An Early-Branching Freshwater Cyanobacterium at the Origin of Plastids

Rafael I Ponce-Toledo, Philippe Deschamps, Purificación López-García, Yvan Zivanovic, Karim Benzerara, David Moreira

An early-branching freshwater cyanobacterium at the origin of chloroplasts

Rafael I. Ponce-Toledo,[1] Philippe Deschamps,[1] Purificación López-García,[1] Yvan Zivanovic,[2] Karim Benzerara,[3] David Moreira[1]*

[1]Unité d'Ecologie, Systématique et Evolution, CNRS UMR 8079, Université Paris-Sud/Paris-Saclay, AgroParisTech, 91400 Orsay, France. [2]Institut de Génétique et Microbiologie, CNRS UMR 8621, Université Paris-Sud/Paris-Saclay, 91405 Orsay, France. [3]Institut de Minéralogie, de Physique des Matériaux, et de Cosmochimie, Sorbonne Universités - UPMC Univ Paris 06, CNRS UMR 7590, MNHN, IRD UMR 206 Paris, France

***Corresponding author. Email: david.moreira@u-psud.fr**

Eukaryotic photosynthesis evolved from the endosymbiosis of a cyanobacterium, the future chloroplast, within a heterotrophic host. However, while the endosymbiotic origin of chloroplasts from a single cyanobacterial ancestor is firmly established, the nature of that ancestor remains controversial: chloroplasts have been proposed to derive from either early- or late-branching cyanobacterial lineages. We carried out phylogenomic and super-network analyses of a comprehensive dataset of chloroplast-encoded proteins, including genomes from novel cyanobacteria. These strongly support that chloroplasts evolved from deep-branching cyanobacteria related to a recently discovered clade widespread in freshwater microbialites and microbial mats, from which the only cultured representative is *Gloeomargarita lithophora*. Therefore, chloroplasts not only emerged early within cyanobacteria, but the first photosynthetic eukaryotes likely evolved in terrestrial-freshwater settings, not in oceans as commonly thought.

That symbiosis had played a crucial role in the origin of photosynthetic eukaryotes was first argued by the botanist Konstantin Mereschkowky in 1905 (*1*). He proposed that modern chloroplasts were once free-living cyanobacteria that established a symbiotic relationship with a heterotrophic eukaryotic host. Molecular support for the cyanobacterial origin of chloroplasts

based on DNA and protein phylogenetic analyses came much later at the end of the 1970s (*2, 3*). It is now widely accepted that a heterotrophic eukaryote engulfed a cyanobacterium and that this endosymbiotic relationship resulted in the first photosynthetic eukaryote, which evolved and diversified to originate the present-day supergroup Archaeplastida (*4*). Archaeplastida group three lineages, all of them possessing double-membrane chloroplasts: Glaucophyta, Rhodophyta (red algae), and Viridiplantae (green algae and land plants). Later, green and red algae were independently recruited via secondary endosymbiosis by diverse heterotrophic eukaryotes, giving rise to a vast variety of photosynthetic lineages (e.g., chlorarachniophytes, euglenids, diatoms, dinoflagellates) (*5, 6*). The monophyly of Archaeplastida is widely accepted, as both nuclear and plastid gene phylogenies point to a single endosymbiotic event at the origin of this supergroup(*7, 8*). However, the identity of the cyanobacterial and eukaryotic partners that established this endosymbiotic relationship remains unsettled. Different studies have tried to identify the closest lineage to that of the chloroplast ancestor within the contemporary phylogenetic diversity of cyanobacteria, with contrasting conclusions. Some studies suggest that chloroplasts derive from an early-branching, yet unidentified, cyanobacterial ancestor (*9-12*), whereas others support a late-branching cyanobacterium closely related to contemporary filamentous heterocyst-bearing N_2-fixing species (*13-16*). Determining the phylogenetic position of chloroplasts in the cyanobacterial tree is critical to better understand the metabolic interactions sustaining the initial endosymbiosis, the environmental setting where it occurred, and the early evolution of the first photosynthetic eukaryotes prior to the tripartite diversification of the Archaeplastida.

To address this question, we have carried out phylogenomic analyses upon a comprehensive dataset of conserved chloroplast-encoded proteins and the richest sampling of cyanobacterial genome sequences used to date. During their long endosymbiotic history, many genes necessary for chloroplast function were transferred into the host nuclear genome and others were lost, such that chloroplast genomes are extremely reduced. With sizes between 100-200 kb, they encode between 80 and 200 proteins, whereas free-living cyanobacterial genomes encode between 1,800 and 12,000 proteins (*14*). Remaining genes in organelle genomes are responsible for essential chloroplast functions (e.g., protein translation, photosystem structure) and their sequences are highly conserved. Therefore, despite the relatively small number of chloroplast-encoded proteins, they are good phylogenetic markers because they i) are direct remnants of the cyanobacterial endosymbiont and ii) exhibit remarkable sequence and functional conservation.

To have a broad and balanced representation of primary chloroplasts and cyanobacterial groups, we mined a large sequence database containing all ribosomal RNAs (rRNAs) and proteins encoded in 19 chloroplasts of Archaeplastida (1 glaucophyte, 8 green algae/plants, and 9 red algae) and 101 cyanobacterial genomes, including recently sequenced members (*11, 14, 17*). We excluded the *Synechococcus-Prochlorococcus* (SynPro) clade from our study for two reasons. First, previous analyses never showed a relationship of SynPro with chloroplasts. Second, this group experienced a severe genome reduction and acceleration of evolutionary rate (*18*). This would imply a considerable amount of missing data and the presence of fast-evolving and compositionally-biased sequences that might be problematic for phylogenomic analyses.

Sequence similarity searches in the genome dataset allowed the identification of 97 widespread conserved proteins, after exclusion of those only present in a restricted number of chloroplasts and/or cyanobacteria and those showing evidence of horizontal gene transfer (HGT) among cyanobacterial species (table S1). We analyzed the dataset of 97 conserved proteins by a supermatrix approach (21,942 concatenated amino acid sites) using probabilistic phylogenetic methods with site-heterogeneous substitution models. In addition, we analyzed an rRNA dataset containing chloroplast and cyanobacterial 16S+23S rRNA concatenated sequences. Phylogenetic trees reconstructed for both datasets by maximum likelihood and Bayesian inference provide full support for the early divergence of chloroplasts among the cyanobacterial species (Fig. 1 and fig. S1, respectively). Moreover, our trees show that the closest present-day relative of chloroplasts is the recently described deep-branching cyanobacterium *Gloeomargarita lithophora* (*17, 19, 20*). This biofilm-forming benthic cyanobacterium has attracted attention by its unusual capacity to accumulate intracellular amorphous calcium-magnesium-strontium-barium carbonates. Phylogenetic analysis based on environmental 16S rRNA sequences has shown that *G. lithophora* belongs to a diverse early-branching cyanobacterial lineage, the *Gloeomargaritales* (*17, 20*), for which it is the only species isolated so far. Although initially found in an alkaline crater lake in Mexico, environmental studies have demonstrated that *G. lithophora* and related species have a widespread terrestrial distribution ranging from freshwater alkaline lakes to thermophilic microbial mats (*20*). Interestingly, it has never been observed in marine samples (*20*).

To test the robustness of the phylogenetic relation between *Gloeomargarita* and chloroplasts, we investigated possible biases that might lead to an artefactual placement of chloroplasts in our trees. First, we recoded the amino acid sequences by grouping amino

acids of similar physicochemical characteristics into four families, a procedure which is known to alleviate possible compositional biases (*21*). Phylogenetic trees based on the recoded alignment still retrieve the *Gloeomargarita*-chloroplasts sister relation (fig. S2). It has been shown that fast-evolving sites in chloroplasts have a very poor fit to evolutionary models (*22*). Therefore, we tested whether the *Gloeomargarita*-chloroplasts relation could be due to the accumulation of fast-evolving sites leading to sequence evolution model violation and long-branch attraction (LBA) artefacts. For that, we calculated the evolutionary rate for each of the 21,942 sites of the 97-protein concatenation and divided them into 10 categories, from the slowest- to the fastest-evolving ones. We then reconstructed phylogenetic trees with a progressive inclusion of fast-evolving sites (the so-called Slow-Fast method, aimed at increasing the signal/noise ratio of sequence datasets (*23*)). All the trees show the *Gloeomargarita*-chloroplasts relation with full statistical support (figs. S3-S11). Interestingly, the trees based on the slowest-evolving positions show a remarkable reduction of the branch length of the chloroplast sequences (especially those of green algae and land plants; figs. S9-S11), which become increasingly longer with the addition of fast-evolving positions. This reflects the well-known acceleration of evolutionary rate that chloroplasts have experienced, in particular in Viridiplantae (*24*). These results argue against the possibility that the *Gloeomargarita*-chloroplasts relation arises from an LBA artefact due to the accumulation of noise in fast-evolving sites. Finally, we tested if the supermatrix approach might have generated artefactual results due to the concatenation of markers with potentially incompatible evolutionary histories (because of HGT, hidden paralogy, etc.). We addressed this issue through the application of a phylogenetic network approach, which can cope with those contradictory histories (*25*), to analyze the set of 97 phylogenetic trees reconstructed with the individual proteins. The supernetwork based on those trees confirms again *G. lithophora* as the closest cyanobacterial relative of chloroplasts (Fig. 2), in agreement with the phylogenomic results.

A deep origin of plastids within the cyanobacterial phylogeny was inferred in past studies based on 16S rRNA gene sequences, chloroplast proteins, and nucleus-encoded proteins of chloroplast origin (*9-12*). However, those studies did not retrieve any close relationship between chloroplasts and any extant cyanobacterial lineage, a result that could be attributed either to lack of phylogenetic resolution or to incomplete taxonomic sampling (*10*). Our results, after inclusion of the new species *G. lithophora,* clearly advocate for the latter. Nevertheless, some studies have proposed that chloroplasts emerged from the apical part of

the cyanobacterial tree, being closely related to late-branching filamentous (Nostocales and Stigonomatales) (*13-15*) or unicellular (Chroococcales) (*16*) N_2-fixing cyanobacteria. However, in agreement with our results, it has been shown that these phylogenies can be explained by similarities in G+C content due to convergent nucleotide composition between chloroplasts and late-branching cyanobacteria and that the use of codon recoding techniques suppresses the compositional bias and recovers a deep origin of chloroplasts (*12*).

Studies supporting a late chloroplast origin did not rely on phylogenomic analyses only but also used other methods, such as quantifying the number and sequence similarity of proteins shared between Archaeplastida and different cyanobacterial species. They showed that the heterocyst-forming filamentous cyanobacterial genera *Nostoc* and *Anabaena* appear to possess the largest and most similar set of proteins possibly present in the chloroplast ancestor (*13, 14*). However, these approaches have two important limitations: 1) the number of proteins is highly dependent on the genome size of the different cyanobacteria, which can vary by more than one order of magnitude; and 2) it is well known that sequence similarity is a poor proxy for evolutionary relationships (*26*). Indeed, if we apply a similar procedure but include several chloroplast representatives in addition to cyanobacteria, we observe that individual chloroplasts can have sequences more similar to cyanobacterial homologues than to those of other chloroplasts (figs. S12-S13). This similarity-based approach would thus suggest that chloroplasts are polyphyletic and have multiple origins within cyanobacteria, a clearly artefactual result. Chloroplast proteins, including those that have been transferred to the nucleus (*24*), have evolved much faster than their cyanobacterial counterparts, such that sequences of one particular primary photosynthetic eukaryote can be more similar to cyanobacterial ones than to those of other distantly-related photosynthetic eukaryotes. Phylogenomic analysis is therefore more suitable than crude sequence similarity to study the cyanobacterial origin of chloroplast sequences.

Proponents of the recent origin of primary chloroplasts from N_2-fixing filamentous cyanobacteria suggest that the dearth of biologically available nitrogen during most of the Proterozoic and the ability of these organisms to fix nitrogen played a key role in the early establishment of the endosymbiosis (*13, 14*). However, there is no trace of such past ability to fix nitrogen in modern chloroplasts. Furthermore, this metabolic ability is also widespread in many cyanobacterial lineages, including basal clades, which has led to propose that the cyanobacterial ancestor was able to fix nitrogen (*27*). Therefore, if N_2 fixation did actually play a role in the establishment of the chloroplast, it cannot discriminate in favor of a late- versus

early-branching cyanobacterial endosymbiont as it is widespread in this bacterial clade. The closest relative of chloroplasts, *G. lithophora*, lacks the genes necessary for N_2 fixation, which further argues against the implication of this metabolism in the origin of chloroplasts. Other hypotheses have proposed a symbiotic interaction between the cyanobacterial ancestor of chloroplasts and its eukaryotic host based on the metabolism of storage polysaccharides. In that scenario, the cyanobacterium would have exported ADP-glucose in exchange for the import of reduced nitrogen from the host (*28*).

Although the nature of the metabolic exchanges between the two partners remains to be elucidated, the exclusive distribution of the *Gloeomargarita* lineage in terrestrial habitats (*20*) provides important clues about the ancient environment where a *Gloeomargarita*-like cyanobacterium established the endosymbiotic relationship with the heterotrophic eukaryote at the origin of primary photosynthetic eukaryotes. Like *G. lithophora*, the first cyanobacteria most likely lived in terrestrial/freshwater habitats (*29, 30*), in agreement with the fact that most basal-branching cyanobacterial lineages thrive in terrestrial/freshwater habitats. Colonization of open oceans and diversification of marine planktonic cyanobacteria came later on evolution, mainly during the Neoproterozoic (1000-541 Mya) with consequential effects on ocean and atmosphere oxygenation (*30*). Notwithstanding their large error intervals, molecular clock analyses infer the origin of Archaeplastida during the mid-Proterozoic (*31, 32*), well before the estimated Neoproterozoic cyanobacterial colonization of oceans. Interestingly, Glaucophyta, the first lineage to diverge within the Archaeplastida (Fig. 1) has been exclusively found in freshwater habitats (*33*), which is also the case for the most basal clade of red algae, the Cyanidiales (*34*). Altogether, these data strongly suggest that chloroplasts, and hence the first photosynthetic eukaryotes, arose in a freshwater/terrestrial environment on the early Earth with low atmospheric and marine oxygen concentrations (*35*).

In conclusion, we provide phylogenomic evidence for a deep origin of chloroplasts within the phylogenetic tree of Cyanobacteria, with the *Gloeomargarita* lineage as the closest extant relative of the chloroplast ancestor. The ecological distribution of this cyanobacterial lineage and extant early-branching eukaryotic algae suggests that the first photosynthetic eukaryotes evolved on the continent, probably in freshwater biofilms or microbial mats. Microbial mats are complex communities where physical and genetic interactions between cyanobacteria and heterotrophic eukaryotes may have been facilitated. Our work highlights the importance of environmental exploration to characterize new organisms that can, in turn, be crucial to resolve unsettled evolutionary questions.

REFERENCES AND NOTES

1. C. Mereschkowsky, *Biologisches Centralblatt* **25**, 593 (1905).
2. R. M. Schwartz, M. O. Dayhoff, *Science* **199**, 395 (1978).
3. M. W. Gray, W. F. Doolittle, *Microbiol Rev* **46**, 1 (1982).
4. S. M. Adl *et al.*, *J Eukaryot Microbiol* **52**, 399 (Sep-Oct, 2005).
5. J. M. Archibald, *Curr Biol.* **19**, R81 (2009).
6. P. J. Keeling, *Annu Rev Plant Biol* **64**, 583 (2013).
7. D. Moreira, H. Le Guyader, H. Philippe, *Nature* **405**, 69 (2000).
8. N. Rodríguez-Ezpeleta *et al.*, *Curr Biol* **15**, 1325 (2005).
9. S. Turner, K. M. Pryer, V. P. Miao, J. D. Palmer, *J Eukaryot Microbiol* **46**, 327 (1999).
10. A. Criscuolo, S. Gribaldo, *Mol Biol Evol* **28**, 3019 (2011).
11. P. M. Shih *et al.*, *Proc Natl Acad Sci U S A* **110**, 1053 (2013).
12. B. Li, J. S. Lopes, P. G. Foster, T. M. Embley, C. J. Cox, *Mol Biol Evol* **31**, 1697 (2014).
13. O. Deusch *et al.*, *Mol Biol Evol.* **25**, 748 (2008).
14. T. Dagan *et al.*, *Genome Biol Evol* **5**, 31 (2013).
15. J. A. Ochoa de Alda, R. Esteban, M. L. Diago, J. Houmard, *Nat Commun* **5**, 4937. (2014).
16. L. I. Falcon, S. Magallon, A. Castillo, *Isme J* **4**, 777 (2010).
17. D. Moreira *et al.*, *Int J Syst Evol Microbiol* (In press.).
18. B. Batut, C. Knibbe, G. Marais, V. Daubin, *Nat Rev Microbiol* **12**, 841 (2014).
19. E. Couradeau *et al.*, *Science* **336**, 459 (2012).
20. M. Ragon, K. Benzerara, D. Moreira, R. Tavera, P. Lopez-Garcia, *Front Microbiol* **5**, 331 (2014).
21. N. Rodríguez-Ezpeleta *et al.*, *Syst Biol* **56**, 389 (Jun, 2007).
22. L. Sun *et al.*, *Sci Rep* **6**, 20528. (2016).
23. H. Philippe *et al.*, *Proc R Soc Lond B Biol Sci* **267**, 1213 (2000).
24. P. Deschamps, D. Moreira, *Mol Biol Evol.* **26**, 2745 (2009).
25. D. H. Huson, D. Bryant, *Mol Biol Evol.* **23**, 254 (2006).
26. L. B. Koski, G. B. Golding, *J Mol Evol* **52**, 540 (2001).
27. N. Latysheva, V. L. Junker, W. J. Palmer, G. A. Codd, D. Barker, *Bioinformatics* **28**, 603 (2012).
28. P. Deschamps *et al.*, *Mol Biol Evol.* **25**, 536 (2008).
29. F. U. Battistuzzi, S. B. Hedges, *Mol Biol Evol.* **26**, 335 (2009).
30. P. Sanchez-Baracaldo, *Sci Rep* **5** (2015).
31. L. W. Parfrey, D. J. Lahr, A. H. Knoll, L. A. Katz, *Proc Natl Acad Sci U S A* **108**, 13624 (2011).
32. L. Eme, S. C. Sharpe, M. W. Brown, A. J. Roger, *Cold Spring Harb Perspect Biol* **6**, 8 (2014).
33. L. Kies, B. P. Kremer, in *Handbook of protoctista* L. Margulis, J. O. Corliss, M. Melkonian, D. J. Chapman, Eds. (Jones and Barlett, Boston, MA, 1990) pp. 152-166.
34. H. S. Yoon, K. M. Müller, R. G. Sheath, F. D. Ott, D. Bhattacharya, *J Phycol* **42**, 482 (2006).
35. T. W. Lyons, C. T. Reinhard, N. J. Planavsky, *Nature* **506**, 307 (2014).

ACKNOWLEDGEMENTS

This study was supported by European Research Council grants ProtistWorld (P.L.-G., Grant Agreement no. 322669) and CALCYAN (K.B., Grant Agreement no. 307110) under the European Union's Seventh Framework Program and the RTP CNRS Génomique Environnementale project MetaStrom (D.M.).

SUPPLEMENTARY MATERIALS

Materials and Methods

Figs. S1 to S13

Table S1

References (36-45)

Figure Legends

Fig. 1. The position of chloroplasts in the cyanobacterial phylogeny. This Bayesian phylogenetic tree is based on the concatenation of 97 chloroplast-encoded proteins and their cyanobacterial homologues. Branches supported by posterior probability 1 are labeled with black circles. Maximum likelihood bootstrap value is also indicated for the branch uniting chloroplasts with the cyanobacterium *Gloeomargarita lithophora*. A false-colored scanning electron microscopy image of this cyanobacterium is shown in the center of the tree. Information about the habitat and morphology of the cyanobacterial species is provided. For the complete tree, see Supplementary Fig. S3).

Fig. 2. Supernetwork analysis of chloroplast-encoded proteins and cyanobacterial homologues. This phylogenetic network is based on the maximum likelihood trees of 97 proteins.

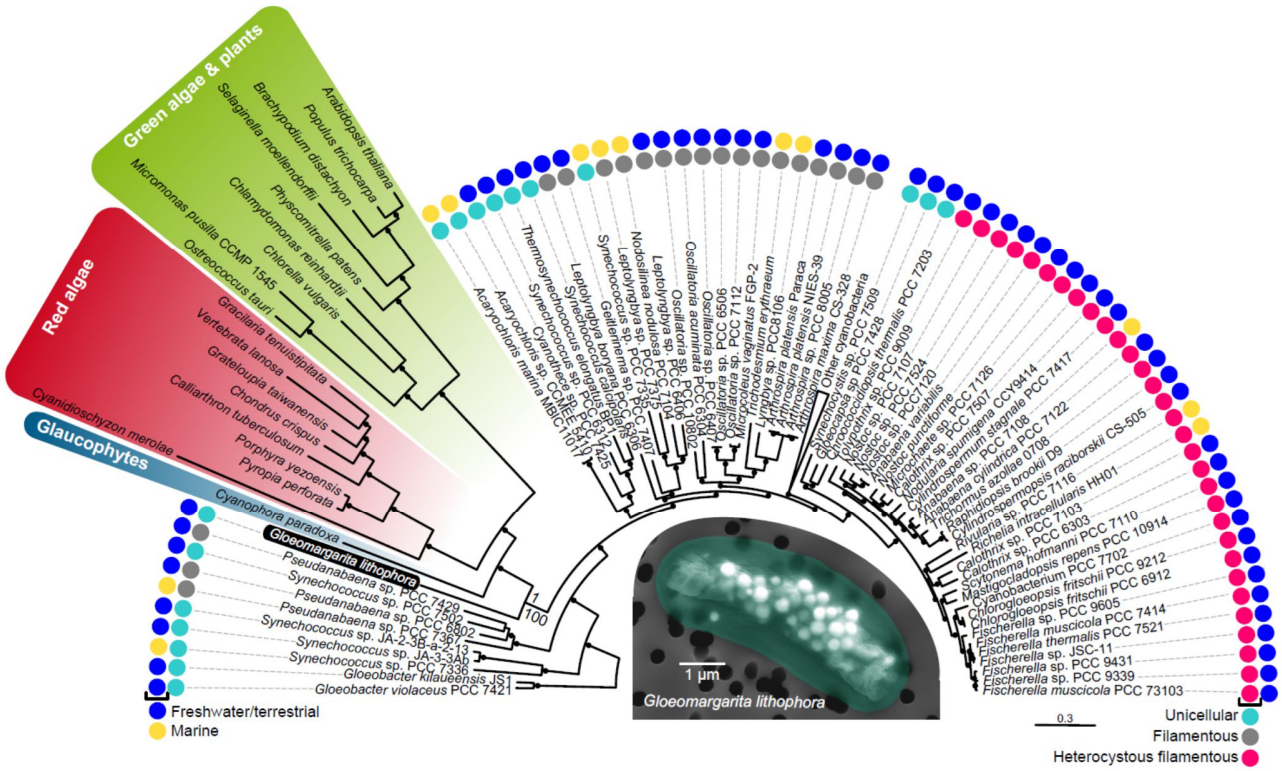
Green algae & plants
Arabidopsis thaliana
Brachypodium distachyon
Populus trichocarpa
Selaginella moellendorffii
Physcomitrella patens
Chlamydomonas reinhardtii
Micromonas pusilla CCMP 1545
Chlorella vulgaris
Ostreococcus tauri
Red algae
Gracilaria tenuistipitata
Vertebrata lanosa
Gelidiocopia taiwanensis
Chondrus crispus
Calliarthron tuberculosum
Porphyra yezoensis
Pyropia perforata
Glaucophytes
Cyanidioschyzon merolae
Cyanophora paradoxa
Gloeomargarita lithophora
Pseudanabaena sp. PCC 7429
Synechococcus sp. PCC 7502
Pseudanabaena sp. PCC 7402
Pseudanabaena sp. PCC 6802
Synechococcus sp. JA-2-3B-a-2(13
Synechococcus sp. JA-3-3Ab
Synechococcus sp. PCC 7336
Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Thermosynechococcus sp. CL-1
Synechococcus sp. CCAP 1479/1
Acaryochloris sp. CCMEE 5410
Acaryochloris marina MBIC10699
Synechococcus sp. PCC 6312
Leptolyngbya sp. PCC 7375
Synechococcus elongatus PCC 6301
Synechococcus elongatus PCC 7942
Gallenema sp. PCC 7407
Leptolyngbya boryana PCC 6306
Leptolyngbya sp. PCC 7376
Nodosilinea sp. PCC 7104
Leptolyngbya sp. PCC 7375
Oscillatoria sp. PCC 6506
Oscillatoria acuminata PCC 6304
Oscillatoria sp. PCC 6407
Oscillatoria sp. PCC 7112
Microcoleus sp. PCC 7113
Trichodesmium erythraeum
Limnoraphis robusta
Arthrospira platensis Paraca
Arthrospira platensis NIES-39
Arthrospira maxima CS-328
Other cyanobacteria
Synechocystis sp. PCC 7509
Chroococcidiopsis thermalis PCC 7203
Gloeocapsa sp. PCC 7428
Oscillatoria sp. PCC 9009
Nostoc sp. PCC 7107
Anabaena variabilis
Nostoc sp. PCC 7524
Microchaete punctiforme
Calothrix sp. PCC 7507
Nodularia spumigena CCY9414
Cylindrospermum stagnale PCC 7417
Anabaena sp. PCC 7108
Trichormus azollae 0708
Raphidiopsis brookii D9
Cylindrospermopsis racborskii CS-505
Rivularia sp. PCC 7116
Richelia intracellularis HH01
Calothrix sp. PCC 6303
Scytonema hofmanni PCC 7110
Mastigocladopsis repens PCC 10914
Cyanobacterium PCC 7702
Chlorogloeopsis fritschii PCC 9212
Chlorogloeopsis fritschii PCC 6912
Fischerella sp. PCC 9605
Fischerella muscicola PCC 7414
Fischerella thermalis PCC 7521
Fischerella sp. JSC-11
Fischerella sp. PCC 9431
Fischerella sp. PCC 9339
Fischerella muscicola PCC 73103
1
100
1 µm
Gloeomargarita lithophora
0.3
Freshwater/terrestrial
Marine
Unicellular
Filamentous
Heterocystous filamentous

Figure 1.

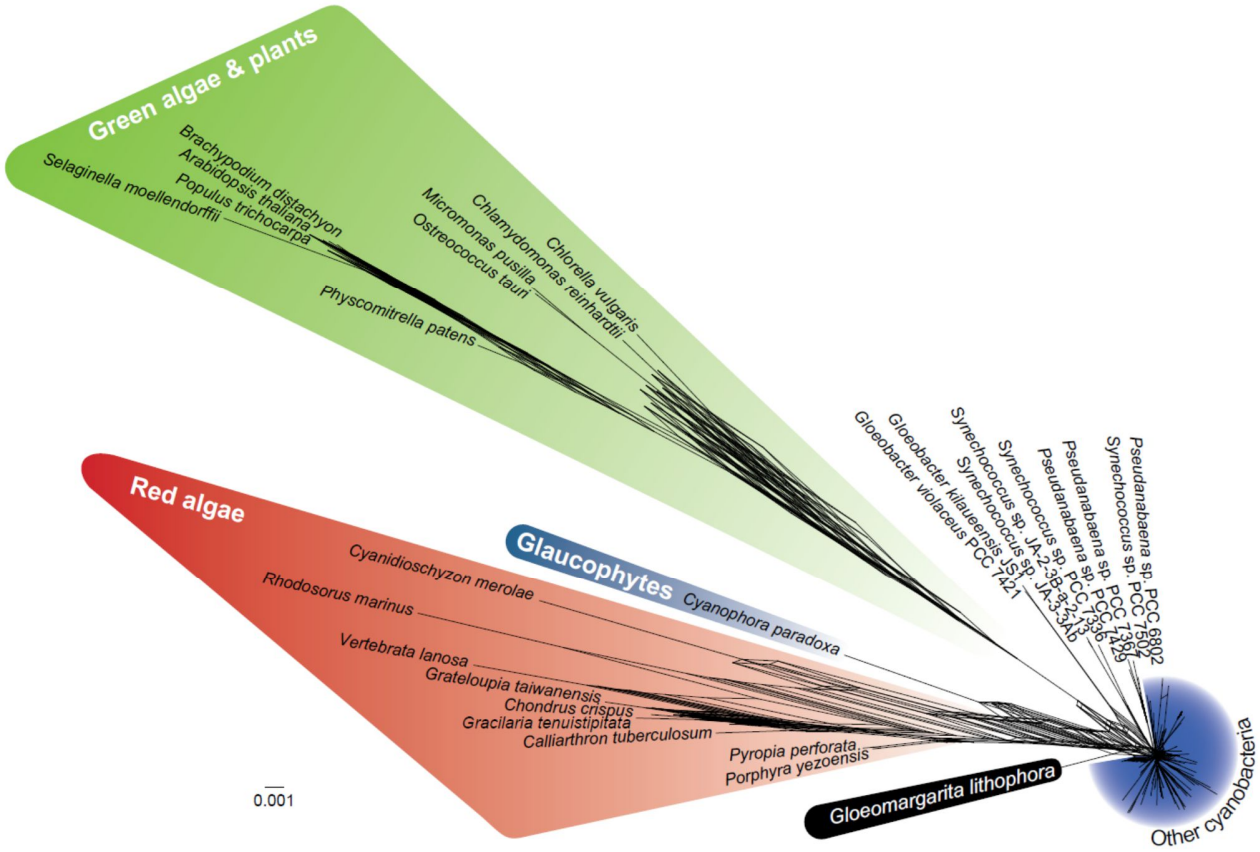
Green algae & plants
Brachypodium distachyon
Arabidopsis thaliana
Populus trichocarpa
Selaginella moellendorffii
Physcomitrella patens
Chlamydomonas reinhardtii
Micromonas pusilla
Ostreococcus tauri
Chlorella vulgaris
Red algae
Cyanidioschyzon merolae
Rhodosorus marinus
Vertebrata lanosa
Grateloupia taiwanensis
Chondrus crispus
Gracilaria tenuistipitata
Calliarthron tuberculosum
Pyropia perforata
Porphyra yezoensis
Glaucophytes
Cyanophora paradoxa
Gloeobacter violaceus PCC 7421
Gloeobacter kilaueensis JS1
Synechococcus sp. JA-2-3B-a(2-13)
Synechococcus sp. JA-3-3Ab
Pseudanabaena sp. PCC 7336
Pseudanabaena sp. PCC 7429
Synechococcus sp. PCC 7502
Pseudanabaena sp. PCC 6802
Gloeomargarita lithophora
Other cyanobacteria
0.001

Figure 2.