Cornet et al. BMC Res Notes (2018) 11:435
https://doi.org/10.1186/s13104-018-3543-y

BMC Research Notes

RESEARCH NOTE

Open Access

# A constrained SSU-rRNA phylogeny reveals the unsequenced diversity of photosynthetic Cyanobacteria (Oxyphotobacteria)

Luc Cornet[1,2], Annick Wilmotte[3,4], Emmanuelle J. Javaux[2] and Denis Baurain[1*]

## Abstract

**Objective:** Cyanobacteria are an ancient phylum of prokaryotes that contain the class Oxyphotobacteria. This group has been extensively studied by phylogenomics notably because it is widely accepted that Cyanobacteria were responsible for the spread of photosynthesis to the eukaryotic domain. The aim of this study was to evaluate the fraction of the oxyphotobacterial diversity for which sequenced genomes are available for genomic studies. For this, we built a phylogenomic-constrained SSU rRNA (16S) tree to pinpoint unexploited clusters of Oxyphotobacteria that should be targeted for future genome sequencing, so as to improve our understanding of Oxyphotobacteria evolution.

**Results:** We show that only a little fraction of the oxyphotobacterial diversity has been sequenced so far. Indeed 31 rRNA clusters of the 60 composing the photosynthetic Cyanobacteria have a fraction of sequenced genomes < 1%. This fraction remains low (min = 1%, median = 11.1%, IQR = 7.3%) within the remaining "sequenced" clusters that already contain some representative genomes. The "unsequenced" clusters are scattered across the whole Oxyphotobacteria tree, at the exception of very basal clades. Yet, these clades still feature some (sub)clusters without any representative genome. This last result is especially important, as these basal clades are prime candidate for plastid emergence.

**Keywords:** Cyanobacteria, Phylogeny, Diversity, Genomics, SSU rRNA (16S), Sequenced genome fraction

## Introduction

Oxyphotobacteria are a class of the phylum Cyanobacteria, one of the most important groups of prokaryotes [1]. This morphologically diversified class is the only group of Bacteria that can carry out oxygenic photosynthesis. It is commonly assumed that this bioenergetic process appeared in Oxyphotobacteria (or their ancestor) and caused the Great Oxidation Event (GOE) 2.4 billion years ago [2–5]. The atmospheric increase in free oxygen had a critical impact on early Earth evolution, including on Oxyphotobacteria themselves, with the appearance of new morphological features (e.g., akinetes) [6]. Oxyphotobacteria were also responsible for the spread of photosynthesis to eukaryotic lineages, through an initial event of primary endosymbiosis, followed by higher-order endosymbiosis [7]. In consequence of their great scientific interest, numerous phylogenies of Oxyphotobacteria have been published during the last decade (e.g., [5, 8–19]), using ever more genomes as it appears in public repositories. The diversity within Oxyphotobacteria is very high, both morphologically and genetically, and their GC-content ranges from 35 to 71 [20]. Therefore, it is important to assess how well the existing genomes represent the whole diversity of the class and detect the lineages for which representative genomes are not yet

*Correspondence: denis.baurain@uliege.be
[1] InBioS-PhytoSYSTEMS, Eukaryotic Phylogenomics, University of Liège, 4000 Liège, Belgium
Full list of author information is available at the end of the article

© The Author(s) 2018. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

available. Indeed, only 659 genomes of Oxyphotobacteria have been sequenced, whereas 6997 non-redundant rRNA sequences can be found in the SILVA database (as of October 2017). In this short communication, we built a SSU rRNA (16S) phylogeny to investigate the distribution of representative genomes across oxyphotobacterial diversity, so as to guide the selection of taxa for future sequencing projects.

## Main text

We took advantage of 7603 different SSU rRNA (16S) sequences [6997 from SILVA 128 Ref NR [21], 466 from genome assemblies publicly available on the NCBI [22] and 140 from the BCCM/ULC public collection of Cyanobacteria (http://bccm.belspo.be/about-us/bccm-ulc)], leading to a dataset of 2302 unique Operational Taxonomic Units (OTUs) after dereplication at an identity threshold of 97.5%. Because multiple sequence alignment (MSA) quality is essential in single-gene analyses [23–25], we used an in-house hand-curated MSA as a starting point to carefully align all the OTUs. This MSA represented a high fraction of known Oxyphotobacteria diversity, as it completely included the broadly sampled dataset of Schirrmeister et al. [26]. To alleviate the stochastic error plaguing single-gene analyses [27], we constrained our SSU rRNA (16S) phylogeny using a more reliable multi-gene phylogeny computed on a subset of the OTUs at hand. To this end, we used the phylogeny of Oxyphotobacteria developed in Cornet et al. [28], the largest ever-built in terms of conserved sites (> 170,000 unambiguously aligned amino-acid positions from 675 genes). This strategy allowed us to easily compare the genome diversity available for phylogenomic studies to the more completely sampled rRNA diversity of Oxyphotobacteria.

As shown in Fig. 1, 31 rRNA clusters have a fraction of sequenced genomes < 1% and are thus nearly completely ignored in phylogenomic studies. These

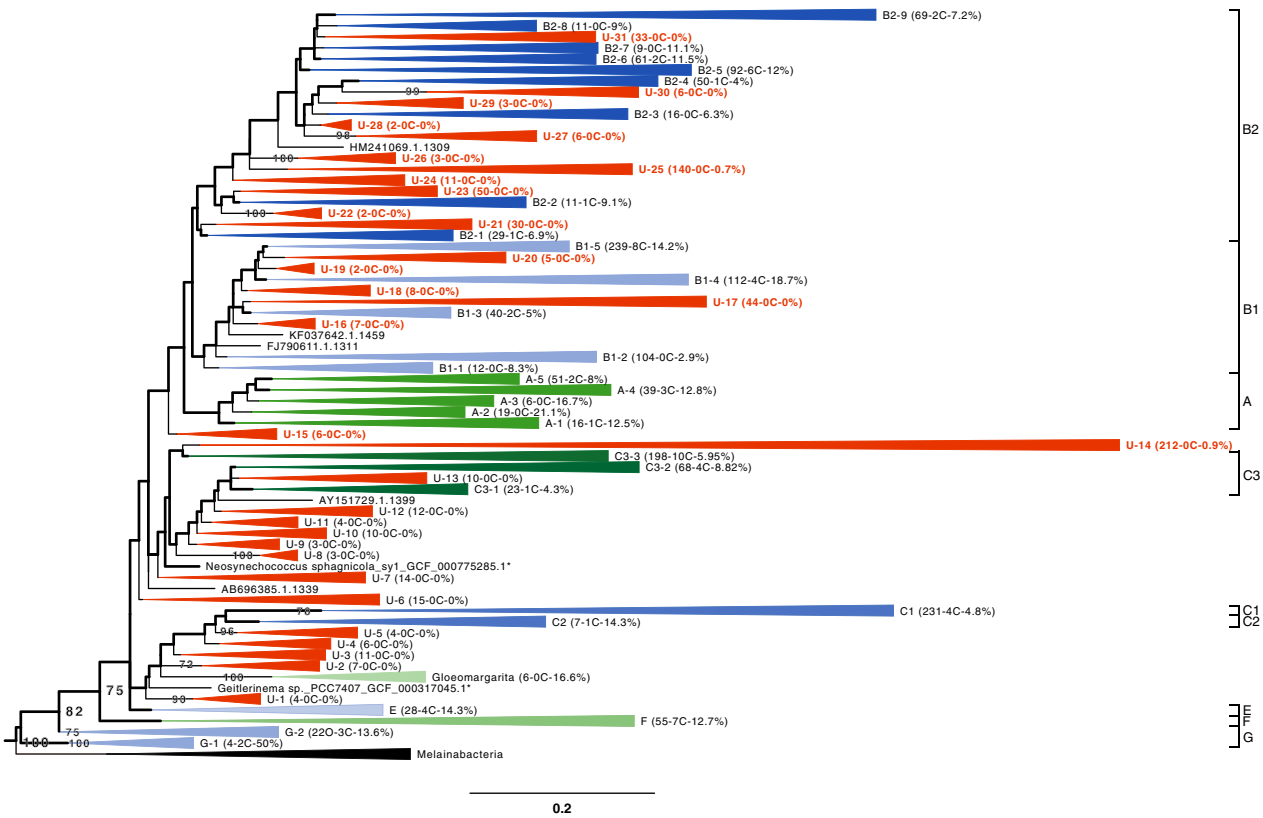


**Fig. 1** Collapsed constrained phylogeny of Oxyphotobacteria based on an alignment of 2302 SSU rRNA (16S) sequences (1233 unambiguously aligned nucleotide positions). The backbone of the tree shown in thick lines corresponds to the branches constrained by a phylogenomic tree of 71 OTUs × 170,983 positions [28]. The tree is rooted on Melainabacteria. Clusters on a red background have a fraction of sequenced genomes < 1% and can be considered as "unsequenced" due to the lack of representative genomes. Clusters on another coloured background have a higher fraction of sequenced genomes ("sequenced" clusters) and were annotated using the clade names of Shih et al. [12]. For each cluster (at the exception of Melainabacteria), we report in brackets the number of OTUs, the number of constrained genome sequences and the fraction of sequenced genomes

"unsequenced" clusters are scattered across the whole oxyphotobacterial tree, at the exception of very basal clades (G, F, E) and the Oscillatoriales clade (A), which have higher fractions of representative genomes. The group C (C1, C2, C3) appears paraphyletic in our tree, and is especially rich in "unsequenced" clusters (14 out of 31), including the largest one (212 OTUs). In all the remaining "'sequenced" clusters that contain representative genomes, the fraction of sequenced genomes remains low (min = 1%, median = 11.1%, IQR = 7.3%). It is thus clear that genomes have not been evenly sampled across the tree. To help researchers to select organisms for future sequencing projects, our final dataset and tree are made available (see below). In these files, SSU rRNA (16S) genes can be readily identified by their accession number, whether for SILVA/ULC strains or for rRNA genes predicted from genome assemblies.

The emergence of eukaryotic plastids from Oxyphotobacteria has been the subject of different hypotheses during the last decade, the most frequent being an early origin around clade E, composed of *Acaryochloris* and *Thermosynechococcus* species [17, 18]. By focusing on the base of our oxyphotobacterial phylogeny, Fig. 2 shows that in spite of the presence of a number of representative genomes, very basal clades (G, F, E) still feature some clusters without any genome (highlighted in yellow in

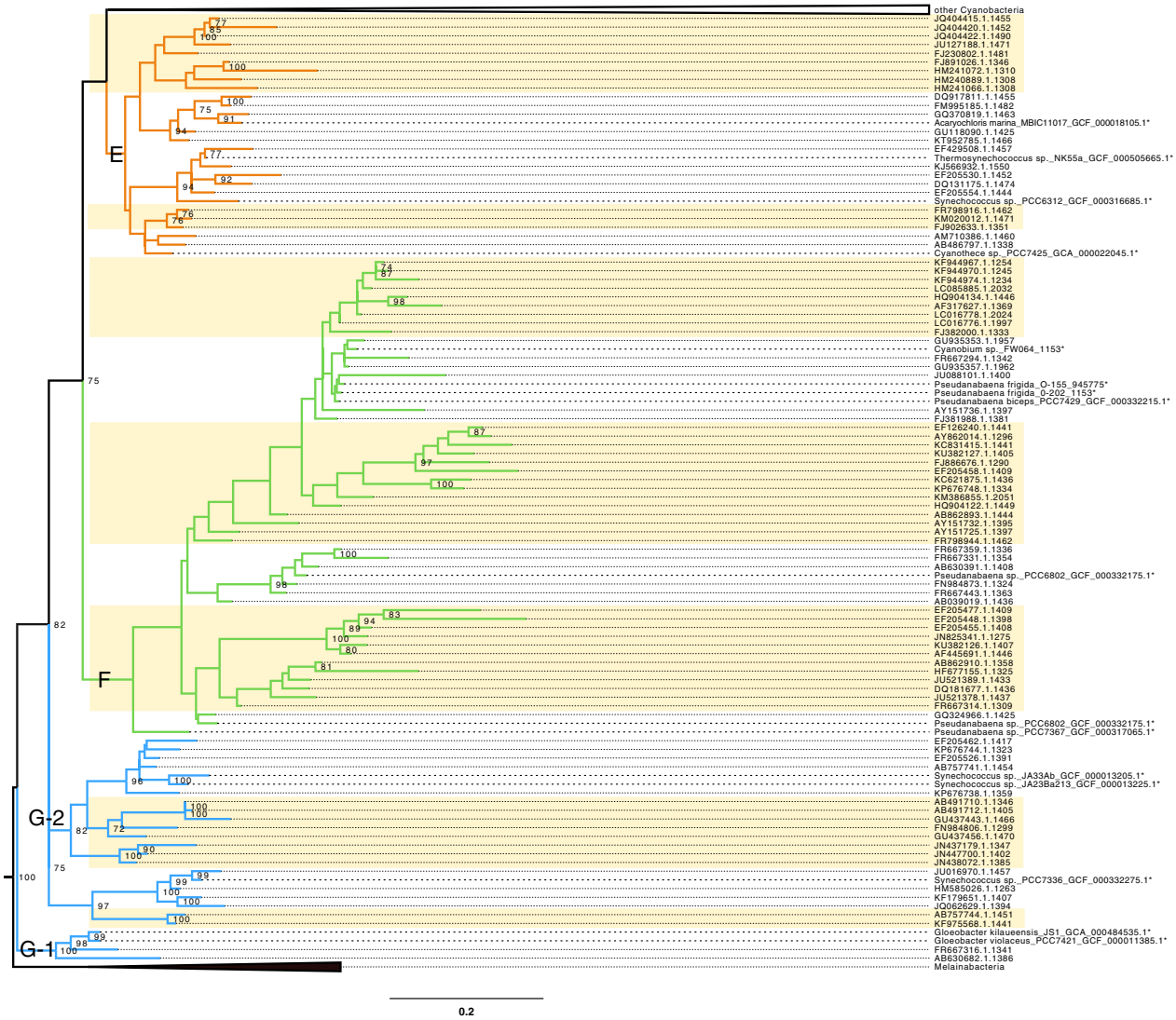


**Fig. 2** Expanded basal part of the constrained rRNA phylogeny of Oxyphotobacteria, with clusters E, F, G1 and G2. The tree is rooted on Melainabacteria. Asterisk indicate constrained genome sequences. The colour of the backbone corresponds to the groups of Shih et al. [12]. "Unsequenced" clusters (both monophyletic and paraphyletic) are shown on a yellow background

Fig. 2). Therefore, these clusters should be targeted in priority for genome sequencing if one wants to resolve the early diversification of Oxyphotobacteria and the origin of plastids by assembling phylogenomic datasets more faithful to the true diversity.

## Methods

Cyanobacterial (651 Oxyphotobacteria and 8 Melainabacteria) genome assemblies were downloaded from the NCBI FTP server on October 1st, 2017. SSU rRNA (16S) genes were predicted for all assemblies with RNAmmer v1.2 [29]. Predicted rRNA genes were taxonomically classified with SINA v1.2.11 [30], using the release 128 of SILVA database composed of 1,922,213 SSU rRNA (16S) reference sequences [21]. Among the 651 Oxyphotobacteria assemblies, 193 were devoid of Oxyphotobacteria rRNA genes (176 without any prediction and 17 with only non-cyanobacterial predicted genes). All Oxyphotobacteria rRNA genes (7001 sequences) of SILVA 128 Ref NR (non-redundant) were downloaded on October 1st, 2017. Four SSU rRNA sequences with a SILVA-provided incongruent lineage (e.g., Proteobacteria for *Prochlorothrix hollandica* PCC 9006) were deleted from the dataset, whereas the 466 sequences predicted with RNAmmer from genome assemblies were added, along with 140 new sequences from strains of the BBMC/ULC Cyanobacteria collection (http://bccm.belspo.be/about-us/bccm-ulc), hence leading to a total of 7603 SSU rRNA (16S) sequences (including 8 Melainabacteria).

The dataset was dereplicated with CD-HIT (cd-hit-est) v4.6 [31], using an identity threshold of 97.5%, an alignment coverage control of 0.0 for the longest sequence and 0.9 for the shortest sequence. The value of 97.5% for the identity threshold was chosen by following the OTU definition of Taton et al. [32]. It probably underestimates the true number of different species, since Edgar [33] has recently shown that the bacterial species delineation threshold is closer to 99% for full-length sequences. However, for the purpose of this study aimed at pinpointing large patches of unsequenced diversity, a coarse-grained clustering was sufficient. The cluster file produced by CD-HIT was retrieved and 13 representative sequences initially selected by CD-HIT were replaced by genomic sequences from the same CD-HIT clusters, so as to be able to enforce the phylogenomic constraints from Cornet et al. [28] when inferring the rRNA tree. This process led to a dataset of 2302 OTUs.

The 2302 sequences were aligned using the BLASTN-based software "two-scalp" (D. Baurain; https://metacpan.org/release/Bio-MUST-Apps-TwoScalp), starting from a hand-curated MSA (1252 sites post-BMGE; 1.52% missing character states) of 1128 SSU rRNA (16S) sequences (L. Cornet; unpublished). The 1128 sequences had been selected in order to be as close as possible to the dataset of Schirrmeister et al. [26], which represented all the Oxyphotobacteria diversity known at that time. Unambiguously aligned sites were selected with BMGE v1.12 [34], using moderately severe settings (entropy cut-off 0.5, gap cut-off 0.2) and yielding a high-quality MSA of 1233 conserved sites with only 1.83% missing character states. In comparison, building a MSA from scratch on the same dataset with MAFFT v7.273 [35], using its automatic mode, led to only 1013 conserved sites (2.52% missing character states) after applying the same protocol to select the unambiguously aligned sites.

The tree was inferred with RAxML v8.1.17 [36] under a GTR + $\Gamma_4$ model with a 100× rapid bootstrap analysis, using a constraining tree of 75 different Oxyphotobacteria. The constraining tree was a phylogenomic tree based on a concatenation of 675 genes (> 170,000 unambiguously aligned amino-acid positions [28]). However, four organisms (*Leptolyngbya* sp. GCF_000733415.1, Oscillatoriales cyanobacterium GCF_000309945.1, *Leptolyngbya* sp. OTC1/1, *Phormidium priestleyi* ANT. PROGRESS2.5) had to be pruned before inference due to the lack of genuine rRNA genes in their assemblies. The rRNA tree was then rooted on the Melainabacteria clan (sister-group of Oxyphotobacteria, composed of eight sequences), manually collapsed using FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree) and further arranged in InkScape v0.92 [37]. For each collapsed subtree, a "sequenced genome fraction" statistic was computed by taking the ratio between the number of OTUs having a publicly available genome in their CD-HIT cluster to the total number of OTUs in the subtree.

## Limitations

Among the 659 available Oxyphotobacteria genomes, 193 genomes were either completely devoid of rRNA genes (or only featured taxonomically incongruent rRNA genes) and were thus unusable for phylogeny (see "Methods" for details). Obviously, the results of this study might have been different if we had used these genomes to evaluate the sequenced genome fraction. The absence of rRNA genes is explained by the fact that the majority of these genomes (143 out of 193) were assembled with a metagenomic pipeline, according to NCBI metadata. This phenomenon was noticed in Cornet et al. [38] and is due to the frequent loss of rRNA sequences during metagenomic assembly [28]. Given that neither the isolation source nor the assembly pipeline was easy to determine, we decided not to consider these 193 genomes in our analysis by precautionary principle. Hence, a significant number of these metagenomes were found in the genome section of RefSeq, even though it is not allowed. This decision is moreover consolidated by the observation

that only a small number (about 20) of SSU rRNA (16S) sequences could be recovered after extensive searches on the NCBI using the "strain" names of these 193 genomes.

### Abbreviations
OTUs: Operational Taxonomic Units; MSA: multiple sequence alignment.

### Authors' contributions
LC and DB designed the experiments, AW generated new SSU rRNA (16S) sequences for 140 strains from the BCCM/ULC public collection of Cyanobacteria, LC performed the computational analyses and drew the figures, EJJ and AW helped with the biogeological interpretations of the results, LC and DB wrote the manuscript, with the assistance of AW and EJJ. All authors read and approved the final manuscript.

### Author details
[1] InBioS-PhytoSYSTEMS, Eukaryotic Phylogenomics, University of Liège, 4000 Liège, Belgium. [2] UR Geology-Palaeobiogeology-Palaeobotany-Palaeopalynology, University of Liège, 4000 Liège, Belgium. [3] InBioS-CIP, Centre for Protein Engineering, University of Liège, 4000 Liège, Belgium. [4] BCCM/ULC Collection of Cyanobacteria, University of Liège, 4000 Liège, Belgium.

### Acknowledgements
We thank Yannick Lara (ULiège) for the cultivation and sequencing of some the BCCM/ULC strains. Rosa Gago is acknowledged for her help when drawing the figures. Finally, we are grateful to the editor and the anonymous reviewer for their insightful comments.

### Competing interests
The authors declare that they have no competing interests.

### Availability of data and materials
The dataset (hand-curated MSA, final raw MSA, final MSA post-processed by BMGE, phylogenomic constraining tree, and final SSU rRNA ML tree) supporting the conclusions of this article is available as a compressed archive in the figshare repository through the following link (https://doi.org/10.6084/m9.figshare.6225641).

### Consent for publication
Not applicable.

### Ethics approval and consent to participate
Not applicable.

### Funding
This work was supported by operating funds from FRS-FNRS (National Fund for Scientific Research of Belgium) (AW), the European Research Council Stg ELITE FP7/308074 (EJJ), the BELSPO project CCAMBIO (SD/BA/03A) (AW), the BELSPO Interuniversity Attraction Pole Planet TOPERS (EJJ and LC). LC was a FRIA fellow of the FRS-FNRS. AW is a Research Associate of the FRS-FNRS. Computational resources were provided through two grants to DB (University of Liège "Crédit de démarrage 2012" SFRD-12/04; FRS-FNRS "Crédit de recherche 2014" CDR J.0080.15).

## Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 10 May 2018 Accepted: 26 June 2018
Published online: 03 July 2018

### References
1. Soo RM, Hemp J, Parks DH, Fischer WW, Hugenholtz P. On the origins of oxygenic photosynthesis and aerobic respiration in Cyanobacteria. Science. 2017;355:1436–40.
2. Bekker A, Holland HD, Wang P-L, Rumble D, Stein HJ, Hannah JL, et al. Dating the rise of atmospheric oxygen. Nature. 2004;427:117–20.
3. Knoll AH. The geological consequences of evolution. Geobiology. 2003;1:3–14.
4. Kopp RE, Kirschvink JL, Hilburn IA, Nash CZ. The Paleoproterozoic snowball Earth: a climate disaster triggered by the evolution of oxygenic photosynthesis. Proc Natl Acad Sci USA. 2005;102:11131–6.
5. de Alda JAGO, Esteban R, Diago ML, Houmard J. The plastid ancestor originated among one of the major cyanobacterial lineages. Nat Commun. 2014;5:4937.
6. Gonzalez-Esquer CR, Smarda J, Rippka R, Axen SD, Guglielmi G, Gugger M, et al. Cyanobacterial ultrastructure in light of genomic sequence data. Photosynth Res. 2016;129:147–57.
7. Archibald JM. The puzzle of plastid evolution. Curr Biol. 2009;19:R81–8.
8. Rodríguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Löffelhardt W, et al. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. Curr Biol. 2005;15:1325–30.
9. Blank CE, Sánchez-Baracaldo P. Timing of morphological and ecological innovations in the cyanobacteria—a key to understanding the rise in atmospheric oxygen. Geobiology. 2010;8:1–23.
10. Criscuolo A, Gribaldo S. Large-scale phylogenomic analyses indicate a deep origin of primary plastids within Cyanobacteria. Mol Biol Evol. 2011;28:3019–32.
11. Dagan T, Roettger M, Stucken K, Landan G, Koch R, Major P, et al. Genomes of Stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. Genome Biol Evol. 2013;5:31–44.
12. Shih PM, Wu D, Latifi A, Axen SD, Fewer DP, Talla E, et al. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. Proc Natl Acad Sci. 2013;110:1053–8.
13. Blank CE. Origin and early evolution of photosynthetic eukaryotes in freshwater environments: reinterpreting proterozoic paleobiology and biogeochemical processes in light of trait evolution. J Phycol. 2013;49:1040–55.
14. Li B, Lopes JS, Foster PG, Embley TM, Cox CJ. Compositional biases among synonymous substitutions cause conflict between gene and protein trees for plastid origins. Mol Biol Evol. 2014;31:1697–709.
15. Sánchez-Baracaldo P. Origin of marine planktonic cyanobacteria. Sci Rep. 2015;5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4665016/. Accessed 7 Nov 2017.
16. Uyeda JC, Harmon LJ, Blank CE. A comprehensive study of cyanobacterial morphological and ecological evolutionary dynamics through deep geologic time. PLoS ONE. 2016;11:e0162539.
17. Ponce-Toledo RI, Deschamps P, López-García P, Zivanovic Y, Benzerara K, Moreira D. An early-branching freshwater cyanobacterium at the origin of plastids. Curr Biol. 2017;27:386–91.
18. Sánchez-Baracaldo P, Raven JA, Pisani D, Knoll AH. Early photosynthetic eukaryotes inhabited low-salinity habitats. Proc Natl Acad Sci. 2017;114:E7737–45.
19. Walter JM, Coutinho FH, Dutilh BE, Swings J, Thompson FL, Thompson CC. Ecogenomics and taxonomy of Cyanobacteria phylum. Front Microbiol. 2017;8:2132. https://doi.org/10.3389/fmicb.2017.02132.
20. Herdman M, Janvier M, Waterbury JB, Rippka R, Stanier RY, Mandel M. Deoxyribonucleic acid base composition of cyanobacteria. Microbiology. 1979;111:63–71.
21. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41:D590–6.
22. Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, et al. The NCBI biosystems database. Nucleic Acids Res. 2010;38:D492–6.
23. Lunter G, Rocco A, Mimouni N, Heger A, Caldeira A, Hein J. Uncertainty in homology inferences: assessing and improving genomic sequence alignment. Genome Res. 2008;18:298–309.
24. Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. Science. 2008;319:473–6.
25. Dessimoz C, Gil M. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol. 2010;11:R37.
26. Schirrmeister BE, Anisimova M, Antonelli A, Bagheri HC. Evolution of cyanobacterial morphotypes. Commun Integr Biol. 2011;4:424–7.
27. Delsuc F, Brinkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 2005;6:nrg1603.

28. Cornet L, Bertrand AR, Hanikenne M, Javaux EJ, Wilmotte A, Baurain D. Metagenomic assembly of new (sub)arctic Cyanobacteria and their associated microbiome from non-axenic cultures. bioRxiv. 2018. https://doi.org/10.1101/287730.
29. Lagesen K, Hallin P, Rødland EA, Stærfeldt H-H, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35:3100–8.
30. Pruesse E, Peplies J, Glöckner FO. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics. 2012;28:1823–9.
31. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.
32. Taton A, Grubisic S, Brambilla E, Wit RD, Wilmotte A. Cyanobacterial diversity in natural and artificial microbial mats of Lake Fryxell (McMurdo Dry Valleys, Antarctica): a morphological and molecular approach. Appl Environ Microbiol. 2003;69:5157–69.
33. Edgar RC. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. PeerJ. 2018;6:e4652.
34. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol. 2010;10:210.
35. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
36. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22:2688–90.
37. Bah T. Inkscape: guide to a vector drawing program (Digital Short Cut). New York: Pearson Education; 2009.
38. Cornet L, Meunier L, Vlierberghe MV, Leonard RR, Durieu B, Lara Y, et al. Consensus assessment of the contamination level of publicly available cyanobacterial genomes. PLOS One. 2018. https://doi.org/10.1101/301788.



Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

**Learn more** biomedcentral.com/submissions

BMC