cannot be definitively proven without the use of a time machine, and thus the scientific community must do its best to interpret the evidence using the current methods available.

One particular bacterial phylum that has played a role in many hypotheses of early life is that of Cyanobacteria. This phylum is intimately intertwined with a unique and defining feature of Earth: the presence of both water and atmospheric oxygen. It is believed that the oxygen in our atmosphere originated from the biological process of oxygenic photosynthesis, where electrons are stripped from water using absorbed light energy, resulting in the release of molecular oxygen. Photosynthesis converts light energy into chemical energy, which is then used to generate carbohydrates for autotrophic organisms. Oxygenic photosynthesis is thought to have originated within the cyanobacterial phylum, as this is the only extant group of prokaryotes which can carry out the process. The drastic change in atmospheric content has profoundly affected the Tree of Life (TOL), as the arrival of aerobic respiration most likely succeeded the oxygenated atmosphere. The presence of aerobic respiration in all three domains of life (Bacteria, Archaea, and Eukarya) strongly suggests that this is indeed an ancient metabolic process, attesting to the direct effect of oxygen and cyanobacteria on all branches of life.

Another profound way the cyanobacterial phylum has affected the course of the TOL is through endosymbiosis. Eukaryotic phototrophs, such as plants, can trace their ability to perform photosynthesis to a cyanobacterial origin due to an elaborate history of endosymbiosis events. Plastid endosymbiosis has been a large driver of eukaryotic diversity, as whole supergroups within the eukaryotic TOL, including Archaeplastida and Chromalveolata, have successfully diversified due to endosymbiosis events. The functions of plastid endosymbionts are correspondingly diverse. Although the advent of eukaryotic photosynthesis is one of the most important and widespread outcomes of plastid endosymbiosis, some species have lost this ability but have retained the plastid organelle. For example, *Plasmodium*, the apicomplexan that is responsible for causing the disease malaria, cannot survive without its apicoplast, a non-photosynthetic plastid.

As cyanobacteria have clearly affected evolution on many branches along the TOL, it has also been suggested that they are near the base of the TOL as one of the oldest forms of life. This is commonly mentioned in the context of their ancient geological role in shaping the Earth's biogeochemical landscape using oxygenic photosynthesis. Microfossil records have suggested that cyanobacteria were present near the time of origin of life 3.5 billion years ago (Gya) (1).

Many of these events occurred well over 1 Gya, and the dating and reconstruction of these evolutionary relationships are inherently difficult given their antiquity. In this chapter, I present the evidence supporting most of these ideas and events, but more importantly, the ambiguity surrounding the interpretations of these key events in the history of life on Earth. I then focus on potential avenues of research that may help improve our understanding of the evolutionary history of cyanobacteria and their role in shaping Earth over billions of years.

Controversy with the interpretation of non-molecular systematic approaches

Simply put, we don't know when life began. Theories based on everything from a RNA world (2) to extraterrestrial origins have been put forward. However, actual hard evidence can only be obtained from the fossil record. Fossils are our only glimpses into the past, and it would be an understatement to say we have an incomplete fossil record of microbial life. Historically, cyanobacterial-like fossils and biomarkers have been implicated as evidence of the ancient presence of cyanobacteria near the base of the TOL. However, issues with convergent evolution – or homoplasy – bias and skew our understanding of events at these geological timescales.

The oldest microfossils

Earth is 4.6 billion year old. It is largely assumed that life could not have begun prior to the hypothetical Late Heavy Bombardment of Earth 3.9 Gya (3), when a significant number of cometary collisions in a short period time would have made the existence of life impossible. However, it has been proposed that life on Earth began ~3.5 Gya (1, 4), based on microfossil records. Most notably, structures resembling filamentous cyanobacteria from Apex cherts of northwestern Australian Pilbara Craton have been heavily debated (1). Conversely, the biological origin and authenticity of these microfossils have been interpreted as artifactual, as they may have formed from amorphous graphite (5, 6).

The interpretation of these fossils as ancestors of extant trichomic cyanobacteria (1) has had great implications on the speculation of the Archean role of cyanobacteria and oxygenic photosynthesis. The conjecture that these fossils are the remnants of oxygen-producing cyanobacteria may be presumptuous for two reasons. First, the highly complex process of oxygenic photosynthesis involves the coordination of more than 100 proteins (7), several multi-subunit complexes, and the incorporation of many different cofactors. It is difficult to imagine that soon after the de novo origin of life, one of the most complex metabolic processes arose within 400 million years; this line of logic may suggest an extraterrestrial origin of life. Second, given the well-described presence of anoxygenic photosynthesis in various cyanobacteria (8, 9), which has been described to use hydrogen sulfide as an electron donor, it is equally likely that the initial forms of photosynthesis in cyanobacteria were anoxygenic. This is consistent with the theory that early life arose near hydrothermal vents which produce hydrogen sulfide (10). Early environments containing hydrogen sulfide may have been where photosynthesis originated.

It is important to emphasize the ambiguity of interpreting these Archean fossils as cyanobacteria. If of biological origin, the filamentous nature of these microfossils is difficult to definitively assign to cyanobacteria. Given the unknown but most likely large amount of extinction that has occurred during the supposed 3.5 billion years of life, there have most likely been filamentous life forms that have arisen other than cyanobacteria; other extant non-cyanobacterial filamentous prokaryotes exist in the present day- *Bacillus*, *Frankia*, *Streptomyces*, and *Chloroflexus* to name just a few genera that contain filamentous species.

The evidence presented by putative microfossils makes it difficult to strongly argue for or against the cyanobacterial nature of these fossils given their age and the incompleteness of the fossil record. For these reasons, I would like to emphasize the uncertainty in the dating of the oldest cyanobacteria, let alone the oldest forms of life.

Debunking hopanoid biomarkers

Besides morphological fossils, indirect molecular fossils have been proposed to indicate the presence of specific taxa. The identification of complex organic biological molecules (biomarkers) in sediments is used to date indirectly the existence of certain organisms. Classically, hopanoids have been used to indicate the presence of cyanobacteria, because hopanoid biosynthesis was thought to be specific to the cyanobacterial phylum (11). This aligned well with the hypothesis that cyanobacteria were present prior to the Great Oxidation Event (GOE) because methylhopane hydrocarbon derivatives could be detected in sediments dating to more than 2.5 Gya (11, 12). This appealing second line of evidence, independent of the morphological cyanobacterial-like fossils, provided further evidence for the existence of cyanobacteria prior to the GOE. In all, a clear story was seemingly emerging in which cyanobacteria and the advent of oxygenic photosynthesis occurred during the early Paleoarchean (3.6-3.2 Gya), thus providing the biological catalyst necessary for the GOE.

However attractive the story, one crucial assumption to this theory was the specificity of hopanoid biosynthesis to cyanobacteria; however, the enzymes involved in the pathway were not yet known. It was later shown that an alpha-proteobacterium, *Rhodopseudomonas palustris*, has the ability to produce significant amounts of 2-methylhopanes (13). Subsequently, the S-adenosylmethionine methylase necessary for hopanoid methylation was identified (14), and more importantly, the methylase could be found in a variety of alpha-proteobacteria and an acidobacterium species, as well as cyanobacteria. Thus, the origin of methylhopanes could not be specifically assigned to cyanobacteria, and therefore, its use as a biomarker for cyanobacteria and oxygenic photosynthesis could not be fully supported.

The use of biomarkers highlights the dramatic influence homoplasy plays in the interpretation of the fossil record. Macroscopic fossils are scored based on many specific morphological characteristics, which can be used to infer their evolutionary relationship to other fossils or extant lineages. However, biomarkers can be thought of as a fossil containing only one "fossil characteristic," the presence or absence of the biomarker. With one very simple homoplastic event (whether it be due to convergent evolution, horizontal gene transfer, or a common ancient origin of the hopanoid biosynthesis between the groups), the interpretation of the evolutionary history breaks down. The comparatively small amount of information that is available with the use of fossils emphasizes a strength of molecular systematic methods which have the luxury of including many positions (characters) from either nucleotide or amino acid sequences, as well as more sophisticated models.

Morphological homoplasy and extinction within the cyanobacterial phylum

The prolific morphological diversity displayed in the cyanobacterial phylum (and the cyanobacterial fossil record) provides yet another potential misinterpretation of the fossil record due to homoplasy. The broad range of cyanobacterial morphological characteristics has enabled the assignment of many Precambrian cyanobacterial-like microfossils to the phylum.

Cyanobacterial taxonomy has classically been divided into five morphologically distinct subsections (15). Subsection I (Chroococcales) consists of unicellular cells that divide through binary fission. Subsection II (Pleurocapsales) consists of unicellular cells that can divide through multiple fissions and can divide in different planes to produce cells called baeocytes. Subsection III (Oscillatoriales) cyanobacteria are filamentous. Subsection IV (Nostocales) and Subsection V (Stigonematales) have the unique ability to undergo cellular differentiation; these cells can form heterocysts (cells necessary to carry out nitrogen fixation), akinetes (dormant cells, analogous to spores), and hormogonia (cell types used in dispersal and niche colonization). Members of Subsection V differ from those of Subsection IV in their unique ability to produce branched filaments.

The cyanobacterial phylum presents an archetypal history of systematics; prior to the use of molecular markers for phylogenetic studies, morphological characters were all that was available. With the advent of DNA sequencing and later genome sequencing, the classical taxonomy broke down in several places within the cyanobacterial phylum. Notably, various phylogenetic studies have conclusively shown that the Chroococcales, Pleurocapsales, and Oscillatoriales subsections are all polyphyletic (16-20). The ability to undergo cellular differentiation (Subsection IV and V) appears to have only arisen once within the cyanobacterial clade, as members of these two subsections form a monophyletic clade (16, 17). Subsection V cyanobacteria also form a monophyletic clade (16, 17). Next generation sequencing technologies will enable more than just the glimpse of evolution of single markers, such as the highly conserved 16S ribosomal RNA gene universally found in bacteria. Genome sequencing from a broad range of morphologically diverse cyanobacteria will enable a more complete understanding of the genetic basis underlying these various morphological characteristics in this diverse phylum.

The large amount of morphological homoplasy within the cyanobacterial phylum is a source of great concern when assigning microfossils as cyanobacterial organisms and interpreting the timing of the emergence of various subclades. This is most recently highlighted by studies that attempt to address the dating and origin of multicellularity in

Cyanobacteria (17, 21). Schirrmeister et al present an exhaustive 16S phylogenetic analysis, which is used to predict ancestral character states based on morphology (17). They conclude that the large majority of the phylum has descended from a 'multicellular' ancestor and infer that multicellularity must have arisen near the time of the GOE, 2.4 Gya. Schirrmeister et al define 'multicellularity' to encompass basic filamentous forms and more complex morphologies. Given the polyphyletic and convergent evolution of the filamentous phenotype (e.g. Oscillatoriales lineages), this study highlights the potential pitfalls in the interpretation of filamentous cyanobacterial-like microfossils. Furthermore, it also highlights the possible homoplasy that could very easily lead to the incorrect assignment of extant lineages to extinct lineages.

Another less-discussed issue is the characteristics of and distinction between stem and crown group cyanobacteria. Crown group cyanobacteria include all present-day (extant) lineages, including all their ancestors back to their most recent common ancestor. Stem group cyanobacteria consist of the ancestral lineages that fit outside of the crown-group distinction and have already gone extinct. The cyanobacterial species existing >2.5 Gya may have been responsible for the GOE and their direct descendants may be today's extant lineages, thus making the ancestors that caused the GOE part of the crown group (Figure 1, Scenario 1). However, it is also possible that these ancient cyanobacterial species may have gone extinct, and therefore these extinct lineages are not part of the crown group cyanobacteria, but rather are stem group lineages (Figure 1, Scenario 2).

It is imperative to make clear that it is impossible to definitively know if the cyanobacteria responsible for the GOE are part of stem group or crown group cyanobacteria. The inherent uncertainty in the nature of these cyanobacterial ancestors highlights the level of discretion and skeptical realism one must have when interpreting events happening billions of years ago. The implications of stem group cyanobacteria potentially having different metabolic properties, such as the presence or absence of oxygenic photosynthesis, greatly affect the hypotheses put forward, as these assumptions have been the crux of many molecular clock analyses. Recently, a molecular clock analysis based on cyanobacterial 16S phylogeny (21) used the GOE as a dating calibration. By doing so, the assumption was placed that crown-group cyanobacteria were present during the GOE. This is disturbing because of the large gaps in our understanding of the cyanobacterial fossil records. Given the controversy surrounding the main fossil evidence for an Archean origin of oxygenic photosynthesis and cyanobacteria, mere anecdotes and weakly supported hypotheses have been put forward concerning the timing and characteristics of possible cyanobacteria existing in this time period.

One key feature that is speculated upon is the composition of photosystems in stem cyanobacteria, as it is not known if Type II and Type I coexisted in one organism first and were subsequently transferred to different lineages to only contain one of two photosystems (7). Conversely, each photosystem may have arisen independently of one another, and both were transferred into one organism, which gave rise to oxygenic photosynthesis in cyanobacteria (22, 23). In this second scenario, stem cyanobacteria could not perform oxygenic photosynthesis, and would most likely perform some sort of anoxygenic photosynthesis, as the role of anoxygenic photosynthesis has been proposed to be crucial for Proterozoic biogeochemical cycling in the oceans (24). Given the large geochemical changes in the atmosphere and oceans from the Archean through the Proterozoic Eons, one could reasonably assume large amounts of extinction of lineages that were unable to adapt to the transition from anaerobic to aerobic or decreasing sulfide levels (24). Thus, it is not out of the question that crown cyanobacteria – and perhaps oxygenic photosynthesis – arose subsequent to the extinct microfossil lineages that are dated upwards of 2 Gya. It may equally be justifiable to assume that oxygenic photosynthesis arose in stem-group cyanobacteria that subsequently went extinct, and crown-group lineages actually arose after the GOE.

Finally, the use of the GOE as a calibration point for crown cyanobacteria is another example of the ramifications of the biomarker and Archean microfossil studies which concluded that the rise of oxygenic photosynthesis occurred via cyanobacteria more than 2.5 Gya. Other studies have incorporated the use of the same calibrations (25); however, it will be important for future phylogenetic efforts addressing cyanobacterial evolution and multicellularity to be more cautious when choosing calibration points. A good example is demonstrated in the molecular clock analyses dating plastid endosymbiosis events by Yoon et al (26). This study includes and omits a calibration of the contentious red algal fossil, *Bangiomorpha*, dated to 1.2 Gya (27), in two separate molecular clock analyses. The results present two dates for the red-green algal divergence, 1.452 and 1.156 Gya, with and without the calibration respectively. These types of studies are more informative as they directly address the inherent ambiguity of Precambrian molecular clock studies.

Issues and potential improvements on molecular systematic approaches

Given the difficulties of using non-molecular systematic approaches to interpret the dating of Precambrian events relating to the rise of cyanobacteria and oxygenic photosynthesis, molecular systematic methods provide alternative techniques to answer the same questions. Although some of these studies may be inadequate to answer evolutionary questions on this timescale given the current data available, a combination of increased coverage as well as improvements in current methods will drastically decrease our reliance on purely fossil-based dating of biological events.

Improving phylogenetic coverage

Increased sampling reduces phylogenetic error (28, 29). If we could theoretically sample every living organism that ever existed (including extinct lineages), one can imagine the ease in reconstructing all phylogenetic relationships. The harsh truth is we have access only to extant lineages; even with these organisms, we have a poor sampling (albeit slowly improving with the advent of newer sequencing technologies) of all present-day lineages, especially those of microbial life. Thus, a critical way to advance our understanding of evolution in general is to expand our knowledge of all extant taxa.

Various efforts have focused on addressing the issue of sampling. Metagenomic studies of as many environments as possible began the sequencing of many unculturable organisms. The Genomic Encyclopedia of Bacteria and Archaea (GEBA) was initially started as an initiative to sequence genomes of microbial organisms that were found in very sparsely covered regions of the TOL with no closely related sequenced genome (30). 16S databases devoted to inferring the evolutionary relationships of all organisms have also played a key role (31).

With an expanded, or even perfect coverage, of extant cyanobacterial species, we would be faced with biases stemming from homoplasy, which could potentially distort our interpretation of the phylogenetic relationships between taxa. It is important to recognize these potential issues when addressing ancient divergence events, because small biases will be compounded over long stretches of time resulting in misleading interpretations. Importantly, strong phylogenetic signal is necessary for inferring ancient events (32). Biases from molecular markers can stem from potentially fewer alternative states/characters leading to homoplasy and skewing the phylogenetic signal. Such phenomena have been observed in various *Prochlorococcus* lineages, where genome sequencing has revealed huge ranges in GC content between different species, ranging from 29%-55% GC content (33). This means that some organisms are more limited in base substitutions than others, leading to different rates of substitution in different lineages. Other facts may also lead to unequal rates of evolution that skew phylogenetic analysis. Symbiotic lifestyles have been shown to affect rates of evolution (34, 35), for example, and many Nostocales are symbionts of various plant lineages (36). Cyanobacteria have even been found to be symbionts of marine algae and diatoms (37, 38). Moreover, variable cladogenesis and extinction events across a tree may contribute to the resolvability of the phylogeny (39). Sudden radiation events, such as the one speculated near the base of the cyanobacterial phylogeny (19), could contribute to these variable cladogenesis events. Despite the many potential ways phylogenetic noise can be introduced by homoplasy, expanding the distribution of taxa studied will help address these issues.

Newer sequencing technologies, such as single cell genomics, will facilitate the sequencing of phylogenetically distinct lineages of cyanobacteria to expand our understanding of the phylum. The first sequenced genome of a cyanobacterium, *Synechocystis* sp. strain PCC 6803, was published in 1996 (40). To date, there are more than 130 sequenced cyanobacterial genomes. One elusive but widespread putative cyanobacterium has been identified from multiple 16S environmental surveys and metagenomic studies, ranging from cow rumen, mouse gut microbiome, landfill, bioreactor, and hot spring samples (41). This putative cyanobacterium is more deeply branching than *Gloeobacter violaceus* PCC 7421, which is thought to be the most deeply-branching characterized cyanobacterium of the phylum, based on phylogenetic analysis and its simpler (and assumed more primitive) membrane structure/ morphology (42). It has been proposed that that the putative cyanobacterium may be non-photosynthetic, based on the fact that it is found in various microbiomes which would not have access to light (41). This is not out of the realm of possibility; it was recently discovered that a symbiotic, non-photosystem II-containing cyanobacterium was recently identified (37, 43).

Because only extant sequence information is available, our interpretation of evolutionary history is inherently flawed and biased. However, in comparison to the fossil records, the nascent field of molecular clock analyses presents a promising alternative to purely paleontological and geological approaches to dating ancient events along the TOL.

Molecular Clock methods

Molecular clock methods make use of rates of evolution and fossil calibration to date divergence events on a phylogenetic tree (chronogram). The bulk of molecular clock studies have focused on events within the Phanerozoic Eon, due to a combination of reliable macroscopic fossil constraints and more reliable accuracy in estimating divergences. It becomes increasingly difficult to accurately estimate more ancient divergences due to a lack of understanding the rates of evolution further back in time. However, improvements in molecular clock studies may provide a means to dating ancient events, such as the rise of cyanobacteria, the origin of oxygenic photosynthesis, and the TOL. A significant amount of attention has been focused on dating the origins of metazoans, plants, and eukaryotes; however, very little has been done with dating microbial (including cyanobacterial) divergence events, most likely because of the difficulty in finding microfossils that can confidently be attributed to an extant lineage.

The concept of a strict molecular clock was first introduced by Zuckerkandl and Pauling in 1965 (44). Although the pervasive problem of non-clocklike rates of evolution was well-known, it was not until fewer than two decades ago that relaxed clock methods were implemented which enabled rate variation across branches on a phylogenetic tree (45, 46). Even with these more sophisticated relaxed clock methods, it has been proposed that there may not be enough phylogenetic signal to date the TOL, let alone reconstruct the phylogenetic relationships at the deepest nodes of the tree (47, 48). Although this may be good reason to abandon the quest for a molecular clock solution to dating divergences, the only other viable alternative would be to rely solely on highly controversial microfossils. Thus, it will be crucial in the future to develop new algorithms and methods to improve estimation of divergence events, especially in Precambrian eras.

Ancestral Reconstruction methods

Once a well-resolved phylogeny is established, ancestral reconstruction methods can be used to infer the characteristics of internal nodes within the phylogeny, providing a more thorough understanding of the evolutionary history. These methods provide one key advantage over paleontological methods, the ability to infer characteristics that cannot be preserved by the fossil record. One example of this is the characterization of traits which cannot be fossilized, such as pigmentation patterns of ancestral *Conus* shells (49). Ancestral character reconstruction methods have been used to infer the morphological characteristics of ancestral cyanobacteria (17). However, the convergence of the Oscillatoriales and Pleurocapsales morphologies within the cyanobacterial phylum complicates the interpretation of the reconstructions of internal nodes. Despite the possibility of homoplasy, these are the only methods available to infer the properties of ancestral species.

Moreover, ancestral reconstruction methods can infer not only simple morphological characteristics, but also ancestral molecular sequences. In contrast to phylogenetic or molecular clock analyses, ancestral sequence reconstruction allows us not only to hypothesize the properties of ancient molecular sequences, but also to ‘resurrect’ and test them in the present day. Ancestral sequence reconstruction studies have been used to infer the temperature of paleoenvironments (50, 51) and enzymatic (52, 53), biochemical (54, 55), and structural properties (56) of ancestral proteins. Although the timing of major divergences within the cyanobacterial phylum are controversial, the properties of resurrected proteins from deep nodes may enable us to infer the likely environment of these proteins, thus allowing us to extrapolate potential geological windows in which these divergences occurred.

The use of comparative genomic methods has also been utilized to infer the ancestral genome composition of ancient organisms. Various studies have identified sets of core and signature genes of cyanobacteria (7, 57, 58) and phototrophic bacteria (59), and thus infer either the core set of genes necessary for a phototrophic organism or the set of genes most likely found in the last common ancestor of these various species. It is assumed that all extant cyanobacteria (except UCYN-A) can perform oxygenic photosynthesis. However, it is still not clear when this common ancestor existed, because crown cyanobacteria may have emerged much later than fossil records suggest.

Finally, it is worth noting that the only alternative to directly study ancient organisms or proteins other than ancestral reconstruction methods is to actually find samples from those time periods. Studies have described the resurrection of ancient samples that have been frozen for long periods of time (60, 61). However, none come close to the geological timescale necessary to address Precambrian events.

Conclusion

The questions surrounding the origins of oxygenic photosynthesis and the rise of cyanobacteria are incredibly difficult to answer. However, this has not and should not prevent researchers from continually re-evaluating prior studies and developing novel ways to address the issues at hand. Although the daunting problems caused by homoplasy are quite pervasive with such ancient evolutionary relationships, it is crucial to be candid about these issues and present data accordingly, as the characterization of Precambrian events should be taken with much more than a grain of salt. Although molecular systematic methods are still controversial, they are likely to improve based on advances in algorithms and methods. With increased confidence in these methods, we will be able to consider them as an independent means by which to validate the equally controversial interpretations of the fossil record.

References:

1. Schopf JW (1993) Microfossils of the Early Archean Apex Chert: New Evidence of the Antiquity of Life. *Science* 260(5108):640-646.
2. Gilbert W (1986) Origin of life: The RNA world. *Nature* 319(6055):618-618.
3. Cohen BA, Swindle TD, & Kring DA (2000) Support for the Lunar Cataclysm Hypothesis from Lunar Meteorite Impact Melt Ages. *Science* 290(5497):1754-1756.
4. Walter MR, Buick R, & Dunlop JSR (1980) Stromatolites 3,400-3,500 Myr old from the North Pole area, Western Australia. *Nature* 284(5755):443-445.
5. Brasier MD*, et al.* (2002) Questioning the evidence for Earth's oldest fossils. *Nature* 416(6876):76-81.
6. Brasier MD*, et al.* (2005) Critical testing of Earth's oldest putative fossil assemblage from the 3.5Ga Apex chert, Chinaman Creek, Western Australia. *Precambrian Research* 140(1–2):55-102.

7. Mulkidjanian AY*, et al.* (2006) The cyanobacterial genome core and the origin of photosynthesis. *Proc Natl Acad Sci* 103(35):13126-13131.

8. Padan E (1979) Facultative Anoxygenic Photosynthesis in Cyanobacteria. *Annu Rev Plant Physiol* 30(1):27-40.

9. Arieli B, Shahak Y, Taglicht D, Hauska G, & Padan E (1994) Purification and characterization of sulfide-quinone reductase, a novel enzyme driving anoxygenic photosynthesis in Oscillatoria limnetica. *J Biol Chem* 269(8):5705-5711.

10. Lane N & Martin WF (2012) The Origin of Membrane Bioenergetics. *Cell* 151(7):1406-1416.

11. Summons RE, Jahnke LL, Hope JM, & Logan GA (1999) 2-Methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis. *Nature* 400(6744):554-557.

12. Brocks JJ, Logan GA, Buick R, & Summons RE (1999) Archean Molecular Fossils and the Early Rise of Eukaryotes. *Science* 285(5430):1033-1036.

13. Rashby SE, Sessions AL, Summons RE, & Newman DK (2007) Biosynthesis of 2-methylbacteriohopanepolyols by an anoxygenic phototroph. *Proc Natl Acad Sci* 104(38):15099-15104.

14. Welander PV, Coleman ML, Sessions AL, Summons RE, & Newman DK (2010) Identification of a methylase required for 2-methylhopanoid production and implications for the interpretation of sedimentary hopanes. *Proc Natl Acad Sci* 107(19):8537-8542.

15. Rippka R, Deruelles J, Waterbury JB, Herdman M, & Stanier RY (1979) Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *J Gen Microbiol* 111:1-61.

16. Turner S, Pryer KM, Miao VPW, & Palmer JD (1999) Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. *J Eukaryot Microbiol* 46(4):327-338.

17. Schirrmeister B, Antonelli A, & Bagheri H (2011) The origin of multicellularity in cyanobacteria. *BMC Evol Biol* 11(1):45.

18. Tomitani A, Knoll AH, Cavanaugh CM, & Ohno T (2006) The evolutionary diversification of cyanobacteria: Molecular–phylogenetic and paleontological perspectives. *Proc Natl Acad Sci* 103(14):5442-5447.

19. Criscuolo A & Gribaldo S (2011) Large-scale phylogenomic analyses indicate a deep origin of primary plastids within Cyanobacteria. *Mol Biol Evol* 28(11):3019-3032.

20. Shih PM*, et al.* (2013) Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc Natl Acad Sci* 110(3):1053-1058.

21. Schirrmeister BE, de Vos JM, Antonelli A, & Bagheri HC (2013) Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event. *Proc Natl Acad Sci* 110(5):1791-1796.

22. Hohmann-Marriott MF & Blankenship RE (2011) Evolution of Photosynthesis. *Annu Rev Plant Biol* 62(1):515-548.

23. Blankenship RE (2010) Early Evolution of Photosynthesis. *Plant Physiol* 154(2):434-438.

24. Johnston DT, Wolfe-Simon F, Pearson A, & Knoll AH (2009) Anoxygenic photosynthesis modulated Proterozoic oxygen and sustained Earth's middle age. *Proc Natl Acad Sci* 106(40):16925-16929.
25. Falcon LI, Magallon S, & Castillo A (2010) Dating the cyanobacterial ancestor of the chloroplast. *ISME J* 4(6):777-783.
26. Yoon HS, Hackett JD, Ciniglia C, Pinto G, & Bhattacharya D (2004) A Molecular Timeline for the Origin of Photosynthetic Eukaryotes. *Mol Biol Evol* 21(5):809-818.
27. Butterfield NJ, Knoll AH, & Swett K (1990) A bangiophyte red alga from the Proterozoic of arctic Canada. *Science* 250:104-107.
28. Zwickl DJ & Hillis DM (2002) Increased taxon sampling greatly reduces phylogenetic error. *Sys Biol* 51(4):588-598.
29. Hillis DM (1996) Inferring complex phylogenies. *Nature* 383(6596):130-131.
30. Wu D*, et al.* (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462(7276):1056-1060.
31. DeSantis TZ*, et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72(7):5069-5072.
32. Salichos L & Rokas A (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497(7449):327-331.
33. Partensky F, Hess WR, & Vaulot D (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol Mol Biol Rev* 63(1):106-127.
34. Lutzoni F & Pagel M (1997) Accelerated evolution as a consequence of transitions to mutualism. *Proc Natl Acad Sci* 94(21):11422-11427.
35. Degnan PH, Lazarus AB, Brock CD, & Wernegreen JJ (2004) Host–Symbiont Stability and Fast Evolutionary Rates in an Ant–Bacterium Association: Cospeciation of Camponotus Species and Their Endosymbionts, Candidatus Blochmannia. *Syst Biol* 53(1):95-110.
36. Rai AN, Soderback E, & Bergman B (2000) Cyanobacterium–plant symbioses. *New Phytol* 147(3):449-481.
37. Thompson AW*, et al.* (2012) Unicellular Cyanobacterium Symbiotic with a Single-Celled Eukaryotic Alga. *Science* 337(6101):1546-1550.
38. Hilton JA*, et al.* (2013) Genomic deletions disrupt nitrogen metabolism pathways of a cyanobacterial diatom symbiont. *Nat Commun* 4:1767.
39. Rokas A & Carroll SB (2006) Bushes in the Tree of Life. *PLoS Biol* 4(11):e352.
40. Kaneko T*, et al.* (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC 6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* 3:109-136.
41. Ley RE*, et al.* (2005) Obesity alters gut microbial ecology. *Proc Natl Acad Sci* 102(31):11070-11075.
42. Nakamura Y*, et al.* (2003) Complete Genome Structure of Gloeobacter violaceus PCC 7421, a Cyanobacterium that Lacks Thylakoids. *DNA Res* 10(4):137-145.
43. Zehr JP*, et al.* (2008) Globally distributed uncultivated oceanic N2-fixing cyanobacteria lack oxygenic photosystem II. *Science* 322(5904):1110-1112.

44. Zuckerkandl E & Pauling L (1965) Molecules as documents of evolutionary history. *J Theor Biol* 8(2):357-366.

45. Rambaut A & Bromham L (1998) Estimating divergence dates from molecular sequences. *Mol Biol Evol* 15(4):442-448.

46. Sanderson MJ (1997) A Nonparametric Approach to Estimating Divergence Times in the Absence of Rate Constancy. *Mol Biol Evol* 14(12):1218.

47. Roger AJ & Hug LA (2006) The origin and diversification of eukaryotes: problems with molecular phylogenetics and molecular clock estimation. *Philos Trans R Soc Lond B Biol Sci* 361(1470):1039-1054.

48. Philippe H & Forterre P (1999) The Rooting of the Universal Tree of Life Is Not Reliable. *J Mol Evol* 49(4):509-523.

49. Gong Z*, et al.* (2012) Evolution of patterns on Conus shells. *Proc Natl Acad Sci* 109(5):E234-241.

50. Boussau B, Blanquart S, Necsulea A, Lartillot N, & Gouy M (2008) Parallel adaptations to high temperatures in the Archaean eon. *Nature* 456(7224):942-945.

51. Gaucher EA, Govindarajan S, & Ganesh OK (2008) Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* 451(7179):704-707.

52. Perez-Jimenez R*, et al.* (2011) Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat Struct Mol Biol* 18(5):592-596.

53. Thomson JM*, et al.* (2005) Resurrecting ancestral alcohol dehydrogenases from yeast. *Nat Genet* 37(6):630-635.

54. Ugalde JA, Chang BSW, & Matz MV (2004) Evolution of Coral Pigments Recreated. *Science* 305(5689):1433-1433.

55. Bridgham JT, Carroll SM, & Thornton JW (2006) Evolution of Hormone-Receptor Complexity by Molecular Exploitation. *Science* 312(5770):97-101.

56. Ortlund EA, Bridgham JT, Redinbo MR, & Thornton JW (2007) Crystal Structure of an Ancient Protein: Evolution by Conformational Epistasis. *Science* 317(5844):1544-1548.

57. Shi T & Falkowski PG (2008) Genome evolution in cyanobacteria: The stable core and the variable shell. *Proc Natl Acad Sci* 105(7):2510-2515.

58. Swingley WD, Blankenship RE, & Raymond J (2008) Integrating Markov Clustering and Molecular Phylogenetics to Reconstruct the Cyanobacterial Species Tree from Conserved Protein Families. *Mol Biol Evol* 25(4):643-654.

59. Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, & Blankenship RE (2002) Whole-Genome Analysis of Photosynthetic Prokaryotes. *Science* 298(5598):1616-1620.

60. Yashina S*, et al.* (2012) Regeneration of whole fertile plants from 30,000-y-old fruit tissue buried in Siberian permafrost. *Proc Natl Acad Sci* 109(10):4008-4013.

61. Murray AE*, et al.* (2012) Microbial life at −13 °C in the brine of an ice-sealed Antarctic lake. *Proc Natl Acad Sci* 109(50):20626-20631.

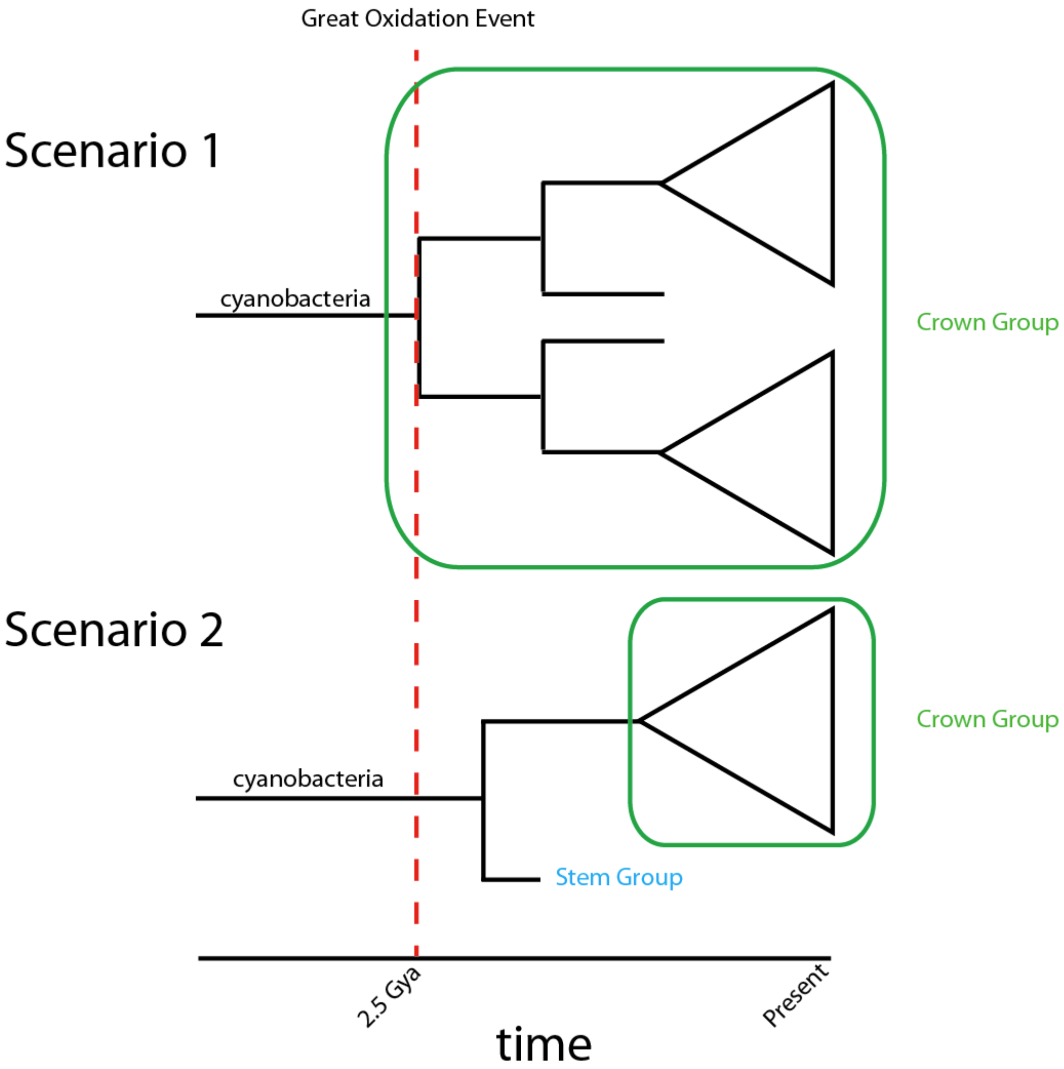


Figure 1. Potential scenarios explaining the relationship of ancestral cyanobacterial lineages responsible for the Great Oxidation Event to crown group cyanobacteria. Lineages/branches that do not make it to the present day have gone extinct. Red dashed line indicates the Great Oxidation Event (2.5 Gya). Crown group versus stem groups are labeled.

Chapter 2

Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing

Abstract

The cyanobacterial phylum encompasses oxygenic photosynthetic prokaryotes of a great breadth of morphologies and ecologies; they play key roles in global carbon and nitrogen cycles. The chloroplasts of all photosynthetic eukaryotes can trace their ancestry to cyanobacteria. Cyanobacteria also attract considerable interest as platforms for "green" biotechnology and biofuels. To explore the molecular basis of their different phenotypes and biochemical capabilities, we sequenced the genomes of 54 phylogenetically and phenotypically diverse cyanobacterial strains. Comparison of cyanobacterial genomes reveals the molecular basis for many aspects of cyanobacterial ecophysiological diversity, as well as the convergence of complex morphologies without the acquisition of novel proteins. This phylum-wide study highlights the benefits of diversity-driven genome sequencing, identifying over 21,000 cyanobacterial proteins with no detectable similarity to known proteins and foregrounds the diversity of light-harvesting proteins and gene clusters for secondary metabolite biosynthesis. Also, our results provide insight into the distribution of genes of cyanobacterial origin in eukaryotic nuclear genomes. Moreover, this study doubles both the amount and the phylogenetic diversity of cyanobacterial genome sequence data. Given the exponentially growing number of sequenced genomes, this diversity-driven study demonstrates the perspective gained by comparing disparate yet related genomes in a phylum-wide context and the insights that are gained from it.

Preface

The contents of this chapter are based on the following publication:

Patrick M. Shih, Dongying Wu, Amel Latifi, Seth D. Axen, David P. Fewer, Emmanuel Talla, Alexandra Calteau, Fei Cai, Nicole Tandeau de Marsac, Rosmarie Rippka, Michael Herdman, Kaarina Sivonen, Therese Coursin, Thierry Laurent, Lynne Goodwin, Matt Nolan, Karen W. Davenport, Cliff S. Han, Edward M. Rubin, Jonathan A. Eisen, Tanja Woyke, Muriel Gugger, Cheryl A. Kerfeld (2013). Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, January 15;110(3):1053-8, doi: 10.1073/pnas.1217107110.

My contributions to this work included designing experiments, bioinformatic analysis, interpreting data, and writing the manuscript. Dr. Dongying Wu, Dr. Amel Latifi, Seth D. Axen, Dr. David P. Fewer, Dr. Emmanuel Talla, Dr. Alexandra Calteau, Dr. Fei Cai, Dr. Nicole Tandeau de Marsac, Dr. Rosmarie Rippka, Dr. Michael Herdman, Dr. Kaarina Sivonen, Therese Coursin, Thierry Laurent, Lynne Goodwin, Matt Nolan, Karen W. Davenport, Cliff S. Han, Dr. Edward M. Rubin, Dr. Jonathan A. Eisen, Dr. Tanja Woyke, and Dr. Muriel Gugger contributed either to the work in this chapter or general project coordination and management. More specific contributions are listed in the 'Author Contribution' section below. Supplemental information for this chapter can be found in Appendix A.

Abbreviations:

GEBA	Genomic Encyclopedia of Bacteria and Archaea
CRISPRs	Clustered Regularly Interspaced Short Palindromic Repeats
EGT	Endosymbiotic Gene Transfer
LHC	Light Harvesting Complex
CBP	Chlorophyll Binding Protein
PSI	Photosystem I
NRPS	Non-ribosomal peptide synthase
PKS	Polyketide synthase

Introduction

The *Cyanobacteria* are one of the most diverse and widely distributed phyla of bacteria. Among photosynthetic prokaryotes, they uniquely have the ability to perform oxygenic photosynthesis; they are considered to be the progenitor of the chloroplast, the photosynthetic organelle found in eukaryotes. Cyanobacteria contribute greatly to global primary production, fixing a substantial amount of biologically available carbon, especially in nutrient-limited environmental niches, from oligotrophic marine surfaces to desert crusts (1, 2). In addition, cyanobacteria are key contributors to global nitrogen fixation (3) and many produce novel secondary metabolites (4). Despite these important traits and substantial interest in developing cyanobacterial strains for biotechnology, there is a paucity and unbalanced distribution of publicly available genomic information from the *Cyanobacteria* as 40% (29/72 species) of the available genomes fall within the closely related marine *Prochlorococcus*/*Synechococcus* subclade. Improvements in coverage of sequenced genomes will enable a more accurate and comprehensive understanding of cyanobacterial morphology, niche-adaptation, and evolution.

Taxonomic studies organized the *Cyanobacteria* into five Subsections based on morophological complexity (5). Unicellular forms are split between those that undergo solely binary fission (Subsection I, Chroococcales) and those that reproduce through multiple fissions in three planes to create smaller daughter cells, baeocytes (Subsection II, Pleurocapsales). Strains in Subsection III (Oscillatoriales) divide the vegetative cell solely perpendicular to the growing axis. Organisms in Subsections IV (Nostocales) and V (Stigonematales) are able to differentiate specific cells, i.e. heterocysts (for nitrogen fixation), and may form akinetes (dormant cells) and hormogonia (for dispersal and symbiosis competence). Subsection V is further distinguished by the ability to form branching filaments. Prior to this study, two Subsections (II and V) had no representative genomes, underscoring the dearth in our understanding of these more complex morphological phenotypes.

In this study, 54 strains of cyanobacteria were chosen to improve the distribution of sequenced genomes. The approach is modeled on the phylogenetically-driven Genomic Encyclopedia of Bacteria and Archaea (GEBA) (6) so we refer to our data as the CyanoGEBA dataset (*Appendix A*, Tables S1). The results highlight the value of phylum-wide genome sequencing based on phylogenetic coverage.

Results

Increased Coverage and Diversity of Cyanobacterial Genomes. Strains were chosen for genome sequencing based on their phylogenetic placement and their physiological relevance to the cyanobacterial research community (e.g. type strains). Beginning with a phylogenetic tree of cyanobacterial small subunit rRNA genes gathered from the greengenes database (7), cultured strains representative of major cyanobacterial branches for which genome sequences were not yet available were chosen for this study. 54 genomes, sequenced using Illumina and 454 technologies, were annotated and assembled, resulting in a collective total of 332 Mb, of which 29 are complete genomes and 25 are assembled to draft genome status (*Appendix A*, Table S1).

The cyanobacteria sequenced in this study cover a broad range of morphologies, lifestyles, and metabolisms. The CyanoGEBA dataset includes genomes from six baeocytous (Subsection II) and five ramified (Subsection V) morphotypes in addition to doubling the number of sequenced genomes from the heterocystous (11 out of 18) and filamentous (19 out of 29) strains. Diverse types of physiology are also encompassed in our dataset; highly halotolerant cyanobacteria (*Halothece* sp. PCC 7418 and *Dactylococcopsis* sp. PCC 8305), a fresh water picocyanobacterium (*Cyanobium* sp. PCC 6307), and a filamentous chlorophyll a and b containing cyanobacterium (*Prochlorothrix hollandica* PCC 9006) are represented at the genomic level. The CyanoGEBA data set also includes the largest cyanobacterial genome to date, *Calothrix* sp. PCC 7103 of 11.6Mb.

To evaluate the degree to which the 54 genomes improved coverage of the phylum, a species tree was generated using phylogenomic methods by concatenating 31 conserved proteins (8) (Fig. 1A, *Appendix A*, Fig. S1). The major subclades of the cyanobacterial tree were highly congruent with the 16S rRNA phylogeny (*Appendix A*, Fig. S2) and previous studies which have primarily used this molecular marker (9). A widely-used method to measure the diversity in a sample is the phylogenetic diversity metric, which takes branch lengths on a phylogeny as a proxy of diversity. This study's contribution to phylogenetic diversity was measured by the sum of the length of the 54 branches added by the CyanoGEBA genomes (10.82). To compare this value, randomly sampled subsets of 54 branches across all genomes were averaged (5.28 ±0.37). Thus, our dataset improves the diversity of the phylum approximately two-fold (1.92-2.20) (*Appendix A*, Table S2). A complementary method to show an improvement in coverage of the phylum is Tree Imbalance, specifically Colless's Imbalance, which measures how equally distributed branches are on a tree. Again, we observe a decrease in tree imbalance, indicative of a more even distribution of sequenced genomes across the cyanobacterial phylum (*Appendix A*, Table S2).

Surprisingly, out of the 292,935 proteins added from this dataset, 21,107 (7.2%) have no detectable similarity to any known protein sequence. Notably, 13% of the proteins from the *Leptolyngbya* sp. PCC 7375 draft genome are in this sense novel proteins (*Appendix A*, Table S3). Likewise, the CyanoGEBA data set contains a large number of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs). 50 of the 54 genomes sequenced in this study contain CRISPRs (*Appendix A*, Table S4); one these genomes, *Geitlerinema* sp. PCC 7105, contains the highest number of repeat-spacer units observed in cyanobacteria, with 650 units in a total of fifteen CRISPR loci.

Morphological Complexity. Examination of our cyanobacterial tree confirms the multiple and independent acquisition of the filamentous morphology (Subsection III), as well as of the ability to form baeocytes (Subsection II) (10); three unambiguous reversions and five gains in morphological complexity were revealed (Fig. 1, *Appendix A*, Table S5). Using comparative genomics, we searched for differences in the lineages bracketing these evolutionary transitions, which may represent proteins necessary for these morphological differences. Notably, there is an overlap in the sets of genes that are lost in two of the reversions from filamentous to unicellular morphology; 29 out of the 32 proteins (most annotated as hypothetical) that are lost in Event 2 correspond to the set of proteins lost in Event 3 (*Appendix A*, Table S6); this may reflect a similar convergence in the gene loss responsible for these two transitions from filamentous to unicellular phenotypes.

Surprisingly, we find no signature proteins specific to any of the complex morphologies. This also strongly argues for distinct convergences of Subsections II and III morphologies. The same holds true when considering the acquisition of the ability to form branching filaments (Subsection V within subclade B1). On the contrary, within the monophyletic heterocystous group within subclade B1 (Subsections IV and V), the morphological differentiation may be predicated on the concomitant presence of a set of genes, such as the twelve defined for heterocyst formation. The ability to undergo this unique cellular differentiation may be due to the presence of regulatory proteins in a common ancestor that lacked the ability to differentiate. This is consistent with previous studies (11) that noted the presence of essential genes for heterocyst development in non-heterocystous cyanobacteria. Similarly, this could explain why several genes previously proposed to underlie other morphological attributes (e.g. hormogonium or akinete formation) (12, 13) are also found spread across the phylum, suggesting they have lineage-specific functions. Overall, comparison of the COG functional categories of the five morphological Subsections shows that, in general, more complex morphologies are enriched in genes found in Signal Transduction and Transcription-related functional categories (*Appendix A*, Fig. S3), which may be indicative of the importance of regulatory elements in establishing morphological transitions.

Plastid Evolution. Cyanobacteria have greatly contributed to eukaryotic diversity, most notably in the plant kingdom, by giving rise to photosynthetic organelles via one or more endosymbiotic events. Many studies have attempted to find the closest relative to the original plastid endosymbiont leading to the Archaeplastida lineage. Poor phylogenetic sampling has also yielded conflicting conclusions on the identity of the most closely related extant cyanobacteria to the original endosymbiont; some studies claim the absence of a closest relative based on phylogenetic placement, whereas other studies have suggested heterocystous cyanobacteria to be the closest relatives to plastids (14, 15). We investigated the placement of the Archaeplastida lineage within the cyanobacterial phylum by building a 'plastidome tree' using a concatenation of 25 conserved plastid proteins. Although most studies support the monophyly of primary plastids (16, 17), others have reported a polyphyletic origin (18, 19). We find strong support for the monophyletic placement of plastids near the base of the cyanobacterial tree (Fig. 2a, *Appendix A*, Fig. S4) as previously observed by single loci phylogenetic analysis (9, 20, 21). The short branches near this node imply a possible large radiation event that occurred near the primary endosymbiosis event, as suggested previously (15). Despite the increased coverage through the inclusion of the CyanoGEBA dataset, we cannot identify which lineage was most closely related to the original plastid endosymbiont, finding no support for the claim that heterocystous cyanobacteria are most similar to the original endosymbiont (14). Criscuolo et al. (15) have previously reported the importance of investigating, at a phylogenomic level, the relation of plastids to the deep-branching *Pseudanabaena* lineage represented in our clade F (Fig. 1). This clade along with a small subset of unicellular cyanobacteria (clade G) is indeed basal to the plastid branch; however, neither of these two clades represents a distinct sister lineage most closely related to plastids, which makes it difficult to propose them as the original endosymbiont. Our phylogenetic analysis does not definitively reject the hypothesis that plastids emerged from clade F. However, considering that clades A-E are monophyletic, show a sister relationship to plastids, and share a common ancestor, all extant cyanobacteria of these clades are just as closely related to modern day plastids. Given that clades A-E cover representatives of all morphological subsections, with highly diverse physiologies, it is clear that it would be difficult to predict the morphological or metabolic traits of the original endosymbiont with any certainty.

Plastids have profoundly changed their eukaryotic hosts through endosymbiotic gene transfer (EGT), the relocation of genes from the endosymbiont genome to the host nuclear genome. Because we lack a close relative of the original endosymbiont and because the primary plastid endosymbiosis happened early within the crown cyanobacteria, only an improved coverage of cyanobacterial genomes can increase our ability to predict which genes underwent EGT. We compared predictions of EGT of nuclear genes from plastid-containing eukaryotes before and after the addition of the CyanoGEBA genomes. Nuclear proteins ascribed as the result of plastid EGT were defined as proteins with top BLAST hits from cyanobacterial, in contrast to other bacterial, archaeal, and non-plastid containing eukaryotic genomes. Given these criteria, we can now assign a cyanobacterial origin to many more genes (an average of 13% per genome) in the nuclear genomes of photosynthetic eukaryotes (Fig. 2b, *Appendix A*, Tables S7, S8).

Distribution of Membrane-bound Light Harvesting Complexes. Because oxygenic photosynthesis is the defining characteristic of cyanobacteria, we investigated the contribution of the CyanoGEBA dataset to surveying the diversity of photosynthetic light harvesting strategies. The majority of cyanobacteria absorb light mainly with soluble pigment-protein complexes called phycobilisomes, in contrast to eukaryotes which use membrane-bound Light Harvesting Complexes (LHCs). However, there is an increasing number of transmembrane proteins involved in cyanobacterial light-harvesting being identified, such as Pcb and IsiA (22, 23). These proteins are analogous in function to eukaryotic LHCs. Due to the growing number of proteins and names, an overarching nomenclature has been proposed to name this protein family the Chlorophyll Binding Proteins (CBPs), which are characterized by six transmembrane helices and the ability to bind chlorophyll (24).

With the increase in number and diversity of genomes, we find that CBPs are widely distributed across the cyanobacterial phylum: 67% (84 out of 126) of cyanobacterial genomes have, in addition to the phycobilisomes, genes that putatively function as membrane-bound light-harvesting proteins. In our phylogenetic analysis, the increase in sequence diversity reveals strong support for various subclades that we have provisionally named CBPIV, V, and VI (Fig. 3a, *Appendix A*, Fig. S5). Although not yet experimentally demonstrated, members of CBPIV, V and VI are expected to bind chlorophyll because they contain positionally-conserved histidine and glutamine residues that ligate chlorophyll in confirmed chlorophyll-binding CBPs (*Appendix A*, Fig. S6). Some of these proteins, such as CBPIV, have previously been annotated as PsbC homologs (25), because all CBP proteins are thought to have an common evolutionary origin with the *psbC* gene (24). Due to the vast enrichment of cyanobacterial protein sequences, the increase from two to six known CBPVI sequences augments phylogenetic resolution (bootstrap support of 85%), allowing us to more confidently assert that there is a separate and distinct CBPVI subfamily. Based on our phylogenetic analysis of the CBP family and consistent with previous studies (26), there appears to be a substantial amount of gene duplication and horizontal gene transfer among CBP IV, V and VI. In some genomes, CBPIV and CBPV are found in a gene cluster with other CBP proteins, including IsiA (Fig. 3c), suggestive of the potential for lateral transfer of gene clusters encoding light-harvesting proteins, as documented in marine cyanobacteria (27). Interestingly, many proteins of the CBPV clade also contain a C-terminal extension (*Appendix A*, Fig. S7) with homology to the PsaL subunit of Photosystem I (PSI). Notably, two distinct subclades within the CBPV family seem to have independently lost the PsaL domains, reflecting the modularity of this C-terminal extension. Homology modeling and insertion of the PsaL-like domain into the PSI structure (Fig. 3b, *Appendix A*, Fig. S8) suggests how the CBPV protein could theoretically be incorporated as an ancillary light harvesting polypeptide into a monomeric, but not trimeric, PSI. Although scattered observations of members of these CBP protein clades have been made in previously sequenced genomes (predominantly IsiA (CBPIII) and Pcb genes (dCBP, CBPI and CBPII)) it is clear that the contribution of the 54 genomes included in this study substantially increases the number of homologs within the CBP family allowing for a more thorough understanding of the distribution of distinct subclades within this large membrane-bound light harvesting protein family.

Secondary Metabolite Analysis. Much of the natural product chemical diversity observed in nature is attributed to versatile non-ribosomal peptide synthase (NRPS) and polyketide synthase (PKS) biosynthetic pathways (4). However, the extent and distribution of the capacity for secondary metabolite synthesis in cyanobacteria has nevertheless been underestimated. We retrieved 384 non-ribosomal gene clusters from 126 genomes, 61% from the CyanoGEBA dataset. Our results reveal that 70% of cyanobacterial genomes encode NRPS or PKS gene clusters (Fig. 1B, *Appendix A*, Fig. S9), their presence is partly correlated to the genome size (Pearson correlation on total number of NRPS/PKS gene clusters, or on total KS domains, as well as on total C domains: $R^2 = 0.3$, $p<0.0001$). Moreover, the distribution is uneven by a skewed frequency of NRPS/PKS in the late branches of our cyanobacterial tree (clades A and B), including all genomes from baeocystous, heterocystous and ramified morphotypes. Notably, 5.2% of the *Fischerella* sp. PCC 9339 genome is devoted to NRPS/PKS clusters, and contains an unexpected diversity with up to 22 NRPS/PKS clusters (5 NRPS, 10 PKS and 7 NRPS/PKS hybrids). Although the PCC 9339 genome is in a draft stage, nine of these clusters are located at a contig border and thus partial. Most of the clusters await characterization, however, the potential for the production of microcystins, shinorine or heterocyst glycolipids can already be predicted (*Appendix A*, Fig. S10).

Likewise, gene clusters involved in ribosome-dependent synthesis of diverse peptides through the post-translational modification of short precursor proteins (28-30), are even more broadly distributed across the phylum (*Appendix A*, Fig. S9). The most abundant corresponds to the newly discovered bacteriocin family (30, 31), whereas the terpenes (32) are present in almost all the 126 genomes. Even the genes encoding cyanobactins (33) were recovered from 10% of the dataset. Strikingly, the *Prochlorococcus*/*Synechococcus* subclade appears to lack NRPS gene clusters and harbors only type III PKS; however, they contain an abundance of bacteriocin clusters.

Discussion

With the exponentially growing capacity for sequencing genomes it is becoming increasingly important to focus sequencing efforts so as to obtain a high value return. Here, we show the benefits of genome sequencing based on a more representative phylogenetic coverage, with the objective of better understanding general characteristics of the phylum, as well as uncovering unique and novel traits of cyanobacterial genomes and subclades.

The addition of the CyanoGEBA genomes lays the foundation for the cyanobacterial phylum to become a model comparative genomic system for understanding the gain and loss of morphological complexity. Given close relationship between morphology and taxonomy for the *Cyanobacteria*, the genome sequence data now available from all five morphological subsections has revealed the lack of specific and unique genes that are the genetic determinants underlying these major phenotypes; a similar result emerged from comparative studies of eukaryotic genomes (34).

An increased distribution of sequenced cyanobacterial genomes has also corrected previous biases, such as the limited occurrence and diversity among CBPs. The addition of the CyanoGEBA dataset clearly shows that two-thirds of cyanobacterial genomes actually have membrane-bound CBPs encoded in their genomes, potentially allowing for alternative light-harvesting strategies other than phycobilisomes. Furthermore, the addition of these diverse CBPs has also enabled the placement of phylogenetically well-supported and distinct CBP subclades.

Our results likewise reveal an unexpectedly high frequency and diversity of NRPS/PKS enzyme systems for the production of secondary metabolites. Furthermore, we found that the known ribosomal dependent pathways for production of small peptides are also frequent and found throughout the lineage. Cyanobacteria have thus adopted multiple parallel strategies for the production of peptides through the modification of short precursors. Ultimately, their chemical diversity may rival or exceed that of the better-known non-ribosomal peptides and polyketides. The increased diversity of NRPS/PKS genes now apparent in the cyanobacterial phylum emphasizes one of the many benefits that are gained when using diversity-driven genome sequencing, which the previously biased genome representation of cyanobacteria failed to reveal.

Despite the global importance of the *Cyanobacteria*, there has been an unbalanced sequence distribution of the phylum, resulting in a lack of understanding at a genome-level of major clades and morphological subsections. The extensive phylogenetically-based survey of this single phylum has refined and extended our understanding of plastid evolution, phenotypic differences in morphology, light harvesting complexes, and secondary metabolisms in cyanobacteria. This study demonstrates the benefits gained from a more balanced representation of sequenced genomes within a phylum.

Materials and Methods

Genome Sequencing and Assembly. The 54 CyanoGEBA genomes were generated at the DOE Joint Genome Institute (JGI) using either a combination of Illumina (35) and 454 technologies (36) or only the Illumina technology (*Appendix A*, Table S9). Sequence data was assembled using an array of assemblers pending the data generated for a given genome. Assemblers included Newbler, Velvet and parallel Phrap (High Performance Software, LLC). The software Consed was used in the finishing process.

Phylogenetic Analysis. The species tree was generated by a concatenation of thirty-one conserved proteins (8). The plastidome tree was generated the same way using twenty-five conserved plastid proteins (*atpH, atpA, atpB, petB, psbA, psbB, psbC, psbD, psbE, psbL, psbH, psaA, psaB, psaC, rpl2, rpl14, rpl16, rps2, rps3, rps4, rps7, rps11, rps19, rpoB, and rpoC2*). Maximum likelihood phylogenetic trees were generated with PhyML 3.0 (37).

Measuring Improved Phylum Sampling. The phylogenetic diversity metric was measured as described by Wu et al. (6). A maximum likelihood tree omitting the four outgroup genomes with 51 resamplings of the random set of taxa was used to estimate the contribution of the CyanoGEBA genomes in increasing the phylogenetic diversity of the overall tree. Methods used for the Tree Imbalance analysis are described in the *Appendix A*.

Comparative Genomic Analysis. An 'all vs all' BLASTP search was conducted for all cyanobacterial proteins used in this study with an e-value threshold of 1e-10 and a span cutoff of 80%, which was then used to build protein families using the Markov clustering algorithm (MCL) where each cluster was considered a protein family (38). Comparative genomics to characterize the protein families lost or gained in specific morphological lineages were based upon the MCL protein families.

Prediction of Endosymbiotic Gene Transfer. Nuclear-encoded proteins from plastid-containing eukaryotes (*Appendix A*, Tables S7 and S8) were used as queries to BLASTP against two databases: 1) containing all cyanobacteria, representatives from other bacterial and archaeal phyla, and representatives from non-plastid containing eukaryotes, and 2) the same above, however using only cyanobacterial genomes available prior to the CyanoGEBA study. Top-BLAST hits to cyanobacterial proteins were considered genes of cyanobacterial descent, and the total counts for each of the nuclear genomes are presented in *Appendix A*, Tables S7.

Secondary Metabolite Analysis. Secondary metabolite biosynthesis gene clusters were identified using met2db (39), antiSMASH (40), and NaPDoS (41). Adenylation domain substrate specificity predictions for NRPS enzymes were made using NRPSpreditor2 (42). Annotations were refined manually using CD-search, BLASTP and InterProScan to identify conserved domains. We estimated the number of gene clusters for each genome using the three methods and containing the minimum of domains needed to perform synthesis. Pearson correlation tests were performed using XLSTAT, v2007-4 (Addinsoft, France).

Author contributions. P.M.S. (evolutionary analysis, novel proteins, endosymbiotic transfer, CBP analysis and modeling, manuscript preparation), D.W. (evolutionary analysis, novel proteins), A.L. (cell division and cell differentiation analysis), S.A. (data analysis), D.P.F. (secondary metabolite analysis, manuscript preparation), E.T. (cell division and cell differentiation analysis), A.C. (secondary metabolite analysis), F.C. (CRISPR analysis), N.T.M. (selection of strains, data analysis), R.R. (selection of strains, strain growth, purity checks, metadatabase compilation), M.H. (metadatabase compilation), K.S. (secondary metabolite analysis), T.C. (strain growth, DNA preparation), T.L. (strain growth), L.G. (project management), A.C., M.N., K.D., C.S.H. (finishing), EMR (manuscript preparation), J.A.E. (data analysis), T.W. (data analysis, project coordination), M.G. (selection of strains, strain curation, DNA preparation, secondary metabolite analysis, metadabase, manuscript preparation, European team coordination), C.A.K (project lead and overall coordination, data analysis, strain selection, manuscript preparation).

References:

1. Partensky F, Hess WR, & Vaulot D (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol Mol Biol Rev* 63:106-127.

2. Garcia-Pichel F, Belnap J, Neuer S, & Schanz F (2003) Estimates of global cyanobacterial biomass and its distribution. *Algol Stud* 109:213-227.
3. Zehr JP*, et al.* (2008) Globally distributed uncultivated oceanic N2-fixing cyanobacteria lack oxygenic photosystem II. *Science* 322:1110-1112.
4. Welker M & Von Döhren H (2006) Cyanobacterial peptides – Nature's own combinatorial biosynthesis. *FEMS Microbiol Rev* 30:530-563.
5. Rippka R, Deruelles J, Waterbury JB, Herdman M, & Stanier RY (1979) Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *J Gen Microbiol* 111:1-61.
6. Wu D*, et al.* (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462:1056-1060.
7. DeSantis TZ*, et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72:5069-5072.
8. Wu M & Eisen JA (2008) A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* 9:R151.
9. Turner S, Pryer KM, Miao VPW, & Palmer JD (1999) Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. *J Eukaryot Microbiol* 46:327-338.
10. Schirrmeister B, Antonelli A, & Bagheri H (2011) The origin of multicellularity in cyanobacteria. *BMC Evol Biol* 11:45.
11. Zhang J-Y, Chen W-L, & Zhang C-C (2009) hetR and patS, two genes necessary for heterocyst pattern formation, are widespread in filamentous nonheterocyst-forming cyanobacteria. *Microbiology* 155:1418-1426.
12. Campbell EL, Wong FCY, & Meeks JC (2003) DNA binding properties of the HrmR protein of *Nostoc punctiforme* responsible for transcriptional regulation of genes involved in the differentiation of hormogonia. *Mol Microbiology* 47:573-582.
13. Zhou R & Wolk CP (2002) Identification of an akinete marker gene in *Anabaena variabilis*. *J Bacteriol* 184:2529-2532.
14. Deusch O*, et al.* (2008) Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Mol Biol Evol* 25:748-761.
15. Criscuolo A & Gribaldo S (2011) Large-scale phylogenomic analyses indicate a deep origin of primary plastids within Cyanobacteria. *Mol Biol Evol* 28:3019-3032.
16. Rodríguez-Ezpeleta N*, et al.* (2005) Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Curr Biol* 15:1325-1330.
17. Chan CX, Gross J, Yoon HS, & Bhattacharya D (2011) Plastid Origin and Evolution: New Models Provide Insights into Old Problems. *Plant Physiol* 155:1552-1560.
18. Nozaki H*, et al.* (2009) Phylogenetic positions of Glaucophyta, green plants (Archaeplastida) and Haptophyta (Chromalveolata) as deduced from slowly evolving nuclear genes. *Mol Phylogenet Evol* 53:872-880.
19. Baurain D*, et al.* (2010) Phylogenomic Evidence for Separate Acquisition of Plastids in Cryptophytes, Haptophytes, and Stramenopiles. *Mol Biol Evol* 27:1698-1709.

20. Marin B, M. Nowack EC, & Melkonian M (2005) A plastid in the making: Evidence for a second primary endosymbiosis. *Protist* 156:425-432.

21. Nelissen B, Van de Peer Y, Wilmotte A, & De Wachter R (1995) An early origin of plastids within the cyanobacterial divergence is suggested by evolutionary trees based on complete 16S rRNA sequences. *Mol Biol Evol* 12:1166-1173.

22. Roche JL*, et al.* (1996) Independent evolution of the prochlorophyte and green plant chlorophyll a/b light-harvesting proteins. *Proc Natl Acad Sci USA* 93:15244-15248.

23. Laudenbach DE & Straus NA (1988) Characterization of a cyanobacterial iron stress-induced gene similar to psbC. *J Bacteriol* 170:5018-5026.

24. Chen M, Zhang Y, & Blankenship R (2008) Nomenclature for membrane-bound light-harvesting complexes of cyanobacteria. *Photosynth Res* 95:147-154.

25. Kaneko T*, et al.* (2001) Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium anabaena sp. strain PCC 7120. *DNA Res* 8:205-213.

26. Chen M, Hiller RG, Howe CJ, & Larkum AWD (2005) Unique origin and lateral transfer of prokaryotic chlorophyll-b and chlorophyll-d light-harvesting systems. *Mol Biol Evol* 22:21-28.

27. Rocap G*, et al.* (2003) Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* 424:1042-1047.

28. Schmidt EW*, et al.* (2005) Patellamide A and C biosynthesis by a microcin-like pathway in *Prochloron didemni*, the cyanobacterial symbiont of Lissoclinum patella. *Proc Natl Acad Sci USA* 102:7315-7320.

29. Ziemert N*, et al.* (2008) Microcyclamide biosynthesis in two strains of *Microcystis aeruginosa*: from structure to genes and vice versa. *Appl Environ Microbiol* 74:1791-1797.

30. Li B*, et al.* (2010) Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria. *Proc Natl Acad Sci USA* 107:10430-10435.

31. Wang H, Fewer DP, & Sivonen K (2011) Genome mining demonstrates the widespread occurrence of gene clusters encoding bacteriocins in cyanobacteria. *PLoS ONE* 6:e22384.

32. Agger SA, Lopez-Gallego F, Hoye TR, & Schmidt-Dannert C (2008) Identification of sesquiterpene synthases from *Nostoc punctiforme* PCC 73102 and *Nostoc* sp. strain PCC 7120. *J Bacteriol* 190:6084-6096.

33. Sivonen K, Leikoski N, Fewer D, & Jokela J (2010) Cyanobactins—ribosomal cyclic peptides produced by cyanobacteria. *Appl Microbiol Biotechnol* 86:1213-1225.

34. Prochnik SE*, et al.* (2010) Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science* 329:223-226.

35. Bennett S (2004) Solexa Ltd. *Pharmacogenomics* 5:433-438.

36. Margulies M*, et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.

37. Guindon S*, et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Sys Biol* 59:307-321.

38. Enright AJ, Van Dongen S, & Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30:1575-1584.

39. Bachmann B & Ravel J (2009) Chapter 8. Methods for In silico prediction of microbial secondary metabolic pathways from DNA sequence data. *Methods Enzymol* 458:181-217.

40. Medema M*, et al.* (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* 39:W339-346.

41. Ziemert N*, et al.* (2012) The natural product domain seeker NaPDoS: A phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS ONE* 7:e34064.

42. Röttig M*, et al.* (2011) NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res* 39:W362-367.

Figures:

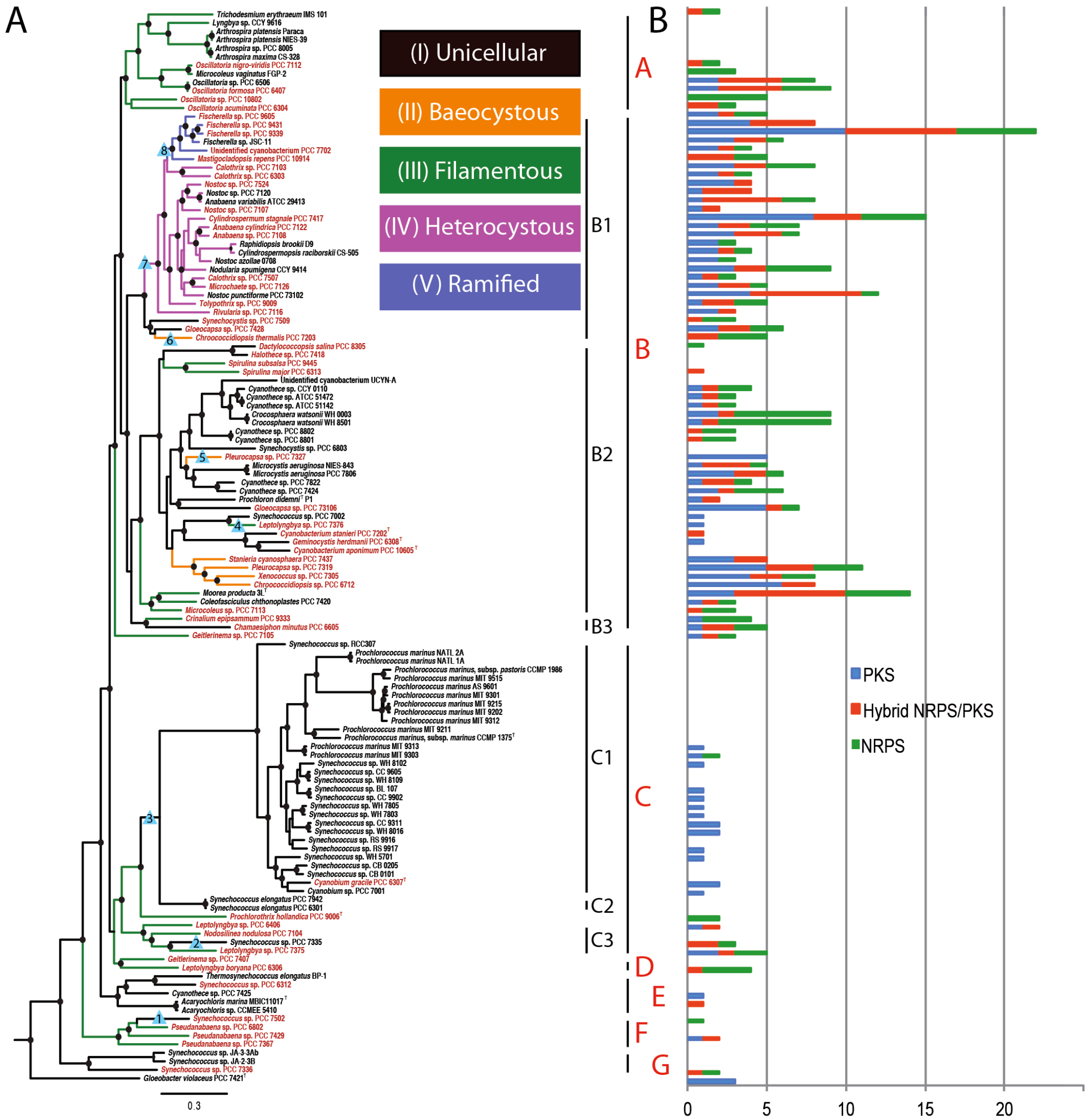


Figure 1. **Cyanobacterial species tree and the distribution of secondary metabolite biosynthesis. A,** Maximum-likelihood phylogeny of cyanobacteria included in this study (outgroup shown in *Appendix A*, Fig. S1). Branches are color-coded based on morphological Subsection. Taxa names in red are genomes sequenced in this study. Nodes supported with a bootstrap of ≥70% are indicated by a black dot. Morphological transitions that were investigated are denoted by blue triangles, annotated by events 1-8. Phylogenetic subclades are grouped into 7 major subclades (A-G), some of which are made up of smaller subgroups. See *Appendix A*, Tables S1 for reference information for genomes used in this analysis. **B**, Distribution of the non-ribosomal peptide and polyketide gene clusters.

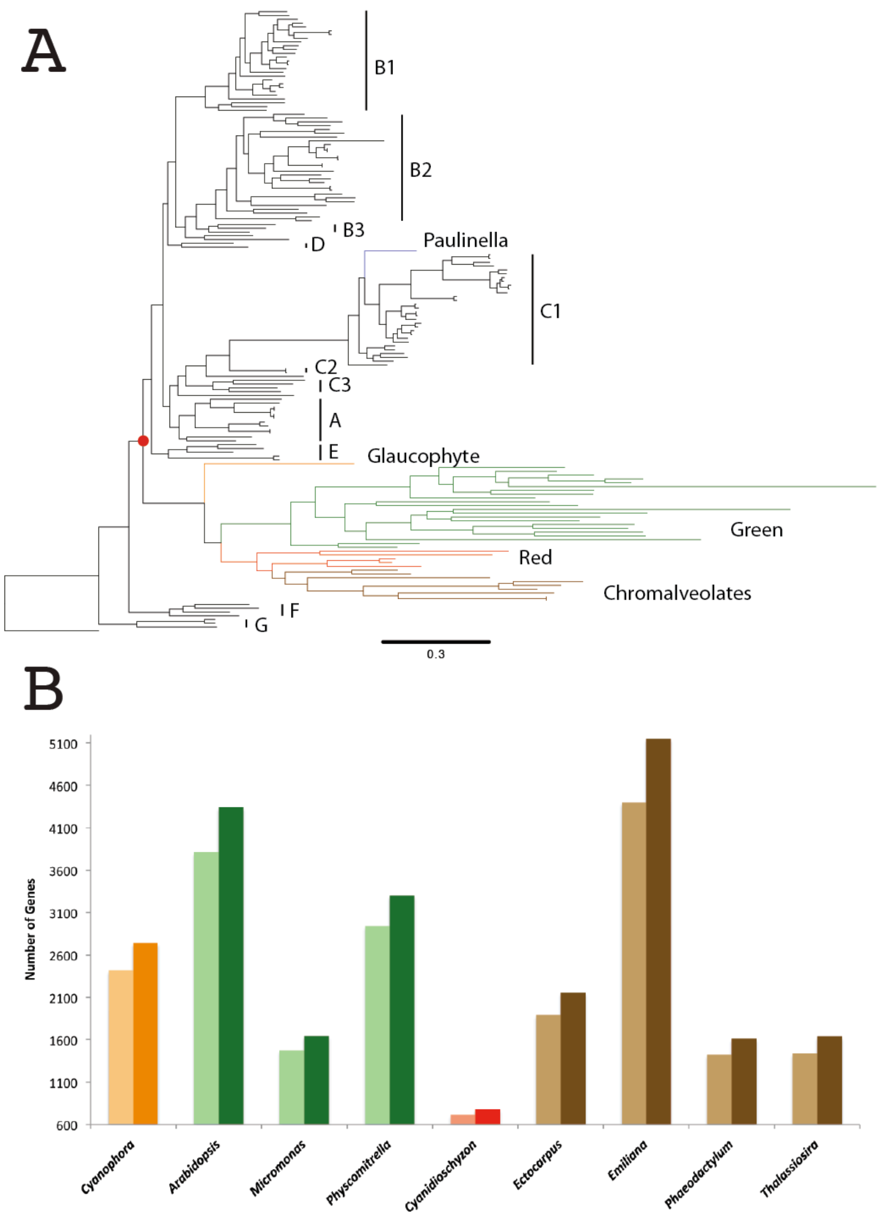


Figure 2. Implications on plastid evolution. A, Maximum-likelihood phylogenetic tree of plastids and cyanobacteria, grouped by subclades (Fig. 1). The red dot (bootstrap support = 97%) represents the primary endosymbiosis event that gave rise to the Archaeplastida lineage, made up of Glaucophytes (orange), Rhodophytes (red), and Viridiplantae (green), and Chromaleveolates (brown). The independent primary endosymbiosis in the amoeba *Paulinella chromatophora* is shown in purple. **B,** Number of predicted eukaryotic, nuclear genes transferred from a cyanobacterial endosymbiont. Colors correspond to the lineage organisms as above. Light and dark shades of colors represent before and after adding the CyanoGEBA genomes, respectively.

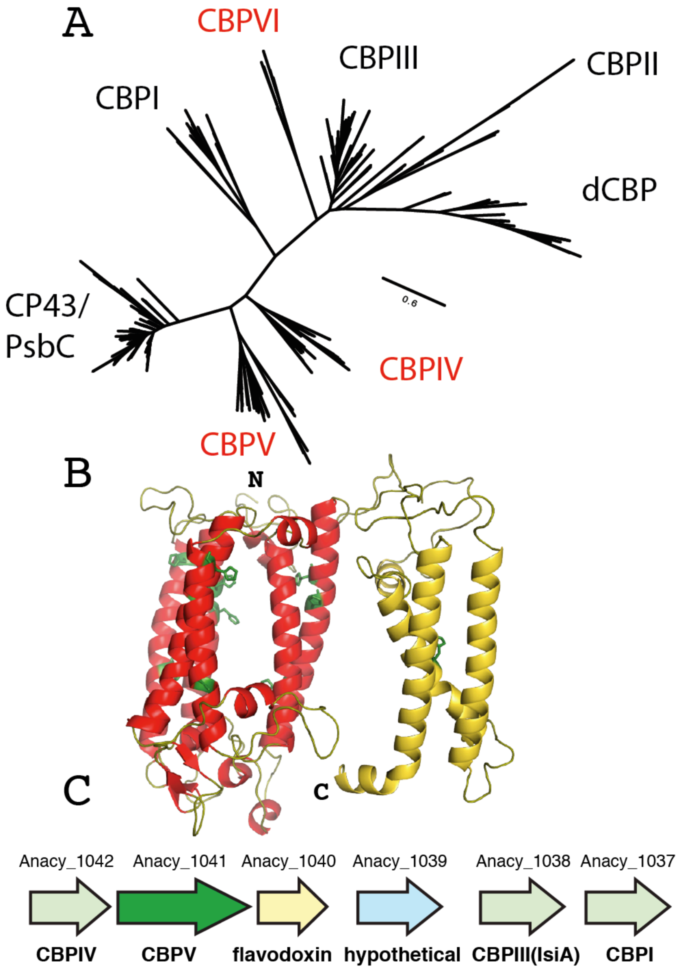


Figure 3. Increased sequence coverage reveals distinct and highly-supported subclades of putative Chlorophyll Binding Proteins. A, Unrooted maximum-likelihood tree of CBP sequences. Putative CBP clades that have emerged as distinct and phylogenetically well-supported are labeled in red, and previously described CBP clades are labeled in black. CP43 protein sequences (encoded by the PsbC gene) are provided as an outgroup. **B,** Cartoon representation of novel domain architecture of CBPV from *Chroococcidiopsis thermalis* PCC 7203 (Chro_2988), based on the two separate homology models of 1) the N-terminal CBP domain (red) and 2) the C-terminal PsaL-like domain (yellow). Potentially chlorophyll-binding histidine residues are shown in green sticks. **C,** Gene cluster containing multiple CBP genes from *Anabaena cylindrica* PCC 7122 (locus tags labeled above, annotations labeled below).

Chapter 3

Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins

Abstract

Chloroplasts and mitochondria descend from bacterial ancestors, but the dating of these primary endosymbiosis events remains very uncertain despite their importance for our understanding of the evolution of both bacteria and eukaryotes. All phylogenetic dating in the Proterozoic and before is difficult: Significant debates surround potential fossil calibration points based on the interpretation of the Precambrian microbial fossil record. Strict molecular clock methods cannot be expected to yield accurate dates; even with more sophisticated relaxed-clock analyses, nodes that are distant from fossil calibrations will have a very high uncertainty in dating. However, endosymbiosis events and gene duplications provide some additional information that has never been exploited in dating, namely that certain nodes on a gene tree must represent the same events and thus must have the same or very similar dates, even if the exact date is uncertain. We devised techniques to exploit this information: cross-calibration, in which node date calibrations are re-used across a phylogeny, and cross-bracing, where node date calibrations are formally linked in a hierarchical Bayesian model. We apply these methods to proteins with ancient duplications that have remained associated and originated from plastid and mitochondrial endosymbionts: the α and β subunits of F_1-ATPase and its relatives, and the elongation factor Ef-Tu. The methods yield reductions in dating uncertainty of 14-26%, while using only date calibrations derived from phylogenetically unambiguous Phanerozoic fossils of multicellular plants and animals. Our results are suggestive that primary plastid endosymbiosis occurred ~900 Mya and mitochondrial endosymbiosis occurred ~1200 Mya.

Preface

The contents of this chapter are based on the following publication (published online before print):

Patrick M. Shih and Nicholas J. Matzke (2013). Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins. *Proceedings of the National Academy of Sciences of the United States of America.* July 23; 110(30):12355-60, doi: 10.1073/pnas.1305813110

My contributions to this work included designing experiments, interpreting data, and writing the manuscript. Nicholas Matzke equally contributed to the design, analysis, and work in this chapter. Supplemental information for this chapter can be found in Appendix A.

Abbreviations:

TOL	Tree of Life

Mya	Million years ago
LUCA	Last universal common ancestor
MCMC	Markov chain Monte Carlo
HPD	Highest posterior density
CV	Coefficient of variation
GOE	Great Oxidation Event
LCA	Last common ancestor

Introduction

Biologists have often attempted to estimate when key events on the Tree of Life (TOL) occurred. This approach has experienced substantial success for dating events in the Phanerozoic [543 million years ago (Mya) to the present], but when trying to date deep events on the TOL, such as endosymbiosis events in the Proterozoic (2500-543 Mya), it becomes increasingly difficult to find reliable fossil calibrations. Molecular dating analysis is performed by calibrating a phylogenetic tree with known dates, usually based on fossil calibration points. Ideally, the dating of phylogenetic events deep in the Precambrian would be well-constrained by fossil calibrations; however, many of the fossil calibrations that have been proposed for Precambrian microorganisms are controversial due to the difficulty in identifying the clade memberships of these groups.

Although the timing of the origin of eukaryotes is heavily studied and debated, the endosymbiosis events involved in the origin and diversification of many eukaryotic lineages are arguably equally contentious. Fossil records for eukaryotes have been claimed back to 2700 Mya (1), while others have speculated that "Snowball Earth" events postponed the origin and/or diversification of eukaryotes until as recently as 850-580 Mya (2-4). Interpretation of microfossils is inherently difficult due to difficult preservation, taphonomic, and interpretive issues [e.g., (5, 6)]. A less-recognized problem is that fossil calibrations are best done via a phylogenetic analysis of characters, which allows objective placement of fossils on a tree, and a measurement of the uncertainty of this placement (7). General similarity to an extant group is an insufficient basis for using a fossil as a date calibration: characters must place the fossil in the crown group rather than a stem group [sometimes an insufficiently appreciated distinction (8)], to constrain the date of the last common ancestor of the crown group (7). However, microfossils typically have a very small number of diagnosable characters (9), thus running the risk of misclassification, especially due to homoplasy. Chemical biomarkers, another much-used strategy to date Precambrian lineages, are equally problematic because, fundamentally, each biomarker constitutes a single character unassociated with other fossil characters. To be used for dating it must be assumed that the character evolved only once and is unique to one extant clade, but this is not always a safe assumption as demonstrated by the recent finding that the methylhopane biomarker, once used specifically for cyanobacteria (10), can also be found in a broad range of other bacterial phyla (11, 12).

Apart from uncertainty in fossil calibrations, molecular dating imposes additional uncertainties. Early attempts at molecular dating, starting with Zuckerkandl and Pauling

(13), invoked a strict molecular clock to date divergences. Subsequent attempts to date deep nodes in the TOL have given wildly varying results, many of which clearly do not agree with fossil, let alone geological, histories primarily due to rate variation not accounted for by strict clock models (14, 15). More sophisticated models allow for rate variation and thus provide a more realistic assessment of uncertainty. However, the uncertainty that results can be vast – the origin of crown eukaryotes has been dated between 3970-1100 Mya throughout various studies (16).

Uncorrelated relaxed-clock methods, available in Bayesian phylogenetic dating methods, allow the rate of evolution on each branch to be drawn from a "common distribution", the parameters of which are themselves estimated during the analysis. One advantage of Bayesian analysis is that it takes into account diverse sources of known prior information. Another technique used in several studies relies on the concatenation of protein sequences in order to increase phylogenetic signal for estimations of deeply rooted events. However, this strategy does nothing to remedy the problem of scarce and ambiguous fossil calibrations for deep nodes.

Given the difficulty of dating deep nodes in the Proterozoic as well as the lack of studies dating Precambrian events with newer methods, it is useful to explore possible improvements in relaxed clock analyses. We hypothesize that better estimates of rates and rate variability, and thus better estimates of dates and dating uncertainty, would occur if more prior information and more date calibrations were input into analyses. Date calibrations are typically scarce, but we suggest they can be multiplied in cases where one or more ancient duplications have been universally or near-universally inherited. In such cases, a single fossil calibration can date not just one node in the tree, but several. An example where this is possible is the protein family of ATPases found within the F_1 portion of the F_1F_o-ATP synthase system and its relatives, the vacuolar V_1V_o-ATPases and archaeal A_1A_o-ATPases (17). The α and β subunits of F_1-ATPase duplicated before the Last Universal Common Ancestor (LUCA) (18, 19) and have been almost universally inherited as a pair since then (Figure 1A). Furthermore, the core function of the ATPases in energy production has resulted in high conservation and a lower probability of extreme rate variation.

The fact that mitochondria and plastids have retained these ATPase proteins (whether they are encoded by the organellar or the nuclear genome) means that many homologs may coexist in one organism. For example, plant genomes contain six homologous copies of this ATPase subunit: both homologous α and β subunits targeted to the mitochondria, chloroplasts, and vacuoles. Therefore, one plant fossil, which calibrates the date of the divergence of monocots and eudicots, can actually provide calibration dates for up to six nodes on the ATPase α and β subunit phylogeny. We propose two methods for use of these calibrations (Appendix B, Figure S1). In the first strategy, which we dub "cross-calibration", the date calibrations are simply re-used at each node, and the dates of these nodes are subsequently sampled independently during the Markov chain Monte Carlo (MCMC) search. Cross-calibration is simple to implement, but neglects the fact that nodes representing the same event should have the same date, even if that date is uncertain. We therefore also propose a second strategy, "cross-bracing," in which the dates of calibrated nodes representing the same speciation events are linked, and thus co-vary during MCMC sampling. This is a more accurate representation of our prior knowledge that a single speciation event led to the simultaneous divergence of the nuclear, mitochondrial, plastid ATPase genes (although some variability could be caused by lineage sorting processes).

Iwabe et al (18) and Gogarten et al (19) attempted to use ancient duplicated genes in inferring distant evolutionary relationships between the three domains of life using α and β subunits of ATP synthase (ATPase) and elongation factor Tu (Ef-Tu). Their rooting of the TOL has been much debated due to problems with saturation of phylogenetic signal at the very deepest nodes of the tree (20), and the possible breakdown of the tree concept itself when it comes to the origin and rooting of the three domains (21). Owing to these issues, we do not attempt to revisit the question of the root of the TOL or its date in this study; instead, we focus on the much more recent, but still Precambrian, endosymbiosis events which gave rise to mitochondria and chloroplasts. The root and the date of the TOL will be treated as highly uncertain nuisance parameters, over which our Bayesian analysis will integrate, due to the numerous hazards involved in extrapolative dating at the base of the TOL, including but not limited to HGT for some ATPases (22). In this study, we augment a standard Bayesian relaxed molecular clock approach with our new cross-calibration and cross-bracing methods and show the influence of these methods on the estimates and precision of dates for major endosymbiosis events within the Eukaryotes.

Results and Discussion

BEAST analyses. In order to measure the effect of cross-calibration and cross-bracing on an overall dating analysis and the effect of different amounts of prior dating information, nine separate relaxed-clock dating analyses using ATPase sequences (*Appendix B*, Table S1) were performed using the program BEAST (23) (24). Six analyses used only α-subunit sequences, each cross-calibrated using some or all of the available node date calibrations (α-cross-calibrated); one analysis conducted cross-calibration with all node date calibrations using only β-subunit sequences (β-cross-calibrated); one analysis conducted cross-calibration with all node date calibrations applied simultaneously to a tree containing all α- and β-subunits (α/β cross-calibrated); and the last analysis used all calibrations and both α- and β-subunits, but used the cross-bracing approach to link node dates (α/β cross-braced). Consensus trees from these analyses are shown in *Appendix B*, Figures S2-S6.

Effect of cross-calibration methods on age, rates, and uncertainty. The change in precision of date estimates between calibration methods was measured by comparing the width of the 95% Highest Posterior Density (HPD) of node age between analyses (only nodes in the α-subunit portion of the tree, which existed in all analyses, were compared). The null hypothesis, indicating no difference, predicts a 1:1 relationship in node uncertainty between methods. Regression was used to test for statistically significant departure from a 1:1 relationship. The increased amount of dating information incorporated into the α/β-cross-calibrated analysis and α/β-cross-braced analysis yielded a decrease in uncertainty (14-26%) for both the α/β-cross-calibrated and α/β-cross-braced runs (Figure 2; *Appendix B*, Table S2). This was a significant result (*p*-value always <0.0025; the F-test was used for all regressions). There was no significant difference in uncertainty when comparing α/β-cross-calibrated and α/β-cross-braced runs. (*Appendix B*, Table S2, Figure S7).

Branch rates were also estimated with more precision using the cross-calibration and cross-bracing methods, where regressions indicate a 42-57% decrease in uncertainty in rate for the α/β-cross-calibrated tree compared to α- and β- cross-calibrated trees (*Appendix B*, Table S2). Some of this decrease is due to the fact that the mean rates as estimated by α/β-cross-calibration were also on average slightly lower (6-12%) than in α-only or β-only analyses, and the mean rate and the uncertainty in rate are strongly correlated. However, the effect remains when the coefficients of variation (CV) in rate estimates were observed; a 14-29% reduction in uncertainty is observed. Further examination of the effects of cross-calibration and cross-bracing on node age and branch rate uncertainty is elaborated in the *Appendix B*, Supplemental Analysis of BEAST runs section.

The α/β-cross-braced run produced node dates that averaged about 5% younger than the corresponding node dates in the α/β-cross-calibrated, α-, and β-cross-calibrated analyses. The differences were statistically significant (vs. α-cross-calibrated: *p*=2.97E-06; vs. β-cross-calibrated: *p*=1.19E-05; vs. α/β-cross-calibrated: 5.75E-08). In addition, the intercept term was significantly negative (vs. α-cross-calibrated: *p*=4E-06; vs. β-cross-calibrated: *p*=0.018; vs. α/β-cross-calibrated: *p*=8.64E-12), indicating that in addition to the 5% average difference, cross-braced node ages tended to be lower by a fixed amount of 37-65 million years (*Appendix B*, Table S2, Figure S8).

To further investigate the effect of reducing the number of prior calibration dates, the α-cross-calibrated analysis using all node calibrations was compared to the α-cross-calibrated analyses using fewer calibration priors (*Appendix B*, Table S3). Uncertainty in node age was not dramatically different between the α-cross-calibrated dataset with all date calibrations versus subsets of these calibrations (*Appendix B*, Table S2, Table S3). However, when the variance between node age and uncertainty is accounted for by calculating the coefficient of variation (CV), comparison of CVs showed a significant decrease in CV (23-44%) when all calibration nodes were used, suggesting that increasing the number of calibration points decreases relative uncertainty in the estimates of node age in α-only analyses. Moreover, branch rate uncertainty significantly increased for runs with fewer calibrations except Run 5 (*Appendix B*, Table S4, Figure S9). Further comparisons of all runs, including the α/β-cross-calibrated and α/β-cross-braced runs, are summarized in the *Appendix B*, Supplemental Analysis of BEAST runs section.

Dating symbiosis events: ATPases. Because the α/β-cross-calibrated and α/β-cross-braced runs were shown to decrease rate and age uncertainty, but neither method yielded significantly more robust results when compared against each other. For simplicity, we will henceforth refer to only the α/β-cross-calibrated analysis (summarized in Table 1).

The timing of plastid endosymbiosis has been as contentious as dating the rise of eukaryotes. The hypothesis that cyanobacteria are responsible for the Great Oxidation Event (GOE) has led to many studies extrapolating divergence time points for a broad range of uses, from dating endosymbiosis events to events of multicellularity (25-28). However, this approach assumes that all crown cyanobacterial lineages emerged at the time of the GOE (29). Our study was aimed at dating the plastid endosymbiosis event agnostic of the GOE, microfossils, or biomarker data, and instead calibrated only by well-accepted Phanerozoic divergence events. Our cross-calibrated analysis estimates primary plastid endosymbiosis and the birth of the Archaeplastida lineage at 857 and 1055 Mya (857/1055 Mya), based on F-type α and β subunits of the tree, respectively. These dates are remarkably similar to the dates estimated by Douzery et al who predicted the plastid endosymbiosis occurred between 825 and 1,162 Mya using 129 concatenated protein sequences, as well as to other previous large-scale and broadly-sampled molecular clock studies (30).

Although younger than other predicted estimated divergence dates (31, 32), our dates present a plausible scenario for the changing geochemical properties of the ocean. The rise of photosynthetic eukaryotes through the acquisition of plastids ~900 Mya most likely dramatically added to primary productivity in the sea, which may have significantly contributed to the conversion of euxinic oceans during the Neoproterozoic to its oxygenated state which persists today (33). This is further supported by the dramatic increase in atmospheric oxygen between 1005-640 Mya (34). Our analysis is suggestive that the diversification of Archaeplastida occurred near or during the time of the transformation of euxinic conditions to its modern day properties, and that there was very little lag time between the origin and diversification of photosynthetic eukaryotes.

Numerous phylogenetic studies have placed the plastid endosymbiosis event near the base of the extant cyanobacterial tree (35-37). Assuming that crown cyanobacteria were responsible for the GOE, this would place the plastid endosymbiosis near the time of the GOE. This is in contradiction to our study and many concatenated, multi-loci molecular clock studies (30-32), which have conservatively dated the origin of crown eukaryotes well after 2 Gya ago. It is therefore difficult to reconcile these dates, because plastid endosymbiosis could not have occurred prior to the origin of eukaryotes. Moreover, all bacterial phyla in our analysis (including cyanobacteria) have diversified after the GOE, thus suggesting that extant crown cyanobacteria were not responsible for the GOE. Our findings are in contrast with Schirrmeister et al (28) who date the origin of crown cyanobacteria prior to the GOE. These findings are attributable to their assignment of ancient (>2 Gya) cyanobacterial-like fossils to extant clades, despite the possibility that the few available morphological characters may be homoplastic and may have evolved several times convergently. Assuming that the GOE was of biological origin, our results imply that crown cyanobacteria may not have been responsible for the GOE. However, this does not rule out the possibility of its origin from stem group cyanobacteria, which may have gone extinct during the major transition from euxinic to oxic oceans (33). In line with this idea, the phylogeny of crown cyanobacteria has been interpreted as a large radiation event (35, 37), which may have occurred following the extinction of stem groups and adaptation of crown lineages to the changing ocean surfaces. These extinct lineages may be the Proterozoic cyanobacterial-like fossils described in previous studies (27, 38-40) and used as fossil calibrations by Schirrmeister et al (28). Our analysis reflects the controversial nature of contrasting molecular and fossil studies and thus emphasizes the need to improve existing phylogenetic techniques to more accurately examine the dating of these Precambrian events.

Our cross-calibration analysis dates the rise of modern-day mitochondria through the endosymbiosis of an α-proteobacterium to be 1176/1248 Mya. Although the vacuolar subclades display an earlier date for the last common ancestor of eukaryotes, our interest was in dating the actual divergence between bacteria and mitochondria; other dates in the analyses were treated as nuisance variables. Given that the MRCA of eukaryotes most likely is younger than the mitochondrial endosymbiosis, we recognize the contradiction between the dates in the two parts of the tree, which is probably caused by fewer calibration points and an accelerated rate of evolution at the base of the V-ATPase tree. However, the only methodological remedy would be to use the cross-bracing technique on those nodes we want to infer, whereas in this study we are examining the potential of cross-linking date calibration nodes. Cross-bracing nodes with dates that are to be inferred rather than used as calibrations should be explored in the future, but issues of extended autocorrelation in the posterior distribution and low estimated sample size become much more pressing if the nodes targeted for inference are cross-braced.

Parfrey et al estimate the last common eukaryotic ancestor to be more than 1600 Mya (31), notably older than our analysis; however, when excluding Proterozoic fossil calibrations, they observed shifts in all major clades to be 300 million years younger-nearly comparable with our results. The effects of excluding Proterozoic microfossil calibrations may explain the incongruence in estimated dates between studies; however, for the purposes of our study, our focus on cross-calibration methods was to increase the amount of dating prior information with younger and less controversial Phanerozoic fossils. Finally, our analysis does not find evidence for the hypothesis that crown eukaryotes originated ~850 Mya and postdate the hypothesized Snowball Earth.

Although earlier Proterozoic and Archean events are not the primary focus of this study and uncertainties this far back are large, we observe long branches leading to the Eukarya/Archaea split, followed by a radiation of extant Eukarya/Archaea (V-type ATPases) and Eubacteria (F-type ATPases) around 2000-2500 Mya. Because the rise in molecular oxygen in the atmosphere occurred around the same time, it is tempting to speculate that this synchronized radiation of extant life across all three kingdoms was somehow facilitated by the GOE and that all extant life forms are the descendants of lineages that most successfully adapted to the changing biogeochemistry in ocean surfaces.

Dating symbiosis events: Ef-Tu. Because there may be inherent biases between particular markers used for any phylogenetic analysis, we extended our cross-calibration study to elongation factor Tu (Ef-Tu) because of its similar evolutionary history to ATPases, which allows for cross-calibration. Bacterial Ef-Tu and its eukaryotic/archaeal homolog, translation elongation factor 1α (EF-1α), allow for entry of aminoacyl tRNAs into the ribosome, and thus are considered conserved, slowly-evolving proteins, decreasing the chance of saturation and high rate variability. The dates estimated from the Ef-Tu chronogram were similar to the dates attained from the ATPase analysis: plastid endosymbiosis (1188 Mya) and mitochondrial endosymbiosis (1196 Mya) (Table 1; *Appendix B*, Figure S10). Estimations of deeper nodes such as the split between Archaea and Eukarya (1528 Mya) differed from the ATPase results by almost 800 Mya. This is not surprising, because many of these nodes may inherently be difficult to estimate due to lack of signal from a saturation of amino acid substitutions (20).

Conclusion

Cross-calibration and cross-bracing, using duplication or endosymbiosis events, provide useful advantages compared to conventional molecular dating. First, they increase the sampling and sequence data used, which improves accuracy of the dating of internal nodes (41, 42). Secondly, by increasing the number of sequences that are cross-calibrated, they decrease the chance of artifacts being introduced by underestimated rate variation. Just as there are multiple calibration points for a given divergence event, a divergence event will be estimated multiple times on the tree. Third, the increase in calibration points allows for the use of more well-accepted calibration points closer to the tips of the tree, rather than relying on older and more contentious microscopic, Precambrian fossils.

The flexibility of the BEAST XML input allows unconventional strategies such as ours to be employed. However, the cross-bracing technique could be improved. Future efforts should develop algorithms that redesign the MCMC tree search such that nodes with linked dates can be specified, and linked nodes can be allowed to share identical dates during sampling. This should eliminate all or most of the need for longer runs to account for increased autocorrelation in the posterior sample. The cross-bracing strategy might also improve inference in another way: nodes which have dates which are unknown, but which represent the same event, could be linked as we have done here for calibration nodes. For example, the nodes representing the divergence of the chloroplasts should have the same or nearly the same date between the α- and β- subunit gene trees, instead of two individually estimated dates. Further refinements could include linking rates for genes when they are inhabiting the same species; this would avoid the assumption, made here by necessity, that rates and rate variation are independent across the tree.

It is important to note that our approach is different from the common technique of concatenation of gene duplicates into a larger alignment. For example, if a researcher were only interested in dating nodes within plants, to increase signal they might concatenate the α- and β -subunit sequences from vacuolar, chloroplast, and mitochondrial ATPases. However, this conventional strategy would be useless when the goal is to date nodes in the gene tree that are not represented by nodes in the species tree – for example, the date of a gene duplication itself, or, as in this study, the date of endosymbiosis events.

Although we observed similar dates between ATPase and Ef-Tu, it will be interesting to determine whether other molecular markers that have undergone duplications or endosymbiotic transfers and can be used in cross-calibration will also yield similar dates. Possible examples include aminoacyl-tRNA synthetases (43), translation initiation factors (44), and phytochrome (45) datasets. Cross-calibration could also be extended to large concatenated datasets if all proteins display similar histories.

Regardless of the detailed method employed, we argue that, due to the difficulty in estimating the timing of Precambrian events, every possible source of information should be included. As we show here, this information is not merely found in the dates of fossil calibrations– it can also include linkages between nodes that represent the same speciation or duplication events. Information about the relative timing of events could also be included– for example, the origin of crown chloroplasts must equal or postdate the origin of crown eukaryotes. Hierarchical Bayesian models excel in the incorporation of such diverse sources of information, and should be exploited wherever possible, along with other attempts to ameliorate dependence on controversial date calibrations based on ancient, microscopic fossils that are difficult to interpret and rigorously place on phylogenies.

Materials and Methods

Alignments. ATPase α and β subunit and EF-Tu/1α protein sequences were all gathered from the Uniprot database and are listed in *Appendix B*, Table S5. Sequences were chosen to cover a broad range of bacterial, archaeal, and eukaryotic phyla. Alignments were generated using the –maxiterate strategy in MAFFT (46).

Dating programs. Estimation of dated phylogenies was conducted with BEAST 1.7.3 (23, 24). BEAST XML input files were started using BEAUTi 1.7.3 (23, 24), but our novel calibration strategies, described below, required custom modifications to the XML code. The WAG model was chosen as the best-fitting amino acid substitution matrix available in BEAST based on ProtTest analysis for all datasets (47). Production of the final BEAST XML files for the different combinations of datasets and calibration methods was done via custom programs in R 2.15 (48). BEAST XML files implementing the cross-calibration and cross-bracing methods are available in *Appendix B,* Supplementary Materials and Methods. All BEAST runs were inspected for convergence and completeness of sampling the posterior distribution using Tracer (49).

Node date calibrations. Dating calibration distributions were based on macroscopic fossils of Phanerozoic plants and animals that provide well-accepted calibration points used in previous molecular dating studies of Phanerozoic groups (50, 51) (*Appendix B*, Table S6). Although the origin of crown angiosperms estimated by Smith *et al.* (50) was older than previous studies and fossil records (52), we found the discrepancy of ~80 Mya negligible in comparison to the divergence estimates we were focused on in this study. More importantly, the other estimated dates used as calibration points from Smith *et al.* aligned well with the current estimates of divergences within land plants (53, 54). Plant calibration points were used for the plant vacuolar, mitochondrial, and plastid ATPases.

The human/chicken and fly/mosquito divergences were used as metazoan calibration points (51). To maintain maximum agnosticism about the date of the Last Common Ancestor (LCA) and the divergence of the ATPase α and β subunits, which occurred before the LCA, a uniform distribution prior between 3800-2500 Mya was set at the base of the tree (the split between α and β subunits), assuming a biological origin of the Great Oxidation Event 2500 Mya (55), and that life most likely could not have began before the Late Heavy Bombardment of Earth ca. 3800 Mya (56).

Cross-calibration and cross-bracing methods. In the cross-calibration method, each node in the gene tree corresponding to the same speciation event is assigned the same prior distribution on the date, i.e., the distribution given in *Appendix B*, Table S6. These distributions are cross-calibrated, or "unlinked": that is, during MCMC sampling, the date of each node is sampled independently from the prior distribution.

As with cross-calibration, in the cross-bracing method, each node in the gene tree corresponding to the same speciation event is assigned the same prior distribution on the date. However, in the cross-bracing method, the dates of nodes corresponding to the same speciation event are "linked." As BEAST cannot formally do joint sampling of node dates, we achieved the same effect by coding into the BEAST XML an additional prior on the differences between the dates of linked nodes and the mean of the linked nodes. This prior was a normal distribution, with a mean of 0 (as any prior on the difference from the mean must have) and a standard deviation set to 1% of the mean of the prior distribution of the date of the speciation event. Thus, while BEAST samples each "linked" node independently during the actual MCMC sampling, samples in which the linked nodes are far apart will have a low posterior probability and will be rejected more often than in the cross-calibration approach. Inspection of linked node dates in Tracer (57) showed that they were indeed highly correlated to each other, unlike in the cross-calibration approach.

The 1% standard deviation value on the distribution of differences from the mean date was chosen to indicate our prior high confidence that nodes corresponding to the same speciation event should have approximately the same date. The distribution on differences from the mean date was not set even more tightly for two reasons. First, lineage-sorting processes can cause some degree of difference in the divergence dates of gene trees during speciation. Second, it was important to give BEAST's MCMC sampler "breathing room" to sample the date of one linked node, then another, then another, etc., without too many of these moves being rejected, so that the full posterior distribution could be explored. Further analysis was conducted as described in the *Appendix B*, Supplemental Material and Methods and Supplemental Analysis of BEAST Runs section.

Author Contributions. P.M.S. and N.J.M both contributed equally to this work.

References:

1. Brocks JJ, Logan GA, Buick R, & Summons RE (1999) Archean Molecular Fossils and the Early Rise of Eukaryotes. *Science* 285(5430):1033-1036.

2. Hoffman PF, Kaufman AJ, Halverson GP, & Schrag DP (1998) A Neoproterozoic Snowball Earth. *Science* 281(5381):1342-1346.
3. Cavalier-Smith T (2006) Cell evolution and Earth history: stasis and revolution. *Philos Trans R Soc Lond B Biol Sci* 361(1470):969-1006.
4. Cavalier-Smith T (2010) Deep phylogeny, ancestral groups and the four ages of life. *Philos Trans R Soc Lond B Biol Sci* 365(1537):111-132.
5. Schopf JW & Kudryavtsev AB (2012) Biogenicity of Earth's earliest fossils: A resolution of the controversy. *Gondwana Research* 22(3–4):761-771.
6. Brasier MD*, et al.* (2002) Questioning the evidence for Earth's oldest fossils. *Nature* 416(6876):76-81.
7. Parham JF*, et al.* (2012) Best Practices for Justifying Fossil Calibrations. *Syst. Biol.* 61(2):346-359.
8. Budd GE (2003) The Cambrian Fossil Record and the Origin of the Phyla. *Integr. Comp. Biol.* 43(1):157-165.
9. Diver W & Peat C (1979) On the interpretation and classification of Precambrian organic-walled microfossils. *Geology* 7(8):401-404.
10. Summons RE, Jahnke LL, Hope JM, & Logan GA (1999) 2-Methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis. *Nature* 400(6744):554-557.
11. Rashby SE, Sessions AL, Summons RE, & Newman DK (2007) Biosynthesis of 2-methylbacteriohopanepolyols by an anoxygenic phototroph. *Proc Natl Acad Sci* 104(38):15099-15104.
12. Welander PV, Coleman ML, Sessions AL, Summons RE, & Newman DK (2010) Identification of a methylase required for 2-methylhopanoid production and implications for the interpretation of sedimentary hopanes. *Proc Natl Acad Sci* 107(19):8537-8542.
13. Zuckerkandl E & Pauling L (1965) Molecules as documents of evolutionary history. *Journal of Theoretical Biology* 8(2):357-366.
14. Martin W, Gierl A, & Saedler H (1989) Molecular evidence for pre-Cretaceous angiosperm origins. *Nature* 339(6219):46-48.
15. Doolittle RF (1992) Reconstructing history with amino acid sequences. *Protein Science* 1(2):191-200.
16. Roger AJ & Hug LA (2006) The origin and diversification of eukaryotes: problems with molecular phylogenetics and molecular clock estimation. *Philos Trans R Soc Lond B Biol Sci* 361(1470):1039-1054.
17. Mulkidjanian AY, Makarova KS, Galperin MY, & Koonin EV (2007) Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. *Nat Rev Micro* 5(11):892-899.
18. Iwabe N, Kuma K, Hasegawa M, Osawa S, & Miyata T (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci* 86(23):9355-9359.
19. Gogarten JP*, et al.* (1989) Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc Natl Acad Sci* 86(17):6661-6665.
20. Philippe H & Forterre P (1999) The Rooting of the Universal Tree of Life Is Not Reliable. *J Mol Evol* 49(4):509-523.

21. Doolittle WF & Bapteste E (2007) Pattern pluralism and the Tree of Life hypothesis. *Proc Natl Acad Sci* 104(7):2043-2049.

22. Hilario E & Gogarten JP (1993) Horizontal transfer of ATPase genes--the tree of life becomes a net of life. *BioSystems* 31:111-119.

23. Drummond AJ & Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7(1):214.

24. Drummond AJ, Suchard MA, Xie D, & Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.*

25. Falcon LI, Magallon S, & Castillo A (2010) Dating the cyanobacterial ancestor of the chloroplast. *ISME J* 4(6):777-783.

26. Schirrmeister B, Antonelli A, & Bagheri H (2011) The origin of multicellularity in cyanobacteria. *BMC Evol Biol* 11(1):45.

27. Tomitani A, Knoll AH, Cavanaugh CM, & Ohno T (2006) The evolutionary diversification of cyanobacteria: Molecular–phylogenetic and paleontological perspectives. *Proc Natl Acad Sci* 103(14):5442-5447.

28. Schirrmeister BE, de Vos JM, Antonelli A, & Bagheri HC (2013) Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event. *Proc Natl Acad Sci*.

29. Sato N (2006) Origin and Evolution of Plastids: Genomic View on the Unification and Diversity of Plastids. *The Structure and Function of Plastids,* Advances in Photosynthesis and Respiration, eds Wise R & Hoober JK (Springer Netherlands), Vol 23, pp 75-102.

30. Douzery EJP, Snell EA, Bapteste E, Delsuc F, & Philippe H (2004) The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils? *Proc Natl Acad Sci* 101(43):15386-15391.

31. Parfrey LW, Lahr DJG, Knoll AH, & Katz LA (2011) Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc Natl Acad Sci* 108(33):13624-13629.

32. Yoon HS, Hackett JD, Ciniglia C, Pinto G, & Bhattacharya D (2004) A Molecular Timeline for the Origin of Photosynthetic Eukaryotes. *Mol Biol Evol* 21(5):809-818.

33. Johnston DT, Wolfe-Simon F, Pearson A, & Knoll AH (2009) Anoxygenic photosynthesis modulated Proterozoic oxygen and sustained Earth's middle age. *Proc Natl Acad Sci* 106(40):16925-16929.

34. Canfield DE & Teske A (1996) Late Proterozoic rise in atmospheric oxygen concentration inferred from phylogenetic and sulphur-isotope studies. *Nature* 382(6587):127-132.

35. Criscuolo A & Gribaldo S (2011) Large-scale phylogenomic analyses indicate a deep origin of primary plastids within Cyanobacteria. *Mol Biol Evol* 28(11):3019-3032.

36. Turner S, Pryer KM, Miao VPW, & Palmer JD (1999) Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. *J Eukaryot Microbiol* 46(4):327-338.

37. Shih PM*, et al.* (2013) Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc Natl Acad Sci* 110(3):1053-1058.

38. Schopf JW (1993) Microfossils of the Early Archean Apex Chert: New Evidence of the Antiquity of Life. *Science* 260(5108):640-646.
39. Schopf JW (2012) The Fossil Record of Cyanobacteria. *Ecology of Cyanobacteria II*, ed Whitton BA (Springer Netherlands), pp 15-36.
40. Hofmann HJ (1976) Precambrian microflora, Belcher Islands, Canada; significance and systematics. *J. Paleontol.* 50(6):1040-1073.
41. Zwickl DJ & Hillis DM (2002) Increased Taxon Sampling Greatly Reduces Phylogenetic Error. *Syst. Biol.* 51(4):588-598.
42. Wertheim JO & Sanderson MJ (2011) Estimating diversification rates: How useful are divergence times? *Evolution* 65(2):309-320.
43. Brown JR & Doolittle WF (1995) Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc Natl Acad Sci* 92(7):2441-2445.
44. Keeling PJ, Fast NM, & McFadden GI (1998) Evolutionary Relationship Between Translation Initiation Factor eIF-2γ and Selenocysteine-Specific Elongation Factor SELB: Change of Function in Translation Factors. *J Mol Evol* 47(6):649-655.
45. Mathews S, Clements MD, & Beilstein MA (2010) A duplicate gene rooting of seed plants and the phylogenetic position of flowering plants. *Philos Trans R Soc Lond B Biol Sci* 365(1539):383-395.
46. Katoh K, Kuma K-i, Toh H, & Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33(2):511-518.
47. Abascal F, Zardoya R, & Posada D (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21(9):2104-2105.
48. R Development Core Team (2012) R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria).
49. Rambaut A & Drummond AJ (2007) Tracer version 1.5).
50. Smith SA, Beaulieu JM, & Donoghue MJ (2010) An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. *Proc Natl Acad Sci*.
51. Berbee ML & Taylor JW (2010) Dating the molecular clock in fungi – how close are we? *Fungal Biol Rev* 24(1–2):1-16.
52. Sanderson MJ, Thorne JL, Wikström N, & Bremer K (2004) Molecular evidence on plant divergence times. *American Journal of Botany* 91(10):1656-1665.
53. Wellman CH & Gray J (2000) The microfossil record of early land plants. *Philos Trans R Soc Lond B Biol Sci* 355(1398):717-732.
54. Doyle JA (1998) Molecules, Morphology, Fossils, and the Relationship of Angiosperms and Gnetales. *Molecular Phylogenetics and Evolution* 9(3):448-462.
55. Kopp RE, Kirschvink JL, Hilburn IA, & Nash CZ (2005) The Paleoproterozoic snowball Earth: A climate disaster triggered by the evolution of oxygenic photosynthesis. *Proc Natl Acad Sci* 102(32):11131-11136.
56. Cohen BA, Swindle TD, & Kring DA (2000) Support for the Lunar Cataclysm Hypothesis from Lunar Meteorite Impact Melt Ages. *Science* 290(5497):1754-1756.
57. Rambaut A & Drummond A (2007) Tracer v1.4. *Available from* http://beast.bio.ed.ac.uk/Tracer.

Figures:

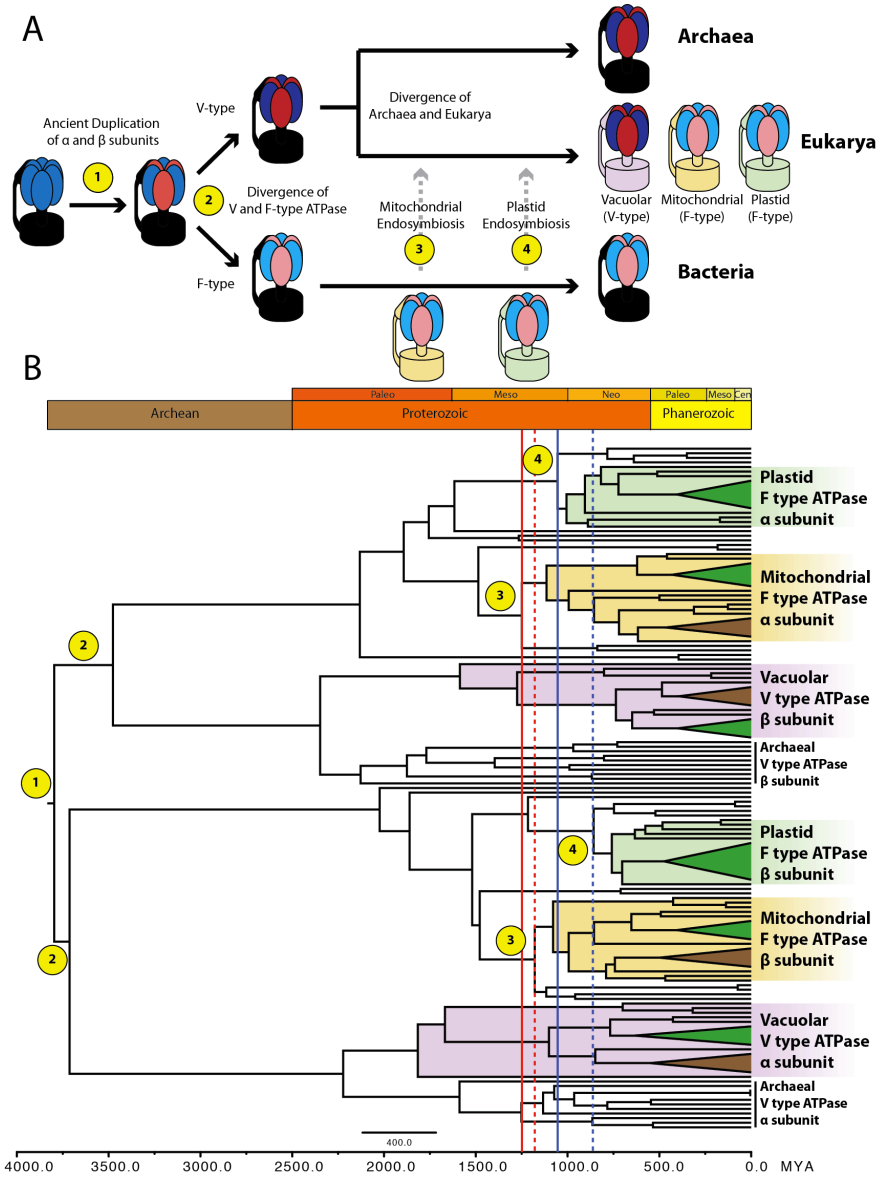

A
Ancient Duplication of α and β subunits
1
2
V-type
F-type
Divergence of V and F-type ATPase
Divergence of Archaea and Eukarya
Mitochondrial Endosymbiosis
3
Plastid Endosymbiosis
4
Archaea
Eukarya
Vacuolar (V-type)
Mitochondrial (F-type)
Plastid (F-type)
Bacteria
B
Paleo
Meso
Neo
Paleo
Meso
Cen
Archean
Proterozoic
Phanerozoic
4
Plastid F type ATPase α subunit
3
Mitochondrial F type ATPase α subunit
2
Vacuolar V type ATPase β subunit
1
Archaeal V type ATPase β subunit
Plastid F type ATPase β subunit
4
Mitochondrial F type ATPase β subunit
3
2
Vacuolar V type ATPase α subunit
Archaeal V type ATPase α subunit
400.0
4000.0
3500.0
3000.0
2500.0
2000.0
1500.0
1000.0
500.0
0.0
MYA

Figure 1. **Evolutionary history of the ATPase α- and β-subunits and divergence time estimates inferred from cross-calibration analysis. A,** Cartoon schematic that demonstrates the common origin of both α- and β-subunits, followed by both the mitochondrial and plastid endosymbiosis events, all of which enable the utilization of cross-calibration methods. Evolutionary events of interest are numbered and labeled onto the subsequent chronogram generated from cross-calibration of the α- and β-subunits. **B**, Time-scale phylogeny generated from Bayesian analysis of cross-calibrated ATPase α- and β-subunits (*Appendix B*, Fig. SB). Blue lines denote the dates estimated for the primary plastid endosymbiosis event. Red lines denote the dates estimated for mitochondrial endosymbiosis. Solid lines represent dates that were inferred from the α-subunit subsection of the phylogeny, whereas dashed lines were inferred from the β-subunit subclade.

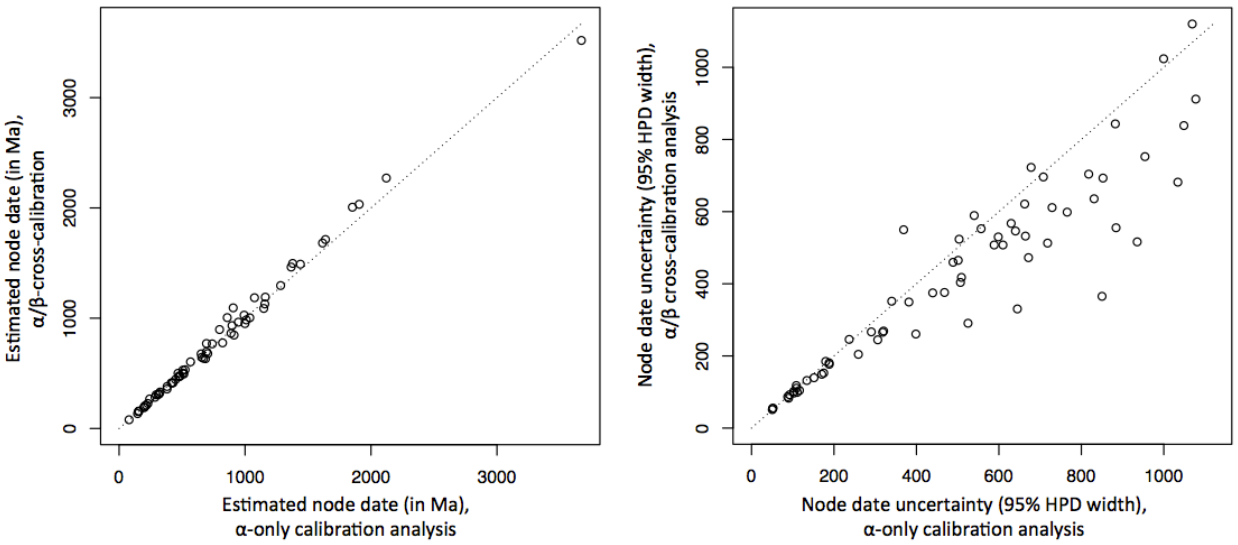


Figure 2. Cross-calibration decreases dating uncertainty. Comparison (regression analysis) of estimates of node age in F_1-ATPase proteins under BEAST runs with two different calibration methods, namely dated calibrations only within the α-subunit gene tree (α-cross-calibrated) (x-axis) and cross-calibration across the ATPase phylogeny of α- and β-subunits (α/β-cross-calibrated) (y-axis). Each dot represents a corresponding node-date estimate from the α-portion of the tree. The left panel shows the mean estimates of node age, which are not statistically significantly affected by calibration strategy (*p*=0.145, F-test). The right panel compares precisions between the two analyses (the width of the 95% highest posterior density (HPD) on node age). Average uncertainty in node age estimates is decreased by about 22% by the cross-calibration strategy, a statistically-significant result (*p*=2.96e-07, F-test).

Table 1: Divergence-time estimates (in Mya) for major endosymbiosis or domain divergence events. Dates in parentheses denote the 95% HPD.

Divergence Event	Cross-calibrated ATPase α/β subunits	Cross-calibrated EfTu
Plastid endosymbiosis	α subunit: 1055 (1278-913)	1188 (896-1613)
	β subunit: 857 (1098-720)	
Mitochondrial endosymbiosis	α subunit: 1248 (1838-1217)	1196 (909-1551)
	β subunit: 1176 (1524-1053)	

Chapter 4

Resurrected Precambrian RuBisCO enzymes reflect changes in atmospheric composition

Abstract:

The vast majority of global life is sustained by carbon assimilation catalyzed by the enzyme ribulose-1,5-bisphosphate carboxylase oxygenase (RuBisCO), the most abundant protein on Earth (1). RuBisCO is an ancient protein that has evolved over billions of years amidst the drastically changing geochemical landscape of Earth, from the anoxic Archean atmosphere to the present atmosphere. A confounding property of RuBisCO is its dual carboxylase and oxygenase activity. It is still debated why the counterproductive oxygenase activity has persisted over billions of years of evolution, as it competes with the carboxylase reaction necessary for carbon fixation. Hypotheses about the selective pressures governing RuBisCO evolution have been confined to mere speculation (2, 3). Here we report the resurrection and characterization of ancestral RuBisCOs, dating back over one billion years ago (Gya). The Precambrian RuBisCOs display slower reaction kinetics, likely due to the high CO_2 and low O_2 Precambrian atmosphere, but exhibit similar specificity factors to extant bacterial homologs. Our findings provide an ancestral point of reference revealing that eukaryotic homologs were driven toward improved specificity for CO_2. In comparison, cyanobacterial homologs evolved to increase rates of carboxylation. Consistent with this, in vivo analysis reveals the propensity of ancestral RuBisCOs to be encapsulated into modern-day carboxysomes, bacterial organelles central to the cyanobacterial CO_2 concentrating mechanism (CCM). This study demonstrates the use of resurrected proteins to probe paleoenvironments and the selective pressures that dictate the evolutionary trajectory of enzymes.

Preface:

My contribution to this work includes designing the experiments, interpreting the data, and writing the manuscript. Dr. Alessandro Occhialini, Dr. John Andralojc, and Dr. Martin Parry carried out the enzyme purification and kinetic measurements. Dr. Jeff Cameron performed microscopy. More specific contributions are listed in the 'Author Contribution' section below. Supplemental information for this chapter can be found in Appendix C.

Abbreviations:

RuBisCO	Ribulose-1,5-bisphosphate carboxylase oxygenase
Gya	Billion years ago
CCM	CO_2 concentrating mechanism
GOE	Great Oxidation Event
PAL	Present atmospheric levels
MRCA	Most recent common ancestor
α-MRCA	MRCA of the Form 1A RuBisCO clade
β-MRCA	MRCA of the Form 1B RuBisCO clade

α/β-MRCA	MRCA of both Form 1A and 1B RuBisCO clades
V_c	Carboxylation turnover rate
τ	Specificity factor
CFP	Cerulean Fluorescent Protein
EYFP	Enhanced Yellow Fluorescent Protein

Introduction:

Our understanding of the Precambrian (4 - 0.542 Gya) environment is limited to the indirect observations of geological proxies to infer the ancient landscape. However, the inherent property of evolution in biological systems offers a means by which to directly probe ancient proteins to infer paleoenvironments (4). Using a combination of phylogenetic and molecular biology techniques, it is possible to predict and 'resurrect' ancient proteins with ancestral sequence reconstruction methods. One particular enzyme that has had profound implications for life on Earth is RuBisCO, the predominant enzyme used to fix CO_2 into organic compounds, linking biological systems to their inorganic environment through global biogeochemical cycles.

Form 1 RuBisCO is composed of eight large (RbcL) and eight small (RbcS) subunits, forming a hexadecameric (L_8S_8) holocomplex. The enzyme has two reactions: 1) the carboxylase activity, which is responsible for photosynthetic carbon fixation and 2) the competing oxygenase activity, which instead fixes O_2 leading to photorespiration. Although the oxygenase activity decreases the productivity of carbon fixation, its activity has persisted across all known RuBisCOs (5). Plants and the majority of cyanobacteria contain Form 1B RuBisCO, whereas proteobacteria and a subgroup of marine cyanobacteria make up the sister Form 1A RuBisCO clade (Figure 1). Although the dating of this divergence event is ambiguous, it is most likely more than a billion years old, given that it predates the origins of plastid-containing eukaryotes (Archaeplastida), which fossil records and molecular clock studies have consistently dated between 1.8 – 1.2 Gya (6-9).

Based on previous studies (10-12) on the decreasing CO_2 and increasing O_2 atmospheric composition over Earth's history (Figure 1), it has been proposed to be the primary and most direct selective pressure on RuBisCO driving the evolution and improvements on carboxylase activity (13). Even after the Great Oxidation Event (GOE) (2.5 Gya), the vast majority of the Proterozoic Era experienced relatively low O_2 concentrations, 5-18% of present atmospheric levels (PAL) (12). Based on solar luminosity models, CO_2 concentrations have been predicted to be between 300-600 times higher than present atmospheric levels (11). Thus, selective pressure to improve carboxylase activity most likely would not have occurred until the late Neoproterozoic (0.6-0.8 Gya) when a second significant increase in O_2 levels to current PAL (14) occurred simultaneously with decreasing CO_2 concentrations.

Results and Discussion:

For both the RbcL and RbcS subunits, we reconstructed the most recent common ancestor of the Form 1A clade (α-MRCA), the Form 1B clade (β-MRCA), and both Form 1A and 1B clades (α/β-MRCA) (Appendix C, Fig. S1 & S2). The ancestral proteins were predicted from independently derived phylogenetic trees for RbcL and RbcS containing a broad diversity of Form 1A and 1B RuBisCO (>100 sequences) (Appendix C, Fig. S3 & S4). Maximum likelihood algorithms (15) were used to reconstruct the most probabilistic ancestral sequence for each ancestral node, as well as for determining the posterior probability for each amino acid position (Appendix C, Fig. S5). The origin of these ancestral proteins can most reliably be traced back to the Mesoproterozoic Era or earlier, given their relationship to the plastid endosymbiosis event.

To investigate the biochemical properties of ancestral RuBisCO, the genes were synthesized and expressed in *Escherichia coli* (*E. coli*). Although most ancestral sequence reconstruction studies have primarily studied monomeric proteins avoiding the challenge to reconstitute ancestral protein-protein interactions, the 550 kiloDalton hexadecameric L_8S_8 holocomplex of the β-MRCA and α-MRCA RuBisCO could still assemble, as determined by size exclusion chromatography (Appendix C, Fig. S6). The enzymes were subsequently purified and biochemically characterized.

Ancestral enzymes displayed slower carboxylation turnover rates (V_c) than those of extant bacterial RuBisCos (Table 1). The Proterozoic climate would have provided little selective pressure from the atmosphere to drive RuBisCO to evolve improved kinetic properties, given that the CO_2 / O_2 ratios would have been orders of magnitude larger than PAL. Correspondingly, the V_c/K_c^{air} parameter - which represents the ability of RuBisCO to function in low CO_2 concentrations (16) - of ancestral RuBisCOs is substantially lower than those of extant RuBisCOs (Table 1), indicating the high CO_2 conditions of the Proterozoic atmosphere were necessary for the ancestral enzymes to properly function. Moreover, the ability to discriminate between CO_2 and O_2, also known as the specificity factor (τ), would not have played a large role in the Proterozoic atmosphere. Interestingly, purified β-MRCA and α-MRCA enzymes display τ values similar to that of extant bacterial RuBisCOs (~50) (Table 1). Although this is a relatively low value in comparison to other extant eukaryotic RuBisCOs, ancestral τ values suggest that the ancestral enzymes began at a low baseline τ and subsequently evolved as separate RuBisCO lineages diverged. In response to the significant atmospheric changes during the Neoproterozoic, bacterial Form 1A and 1B RuBisCOs continued an evolutionary path which maintained a low τ, however increased their V_c. In contrast, eukaryotic RuBisCOs diverged and continued along a separate evolutionary path focusing on improving specificity rather than V_c (Figure 2).

A tradeoff between V_c and τ has been observed with the broad range of characterized extant RuBisCOs (3, 17). Considering the protein fitness landscape of RuBisCO, it has been hypothesized that the best-fit curve encompassing the inverse relationship of the two parameters represents the upper limit of RuBisCO activity constrained by the physiochemical and structural properties of the enzyme (18). Ancestral RuBisCOs display relatively slow V_c and low τ when compared to their extant counterparts (Figure 2A). These results suggest that the changing atmospheric conditions provided the selective pressure needed to push divergent RuBisCO lineages towards the upper limit of carboxylase activity, given the tradeoff between V_c and τ (Figure 2B). Comparison of ancestral RuBisCO kinetics to those of single point mutant and chimeric RuBisCOs show that the ancestral RuBisCOs display higher kinetic parameters than many of their modified extant counterparts, suggesting that the predicted sequences are reasonably accurate, as they still have functional activity and can reconstitute the 550 kDa hexadecameric holocomplex (Figure 2).

Based on their biochemical properties, extant cyanobacterial RuBisCOs could still perform photosynthesis efficiently under early Phanerozoic CO_2 (15-20 times PAL) and O_2 (at PAL) concentrations (13). Thus, it is reasoned that CCMs were not necessary and did not evolve until ~0.4 Gya. Due to the complexity in evolving the multiple CCM components (e.g. transporters, carboxysomes), the factor most directly selected upon and thus most likely to evolve first due to the drastically changing atmosphere was the catalytic properties of RuBisCO. To examine the subsequent rise of various CCMs, we focused on the carboxysome, a distinctive component of the bacterial CCM. The carboxysome is an organelle composed of an array of proteins, including RuBisCO and carbonic anhydrase; encapsulation of these enzymes in a protein shell increases the local CO_2 concentration around RuBisCO. All known cyanobacteria contain carboxysomes and encapsulate the majority of the cellular RuBisCO within; α-carboxysomes contain Form 1A RuBisCO and β-carboxysomes for Form 1B RuBisCO.

Because the large subunit has been shown to determine holoenzyme encapsulation within the carboxysome (19), we co-localized fluorescently-tagged ancestral RbcL with the β-carboxysome in an extant cyanobacterium, *Synechococcus elongatus* PCC 7942 (*Synechococcus*) (Figure 3). All ancestral RuBisCOs displayed spatially distributed puncta across the cell, a previously described signature of carboxysome organization (20). The potential for encapsulation is known to be a specific feature of only a subset of Form 1 enzymes. For example, some carboxysome-containing organisms contain two Form 1 RuBisCOs – a carboxysomal and noncarboxysomal version; the latter lacks the ability to incorporate in the carboxyosme (19). Our results indicate that the ancestral large subunits have the propensity for encapsulation – a property whose selection may have allowed successful lineages to avoid extinction during the changing Proterozoic atmosphere.

RuBisCO encapsulation may have been the first step towards the evolution of the carboxysome, just as RuBisCO aggregation increases the local enzyme concentration thus increasing overall carbon fixation (e.g. algal pyrenoids). After RuBisCO aggregation, the association with carbonic anhydrase (CA) was likely a subsequent step in the evolution of the carboxysome; this hypothesis is supported by the presence of independent and distinct classes of CA homologs, (β-CA in α-carboxysomes and γ-CA in β-carboxysomes), encoded in all core carboxysome operons (21). The origins of the α- and β- carboxysome most likely occurred during the Phanerozoic due to the dramatic decrease in CO_2 and increase in O_2. Prior to this, the gradual changes during the later Proterozoic most likely provided weak selective pressure on ancestral RuBisCOs to improve their enzymatic properties towards the protein landscape optimum, which all extant RuBisCOs lie upon today. Subsequently, stronger selective pressures during the Phanerozoic forced CCMs, including carboxysomes, to evolve, as solely relying upon improved RuBisCO kinetics was no longer sufficient. Finally, our assumptions on the timing of events would suggest that plants and eukaryotes within the Archaeplastida lineage lack carboxysomes, because carboxysomes arose 1 – 0.5 Gya after the primary plastid endosymbiosis, consistent with the fact that no Archaeplastidal genome has been shown to contain any carboxysome components. Nonetheless, our data do not exclude the possibility of a more ancient origin of the carboxysome, as previously suggested (22), given the ability of ancestral RuBisCOs to be encapsulated in extant carboxysomes.

Efforts to engineer and improve upon extant RuBisCO have been largely unsuccessful (23); this may be due to extant RuBisCOs stalled in a local optimum of the protein fitness landscape. Ancestral sequence reconstruction provides a unique platform to go back in time, opening the possibilities of forward and reverse engineering on an enzyme that has not yet been subjected to the selective pressures of history. It will be interesting to introduce ancestral RuBisCO back into organisms in long-term adaptive evolution studies and observe if the proteins follow the same evolutionary trajectory, essentially “replaying life’s tape” (24).

Methods:

Alignment and Phylogeny. 125 RbcL and 131 RbcS sequences spanning the Form 1 RuBisCO clade were aligned using structural information from a variety of RuBisCO crystal structures in the Protein Data Bank (PDB IDs: 1SVD, 1IR1, 1GK8, 1RBL) using the PROMALS3D (25). Maximum-likelihood phylogenies were generated using PhyML (26) with 100 bootstrap replicates. The LG amino acid substitution model was chosen based on ProtTest (27) with gamma-distributed variation (four categories) and estimation of a proportion of variable sites. The tree was rooted to the Form 1C and 1D monophyletic subclade.

Ancestral Sequence Reconstruction. Maximum-likelihood methods implemented in PAML were used to resurrect ancestral sequences (15). Posterior probabilities were calculated for all amino acid residues across the sequence, and the residue with the highest probability was assigned to each site. Estimation of the positions of ancestral gaps due to insertions and deletions was predicted as described by Hall (2006) (28).

Protein Purification. Reconstructed sequences were synthesized and codon-optimized for expression in E. coli (Genscript). Corresponding RbcL and RbcS sequences were synthesized and subcloned into the pET11a vector as previously described (29). Modified pET11a vectors and pBAD33*ES*/*EL* were co-transformed into *E. coli* BL21 (DE3) cells. Protein expression was performed as previously described (29). Fully assembled L_8S_8 complexes from the α/β-MRCA construct could not be isolated, possibly due to expression issues.

For determination of kinetic activity, preparations of ancestral α-MRCA and β-MRCA RuBisCO were obtained from *E.coli* cultures after one step of extraction by sonication followed by one step of size exclusion chromatography. The harvested cultures expressing α-MRCA and β-MRCA RuBisCOs were resuspended in buffer containing 0.1M bicine-NaOH pH 8.0, 20 mM $MgCl_2$, 50 mM $NaHCO_3$, 2mM PMSF, bacterial protease inhibitor cocktail (Sigma-Aldrich) and sonicated 6 times for 15 seconds on ice. Fractions containing RuBisCO were then selected from the sonicated-clarify supernatant using PD-10 columns (GE-Healthcare) pre-equilibrated with buffer containing 0.1M bicine-NaOH pH 8.1, 10 mM $MgCl_2$, 1 mM EDTA, 1 mM ε-aminocaproic acid, 1 mM benzamidine, 1 mM KH_2PO_4, 2 % (w/v) PEG-4000, 10 mM $NaHCO_3$ and 5 mM DTT.

For determination of specificity factor, pure preparations of α-MRCA and β-MRCA RuBisCO were obtained from *E.coli* cultures after several step of purification. The harvested cultures expressing α-MRCA and β-MRCA RuBisCO were sonicated as before, obtaining clarify-supernatants containing RuBisCO. The supernatants were diluted to have a final concentration of 20.5 % PEG-4000 and 20 mM $MgCl_2$, incubated for 30 min on ice and centrifuged for 20 min at 12,000 rpm. At this concentration of PEG-4000 and $MgCl_2$ RuBisCO can be precipitated. The pellets were subjected to a first step of anion-exchange chromatography using HiTrap Q-5 ml columns (GE-Healthcare) pre-equilibrated with Q-buffer pH 8 containing 25 mM TEA, 5 mM $MgCl_2$, 0.5 mM EDTA 1 mM ε-aminocaproic, 1 mM benzamidine, 12.5 % (v/v) glycerol, 2 mM DTT, 5 mM $NaHCO_3$. The column was then developed with a 0 – 600 mM NaCl gradient in Q-buffer and the fractions containing the highest RuBisCO activity were selected. In a second step, fractions-containing RuBisCO were desalted by size-exclusion chromatography using Sephacryl S-200 columns (GE-Healthcare) pre-equilibrated with S-200 buffer (50 mM bicine-NaOH pH 8, 20 mM $MgCl_2$, 0.2 mM EDTA, 2 mM DTT). Finally the preparations containing RuBisCO were subjected to a last step of ultrafiltration and desalting using 20 ml/150 K concentrator (Thermo Pierce).

RuBisCO activity assays. The carboxylase and oxygenase kinetic parameters (V_{max} and K_m) were determined simultaneously using Michaelis-Menten equations knowing the amount of ^{14}C incorporated in the substrate RuBP at known O_2 concentrations present in reaction buffer at 25°C.

4 sets of 6 Pico vials each containing 0.85 ml of CO_2-free assay buffer (235 mM bicine-NaOH pH 8.1, 23.5 mM $MgCl_2$, 20 μg/ml carbonic anhydrase) and different concentrations of $NaH^{14}CO_3$ were connected to 4 gas lines for simultaneous supply of nitrogen containing 21 %, 0 %, 60 % and 100 % O_2. After 60 minutes of incubation at 25°C to encourage equilibration of CO_2 between liquid and gas phases, 15 μl of 26.7 mM RuBP were added to each set of vials. Each reaction was then started adding one-by-one 25 μl of previously activated RuBisCO (incubated with the substrate $NaH^{14}CO_3$) and after 1 minute quenched by adding 100 μl of formic acid. The amount of ^{14}C incorporated in the substrate RuBP was measured using a scintillation spectrometer. 2 controls for monitoring change in activity during course of experiment and 2 negative controls without RuBP and using RuBisCO previously incubated with the inhibitor CABP were performed to validate the experiments.

The specificity factor for α-MRCA and β-MRCA RuBisCO was determined at 25°C by the total consumption of RuBP in an oxygen electrode vessel. Purified RuBisCO preparations were dissolved and desalted by spin-desalt protocol using G50 Sephadex columns (Helmerhorst and Stokes 1980) previously equilibrated with a buffer CO_2-free containing 0.1 M bicine-NaOH pH 8.2 and 20 mM $MgCl_2$. An adequate volume of desalted RuBisCO preparations were resuspended to a final concentration of 10 mM $NaH^{14}CO_3$ and 4 mM orthophosphate solution pH 8.2 and then incubated at room temperature for 40 minutes for total RuBisCO activation. For each essay 1 ml of reaction mixture was prepared directly in an oxygen electrode vessel (Model DW1; Hansatech, Kings Lynn., UK) adding in order the following components: 0.95 ml of CO_2-free buffer equilibrated with CO_2-free air at 25°C and containing 0.1 M bicine-NaOH pH 8.2, 20 mM $MgCl_2$ and 1.5 mg (7000 W-A units) per 100 ml of carbonic anhydrase; 20 μl ml of 0.1 M $NaH^{14}CO_3$; 10 μl of 15 mM RuBP; finally the reaction was started adding 20 μl of activated RuBisCO preparation containing enough enzyme. RuBisCO is added to start the reaction in order to avoid the decarbamylation of the enzyme before starting the reaction. 0.1 ml of the reaction mixture was quenched adding 0.1 ml of formic acid and used for estimation of RuBP carboxylation by the amount of ^{14}C incorporated in the product. The amount of RuBP oxygenation was then calculated from the electrode trace of oxygen consumption. A series of assays containing wheat RuBisCO were performed in parallel and the results obtained for α-MRCA and β-MRCA RuBisCO where normalized to the overage value of wheat RuBisCO. Comparisons of kinetic data to extant RuBisCO are summarized in Appendix C, Table S1.

***Synechococccus* strains.** All constructs were cloned using BioBrick Assembly standard 21 (BglBrick assembly) format (30) in *E. coli* and subsequently cloned into neutral site vector pAM1573PMS for genomic integration into the *Synechococcus* genome at Neutral Site 2. pAM1573PMS was modified from pAM1573 (31) to be BglBrick compatible. Ancestral RbcL sequences were fused to a C-terminal Cerulean Fluorescent Protein (CFP) and constitutively expressed with the rplC promoter, as described by Savage et al (2010) (20). Wild-type *Synechococcus* was grown in BG11 medium with constant light at 30° Celsius. Wild-type cells were transformed and selected for on BG11 plates with antibiotics. For co-localization studies, strains expressing CFP-tagged ancestral RbcL were subsequently transformed with another vector for expressing carboxysomal protein (CcmN) fused to a C-terminal Enhanced Yellow Fluorescent Protein (EYFP) driven with a CcmK2 promoter. This vector (pAM2314PMS) is a modified version of a BglBrick modified version of pAM2314 (31) and mediates genomic integration into the *Synechococcus* genome at Neutral Site 1. Appendix C, Table S2 describes the various strains used in this study.

Fluorescence Microscopy. Cells grown on solid BG11 media were spotted on to 1% agarose pads (w/v in BG11) in a 16 well chamber slide (Lab-Tek, Scotts Valley, CA) and covered with a 0.17 mm coverglass. Images were acquired on a Zeiss LSM 710 inverted confocal microscope (Carl Zeiss Inc, Thornwood, NY) using laser lines at 405, 514, and 633 nm and a 63x/1.4 NA oil-immersion objective. Images were captured using Zen 2010 (Carl Zeiss, Inc.) and analyzed using ImageJ(32).

Acknowledgments:
We thank the Hayer-Hartl lab for providing the pBAD33*ES*/*EL* vector. P.M.S., J.C.C. and C.A.K. were supported by the NSF (MCB0851054, C.A.K. and P.M.S.; EF 1105897, C.A.K. and J.C.C.) and Arpa-E DE-0000200 A.O., J.A., and M.A.J.P. are supported by the BBSRC 20:20 Wheat Institute Strategic Program (BBSRC BB/J/00426X/1 20:20 Wheat®) and BBSRC BB/I002545/1, BB/I017372/1. P.J.A, M.A.J.P and A. O. acknowledge support through BBSRC BB/1024488/1.

Author contributions:
P.M.S. designed and conducted the experiments, interpreted the data and wrote the paper. J.C.C performed microscopy. A.O., J.A., and M.A.J.P characterized the ancestral assemblies, measured kinetics and interpreted the data. C.A.K supervised the project, interpreted the data and wrote the paper.

References:

1. Ellis RJ (1979) The most abundant protein in the world. *Trends Biochem Sci* 4(11):241-244.
2. Jordan DB & Ogren WL (1983) Species variation in kinetic properties of ribulose 1,5-bisphosphate carboxylase/oxygenase. *Arch Biochem Biophys* 227(2):425-433.
3. Tcherkez GGB, Farquhar GD, & Andrews TJ (2006) Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized. *Proc Natl Acad Sci* 103(19):7246-7251.
4. Gaucher EA, Thomson JM, Burgan MF, & Benner SA (2003) Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* 425(6955):285-288.
5. Tabita FR*, et al.* (2007) Function, Structure, and Evolution of the RubisCO-Like Proteins and Their RubisCO Homologs. *Microbiol Mol Biol Rev* 71(4):576-599.
6. Butterfield NJ, Knoll AH, & Swett K (1990) A bangiophyte red alga from the Proterozoic of arctic Canada. *Science* 250:104-107.
7. Douzery EJP, Snell EA, Bapteste E, Delsuc F, & Philippe H (2004) The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils? *Proc Natl Acad Sci* 101(43):15386-15391.
8. Yoon HS, Hackett JD, Ciniglia C, Pinto G, & Bhattacharya D (2004) A Molecular Timeline for the Origin of Photosynthetic Eukaryotes. *Mol Biol Evol* 21(5):809-818.
9. Parfrey LW, Lahr DJG, Knoll AH, & Katz LA (2011) Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc Natl Acad Sci* 108(33):13624-13629.
10. Berner RA & Kothavala Z (2001) Geocarb III: A Revised Model of Atmospheric CO2 over Phanerozoic Time. *Am J Sci* 301(2):182-204.
11. Kasting JF (1993) Earth's early atmosphere. *Science* 259(5097):920-926.

12. Sessions AL, Doughty DM, Welander PV, Summons RE, & Newman DK (2009) The Continuing Puzzle of the Great Oxidation Event. *Curr Biol* 19(14):R567-R574.
13. Badger MR, Hanson D, & Price GD (2002) Evolution and diversity of CO2 concentrating mechanisms in cyanobacteria. *Func Plant Biol* 29(3):161-173.
14. Canfield DE & Teske A (1996) Late Proterozoic rise in atmospheric oxygen concentration inferred from phylogenetic and sulphur-isotope studies. *Nature* 382(6587):127-132.
15. Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* 24(8):1586-1591.
16. Whitney SM, Baldet P, Hudson GS, & Andrews TJ (2001) Form I Rubiscos from non-green algae are expressed abundantly but not assembled in tobacco chloroplasts. *Plant J* 26(5):535-547.
17. Bainbridge G*, et al.* (1995) Engineering Rubisco to change its catalytic properties. *J Exp Bot* 46(special issue):1269-1276.
18. Savir Y, Noor E, Milo R, & Tlusty T (2010) Cross-species analysis traces adaptation of Rubisco toward optimality in a low-dimensional landscape. *Proc Natl Acad Sci* 107(8):3475-3480.
19. Menon BB, Dou Z, Heinhorst S, Shively JM, & Cannon GC (2008) Halothiobacillus neapolitanus Carboxysomes Sequester Heterologous and Chimeric RubisCO Species. *PLoS ONE* 3(10):e3570.
20. Savage DF, Afonso B, Chen AH, & Silver PA (2010) Spatially Ordered Dynamics of the Bacterial Carbon Fixation Machinery. *Science* 327(5970):1258-1261.
21. Zarzycki J, Axen SD, Kinney JN, & Kerfeld CA (2013) Cyanobacterial-based approaches to improving photosynthesis in plants. *J Exp Bot* 64(3):787-798.
22. Giordano M, Beardall J, & Raven JA (2005) CO2 Concentrating Mechanisms in Algae: Mechanisms, Environmental Modulation, and Evolution. *Annu Rev Plant Biol* 56(1):99-131.
23. Zhu XG, Portis AR, & Long SP (2004) Would transformation of C3 crop plants with foreign Rubisco increase productivity? A computational analysis extrapolating from kinetic properties to canopy photosynthesis. *Plant Cell Environ* 27(2):155-165.
24. Gould SJ (1989) *Wonderful life: the Burgess Shale and the nature of history* (W W Norton & Company Incorporated) p 347.
25. Pei J, Kim B-H, & Grishin NV (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res* 36(7):2295-2300.
26. Guindon S*, et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Sys Biol* 59(3):307-321.
27. Abascal F, Zardoya R, & Posada D (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21(9):2104-2105.
28. Hall BG (2006) Simple and accurate estimation of ancestral protein sequences. *Proc Natl Acad Sci* 103(14):5431-5436.
29. Saschenbrecker S*, et al.* (2007) Structure and Function of RbcX, and Assembly Chaperone for Hexadecameric Rubisco. *Cell* 129(6):1189-1200.

30. Anderson JC*, et al.* (2010) BglBricks: A flexible standard for biological part assembly. *J Biol Eng* 4(1):1.

31. Mackey SR, Ditty JL, Clerico EM, & Golden SS (2007) Detection of Rhythmic Bioluminescence From Luciferase Reporters in Cyanobacteria. *Circadian Rhythms,* Methods in Molecular Biology™, ed Rosato E (Humana Press), Vol 362, pp 115-129.

32. Abramoff MD, Magalhaes PJ, & Ram SJ (2004) Image processing with ImageJ. *Biophotonics International* 11:36-42.

Figures:

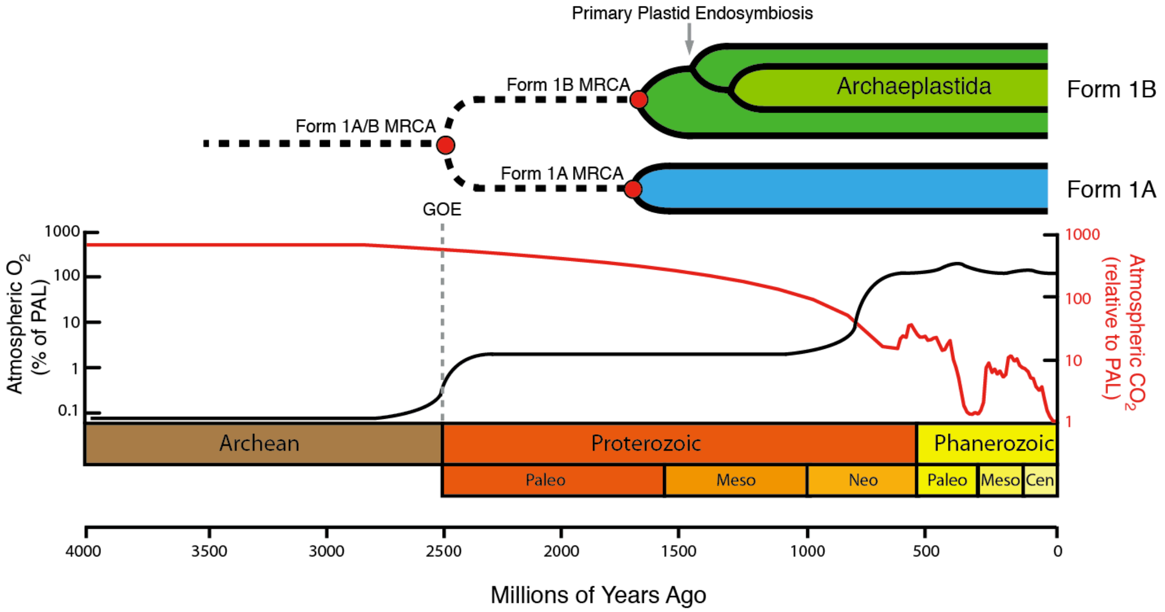


Figure 1. Model of the evolutionary timeline of RuBisCO and corresponding O_2 and CO_2 atmospheric concentrations. Estimated O_2 concentrations (black line) over geologic time are based on the percent of PAL. Grey dashed line below represents the Great Oxidation Event (GOE).

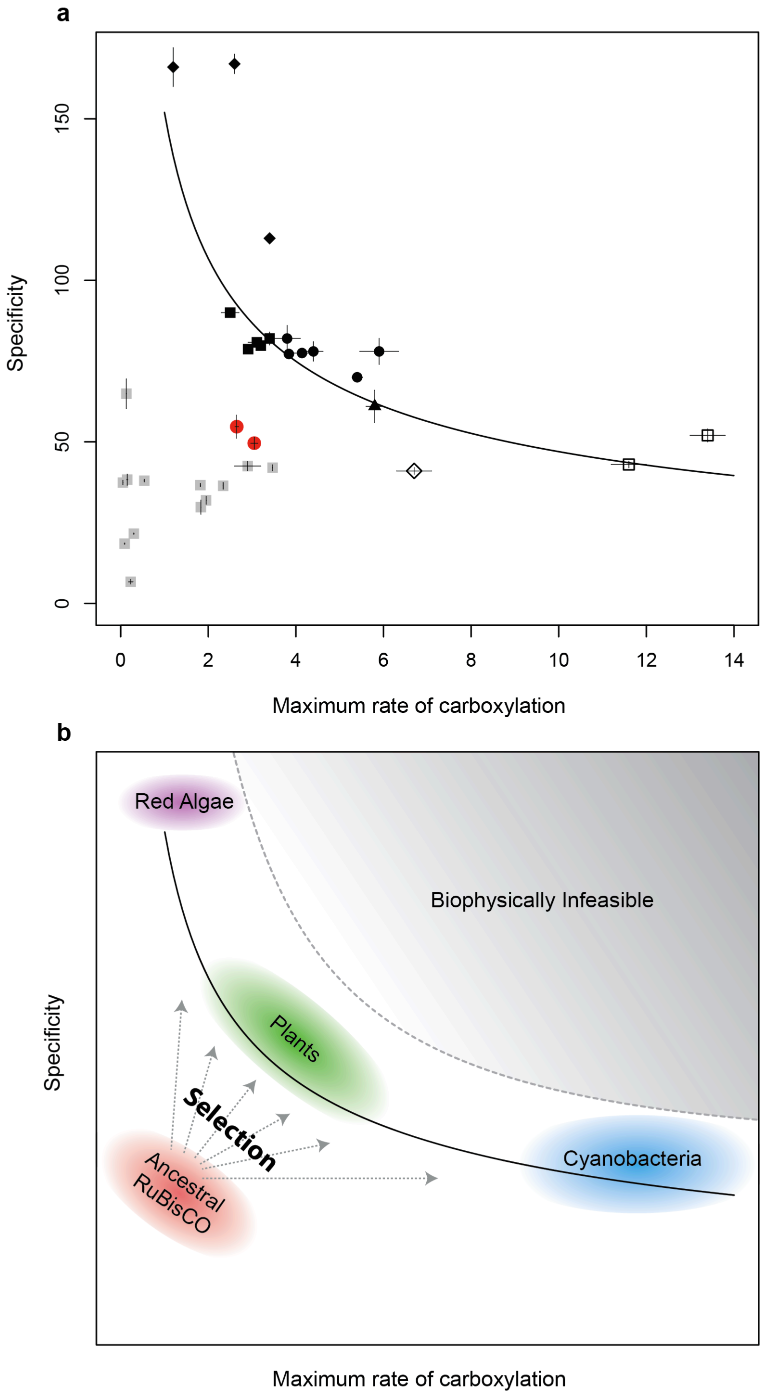


Figure 2. Comparison of ancestral and extant RuBisCO. a, Specificity factor (τ) versus carboxylation rate (V_c) for characterized RuBisCOs. Best-fit curve for extant RuBisCOs (black line) previously described(18). [Form 1A (open diamonds); Form 1B Cyanobacteria (open squares); eukaryotic green algae (filled triangles); eukaryotic non-green algae (filled diamonds); C4 plants (filled circles); C3 plants (filled squares); ancestral (red circles); point mutant or chimeric (grey squares)]. Values and error bars summarized in Appendix C, Table S1. **b**, Model of selective pressures pushing properties of RuBisCO towards the hypothesized protein landscape optimum – upper limits of the kinetic parameters – represented by the best-fit curve (black line).

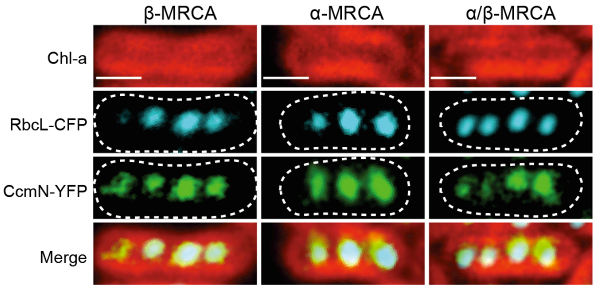


Figure 3. Encapsulation of ancestral RuBisCO in carboxysomes of extant cyanobacteria. Ancestral β-MRCA, α-MRCA, and α/β-MRCA RbcL subunits fused to CFP (blue) co-localize with carboxysomal subunit, CcmN, fused to YFP (green) in *Synechococcus elongatus* PCC 7942. Chlorophyll-a (Chl-a) fluorescence from the thylakoid membrane is shown in red. All strains exhibit spatially distributed fluorescent puncta typical of carboxysome localization. Scale bars, 1 μm.

Table 1. Characterization of Ancestral Form 1A and Form 1B RuBisCO. Data collected at 25°C. Data further summarized in Appendix C, Table S1.

RuBisCO	Vcmax	Vomax	K_Mc	K_mo	Specificity factor		Vc/K_Cair
	(μmol / min / mg Rubisco)		(μM)				
Ancestral Form 1A	2.31 ± 0.04	0.87 ± 0.08	113 ± 6	2,329 ± 208	54.7 ± 3.5	n = 6	21.07
Ancestral Form 1B	2.66 ± 0.08	0.28 ± 0.02	120 ± 10	641 ± 49	49.6 ± 1.8	n = 6	18.02
Extant Form 1A (Chromatium)	6.7 ± .4[a]	-	37 ± 2[a]	290 ± 25[a]	41 ± 1[a]	-	94.96
Extant Form 1B (Synechococcus)	13.4 ± .4[a]	-	246 ± 20[a]	1300 ± 130[a]	52 ± 2[a]	-	51.25

Chapter 5

Conclusions

The rise of oxygenic photosynthesis changed the primordial biogeochemical landscape of Earth and continues to sustain a majority of global life. The works presented in this thesis are a small contribution to the largely open-ended and highly speculative field of understanding the origins of oxygenic photosynthesis and their supposed cyanobacterial inventors. The ambiguity in the field attests to the difficulty in studying such ancient events, whereas the importance of the subject emphasizes the necessity of major improvements in the datasets and methods used to address these questions.

Many impending global crises concern agriculture, such as food security and climate change. Thus, the pivotal and potential role of photosynthesis in addressing these issues has been of great interest to both the popular and scientific communities. However, in times when there is growing pressure to deliver sustainable energies, I believe it will be pertinent to not only invest in applied engineering efforts of globally relevant biochemical metabolisms, such as photosynthesis, but also the basic sciences. Fundamental knowledge of the evolution of these organisms and enzymes will have an equal role in contributing to these efforts by providing insightful perspective.

Humans have successfully domesticated plant species to meet our needs for over ten thousand years. Thus, it is not shocking that in the face of our current problems, we may one day be able to go beyond domestication and directly engineer desired traits into our crops. There are clear examples where this has already been successful, such as Roundup Ready crops. However, when translating these technologies into more complex traits such as improving photosynthesis, a more comprehensive understanding of the basic biology and evolution will be essential. Fundamental questions in the field are still unanswered or highly controversial, such as the role of photorespiration, the capacity of introducing synthetic carbon fixation pathways, etc. I believe that shedding light on the origin and evolution of these topics will provide answers to some of these controversial questions, but more importantly, the insight gained may guide future efforts in improving photosynthetic yield and addressing our future agricultural needs.

Appendix A

Supplemental Information for Chapter 2

Supplementary Materials and Methods:

The 54 strains used for genome sequencing in this study are available at Pasteur Culture collection of Cyanobacteria (http://www.pasteur.fr/pcc_cyanobacteria). The 54 sequenced genomes in this study were compared to 72 publicly available cyanobacterial genomes (Table S1).

A sequence similarity matrix was calculated for alignments of 1,813 16S small subunit rRNA sequences of cyanobacterial isolates from the greengenes database, excluding sequences from environmental samples (December 2008). The cyanobacterial isolates were grouped into 104 clusters by MCL clustering performed on the sequence similarity matrix at similarity cutoff of 95% and inflation value of 2. Type strains, PCC identification numbers and the status of previous sequencing efforts were highlighted for all the isolates in the 104 clusters. This analysis, interest of the strains to the research community and their availability at the Pasteur Culture Collection, was used as guide to choose the strains for genome sequencing. For strains chosen, 1.25 L of liquid cultures in late exponential to linear growth phase were centrifuged at 12,000*g* for 10min at 20°C. After washing twice with sterile distilled water or sterile saline solution (1% NaCl) for marine strains, the pellets were immediately frozen in liquid N2 prior to being lyophilized. DNA of the lyophilized pellets was extracted using Genomic DNA isolation - NucleoBond ® AX (Macherey-Nagel, Hoerdt, France) according manufacturer's instructions for bacterial DNA using the columns Nucleobond AX-G 500.

Genome sequencing and assembly

The 54 CyanoGEBA draft genomes were generated at the DOE Joint Genome Institute (JGI) using either a combination of Illumina (1) and 454 technologies (2) or the Illumina technology (Table S10). The 454 Titanium standard data and the 454 paired end data were assembled using Newbler, versions 2.3 to 2.6, and the resulting consensus sequences were computationally shredded into 2 Kbp overlapping fake reads (shreds). Illumina sequencing data was assembled with Velvet, versions 0.7.55 to 1.105 (3) , and the consensus sequence computationally shredded into 1.5 Kbp overlapping fake reads (shreds). The 454 Newbler consensus shreds, the Illumina Velvet consensus shreds and the read pairs in the 454 paired end library were then integrated using parallel Phrap, version SPS - 4.24 (High Performance Software, LLC). The software Consed (4),(5) (6) was used in the following finishing process. Illumina data were used to correct potential base errors and increase consensus quality using the software Polisher developed at JGI. Possible mis-assemblies were corrected using gapResolution, Dupfinisher (7), or sequencing cloned bridging PCR fragments with subcloning. Gaps between contigs were closed by editing in Consed, by PCR and by Bubble PCR primer walks. All general aspects of library construction and sequencing performed at the JGI can be found at http://www.jgi.doe.gov/. At Los Alamos National Laboratory (LANL), 25 of these genomes underwent manual finishing efforts, while 20 others underwent autofinishing.

Gap closure in autofinishing is fully automated and thus less extensive as compared to manually finishing. The 9 remaining CyanoGEBA genomes were not subjected to finishing efforts. For PCC 9605 and PCC 10914, all raw Illumina sequence data were passed through DUK, a filtering program developed at JGI, which removes known Illumina sequencing and library preparation artifacts. Illumina sequence reads were assembled using Allpaths-LG versions 38118 (PCC 9339), 38445 (PCC 9431, PCC 10914, PCC 7702) and 39750 (PCC 9605). For PCC 73106, PCC 7509, and PCC 6406, following steps were performed for genome assembly: 1) filtered Illumina reads were assembled using Velvet (3), 2) 1-3 Kbp simulated paired end reads were created from Velvet contigs using wgsim (https://github.com/lh3/wgsim), 3) Illumina reads were assembled with simulated read pairs using Allpaths-LG (versions 37843 and 38118) (8).

Genome annotation

Genes were identified using Prodigal (9), followed by a round of manual curation using GenePRIMP (10) for finished genomes and draft genomes in fewer than 10 scaffolds. The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) nonredundant database, UniProt, TIGRFam, Pfam, KEGG, COG, and InterPro databases. The tRNAScanSE tool (11) was used to find tRNA genes, whereas ribosomal RNA genes were found by searches against models of the ribosomal RNA genes built from SILVA (12). Other non-coding RNAs such as the RNA components of the protein secretion complex and the RNase P were identified by searching the genome for the corresponding Rfam profiles using INFERNAL (http://infernal.janelia.org). Additional gene prediction analysis and manual functional annotation was performed within the Integrated Microbial Genomes (IMG) platform (http://img.jgi.doe.gov) developed by the Joint Genome Institute, Walnut Creek, CA, USA (13).

Species tree

The Species tree was generated by a concatenation of thirty-one conserved proteins as described by Wu et al. (14). Homologs of each ribosomal protein were identified using reciprocal BLAST of the 49 publicly available cyanobacterial genomes in IMG at the end of 2009. These gene families were aligned using MAFFT, using the maxiterative function (15). The subsequent alignment was used to create Hidden Markov Models (HMMs) for the respective ribosomal protein using HMMer v.2.0 (16). Total protein coding sequences for each cyanobacterial genome, and of four outgroups (*Chloroflexus auranticus* J-10, *Rhodobacter sphaeroides* 2.4.1, *Heliobacterium modesticaldum* Ice1, and *Chlorobium tepidum* TLS) were retrieved using IMG (13). Using HMMer, the hmmsearch function was used to identify orthologs and align them using the hmmalign function. The resulting thirty-one alignments were then concatenated. The default setting to omit gappy columns was used with the software Belvu (17). A phylogenetic tree was generated with the alignment using PhyML (18). The LG amino acid substitution model was chosen using ProtTest with gamma-distributed rate variation (four categories) and estimation of a proportion of invariable sites (19).

Tree Imbalance study

Two trees, one with all cyanobacterial genomes (126 species) and one with only the 72 publicly available were generated. The alignments and the phylogenetic trees were generated using the same methods described to construct the Species Tree. *Gloeobacter violaceus* PCC 7421 was set as the outgroup in both trees. The tree imbalance of both trees was measured using Colless' Imbalance in the software Mesquite (20, 21). The tree depth was set to 10 and 1000 simulations of both uniform and equiprobable speciations were conducted.

16S rRNA phylogeny

A phylogeny using 16S rRNA sequences retrieved from IMG for all cyanobacteria of this study was generated to compare to the Species tree. Due to incomplete or partial sequences, *Arthrospira* sp. PCC 8005, *Synechococcus* sp. CB0101, *Synechococcus* sp. CB0205, and *Crocosphaera watsonii* WH 0003 were omitted from this phylogeny. Sequences were aligned in MAFFT. A maximum likelihood tree was generated using PhyML, using the GTR model with gamma-distributed rate variation (four categories) and an estimation of proportion of invariable sites.

Identification of novel proteins

All 292,935 proteins from the CyanoGEBA genomes were searched against the entire amino acid non-redundant (nr) database downloaded from NCBI, updated April 2nd, 2012, using BLASTP set at an e-value cutoff of 1e-2. The 21,107 proteins with no hits were considered 'novel' as they have no homology to the nr database.

Morphological transitions analysis

Protein families generated from MCL analysis was used. The specific nodes tested for morphological transitions are indicated in Fig. 1. A set of genes involved in the morphological transition were defined by comparison of presence in one genome or a set of genome belonging to a subsection and their absence in another genome or a set of genomes as reported in Table S5. Moreover, a BLASTP search of the 32 proteins from *Prochlorothrix hollandica* PCC 9006 from Event 2 against the 674 proteins from Event 3 was done, yielding 29 out of the 32 hits. We generated a null hypothesis to verify the enrichment in 29 out of the 32 homologous proteins by randomly sampling the *Prochlorothrix hollandica* PCC 9006 genome against the 674 proteins from Event 3 with BLASTP, 10,000 times, which showed that the value (29 out of 32 proteins) was significant (p-value = 0).

Heterocyst, hormogonium, and akinete related gene distribution analysis

Seed proteins (29 and 20 are involved in cell division and cell differentiation, respectively) were downloaded from the cyanobase (http://genome.kazusa.or.jp/cyanobase) and used for BLAST comparison searches. Putative orthology relationships between a seed protein and other cyanobacterial proteins were defined by an alignment threshold of at least 30% sequence identity with an e-value lower than 1e-10.

COG functional categories

COG functional category data was downloaded by Morphological Subsection from the IMG database.

Plastidome tree

The plastidome tree was generated by a concatenation of twenty-five conserved plastid proteins using the same method to generate the *Cyanobacteria* tree. Proteins from fully sequenced plastid genomes were downloaded from the High-quality Automated and Manual Annotation of microbial Proteins (HAMAP) database (22). Plastids downloaded from HAMAP were: *Cyanophora paradoxa*, *Chaetosphaeridium globosum*, *Anthoceros formosae*, *Cycas taitungensis*, *Arabidopsis thaliana*, *Amborella trichopoda*, *Selaginella uncinata*, *Zygnema circumcarinatum*, *Staurastrum punctulatum*, *Chara vulgaris*, *Nephroselmis olivacea*, *Ostreococcus tauri*, *Bigelowiella natans*, *Chlorella vulgaris*, *Pseudendoclonium akinetum*, *Pseudendoclonium akinetum*, *Oltmannsiellopsis viridis*, *Scenedesmus obliquus*, *Chlamydomonas reinhardtii*, *Stigeoclonium helveticum*, *Oedogonium cardiacum*, *Euglena gracilis*, *Mesostigma viride*, *Chlorokybus atmophyticus*, *Cyanidioschyzon merolae*, *Cyanidium caldarium*, *Porphyra yezoensis*, *Porphyra purpurea, Gracilaria tenuistipitata* var. liui, *Rhodomonas salina*, *Guillardia theta*, *Emiliania huxleyi*, *Phaeodactylum tricornutum*, *Odontella sinensis*, *Thalassiosira pseudonana*, *Vaucheria litorea*, *Heterosigma akashiwo* CCMP452, *Heterosigma akashiwo* NIES293, and the chromatophore of *Paulinella chromatophora*. A phylogenetic tree was generated with the alignment using PhyML 3.0. The LG amino acid substitution model was chosen by ProtTest and with gamma-distributed rate variation (four categories) and estimation of a proportion of invariable sites. The tree was rooted to *Gloeobacter violaceus* PCC 7421.

Prediction of Endosymbiotic Gene Transfer.

Proteins from the genomes used in this study were divided into four groups: 1) Nuclear genomes from plastid-containing eukaryotes (Table S8), 2) Bacteria not from the phylum *Cyanobacteria* (*Agrobacterium tumefaciens* C58-Cereon, *Aquifex aeolicus* VF5, *Bacillus subtilis subtilis* 168, *Caulobacter crescentus* CB15, *Chlamydia trachomatis* E/150, *Chlorobium limicola* DSM 245, *Chloroflexus aurantiacus* J-10-fl, *Heliobacterium modesticaldum* Ice1, Candidatus *Kuenenia stuttgartiensis*, *Rickettsia peacockii* Rustic, *Thermotoga maritima* MSB8), 3) Archaea (*Archaeoglobus fulgidus* VC-16, DSM 4304, *Cenarchaeum symbiosum* A, *Methanocaldococcus jannaschii* DSM 2661, *Nanoarchaeum equitans* Kin4-M, *Sulfolobus acidocaldarius* DSM 639), 4) Eukaryotes presumably not containing plastids derived from endosymbiosis (*Caenorhabditis elegans* Bristol N2, *Cryptococcus neoformans* var. neoformans JEC 21, *Drosophila melanogaster*, *Monosiga brevicollis* MX1, *Saccharomyces cerevisiae* S288C). The nuclear proteins from Group 1 were used as queries to BLASTP against two databases: 1) all proteins from Groups 2-4 and all cyanobacterial proteins in this study (CyanoGEBA and publicly-available genomes), and 2) all proteins from Groups 2-4 and cyanobacterial proteins from only publicly-available genomes. Those with top-hits to cyanobacterial proteins were considered genes of cyanobacterial descent, and the total counts for each of the nuclear genomes from Group 1 are described in Table S8 and Table S11. COGs for all proteins were assigned using the same methods as in the IMG pipeline (13).

Chlorophyll Binding Protein (CBP) studies

Phylogenetic analysis

CBP homologs were collected by performing a BLASTP search on all cyanobacteria in the IMG database using the inner chlorophyll-binding antenna protein CP43 of PSII from *Thermosynechococcus elongatus* BP-1 as the query, setting and e-value threshold of 1e-10. All homologs were aligned using MAFFT. The alignment was used to build a maximum likelihood phylogenetic tree in PhyML, under the LG model with gamma-distributed rate variation (four categories) and an estimation of a proportion of invariable sites, after choosing the best-suited model in ProtTest.

Alignment and analysis of chlorophyll binding amino acids

An alignment of a subset of CBP proteins was generated in order to investigate the presence of conserved amino acids that are known to ligate chlorophyll to the protein. The amino acid sequences for the N- and C- termini of PsaA and PsaB, PsbB, PsbC, and IsiA from *Thermosynechococcus elongatus* were aligned to various CBP; the sequences were aligned using MAFFT, followed by manual curation of the alignment, using only the alignments of the first six helices (Fig. S6).

Further analysis of the C-terminal PsaL-like domain of the CBPV was carried out by truncating CBPV sequences to examine specifically the ladder domain. PsaL subunits from *Synechococcus elongatus* PCC 7942 *Synechocystis* sp. PCC 6803 and *Thermosynechococcus elongatus* BP1 were aligned with the truncated CPBV sequences using MAFFT (Fig. S7).

Homology model

The CBPV homolog from *Chroococcidiopsis thermalis* PCC 7203 (Chro_2988) was submitted to the SWISS-MODEL web server (http://swissmodel.expasy.org/) for three-dimensional structural homology modeling. Two homology models were made. 1) The N-terminal domain (first six transmembrane helices) was homology modeled off the template from the Protein Data Bank, 3ARC_C (the PsbC subunit of Photosystem II from *Thermosynechococcus vulcanus* modeling amino acid positions 6-346). The C-terminal domain (last three transmembrane helices) was modeled off the template, 1JB0_L (the PsaL subunit of Photosystem I from *Synechococcus elongatus* modeling amino acid positions 342-504). The last five amino acids were removed from the N-terminal domain, and the C-terminal domain was positioned near it using PyMol (http://www.pymol.org/). A monomeric subunit of the Photosystem I structure, 1JB0, was used to model the CBPV homolog interaction when replacing the PsaL subunit (Fig. S8).

CRISPR analysis

CRISPRs (Clustered Regularly Interspaced Short Palindromic Repeats) loci were predicted using both CRISPRfinder (23) and CRISPR Recognition Tool (24) (CRT, which is integrated into the IMG pipeline). The presence of CRISPR/Cas systems was confirmed by examining the co-existence of CRISPR loci and the ubiquitous CRISPR-associated (*cas*) genes, namely *cas1* and *cas2*, within one genome.

Figures S1-S9

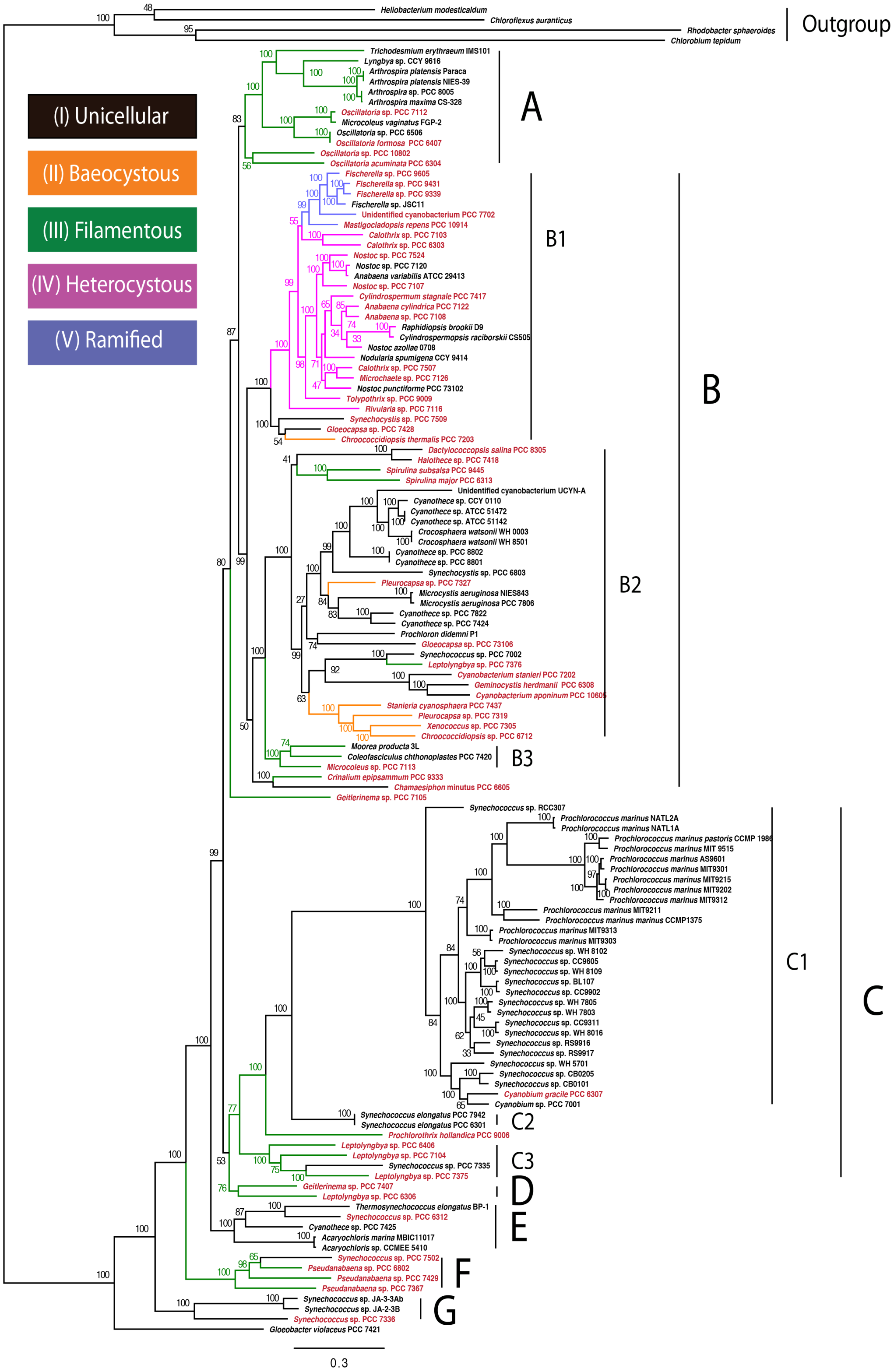

(I) Unicellular
(II) Baeocystous
(III) Filamentous
(IV) Heterocystous
(V) Ramified
Outgroup
Heliobacterium modesticaldum
Chloroflexus auranticus
Rhodobacter sphaeroides
Chlorobium tepidum
A
Trichodesmium erythraeum IMS101
Lyngbya sp. CCY 9616
Arthrospira platensis Paraca
Arthrospira platensis NIES-39
Arthrospira sp. PCC 8005
Arthrospira maxima CS-328
Oscillatoria sp. PCC 7112
Microcoleus vaginatus FGP-2
Oscillatoria sp. PCC 6506
Oscillatoria formosa PCC 6407
Oscillatoria sp. PCC 10802
Oscillatoria acuminata PCC 6304
B
B1
Fischerella sp. PCC 9605
Fischerella sp. PCC 9431
Fischerella sp. PCC 9339
Fischerella sp. JSC11
Unidentified cyanobacterium PCC 7702
Mastigocladopsis repens PCC 10914
Calothrix sp. PCC 7103
Calothrix sp. PCC 6303
Nostoc sp. PCC 7524
Nostoc sp. PCC 7120
Anabaena variabilis ATCC 29413
Nostoc sp. PCC 7107
Cylindrospermum stagnale PCC 7417
Anabaena cylindrica PCC 7122
Anabaena sp. PCC 7108
Raphidiopsis brookii D9
Cylindrospermopsis raciborskii CS505
Nostoc azollae 0708
Nodularia spumigena CCY 9414
Calothrix sp. PCC 7507
Microchaete sp. PCC 7126
Nostoc punctiforme PCC 73102
Tolypothrix sp. PCC 9009
Rivularia sp. PCC 7116
Synechocystis sp. PCC 7509
Gloeocapsa sp. PCC 7428
Chroococcidiopsis thermalis PCC 7203
B2
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
Spirulina subsalsa PCC 9445
Spirulina major PCC 6313
Unidentified cyanobacterium UCYN-A
Cyanothece sp. CCY 0110
Cyanothece sp. ATCC 51472
Cyanothece sp. ATCC 51142
Crocosphaera watsonii WH 0003
Crocosphaera watsonii WH 8501
Cyanothece sp. PCC 8802
Cyanothece sp. PCC 8801
Synechocystis sp. PCC 6803
Pleurocapsa sp. PCC 7327
Microcystis aeruginosa NIES843
Microcystis aeruginosa PCC 7806
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Prochloron didemni P1
Gloeocapsa sp. PCC 73106
Synechococcus sp. PCC 7002
Leptolyngbya sp. PCC 7376
Cyanobacterium stanieri PCC 7202
Geminocystis herdmanii PCC 6308
Cyanobacterium aponinum PCC 10605
Stanieria cyanosphaera PCC 7437
Pleurocapsa sp. PCC 7319
Xenococcus sp. PCC 7305
Chroococcidiopsis sp. PCC 6712
B3
Moorea producta 3L
Coleofasciculus chthonoplastes PCC 7420
Microcoleus sp. PCC 7113
Crinalium epipsammum PCC 9333
Chamaesiphon minutus PCC 6605
Geitlerinema sp. PCC 7105
C
C1
Synechococcus sp. RCC307
Prochlorococcus marinus NATL2A
Prochlorococcus marinus NATL1A
Prochlorococcus marinus pastoris CCMP 1986
Prochlorococcus marinus MIT 9515
Prochlorococcus marinus AS9601
Prochlorococcus marinus MIT9301
Prochlorococcus marinus MIT9215
Prochlorococcus marinus MIT9202
Prochlorococcus marinus MIT9312
Prochlorococcus marinus MIT9211
Prochlorococcus marinus marinus CCMP1375
Prochlorococcus marinus MIT9313
Prochlorococcus marinus MIT9303
Synechococcus sp. WH 8102
Synechococcus sp. CC9605
Synechococcus sp. WH 8109
Synechococcus sp. BL107
Synechococcus sp. CC9902
Synechococcus sp. WH 7805
Synechococcus sp. WH 7803
Synechococcus sp. CC9311
Synechococcus sp. WH 8016
Synechococcus sp. RS9916
Synechococcus sp. RS9917
Synechococcus sp. WH 5701
Synechococcus sp. CB0205
Synechococcus sp. CB0101
Cyanobium gracile PCC 6307
Cyanobium sp. PCC 7001
C2
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301
Prochlorothrix hollandica PCC 9006
C3
Leptolyngbya sp. PCC 6406
Leptolyngbya sp. PCC 7104
Synechococcus sp. PCC 7335
Leptolyngbya sp. PCC 7375
D
Geitlerinema sp. PCC 7407
Leptolyngbya sp. PCC 6306
E
Thermosynechococcus elongatus BP-1
Synechococcus sp. PCC 6312
Cyanothece sp. PCC 7425
Acaryochloris marina MBIC11017
Acaryochloris sp. CCMEE 5410
F
Synechococcus sp. PCC 7502
Pseudanabaena sp. PCC 6802
Pseudanabaena sp. PCC 7429
Pseudanabaena sp. PCC 7367
G
Synechococcus sp. JA-3-3Ab
Synechococcus sp. JA-2-3B
Synechococcus sp. PCC 7336
Gloeobacter violaceus PCC 7421
0.3

Figure S1. Maximum likelihood tree of *Cyanobacteria* with bootstrap support

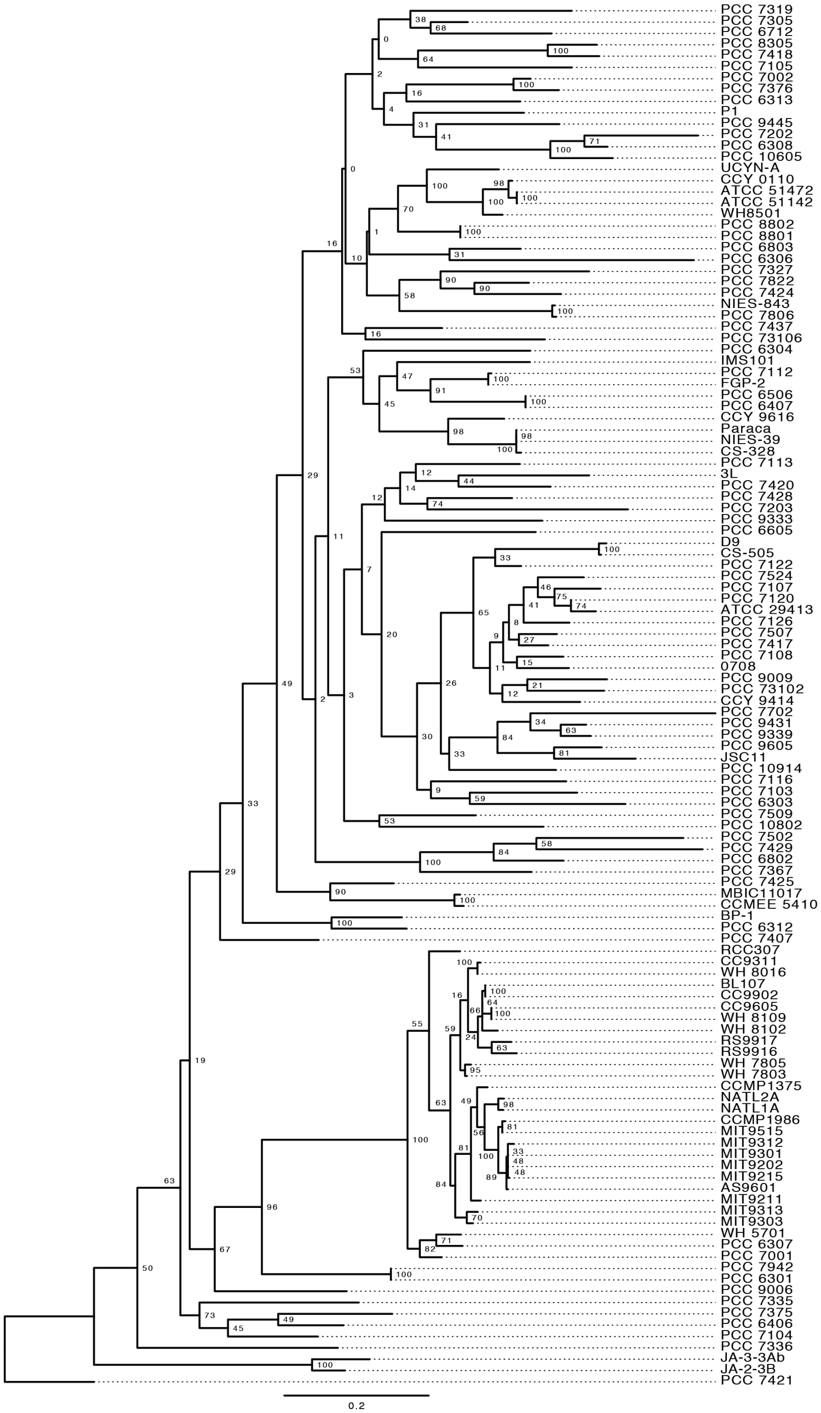


Figure S2. 16S rRNA gene phylogeny of *Cyanobacteria*. Maximum Likelihood tree based on 16S rRNA gene sequences from cyanobacteria included in this study and named accordingly to the Strain_ID in Table S1. Many of the clades defined in Fig. 1 are retrieved in 16S rRNA gene phylogeny. However, given poor bootstrap supports in the latter, there are incongruences between the topologies of the two trees.

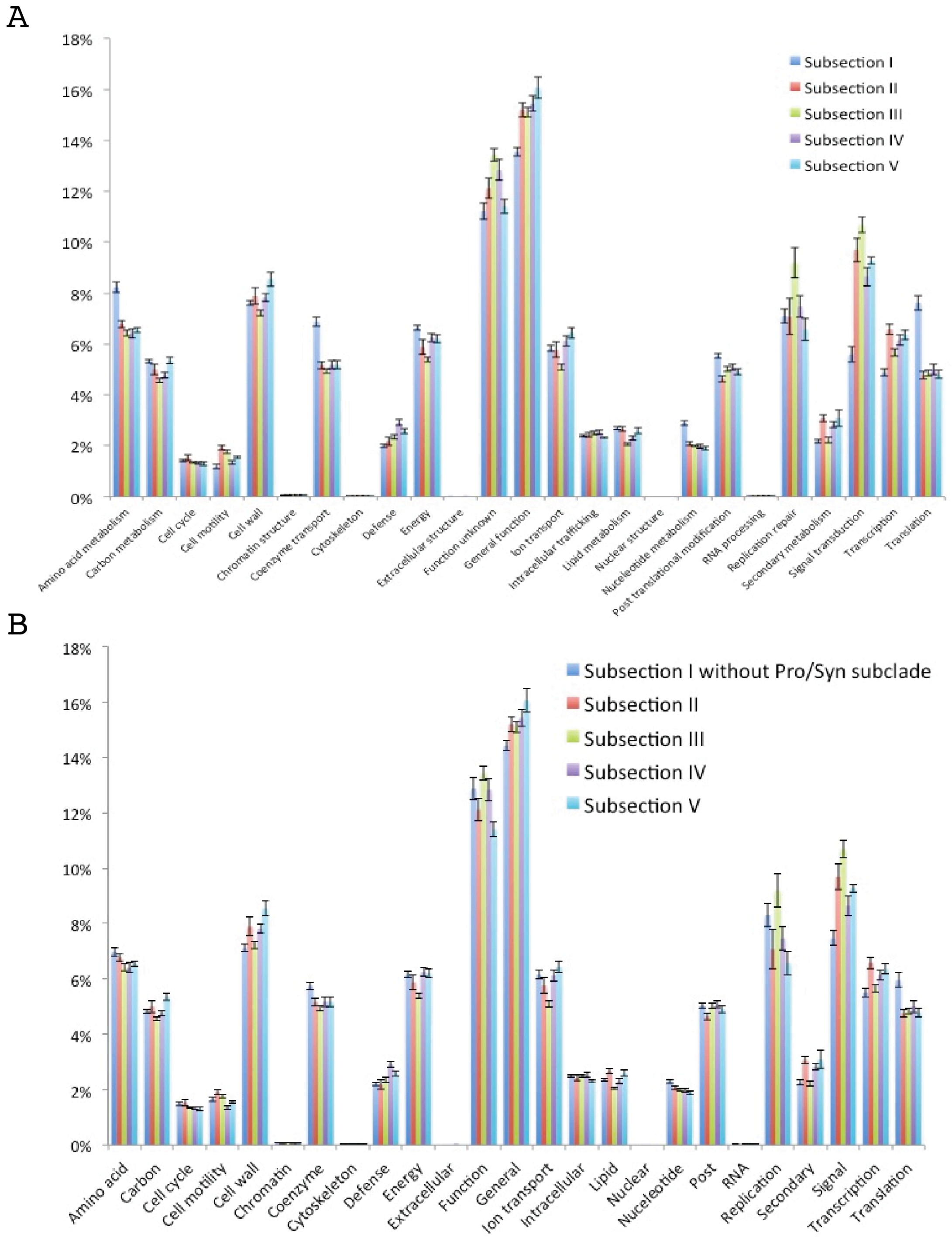


Figure S3. COG functional categories within morphological subsections. Bars represent the standard error given the sampling size of each morphological Subsection. **A**, COG analysis of all cyanobacteria included in this study. **B**, COG analysis of all cyanobacteria, excluding the *Prochlorococcus/Synechococcus* subclade in order to decrease bias within Subsection I.

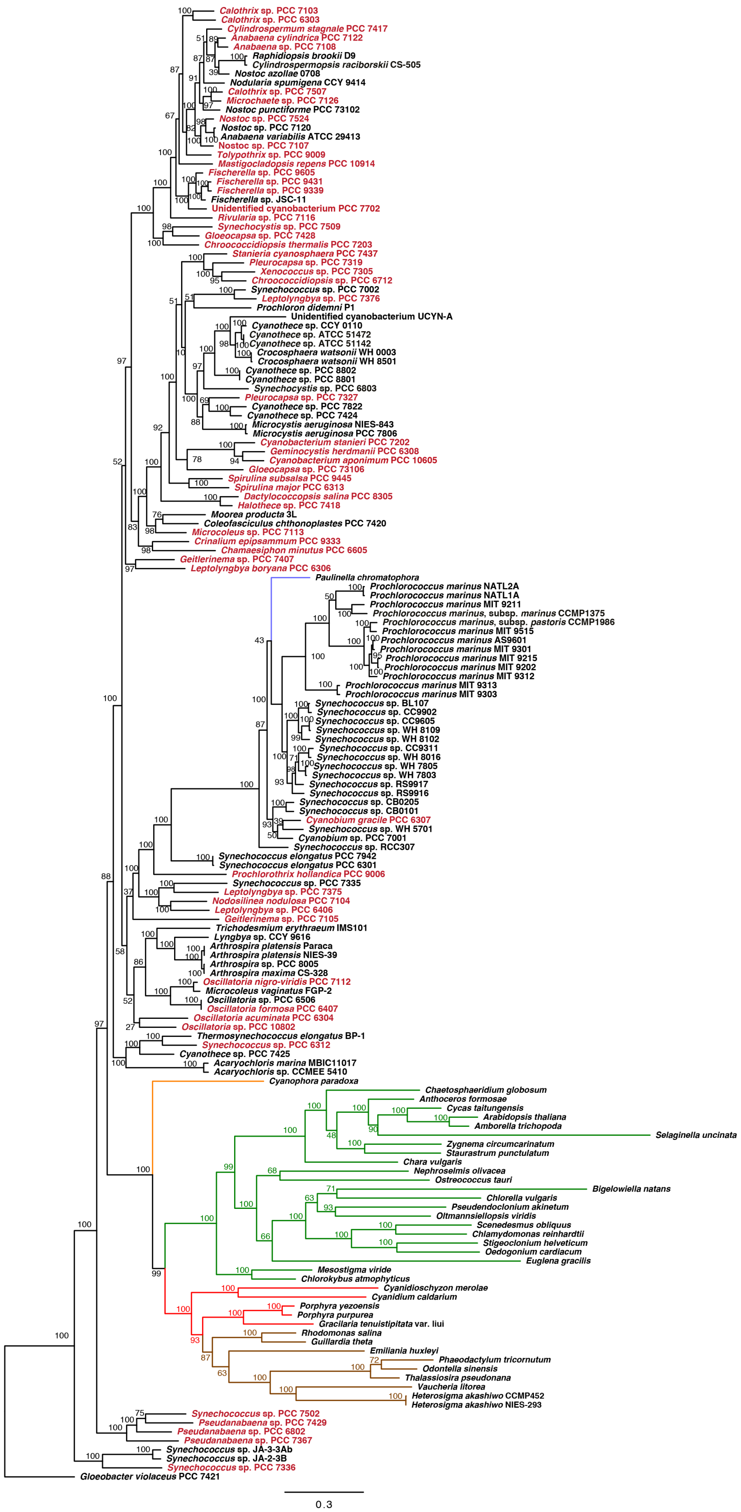

Calothrix sp. PCC 7103
Calothrix sp. PCC 6303
Cylindrospermum stagnale PCC 7417
Anabaena cylindrica PCC 7122
Anabaena sp. PCC 7108
Raphidiopsis brookii D9
Cylindrospermopsis raciborskii CS-505
Nostoc azollae 0708
Nodularia spumigena CCY 9414
Calothrix sp. PCC 7507
Microchaete sp. PCC 7126
Nostoc punctiforme PCC 73102
Nostoc sp. PCC 7524
Nostoc sp. PCC 7120
Anabaena variabilis ATCC 29413
Nostoc sp. PCC 7107
Tolypothrix sp. PCC 9009
Mastigocladopsis repens PCC 10914
Fischerella sp. PCC 9605
Fischerella sp. PCC 9431
Fischerella sp. PCC 9339
Fischerella sp. JSC-11
Unidentified cyanobacterium PCC 7702
Rivularia sp. PCC 7116
Synechocystis sp. PCC 7509
Gloeocapsa sp. PCC 7428
Chroococcidiopsis thermalis PCC 7203
Stanieria cyanosphaera PCC 7437
Pleurocapsa sp. PCC 7319
Xenococcus sp. PCC 7305
Chroococcidiopsis sp. PCC 6712
Synechococcus sp. PCC 7002
Leptolyngbya sp. PCC 7376
Prochloron didemni P1
Unidentified cyanobacterium UCYN-A
Cyanothece sp. CCY 0110
Cyanothece sp. ATCC 51472
Cyanothece sp. ATCC 51142
Crocosphaera watsonii WH 0003
Crocosphaera watsonii WH 8501
Cyanothece sp. PCC 8802
Cyanothece sp. PCC 8801
Synechocystis sp. PCC 6803
Pleurocapsa sp. PCC 7327
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Microcystis aeruginosa NIES-843
Microcystis aeruginosa PCC 7806
Cyanobacterium stanieri PCC 7202
Geminocystis herdmanii PCC 6308
Cyanobacterium aponimum PCC 10605
Gloeocapsa sp. PCC 73106
Spirulina subsalsa PCC 9445
Spirulina major PCC 6313
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
Moorea producta 3L
Coleofasciculus chthonoplastes PCC 7420
Microcoleus sp. PCC 7113
Crinalium epipsammum PCC 9333
Chamaesiphon minutus PCC 6605
Geitlerinema sp. PCC 7407
Leptolyngbya boryana PCC 6306
Paulinella chromatophora
Prochlorococcus marinus NATL2A
Prochlorococcus marinus NATL1A
Prochlorococcus marinus MIT 9211
Prochlorococcus marinus, subsp. marinus CCMP1375
Prochlorococcus marinus, subsp. pastoris CCMP1986
Prochlorococcus marinus MIT 9515
Prochlorococcus marinus AS9601
Prochlorococcus marinus MIT 9301
Prochlorococcus marinus MIT 9215
Prochlorococcus marinus MIT 9202
Prochlorococcus marinus MIT 9312
Prochlorococcus marinus MIT 9313
Prochlorococcus marinus MIT 9303
Synechococcus sp. BL107
Synechococcus sp. CC9902
Synechococcus sp. CC9605
Synechococcus sp. WH 8109
Synechococcus sp. WH 8102
Synechococcus sp. CC9311
Synechococcus sp. WH 8016
Synechococcus sp. WH 7805
Synechococcus sp. WH 7803
Synechococcus sp. RS9917
Synechococcus sp. RS9916
Synechococcus sp. CB0205
Synechococcus sp. CB0101
Cyanobium gracile PCC 6307
Synechococcus sp. WH 5701
Cyanobium sp. PCC 7001
Synechococcus sp. RCC307
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301
Prochlorothrix hollandica PCC 9006
Synechococcus sp. PCC 7335
Leptolyngbya sp. PCC 7375
Nodosilinea nodulosa PCC 7104
Leptolyngbya sp. PCC 6406
Geitlerinema sp. PCC 7105
Trichodesmium erythraeum IMS101
Lyngbya sp. CCY 9616
Arthrospira platensis Paraca
Arthrospira platensis NIES-39
Arthrospira sp. PCC 8005
Arthrospira maxima CS-328
Oscillatoria nigro-viridis PCC 7112
Microcoleus vaginatus FGP-2
Oscillatoria sp. PCC 6506
Oscillatoria formosa PCC 6407
Oscillatoria acuminata PCC 6304
Oscillatoria sp. PCC 10802
Thermosynechococcus elongatus BP-1
Synechococcus sp. PCC 6312
Cyanothece sp. PCC 7425
Acaryochloris marina MBIC11017
Acaryochloris sp. CCMEE 5410
Cyanophora paradoxa
Chaetosphaeridium globosum
Anthoceros formosae
Cycas taitungensis
Arabidopsis thaliana
Amborella trichopoda
Selaginella uncinata
Zygnema circumcarinatum
Staurastrum punctulatum
Chara vulgaris
Nephroselmis olivacea
Ostreococcus tauri
Bigelowiella natans
Chlorella vulgaris
Pseudendoclonium akinetum
Oltmannsiellopsis viridis
Scenedesmus obliquus
Chlamydomonas reinhardtii
Stigeoclonium helveticum
Oedogonium cardiacum
Euglena gracilis
Mesostigma viride
Chlorokybus atmophyticus
Cyanidioschyzon merolae
Cyanidium caldarium
Porphyra yezoensis
Porphyra purpurea
Gracilaria tenuistipitata var. liui
Rhodomonas salina
Guillardia theta
Emiliania huxleyi
Phaeodactylum tricornutum
Odontella sinensis
Thalassiosira pseudonana
Vaucheria litorea
Heterosigma akashiwo CCMP452
Heterosigma akashiwo NIES-293
Synechococcus sp. PCC 7502
Pseudanabaena sp. PCC 7429
Pseudanabaena sp. PCC 6802
Pseudanabaena sp. PCC 7367
Synechococcus sp. JA-3-3Ab
Synechococcus sp. JA-2-3B
Synechococcus sp. PCC 7336
Gloeobacter violaceus PCC 7421
0.3

Figure S4. Maximum likelihood plastidome tree with full names and bootstrap support. Cyanobacteria are named accordingly to the Strain ID in Table S1.

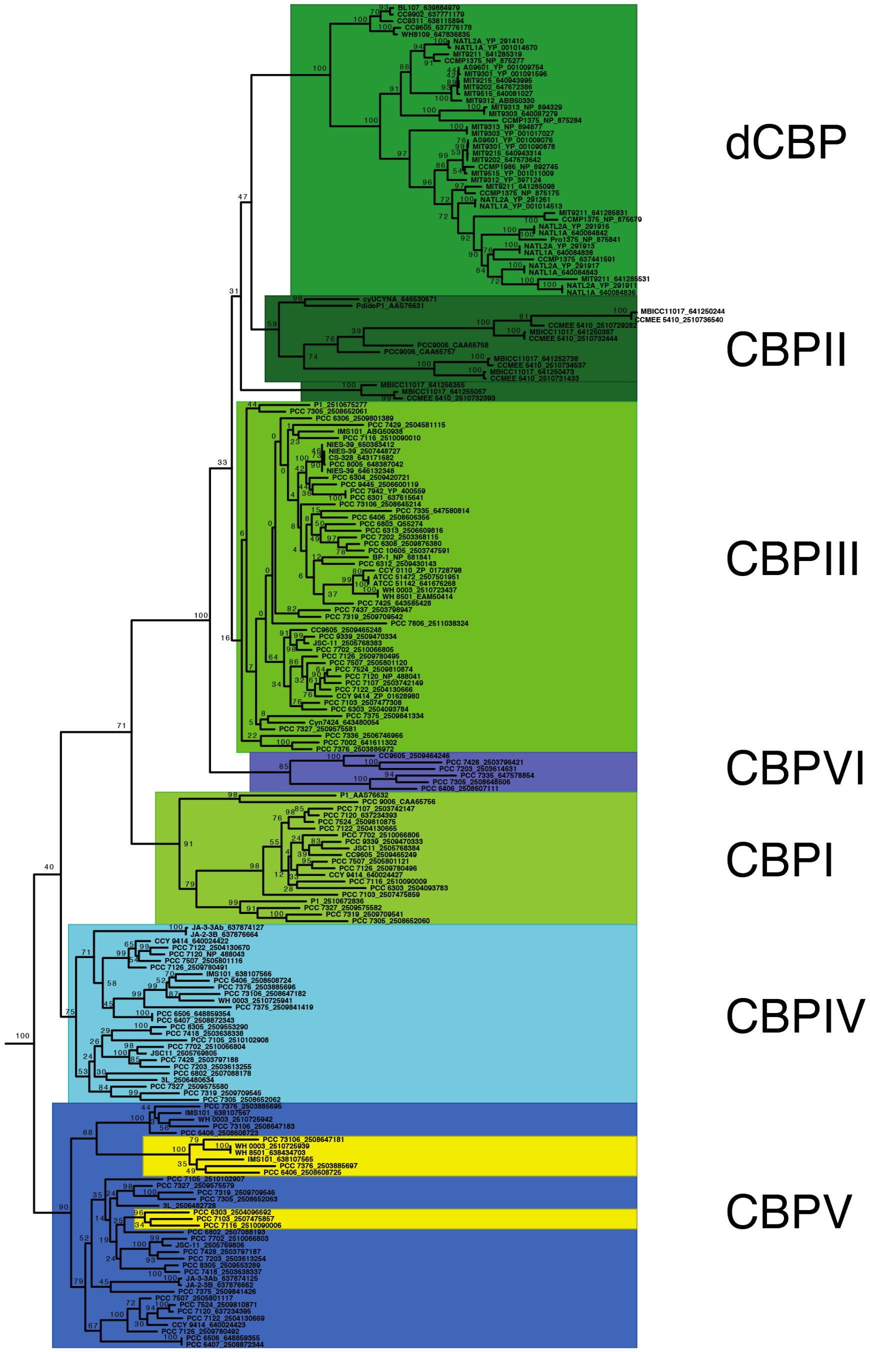


Figure S5. Maximum-likelihood CBP phylogeny reveals a diversity of previously uncharacterized clades. CP43 sequences are used as an outgroup (not shown, Newick file is available upon request), while the major CBP clades are color-coded. Shades of green represent previously characterized CBP clades (divinyl CBP = dCBP for their use of divinyl chlorophyll), whereas shades of blue represent new clades distinctly supported with the addition of CyanoGEBA genomes. Yellow subclades indicate CBPV proteins that lack the C-terminal PsaL-like domain. We find very little support for subclades CBPIII and CBPII. Taxa are named by their strain IDs abbreviation and followed by their IMG Gene Object ID or their GenBank Accessions.

```
                               Helix 1                         Helix 2
BP-1_CP43              LLGAHVAHAGLIVFWAGAMTLFEL---------VGVVHLISSAVLGFGGVYHAIRGP---
BP-1_CP47              LIAAHLMHTALVAGWAGSMALYEL---------VALAHIVLSGLLFLAACWHWVYWD---
BP-1_PsaA              IFSAHFGHLAVVFIWLSGMYFHGA---------TAIGGLVMAGLMLFAGWFHYHKRA---
BP-1_PsaB              IFASHFGHLAIIFLWVSGSLFHVA---------GAIFLLILASLALFAGWLHLQPKF---
BP-1_IsiA_CBPIII       FIAAHVAQAALSVFWAGAFTLYEI---------IGAVHLISSAVLGAGALFHTFRAP---
CCMP1375_PcbC_dCBP     FIAAHAAHAGLMMFWAGAFTLFEL---------IAVLHLIFSGVLGAGGLLHSMRYE---
MIT9313_PcbA_dCBP      FIASHIGHTGLICFGAGANTLFEL---------VAVFHLIFSAVYAGGAMLHSFRYK---
CCMP1986_PcbA_dCBP     FIAAHVAHAGLIVFWAGAFTLFEL---------IAIVHLVSSMVLAAGGLLHSLLLP---
PCC 9006_PcbC_CBPI     LLGAHIAHAGLIAFWAGSITVLEV---------IGILHLVTSAVLGAGGLFHTFKGP---
MBIC11017_PcbC_CBPI    LLGAHLCHAALMSVVPGAFIVQEV---------IGVLHFFIAAVCCAAGLFHTFRGE---
P1_PcbA_CBPII          WLAAHVAQAALIVFWAGAICLFEV---------VGVVHLVSSAVIGAGGLYHSLRGP---
PCC 7120_CBPIV         LLGAHVAHAGLIVLWAGATTLFEL---------IGVLHLISSAFLGLGGIFHALLGP---
PCC 6406_CBPIV         LLGAHVAHGGLIVFWAGAITLFEV---------IGALHLISSAFLGAGGIFHALRGP---
PCC 7375_CBPIV         LLGAHVAHAGLIVLWAGLITLFEV---------VGAVHLISSAFLGFGGIFHTLKGP---
PCC 7203_CBPIV         LLGAHVAHAGLIVFWAGAMTLFEL---------IGSIHLISSAFLGYGGIFHALRGP---
PCC 7120_CBPV          LLGAHIAHAGLIILWAGAMTLFEI---------IGVVHLVSSAVLAAGGIYHALLGP---
PCC 7203_CBPV          LLGAHIAHSALILLWAGGMTLFEL---------ISVLHLIPSVILAAGGIYHSLLGP---
PCC 7375_CBPV          LLGAHVAHAGLITLWAGAMTLFEL---------VGMFHLVASAVLGAGGLYHSFLGP---
PCC 6406_CBPV          LLGAHVAHAGLIVFWAGAMTLFEL---------IGMVHLISAAVLGAGGIYHAVLGP---
PCC 6406_CPBPV         LLGAHIAHAGLIVLWAGAMTLFEL---------IGVVHLISSAVLGAGGLYHTVLGP---
PCC 6406_CBPVI         FIVAHVAQAALIMFWAGAFTLFEL---------IGVIHIVAAGVLAGGAYFHRERLG---
PCC 7203_CBPVI         FLTAHIAHAAIVSFSIGALILLEI---------FGVVLLVSAAVFTAGTLFHRSQVP---

                                Helix 3                       Helix 4
BP-1_CP43              ---ILGFHLI-VLGIGALLLVAKAMFFG------VVGGHIWIGLICIAGGIWHIL-----
BP-1_CP47              ---MFGIHLF-LAGL--LCFGFGAFHLT------VVAHHIAAGIVGIIAGLFHIL-----
BP-1_PsaA              ---MLNHHLAGLLGLGSLAWAGHQIHVS------TAHHHLAIAVLFIIAG--HMY-----
BP-1_PsaB              ---RLNHHLAGLFGVSSLAWAGHLIHVA------MAHHHLAIAVLFIVAG--HMY-----
BP-1_IsiA_CBPIII       ---ILGHHLL-FLGFGALLLVLKATIWG------LVGGHIYIAILLIAGGIWHIL-----
CCMP1375_PcbC_dCBP     ---ILGHHLI-FLGLGNIQFVEWARIH-------VMGGHAFLAFFLIIGGAFHIA-----
MIT9313_PcbA_dCBP      ---ILGHHLL-FLGLGCVQFVEWAKYH-------VMGGHAFLAFFLSAGAIWHIF-----
CCMP1986_PcbA_dCBP     ---ILGHHLI-ILGFAVILLVEWARVH-------VMGGHAFLAFVLITGGAWHIA-----
PCC 9006_PcbC_CBPI     ---ILGHHLL-LLGILCLAFVAKAMFWG------IIGGHVYIGILELIGGTWHIL-----
MBIC11017_PcbC_CBPI    ---IVGHHLV-FISVACLIFAVNATYGT------VIGGHFLIGVIDLLGAAFHIL-----
P1_PcbA_CBPII          ---ILGHHLI-LLGLGALFLVLWAVFF-------LIGGHVYVAIIEISGGLWHIF-----
PCC 7120_CBPIV         ---ILGIHLV-LLGLGAGLLVAKAVFFG------LVGGHIWVSILCIAGGLWHIT-----
PCC 6406_CBPIV         ---ILGIHLV-LLGLGTFLLVTKAMIFG------AVGGHIWVGLMCMLGGIWHMR-----
PCC 7375_CBPIV         ---ILGSHLV-LLGGGALLLVAKAIFLG------VVGGHLYIGIVLILGGLWHIF-----
PCC 7203_CBPIV         ---ILGIHLV-LLGIGAFLLVAKAMYFG------VVGGHIWVGGILILGGLFHIA-----
PCC 7120_CBPV          ---IIGIHLL-LLGAGAWLLVAKALFWG------VVGGHIWVGILCIGGGFWHIL-----
PCC 7203_CBPV          ---ILGIHLM-LLGLGALLLVAKAMFWG------IVGGHIWVGGILIGGGIFHIL-----
PCC 7375_CBPV          ---IIGIHLV-LLGLGAWLLVAKAMFWG------IVGGHLWVGLMCVLGGIWHIA-----
PCC 6406_CBPV          ---ILGIHLM-LLGIGALLLVLKGAYFG------LVGGHFWVALLCLGGGFFHIM-----
PCC 6406_CPBPV         ---ILGIHLV-LLGVGALLLVVKATTFG------VVGGHLWIGAIAILGGIWHIR-----
PCC 6406_CBPVI         ---ILGHHLA-ILGLGALLLVVKATAFG------LVGGHIYVAVLLLLGGAWHIL-----
PCC 7203_CBPVI         ---ILGNHLI-FLGIGALLLVAKAMFFG------VVGGHIYVGALLIVAGIWHMI-----

                              Helix 5                          Helix 6
BP-1_CP43              -LSYSLGA-----LSMMGFIATCFVWFN------WLATSHFVLAFF-FLVGHLWHAG
BP-1_CP47              -LSSSIAA-----VFFAAFVVAGTMWYG------WFTFAHAVFALL-FFFGHIWHGA
BP-1_PsaA              -LTTSWHAQLAINLAMMGSLSIIVAQHM------SLFTHHMWIGGF-LVVGGAAHGA
BP-1_PsaB              -YNNSLHFQLGWHLACLGVITSLVAQHM------ALYTHHQYIAGF-LMVGAFAHGA
BP-1_IsiA_CBPIII       -LSYSLGG-----IALAGFVAAYFCAVN------WLANAHFFLAFF-FLQGHLWHAL
CCMP1375_PcbC_dCBP     -LSYSLAG-----VAYCAFVAAFWCATN------WLSNVHFYLGFF-FLQGHLWHAL
MIT9313_PcbA_dCBP      -LSTSLAG-----AAFIAFVAAFWASMN------WLSNFHFYLGFF-YLQGHFWHGL
CCMP1986_PcbA_dCBP     -LSWSLAG-----IGWMAIIAAFWSASN------WLANVHYYFGFF-FIQGHLWHAL
PCC 9006_PcbC_CBPI     -LSYSLGA-----VGWMGLLSGFFVRYC------GAAAVQYILGVL-LLVGHVWHAT
MBIC11017_PcbC_CBPI    -LSWSVAS-----VGFMGISSSLFIRYC------GAATLQLILGLVWMLGGGLWHGL
P1_PcbA_CBPII          -LAYALGG-----LAIMGFTAAVYCAFN------WLCNVHFFLAFF-VLQGHLWHAL
PCC 7120_CBPIV         -LSYSLGA-----LSLMSLIAAYFVSIN------WLANAHFWLGFF-FLQGHLWHAL
PCC 6406_CBPIV         -LSYSIGA-----VSLMAFVATLFVSVN------WLANAHFWLGFF-FLQGHLFHAL
PCC 7375_CBPIV         -LAYSLGA-----LSLMTFVATLFVSVN------WLANAHFWLGFF-FLQGHLWHTL
PCC 7203_CBPIV         -LSYSLGA-----LALMGFIATLFVSVN------WLANTHFWLAFF-FLQGHIWHAL
PCC 7120_CBPV          -LSYSLAA-----LAYMGLLAAYFVTVN------WLATSHFALAIV-FLSGHIWHAL
PCC 7203_CBPV          -LAYSIGA-----VAYMGFFAAYFASVN------WLVSFHFVLAVI-FLLGHIWHAL
PCC 7375_CBPV          -LSYSLGA-----LAYMGIFAGYFVTVN------WLAAFHFAFGGL-LLAGHLWHAI
PCC 6406_CBPV          -LSYSLGA-----LAIAGLSVAVFVSVN------ALASVHAGLGFL-ALLGHLWHAC
PCC 6406_CPBPV         -LAYSQAA-----LAYMGFFAAYFVWVN------WLMLFHVVFASL-LLAGHFWHGL
PCC 6406_CBPVI         -LSYSLFG-----IALAGFAASYYCGFN------WLANAHFYLAFF-FLQGGLWHFQ
PCC 7203_CBPVI         -LSYSLFS-----LALTGFAGSYFCGFN------WLANTYFYLSFF-TLQGSLWHFG
```

Figure S6. Newly characterized CBP clades have conserved residues for potentially binding chlorophyll. Alignment of the transmembrane helices of CBP proteins and similar light-harvesting proteins. Amino acids highlighted in green (histidine) and yellow (glutamine) correspond to putative chlorophyll-binding residues. Organisms are named accordingly to the Strain ID of Table S1.

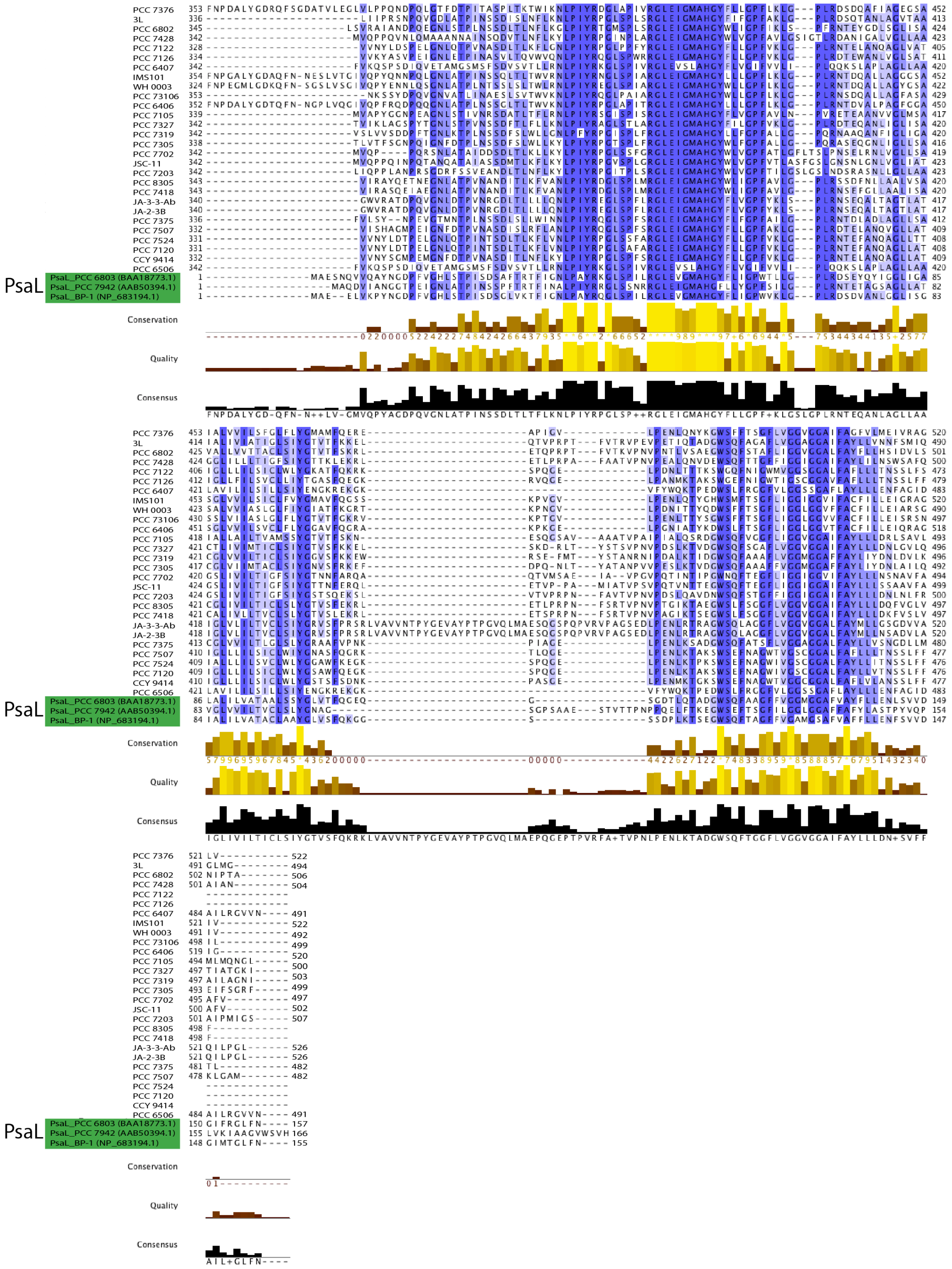


Figure S7. The C-terminal PsaL-like domain of CBPV proteins is homologous to PsaL. Alignment of the C-terminal PsaL-like domain of CBPV proteins containing full-length PsaL domains to the canonical PsaL of PSI (highlighted green accessions mark the amino acid sequences of the PsaL subunits of *Synechoccoccus elongatus* PCC 7942, *Synechocystis* sp. PCC 6803, and *Thermosynechococcus elongatus* BP-1. Organisms are named accordingly to the Strain ID of Table S1.

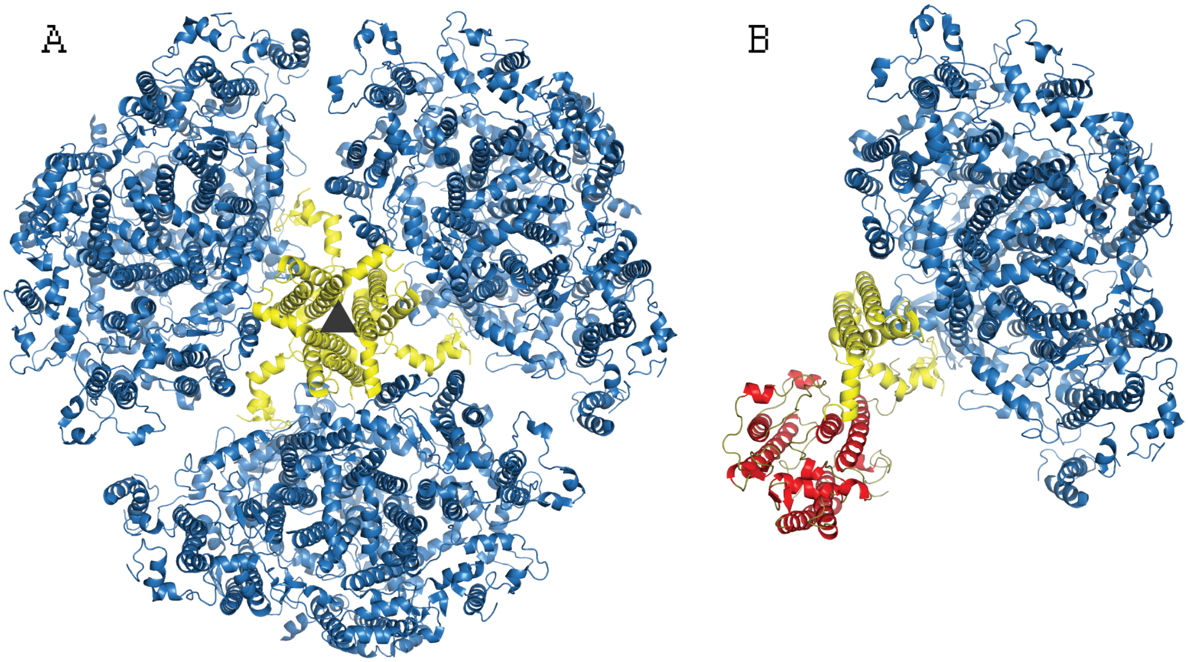


Figure S8. Comparison of trimeric Photosystem I to proposed CBPV-Photosystem I complex model. A, Top view of trimeric Photosystem I structure of *Thermosynechococcus elongatus* from the Protein Data Bank structure, 1JB0 (PDB ID). The threefold symmetry axis is denoted by the black triangle in the center. PsaL subunits are highlighted in yellow. **B,** Top view of proposed model of CBPV from *Chroococcidiopsis thermalis* PCC 7203 (Chro_2988) interacting with the Photosystem I monomer from the upper right of the trimer. Replacing the PsaL subunit (yellow) of a monomeric PSI with the PsaL-like domain of CBPV would preclude trimer formation, potentially resulting in monomerization of Photosystem I. The CBP domain (first six helices) is highlighted in red, whereas the monomeric Photosystem I, excluding the PsaL subunit, is highlighted in yellow.

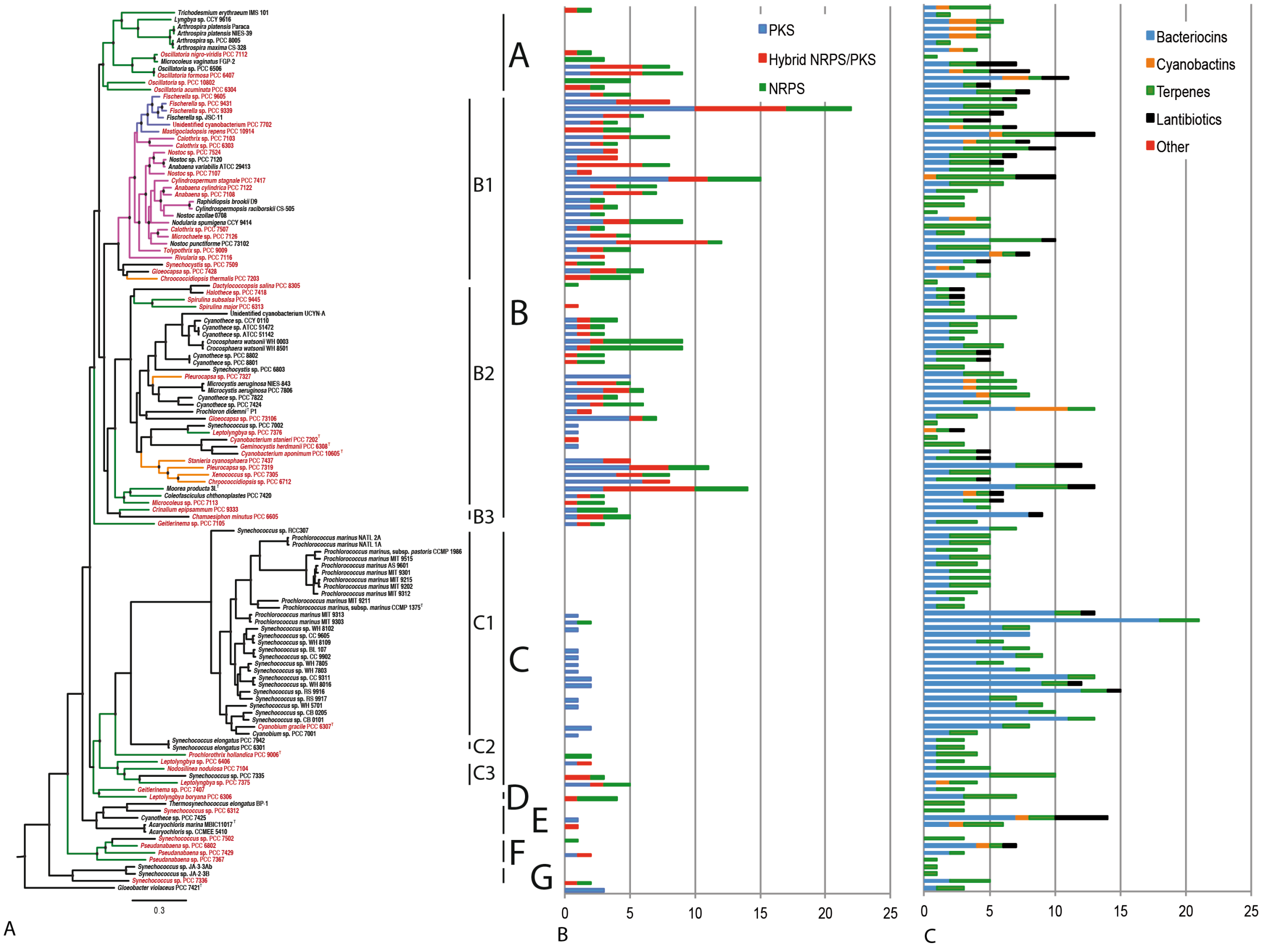


Figure S9. | Distribution of the ribosome dependent and non-ribosomal encoded peptide and polyketide biosynthetic pathways in Cyanobacteria. A, Cyanobacterial Tree as in Fig. 1, **B**, Distribution of the non-ribosomal peptide and polyketide gene clusters (number and occurrence within each genome), **C**, Distribution of the gene clusters involved in ribosome-dependent synthesis of diverse peptides (number and occurrence within each genome).

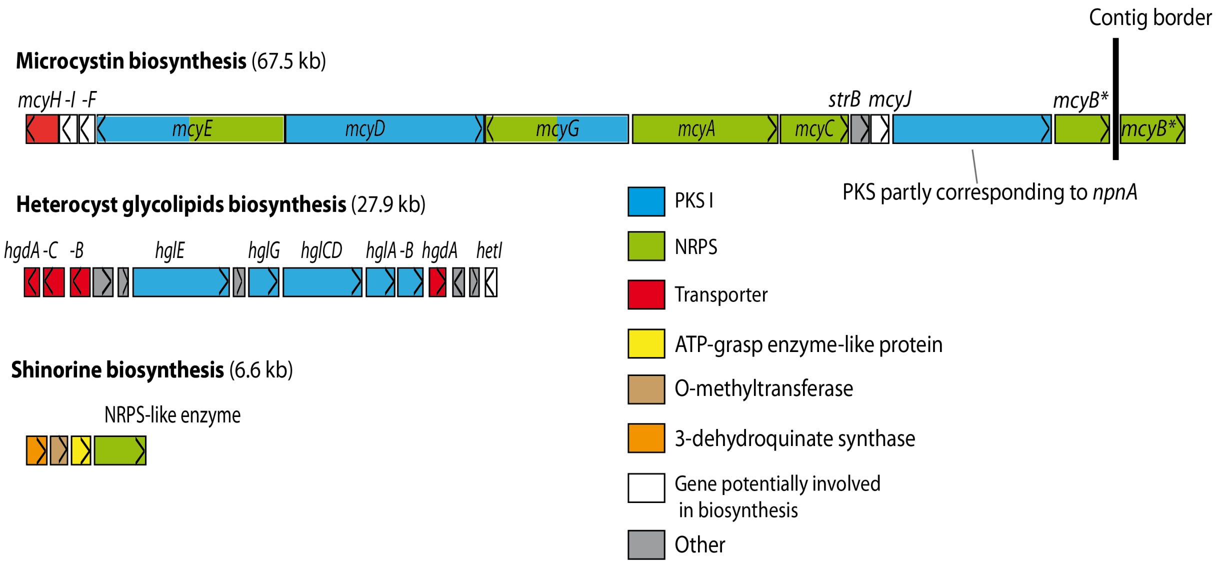

Contig border
Microcystin biosynthesis (67.5 kb)
mcyH -I -F
mcyE
mcyD
mcyG
mcyA
mcyC
strB mcyJ
mcyB*
mcyB*
PKS partly corresponding to npnA
Heterocyst glycolipids biosynthesis (27.9 kb)
hgdA -C -B
hglE
hglG
hglCD
hglA -B
hgdA
hetI
Shinorine biosynthesis (6.6 kb)
NRPS-like enzyme
PKS I
NRPS
Transporter
ATP-grasp enzyme-like protein
O-methyltransferase
3-dehydroquinate synthase
Gene potentially involved in biosynthesis
Other

Figure S10. Predicted genetic potential for production of already kwon secondary metabolites found in the genome of *Fischerella* sp. PCC 9339. The identities of the sequence are estimated at the amino-acid level (% AASI). The putative microcystin gene cluster has 79.8% AASI to the one of *Anabaena* sp. 90 (25) and 88.5% AASI to the partial one retrieved from *Hapalosiphon hibernicus* BZ-3-1(26). Note the additional PKS gene, which on 2/3 of its length with 77.5% AASI corresponds to *NpnA* gene of the nostophycin gene cluster in *Nostoc* sp. 152 (27). The putative heterocyst glycolipids gene cluster has 67% AASI to the gene cluster required for synthesis and deposition of envelope glycolipids in *Nostoc* sp. PCC 7120 (28). Note the presence of two *hgdA* and the combination of *hglC* and *hglD* into a single gene in the heterocyst producing *Fischerella* sp. PCC 9339. The putative shinorine gene cluster is 70% AASI to the one identified in *Anabaena variabilis* ATCC 29413 (29).

Tables S1-S11

Table S1. 126 Cyanobacteria included in this study

Details on the strains are available in the Table S2.
^{T} indicates Type strain or Type species, for genome status: F, finished, D, draft, P, permanent draft.

Strain	Strain ID	Genome size (Mb)	% mol GC	No of scaffolds (chromosome / plasmid) - status	NCBI Project ID	References
Subsection I						
Acaryochloris sp.	CCMEE 5410	7.88	47.1	511 - D	16707	(30)
Acaryochloris marina	MBIC11017^T	8.36	47	10 (1/9) - F	12997	(31)
Chamaesiphon minutus	PCC 6605	6.76	45.7	3 - P	158825	This study
Crocosphaera watsonii	WH 0003	5.89	37.7	1126 - D	61839	(32)
Crocosphaera watsonii	WH 8501	6.24	37.1	323 - D	10651	(33)
Cyanobacterium aponinum	PCC 10605^T	4.18	34.9	2 - F	158691	This study
Cyanobacterium stanieri	PCC 7202 ^{T}	3.16	38.7	1 - F	39697	This study
Cyanobium gracile	PCC 6307 ^{T}	3.34	68.7	1 - F	158695	This study
Cyanobium sp.	PCC 7001	2.83	68.7	2 - D	19301	
Cyanothece sp.	ATCC 51142	5.46	37.9	6 (2/4) - F	20319	(34)
Cyanothece sp.	ATCC 51472	5.46	37.9	7 - F	59973	(35)
Cyanothece sp.	CCY 0110	5.88	36.7	163 - D	18951	
Cyanothece sp.	PCC 7424	6.55	38.5	7 (1/6) - F	20479	(35)
Cyanothece sp.	PCC 7425	5.79	50.7	4 (1/3) - F	28337	(35)
Cyanothece sp.	PCC 7822	7.84	39.9	7 (1/6) - F	28535	(35)
Cyanothece sp.	PCC 8801	4.79	39.8	4 (1/3) - F	20503	(35)
Cyanothece sp.	PCC 8802	4.8	39.8	5 (1/4) - F	28339	(35)
Dactylococcopsis salina	PCC 8305	3.78	42.4	1 - F	158703	This study
Geminocystis herdmanii	PCC 6308 ^{T}	4.26	34.3	1 - P	62511	This study
Gloeobacter violaceus	PCC 7421 ^{T}	4.66	62	1 - F	9606	(36)

Gloeocapsa sp.	PCC 73106	4.03	41.1	228 - D	159497	This study
Gloeocapsa sp.	PCC 7428	5.88	43.4	5 - F	158831	This study
Halothece sp.	PCC 7418	4.18	42.9	1 - F	40817	This study
Microcystis aeruginosa	NIES-843	5.84	42.3	1 - F	27835	(37)
Microcystis aeruginosa	PCC 7806	5.2	42	118 - D	15702	(38)
Prochlorococcus marinus	AS9601	1.67	31.3	1 - F	13548	(39)
Prochlorococcus marinus	MIT9202	1.69	31.1	1 - D	19343	
Prochlorococcus marinus	MIT9211	1.69	38	1 - F	13551	(39)
Prochlorococcus marinus	MIT9215	1.74	31.2	1 - F	18633	(39)
Prochlorococcus marinus	MIT9301	1.64	31.3	1 - F	15746	(39)
Prochlorococcus marinus	MIT9303	2.68	50	1 - F	13496	(39)
Prochlorococcus marinus	MIT9312	1.71	31.2	1 - F	13910	(40)
Prochlorococcus marinus	MIT9313	2.41	50.7	1 - F	220	(41)
Prochlorococcus marinus	MIT9515	1.7	31	1 - F	13617	(39)
Prochlorococcus marinus	NATL1A	1.86	35	1 - F	15660	(39)
Prochlorococcus marinus	NATL2A	1.84	35.1	1 - F	13911	(39)
Prochlorococcus marinus, subsp. *marinus*	CCMP1375 ^{T}	1.75	36.4	1 - F	419	(42)
Prochlorococcus marinus, subsp. *pastoris*	CCMP1986	1.66	30.8	1 - F	213	(41)
Prochloron didemni (metagenome)	P1	6.2	42	100 - D	13452	(43)
*Synechococcus elongatu*s	PCC 6301	2.7	55.5	1 - F	13282	(44)
Synechococcus elongatus	PCC 7942	2.74	55.4	2 (1/1) - F	10645	
Synechococcus sp.	BL107	2.28	54.2	6 - D	13559	(45)
Synechococcus sp.	CB0101	2.69	64.2	94 - D	46501	
Synechococcus sp.	CB0205	2.43	63	78 - D	46503	
Synechococcus sp.	CC9311	2.61	52.5	1 - F	12530	(46)
Synechococcus sp.	CC9605	2.51	59.2	1 - F	13643	(45)
Synechococcus sp.	CC9902	2.23	54.2	1 - F	13655	(45)
Synechococcus sp.	JA-2-3B	3.05	58.5	1 - F	16252	(47)

Synechococcus sp.	JA-3-3Ab	2.93	60.2	1 - F	16251	(47)
Synechococcus sp.	PCC 6312	3.72	48.5	2 - F	158717	This study
Synechococcus sp.	PCC 7002	3.41	49.2	7 (1/6) - F	28247	
Synechococcus sp.	PCC 7335	5.97	48.2	11 - F	19377	
Synechococcus sp.	PCC 7336	5.14	53.7	2 - F	158719	This study
Synechococcus sp.	PCC 7502	3.58	40.6	3 - F	159509	This study
Synechococcus sp.	RCC307	2.22	60.8	1 - F	13654	(45)
Synechococcus sp.	RS9916	2.66	59.8	4 - D	13557	(45)
Synechococcus sp.	RS9917	2.58	64.5	9 - D	13555	(45)
Synechococcus sp.	WH 5701	3.04	65.4	135 - D	13554	(45)
Synechococcus sp.	WH 7803	2.37	60.2	1 - F	13642	(45)
Synechococcus sp.	WH 7805	2.62	57.6	13 - F	13553	(45)
Synechococcus sp.	WH 8016	2.71	54.1	1 - F	61805	
Synechococcus sp.	WH 8102	2.43	59.4	1 - F	230	(48)
Synechococcus sp.	WH 8109	2.12	60.1	1 - F	37911	
Synechocystis sp.	PCC 6803	3.95	47.4	5 (1/4) - F	60	(49)
Synechocystis sp.	PCC 7509	4.77	41.6	174 - D	159501	This study
Thermosynechococcus elongatus	BP-1	2.59	53.9	1 - F	308	(50)
Unidentified cyanobacterium (symbiont)	UCYN-A	1.44	31.1	1 - F	30917	(51)
Subsection II						
Chroococcidiopsis sp.	PCC 6712	5.7	35.3	3 - F	158687	This study
Chroococcidiopsis thermalis	PCC 7203	6.69	44.5	3 - F	38119	This study
Pleurocapsa sp.	PCC 7319	7.39	38.7	10 - P	158813	This study
Pleurocapsa sp.	PCC 7327	4.99	45.2	1 - F	158829	This study
Stanieria cyanosphaera	PCC 7437	5.55	36.2	6 - F	158877	This study
Xenococcus sp.	PCC 7305	5.93	39.7	234 - D	159499	This study
Subsection III						
Arthrospira maxima	CS-328	6	44.8	129 - D	29085	

Arthrospira platensis	NIES-39	6.79	44.3	1 - F	42161	(52)
Arthrospira platensis	Paraca	5,00	44.3	1820 - D	34793	
Arthrospira sp.	PCC 8005	6.15	44.7	119 - D	40633	(53)
Coleofasciculus chthonoplastes	PCC 7420	8.68	45.4	57 - D	19325	
Crinalium epipsammum	PCC 9333	5.62	40.2	9 - F	158835	This study
Geitlerinema sp.	PCC 7105	6.15	51.6	8 - P	158727	This study
Geitlerinema sp.	PCC 7407	4.68	58.5	1 - F	158833	This study
Leptolyngbya boryana	PCC 6306	7.26	47	5 - P	158729	This study
Leptolyngbya sp.	PCC 6406	5.61	55.2	377 - P	159511	This study
Leptolyngbya sp.	PCC 7375	9.42	47.6	5 - P	43137	This study
Leptolyngbya sp.	PCC 7376	5.13	43.9	1 - F	43487	This study
Lyngbya sp.	CCY 9616	7.04	41.1	110 - D	13409	
Microcoleus sp.	PCC 7113	7.97	46.2	9 - F	158839	This study
Microcoleus vaginatus	FGP-2	6.7	46	40 - P	47601	(54)
Moorea producta	3L^{T}	8.48	43.7	161 - D	60895	(55)
Nodosilinea nodulosa	PCC 7104	6.89	57.7	2 - P	62311	This study
Oscillatoria acuminata	PCC 6304	7.8	47.6	3 - F	158709	This study
Oscillatoria formosa	PCC 6407	6.89	43.4	12 - P	158733	This study
Oscillatoria nigro-viridis	PCC 7112	8.27	45.8	6 - F	158711	This study
Oscillatoria sp.	PCC 10802	8.59	54.1	9 - P	158815	This study
Oscillatoria sp.	PCC 6506	6.68	43.4	377 - D	49445	(56)
Prochlorothrix hollandica	PCC 9006^{T}	5.65	54.4	13 - P	158811	This study
Pseudanabaena sp.	PCC 6802	5.62	47.8	6 - P	158731	This study
Pseudanabaena sp.	PCC 7367	4.89	46.2	2 - F	158713	This study
Pseudanabaena sp.	PCC 7429	5.48	43.2	464 - D	158837	This study
Spirulina major	PCC 6313	5.05	53.5	2 - F	158715	This study
Spirulina subsalsa	PCC 9445	5.32	47.4	2 - F	158827	This study
Trichodesmium erythraeum	IMS101	7.75	34.1	1 - F	318	

Subsection IV						
Anabaena cylindrica	PCC 7122	7.06	38.8	7 - F	43355	This study
Anabaena sp.	PCC 7108	5.89	38.8	3 - F	158737	This study
Anabaena variabilis	ATCC 29413	7.11	41.4	5 (2/3) - F	10642	
Calothrix sp.	PCC 6303	6.96	39.8	4 - F	158041	This study
Calothrix sp.	PCC 7103	11.58	38.6	12 - P	159495	This study
Calothrix sp.	PCC 7507	7.02	42.3	1 - F	158683	This study
Cylindrospermopsis raciborskii	CS-505	3.88	40.2	93 - D	40109	(57)
Cylindrospermum stagnale	PCC 7417	7.61	42.2	4 - P	158809	This study
Microchaete sp.	PCC 7126	5.74	42.2	3 - P	158817	This study
Nodularia spumigena	CCY 9414	5.32	41.3	204 - D	13447	
Nostoc azollae (endosymbiont)	708	5.49	38.4	3 (1/2) - F	30807	(58)
Nostoc punctiforme	PCC 73102	9.06	41.4	6 (1/5) - F	216	
Nostoc sp.	PCC 7107	6.33	40.4	1 - F	158705	This study
Nostoc sp.	PCC 7120	7.21	41.3	7 (1/6) - F	244	(59)
Nostoc sp.	PCC 7524	6.72	41.5	3 - F	158707	This study
Raphidiopsis brookii	D9	3.19	40.1	47 - D	40111	(57)
Rivularia sp.	PCC 7116	8.73	37.5	3 - F	63147	This study
Tolypothrix sp.	PCC 9009	8.18	41.2	204 - D	Submit	This study
Subsection V						
Fischerella sp.	JSC-11	5.38	41.1	34 - D	61093	
Fischerella sp.	PCC 9339	8.4	40.1	95 - P	159505	This study
Fischerella sp.	PCC 9431	7.14	40.2	36 - P	158821	This study
Fischerella sp.	PCC 9605	8.2	42.6	36 - P	158819	This study
Mastigocladopsis repens	PCC 10914	6.31	43.5	23 - P	158735	This study
Unidentified cyanobacterium*	PCC 7702	4.9	42.4	4 - P	158823	This study

*PCC 7702 corresponds to the high temperature forms (HTF) of cyanobacteria found in hot springs, at temperatures higher than 50 °C (up to 62°C), and originally thought to be related to "*Mastigocladus laminosus*". The morphology of this HTF strain is variable from unicellular to very short filaments, and consequently, impossible to identify at the genus level. Furthermore, PCC 7702 strain is unable to fix nitrogen under aerobic conditions but contains *nif* genes.

Table S2. Improvement of phylogenetic diversity with the addition of the CyanoGEBA dataset measured by Tree Imbalance

Phylogenetic Diversity Metric

CyanoGEBA set	Random set	Fold Improvement
10.82	5.28±0.37	1.92-2.20

Tree Imbalance

Average Colless's Imbalance (n=1000)	Genomes prior to this study	All Genomes, including CyanoGEBA
Uniform Speciation	0.093	0.059
Equiprobable Speciation	0.30	0.24

Table S3. Novel* proteins in CyanoGEBA genomes
*lacking similarity to any protein in Genbank

CyanoGEBA genome	Number of novel proteins coding genes	% of novel protein coding gene
Anabaena cylindrica PCC 7122	338	5.40
Anabaena sp. PCC 7108	291	5.57
Calothrix sp. PCC 6303	370	6.33
Calothrix sp. PCC 7103	1153	11.16
Calothrix sp. PCC 7507	375	6.00
Chamaesiphon minutus PCC 6605	704	10.94
Chroococcidiopsis sp. PCC 6712	334	6.45
Chroococcidiopsis thermalis PCC 7203	339	5.62
Crinalium epipsammum PCC 9333	372	7.35
Cyanobacterium aponinum PCC 10605	138	3.82
Cyanobacterium stanieri PCC 7202	97	3.30
Cyanobium gracile PCC 6307	212	6.16
Cylindrospermum stagnale PCC 7417	486	7.21
Dactylococcopsis salina PCC 8305	199	5.40
Fischerella sp. PCC 9339	505	7.40
Fischerella sp. PCC 9431	360	5.90
Fischerella sp. PCC 9605	626	8.78
Geitlerinema sp. PCC 7105	412	7.63
Geitlerinema sp. PCC 7407	162	4.14
Geminocystis herdmanii PCC 6308	168	4.00
Gloeocapsa sp. PCC 73106	171	4.12
Gloeocapsa sp. PCC 7428	251	4.73
Halothece sp. PCC 7418	133	3.39
Leptolyngbya boryana PCC 6306	736	10.65
Leptolyngbya sp. PCC 6406	468	8.92
Leptolyngbya sp. PCC 7375	1137	13.46
Leptolyngbya sp. PCC 7376	342	7.35
Mastigocladopsis repens PCC 10914	409	7.17
Microchaete sp. PCC 7126	336	6.37
Microcoleus sp. PCC 7113	458	6.71
Nodosilinea nodulosa PCC 7104	480	7.42
Nostoc sp. PCC 7107	220	3.97
Nostoc sp. PCC 7524	253	4.45

Oscillatoria acuminata PCC 6304	419	6.87
Oscillatoria formosa PCC 6407	110	8.76
Oscillatoria nigro-viridis PCC 7112	508	13.37
Oscillatoria sp. PCC 10802	937	1.55
Pleurocapsa sp. PCC 7319	452	6.70
Pleurocapsa sp. PCC 7327	221	4.73
Prochlorothrix hollandica PCC 9006	492	10.20
Pseudanabaena sp. PCC 6802	525	9.64
Pseudanabaena sp. PCC 7367	357	8.89
Pseudanabaena sp. PCC 7429	406	8.42
Rivularia sp. PCC 7116	437	6.29
Spirulina major PCC 6313	247	5.54
Spirulina subsalsa PCC 9445	216	4.67
Stanieria cyanosphaera PCC 7437	255	5.06
Synechococcus sp. PCC 6312	313	8.25
Synechococcus sp. PCC 7336	472	9.90
Synechococcus sp. PCC 7502	256	6.98
Synechocystis sp. PCC 7509	247	5.19
Tolypothrix sp. PCC 9009	636	8.53
Xenococcus sp. PCC 7305	396	7.30
Unidentified cyanobacterium PCC 7702	170	3.89

Table S4. Prediction of CRISPR loci in CyanoGEBA genomes

CyanoGEBA genome	Number of spacer-direct repeat units	Number of CRISPR loci
Anabaena cylindrica PCC 7122	367	13
Anabaena sp. PCC 7108	95	7
Calothrix sp. PCC 6303	72	6
Calothrix sp. PCC 7103	178	13
Calothrix sp. PCC 7507	336	10
Chamaesiphon minutus PCC 6605	59	3
Chroococcidiopsis sp. PCC 6712	47	5
Chroococcidiopsis thermalis PCC 7203	64	2
Crinalium epipsammum PCC 9333	113	6
Cyanobacterium aponinum PCC 10605	166	10
Cyanobacterium stanieri PCC 7202	15	2
Cyanobium gracile PCC 6307	0	0
Cylindrospermum stagnale PCC 7417	191	10
Dactylococcopsis salina PCC 8305	0	0
Fischerella sp. PCC 9339	26	7
Fischerella sp. PCC 9431	18	4
Fischerella sp. PCC 9605	11	2
Geitlerinema sp. PCC 7105	650	15
Geitlerinema sp. PCC 7407	23	1
Geminocystis herdmanii PCC 6308	33	2
Gloeocapsa sp. PCC 73106 *	50	4
Gloeocapsa sp. PCC 7428	98	3
Halothece sp. PCC 7418	443	4
Leptolyngbya boryana PCC 6306	80	5
Leptolyngbya sp. PCC 6406 *	168	9
Leptolyngbya sp. PCC 7375	188	12
Leptolyngbya sp. PCC 7376	6	1
Mastigocladopsis repens PCC 10914	0	0
Microchaete sp. PCC 7126	88	4
Microcoleus sp. PCC 7113	72	10
Nodosilinea nodulosa PCC 7104	75	4
Nostoc sp. PCC 7107	252	14
Nostoc sp. PCC 7524	278	6
Oscillatoria acuminata PCC 6304	279	10

Oscillatoria formosa PCC 6407	95	10
Oscillatoria nigro-viridis PCC 7112	304	9
Oscillatoria sp. PCC 10802	531	18
Pleurocapsa sp. PCC 7319	68	1
Pleurocapsa sp. PCC 7327	100	4
Prochlorothrix hollandica PCC 9006	237	8
Pseudanabaena sp. PCC 6802	77	2
Pseudanabaena sp. PCC 7367	160	7
Pseudanabaena sp. PCC 7429 *	610	14
Rivularia sp. PCC 7116	256	15
Spirulina major PCC 6313	102	7
Spirulina subsalsa PCC 9445	625	17
Stanieria cyanosphaera PCC 7437	74	4
Synechococcus sp. PCC 6312	154	4
Synechococcus sp. PCC 7336	285	8
Synechococcus sp. PCC 7502	62	2
Synechocystis sp. PCC 7509 *	6	1
Tolypothrix sp. PCC 9009 *	201	15
Xenococcus sp. PCC 7305 *	37	5
Unidentified cyanobacterium PCC 7702	8	2

* These genomes are not finished and currently contain more than 100 scaffolds. The number of spacer-direct repeat units and CRISPR loci therefore may be underestimated.

Table S5. Comparative genomics of morphological transitions
Events of morphological transition are shown in Fig. 1. For each event, the set of genes involved in one genome or in genomes belonging to one subsection (genome in) were compared those of genomes of another subsection (genome out). Genomes are annotated by the Strain ID as in Table S1.

Morphological transition (Genomes in *vs* out)	**Evolutionary transition (Subsection to Subsection)**	**Number of genes**
Event 1 (PCC 7367, PCC 7429, PCC 6802 *vs* PCC 7502)	III to I	88
Event 2 (PCC 6406, PCC 7104, PCC 7375 *vs* PCC 7335)	III to I	674
Event 3 (PCC 9006, PCC 6406, PCC 7104, PCC 7375, PCC 6306, PCC 7407 *vs* subclade C1 and C2)	III to I	32
Event 4 (PCC 7002, PCC 7202, PCC 6308, and PCC 10605 *vs* PCC 7376)	I to III	3172
Event 5 (NIES-843, PCC 7806, PCC 7822, and PCC 7424 *vs* PCC 7327)	I to II	2531
Event 6 (PCC 7428, PCC 7509 *vs* PCC 7203)	I to II	3783
Event 7 (PCC 7203, PCC 7428, PCC 7509 *vs* Subsection IV and V)	I to IV and V	9
Event 8 (Subsection V *vs* Subsection IV)	IV to V	0

Table S6. **Homologous proteins lost during the reversion of filamentous to unicellular morphology in both Event 2 and Event 3.**

Query locus tag in Event 2	Top hit locus tag in Event 3	Query annotation
Pro9006DRAFT_1077	Lepto6406DRAFT_00007290	Arsenite-activated ATPase ArsA
Pro9006DRAFT_3818	Lepto6406DRAFT_00024510	HAS barrel domain.
Pro9006DRAFT_3344	Lepto6406DRAFT_00009900	Hypothetical protein
Pro9006DRAFT_0305	Lepto6406DRAFT_00035530	Hypothetical protein
Pro9006DRAFT_0620	Lepto6406DRAFT_00010140	Hypothetical protein
Pro9006DRAFT_4432	Lepto6406DRAFT_00049190	Highly conserved protein containing a thioredoxin domain
Pro9006DRAFT_1144	Lepto6406DRAFT_00002010	Asparaginase
Pro9006DRAFT_2144	Lepto6406DRAFT_00041660	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain
Pro9006DRAFT_3622	Lepto6406DRAFT_00019670	Iron-sulfur cluster binding protein, putative
Pro9006DRAFT_0863	Lepto6406DRAFT_00033060	Hypothetical protein
Pro9006DRAFT_2707	Lepto6406DRAFT_00005310	Hypothetical protein
Pro9006DRAFT_1113	Lepto6406DRAFT_00025920	Hypothetical protein
Pro9006DRAFT_0892	Lepto6406DRAFT_00019960	Hypothetical protein
Pro9006DRAFT_0326	Lepto6406DRAFT_00025500	Hypothetical protein
Pro9006DRAFT_4045	Lepto6406DRAFT_00040530	Alpha-amylase/alpha-mannosidase
Pro9006DRAFT_1882	Lepto6406DRAFT_00043930	Hypothetical protein
Pro9006DRAFT_4594	Lepto6406DRAFT_00035370	Hypothetical protein
Pro9006DRAFT_1996	Lepto6406DRAFT_00016810	Hypothetical protein
Pro9006DRAFT_1710	Lepto6406DRAFT_00003520	Hypothetical protein
Pro9006DRAFT_25	Lepto6406DRAFT_000313	Polyketide cyclase / dehydrase and

50	50	lipid transport.
Pro9006DRAFT_2845	Lepto6406DRAFT_00014640	Hypothetical protein
Pro9006DRAFT_0040	Lepto6406DRAFT_00005410	Hypothetical protein
Pro9006DRAFT_1334	Lepto6406DRAFT_00025630	Hypothetical protein
Pro9006DRAFT_1711	Lepto6406DRAFT_00003510	Hypothetical protein
Pro9006DRAFT_1895	Lepto6406DRAFT_00032390	Hypothetical protein
Pro9006DRAFT_2407	Lepto6406DRAFT_00028690	FOG: GAF domain
Pro9006DRAFT_4751	Lepto6406DRAFT_00041610	Hypothetical protein
Pro9006DRAFT_1359	Lepto6406DRAFT_00013970	Uncharacterized conserved protein
Pro9006DRAFT_1554	Lepto6406DRAFT_00014960	Uncharacterized protein conserved in bacteria

Table S7. Increase in number of cyanobacterial proteins improves prediction of eukaryotic nuclear genes that resulted from Endosymbiotic Gene Transfer.

Eukaryote	**Number of genes predicted without CyanoGEBA genomes**	**Number of genes predicted including CyanoGEBA genomes**	**% increase with CyanoGEBA**
Arabidopsis (plant)	3811	4339	14%
Physcomitrella (plant)	2941	3300	12%
Micromonas (green algae)	1472	1643	12%
Cyanidioschyzon (red algae)	711	777	9%
Ectocarpus (brown algae)	1891	2156	14%
Emiliana (haptophyte)	4397	5151	17%
Phaeodactylum (diatom)	1425	1610	13%
Thalassiosira (diatom)	1436	1637	14%
Cyanophora (glaucophyte)	2417	2739	13%
Average			**13%**

Table S8. COG functional category distribution of nuclear genes that are of cyanobacterial descent

Functional category of Cluster of Orthologous Group (COG) from cyanobacterial genomes retrieved in the nuclear genomes of diverse photosynthetic eukaryotes. The latter are indicated as followed: 1, *Arabidopsis*; 2, *Physcomitrella;* 3, *Micromonas;* 4, *Cyanidioschyzon;* 5, *Ectocarpus;* 6, *Emiliana;* 7, *Thalassiosira;* 8, *Phaeodactylum;* 9, *Cyanophora*

COG	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
RNA processing and modification	0%	0%	0%	0%	0%	0%	0%	0%	0%
Chromatin structure and dynamics	0%	0%	0%	0%	0%	0%	0%	0%	0%
Energy production and conversion	4%	7%	5%	5%	4%	4%	5%	5%	4%
Cell cycle control, cell division, chromosome partitioning	0%	1%	1%	1%	1%	1%	0%	1%	1%
Amino acid transport and metabolism	5%	6%	6%	9%	4%	4%	6%	7%	5%
Nucleotide transport and metabolism	1%	1%	1%	2%	1%	2%	1%	1%	1%
Carbohydrate transport and metabolism	7%	9%	7%	7%	5%	6%	6%	6%	5%
Coenzyme transport and metabolism	3%	4%	6%	7%	4%	4%	5%	5%	4%
Lipid transport and metabolism	8%	5%	3%	4%	3%	4%	4%	4%	2%
Translation, ribosomal structure and biogenesis	4%	5%	6%	7%	3%	4%	5%	5%	3%
Transcription	2%	3%	2%	2%	3%	2%	2%	2%	3%
Replication, recombination and repair	2%	2%	3%	5%	3%	4%	3%	3%	5%
Cell wall/membrane/ envelope biogenesis	6%	6%	5%	6%	4%	4%	4%	5%	3%
Cell motility	1%	0%	1%	0%	3%	1%	1%	1%	1%
Posttranslational modification, protein turnover, chaperones	7%	7%	9%	8%	8%	7%	9%	8%	6%
Inorganic ion transport and metabolism	4%	4%	4%	5%	4%	6%	4%	5%	4%
Secondary metabolites biosynthesis, transport and catabolism	6%	4%	5%	3%	4%	7%	4%	5%	3%

General function prediction only	21%	18%	20%	16%	25%	19%	23%	19%	23%
Function unknown	12%	11%	12%	9%	13%	13%	11%	11%	10%
Signal transduction mechanisms	4%	4%	3%	2%	5%	3%	4%	4%	14%
Intracellular trafficking, secretion, and vesicular transport	2%	2%	2%	2%	4%	2%	2%	2%	2%
Defense mechanisms	0%	0%	0%	0%	0%	1%	0%	0%	1%
Extracellular structures	0%	0%	0%	0%	0%	0%	0%	0%	0%
Cytoskeleton	0%	0%	0%	0%	0%	0%	0%	0%	0%

Table S9. Sequencing information of CyanoGEBA organisms
The finishing efforts are indicated as followed: MF, manual finishing; AF, autofinishing. Submit indicates that the genome sequence has been submitted to NCBI to obtain the BioProject number.

CyanoGEBA Organism	454 Libraries	454 Total Reads	454 Total Mb	Illumina Libraries	Illumina Total Reads	Illumina Total bp	Finishing efforts	Nb of contigs / scaffolds	IMG Taxon ID
Anabaena cylindrica PCC 7122	(1) 454 STD TIT, (3) 454 PE (9138 kb, 3178 kb, NA)	1,079,579	361.9	(1) ILL STD	180,472,451	6,497,008,236	MF	7 / 7	2503982047
Anabaena sp. PCC 7108	(1) 454 STD TIT, (2) 454 PE (11344kb, 4036 kb)	727,027	181.2	(1) ILL STD	60,554,068	4,602,109,168	AF	13 / 3	2506485002
Calothrix sp. PCC 6303	(1) 454 STD TIT, (2) 454 PE (9829 kb, 4087.8 kb)	1,303,031	461.6	(1) ILL STD	115,161,558	8,752,278,408	MF	4 / 4	2503982036
Calothrix sp.	(0) 454	640,339	216.1	(1) ILL	37,899,348	2,880,350,448	AF	67 / 12	2507262048

PCC 7103	STD TIT, (2) 454 PE (5331 kb, 6844 kb)			STD					
Calothrix sp. PCC 7507	(1) 454 STD TIT, (2) 454 PE (5438 kb, 2730 kb)	672,159	258.3	(1) ILL STD	42,042,292	3,195,214,192	MF	1 / 1	2505679032
Chamaesiphon minutus PCC 6605	(1) 454 STD TIT, (1) 454 PE (6916 kb)	976,084	247.3	(1) ILL STD	60,314,630	4,583,911,880	MF	3 / 3	2510436000
Chroococcidiopsis sp. PCC 6712	(1) 454 STD TIT, (3) 454 PE (2604 kb, 12,305 kb, 2694 kb)	1,269,117	353.5	(1) ILL STD	36,438,868	1,311,799,248	AF	18 / 3	2505679029
Chroococcidiopsis thermalis PCC 7203	(1) 454 STD TIT,	788,934	272.3	(3) ILL STD	32,800,000	1,180,704,000	MF	3 / 3	2503538021

	(1) 454 PE (8583 kb)								
Crinalium epipsammum PCC 9333	(1) 454 STD TIT, (1) 454 PE (8063)	230,731	128.8	(1) ILL STD	30,965,529	1,114,759,044	MF	9 / 9	2504643013
Cyanobacterium aponinum PCC 10605	(2) 454 STD TIT, (2) 454 PE (NA, NA)	519,034	145	(1) ILL STD	43,225,758	3,285,157,608	MF	2 / 2	2503707009
Cyanobacterium stanieri PCC 7202	(1) 454 STD TIT, (1) 454 PE (8540 kb)	754,375	252.4	(1) ILL STD	2,050,270	366,482,655	MF	1 / 1	2503283023
Cyanobium gracile PCC 6307	(1) 454 STD TIT, (1) 454 PE (7784 kb)	356,894	159	(1) ILL STD	66,080,366	5,022,107,816	MF	1 / 1	2508501011

Cylindrospermum stagnale PCC 7417	(1) 454 STD TIT, (2) 454 PE (6956 kb, 4374 kb)	1,662,064	379.2	(1) ILL STD	74,952,294	5,696,374,344	AF	10 / 4	2509601025
Dactylococcopsis salina PCC 8305	(1) 454 STD TIT, (1) 454 PE (7217 kb)	976,293	246.7	(1) ILL STD	29,937,544	1,077,751,584	MF	1 / 1	2509276056
Fischerella sp. PCC 9339	-	-	-	(1) ILL STD, (1) ILL PE	31,117,314	4,667,600,000	none	171 / 95	2516653082
Fischerella sp. PCC 9431	-	-	-	(1) ILL STD, (1) ILL PE (6617 kb)	560,072,428	81,357,230	none	201 / 36	2512875027
Fischerella sp. PCC 9605	-	-	-	(1) ILL STD, (1) ILL PE (2209 kb)	45,267,538	6,790,130,000	none	49 / 36	2516143000
Geitlerinema sp. PCC 7105	(1) 454 STD TIT,	1,285,347	304.2	(1) ILL STD	116,062,307	7,311,925,341	AF	288 / 8	2510065011

	(2) 454 PE (10539 kb, 4458 kb)								
Geitlerinema sp. PCC 7407	(1) 454 STD TIT, (1) 454 PE (4018 kb)	292,666	167.4	(1) ILL STD	37,618,333	2,858,993,308	MF	1 / 1	2503538020
Geminocystis herdmanii PCC 6308	-	-	-	(1) ILL STD	64,203,930	4,882,710,000	AF	11 /1	2509601046
Gloeocapsa sp. PCC 73106	(1) 454 STD TIT, (2) 454 PE / (8550 kb and 7666 kb)	481,442	297.2	(1) ILL STD	62,560,585	4,754,604,460	none	228/ 228	2508501033
Gloeocapsa sp. PCC 7428	(1) 454 STD TIT, (1) 454 PE/ (9786 kb)	129,654	226.5	(1) ILL STD	31,204,529	576,136,120	MF	5 / 5	2503754017
Halothece sp. PCC 7418	(0) 454 STD TIT, (2) 454 PE	902,827	216.1	(1) ILL STD	257,227,056	19,549,256,256	MF	1 / 1	2503538028

	(2627 kb, 9799 kb)								
Leptolyngbya boryana PCC 6306	-	-	-	(1) ILL STD	9,298,704	6,649,250,000	AF	11 / 5	2509601031
Leptolyngbya sp. PCC 6406	(1) 454 STD TIT, (2) 454 PE (8212 kb)	1,049,271	273.4	(1) ILL STD	86,532,372	6,576,460,272	none	377 / 377	2517572073
Leptolyngbya sp. PCC 7375	(1) 454 STD TIT, (1) 454 PE/ (12811 kb)	228,442	170	(1) ILL STD	22,675,741	816,326,676	AF	40 / 5	2509601039
Leptolyngbya sp. PCC 7376	-	-	-	(1) ILL STD, (1) ILL PE (2481 kb)	529,092,128	79,363,820,000	MF	1 / 1	2503754048
Mastigocladopsis repens PCC 10914	(1) 454 STD TIT, (2) 454 PE (9610 kb, 3964 kb)	1,444,337	316.7	(1) ILL STD	25,286,224	910,304,064	none	325 / 23	2517093042

Microchaete sp. PCC 7126	(1) 454 PE/ (117346 kb)	735,764	109.6	(1) ILL STD	69,022,092	5,245,678,992	AF	5 / 3	2509601027
Microcoleus sp. PCC 7113	(2) 454 STD TIT, (3) 454 PE (4283 kb, 7800 kb, NA)	626,176	201.3	(1) ILL STD	57,251,139	4,351,086,564	MF	9 / 9	2509276031
Nodosilinea nodulosa. PCC 7104	(1) 454 STD TIT, (4) 454 PE (2798 kb, 24356 kb, 22893 kb, 11125 kb)	1,921,672	486.1	(1) ILL STD	25,897,163	932,297,868	AF	62 / 2	2509601026
Nostoc sp. PCC 7107	(1) 454 STD TIT, (2) 454 PE (1695 kb, 4068 kb)	2,132,299	546.3	(1) ILL STD	62,447,094	4,745,979,144	MF	1 / 1	2503707008
Nostoc sp. PCC 7524	(1) 454 STD TIT, (2)	681,222	256.3	(1) ILL STD	17,798,114	640,732,104	MF	3 / 3	2509601032

	454 PE (11786 kb, 11762 kb)								
Oscillatoria acuminata PCC 6304	(0) 454 STD TIT, (1) 454 PE (8203 kb)	652,065	129.4	(1) ILL STD	67,180,232	5,105,697,632	MF	3 / 3	2509276028
Oscillatoria formosa PCC 6407	(1) 454 STD TIT, (2) 454 PE	1,050,403	253.9	(1) ILL STD	25,052,472	901,888,992	AF	259 / 12	2508501075
Oscillatoria nigro-viridis PCC 7112	(1) 454 STD TIT, (2) 454 PE (8172 kb , 6631 kb)	1,446,977	433.8	(1) ILL STD	46,329,519	3,521,043,444	AF	108 / 6	2503982035
Oscillatoria sp. PCC 10802	(1) 454 STD TIT, (2) 454 PE	499,658	244.6	(1) ILL STD	70,039,722	5,323,018,872	MF	6 / 9	2509276047
Pleurocapsa sp. PCC 7319	(1) 454 STD TIT, (1) 454	1,020,605	299.4	(1) ILL STD	31,122,538	2,365,312,888	AF	30 / 10	2509601013

	PE (12243 kb)								
Pleurocapsa sp. PCC 7327	(1) 454 STD TIT, (2) 454 PE (1525 kb, 7351 kb)	1,361,678	352.7	(1) ILL STD	145,035,126	11,022,669,576	MF	1 / 1	2509276061
Prochlorothrix hollandica PCC 9006	(1) 454 STD TIT, (2) 454 PE (5749 kb, 8122 kb)	830,913	198.7	(1) ILL STD	112,562,730	8,554,767,480	AF	233 / 13	2509276045
Pseudanabaena sp. PCC 6802	(1) 454 STD TIT, (2) 454 PE (4119 kb, 12050 kb)	1,300,658	271.9	(1) ILL STD	31,942,889	1,149,944,004	AF	28 / 6	2506783054
Pseudanabaena sp. PCC 7367	(0) 454 STD TIT, (1) 454 PE (10442 kb)	396,482	75.9	(1) ILL STD	82,635,242	6,280,278,392	MF	2 / 2	2504643012

Pseudanabaena sp. PCC 7429	(1) 454 STD TIT, (2) 454 PE (9299 kb, 3092 kb)	613,351	198.8	(1) ILL STD	83,683,990	6,359,983,240	AF	517 / 464	2504557005
Rivularia sp. PCC 7116	(1) 454 STD TIT, (3) 454 PE (3104 kb, 22486 kb, 22469 kb)	1,240,665	224	(1) ILL STD	47,209,745	1,699,550,820	MF	3 / 3	2510065008
Spirulina major PCC 6313	(1) 454 STD TIT, (1) 454 PE (5378 kb)	487,235	257.8	(1) ILL STD	87,627,634	6,659,700,184	AF	10 / 2	2506520014
Spirulina subsalsa PCC 9445	(1) 454 STD TIT, (1) 454 PE (13930 kb)	404,680	198	(1) ILL STD	61,669,554	4,686,886,104	AF	10 / 2	2506520011
Stanieria cyanosphaera PCC 7437	(0) 454 STD TIT,	378,359	74.8	(1) ILL STD	86,083,820	6,542,370,320	MF	6 / 6	2503754019

	(1) 454 PE (7497 kb)								
Synechococcus sp. PCC 6312	(1) 454 STD TIT, (1) 454 PE (4402 kb)	823,816	251.4	(1) ILL STD	72,440,844	5,505,504,144	MF	2 / 2	2509276030
Synechococcus sp. PCC 7336	(1) 454 STD TIT, (2) 454 PE (4179 kb, 22856 kb)	949,313	199.2	(1) ILL STD	44,507,806	3,382,593,256	AF	9 / 2	2506520048
Synechococcus sp. PCC 7502	(1) 454 STD TIT, (3) 454 PE (1102 kb, 9022 kb, 9794 kb)	573,805	166.2	(1) ILL STD	86,633,080	6,150,948,680	MF	8 / 3	2508501041
Synechocystis sp. PCC 7509	-	-	-	(1) ILL STD	3,753,429	5,832,004,000	none	174 / 174	2517572074
Tolypothrix sp. PCC 9009	(0) 454 STD TIT, (1)	920,752	178.4	(1) ILL STD	72,204,518	5,487,543,368	AF	167 / 204	2504756053

	454 PE (7854 kb)								
Xenococcus sp. PCC 7305		-	-	(1) ILL STD	9,298,704	7,029,000,000	none	234 / 225	2508501034
Unidentified cyanobacterium PCC 7702	-	-	-	(1) ILL STD, (1) ILL PE	45,267,538	6,790,130,000	none	49 / 4	2512564012

References:

1. Bennett S (2004) Solexa Ltd. *Pharmacogenomics* 5(4):433-438.
2. Margulies M*, et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057):376-380.
3. Zerbino DR & Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18(5):821-829.
4. Ewing B & Green P (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res* 8(3):186-194.
5. Ewing B, Hillier L, Wendl MC, & Green P (1998) Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res* 8(3):175-185.
6. Gordon D, Abajian C, & Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Res* 8(3):195-202.
7. Han C & Chain P (2006) Finishing repeat regions automatically with Dupfinisher. *Proceeding of the 2006 international conference on bioinformatics & computational biology.*, eds Arabnia HR & Valafar H (CSREA Press), pp 141-146.
8. Gnerre S*, et al.* (2010) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* 108(4):1513-1518.
9. Hyatt D*, et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11(1):119.
10. Pati A*, et al.* (2010) GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nat Methods* 7(6):455-457.
11. Lowe TM & Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25(5)(0305-1048):955–964.

12. Pruesse E*, et al.* (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35(21):7188-7196.
13. Markowitz VM*, et al.* (2009) IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* 25(17):2271-2278.
14. Wu M & Eisen JA (2008) A simple, fast, and accurate method of phylogenomic inference. *Genome biology* 9(10):R151.
15. Katoh K, Kuma K-i, Toh H, & Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33(2):511-518.
16. Eddy SR (1998) Profile hidden markov models. *Bioinformatics* 14(9):755-763.
17. Sonnhammer E & Hollich V (2005) Scoredist: A simple and robust protein sequence distance estimator. *BMC Bioinformatics* 6(1):108.
18. Guindon S*, et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Sys Biol* 59(3):307-321.
19. Abascal F, Zardoya R, & Posada D (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21(9):2104-2105.
20. Colless DH (1982) Phylogenetics: The theory and practice of phylogenetic systematics. *Syst Zool* 31(1):100-104.
21. Maddison WP & Maddison DR (2011) Mesquite: A modular system for evolutionary analysis. *Version 2.75* http://mesquiteproject.org.
22. Lima T*, et al.* (2009) HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res* 37(suppl 1):D471-D478.
23. Grissa I, Vergnaud G, & Pourcel C (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* 35(suppl 2):W52-W57.
24. Bland C*, et al.* (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 8(1):1-8.
25. Rouhiainen L*, et al.* (2004) Genes Coding for Hepatotoxic Heptapeptides (Microcystins) in the Cyanobacterium Anabaena Strain 90. *Appl Environ Microbiol* 70(2):686-692.
26. Fewer D*, et al.* (2007) Recurrent adenylation domain replacement in the microcystin synthetase gene cluster. *BMC Evol Biol* 7(1):183.
27. Fewer DP*, et al.* (2011) Nostophycin Biosynthesis Is Directed by a Hybrid Polyketide Synthase-Nonribosomal Peptide Synthetase in the Toxic Cyanobacterium Nostoc sp. Strain 152. *Appl Environ Microbiol* 77(22):8034-8040.
28. Fan Q*, et al.* (2005) Clustered genes required for synthesis and deposition of envelope glycolipids in Anabaena sp. strain PCC 7120. *Mol Microbiol* 58(1):227-243.
29. Balskus EP & Walsh CT (2010) The Genetic and Molecular Basis for Sunscreen Biosynthesis in Cyanobacteria. *Science* 329(5999):1653-1656.

30. Miller S, Wood A, Blankenship RE, Kim M, & Ferriera S (2011) Dynamics of gene duplication in the genomes of chlorophyll *d*-producing cyanobacteria: Implications for the ecological niche. *Genome Biol Evol* 3:601-613.
31. Swingley W*, et al.* (2008) Niche adaptation and genome expansion in the chlorophyll d-producing cyanobacterium *Acaryochloris marina*. *Proc Natl Acad Sci USA* 12:2005-2010.
32. Bench S, Ilikchyan I, Tripp H, & Zehr J (2011) Two strains of *Crocosphaera watsonii* with highly conserved genomes are distinguished by strain-specific features. *Front Microbiol* 2:261.
33. Shi T, Ilikchyan I, Rabouille S, & Zehr J (2010) Genome-wide analysis of diel gene expression in the unicellular N_2-fixing cyanobacterium *Crocosphaera watsonii* WH 8501. *ISME J* 4:621-632.
34. Welsh E*, et al.* (2008) The genome of *Cyanothece* 51142, a unicellular diazotrophic cyanobacterium important in the marine nitrogen cycle. *Proc Natl Acad Sci USA* 105:15094-15099.
35. Bandyopadhyay A*, et al.* (2011) Novel metabolic attributes of the genus *Cyanothece*, comprising a group of unicellular nitrogen-fixing cyanobacteria. *mBio* 2:e00214-00211. doi:00210.01128/mBio.00214-00211.
36. Nakamura Y*, et al.* (2003) Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. *DNA Res* 10:137-145.
37. Kaneko T*, et al.* (2007) Complete genomic structure of the bloom-forming toxic cyanobacterium *Microcystis aeruginosa* NIES-843. *DNA Res* 14:247-256.
38. Frangeul L*, et al.* (2008) Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics* 9:274.
39. Kettler G*, et al.* (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* 3:e231. doi:210.1371/journal.pgen.0030231.
40. Coleman M*, et al.* (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* 311:1768-1770.
41. Rocap G*, et al.* (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* 424:1042-1047.
42. Dufresne A*, et al.* (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci USA* 100:10020-10025.
43. Donia M*, et al.* (2011) Complex microbiome underlying secondary and primary metabolism in the tunicate-*Prochloron* symbiosis. *Proc Natl Acad Sci USA* 108:E1423-1432.
44. Sugita C*, et al.* (2007) Complete nucleotide sequence of the freshwater unicellular cyanobacterium *Synechococcus elongatus* PCC 6301 chromosome: gene content and organization. *Photosynth Res* 93:55-67.
45. Dufresne A*, et al.* (2008) Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome biology* 9:R90. doi:10.1186/gb-2008-1189-1185-r1190.
46. Palenik B*, et al.* (2006) Genome sequence of *Synechococcus* CC9311: Insights into adaptation to a coastal environment. *Proc Natl Acad Sci USA* 103:13555-13559.

47. Bhaya D*, et al.* (2007) Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. *ISME J* 1:703-713.

48. Palenik B*, et al.* (2003) The genome of a motile marine *Synechococcus*. *Nature* 424:1037-1042.

49. Kaneko T*, et al.* (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC 6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* 3:109-136.

50. Nakamura Y*, et al.* (2002) Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1. *DNA Res* 9:123-130.

51. Tripp H*, et al.* (2010) Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* 464:90-94.

52. Fujisawa T*, et al.* (2010) Genomic structure of an economically important cyanobacterium, *Arthrospira* (Spirulina) *platensis* NIES-39. *DNA Res* 17:85-103.

53. Janssen P*, et al.* (2010) Genome sequence of the edible cyanobacterium *Arthrospira* sp. PCC 8005. *Journal of bacteriology* 192:2465-2466.

54. Starkenburg S*, et al.* (2011) Genome of the cyanobacterium *Microcoleus vaginatus* FGP-2, a photosynthetic ecosystem engineer of arid land soil biocrusts worldwide. *Journal of bacteriology* 193:4569-4570.

55. Jones A*, et al.* (2011) Genomic insights into the physiology and ecology of the marine filamentous cyanobacterium *Lyngbya majuscula*. *Proc Natl Acad Sci USA* 108:8815-8820.

56. Méjean A*, et al.* (2010) The genome sequence of the cyanobacterium *Oscillatoria* sp. PCC 6506 reveals several gene clusters responsible for the biosynthesis of toxins and secondary metabolites. *Journal of bacteriology* 192:5264-5265.

57. Stucken K*, et al.* (2010) The smallest known genomes of multicellular and toxic cyanobacteria: comparison, minimal gene sets for linked traits and the evolutionary implications. *PLoS ONE* 5:e9235. doi:9210.1371/journal.pone.0009235.

58. Ran L*, et al.* (2010) Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. *PLoS ONE* 5:e11486. doi:11410.11371/journal.pgen.0030231.

59. Kaneko T*, et al.* (2001) Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res* 8:227-253.

Appendix B

Supplemental Information for Chapter 4

Supplemental Material and Methods:

Application of the methods and comparison of results. To measure the effect of cross-calibration vs. cross-bracing on an overall dating analysis and the effect of different amounts of prior dating information, nine separate relaxed-clock dating analyses (Table S1) were performed using the program BEAST. The first three runs were done with only the α-subunits from the overall alignment (run 1), only β-subunits (run 2), and both α and β subunits, calibrated under the cross-calibration method (run 3, Figure 1). All available node date calibrations were used in these analyses. To examine the effect of systematically removing certain categories of calibration points, runs 4-8 used only α-subunits, with various categories of subunits removed from the calibrations (Table S1). Specifically, run 4 only included metazoan calibration points, and run 5 only included plant calibration points. We also tested the effect of using calibration points that were more broadly sampled from across the α-subunit gene tree, but were symbiont/nuclear-specific; thus run 6 used only the calibration points in the chloroplast subclade, run 7 the mitochondrial subclade, and run 8 used only the calibration points from the vacuolar subclade. Finally, run 9 used the same sequence data as run 3 (both α and β subunits in one large gene tree), but used the cross-bracing method.

All runs except the last were sampled for ten million generations, with samples collected every 1000 generations, and with the first 50% discarded as burnin. The cross-bracing approach induced additional autocorrelation in the runs, requiring much longer runs to assure convergence and adequate ESS (estimated sample size) values for all parameters. Therefore, for cross-bracing, four independent BEAST runs of approximately 40 million generations each were conducted. In each case, the first 20% was discarded as burn-in (as this appeared to be well past the burn-in period), and the remaining samples were concatenated. This resulted in 142,555 samples representing 142.5 million generations of post-burnin sampling.

All runs were inspected in Tracer for convergence and adequate estimated sample size (ESS) values. Sampling was judged to be adequate in all cases. As expected, linked node dates are highly correlated in the cross-bracing approach, resulting in lower ESS values (~50-100) for the linked dates. However, this was not of great concern, as the dates for these nodes are specified in prior distributions, and are not the target of inference, and sample sizes of 50-100 are still adequate to indicate reasonable sampling of the overall distribution. This was confirmed by inspection of trace plots for each parameter, confirming that a reasonable range of values was being stochastically sampled, rather than having a parameter “stuck” on a particular value.

Majority-rule dated consensus trees were generated by using TreeAnnotator v1.73 on the posterior sample of trees for each analysis. The node date and branch rate estimates from each of the nine TreeAnnotator-derived consensus trees were compared to each other using linear regression procedures in R. In addition to comparing the estimates between runs, the uncertainty of these estimates was compared, as measured by the 95% highest posterior density (HPD width) on the dates and rates. As the uncertainty in dates and rates is typically heteroscedastic (i.e. nodes with higher absolute ages will typically have higher absolute uncertainty in age), the runs were also compared by relative uncertainty in dates and rates as measured by CV (coefficient of variation). CV is equal to the standard deviation divided by the mean; here, standard deviation (SD) was approximated as SD=((95% HPD width) / (2 x 1.96)), an approximation which applies well as almost all nodes had approximately normal distributions (the only exception were the LUCA-related nodes, but these were not a target of inference in our study).

In each comparison of two BEAST runs (comparing node ages, branch rates, uncertainty measures, etc.) the null hypothesis was that the linear regression between the two would have a slope of 1:1 and an intercept of 0. This null expectation would be the guaranteed result if two BEAST runs produced e.g. identical estimates of node age for all nodes, or if node ages from a run were regressed against themselves. Bonferroni corrections for multiple testing were applied, multiplying *p*-values by 5 tests for the comparisons between standard and cross-calibration/cross-bracing runs (Table S2).

As BEAST analyses are complex and highly parameterized, priors may interact in unexpected ways to influence results, and this might particularly be an issue as our analyses contain many nodes with dating priors and fixed topology. To ensure that the data rather than the combination of dating priors and tree prior was dominating the inference of node ages for non-calibration nodes deeper in the tree, an alignment lacking data, with all amino acids changed to gaps, was generated and run in BEAST with all calibration points. The no-data run yielded calibration node dates closely following the prior specifications, and non-calibration nodes estimated by the no-data dataset had either dramatically different dates or were not resolved at all, giving strong evidence that the amino acid sequence data are strongly influencing our results.

Supplemental Analysis of BEAST runs:

Node Age Uncertainty. The intercept terms on the regressions were not significantly different than zero (Table S2). Comparing the cross-calibration and cross-bracing analyses indicated that cross-bracing may provide slightly lower uncertainty (3%) than the cross-calibration method (Table S2); however, the effect was not statistically significant (*p*=0.598).

Inference of Coefficient of Variation. Examination of the CV in node age shows that the reduction in node age uncertainty was not due to mere reductions in the average node age (a concern because node age and node age uncertainty are strongly correlated in BEAST analyses). α/β-calibrated analyses had CVs 14% lower than standard analyses, with strongly significant *p*-values (all *p*-values <0.00023). The regression of the CVs of the α-calibrated and β-calibrated analyses against the CVs of cross-braced showed an even greater deviation from a 1:1 slope (22% and 18% lower, respectively), however, this is more than compensated for by a significant positive intercept. A similar, but weaker, effect was found when comparing the α/β-braced and α/β-calibrated runs. In other words, when CVs from a cross-calibrated analysis are used to predict CVs from a cross-braced analysis, the cross-bracing-derived CVs are typically higher by a fixed amount (the intercept), but this effect declines at very high CVs due to a slope below 1:1. The typically slightly higher CVs in the cross-bracing analysis compared to the cross-calibration analysis can be attributed mostly to the fact that the uncertainty in node age declines little between the two methods, but node ages are consistently slightly lower in the cross-bracing analysis.

Inference of node dates. There is no significant difference between the ages estimated by the α-calibrated or β-calibrated analyses (runs 1 and 2) and the cross-calibrated analysis (run 3; Table S2); the slope between mean node ages estimated by the two methods is not significantly different from 1:1.

Inference of branch rates. Only small differences were found in the estimation of branch substitution rates (Table S2) between the all cross-calibrated runs. Intercepts were not significantly different from zero, and the slope was indistinguishable from 1:1 for β-calibrated vs. cross-calibrated analysis. There was a significant difference in the slope between α-calibrated and α/β-calibrated analysis, but the size of the effect was small. Comparison of the rate estimates from the α-calibrated and β-calibrated runs and the α/β-calibrated run with the α/β-braced run showed significantly negative slopes, however the regression also contains a highly significant and positive intercept with a large effect size; the estimated means of substitution rate on each branch of the cross-bracing analysis are very often higher than for the other analyses.

Uncertainty in branch rates. Comparing simple and complex analyses in their estimates of uncertainty in mean per-branch substitution rate estimates (Table S2) appear to show a violation of the linear model used in regression, with the relationship between rate uncertainty in the α-calibrated and β-calibrated runs and rate uncertainty in the α/β-calibrated and α/β-braced runs showing a linear relationship at lower uncertainties, but flattening out at mid-to-high rate uncertainties. This is expressed in the regression statistics as highly significant p-values (p<1.7E-07 for all comparisons) and large effect size with slopes 40-60% below 1:1, and intercepts representing approximately 25% of the overall mean rate uncertainty. When the two more complex analyses (α/β-calibrated and α/β-braced) are compared, it is evident that the cross-bracing analysis produces higher absolute estimates of rate uncertainty than those of the cross-calibration estimate by a substantial margin, mostly due to the large intercept (p=7.43E-13).

The higher absolute uncertainty in the rate estimates for cross-bracing may simply be due to the higher rates estimated by cross-bracing, and heteroscedasticity in rate uncertainty. Examination of the regressions based on rate CV (Table S2) shows that this is indeed the case, but that rate uncertainties increased proportionately less than the rate means in the cross-bracing analysis. Therefore, rate CV is typically lower for cross-braced estimates than for either α-calibrated, β-calibrated, or α/β-calibrated analyses. As with estimates of branch rate uncertainty, there appears to be some evidence of nonlinearity, with the relationship between rate CVs from simpler and more complex analyses flattening out at higher rate CVs. In summary, cross-braced analyses have less relative uncertainty in their estimates of branch substitution rates when rates and rate uncertainty are high.

Effect of reducing the number of node age calibrations. α-calibrated runs using less calibration dates, runs 4-8 (Table S3), were compared against the α/β-calibrated and α/β-braced runs. Datasets with fewer calibration points had systematically slightly lower estimates of node ages than the α-calibrated and α/β-calibrated runs. The differences were statistically significant (Table S3), with slope and intercept both positively elevated. Interestingly, the difference in node age estimates closely matched the difference in node age estimates between α/β-calibrated and α/β-braced runs, such that the node age differences between runs 4-8 and the cross-bracing run are minimal and mostly insignificant.

Uncertainty in node age was not dramatically different between the α-calibrated dataset with all calibration node ages versus subsets of these calibrations (Table S3), with all of the statistical tests being non-significant at the p=0.05 level, or barely significant with small effect. Comparison of the reduced-calibrations (α-calibrated) runs to the α/β-calibrated analysis showed that runs 7 and 8 (only plant-based calibrations, and only vacuolar calibrations) had significantly higher uncertainty on average (slopes respectively 12% and 15% higher than 1:1) than the α/β-calibrated analysis. When compared to the α/β-braced analysis, all reduced-calibrations runs had significantly higher uncertainty, but as found above, higher uncertainty is expected even with the complete set of calibrations used on the α-calibrated dataset (Table S2).

Comparison of the node age CVs from runs 4-8 to the α-calibrated, α/β-calibrated, and α/β-braced runs indicated consistently significantly higher CVs in the reduced-calibration runs. In the case of comparison to α-calibrated and α/β-calibrated runs, this effect is likely due to the consistently lower estimates of node ages in the reduced-calibrations runs, plus perhaps slightly increased absolute uncertainties in node age. The fact that the comparison of node age CV to between runs 4-8 and the cross-bracing run shows highly significant difference in slope (all p-values < 0.001), with the reduced-calibrations having 35%-49% higher CVs, indicates that the effect cannot be solely due to differences in node age. Inspection of the node age CV regression plots shows that certain nodes are dramatically different in CV between the reduced calibrations runs and the all-calibrations runs. These nodes showing a large difference are the ones which are calibrated in one analysis and not in the other. In other words, when a node that was calibrated in one run is uncalibrated in another, the uncertainty in its age increases dramatically, while the mean estimate of its age changes comparatively little.

Similar comparisons of branch rate estimates and uncertainty are shown in Table S4. Branches typically have slightly higher rates in the reduced-calibrations runs, explaining the slightly lower average node ages, however, unlike in the node age comparisons, the differences are not significant (Table S4) in the comparison to α-calibrated (run 1), probably because of greater scatter in the rate estimates between runs compared to the age estimates.

Absolute uncertainty in branch rates is almost always higher in the reduced-calibration runs; of course, this is a partial product of the higher rate estimates in these runs. Comparison to the α/β-braced analysis, which has similar mean estimates of node ages and branch rates, shows dramatically increased CV in the reduced-calibrations runs, indicating that removing calibration points increases relative uncertainty; however, comparison to the α-calibrated run shows no significant differences in rate CV. Overall, these results show that uncertainty is reduced the most when calibration points from both alpha and beta subunits are utilized with a cross-calibration approach, and is reduced further if the cross-bracing approach is used to tie the node dates together. However, the effect of removing node calibrations within subgroups of the alpha subtree is minimal on the overall estimates, although the effect of changing an individual calibrated node to an uncalibrated one can be a significant increase in age uncertainty. Node age uncertainty is strongly affected by nearby date calibrations, but overall mean node ages are determined by the estimates of branch rates. In a BEAST uncorrelated relaxed-clock analysis, the rates of each branch are drawn from a common lognormal distribution, which should be robust to the inclusion or exclusion of small groups of calibration points.

Supplemental Figures S1-S10:

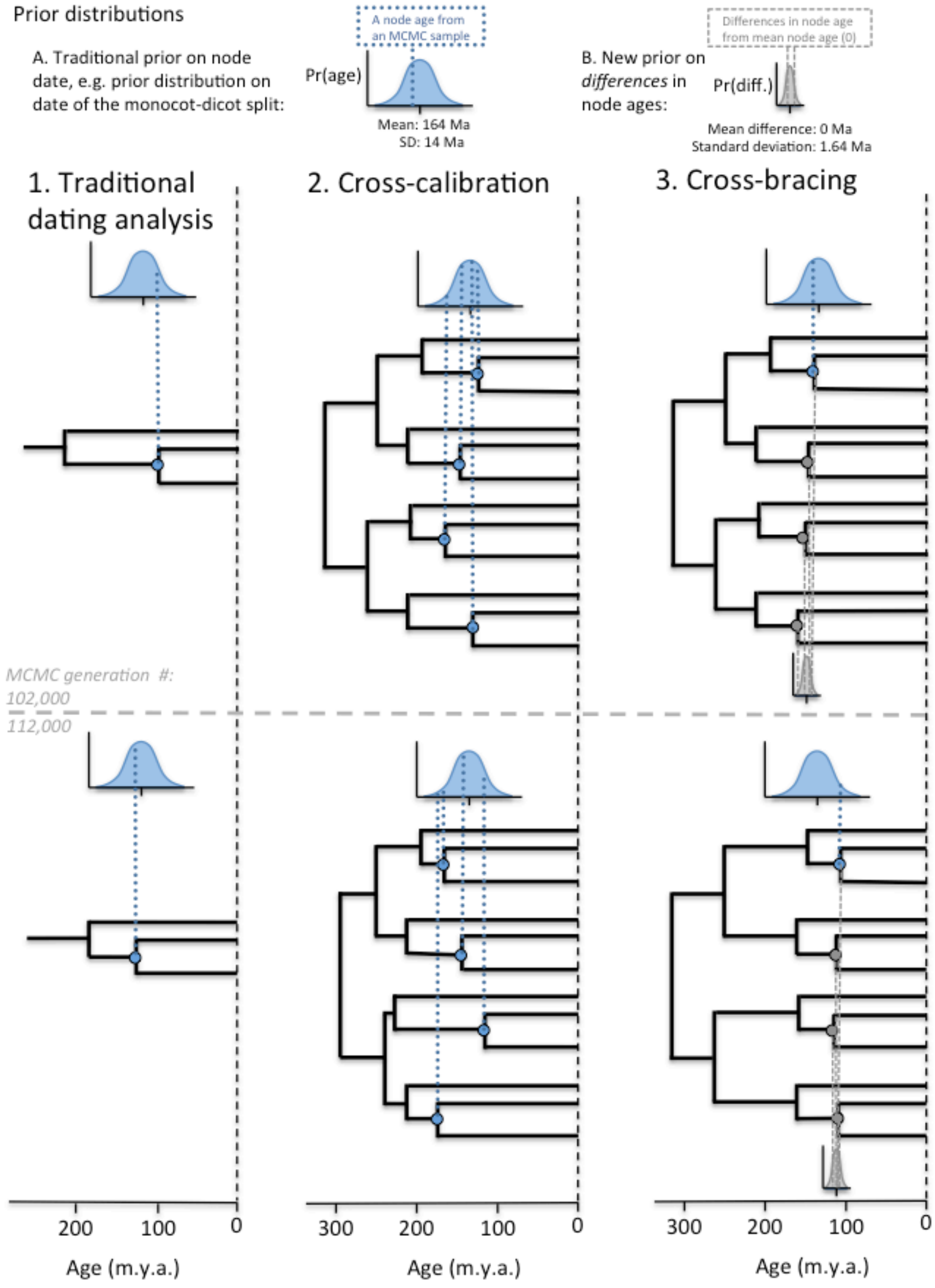

Prior distributions
A. Traditional prior on node date, e.g. prior distribution on date of the monocot-dicot split:
A node age from an MCMC sample
Pr(age)
Mean: 164 Ma
SD: 14 Ma
B. New prior on *differences* in node ages:
Differences in node age from mean node age (0)
Pr(diff.)
Mean difference: 0 Ma
Standard deviation: 1.64 Ma
1. Traditional dating analysis
2. Cross-calibration
3. Cross-bracing
MCMC generation #:
102,000
112,000
200
100
0
Age (m.y.a.)
300
200
100
0
Age (m.y.a.)
300
200
100
0
Age (m.y.a.)

Figure S1: Summary of the cross-calibration and cross-bracing strategies.

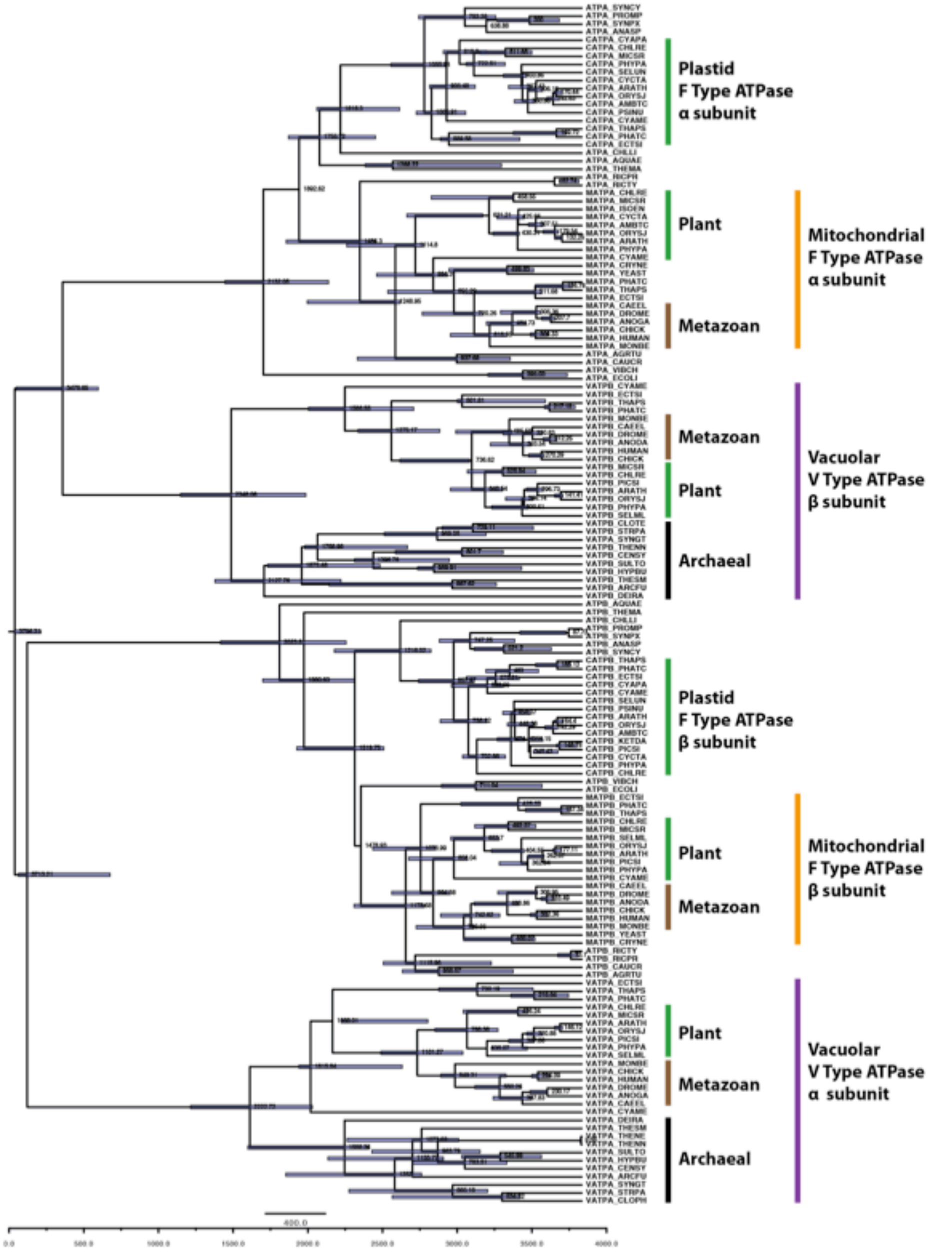


Figure S2: ATPase α and β subunits cross-calibrated chronogram

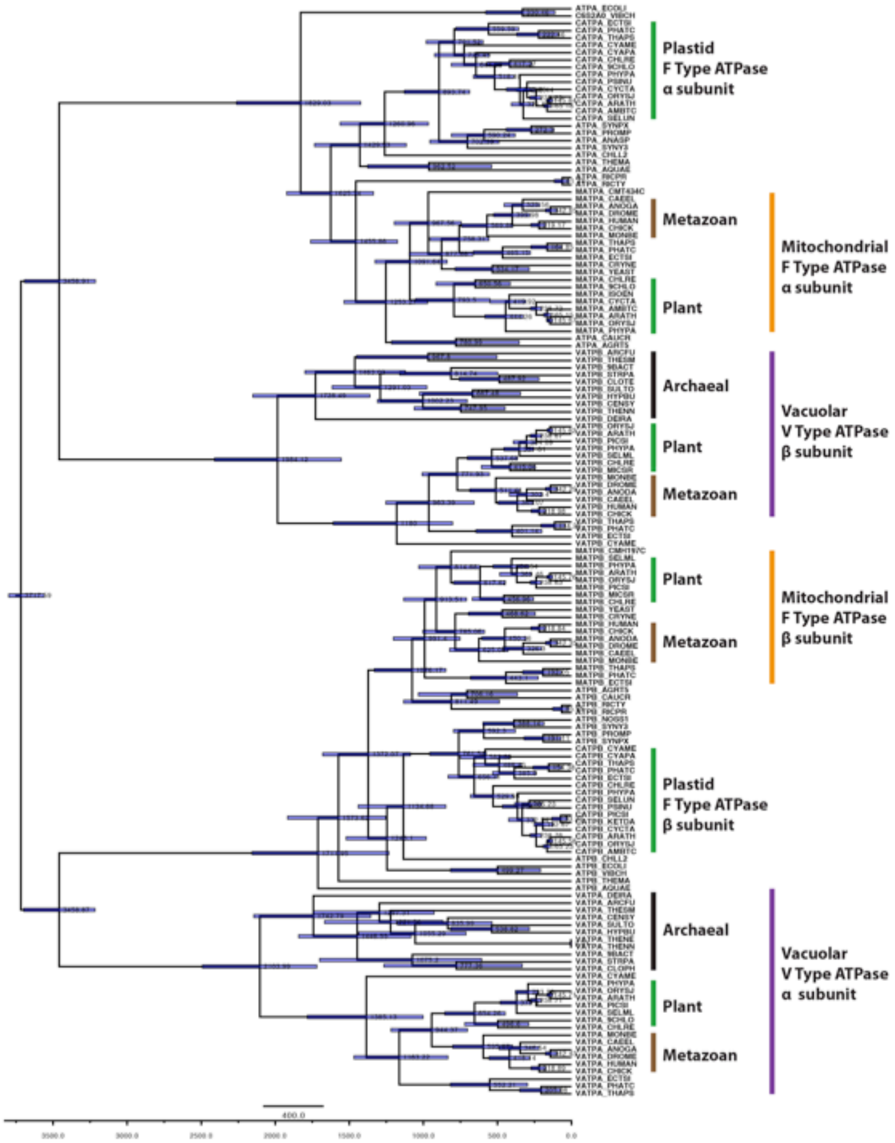


Figure S3: ATPase α and β subunits cross-braced chronogram

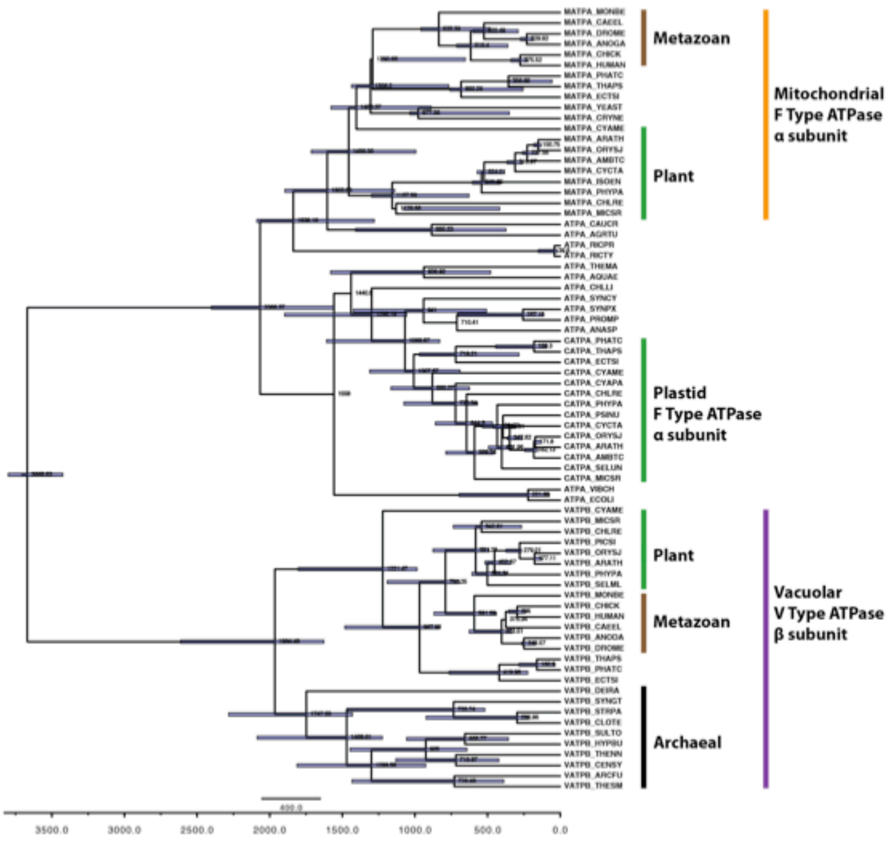


Figure S4: ATPase α-subunit cross-calibrated chronogram

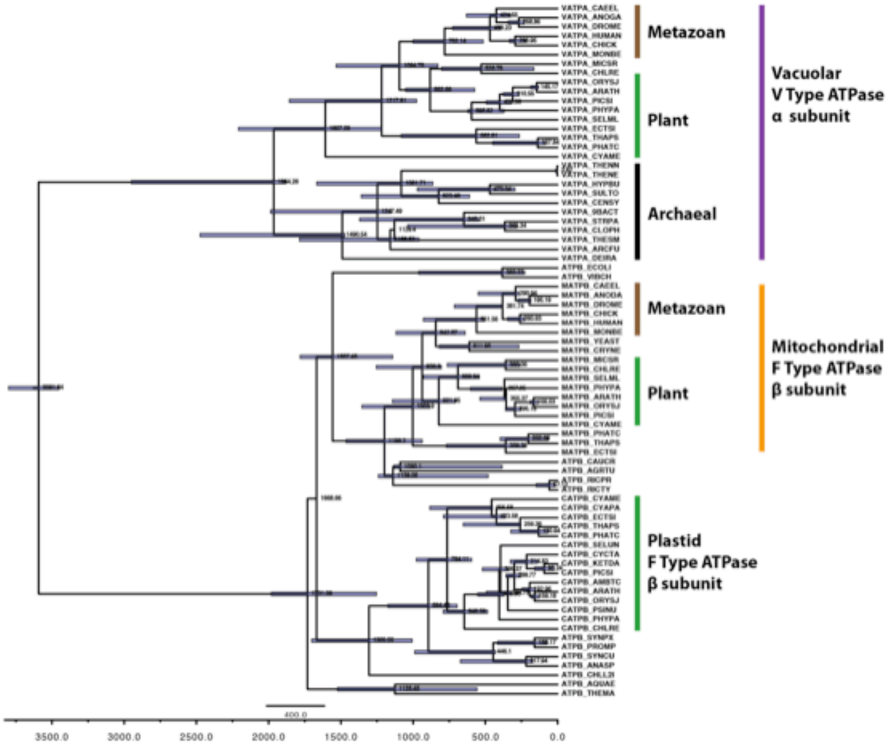


Figure S5: ATPase β subunit cross-calibrated chronogram

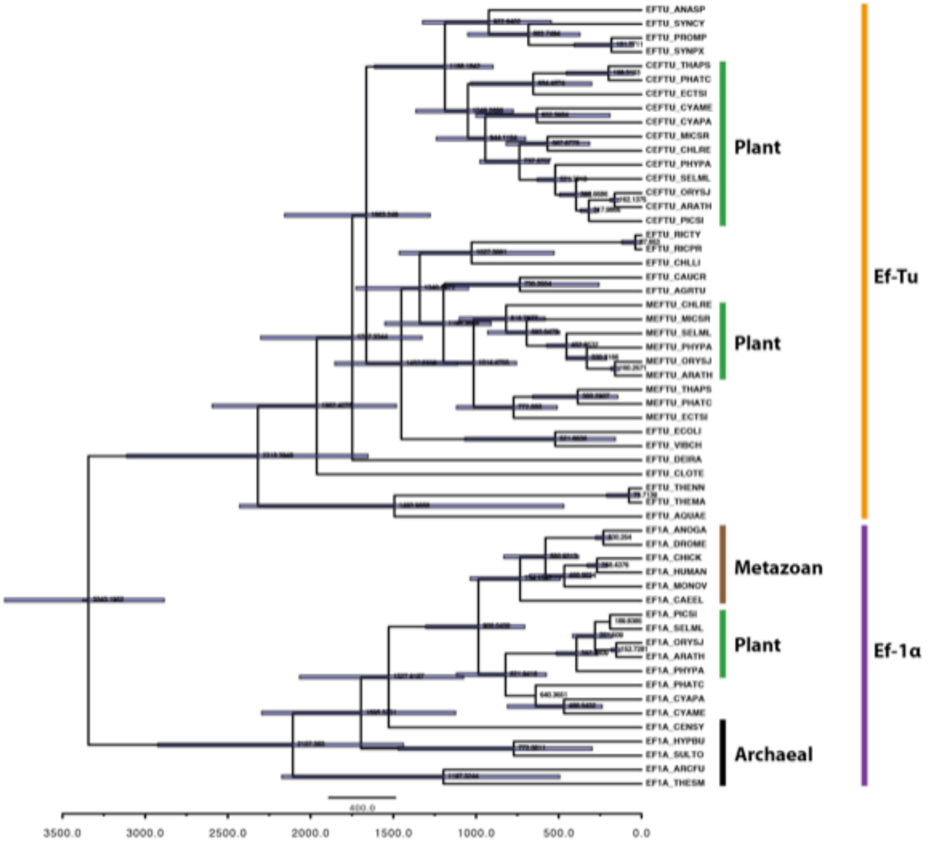


Figure S6: Ef-Tu cross-calibrated chronogram

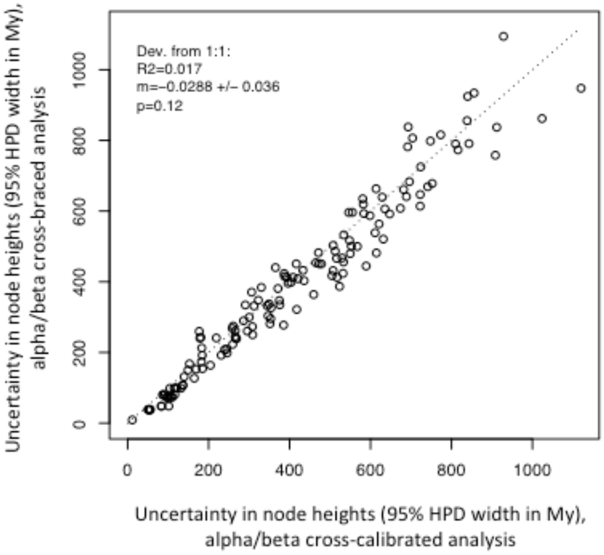


Figure S7: No significant difference in uncertainty (as measured by 95% HPD width of node height estimates) between alpha/beta cross-calibrated and cross-braced analyses. Regression statistics are for deviation from 1:1 line.

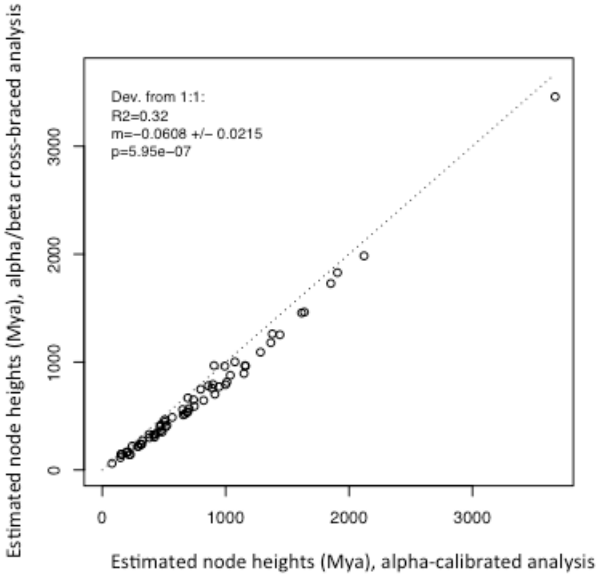


Figure S8: Node heights as estimated by alpha/beta cross-bracing tend to be slightly lower than with other methods. Regression statistics are for deviation from 1:1 line.

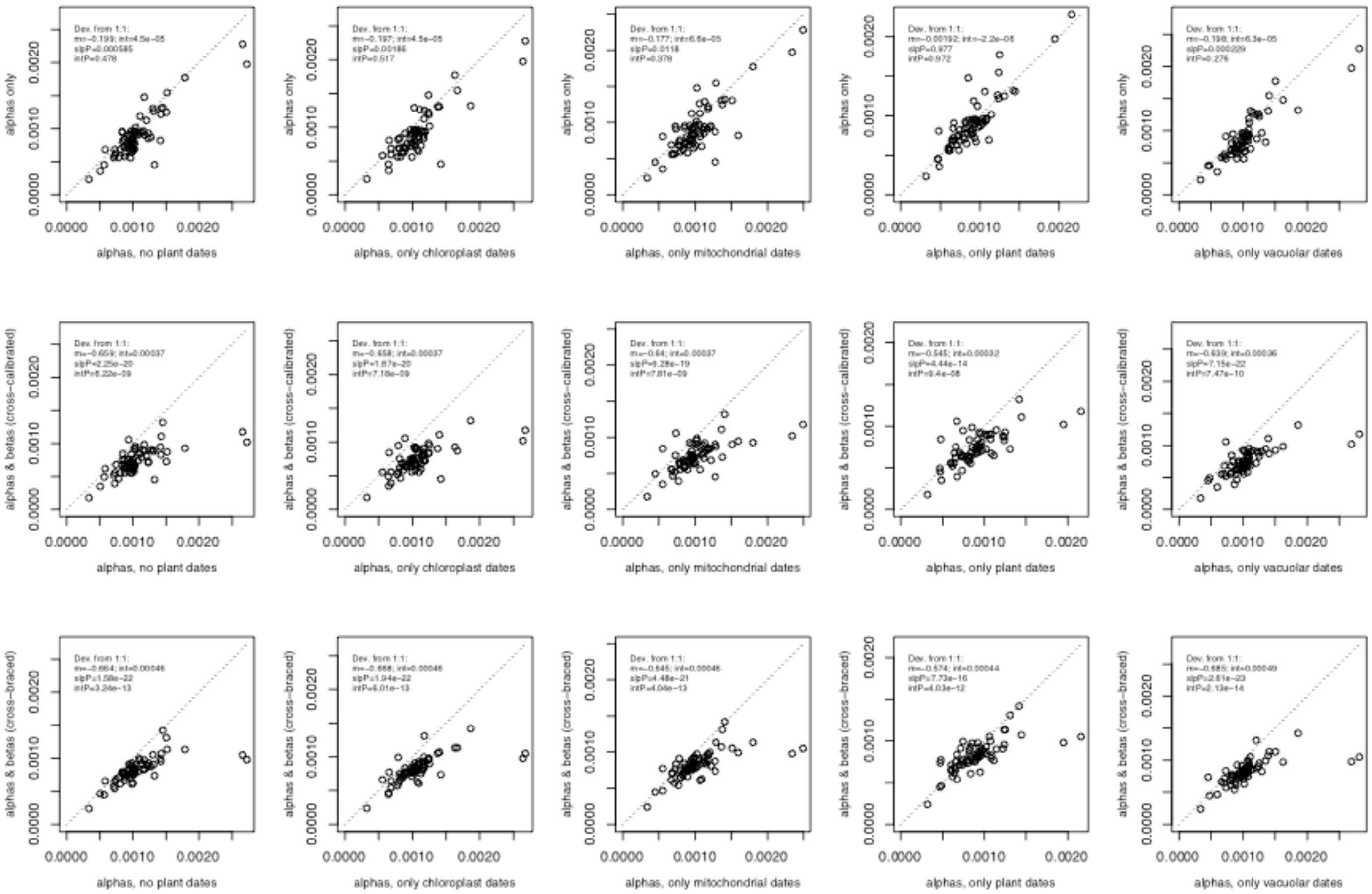


Figure S9: Uncertainty in branch rate estimates as measured by 95% HPD width of branch rates. Five analysis with alpha-only subunits and reduced numbers of calibration points (runs 4-8) are compared with run 1 in row 1 (alpha-only, all calibration points), with run 3 in row 2 (alpha/beta cross-calibrated, all calibration points), and with run 9 in row 3 (alpha/beta cross-braced, all calibration points). The dotted line shows the 1:1 slope between the analyses; points below the line indicate lower uncertainty in the analyses using all calibration points (runs 1,3, and 9). Regression statistics are calculated for deviations from the 1:1 line.

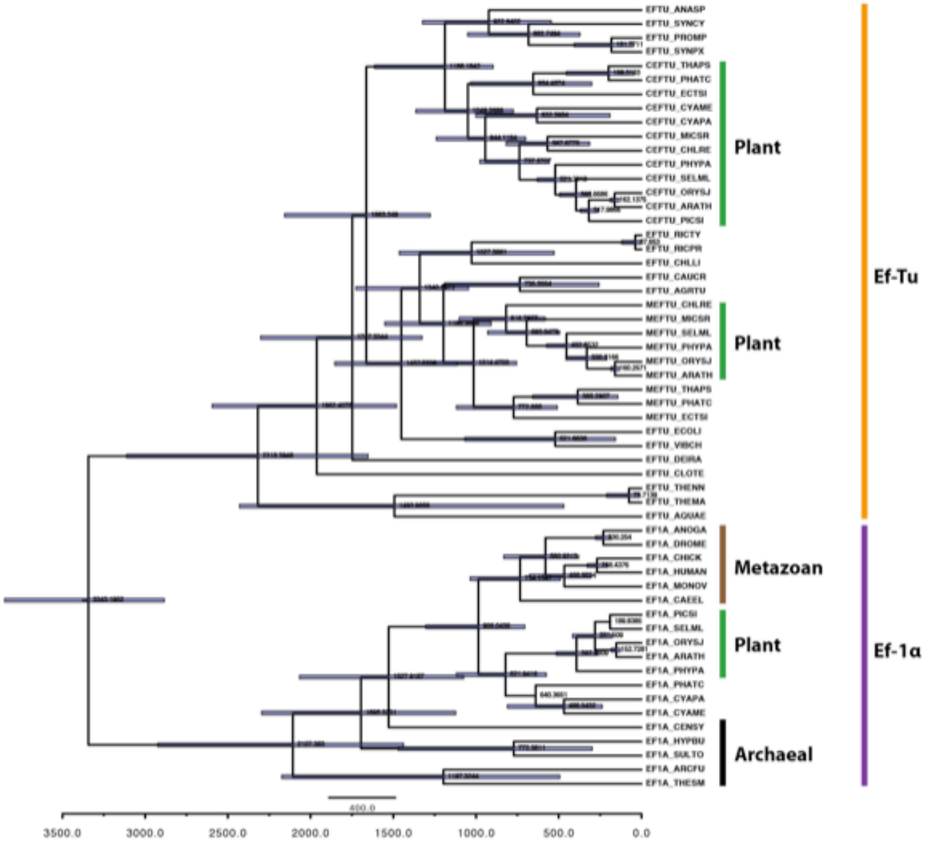


Figure S6: Ef-Tu/1α cross-calibrated chronogram

Supplemental Tables S1-S6:

Table S1. Description of ATPase BEAST runs with varying levels of calibration priors

Run	Name	Calibration Type	Description
1	α-cross-calibrated	cross-calibrated	chronogram of only α subunits; all calibration points are used
2	β-cross-calibrated	cross-calibrated	chronogram of only β subunits; all calibration points are used
3	α/β-cross-calibrated	cross-calibrated	chronogram of both α and β subunits; all calibration points are used
4	α-metazoan-cross-calibrated	cross-calibrated	chronogram of only α subunits; only metazoan calibration points are used
5	α-plant-cross-calibrated	cross-calibrated	chronogram of only α subunits; only plant calibration points are used
6	α-chloroplast-cross-calibrated	cross-calibrated	chronogram of only α subunits; only plastid calibration points are used
7	α-mitochondria-cross-calibrated	cross-calibrated	chronogram of only α subunits; only mitochondrial calibration points are used
8	α-vacuole-cross-calibrated	cross-calibrated	chronogram of only α subunits; only vacuolar calibration points are used
9	α/β-cross-braced	cross-braced	chronogram of both α and β subunits; all calibration points are used

Table S2. Statistical tests for deviations from a 1:1 relationship between calibration analyses. Linear models were built using a simpler analysis as a predictor (x-axis), and a more complex analysis as a response (y-axis). To remove the 1:1 relationship, the response variable was detrended by subtracting the predictor variable. Thus, if the relationship between e.g. node age uncertainty from a simpler analysis and node age uncertainty in a more complex analysis is truly 1:1, then the detrended response variable will be exactly flat (slope and intercept = 0). A negative slope of -0.05 would indicate that uncertainty in node age is on average 5% lower in the more complex analysis (if the intercept is close to 0). The results show that the mean estimates of node age in cross-calibrated analyses are not significantly lower than in non-calibrated, alpha-alone analyses. *P*-values were corrected by Bonferroni correction for 5 tests. *=*p*<0.05; **=*p*<0.01; ***=*p*<0.001.

Comparison	Slope of difference from 1:1 line	95% CI	p		Intercept	95% CI	p	
age mean								
alpha vs. cross-calibrated	0.02	(-0.01,0.04)	0.723		2.5	(-20.1,25.2)	4.13585	
beta vs. cross-calibrated	-0.04	(-0.08,-0.01)	0.054		31.7	(2,61.5)	0.20073	
alpha vs. cross-braced	-0.06	(-0.08,-0.04)	2.97E-06	***	-57.2	(-77.9,-36.5)	4.66E-06	***
beta vs. cross-braced	-0.07	(-0.1,-0.04)	1.19E-05	***	-37.3	(-61.6,-13)	0.01836	*
cross-calibrated vs. cross-braced	-0.05	(-0.07,-0.03)	5.75E-08	***	-65.0	(-81.4,-48.5)	8.64E-12	***
uncertainty in node age (HPD width)								
alpha vs. cross-calibrated	-0.22	(-0.3,-0.15)	1.48E-06	***	30.8	(-12.6,74.2)	0.84297	
beta vs. cross-calibrated	-0.16	(-0.22,-0.09)	9.50E-05	***	26.9	(-6.9,60.7)	0.61656	
alpha vs. cross-braced	-0.26	(-0.32,-0.19)	3.61E-10	***	16.6	(-20.6,53.8)	1.92425	
beta vs. cross-braced	-0.14	(-0.22,-0.07)	2.54E-03	**	3.4	(-35.2,42.1)	4.31047	
cross-calibrated vs. cross-braced	-0.03	(-0.06,0.01)	0.598		-7.9	(-25,9.3)	1.84405	
age coefficient of variation (CV: std. dev. / mean)								
alpha vs. cross-calibrated	-0.14	(-0.2,-0.08)	2.32E-04	***	-0.001	(-0.014,0.012)	4.55538	
beta vs. cross-calibrated	-0.14	(-0.17,-0.11)	2.96E-11	***	0.007	(0,0.015)	0.31751	
alpha vs. cross-braced	-0.22	(-0.28,-0.15)	4.44E-08	***	0.024	(0.01,0.037)	0.00453	**
beta vs. cross-braced	-0.18	(-0.22,-0.15)	2.48E-13	***	0.028	(0.019,0.036)	2.00E-07	***
cross-calibrated vs. cross-braced	-0.08	(-0.11,-0.04)	3.67E-05	***	0.024	(0.018,0.03)	9.21E-12	***
rate mean								
alpha vs. cross-calibrated	-0.12	(-0.16,-0.08)	3.79E-06	***	0.000031	(0.00001,0.00006)	0.10687	
beta vs. cross-calibrated	-0.06	(-0.12,-0.01)	0.138		0.000035	(0.00001,0.00007)	0.12761	
alpha vs. cross-braced	-0.24	(-0.3,-0.19)	1.75E-12	***	0.000167	(0.00014,0.0002)	1.13E-14	***
beta vs. cross-braced	-0.23	(-0.3,-0.16)	2.09E-08	***	0.000171	(0.00014,0.00021)	8.95E-13	***
cross-calibrated vs. cross-braced	-0.16	(-0.2,-0.13)	1.53E-16	***	0.000142	(0.00012,0.00016)	9.54E-31	***
uncertainty in branch rate (HPD width)								
alpha vs. cross-calibrated	-0.57	(-0.66,-0.48)	1.59E-18	***	0.000351	(0.00027,0.00044)	1.43E-10	***
beta vs. cross-calibrated	-0.42	(-0.52,-0.32)	5.56E-11	***	0.000269	(0.00018,0.00035)	1.76E-07	***
alpha vs. cross-braced	-0.61	(-0.7,-0.52)	3.25E-20	***	0.000470	(0.00039,0.00055)	1.46E-15	***
beta vs. cross-braced	-0.51	(-0.61,-0.41)	6.07E-14	***	0.000398	(0.00031,0.00048)	8.35E-13	***
cross-calibrated vs. cross-braced	-0.27	(-0.36,-0.19)	8.75E-09	***	0.000272	(0.00021,0.00034)	7.43E-13	***
rate coefficient of variation (CV: std. dev. / mean)								
alpha vs. cross-calibrated	-0.29	(-0.43,-0.14)	1.50E-03	**	0.078	(0.01,0.15)	0.13667	
beta vs. cross-calibrated	-0.14	(-0.28,0)	0.309		0.026	(-0.04,0.09)	2.10995	
alpha vs. cross-braced	-0.47	(-0.6,-0.33)	1.14E-08	***	0.142	(0.08,0.2)	1.04E-04	***
beta vs. cross-braced	-0.22	(-0.33,-0.12)	3.98E-04	***	0.046	(0,0.09)	0.29495	
cross-calibrated vs. cross-braced	-0.30	(-0.38,-0.23)	1.88E-12	***	0.106	(0.07,0.14)	2.33E-09	***

Table S3: Tests for deviations from a 1:1 relationship between the age-related node statistics from BEAST analyses based on subsets of alpha node calibrations, and (a) the complete list of alpha node calibrations, (b) the cross-calibration method, and (c) the cross-bracing method.

Comparison	Slope of difference from 1:1 line	95% CI	p	Inter-cept	95% CI	p
age mean						
(a) comparisons to alpha with all calibrations						
age mean: alphas, no plant dates vs. alphas only	0.06	(0.04,0.09)	4.849E-07 ***	65.34	(45.95,84.72)	8.034E-09 ***
age mean: alphas, only chloroplast dates vs. alphas only	0.09	(0.05,0.12)	6.160E-06 ***	103.43	(75.3,131.55)	7.455E-10 ***
age mean: alphas, only mitochondrial dates vs. alphas only	0.06	(0.04,0.09)	2.051E-05 ***	82.66	(60.08,105.24)	6.870E-10 ***
age mean: alphas, only plant dates vs. alphas only	0.03	(0.01,0.05)	2.099E-03 **	37.02	(18.82,55.21)	1.695E-04 ***
age mean: alphas, only vacuolar dates vs. alphas only	0.06	(0.03,0.08)	8.355E-06 ***	69.23	(49.57,88.89)	2.264E-09 ***
(b) comparison to cross-calibration						
age mean: alphas, no plant dates vs. alphas & betas (cross-calibrated)	0.08	(0.04,0.12)	4.051E-04 ***	71.72	(37.71,105.72)	1.073E-04 ***
age mean: alphas, only chloroplast dates vs. alphas & betas (cross-calibrated)	0.10	(0.06,0.15)	3.137E-05 ***	108.46	(71.3,145.63)	2.806E-07 ***
age mean: alphas, only mitochondrial dates vs. alphas & betas (cross-calibrated)	0.08	(0.04,0.12)	1.309E-04 ***	85.95	(51.58,120.31)	6.476E-06 ***
age mean: alphas, only plant dates vs. alphas & betas (cross-calibrated)	0.05	(0.01,0.08)	1.033E-02 *	42.54	(11.65,73.43)	0.009 **
age mean: alphas, only vacuolar dates vs. alphas & betas (cross-calibrated)	0.07	(0.04,0.11)	8.403E-05 ***	70.81	(41,100.61)	1.671E-05 ***
(c) comparison to cross-bracing						
age mean: alphas, no plant dates vs. alphas & betas (cross-braced)	0.00	(-0.03,0.03)	0.996	3.11	(-20.12,26.35)	0.794
age mean: alphas, only chloroplast dates vs. alphas & betas (cross-braced)	0.02	(0,0.05)	0.095	36.61	(13.72,59.51)	2.572E-03 **
age mean: alphas, only mitochondrial dates vs. alphas & betas (cross-braced)	0.01	(-0.02,0.03)	0.634	15.60	(-3.32,34.52)	0.111
age mean: alphas, only plant dates vs. alphas & betas (cross-braced)	-0.03	(-0.06,0)	0.031 *	-22.33	(-45.95,1.29)	0.068
age mean: alphas, only vacuolar dates vs. alphas & betas (cross-braced)	-0.01	(-0.03,0.02)	0.662	4.75	(-15.53,25.04)	0.647
uncertainty in node age (HPD width)						
(a) comparisons to alpha with all calibrations						
age HPD width: alphas, no plant dates vs. alphas only	0.11	(0.03,0.2)	0.012 *	-40.91	(-87.65,5.83)	0.091
age HPD width: alphas, only chloroplast dates vs. alphas only	0.10	(-0.01,0.2)	0.074	9.66	(-44.55,63.88)	0.728
age HPD width: alphas, only mitochondrial dates vs. alphas only	0.12	(0.02,0.22)	0.021 *	-11.13	(-63.45,41.18)	0.678
age HPD width: alphas, only plant dates vs. alphas only	0.06	(0,0.13)	0.055	0.25	(-34.65,35.15)	0.989
age HPD width: alphas, only vacuolar dates vs. alphas only	0.01	(-0.08,0.11)	0.771	-7.36	(-64,49.27)	0.800
(b) comparison to cross-calibration						
age HPD width: alphas, no plant dates vs. alphas & betas (cross-calibrated)	-0.09	(-0.19,0)	0.057	-17.62	(-67.78,32.54)	0.494
age HPD width: alphas, only chloroplast dates vs. alphas & betas (cross-calibrated)	-0.09	(-0.18,-0.01)	0.035 *	11.74	(-31.61,55.09)	0.597
age HPD width: alphas, only mitochondrial dates vs. alphas & betas (cross-calibrat	-0.08	(-0.17,0.01)	0.098	-3.78	(-52,44.43)	0.878
age HPD width: alphas, only plant dates vs. alphas & betas (cross-calibrated)	-0.12	(-0.19,-0.05)	0.002 **	9.77	(-27.95,47.48)	0.614
age HPD width: alphas, only vacuolar dates vs. alphas & betas (cross-calibrated)	-0.15	(-0.23,-0.07)	3.222E-04 ***	-8.13	(-52.15,35.88)	0.718
(c) comparison to cross-bracing						
age HPD width: alphas, no plant dates vs. alphas & betas (cross-braced)	-0.11	(-0.18,-0.05)	1.351E-03 **	-39.82	(-74.92,-4.71)	0.030 *
age HPD width: alphas, only chloroplast dates vs. alphas & betas (cross-braced)	-0.11	(-0.17,-0.05)	7.710E-04 ***	-8.06	(-38.46,22.34)	0.605
age HPD width: alphas, only mitochondrial dates vs. alphas & betas (cross-braced)	-0.09	(-0.15,-0.02)	9.042E-03 **	-25.90	(-59.3,7.5)	0.133
age HPD width: alphas, only plant dates vs. alphas & betas (cross-braced)	-0.16	(-0.21,-0.1)	4.198E-07 ***	-5.23	(-34.26,23.8)	0.725
age HPD width: alphas, only vacuolar dates vs. alphas & betas (cross-braced)	-0.18	(-0.24,-0.11)	4.466E-07 ***	-22.86	(-57.78,12.06)	0.204
age coefficient of variation (CV: std. dev. / mean)						
(a) comparisons to alpha with all calibrations						
age CV: alphas, no plant dates vs. alphas only	-0.26	(-0.45,-0.06)	0.014 *	0.01	(-0.033,0.062)	0.548
age CV: alphas, only chloroplast dates vs. alphas only	-0.44	(-0.64,-0.24)	5.201E-05 ***	0.05	(0.005,0.104)	0.033 *
age CV: alphas, only mitochondrial dates vs. alphas only	-0.35	(-0.55,-0.15)	8.868E-04 ***	0.04	(-0.006,0.087)	0.089
age CV: alphas, only plant dates vs. alphas only	-0.28	(-0.43,-0.13)	5.529E-04 ***	0.04	(0.007,0.074)	0.021 *
age CV: alphas, only vacuolar dates vs. alphas only	-0.23	(-0.41,-0.05)	0.013 *	0.01	(-0.036,0.05)	0.743
(b) comparison to cross-calibration						
age CV: alphas, no plant dates vs. alphas & betas (cross-calibrated)	-0.30	(-0.47,-0.13)	1.159E-03 **	0.00	(-0.044,0.039)	0.903
age CV: alphas, only chloroplast dates vs. alphas & betas (cross-calibrated)	-0.42	(-0.58,-0.26)	3.843E-06 ***	0.02	(-0.017,0.063)	0.269
age CV: alphas, only mitochondrial dates vs. alphas & betas (cross-calibrated)	-0.37	(-0.54,-0.2)	6.878E-05 ***	0.02	(-0.024,0.056)	0.430
age CV: alphas, only plant dates vs. alphas & betas (cross-calibrated)	-0.34	(-0.47,-0.21)	4.814E-06 ***	0.03	(-0.004,0.056)	0.092
age CV: alphas, only vacuolar dates vs. alphas & betas (cross-calibrated)	-0.26	(-0.4,-0.11)	9.659E-04 ***	-0.01	(-0.05,0.022)	0.443
(c) comparison to cross-bracing						
age CV: alphas, no plant dates vs. alphas & betas (cross-braced)	-0.44	(-0.62,-0.27)	6.836E-06 ***	0.04	(-0.002,0.084)	0.065
age CV: alphas, only chloroplast dates vs. alphas & betas (cross-braced)	-0.49	(-0.65,-0.33)	4.545E-08 ***	0.05	(0.012,0.088)	0.012 *
age CV: alphas, only mitochondrial dates vs. alphas & betas (cross-braced)	-0.45	(-0.61,-0.29)	7.410E-07 ***	0.05	(0.008,0.085)	0.020 *
age CV: alphas, only plant dates vs. alphas & betas (cross-braced)	-0.37	(-0.48,-0.25)	3.818E-08 ***	0.04	(0.016,0.068)	2.253E-03 **
age CV: alphas, only vacuolar dates vs. alphas & betas (cross-braced)	-0.35	(-0.5,-0.21)	1.206E-05 ***	0.02	(-0.017,0.054)	0.322

Table S4. Tests for deviations from a 1:1 relationship between the rate-related node statistics from BEAST analyses based on subsets of alpha node calibrations, and (a) the complete list of alpha node calibrations, (b) the cross-calibration method, and (c) the cross-bracing method.

rate mean: alphas, no plant dates vs. alphas & betas (cross-braced)	**Slope of difference from 1:1 line**	**95% CI**	**p**	**Intercept**	**95% CI**	**p**
rate mean						
(a) comparisons to alpha with all calibrations						
rate mean: alphas, no plant dates vs. alphas only	0.04	(-0.03,0.11)	0.267	0.00	-0.0001,-0.0001	8.803E-05 ***
rate mean: alphas, only chloroplast dates vs. alphas only	0.02	(-0.06,0.09)	0.682	0.00	-0.0002,-0.0001	1.503E-05 ***
rate mean: alphas, only mitochondrial dates vs. alphas only	0.01	(-0.06,0.08)	0.836	0.00	(-0.0001,0)	7.518E-04 ***
rate mean: alphas, only plant dates vs. alphas only	0.06	(-0.01,0.12)	0.084	0.00	(-0.0001,0)	0.014 *
rate mean: alphas, only vacuolar dates vs. alphas only	0.01	(-0.03,0.06)	0.559	0.00	(-0.0001,0)	1.507E-05 ***
(b) comparison to cross-calibration						
rate mean: alphas, no plant dates vs. alphas & betas (cross-calibrated)	-0.07	(-0.15,0)	0.057	0.00	(-0.0001,0)	0.014 *
rate mean: alphas, only chloroplast dates vs. alphas & betas (cross-calibrated)	-0.10	(-0.17,-0.03)	0.010 *	0.00	(-0.0001,0)	2.742E-03 **
rate mean: alphas, only mitochondrial dates vs. alphas & betas (cross-calibrated)	-0.10	(-0.17,-0.03)	5.140E-03 **	0.00	(-0.0001,0)	0.033 *
rate mean: alphas, only plant dates vs. alphas & betas (cross-calibrated)	-0.06	(-0.13,0.01)	0.075	0.00	(-0.0001,0)	0.535
rate mean: alphas, only vacuolar dates vs. alphas & betas (cross-calibrated)	-0.10	(-0.16,-0.04)	9.051E-04 ***	0.00	(-0.0001,0)	0.052
(c) comparison to cross-bracing						
rate mean: alphas, no plant dates vs. alphas & betas (cross-braced)	-0.19	(-0.25,-0.13)	1.034E-07 ***	0.00	(0,0.0001)	4.192E-04 ***
rate mean: alphas, only chloroplast dates vs. alphas & betas (cross-braced)	-0.21	(-0.27,-0.15)	1.080E-08 ***	0.00	(0,0.0001)	7.984E-03 **
rate mean: alphas, only mitochondrial dates vs. alphas & betas (cross-braced)	-0.20	(-0.25,-0.15)	3.719E-11 ***	0.00	(0,0.0001)	9.974E-06 ***
rate mean: alphas, only plant dates vs. alphas & betas (cross-braced)	-0.21	(-0.29,-0.13)	2.225E-06 ***	0.00	(0.0001,0.0002)	4.819E-07 ***
rate mean: alphas, only vacuolar dates vs. alphas & betas (cross-braced)	-0.22	(-0.28,-0.17)	9.393E-11 ***	0.00	(0.0001,0.0001)	5.259E-07 ***
uncertainty in branch rate (HPD width)						
(a) comparisons to alpha with all calibrations						
rate HPD width: alphas, no plant dates vs. alphas only	-0.20	(-0.31,-0.09)	5.854E-04 ***	0.00	(-0.0001,0.0002)	0.478
rate HPD width: alphas, only chloroplast dates vs. alphas only	-0.20	(-0.32,-0.08)	1.865E-03 **	0.00	(-0.0001,0.0002)	0.517
rate HPD width: alphas, only mitochondrial dates vs. alphas only	-0.18	(-0.31,-0.04)	0.012 *	0.00	(-0.0001,0.0002)	0.378
rate HPD width: alphas, only plant dates vs. alphas only	0.00	(-0.13,0.13)	0.977	0.00	(-0.0001,0.0001)	0.972
rate HPD width: alphas, only vacuolar dates vs. alphas only	-0.20	(-0.3,-0.1)	2.288E-04 ***	0.00	(0,0.0002)	0.276
(b) comparison to cross-calibration						
rate HPD width: alphas, no plant dates vs. alphas & betas (cross-calibrated)	-0.66	(-0.75,-0.56)	2.253E-20 ***	0.00	(0.0003,0.0005)	6.216E-09 ***
rate HPD width: alphas, only chloroplast dates vs. alphas & betas (cross-calibrated	-0.66	(-0.75,-0.56)	1.868E-20 ***	0.00	(0.0003,0.0005)	7.178E-09 ***
rate HPD width: alphas, only mitochondrial dates vs. alphas & betas (cross-calibrat	-0.64	(-0.74,-0.54)	9.277E-19 ***	0.00	(0.0003,0.0005)	7.807E-09 ***
rate HPD width: alphas, only plant dates vs. alphas & betas (cross-calibrated)	-0.54	(-0.65,-0.44)	4.441E-14 ***	0.00	(0.0002,0.0004)	9.405E-08 ***
rate HPD width: alphas, only vacuolar dates vs. alphas & betas (cross-calibrated)	-0.64	(-0.73,-0.55)	7.153E-22 ***	0.00	(0.0003,0.0005)	7.469E-10 ***
(c) comparison to cross-bracing						
rate HPD width: alphas, no plant dates vs. alphas & betas (cross-braced)	-0.66	(-0.75,-0.58)	1.583E-22 ***	0.00	(0.0004,0.0006)	3.241E-13 ***
rate HPD width: alphas, only chloroplast dates vs. alphas & betas (cross-braced)	-0.67	(-0.76,-0.58)	1.940E-22 ***	0.00	(0.0004,0.0006)	6.011E-13 ***
rate HPD width: alphas, only mitochondrial dates vs. alphas & betas (cross-braced	-0.65	(-0.74,-0.55)	4.481E-21 ***	0.00	(0.0004,0.0006)	4.036E-13 ***
rate HPD width: alphas, only plant dates vs. alphas & betas (cross-braced)	-0.57	(-0.68,-0.47)	7.726E-16 ***	0.00	(0.0003,0.0005)	4.027E-12 ***
rate HPD width: alphas, only vacuolar dates vs. alphas & betas (cross-braced)	-0.68	(-0.77,-0.6)	2.606E-23 ***	0.00	(0.0004,0.0006)	2.126E-14 ***
rate coefficient of variation (CV: std. dev. / mean)						
(a) comparisons to alpha with all calibrations						
rate CV: alphas, no plant dates vs. alphas only	-0.05	(-0.22,0.12)	0.561	0.01	(-0.07,0.09)	0.743
rate CV: alphas, only chloroplast dates vs. alphas only	0.06	(-0.17,0.29)	0.598	-0.01	(-0.11,0.09)	0.849
rate CV: alphas, only mitochondrial dates vs. alphas only	-0.01	(-0.21,0.19)	0.946	0.02	(-0.07,0.11)	0.710
rate CV: alphas, only plant dates vs. alphas only	-0.06	(-0.22,0.1)	0.458	0.04	(-0.03,0.11)	0.254
rate CV: alphas, only vacuolar dates vs. alphas only	0.01	(-0.19,0.21)	0.915	-0.01	(-0.11,0.08)	0.774
(b) comparison to cross-calibration						
rate CV: alphas, no plant dates vs. alphas & betas (cross-calibrated)	-0.26	(-0.47,-0.06)	0.016 *	0.06	(-0.04,0.16)	0.232
rate CV: alphas, only chloroplast dates vs. alphas & betas (cross-calibrated)	-0.07	(-0.29,0.15)	0.516	0.00	(-0.1,0.09)	0.930
rate CV: alphas, only mitochondrial dates vs. alphas & betas (cross-calibrated)	-0.09	(-0.29,0.1)	0.344	0.00	(-0.08,0.09)	0.940
rate CV: alphas, only plant dates vs. alphas & betas (cross-calibrated)	-0.22	(-0.41,-0.04)	0.023 *	0.06	(-0.02,0.14)	0.156
rate CV: alphas, only vacuolar dates vs. alphas & betas (cross-calibrated)	-0.09	(-0.28,0.11)	0.396	-0.02	(-0.11,0.07)	0.631
(c) comparison to cross-bracing						
rate CV: alphas, no plant dates vs. alphas & betas (cross-braced)	-0.42	(-0.59,-0.25)	1.053E-05 ***	0.12	(0.04,0.2)	6.351E-03 **
rate CV: alphas, only chloroplast dates vs. alphas & betas (cross-braced)	-0.26	(-0.44,-0.08)	6.080E-03 **	0.06	(-0.02,0.14)	0.146
rate CV: alphas, only mitochondrial dates vs. alphas & betas (cross-braced)	-0.30	(-0.47,-0.14)	6.007E-04 ***	0.08	(0,0.15)	0.042 *
rate CV: alphas, only plant dates vs. alphas & betas (cross-braced)	-0.33	(-0.47,-0.18)	2.962E-05 ***	0.09	(0.03,0.15)	6.652E-03 **
rate CV: alphas, only vacuolar dates vs. alphas & betas (cross-braced)	-0.30	(-0.47,-0.13)	9.526E-04 ***	0.06	(-0.02,0.14)	0.148

Table S5. List of taxa and UniProt accessions for sequences included in this study.

Species	**Taxon Abbreviation**	
Uniprot Accession		
<u>F-type ATPase (α subunit)</u>		
Plastid		
Arabidopsis thaliana	CATPA_ARATH	P56757
Oryza sativa subsp. Japonica	CATPA_ORYSJ	P0C2Z6
Amborella trichopoda	CATPA_AMBTC	Q70XV0
Cycas taitungensis	CATPA_CYCTA	A6H5F1
Psilotum nudum	CATPA_PSINU	Q8WI30
Selaginella uncinata	CATPA_SELUN	Q2WGJ0
Physcomitrella patens subsp. patens	CATPA_PHYPA	Q6YXK3
Chlamydomonas reinhardtii	CATPA_CHLRE	P26526
Micromonas sp. strain RCC299	CATPA_MICSR	C1KR42
Cyanophora paradoxa	CATPA_CYAPA	P48080
Cyanidioschyzon merolae	CATPA_CYAME	Q85FQ8
Thalassiosira pseudonana	CATPA_THAPS	A0T0P4
Ectocarpus siliculosus	CATPA_ECTSI	D1J797
Phaeodactylum tricornutum	CATPA_PHATC	A0T0F1
Mitochondrial		
Arabidopsis thaliana	MATPA_ARATH	P92549
Oryza sativa subsp. japonica	MATPA_ORYSJ	P0C522
Cycas taitungensis	MATPA_CYCTA	B0BLD4
Amborella trichopoda	MATPA_AMBTC	Q9T718
Isoetes engelmannii	MATPA_ISOEN	C6FJG2
Physcomitrella patens subsp. patens	MATPA_PHYPA	Q1XGA4
Chlamydomonas reinhardtii	MATPA_CHLRE	Q96550
Micromonas sp. strain RCC299	MATPA_MICSR	C1EHC0
Cyanidioschyzon merolae	MATPA_CYAME	CMT434C*
Homo sapiens	MATPA_HUMAN	P25705
Caenorhabditis elegans	MATPA_CAEEL	Q9XXK1
Drosophila melanogaster	MATPA_DROME	P35381
Anopheles gambiae	MATPA_ANOGA	P35381
Gallus gallus	MATPA_CHICK	Q8UVX3
Monosiga brevicollis	MATPA_MONBE	A9V9Z0
Saccharomyces cerevisiae	MATPA_YEAST	A9V9Z0
Cryptococcus neoformans var. neoformans	MATPA_CRYNE	P07251
Phaeodactylum tricornutum	MATPA_PHATC	B7G531
Thalassiosira pseudonana	MATPA_THAPS	B8C6C6
Ectocarpus siliculosus	MATPA_ECTSI	D8LJM3
Bacterial		
Synechocystis sp. strain PCC 6803	ATPA_SYNCY	P27179
Nostoc sp. strain PCC 7120	ATPA_ANASP	P12405
Prochlorococcus marinus subsp. pastoris strain CCMP1986	ATPA_PROMP	Q7V037
Synechococcus sp. strain WH8102	ATPA_SYNPX	Q7U8W5

Rickettsia prowazekii strain, Madrid E	ATPA_RICPR	O50288
Rickettsia typhi strain ATCC VR-144	ATPA_RICTY	Q68VU6
Caulobacter crescentus strain ATCC 19089	ATPA_CAUCR	Q9A2V7
Agrobacterium tumefaciens strain C58	ATPA_AGRTU	Q8UC74
Escherichia coli strain K12	ATPA_ECOLI	P0ABB0
Vibrio cholerae strain ATCC 39315	ATPA_VIBCH	Q9KNH3
Thermotoga maritima strain ATCC 43589	ATPA_THEMA	Q9X1U7
Chlorobium limicola strain DSM 245	ATPA_CHLLI	B3EHU6
Aquifex aeolicus strain VF5	ATPA_AQUAE	O66907

V-type ATPase (β subunit)

Arabidopsis thaliana	VATPB_ARATH	P11574
Oryza sativa subsp. japonica	VATPB_ORYSJ	Q9ASE0
Picea sitchensis	VATPB_PICSI	A9NVU9
Selaginella moellendorffii	VATPB_SELML	D8SQC5
Physcomitrella patens subsp. patens	VATPB_PHYPA	A9SP56
Chlamydomonas reinhardtii	VATPB_CHLRE	A8IA45
Micromonas sp. strain RCC299	VATPB_MICSR	C1DYK7
Cyanidioschyzon merolae	VATPB_CYAME	Q84KP2
Homo sapiens	VATPB_HUMAN	P15313
Caenorhabditis elegans	VATPB_CAEEL	Q19626
Drosophila melanogaster	VATPB_DROME	P31409
Anopheles darling	VATPB_ANODA	E3XA70
Gallus gallus	VATPB_CHICK	P49712
Monosiga brevicollis	VATPB_MONBE	A9V6U8
Phaeodactylum tricornutum	VATPB_PHATC	B7FQQ8
Thalassiosira pseudonana	VATPB_THAPS	B8C0L1
Ectocarpus siliculosus	VATPB_ECTSI	D8LCT6
Archaeoglobus fulgidus strain ATCC 49558	VATPB_ARCFU	O29100
Thermococcus sibiricus DSM 12597	VATPB_THESM	C6A5E7
Cenarchaeum symbiosum strain A	VATPB_CENSY	A0RXK0
Sulfolobus tokodaii strain DSM 16993	VATPB_SULTO	Q971B6
Hyperthermus butylicus strain DSM 5456	VATPB_HYPBU	A2BKX5
Thermotoga neapolitana strain ATCC 49049	VATPB_THENN	B9K814
Deinococcus radiodurans strain ATCC 13939	VATPB_DEIRA	Q9RWG7
Clostridium tetani strain Massachusetts / E88	VATPB_CLOTE	Q896K3
Streptococcus parasanguinis FW213	VATPB_STRPA	I1ZJ86
Synergistetes bacterium SGP1	VATPB_SYNGT	D4M879

F-type ATPase (β subunit)

Plastid

Arabidopsis thaliana	CATPB_ARATH	P19366
Oryza sativa subsp. japonica	CATPB_ORYSJ	P12085
Amborella trichopoda	CATPB_AMBTC	Q70XZ6
Cycas taitungensis	CATPB_CYCTA	A6H5I4
Picea sitchensis	CATPB_PICSI	C1IXH0
Keteleeria davidiana	CATPB_KETDA	B7ZIP2
Psilotum nudum	CATPB_PSINU	O03081
Selaginella uncinata	CATPB_SELUN	Q2WGH4

Physcomitrella patens subsp. patens	CATPB_PHYPA	P80658
Chlamydomonas reinhardtii	CATPB_CHLRE	P06541
Cyanophora paradoxa	CATPB_CYAPA	P48081
Cyanidioschyzon merolae	CATPB_CYAME	Q85FT2
Thalassiosira pseudonana	CATPB_THAPS	A0T0R6
Ectocarpus siliculosus	CATPB_ECTSI	D1J7B4
Phaeodactylum tricornutum	CATPB_PHATC	A0T0D2

Mitochondrial

Arabidopsis thaliana	MATPB_ARATH	P83483
Oryza sativa subsp. japonica	MATPB_ORYSJ	Q01859
Picea sitchensis	MATPB_PICSI	A9NUR7
Selaginella moellendorffii	MATPB_SELML	D8QQX5
Physcomitrella patens subsp. patens	MATPB_PHYPA	A9T281
Chlamydomonas reinhardtii	MATPB_CHLRE	P38482
Micromonas sp. strain RCC299	MATPB_MICSR	C1FE16
Cyanidioschyzon merolae	MATPB_CMH197C	CMH197C*
Homo sapiens	MATPB_HUMAN	P06576
Caenorhabditis elegans	MATPB_CAEEL	P46561
Drosophila melanogaster	MATPB_DROME	Q05825
Anopheles darling	MATPB_ANODA	E3XEC7
Gallus gallus	MATPB_CHICK	Q5ZLC5
Monosiga brevicollis	MATPB_MONBE	A9UYC5
Saccharomyces cerevisiae	MATPB_YEAST	P00830
Cryptococcus neoformans var. neoformans	MATPB_CRYNE	Q5KFU0
Ectocarpus siliculosus	MATPB_ECTSI	D7FXG1
Phaeodactylum tricornutum	MATPB_PHATC	B7FS46
Thalassiosira pseudonana	MATPB_THAPS	B5YP88

Bacterial

Synechocystis sp. strain PCC 6803	ATPB_SYNCY	P26527
Nostoc sp. strain PCC 7120	ATPB_ANASP	P06540
Prochlorococcus marinus subsp. pastoris strain CCMP1986	ATPB_PROMP	Q7V049
Synechococcus sp. strain WH8102	ATPB_SYNPX	Q7U8U7
Rickettsia prowazekii strain Madrid E	ATPB_RICPR	O50290
Rickettsia typhi strain ATCC VR-144	ATPB_RICTY	Q68VU8
Caulobacter crescentus strain ATCC 19089	ATPB_CAUCR	Q9A2V9
Agrobacterium tumefaciens strain C58	ATPB_AGRTU	Q8UC76
Escherichia coli strain K12	ATPB_ECOLI	P0ABB4
Vibrio cholerae strain ATCC 39315	ATPB_VIBCH	Q9KNH5
Thermotoga maritima strain ATCC 43589	ATPB_THEMA	O50550
Chlorobium limicola strain DSM 245	ATPB_CHLLI	B3EDQ7
Aquifex aeolicus strain VF5	ATPB_AQUAE	O67828

V-type ATPase (α subunit)

Arabidopsis thaliana	VATPA_ARATH	O23654
Oryza sativa subsp. japonica	VATPA_ORYSJ	Q651T8
Picea sitchensis	VATPA_PICSI	D5A887
Selaginella moellendorffii	VATPA_SELML	D8R3X7

Physcomitrella patens subsp. patens	VATPA_PHYPA	A9RGW5
Chlamydomonas reinhardtii	VATPA_CHLRE	A8I164
Micromonas sp. strain RCC299	VATPA_MICSR	C1E9Q8
Cyanidioschyzon merolae	VATPA_CYAME	Q84KP3
Homo sapiens	VATPA_HUMAN	P38606
Caenorhabditis elegans	VATPA_CAEEL	Q9XW92
Drosophila melanogaster	VATPA_DROME	P48602
Anopheles gambiae	VATPA_ANOGA	Q5TTG1
Gallus gallus	VATPA_CHICK	Q90647
Monosiga brevicollis	VATPA_MONBE	A9V438
Phaeodactylum tricornutum	VATPA_PHATC	B7G162
Thalassiosira pseudonana	VATPA_THAPS	B8CBV3
Ectocarpus siliculosus	VATPA_ECTSI	D8LGA9
Archaeoglobus fulgidus strain ATCC 49558	VATPA_ARCFU	O29101
Thermococcus sibiricus DSM 12597	VATPA_THESM	C6A5E8
Cenarchaeum symbiosum strain A	VATPA_CENSY	A0RXK1
Sulfolobus tokodaii strain DSM 16993	VATPA_SULTO	Q971B7
Hyperthermus butylicus strain DSM 5456	VATPA_HYPBU	A2BKX6
Thermotoga neapolitana strain ATCC 49049	VATPA_THENN	B9K813
Thermotoga neapolitana	VATPA_THENE	Q8GB11
Deinococcus radiodurans strain ATCC 13939	VATPA_DEIRA	Q9RWG8
Clostridium phytofermentans strain ATCC 700394	VATPA_CLOPH	A9KQV0
Streptococcus parasanguinis strain ATCC 15912	VATPA_STRPA	F8DGA3
Synergistetes bacterium SGP1	VATPA_SYNGT	D4M878

<u>Elongation factor Tu (Ef-Tu)</u>

Plastid

Arabidopsis thaliana	CEFTU _ARATH	P17745
Oryza sativa subsp. japonica	CEFTU _ORYSJ	Q6ZI53
Picea sitchensis	CEFTU _PICSI	C0PQG8
Selaginella moellendorffii	CEFTU _SELML	D8T8L9
Physcomitrella patens subsp. patens	CEFTU _PHYPA	A9T0S0
Micromonas sp. strain RCC299	CEFTU _MICSR	C1KR64
Chlamydomonas reinhardtii	CEFTU _CHLRE	P17746
Cyanidioschyzon merolae	CEFTU _CYAME	Q85FT7
Cyanophora paradoxa	CEFTU_CYAPA	P17245
Ectocarpus siliculosus	CEFTU _ECTSI	D1J725
Thalassiosira pseudonana	CEFTU _THAPS	A0T100
Phaeodactylum tricornutum	CEFTU _PHATC	A0T0K6

Mitochondrial

Arabidopsis thaliana	MEFTU _ARATH	Q9ZT91
Oryza sativa subsp. japonica	MEFTU _ORYSJ	Q851Y8
Selaginella moellendorffii	MEFTU _SELML	D8S6G0
Physcomitrella patens subsp. patens	MEFTU _PHYPA	A9T9Z0
Micromonas sp. strain RCC299	MEFTU _MICSR	C1E231
Chlamydomonas reinhardtii	MEFTU _CHLRE	A8HXR2
Ectocarpus siliculosus	MEFTU _ECTSI	D8LDT2

Thalassiosira pseudonana	MEFTU _THAPS	B8CA96
Phaeodactylum tricornutum	MEFTU _PHATC	B7GA11

Bacterial

Escherichia coli strain K12	EFTU_ECOLI	P0CE47
Caulobacter crescentus strain ATCC 19089	EFTU _CAUCR	Q99QM0
Chlorobium limicola strain DSM 245	EFTU _CHLLI	B3EH93
Vibrio cholerae strain ATCC 39315	EFTU _VIBCH	Q9KV37
Rickettsia typhi strain ATCC VR-144	EFTU _RICTY	Q8KT95
Rickettsia prowazekii strain Madrid E	EFTU _RICPR	P48865
Agrobacterium tumefaciens strain C58	EFTU _AGRTU	Q8UE16
Synechococcus sp. strain WH8102	EFTU _SYNPX	Q7U4D1
Aquifex aeolicus strain VF5	EFTU _AQUAE	O66429
Thermotoga maritima strain ATCC 43589	EFTU _THEMA	P13537
Prochlorococcus marinus subsp. pastoris strain CCMP1986	EFTU _PROMP	Q7UZY7
Synechocystis sp. strain PCC 6803	EFTU _SYNCY	P74227
Nostoc sp. strain PCC 7120	EFTU _ANASP	Q8YP63

Elongation factor 1 α (Ef-1 α)

Sulfolobus tokodaii strain DSM 16993	EF1A_SULTO	Q976B1
Hyperthermus butylicus strain DSM 5456	EF1A_HYPBU	A2BN41
Thermotoga neapolitana strain ATCC 49049	EF1A _THENN	B9K884
Cenarchaeum symbiosum strain A	EF1A_CENSY	A0RUM4
Clostridium tetani strain Massachusetts / E88	EF1A _CLOTE	Q877L9
Deinococcus radiodurans strain ATCC 13939	EF1A _DEIRA	Q9R342
Archaeoglobus fulgidus strain ATCC 49558	EF1A_ARCFU	O29325
Thermococcus sibiricus strain MM 739	EF1A_THESM	C6A4R7
Arabidopsis thaliana	EF1A _ARATH	Q8GTY0
Oryza sativa subsp. japonica	EF1A _ORYSJ	O64937
Picea sitchensis	EF1A _PICSI	C0PSF0
Selaginella moellendorffii	EF1A _SELML	D8RAR5
Physcomitrella patens subsp. patens	EF1A _PHYPA	A9SJB4
Cyanophora paradoxa	EF1A _CYAPA	Q9ZSW2
Cyanidioschyzon merolae	EF1A _CYAME	Q84KQ1
Phaeodactylum tricornutum	EF1A _PHATC	B5Y4J2
Caenorhabditis elegans	EF1A _CAEEL	P53013
Drosophila melanogaster	EF1A _DROME	P08736
Anopheles gambiae	EF1A _ANOGA	Q7PT29
Homo sapiens	EF1A _HUMAN	P68104
Gallus gallus	EF1A _CHICK	Q90835
Monosiga ovate	EF1A _MONOV	Q2TTF7

*Accessions correspond to Cyanidioschyzon merolae genome sequencing project, http://merolae.biol.s.u-tokyo.ac.jp (Matsuzaki et al, 2004 Nature)

Table S6. Divergence-time calibration points used in this study

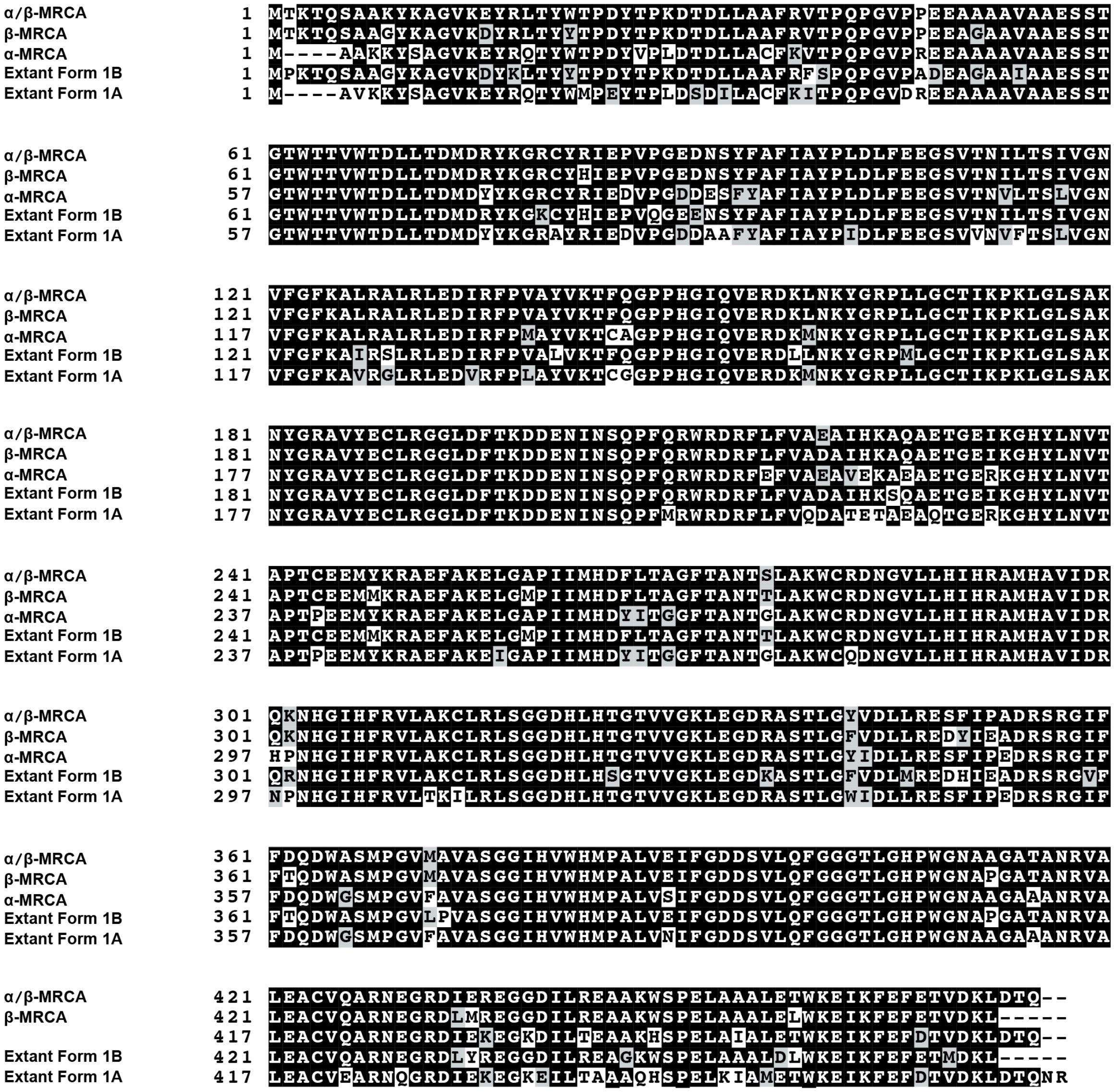
Taxon	Constraint in Mya (±std dev)
Monocotyledoneae†	156(±14).
Angiospermae†	217(±40)
Gymnospermatophyta†	327(±30)
Tracheophyta†	432(±30).
Land Plants†	477(±70)
Human/Chicken§	300(±30)
Fly/Mosquito§	235(±24)

†Adapted from SA Smith et al (2009)
§Adapted from ML Berbee and JW Taylor (2010)

Appendix C

Supplemental Information for Chapter 4

Supplemental Figures:



```
α/β-MRCA            1  MTKTQSAAKYKAGVKEYRLTYWTPDYTPKDTDLLAAFRVTPQPGVPPEEAAAAVAAESST
β-MRCA              1  MTKTQSAAGYKAGVKDYRLTYYTPDYTPKDTDLLAAFRVTPQPGVPPEEAGAAVAAESST
α-MRCA              1  M----AAKKYSAGVKEYRQTYWTPDYVPLDTDLLACFKVTPQPGVPREEAAAAVAAESST
Extant Form 1B      1  MPKTQSAAGYKAGVKDYKLTYYTPDYTPKDTDLLAAFRFSPQPGVPADEAGAAIAAESST
Extant Form 1A      1  M----AVKKYSAGVKEYRQTYWMPEYTPLDSDILACFKITPQPGVDREEAAAAVAAESST

α/β-MRCA           61  GTWTTVWTDLLTDMDRYKGRCYRIEPVPGEDNSYFAFIAYPLDLFEEGSVTNILTSIVGN
β-MRCA             61  GTWTTVWTDLLTDMDRYKGRCYHIEPVPGEDNSYFAFIAYPLDLFEEGSVTNILTSIVGN
α-MRCA             57  GTWTTVWTDLLTDMDYYKGRCYRIEDVPGDDESFYAFIAYPLDLFEEGSVTNVLTSLVGN
Extant Form 1B     61  GTWTTVWTDLLTDMDRYKGKCYHIEPVQGEENSYFAFIAYPLDLFEEGSVTNILTSIVGN
Extant Form 1A     57  GTWTTVWTDLLTDMDYYKGRAYRIEDVPGDDAAFYAFIAYPIDLFEEGSVVNVFTSLVGN

α/β-MRCA          121  VFGFKALRALRLEDIRFPVAYVKTFQGPPHGIQVERDKLNKYGRPLLGCTIKPKLGLSAK
β-MRCA            121  VFGFKALRALRLEDIRFPVAYVKTFQGPPHGIQVERDKLNKYGRPLLGCTIKPKLGLSAK
α-MRCA            117  VFGFKALRALRLEDIRFPMAYVKTCAGPPHGIQVERDKMNKYGRPLLGCTIKPKLGLSAK
Extant Form 1B    121  VFGFKAIRSLRLEDIRFPVALVKTFQGPPHGIQVERDLLNKYGRPMLGCTIKPKLGLSAK
Extant Form 1A    117  VFGFKAVRGLRLEDVRFPLAYVKTCGGPPHGIQVERDKMNKYGRPLLGCTIKPKLGLSAK

α/β-MRCA          181  NYGRAVYECLRGGLDFTKDDENINSQPFQRWRDRFLFVAEAIHKAQAETGEIKGHYLNVT
β-MRCA            181  NYGRAVYECLRGGLDFTKDDENINSQPFQRWRDRFLFVADAIHKAQAETGEIKGHYLNVT
α-MRCA            177  NYGRAVYECLRGGLDFTKDDENINSQPFQRWRDRFEFVAEAVEKAEAETGERKGHYLNVT
Extant Form 1B    181  NYGRAVYECLRGGLDFTKDDENINSQPFQRWRDRFLFVADAIHKSQAETGEIKGHYLNVT
Extant Form 1A    177  NYGRAVYECLRGGLDFTKDDENINSQPFMRWRDRFLFVQDATETAEAQTGERKGHYLNVT

α/β-MRCA          241  APTCEEMYKRAEFAKELGAPIIMHDFLTAGFTANTSLAKWCRDNGVLLHIHRAMHAVIDR
β-MRCA            241  APTCEEMMKRAEFAKELGMPIIMHDFLTAGFTANTTLAKWCRDNGVLLHIHRAMHAVIDR
α-MRCA            237  APTPEEMYKRAEFAKELGAPIIMHDYITGGFTANTGLAKWCRDNGVLLHIHRAMHAVIDR
Extant Form 1B    241  APTCEEMMKRAEFAKELGMPIIMHDFLTAGFTANTTLAKWCRDNGVLLHIHRAMHAVIDR
Extant Form 1A    237  APTPEEMYKRAEFAKEIGAPIIMHDYITGGFTANTGLAKWCQDNGVLLHIHRAMHAVIDR

α/β-MRCA          301  QKNHGIHFRVLAKCLRLSGGDHLHTGTVVGKLEGDRASTLGYVDLLRESFIPADRSRGIF
β-MRCA            301  QKNHGIHFRVLAKCLRLSGGDHLHTGTVVGKLEGDRASTLGFVDLLREDYIEADRSRGIF
α-MRCA            297  HPNHGIHFRVLAKCLRLSGGDHLHTGTVVGKLEGDRASTLGYIDLLRESFIPEDRSRGIF
Extant Form 1B    301  QRNHGIHFRVLAKCLRLSGGDHLHSGTVVGKLEGDKASTLGFVDLMREDHIEADRSRGVF
Extant Form 1A    297  NPNHGIHFRVLTKILRLSGGDHLHTGTVVGKLEGDRASTLGWIDLLRESFIPEDRSRGIF

α/β-MRCA          361  FDQDWASMPGVMAVASGGIHVWHMPALVEIFGDDSVLQFGGGTLGHPWGNAAGATANRVA
β-MRCA            361  FTQDWASMPGVMAVASGGIHVWHMPALVEIFGDDSVLQFGGGTLGHPWGNAPGATANRVA
α-MRCA            357  FDQDWGSMPGVFAVASGGIHVWHMPALVSIFGDDSVLQFGGGTLGHPWGNAAGAAANRVA
Extant Form 1B    361  FTQDWASMPGVLPVASGGIHVWHMPALVEIFGDDSVLQFGGGTLGHPWGNAPGATANRVA
Extant Form 1A    357  FDQDWGSMPGVFAVASGGIHVWHMPALVNIFGDDSVLQFGGGTLGHPWGNAAGAAANRVA

α/β-MRCA          421  LEACVQARNEGRDIEREGGDILREAAKWSPELAAALETWKEIKFEFETVDKLDTQ--
β-MRCA            421  LEACVQARNEGRDLMREGGDILREAAKWSPELAAALELWKEIKFEFETVDKL-----
                  417  LEACVQARNEGRDIEKEGKDILTEAAKHSPELAIALETWKEIKFEFDTVDKLDTQ--
Extant Form 1B    421  LEACVQARNEGRDLYREGGDILREAGKWSPELAAALDLWKEIKFEFETMDKL-----
Extant Form 1A    417  LEACVEARNQGRDIEKEGKEILTAAAQHSPELKIAMETWKEIKFEFDTVDKLDTQNR
```

Figure S1. Alignment of ancestral RbcL sequences and extant Form 1A and Form 1B counterparts. Extant Form 1B sequence from *Synechococcus elongatus* PCC6301. Extant Form 1A sequence from *Halothiobacillus neapolitanus*. Black boxes indicate fully conserved residues. White and grey boxes denote differences in sequences.

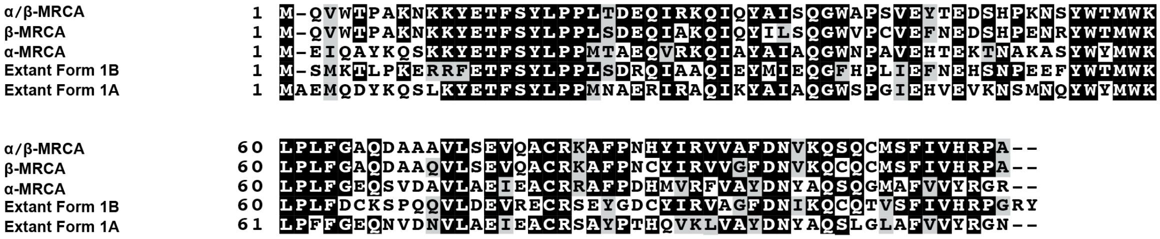

α/β-MRCA 1 M-QVWTPAKNKKYETFSYLPPLTDEQIRKQIQYAISQGWAPSVEYTEDSHPKNSYWTMWK
β-MRCA 1 M-QVWTPAKNKKYETFSYLPPLSDEQIAKQIQYILSQGWVPCVEFNEDSHPENRYWTMWK
α-MRCA 1 M-EIQAYKQSKKYETFSYLPPMTAEQVRKQIAYAIAQGWNPAVEHTEKTNAKASYWYMWK
Extant Form 1B 1 M-SMKTLPKERRFETFSYLPPLSDRQIAAQIEYMIEQGFHPLIEFNEHSNPEEFYWTMWK
Extant Form 1A 1 MAEMQDYKQSLKYETFSYLPPMNAERIRAQIKYAIAQGWSPGIEHVEVKNSMNQYWYMWK
α/β-MRCA 60 LPLFGAQDAAAVLSEVQACRKAFPNHYIRVVAFDNVKQSQCMSFIVHRPA--
β-MRCA 60 LPLFGAQDAAQVLSEVQACRKAFPNCYIRVVGFDNVKQCQCMSFIVHRPA--
α-MRCA 60 LPLFGEQSVDAVLAEIEACRRAFPDHMVRFVAYDNYAQSQGMAFVVYRGR--
Extant Form 1B 60 LPLFDCKSPQQVLDEVRECRSEYGDCYIRVAGFDNIKQCQTVSFIVHRPGRY
Extant Form 1A 61 LPFFGEQNVDNVLAEIEACRSAYPTHQVKLVAYDNYAQSLGLAFVVYRGN--

Figure S2. Alignment of ancestral RbcL sequences and extant Form 1A and Form 1B counterparts. Extant Form 1B sequence from *Synechococcus elongatus* PCC6301. Extant Form 1A sequence from *Halothiobacillus neapolitanus*.

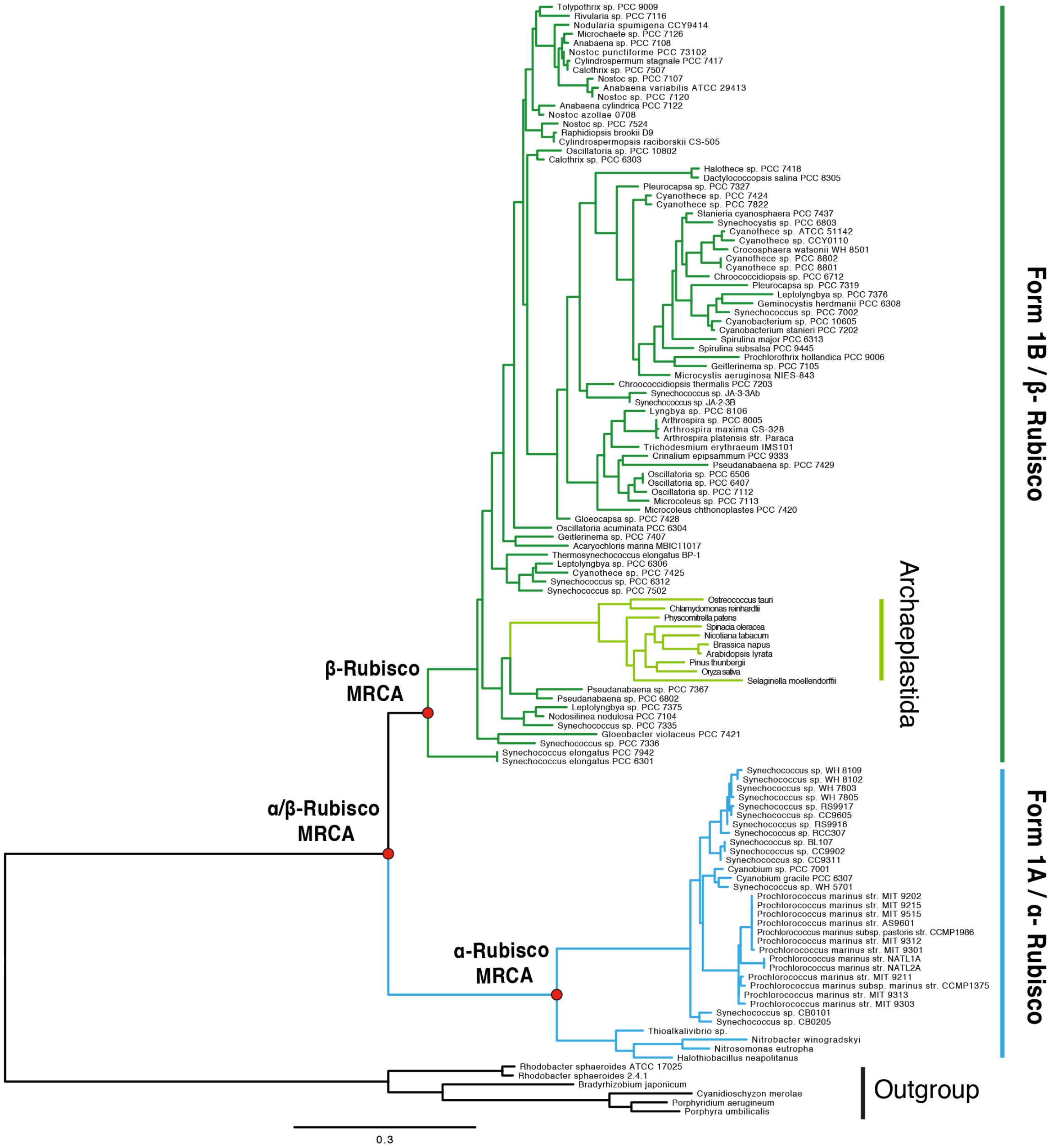

Tolypothrix sp. PCC 9009
Rivularia sp. PCC 7116
Nodularia spumigena CCY9414
Microchaete sp. PCC 7126
Anabaena sp. PCC 7108
Nostoc punctiforme PCC 73102
Cylindrospermum stagnale PCC 7417
Calothrix sp. PCC 7507
Nostoc sp. PCC 7107
Anabaena variabilis ATCC 29413
Nostoc sp. PCC 7120
Anabaena cylindrica PCC 7122
Nostoc azollae 0708
Nostoc sp. PCC 7524
Raphidiopsis brookii D9
Cylindrospermopsis raciborskii CS-505
Oscillatoria sp. PCC 10802
Calothrix sp. PCC 6303
Halothece sp. PCC 7418
Dactylococcopsis salina PCC 8305
Pleurocapsa sp. PCC 7327
Cyanothece sp. PCC 7424
Cyanothece sp. PCC 7822
Stanieria cyanosphaera PCC 7437
Synechocystis sp. PCC 6803
Cyanothece sp. ATCC 51142
Cyanothece sp. CCY0110
Crocosphaera watsonii WH 8501
Cyanothece sp. PCC 8802
Cyanothece sp. PCC 8801
Chroococcidiopsis sp. PCC 6712
Pleurocapsa sp. PCC 7319
Leptolyngbya sp. PCC 7376
Geminocystis herdmanii PCC 6308
Synechococcus sp. PCC 7002
Cyanobacterium sp. PCC 10605
Cyanobacterium stanieri PCC 7202
Spirulina major PCC 6313
Spirulina subsalsa PCC 9445
Prochlorothrix hollandica PCC 9006
Geitlerinema sp. PCC 7105
Microcystis aeruginosa NIES-843
Chroococcidiopsis thermalis PCC 7203
Synechococcus sp. JA-3-3Ab
Synechococcus sp. JA-2-3B
Lyngbya sp. PCC 8106
Arthrospira sp. PCC 8005
Arthrospira maxima CS-328
Arthrospira platensis str. Paraca
Trichodesmium erythraeum IMS101
Crinalium epipsammum PCC 9333
Pseudanabaena sp. PCC 7429
Oscillatoria sp. PCC 6506
Oscillatoria sp. PCC 6407
Oscillatoria sp. PCC 7112
Microcoleus sp. PCC 7113
Microcoleus chthonoplastes PCC 7420
Gloeocapsa sp. PCC 7428
Oscillatoria acuminata PCC 6304
Geitlerinema sp. PCC 7407
Acaryochloris marina MBIC11017
Thermosynechococcus elongatus BP-1
Leptolyngbya sp. PCC 6306
Cyanothece sp. PCC 7425
Synechococcus sp. PCC 6312
Synechococcus sp. PCC 7502
Ostreococcus tauri
Chlamydomonas reinhardtii
Physcomitrella patens
Spinacia oleracea
Nicotiana tabacum
Brassica napus
Arabidopsis lyrata
Pinus thunbergii
Oryza sativa
Selaginella moellendorffii
Pseudanabaena sp. PCC 7367
Pseudanabaena sp. PCC 6802
Leptolyngbya sp. PCC 7375
Nodosilinea nodulosa PCC 7104
Synechococcus sp. PCC 7335
Gloeobacter violaceus PCC 7421
Synechococcus sp. PCC 7336
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301
Synechococcus sp. WH 8109
Synechococcus sp. WH 8102
Synechococcus sp. WH 7803
Synechococcus sp. WH 7805
Synechococcus sp. RS9917
Synechococcus sp. CC9605
Synechococcus sp. RS9916
Synechococcus sp. RCC307
Synechococcus sp. BL107
Synechococcus sp. CC9902
Synechococcus sp. CC9311
Cyanobium sp. PCC 7001
Cyanobium gracile PCC 6307
Synechococcus sp. WH 5701
Prochlorococcus marinus str. MIT 9202
Prochlorococcus marinus str. MIT 9215
Prochlorococcus marinus str. MIT 9515
Prochlorococcus marinus str. AS9601
Prochlorococcus marinus subsp. pastoris str. CCMP1986
Prochlorococcus marinus str. MIT 9312
Prochlorococcus marinus str. MIT 9301
Prochlorococcus marinus str. NATL1A
Prochlorococcus marinus str. NATL2A
Prochlorococcus marinus str. MIT 9211
Prochlorococcus marinus subsp. marinus str. CCMP1375
Prochlorococcus marinus str. MIT 9313
Prochlorococcus marinus str. MIT 9303
Synechococcus sp. CB0101
Synechococcus sp. CB0205
Thioalkalivibrio sp.
Nitrobacter winogradskyi
Nitrosomonas eutropha
Halothiobacillus neapolitanus
Rhodobacter sphaeroides ATCC 17025
Rhodobacter sphaeroides 2.4.1
Bradyrhizobium japonicum
Cyanidioschyzon merolae
Porphyridium aerugineum
Porphyra umbilicalis
β-Rubisco MRCA
α/β-Rubisco MRCA
α-Rubisco MRCA
Form 1B / β- Rubisco
Archaeplastida
Form 1A / α- Rubisco
Outgroup
0.3

Figure S3: Maximum-likelihood phylogeny of RbcL.

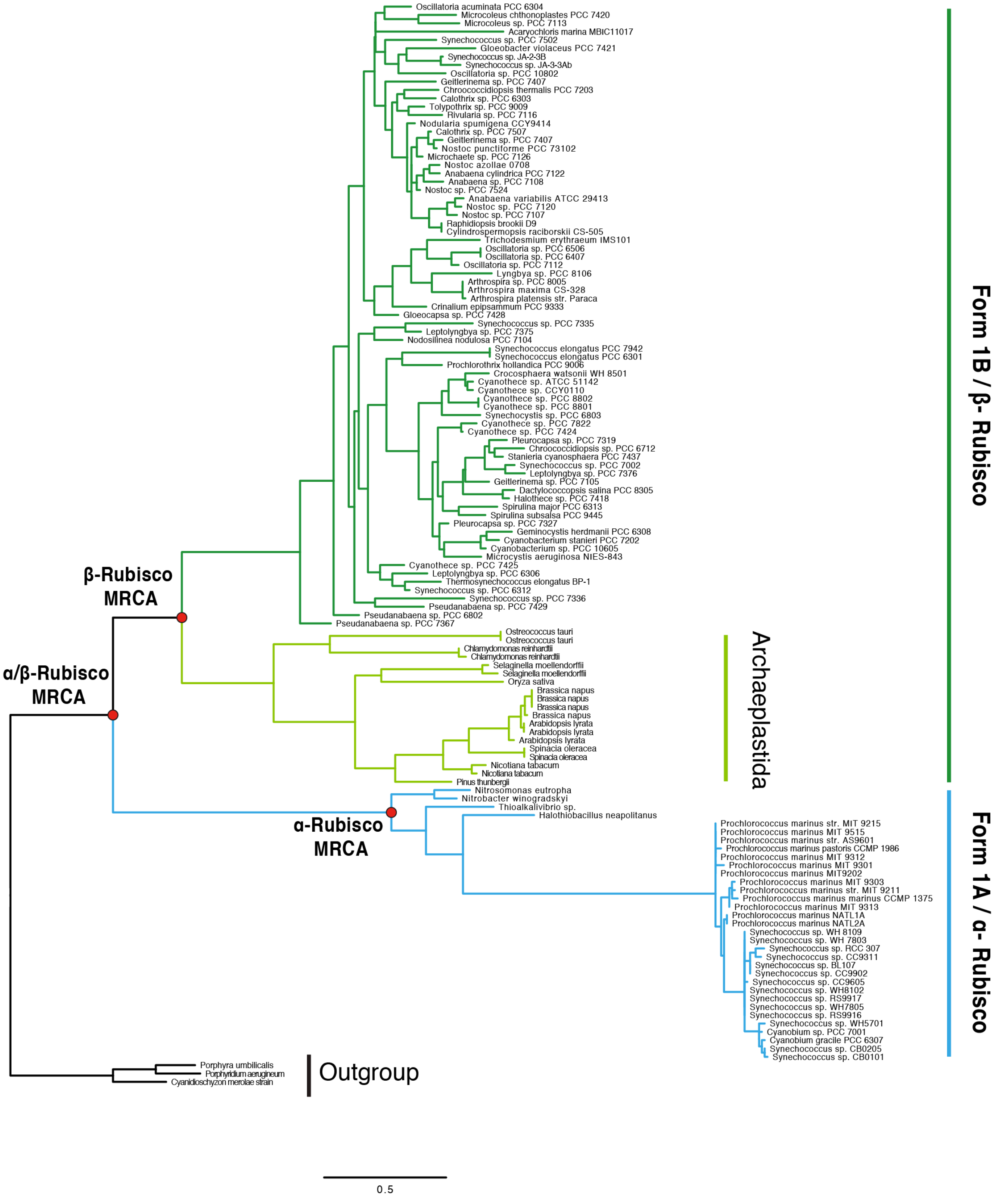

Oscillatoria acuminata PCC 6304
Microcoleus chthonoplastes PCC 7420
Microcoleus sp. PCC 7113
Acaryochloris marina MBIC11017
Synechococcus sp. PCC 7502
Gloeobacter violaceus PCC 7421
Synechococcus sp. JA-2-3B
Synechococcus sp. JA-3-3Ab
Oscillatoria sp. PCC 10802
Geitlerinema sp. PCC 7407
Chroococcidiopsis thermalis PCC 7203
Calothrix sp. PCC 6303
Tolypothrix sp. PCC 9009
Rivularia sp. PCC 7116
Nodularia spumigena CCY9414
Calothrix sp. PCC 7507
Geitlerinema sp. PCC 7407
Nostoc punctiforme PCC 73102
Microchaete sp. PCC 7126
Nostoc azollae 0708
Anabaena cylindrica PCC 7122
Anabaena sp. PCC 7108
Nostoc sp. PCC 7524
Anabaena variabilis ATCC 29413
Nostoc sp. PCC 7120
Nostoc sp. PCC 7107
Raphidiopsis brookii D9
Cylindrospermopsis raciborskii CS-505
Trichodesmium erythraeum IMS101
Oscillatoria sp. PCC 6506
Oscillatoria sp. PCC 6407
Oscillatoria sp. PCC 7112
Lyngbya sp. PCC 8106
Arthrospira sp. PCC 8005
Arthrospira maxima CS-328
Arthrospira platensis str. Paraca
Crinalium epipsammum PCC 9333
Gloeocapsa sp. PCC 7428
Synechococcus sp. PCC 7335
Leptolyngbya sp. PCC 7375
Nodosilinea nodulosa PCC 7104
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301
Prochlorothrix hollandica PCC 9006
Crocosphaera watsonii WH 8501
Cyanothece sp. ATCC 51142
Cyanothece sp. CCY0110
Cyanothece sp. PCC 8802
Cyanothece sp. PCC 8801
Synechocystis sp. PCC 6803
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Pleurocapsa sp. PCC 7319
Chroococcidiopsis sp. PCC 6712
Stanieria cyanosphaera PCC 7437
Synechococcus sp. PCC 7002
Leptolyngbya sp. PCC 7376
Geitlerinema sp. PCC 7105
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
Spirulina major PCC 6313
Spirulina subsalsa PCC 9445
Pleurocapsa sp. PCC 7327
Geminocystis herdmanii PCC 6308
Cyanobacterium stanieri PCC 7202
Cyanobacterium sp. PCC 10605
Microcystis aeruginosa NIES-843
Cyanothece sp. PCC 7425
Leptolyngbya sp. PCC 6306
Thermosynechococcus elongatus BP-1
Synechococcus sp. PCC 6312
Synechococcus sp. PCC 7336
Pseudanabaena sp. PCC 7429
Pseudanabaena sp. PCC 6802
Pseudanabaena sp. PCC 7367
Ostreococcus tauri
Ostreococcus tauri
Chlamydomonas reinhardtii
Chlamydomonas reinhardtii
Selaginella moellendorffii
Selaginella moellendorffii
Oryza sativa
Brassica napus
Brassica napus
Brassica napus
Brassica napus
Arabidopsis lyrata
Arabidopsis lyrata
Arabidopsis lyrata
Spinacia oleracea
Spinacia oleracea
Nicotiana tabacum
Nicotiana tabacum
Pinus thunbergii
Nitrosomonas eutropha
Nitrobacter winogradskyi
Thioalkalivibrio sp.
Halothiobacillus neapolitanus
Prochlorococcus marinus str. MIT 9215
Prochlorococcus marinus MIT 9515
Prochlorococcus marinus str. AS9601
Prochlorococcus marinus pastoris CCMP 1986
Prochlorococcus marinus MIT 9312
Prochlorococcus marinus MIT 9301
Prochlorococcus marinus MIT9202
Prochlorococcus marinus MIT 9303
Prochlorococcus marinus str. MIT 9211
Prochlorococcus marinus marinus CCMP 1375
Prochlorococcus marinus MIT 9313
Prochlorococcus marinus NATL1A
Prochlorococcus marinus NATL2A
Synechococcus sp. WH 8109
Synechococcus sp. WH 7803
Synechococcus sp. RCC 307
Synechococcus sp. CC9311
Synechococcus sp. BL107
Synechococcus sp. CC9902
Synechococcus sp. CC9605
Synechococcus sp. WH8102
Synechococcus sp. RS9917
Synechococcus sp. WH7805
Synechococcus sp. RS9916
Synechococcus sp. WH5701
Cyanobium sp. PCC 7001
Cyanobium gracile PCC 6307
Synechococcus sp. CB0205
Synechococcus sp. CB0101
Porphyra umbilicalis
Porphyridium aeruginosum
Cyanidioschyzon merolae strain
β-Rubisco
MRCA
α/β-Rubisco
MRCA
α-Rubisco
MRCA
Form 1B / β- Rubisco
Archaeplastida
Form 1A / α- Rubisco
Outgroup
0.5

Figure S4: Maximum-likelihood phylogeny of RbcS.

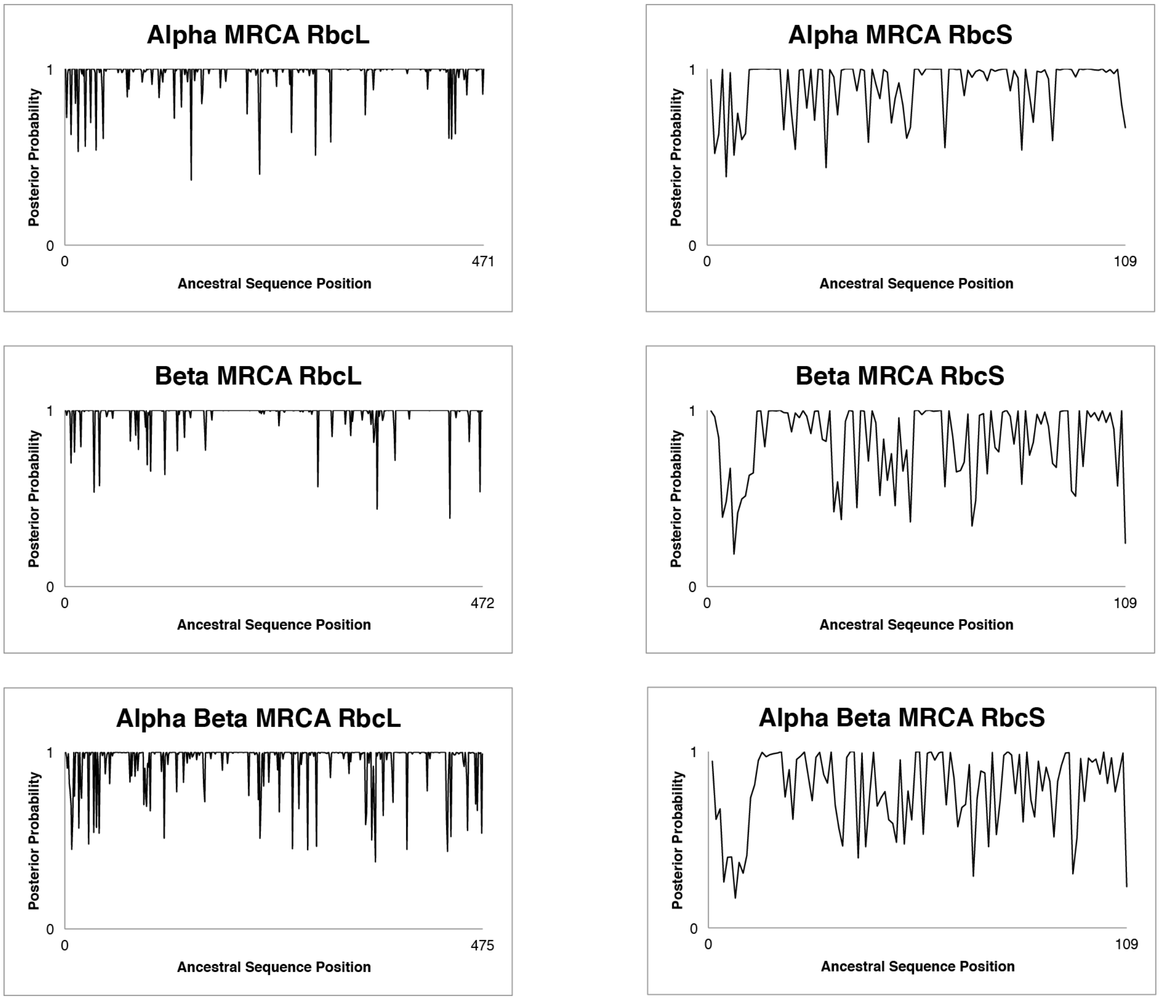


Figure S5. Plot of posterior probabilities of ancestral RbcL and RbcS sequences by sequence position.

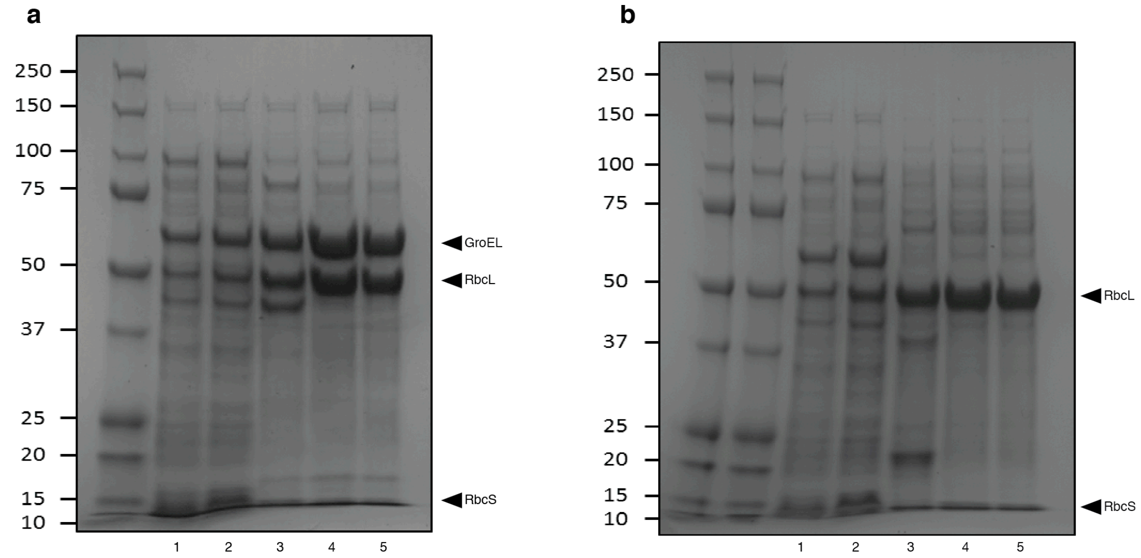


Figure S6: Protein Purification of ancestral α-MRCA and β-MRCA RuBisCO. a. Protein purification of α-MRCA RuBisCO from *E. coli*. Lanes: 1) Crude lysate, 2) PEG precipitation, 3) Anion-Exchange, 4) Size-Exclusion, 5) Ultrafiltation. **b.** Protein purification of β-MRCA RuBisCO from *E. coli*. Lanes: 1) Crude lysate, 2) PEG precipitation, 3) Anion-Exchange, 4) Size-Exclusion, 5) Ultrafiltation.

Table S1. Kinetic parameters of various extant, mutated, chimeric, and ancestral RuBisCO.

	K_c (μM)	V_c (s^{-1})	Specificity	K_o (μM)	Reference
Form 1 A Rubisco					
Chromatium vinosum	37 ±2	6.7 ±0.4	41 ±1	290 ±25	(1)
Form 1B Cyanobacteria					
Synechococcus sp. PCC7002	246 ±20	13.4 ±0.4	52 ±2	1300 ±130	(1)
Synechococcus sp. PCC6301	340 ±12	11.6 ±0.4	43 ±1	972 ±26	(1)
Form 1B Green Algae					
Chlamydomonas reinhardtii	29 ±2	5.8 ±0.2	61 ±5	480 ±58	(1)
Form 1B Non-Green Algae					
Phaeodactylum tricornutum	27.9 ±0.4	3.4 ±0.1	113 ±1	467 ±22	(1)
Galdieria sulfuraria	3.3 ±0.4	1.2 ±0.1	166 ±6	374 ±92	(1)
Griffithsia monilis	9.3 ±0.8	2.6 ±0.1	167 ±3		(1)
Form 1B C4 higher plants					
Zea mays	34 ±2.38	4.4 ±0.22	78 ±3	810 ±97.2	(1)
Amaranthus hybridus	16 ±1.12	3.8 ±0.3	82 ±4	640 ±76	(1)
Flaveria australasica	22 ±4.7	3.84 ±0.03	77.2 ±0.3	309 ±17	(1)
Amaranthus edulis	18.2 ±3.6	4.14 ±0.19	77.5 ±0.2	289 ±9	(1)
Sorghum bicolor	30 ±0.3	5.4 ±0.12	70 ±1		(1)
Potulaca oleraca	13.6 ±0.1	5.9 ±0.44	78 ±4		(1)
Form 1B C3 higher plants					
Triticum aestivum	14 ±3	2.5 ±0.2	90 ±1	730 ±41	(1)
Spinacia oleracea	12.1 ±0.8	3.2 ±0.09	79.8 ±0.5	574 ±19	(1)
Nicotiana tabacum	10.7 ±0.6	3.4 ±0.1	82 ±2	295 ±71	(1)
Flaveria pringlei	12 ±2.1	3.11 ±0.2	80.8 ±1.2	666 ±28	(1)
Chenopodium alba	11.2 ±2.8	2.91 ±0.07	78.7 ±1	415 ±16	(1)
Ancestral RuBisCO					
Ancestral Form 1A	113 ±6	2.65 ±0.04	54.7 ±3.6	2329 ±208	this work
Ancestral Form 1B	120 ±20	3.05 ±0.08	49.6 ±1.8	641 ±49	this work
Mutated/Chimeric RuBisCO					
6301 L euk S hybrid, pVTAC223	85 ±4	0.13	64.9 ±4.6	1878 ±207	(2)
6301 L euk S hybrid, pANOLI	179 ±6	0.05	37.4 ±0.7	1437 ±173	(2)
6301 T342I	169 ±31	1.95	31.9 ±1.2	446 ±66	(3)
6301 T342V	111 ±13	1.83	29.8 ±2.2	364 ±44	(3)
6301 K339P	264 ±46	0.54	38 ±0.4	721 ±105	(3)
6301 A340L	185 ±12	2.34	36.4 ±1	629 ±60	(3)
6301 S341M	155 ±1	3.47	42 ±0.9	1428 ±305	(3)
Anacystis Mutant K128R		2.9 ±0.3	42.5 ±1.4		(4)
Anacystis Mutant K128G		0.15 ±0.01	38.3 ±1.7		(4)
Anacystis Mutant K128Q		0.23 ±0.05	6.7 ±0.7		(4)
6301 T65S	620	1.82	36.6 ±0.5		(5)
6301 T65A	484	0.3	21.6 ±0.2		(5)
6301 T65V	633	0.091	18.5 ±0.2		(5)

Table S2. Strains and plasmids used in this study.

Construct	Host	Description	Reference
pAM1573	*Synechococcus*	Neutral Site 2 genomic integration vector	(6)
pAM1573PMS	*Synechococcus*	BglBrick modified pAM1573 vector	this work
pAM2314	*Synechococcus*	Neutral site 1 genomic integration vector	(6)
pAM2314PMS	*Synechococcus*	BglBrick modified pAM2314 vector	this work
PMS4622	*Synechococcus*	*P_rplC*::AncRbcL_α/βMRCA::CFP cloned into pAM1573PMS	this work
PMS4623	*Synechococcus*	*P_rplC*::AncRbcL_βMRCA::CFP cloned into pAM1573PMS	this work
PMS4624	*Synechococcus*	*P_rplC*::AncRbcL_αMRCA::CFP cloned into pAM1573PMS	this work
JC178	*Synechococcus*	*P_ccmK2*::CcmN::YFP cloned into pAM2314PMS	this work
pET11a	*E. coli*	IPTG inducible expression vector	Novagen
pET11a-AncBetaRbc	*E. coli*	IPTG inducible Ancestral β MRCA RbcL and RbcS	this work
pET11a-AncAlphaRbc	*E. coli*	IPTG inducible Ancestral α MRCA RbcL and RbcS	this work
pBAD33*ES/EL*	*E. coli*	Arabinose inducible GroEL/ES chaperone proteins	(7)

References:

1. Savir Y, Noor E, Milo R, & Tlusty T (2010) Cross-species analysis traces adaptation of Rubisco toward optimality in a low-dimensional landscape. *Proc Natl Acad Sci* 107(8):3475-3480.
2. Read BA & Tabita FR (1992) A hybrid ribulose bisphosphate carboxylase/oxygenase enzyme exhibiting a substantial increase in substrate specificity factor. *Biochemistry* 31(24):5553-5560.
3. Read BA & Tabita FR (1994) High Substrate Specificity Factor Ribulose Bisphosphate Carboxylase/Oxygenase from Eukaryotic Marine Algae and Properties of Recombinant Cyanobacterial Rubisco Containing "Algal" Residue Modifications. *Arch Biochem Biophys* 312(1):210-218.
4. Bainbridge G*, et al.* (1995) Engineering Rubisco to change its catalytic properties. *J Exp Bot* 46(special issue):1269-1276.
5. Morell MK, Paul K, O'Shea NJ, Kane HJ, & Andrews TJ (1994) Mutations of an active site threonyl residue promote beta elimination and other side reactions of the enediol intermediate of the ribulosebisphosphate carboxylase reaction. *J Biol Chem* 269(11):8091-8098.
6. Mackey SR, Ditty JL, Clerico EM, & Golden SS (2007) Detection of Rhythmic Bioluminescence From Luciferase Reporters in Cyanobacteria. *Circadian Rhythms,* Methods in Molecular Biology™, ed Rosato E (Humana Press), Vol 362, pp 115-129.
7. Saschenbrecker S*, et al.* (2007) Structure and Function of RbcX, and Assembly Chaperone for Hexadecameric Rubisco. *Cell* 129(6):1189-1200.