Tracking evolutionary trends towards increasing complexity: a case study in Cyanobacteria

Andrés Moya[1,2,3], José L. Oliver[4,5,*], Miguel Verdú[6,*], Luis Delaye[7,*], Vicente Arnau[1], Pedro Bernaola-Galván[8], Rebeca de la Fuente[9], Wladimiro Díaz[1], Cristina Gómez-Martín[4,5], Francisco M. González[7], Amparo Latorre[1,2,3], Ricardo Lebrón[4,5] and Ramón Román-Roldán[10]

[1] Institute of Integrative Systems Biology (I2Sysbio), University of València and Consejo Superior de Investigaciones Científicas (CSIC), 46980 València, Spain.

[2] Foundation for the Promotion of Sanitary and Biomedical Research of Valencian Community (FISABIO), 46020 València, Spain.

[3] CIBER in Epidemiology and Public Health, 28029 Madrid, Spain

[4] Department of Genetics, Faculty of Sciences, University of Granada, 18071 Granada, Spain

[5] Laboratory of Bioinformatics, Institute of Biotechnology, Center of Biomedical Research, 18100 Granada, Spain.

[6] Centro de Investigaciones sobre Desertificación, Consejo Superior de Investigaciones Científicas (CSIC), University of València and Generalitat Valenciana, 46113 València, Spain

[7] Department of Genetic Engineering, CINVESTAV, 36821 Irapuato, México.

[8] Department of Applied Physics II and Institute Carlos I for Theoretical and Computational Physics, University of Málaga, 29071 Málaga, Spain.

[9] Institute for Cross-Disciplinary Physics and Complex Systems, University of Balearic Islands and CSIC, 07122 Palma de Mallorca, Spain.

[10] Department of Applied Physics, University of Granada, 18071 Granada, Spain.

Key words: evolutionary progress; genome complexity; metrics of genome complexity; phylogenetic signal; phylogenetic regression; Cyanobacteria

Running head: Evolutionary trends in Cyanoabacteria

*These authors contributed equally

Corresponding author: Andrés Moya, andres.moya@uv.es

Abstract

Progressive evolution, the tendency towards increasing complexity, is a controversial issue in Biology, whose resolution requires a proper measurement of complexity. Genomes are the best entities to address this challenge, as they record the history and information gaining of organisms in their ongoing biotic and environmental interactions. Using six metrics of genome complexity, none of which is primarily associated to biological function, we measure genome complexity in 91 genomes from the phylum Cyanobacteria. Several phylogenetic analyses reveal the existence of progressive evolution towards higher genome complexity: 1) all the metrics detect strong phylogenetic signals; 2) ridge regressions detect positive trends towards higher complexity; and 3) classical proofs for progressive evolution (the minimum, the ancestor-descendent and the sub-clade tests), show that some of these positive trends are driven, being mainly due to natural selection. These findings support the existence of progressive genome evolution in this ancient and diverse group of organisms.

Introduction

Treatises on biological evolution reflect a conflict between the relative roles played by contingency and necessity (Moya, 2014). An important tradition in Evolutionary Biology, based on a large amount of empirical evidence, considers that contingency marks the dynamics of evolution in a way that makes it unpredictable. The trend towards the appearance of increasing complexity falls within the frame of contingent evolution insofar as it is inevitable given that, passively, we can expect that sooner or later more complex entities will evolve from the original, simpler entities. This is what Gould (1996) calls the passive tendency towards complexity marked by the minimum initial complexity wall.

Assuming there is an evolutionary trend towards greater complexity, a fundamental question is whether we can prove the existence of a metric accounting for it (McShea and Brandon, 2010; Day, 2012). We first conjecture that it is in the genomes where we can find evidence of such metrics, which may eventually increase over evolutionary time. Genomes record the history and information gained by organisms during their interaction with environmental and biotic factors over time (Adami, 2002, 2016; Krakauer, 2011). However, genome parameters such as genome size, number of genes, gene components (i.e., introns, exons), etc., are insufficient to show any evolutionary trend. We speculate that this is probably because they only partially capture the abovementioned history and information gain (Lieberman-Aiden *et al*., 2009; Dekker *et al*., 2017). Our second conjecture is that the best way to measure genome complexity is by using metrics that are not primarily associated to biological function.

The currently available metrics applied to genomes are very broad and not all of them appropriately capture the information gained by the genome during its evolutionary history (Zurek, 1990; Adami, 2002). Algorithmic complexity (Chaitin, 1977; Li and Vitányi, 2008) is inconveniently maximum for randomness. The effective complexity of Gell-Mann and Lloyd (1996) is recommended for collections or ensembles of sequences, but in many cases, such as in genome sequences, it is not clear how to define an appropriate ensemble. Likewise, those based on mutual information, i.e. statistical dependence (Grassberger, 1986; Adami and Cerf, 2000) also quantify the complexity of sequence ensembles generated by a given source rather than the complexity of a single sequence. Here we consider metrics based on an individual entity (genome). Interestingly, four of them consider the number of different parts composing the genome and the irregularity of their distribution, thus extrapolating to DNA sequences the operational definition of McShea (1993), which measures biological complexity as the degree to which the parts of a morphological structure differ from each other. Two additional metrics are based on *k-mer* statistics. A key point is that none of the six metrics we used to measure genome complexity considers biological function.

A first group of metrics are based on the Sequence Compositional Complexity derived from a four-symbol DNA sequence (*SCC*) or from the binary sequences resulting from grouping the four nucleotides into S(C,G) vs. W(A,T) or R(A,G) vs. Y (T,C), or K(A,C) vs. M(T,G), thus obtaining SCC_{SW}, SCC_{RY} and SCC_{KM} metrics, respectively (Román-Roldán *et al.*, 1998). These four metrics increase with the number of parts (i.e. compositional domains) found in a genome sequence by a segmentation algorithm, and both the length and compositional differences among them.

The fifth metric we used is the Biobit (*BB*), a metric based on the difference between the maximum entropy for a *k-mer* of a random genome of same length as the genome under consideration and the entropy of that genome for such a *k*-mer (Bonnici and Manca, 2016). Lastly, we used the Genomic Signature (*GS*), also a *k-mer*-based metric, which is the value corresponding to the *k-mer* that maximizes the difference between observed and expected equi-frequent classes of *mers*. *GS* is based on the relative abundances of short oligonucleotides (Karlin and Ladunga, 1994) and chaos game representation applied to genomes (Jeffrey, 1990; Almeida *et al*., 2001).

As already stated, the four *SCCs* are the only metrics using genome segmentation (i.e., the partition of the genome into non-overlapping fragments of varying lengths and with homogeneous composition). These metrics parallel the concept of 'pure complexity' of McShea and Brandon (2010), where complexity is more closely related with the number of part types of a living being than with the number of functions. The other two metrics, based on distribution of *k-mers*, do not use genome partitioning.

We test the above-mentioned metrics by analyzing the genome evolution of an ancient and diverse group of organisms: the Cyanobacteria phylum. These microorganisms played a fundamental role in the evolution of life on Earth. The fossil record shows that they were here 2.0 billion years ago (Bya) and molecular clock analyses indicate that the phylum originated over 2.5 Bya (Sergeev *et al*., 2002; Schirrmeister *et al.,* 2013). By releasing oxygen through photosynthesis, Cyanobacteria caused the Great Oxidation Event about 2.3 Bya and changed the history of life on Earth (Bekker *et al.,* 2004). The oxidation of the environment allowed the evolution of complex multicellular life forms (Hedges *et al*., 2004), leading to the origin of eukaryotic crown groups including plants and animals (Knoll, 2014). As it is well known, Cyanobacteria originated plastids through symbiosis with ancient eukaryotes (Sagan, 1967).

Cyanobacteria are morphologically diverse. The group was traditionally classified into five subsections according to several biological criteria (Rippka *et al*., 1979; Rippka, 1988). These subsections of Cyanobacteria are not monophyletic (Schirrmeister *et al.,* 2013; Dagan *et al.,* 2013). More recent classification schemes using phylogenetic analysis from 31 conserved protein sequences divide Cyanobacteria into nine different groups (Komárek *et al*., 2014). These are: Gleobacterales, Synechococcales, Oscillatoriales, Chroococcales, Pleurocapsales, Spirulinales, Rubidibacter/Halothece, Chroococcidiopsidales and Nostocales. Not all these groups are monophyletic. Clearly, taxonomy and evolution of Cyanobacteria is an active area of research and this classification is likely to change in the near future.

As a proof of concept, in the present work we test whether there is a statistically and phylogenetically supported driven tendency towards increasing genome complexity as reflected by the metrics of genome complexity and/or by genome standard parameters (genome size, %GC and number of genes) in the evolution of Cyanobacteria.

Results

Complexity measures throughout Cyanobacteria phylogeny

The values of the four *SCCs*, *BB* and *GS* metrics as well as three standard genome parameters (Genome size, %GC and No. of genes) (see material and methods) for 91 Cyanobacterial genomes are reported in Supplementary file 1. Figure 1 shows a maximum likelihood phylogeny of Cyanobacteria whose branch lengths are proportional to the number of amino acid substitutions (see material and methods). The phylogeny is in general agreement with previous analysis (Dagan *et al*., 2013; Schirrmeister *et al*., 2013; Komárek *et al*., 2014).

Phylogenetic signal

All metrics of complexity and genome parameters showed a significant phylogenetic signal (Table 1), indicating that for all cases genomes of closely related cyanobacterial species tend to resemble more than two randomly selected genomes (Figure 1). However, the magnitude of the phylogenetic signal differs across metrics and parameters, with %GC and *GS* showing the highest values, which probably reflects different forces shaping the evolution of all these metrics and parameters (see Discussion).

Phylogenetic correlations

To gain a better understanding of the metrics, after correction of the phylogenetic signals, we evaluated how they correlate with each other and, particularly, with the genomic parameters (see Table 2). It is worth noticing that two metrics in particular, *SCC* and SCC_{RY} correlate with other metrics and parameters (six correlations each one), accounting for 43% of all significant correlations.

Ridge regression of complexity metrics *versus* age

Using ridge regression of genome complexity metrics and genome parameters *versus* age (distance from the root), we have studied whether there is evidence of evolutionary trends, having detected interesting patterns (Figure 2, panels A and B). Of the complexity metrics, four out of six show a statistically significant positive trend (*SCC,* SSC_{SW}*,* SCC_{RY} and *GS*), indicating that complexity has increased with evolutionary time.

In contrast, SCC_{KM} shows no trend and *BB* reveals a significant negative trend. Interestingly genome parameters, on the other hand, show no evidence of any evolutionary trend.

Driven trends in Cyanobacteria

To unravel whether the positive trends are passive or driven we have applied three types of proofs, called the minimum, the ancestor-descendent and the subclade proofs, respectively (McShea, 1994; see also McShea and Brandon, 2010). These proofs are well known in Paleontology and Evolutionary Biology and, to the best of our knowledge, this is the first time they have been applied to genome evolutionary analyses. To gain a better understanding of the positive trends we have also applied those proofs for comparative purposes to the metrics and genome parameters that do not show evidence of such a positive evolutionary trend.

Minimum proof

Regarding minimum proof, we have applied three types of tests. The first one is the skewness of the entire phylum. In this respect, we observed that the skewness of all metrics (except *SCC* and GC content) exhibit significant and positive skewness (D'Agostino-Pearson test, $n = 91$; see Table 3). This result supports a left wall for these metrics and parameters, which is compatible with either a passive or a driven trend. On the other hand, it is expected that if the minimum value of a given metric or parameter delimiting the left wall increases with evolutionary time, then the trend will probably be driven. To evaluate this, we considered as the minimum the estimated value of the most basal clade, x_b, for each metric/parameter (see Figure 1). However, as it can be observed (Figure 3), there are lower or higher values than the one corresponding to the basal clade for any metric/parameter. A deeper study of the distribution of lower and higher values with respect to x_b could give us evidence of the putative existence of a driven trend. To test this (second test of the minimum proof) we measure $|x_d\text{-}x_b|$, the absolute difference between descendants' clades and the most basal clade. Table 3 shows whether there is a statistical difference (Chi-square test) between those clades (179 in total) that are higher or lower than the basal clade, x_b. As it can be observed, all the tests are significant with four metrics (*SCC*, SCC_{RY}, SCC_{KM} and *BB*) and two parameters (Genome size and No. of genes) showing a significant increase in the metric/parameter with respect to the corresponding basal values. In contrast, two metrics (SCC_{SW} and *GS*) and one parameter (%GC) present a significant decrease. We have also tested by means of a Student's t-test (third test of the minimum proof) whether there is a statistical difference between the average value of the absolute difference ($|x_d\text{-}x_b|$) of a give metric or parameter higher or lower than x_b. In this case, it can be observed (see Table 3) that three metrics (SCC_{SW}, SCC_{RY} and SCC_{KM}) show a statistical difference in favor of a higher value than x_b and one metric (*GS*) and the three parameters (Genome size, %GC content and No. of genes) present a statistical difference in favor of a lower value than x_b.

The ancestor-descendent proof

According to Gould (1996) the ancestor-descendent proof is the most appropriate one to discover whether positive trends are passive or driven. McSchea (1994) indicates that in a passive system, increases and decreases should be the same, whereas in a driven trend the number of increases should occur more often. To test this, we tabulated the derived clades for all possible nodes and whether they present a higher, lower or equal value of the metric/parameter than the ancestral clade corresponding to each node. In order to avoid bias due to proximity to the putative left wall, McSchea (1994) recommends applying the test only to those clades where both ancestor and descendent are higher than the average value of the metric/parameter. As it can be observed (Table 4) this exigent test shows that metrics *SCC* and *GS* and the three genome parameters are in favor of a driven trend. A good visualization of the ancestor-descendant proof on the phylogeny of the Cyanobacteria for each metric/parameter has been obtained by estimating the values of internal nodes using a maximum likelihood function and interpolating the value along each edge. Figure 4 shows the mapping corresponding to the *SCC* metric where the driven positive trend of this metric can be clearly appreciated (see Supplementary file 2 for the mapping of the rest of metrics/parameters).

The sub-clade proof

The final proof applied is the sub-clade proof. According to McSchea (1993) if the parent distribution is skewed (see histograms of Figure 3 and Table 3) and the mean skew of a sub-clade drawn from the right tail is also skewed, the system is probably driven. For this proof we have applied two types of test. First, we tested whether the trend observed at Phylum level is also observed in four selected monophyletic clades and second, we have also applied the skewness test proposed by McSchea (1993) properly to either the entire phylum (results given in Table 3) and to the chosen sub-clades. We have chosen four monophyletic clades formed by clusters 97, 132, 162 and 172 that harbor 18, 22, 11 and 8 species, respectively (four color boxes in Figure 1 and Supplementary file 3). Clade 97 is formed by Synechococcus, Prochlorococcus and Cyanobium; clade 132 corresponds to Nostocales (subsections IV and V, Komarek *et al.*, 2012); clade 162 contains Cyanothece and Microcystis; and clade 172, among others, contains Geminocystis and Cyanobacterium. The most relevant result found was that some metrics of genome complexity show statistically significant positive trends (SCC, SCC_{RY} and GS) and some others show negative trends (SCC_{SW} and SCC_{KM}), whereas the genomes parameters do not show any positive trends (see Supplementary files 4 and 5). Thus, we keep *SCC*, SCC_{RY} and *GS* as the metrics showing positive trends at both levels of phylogenetic resolution.

Regarding the second test for sub-clade proof, we followed the criteria given by McShea, (1994) whereby the monophyletic sub-clade drawn from the right tail of the entire distribution should have a statistically significant average higher value than the one corresponding to the entire phylum. Regarding the skewness of the phylum (Table 3) we observe that all metrics (except *SCC* and %GC) exhibit significant and positive skewness. However, this test of skewness cannot be applied to the four chosen monophyletic sub-clades either because a) the average value (median) of a given metric/parameter for each sub-clade was lower than the median of the phylum (16 cases out of 36) or, b) there was no statistical evidence (the remaining 20 cases) of a higher median (Mood's median test) of a given metric/parameter for each sub-clade than the median of the entire phylum (see Supplementary file 6).

In summary**,** the overall results obtained in relation to the evidence found for a trend in a given metric or parameter (i.e., the phylogenetic signal, the number of significant correlations against the rest of metrics/parameters, as well as whether the trend is driven or not (see Table 5)) show that *SCC*, SCC_{RY} and to a lower extent *GS* present the highest scores, and can thus be considered metrics evidencing progressive evolution of Cyanobacteria.

Discussion

Genomes probably provide the best record of the biological history of species, not only do they enable us to reconstruct their phylogenetic relationships but they also contain information gained from their continuous biotic and environmental interactions over time (Adami, 2002; Krakauer, 2011). This information is an elusive but crucial component of the genome, whose study as a whole deserves deeper attention because it holds clues to answer many biological questions, particularly those of an evolutionary nature. The genome has distinct layers of information encoded in DNA sequences (Dekker *et al*., 2017; Cristadoro *et al*., 2018). The most well-known of these are those involved in biological function, such as the typical genome division into coding and non-coding parts or, within genes, the differential conservation shown by distinct codon positions due to the differential evolutionary constraints acting on them (Ikemura, 1985; Sueoka, 1992; Bernardi, 2004). In the present study we intend to capture or approximate the genome information held in these layers using certain metrics (collectively named 'genome complexity metrics') to determine whether they show phylogenetic signals and indicate - or not - some kind of evolutionary trend. To do so, we use a group of organisms with a long phylogenetic history: the phylum Cyanobacteria. *SCC* accounts for the global compositional complexity of a DNA sequence encoded by the four nucleotides {A, T, C and G} and shares similarity with McShea's (1993) operational definition of biological complexity: the degree to which the parts of a morphological structure differ from each other. SCC_{SW} may account for the complexity due to the partition of the genome into GC-rich and GC-poor segments (as for example, the isochores), which are known to be associated to many functionally relevant properties such as gene density, gene length, retrotransposon density, recombination frequency, etc. (Bernardi *et al*., 1985; Mouchiroud *et al*., 1988; Zoubak *et a*l., 1996; Oliver *et al.*, 2004; Bernardi, 2015; Jabbari and Bernardi, 2017). Thus, SCC_{SW} might capture the genome information gained throughout evolution by the selective forces acting on these important functional elements. On the other hand, SCC_{RY} accounts for the complexity due to the partition of the genome into segments of different purine/pyrimidine richness. Such strand asymmetries are less directly related to biological function, but this alphabet has been useful to uncover long-range correlations and analyze the evolution of fractal structure in the genomes (Li and Kaneko, 1992; Peng *et al.*, 1992; Voss, 1992). Recently, a connection has been found between strand symmetry and the repetitive action of transposable elements during evolution (Cristadoro *et al.*, 2018; see also Koonin, 2016a and his concept of 'fuzzy meaning' of sequences). The partition given by SCC_{KM}, to our knowledge, has not been associated to any biological function. Finally, *GS* and *BB* explore the maximum deviation for a given *k-mer* between a real and a random genome. *GS* directly compares the observed distribution of *k-mer* classes of a real genome with respect to that corresponding to a random one. On the other hand, by calculating the entropy differences between both groups, *BB* measures the relative entropic and anti-entropic fraction of a real genome (Bonnici and Manca, 2016).

From a population genetics perspective, Cyanobacteria can be considered proto-typical bacterial species whose populations are evolving under high effective population sizes (Lynch and Connery, 2003), with intermediate mutation rates between those of RNA viruses (higher mutation rate) and lower or higher eukaryotes (lower mutation rates) (Gago *et al.*, 2009). Therefore, natural selection is expected to play a major role in the evolution of these organisms. Irrespective of whether mutations (or any source of genetic novelty) are deleterious or beneficial, their destiny will be dictated by the deterministic action of purifying or positive selection, respectively (Lynch, 2007; Koonin, 2016b). This observation is highly pertinent when it comes to appropriately interpreting the phylogenetic signals observed in the metrics of complexity measures and genome parameters following the *in silico* evolutionary processes described by Revell *et al*. (2008). Considering, thus, that selection is a key force in the evolution of

Cyanobacteria, most of the *K* values estimated for the metrics may reflect the action of purifying or stabilizing selection, particularly those that are below 1 (all metrics and parameters, except *GS* and %GC). *K* from *GS* is 1, which could be interpreted either as a random drift effect or, more convincingly for this type of organism, as fluctuating selection for a relatively high rate of movement of the optimum (Revell *et al*., 2008). Finally, *K* associated to %GC is much higher than one, which can also be interpreted as the result of an evolutionary process with heterogeneous peak shifts.

Importantly, our study of the evolutionary trends in Cyanobacteria by means of ridge regression found clear differences between metrics of complexity and genome parameters. Four metrics (*SCC, SCC_{RY}, SCC_{SW}* and *GS*) indicate changes toward higher complexity in more evolved clades (long-branch distance with respect to the root of the tree), while *SCC_{KM}* does not show any signs of a trend and *BB* shows a negative trend. However, the genome parameters show no evidence of any trend (Figure 2). These results are reinforced when comparatively analyzing trends between metrics and parameters at a lower phylogenetic resolution (Supplementary files 4, 5 and 6). Although metrics used in this work capture different aspects of the evolution of genome complexity in Cyanobacteria (positive trends in *SCC, SCC_{RY}* and *GS* vs. negative trends in *SCC_{SW}* and *SCC_{KM}*), the genome parameters never present any positive trends (Supplementary files 3 and 4). In that respect, although some metrics capture increasing complexity, genome parameters do not. It is interesting to point out the process of selection and genome streamlining of *Synechococcus* and *Prochlorococcus* in clade 97, giving rise to more evolved shorter genomes, which are AT-rich and show a lower number of genes than the rest of Cyanobacteria (Supplementary file 1). As it can be observed, there are statistically significant negative trends in the three genome parameters but also positive trends of *SCC* and *SCC_{RY}* metrics (Supplementary files 3 and 4). Therefore, genome reduction in this clade does not imply loss of genome complexity; on the contrary, our study shows that this younger clade also has the most complex genomes of Cyanobacteria, a result in agreement with the high density of functional sequences observed in these free-living organisms (Batut *et al*., 2014).

In summary, considering that selection is a major driver in the evolution of Cyanobacteria, the observed positive trends towards increasing complexity captured by *SCC, SCC_{RY}* and *GS* metrics cannot be explained, contrary to Gould (1996), as a passive tendency to increase. The three proofs carried out in order to demonstrate whether positive trends are passive or driven show us that the positive trend is driven and is mainly due to the action of natural selection.

As stated by Day (2012) (see also Corominas-Murtra et al., 2018), a necessary condition for progressive and open-ended evolution is the existence of a parameter (metric) that increases with the evolutionary age of the corresponding organisms. This is what we report here for the case of several metrics, which increase during the evolution of Cyanobacteria.

Materials and methods

Phylogenetic Analysis

Ninety-one complete and nearly-complete cyanobacterial genomes were downloaded from GenBank and annotated using Prokka (Seemann, 2014) (see Supplementary file 1). To infer a phylogenomic tree we proceeded first to identify the set of homologous gene families conserved among Cyanobacteria (core genome) using get_homologues.pl pipeline (Contreras-Moreira and Vinuesa, 2013). For this, we used BDBH and OMCL methodologies within get_homologues.pl with the following parameters: a threshold e-value $\leq 10{-}^{10}$ for BLAST searches; a minimum percent amino acid identity > 30% between query and subject sequences; and for OMCL, we set the inflation parameter (I) set to 2.0. The consensus core-genome was inferred by the intersection of BDBH and OMCL gene families. To select high-quality phylogenetic markers from the core-genome (i.e. those gene families not showing recombination and/or horizontal gene transfer), we used the software package get_phylomarkers (Vinuesa *et al.,* 2018). By this procedure, we obtained an alignment of 96 top markers. The multiple alignment was cured by eliminating uninformative sites and misaligned positions with Gblocks (Talavera and Castresana, 2007). Finally, a maximum likelihood phylogeny was reconstructed using PhyML (Guindon and Gascuel, 2003) with LG model + I (estimation of invariant sites) + G (gamma distribution) as selected by prottest3 (Darriba *et al.,* 2011). For branch support we used SH statistic within PhyML.

Genome complexity metrics

SCC. Sequence Compositional Complexity of genomes was calculated by using a two-step process. We first obtained the non-overlapping compositional domains comprising the genome sequence, and then applied an entropic complexity measurement able to account for the heterogeneity of such compositional domains. The compositional domains of a given genome sequence are obtained through a segmentation algorithm that was properly designed (Bernaola-Galván *et al*., 1996) by using the Jensen-Shannon entropic divergence (Grosse *et al*., 2002; Lin, 1991) to split the sequence -and iteratively the sub-sequences- into non-overlapping compositional domains which, at a given statistical significance, *s*, are homogeneous and compositionally different from the neighboring domains. Note that this algorithm does not use any scanner window, thus avoiding introducing an artificial parameter. Note also that the statistical significance level *s*, is the probability that the difference between each pair of adjacent domains is not due to statistical fluctuations. By changing this parameter one can obtain the underlying distribution of segment lengths and nucleotide compositions at different levels of detail (Li, 1997), thus fulfilling one of the key requirements for complexity measures (Gell-Mann and Lloyd, 1996). Recent improvements to this segmentation algorithm also allow to segment long-range correlated sequences (Bernaola-Galván *et al.*, 2012).

Once a genome sequence was segmented into *n* compositional domains, we computed *SCC* as:

$$SCC = H(S) - \sum_{i=1}^{n} \frac{G_i}{G} H(S_i)$$

where *S* denotes the whole genomes and *G* its length, G_i the length of the *i* th domain, S_i. $H(\cdot) = -\sum f log_2 f$ is the Shannon entropy of the distribution of relative frequencies of symbol occurrences {*f*} in the corresponding (sub)sequence (Román-Roldán *et al.*, 1998). It should be noted that the above expression is the same one than used in the segmentation process, applying it to the tentative two new subsequences (*n* = 2) to be obtained in each step. Thus, the two parts of the *SCC* computation are based on the same theoretical background.

We apply the above two-step procedure to each of the entire four-symbol cyanobacterial genomes, thus obtaining a *SCC* complexity value for each of them. In addition, we also apply the same procedure to the binary sequences resulting from grouping the four nucleotides into S(C,G) vs. W(A,T) or R(A,G) vs. Y (T,C), or K(A,C) vs. M(T,G), then obtaining SCC_{SW}, SCC_{RY} and SCC_{KM} metrics, respectively. These three additional metrics are partial complexities providing complementary views of genome complexity to that obtained with the four-symbol sequence (Bernaola-Galván *et al*., 1999; Bernaola-Galván *et al*., 2004).

GS. The Chaos Game Representation (*CGR*, Jeffrey, 1990; Almeida *et al*., 2001) is an image derived from a genome where each point of the image corresponds to a given *k-mer* level of analysis. If the genome sequence is a random collection of bases, the *CGR* will be a uniformly filled square image. On the bases of building a *CGR* for a particular genome, we define a corresponding genomic signature (*GS*) that is a numerical value obtained for a particular *k-mer* level by comparing point-by-point the difference between the *CGR*'s of a real genome and a random genome of the same length. In order to make it comparable, the pixel values of the images are normalized. As stated, the size of the images generated depends on the *k-mer* used. For a given *k-mer*, we have 4^k different words and the corresponding image 4^k pixels too. To build a frequency table for each *k-mer* minus the expected frequency for a random genome is equivalent to the difference between the *CGR* images of a real and a random genome. In fact, if *G* is the size of the genome to analyze, the expected value (*EV*) for a given *k-mer* is given by $EV=G/(4^k)$. This value is used to normalize to 1 the values of the *k-mers* obtained for each of the genomes analyzed. We then define the *GS* as:

$$GS = max_k \sum_{\mathrm{i}=0}^{4^{\exp(k)}} \left|\frac{P_i}{EV} - 1\right|$$

where P_i is the relative frequency of the *k-mer i*.

BB. The biobit is a metric of genome complexity that is derived from the comparison between the *k-mer* that yields the maximum entropy of a given random genome and the corresponding entropy of the real genome of the same length (Bonnici and Manca, 2016). The authors demonstrated that the entropy of a real genome of length *G*, $E_{2L(G)}$ takes a value between the maximum ($2log_4(G)$ or $2L(G)$) and the minimum ($L(G)$) entropy. On the other hand, Bonnici and Manca (2016) define and measure two additional components, that they call entropic ($E(G)$) and anti-entropic ($A(G)$) of a real genome, in such a way that $A(G) + E(G) = L(G)$. Then, the entropy of those components are given by $E(G) = E_{2L(G)} - 2L(G)$ and $A(G) = 2L(G) - E_{2L(G)}$, respectively.

Finally, the biobit *BB* of a genome ($BB(G)$), is a non-linear combination of the two entropic and anti-entropic components given by:

$$BB(G) = \sqrt{L(G)}\sqrt{\frac{A(G)}{L(G)}}\left(1 - 2\frac{A(G)}{L(G)}\right)^3 ,$$

where $\frac{A(G)}{L(G)}$ is the anti-entropic fraction of the genome and $1 - 2\frac{A(G)}{L(G)}$ is the corresponding entropic fraction. Both components vary between 0 and 1.

Standard genome parameters. Finally, we have also included three standard genome parameters: genome size, %*GC* and number of genes.

Phylogenetic signal

We used the phylogenetic tree of Cyanobacteria to test the existence of a phylogenetic signal in the genome complexity metrics and genome parameters through Blomberg *et al.* (2003) *K*-statistic in the picante package for *R* (Kembel *et al.*, 2010). *K* ranges from 0 to ∞. *K* values significantly higher than zero are indicative of the presence of a phylogenetic signal or, in other words, that closely related species resemble more in the study trait than expected by chance. $K = 1$ is the value expected under Brownian evolution.

Phylogenetic correlations

We have examined the correlation between genome parameters and metrics of genome complexity after correcting the phylogenetic signal. Pearson *r* value between variables was computed as the phylogenetic trait variance-covariance matrix between two variables and significance tested against a *t*-distribution with *n-2* degrees of freedom. We used the R code provided by Liam Revell to perform Pearson correlation with phylogenetic data (http://blog.phytools.org/2017/08/pearson-correlation-with-phylogenetic.html). The *p*-value obtained with this procedure is the same as that provided by a phylogenetic generalized linear square model. As we run multiple phylogenetic correlations, we corrected *p*-values by false discovery rates.

Evolutionary trends

We tested the existence of an evolutionary trend in the genomic complexity measures and genome parameters by fitting a ridge regression of each of these genomic values against evolutionary age (i.e., tip-to-root or node-to-root distances). Significance was tested against 10,000 slopes obtained under Brownian motion simulations with the help of the *search.trend* function in the RRphylo package for R (Castiglione *et al.,* 2019).

Acknowledgements

This work was supported by grants from the Spanish Minister of Science, Innovation and Universities (former Spanish Minister of Economy and Competitiveness) to AM (project SAF2015-65878-R), JLO (project AGL2017-88702-C2-2-R) and AL (project PGC2018-099344-B-I00), grant from Generalitat Valenciana to AM (project Prometeo/2018/A/133), and co-financed by the European Regional Development Fund (ERDF). This project was also supported by a Fulbright fellowship (Spanish Minister of Science, Innovation and Universities) to AM for a sabbatical leave at Harvard University.

References

Adami C. 2002. What is complexity? *BioEssays* **24**:1085-1094.

Adami, C. 2016. What is information? *Phil. Trans. R. Soc. A* **374**:20150230.

Adami, C, Cerf NJ. 2000. Physical complexity of symbolic sequences. *Phys. D. Nonlinear. Phenom.* **137**:62-69.

Almeida JS, Carriço JA, Maretzek A, Noble PA, Fletcher M. 2001. Analysis of genomic sequences by Chaos Game Representation. *Bioinformatics* **17**:420-437.

Batut B, Knibbe C, Marais G, Daubin V. 2014. Reductive genome evolution at both ends of the bacterial population size spectrum. *Nature Reviews Microbiology* **12**:841-850.

Bekker A, Holland HD, Wang PL, Rumble D 3rd, Stein HJ, Hannah JL, Coetzee LL, Beukes NJ. 2004. Dating the rise of atmospheric oxygen. *Nature* **427**:117-1200.

Bernaola-Galván P, Oliver JL, Carpena P, Clay O, Bernardi G. 2004. Quantifying intrachromosomal GC heterogeneity in prokaryotic genomes. *Gene* **333**:121-133.

Bernaola-Galván P, Oliver JL, Hackenberg M, Coronado AV, Ivanov PCh, Carpena P. 2012. Segmentation of time series with long-range fractal correlations. *Eur. Phys. J. B* **85**:211.

Bernaola-Galván P, Oliver JL, Román-Roldán R. 1999. Decomposition of DNA sequence complexity. *Phys. Rev. Lett.* **83**:3336-3339.

Bernaola-Galván P, Román-Roldán R, Oliver JL. 1996. Compositional segmentation and long-range fractal correlations in DNA sequences. *Phys. Rev. E.* **53**:5181-5189.

Bernardi G. 2004. Structural and evolutionary genomics. Natural selection in genome evolution. Amsterdam: Elsevier.

Bernardi G. 2015. Chromosome architecture and genome organization. *PLoS One* **10**:e0143739.

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F. 1985. The mosaic genome of warm-blooded vertebrates. Science 228:953-958.

Blomberg SP, Garland JrT, Ives, AR. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* **57**:717-745.

Bonnici V, Manca V. 2016. Informational laws of genome structures. *Scientific Reports* **6**:28840.

Castiglione S, Serio C, Mondanaro A, Di Febbraro M, Profico A, Girardi G, Raia P. 2019. Simultaneous detection of macroevolutionary patterns in phenotypic means and rate of change with and within phylogenetic trees including extinct species. *PLoS One* **14**:e0210101.

Chaitin, GJ. 1977. Algorithmic Information *Theory. IBM J. Res. Dev.* **21**:350–359.

Contreras-Moreira B, Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol*. **79**:7696-7701.

Corominas-Murtra B, Seoane LF, Solé R. 2018. Zipf's, undounded complexity and open-ended evolution. *J.R. Soc. Interface* **15**:20180395.

Cristadoro, G., Degli Esposti, M., and Altmann, E.G. 2018. The common origin of symmetry and structure in genetic sequences. *Scientific Reports* **8**:15817.

Dagan T, Roettger M, Stucken K, Landan G, Koch R, Major, P, Gould SB, Goremykin VV, Rippka R, Tandeau de Marsac N, Gugger M, Lockhart PJ, Allen JF, Brune I, Maus I, Pühler A, Martin WF. 2013. Genomes of Stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. *Genome Biol. Evol*. **5**:31-44.

Darriba D, Taboada GL, Doallo R, Posada, D 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**:1164-1165.

Day T. 2012. Computability, Gödel's incompleteness theorem, and an inherent limit on the predictability of evolution. *J.R. Soc. Interface* **9**:624-639.

Dekker J, Belmont AS, Guttman M, Leshyk VO, Lis JT, Lomvardas S, Mirny LA, O'Shea CC, Park PJ, Ren B, Ritland Politz JC, Shendure J, Zhong S, and the 4D Nucleome Network. 2017. *Nature* **549**: 21-226.

Felsenstein, J. 1985. Phylogenies and the comparative method. *American Naturalist* **125**:1-15.

Gago S, Elena SF, Flores R, Sanjuán R. 2009. Extremely high mutation rate of a hammerhead viroid. *Science* **323**:1308.

Gell-Mann M, Lloyd S. 1996. Information measures, effective complexity, and total information. *Complexity* **2**:44-52.

Gould SJ. 1996. Full house: the spread of excellence from Plato to Darwin. Harmony Books, New York.

Grassberger P. 1986. Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.* **25**:907-938.

Grosse I, Bernaola-Galván P, Carpena P, Román-Roldán R, Oliver J, Stanley HE. 2002. Analysis of symbolic sequences using the Jensen-Shannon divergence. *Phys. Rev. E.* **65**:041905.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**:696-704.

Hedges SB, Blair JE, Venturi ML, Shoe JL. 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol Biol.* **4**:2.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol.* **2**:13-34.

Jabbari K, Bernardi G. 2017. An isochore framework underlies chromatin architecture. *PLoS One* **12**:e0168023.

Jeffrey, H.J. 1990. Chaos game representation of gene structure. *Nucleic Acids Research* **18**:2163-2170.

Karlin S, Ladunga I. 1994. Comparison of eukaryotic genomic sequences. *Proc. Natl. Acad. Sci. USA* **91**:12832-12836.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. 2010. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**:1463-1464.

Knoll AH. 2014. Paleobiological perspectives on early eukaryotic evolution. *Cold Spring Harb. Perspect. Biol.* **6:**016121.

Komárek J, Kaštovský J, Mareš J, Johansen JR. 2014. Taxonomic classification of cyanoprokaryotes (cyanobacterial genera) 2014, using a polyphasic approach. *Preslia* **86**:295-335.

Koonin EV. 2016a. The meaning of biological information. *Phil. Trans. R. Soc A* **374**:20150065.

Koonin EV. 2016. Splendor and misery of adaptation, or the importance of neutral null for understanding evolution. *BMC Biology* **14**:114.

Krakauer DC. 2011. Darwinian demons, evolutionary complexity, and information maximization. *Chaos* **21:**037110.

Li M, Vitányi P. 2008. An Introduction to Kolmogorov Complexity and Its Applications. Springer, New York.

Li W. 1997. The complexity of DNA. *Complexity* **3**:33-38.

Li W, Kaneko K. 1992. DNA correlations. *Nature* **360**:635-636.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289-293.

Lin J. 1991. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **37**:145-151.

Lynch ML. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl. Acad. Sci. USA* **104 Suppl 1**:8597-8604.

Lynch, M.L., and Connery, J.S. 2003. The origins of genome complexity. *Science* **302**:1401-1404.

McShea DN. 1993. Evolutionary change in the morphological complexity of the mammalian vertebral column. *Evolution* **47**:730-740.

McShea DN. 1994. Mechanisms of large-scale evolutionary trends. *Evolution* **48**:1747-1763.

McShea DN, Brandon RN. 2010. Biology's first law. Chicago University Press, Chicago.

Mouchiroud D, Gautier C, Bernardi G. 1988. The compositional distribution of coding sequences and DNA molecules in humans and murids. *J Mol Evol* **27**:311-320.

Moya A. 2014. The calculus of life. Springer, New York.

Oliver JL, Carpena P, Hackenberg M, Bernaola-Galván P. 2004. IsoFinder: Computational prediction of isochores in genome sequences. *Nucleic Acids Res.* **32, Suppl_2**:W287–W292.

Peng CC-KK, Buldyrev SVS, Goldberger AL, Havlin S, Sciortino F, Simons M, Stanley HE. 1992. Long-range correlations in nucleotide sequences. *Nature* **356**:168-170.

Revell, LJ. 2012. phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**:217-223.

Revell LJ, Harmon LJ, and Collar DC. 2008. Phylogenetic signal, evolutionary process, and rate. *Systematic Biology* **57**:591-601.

Rippka R. 1988. Recognition and identification of cyanobacteria. *Methods in Enzymology* **167**:28-67.

Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RY. 1979. Generic assignments, strain histories and properties of pure cultures of Cyanobacteria. *Journal of General Microbiology* **111**:1-61.

Román-Roldán R, Bernaola-Galván P, Oliver JL. 1998. Sequence compositional complexity of DNA through an entropic segmentation method. *Phys. Rev. Lett.* **80**:1344-1347.

Sagan L. 1967. On the origin of mitosing cells. *J. Theor. Biol.* **14**:255-274.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**:2068-2069.

Sergeev VN, Gerasimenko LM, Zavarzin GA. 2002. Proterozoic history and present state of cyanobacteria. *Microbiology* **71**:623-637.

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. USA* **110**:1791-1796.

Sueoka N. 1992. Directional mutation pressure, selective constraints, and genetic equilibria. *J Mol Evol*. **34**:95-114.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* **56**:564-577.

Vinuesa P, Ochoa-Sánchez LE, Contreras-Moreira B. 2018. GET_PHYLOMARKERS, a software package to select optimal orthologous clusters for phylogenomics and inferring pan-genome phylogenies, used for a critical geno-taxonomic revision of the genus Stenotrophomonas. *Frontiers in Microbiology* **9**:771.

Voss R. 1992. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys Rev Lett* **68**:3805-3808.

Zoubak S, Clay O, Bernardi G. 1996. The gene distribution of the human genome. *Gene* **174**:95-102.

Zurek W. (Ed.). 1990. Complexity, Entropy and the Physics of Information. Addison-Wesley Press, Cambridge, MA.

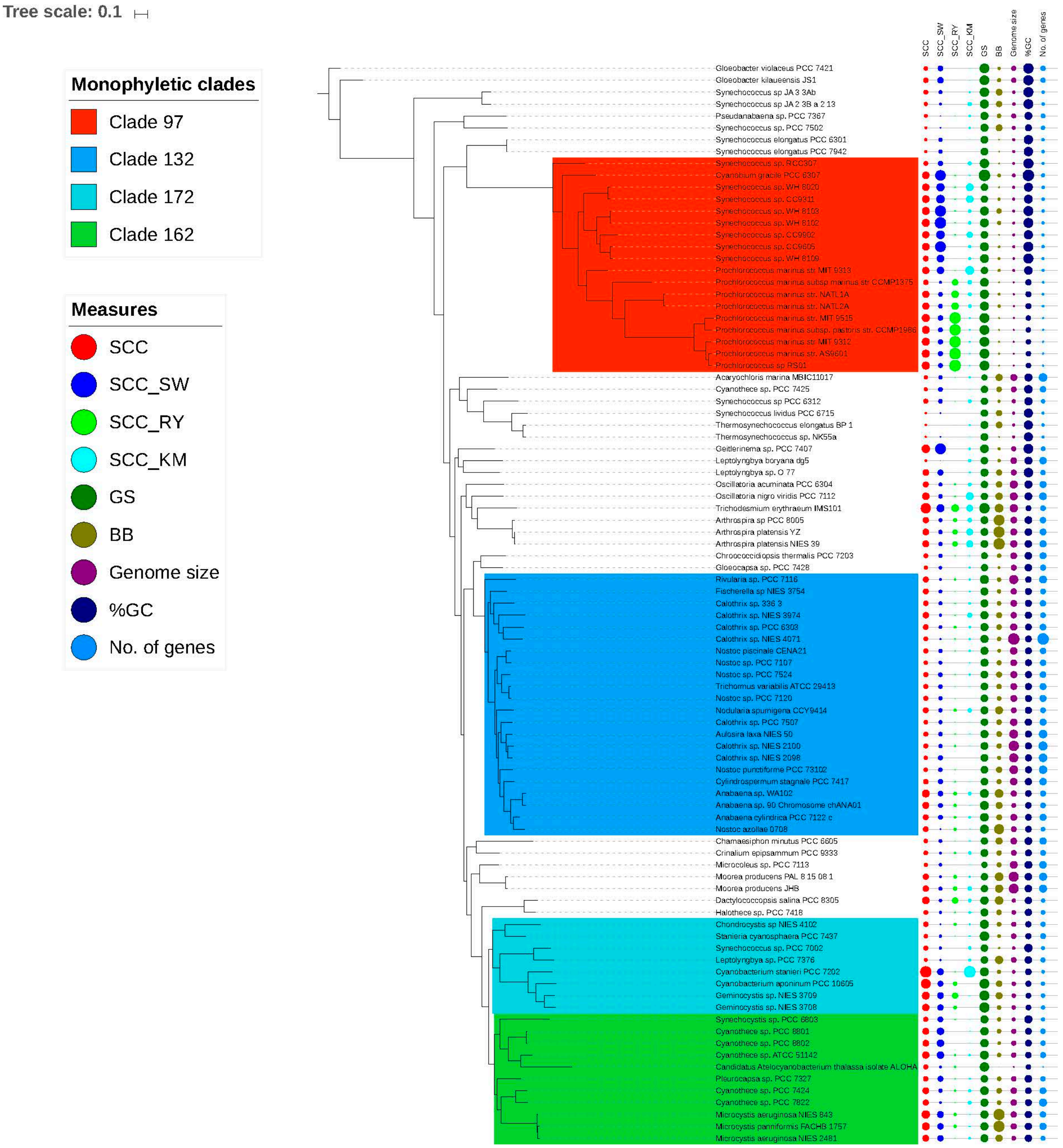
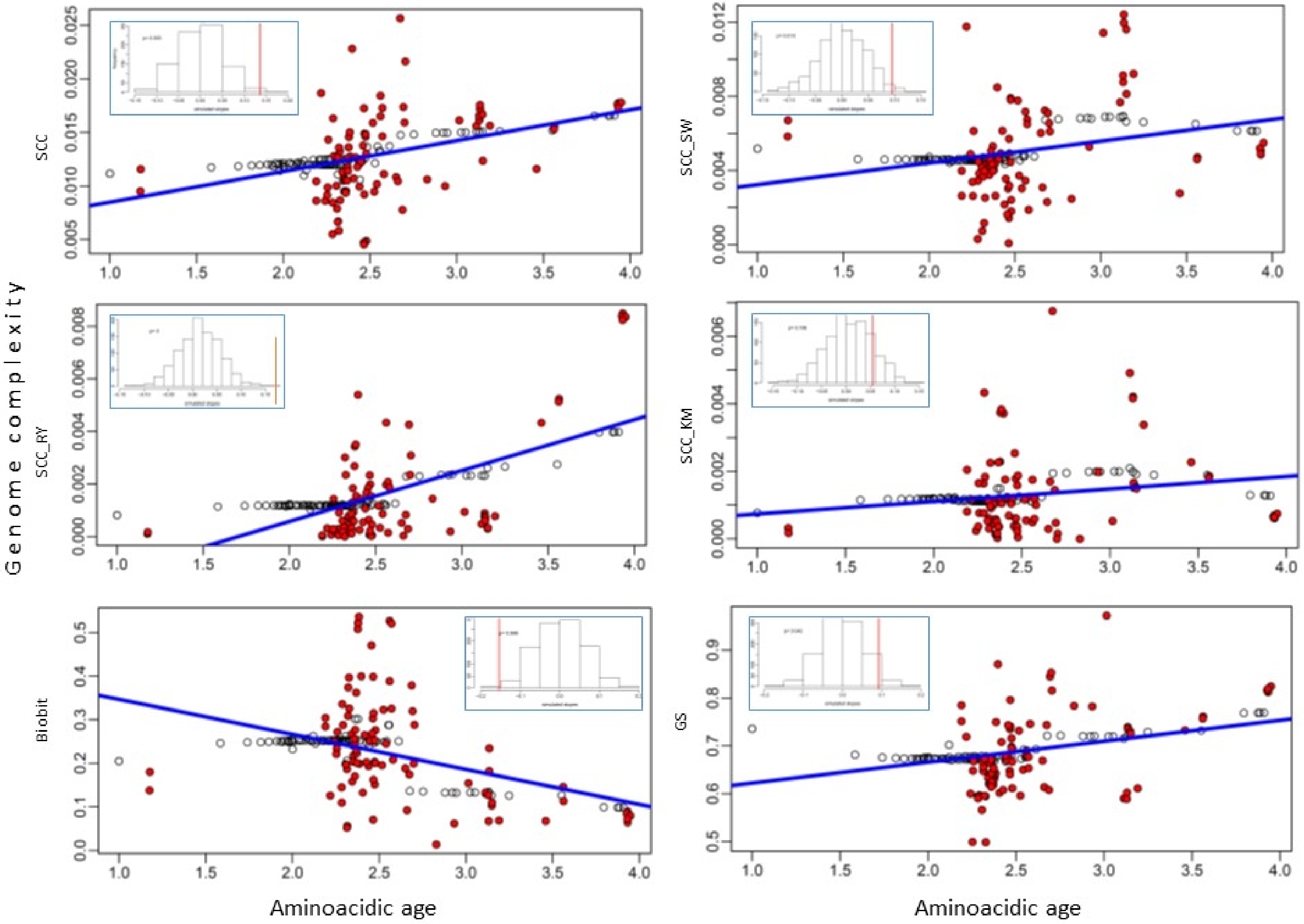

Tree scale: 0.1
Monophyletic clades
🟥 Clade 97
🟦 Clade 132
🟦 Clade 172
🟩 Clade 162
Measures
● SCC
● SCC_SW
● SCC_RY
● SCC_KM
● GS
● BB
● Genome size
● %GC
● No. of genes
SCC
SCC_SW
SCC_RY
SCC_KM
GS
BB
Genome size
%GC
No. of genes
Gloeobacter violaceus PCC 7421
Gloeobacter kilaueensis JS1
Synechococcus sp JA 3 3Ab
Synechococcus sp JA 2 3B a 2 13
Pseudanabaena sp. PCC 7367
Synechococcus sp. PCC 7502
Synechococcus elongatus PCC 6301
Synechococcus elongatus PCC 7942
Synechococcus sp. RCC307
Cyanobium gracile PCC 6307
Synechococcus sp. WH 8020
Synechococcus sp. CC9311
Synechococcus sp. WH 8103
Synechococcus sp. WH 8102
Synechococcus sp. CC9902
Synechococcus sp. CC9605
Synechococcus sp. WH 8109
Prochlorococcus marinus str MIT 9313
Prochlorococcus marinus subsp marinus str CCMP1375
Prochlorococcus marinus str. NATL1A
Prochlorococcus marinus str. NATL2A
Prochlorococcus marinus str. MIT 9515
Prochlorococcus marinus subsp. pastoris str. CCMP1986
Prochlorococcus marinus str MIT 9312
Prochlorococcus marinus str. AS9601
Prochlorococcus sp RS01
Acaryochloris marina MBIC11017
Cyanothece sp. PCC 7425
Synechococcus sp PCC 6312
Synechococcus lividus PCC 6715
Thermosynechococcus elongatus BP 1
Thermosynechococcus sp. NK55a
Geitlerinema sp. PCC 7407
Leptolyngbya boryana dg5
Leptolyngbya sp. O 77
Oscillatoria acuminata PCC 6304
Oscillatoria nigro viridis PCC 7112
Trichodesmium erythraeum IMS101
Arthrospira sp PCC 8005
Arthrospira platensis YZ
Arthrospira platensis NIES 39
Chroococcidiopsis thermalis PCC 7203
Gloeocapsa sp. PCC 7428
Rivularia sp. PCC 7116
Fischerella sp NIES 3754
Calothrix sp. 336 3
Calothrix sp. NIES 3974
Calothrix sp. PCC 6303
Calothrix sp. NIES 4071
Nostoc piscinale CENA21
Nostoc sp. PCC 7107
Nostoc sp. PCC 7524
Trichormus variabilis ATCC 29413
Nostoc sp. PCC 7120
Nodularia spumigena CCY9414
Calothrix sp. PCC 7507
Aulosira laxa NIES 50
Calothrix sp. NIES 2100
Calothrix sp. NIES 2098
Nostoc punctiforme PCC 73102
Cylindrospermum stagnale PCC 7417
Anabaena sp. WA102
Anabaena sp. 90 Chromosome chANA01
Anabaena cylindrica PCC 7122 c
Nostoc azollae 0708
Chamaesiphon minutus PCC 6605
Crinalium epipsammum PCC 9333
Microcoleus sp. PCC 7113
Moorea producens PAL 8 15 08 1
Moorea producens JHB
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
Chondrocystis sp NIES 4102
Stanieria cyanosphaera PCC 7437
Synechococcus sp. PCC 7002
Leptolyngbya sp. PCC 7376
Cyanobacterium stanieri PCC 7202
Cyanobacterium aponinum PCC 10605
Geminocystis sp. NIES 3709
Geminocystis sp. NIES 3708
Synechocystis sp. PCC 6803
Cyanothece sp. PCC 8801
Cyanothece sp. PCC 8802
Cyanothece sp. ATCC 51142
Candidatus Atelocyanobacterium thalassa isolate ALOHA
Pleurocapsa sp. PCC 7327
Cyanothece sp. PCC 7424
Cyanothece sp. PCC 7822
Microcystis aeruginosa NIES 843
Microcystis panniformis FACHB 1757
Microcystis aeruginosa NIES 2481

Figure 1. Phylogeny of Cyanobacteria with the metrics of complexity and genome parameters for each chosen genome. Values of metrics and parameters are proportional to circle size and were standardized to have a mean of zero and variance of one. The four colored boxes represent four monophyletic sub-clades that were used to test evidence of a driven trend for each sub-clade.

Panel A


Genome complexity
SCC
0.005
0.010
0.015
0.020
0.025
SCC_SW
0.000
0.004
0.008
0.012
SCC_RY
0.000
0.002
0.004
0.006
0.008
SCC_KM
0.000
0.002
0.004
0.006
Biobit
0.0
0.1
0.2
0.3
0.4
0.5
GS
0.5
0.6
0.7
0.8
0.9
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Aminoacidic age
Aminoacidic age

Panel B

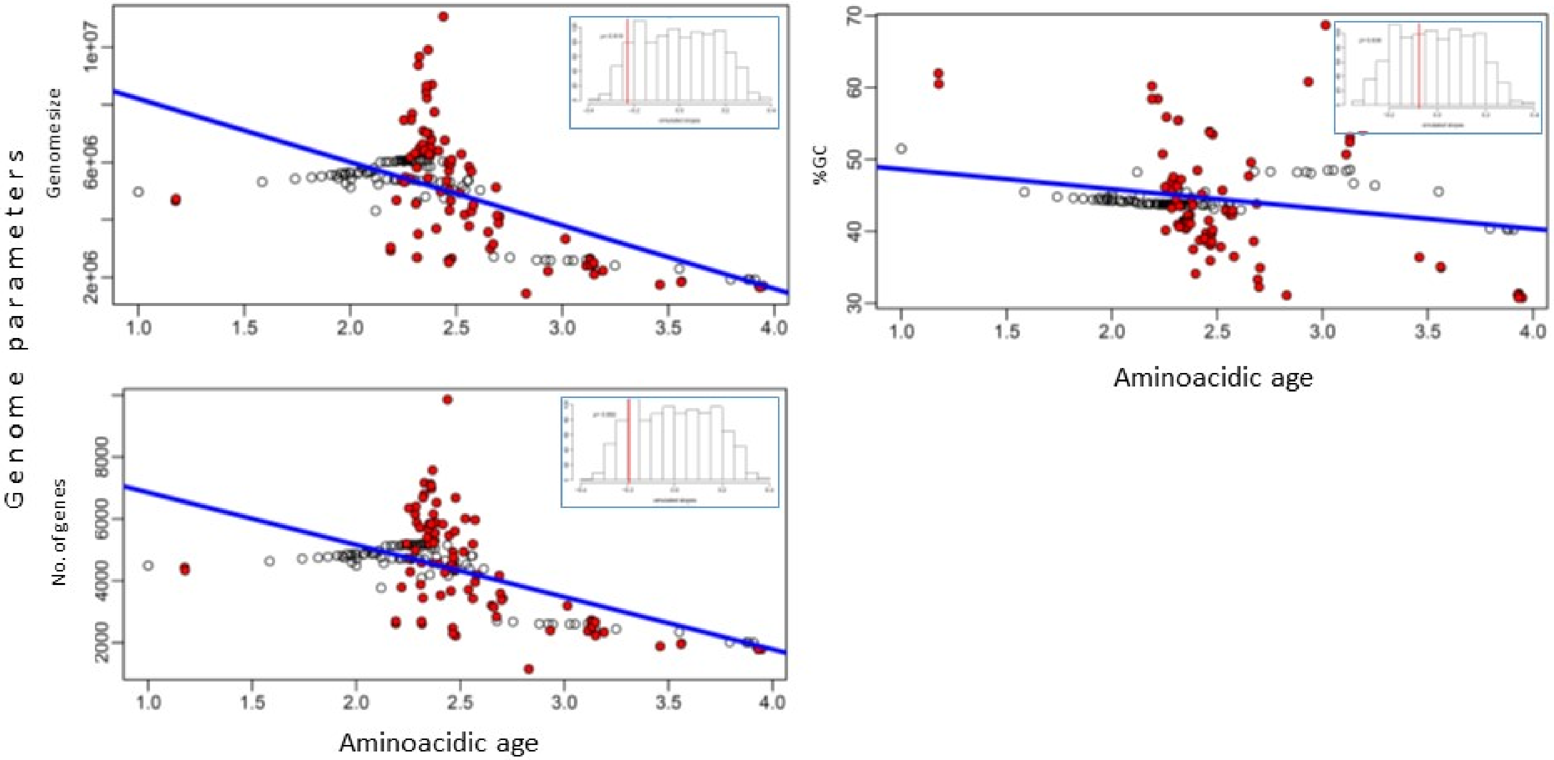
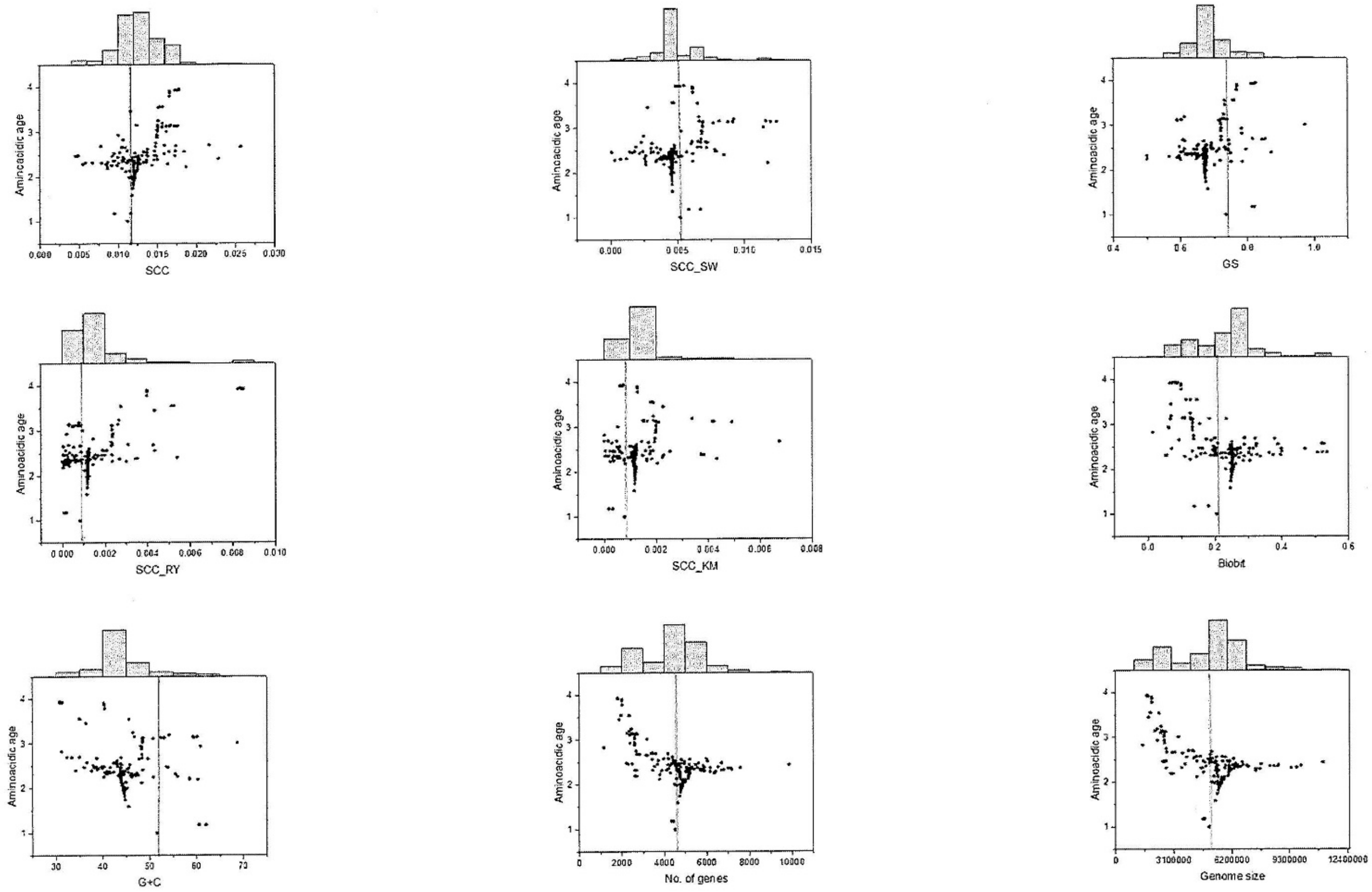

Genome parameters
Genome size
1e+07
6e+06
2e+06
1.0
1.5
2.0
2.5
3.0
3.5
4.0
%GC
70
60
50
40
30
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Aminoacidic age
No. of genes
8000
6000
4000
2000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Aminoacidic age

Figure 2. Phylogenetic trends of genomic complexity metrics (panel A) and standard genome parameters (panel B). The estimated genomic value for each tip (red circles) or node (white circles) in the phylogenetic tree is regressed against its evolutionary age (i.e., distance from the root). The statistical significance of the regression (blue lines) is tested against 10,000 slopes obtained under simulated Brownian evolution. The upper plot shows the frequency distribution of the 10,000 simulated slopes and the red vertical line shows the estimated slope. The slopes and their *p*-values are shown in Supplementary file 4.



Aminoacidic age
1
2
3
4
0.000
0.005
0.010
0.015
0.020
0.025
0.030
SCC
Aminoacidic age
1
2
3
4
0.000
0.005
0.010
0.015
SCC_SW
Aminoacidic age
1
2
3
4
0.4
0.6
0.8
1.0
GS
Aminoacidic age
1
2
3
4
0.000
0.002
0.004
0.006
0.008
0.010
SCC_RY
Aminoacidic age
1
2
3
4
0.000
0.002
0.004
0.006
0.008
SCC_KM
Aminoacidic age
1
2
3
4
0.0
0.2
0.4
0.6
Biobit
Aminoacidic age
1
2
3
4
30
40
50
60
70
G+C
Aminoacidic age
1
2
3
4
0
2000
4000
6000
8000
10000
No. of genes
Aminoacidic age
1
2
3
4
0
3100000
6200000
9300000
12400000
Genome size

Figure 3. Distribution of metrics and parameters according to amino-acidic age. The interior dashed line corresponds to the value of the basal clade, x_b. The histograms that appear above each figure correspond to the number of accumulated values of metrics and parameters (regardless of the age) ranging from lower (left) to higher (right) values than x_b.

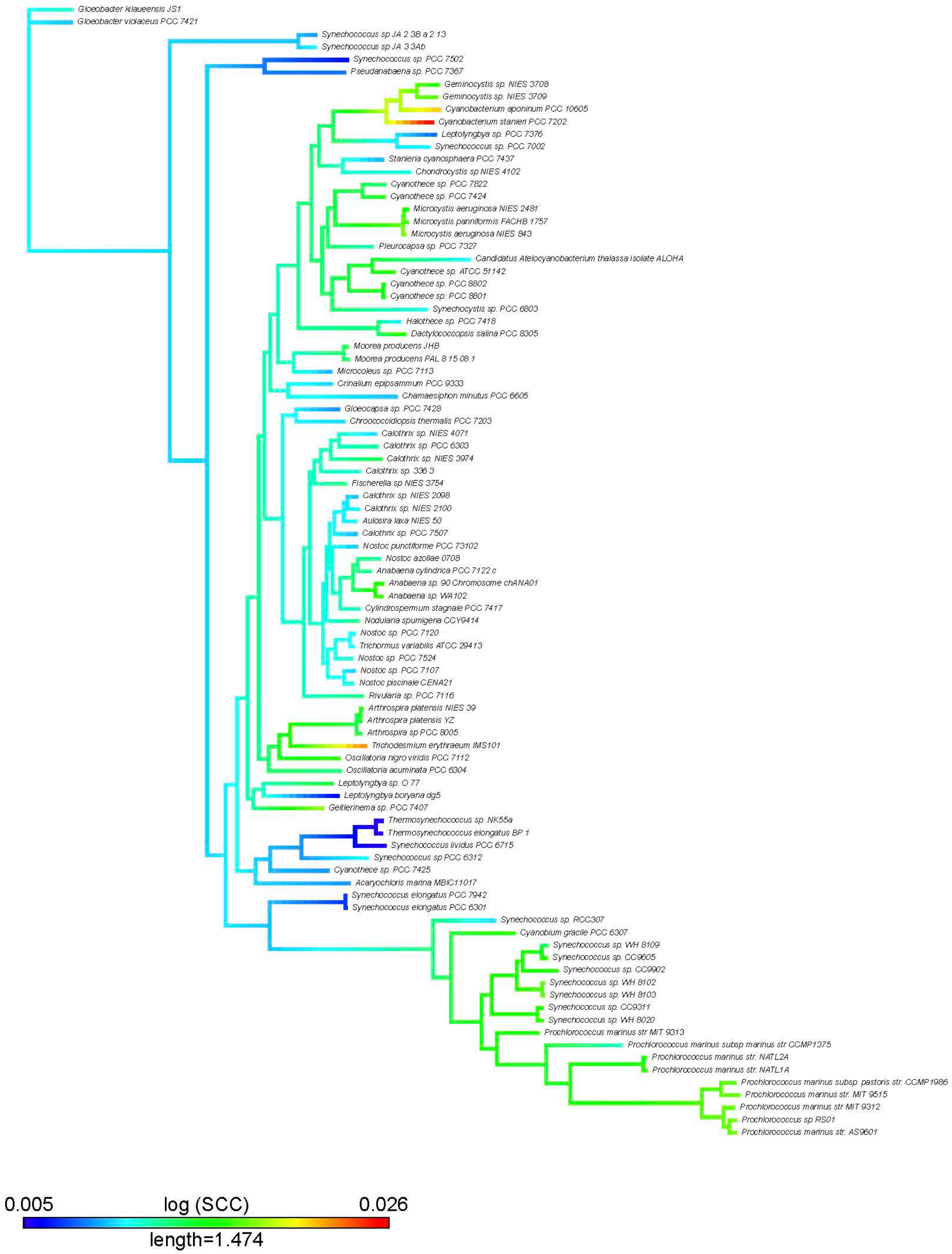

Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Synechococcus sp JA 2 3B a 2 13
Synechococcus sp JA 3 3Ab
Synechococcus sp. PCC 7502
Pseudanabaena sp. PCC 7367
Geminocystis sp. NIES 3708
Geminocystis sp. NIES 3709
Cyanobacterium aponinum PCC 10605
Cyanobacterium stanieri PCC 7202
Leptolyngbya sp. PCC 7376
Synechococcus sp. PCC 7002
Stanieria cyanosphaera PCC 7437
Chondrocystis sp NIES 4102
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Microcystis aeruginosa NIES 2481
Microcystis panniformis FACHB 1757
Microcystis aeruginosa NIES 843
Pleurocapsa sp. PCC 7327
Candidatus Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp. ATCC 51142
Cyanothece sp. PCC 8802
Cyanothece sp. PCC 8801
Synechocystis sp. PCC 6803
Halothece sp. PCC 7418
Dactylococcopsis salina PCC 8305
Moorea producens JHB
Moorea producens PAL 8 15 08 1
Microcoleus sp. PCC 7113
Crinalium epipsammum PCC 9333
Chamaesiphon minutus PCC 6605
Gloeocapsa sp. PCC 7428
Chroococcidiopsis thermalis PCC 7203
Calothrix sp. NIES 4071
Calothrix sp. PCC 6303
Calothrix sp. NIES 3974
Calothrix sp. 336 3
Fischerella sp NIES 3754
Calothrix sp. NIES 2098
Calothrix sp. NIES 2100
Aulosira laxa NIES 50
Calothrix sp. PCC 7507
Nostoc punctiforme PCC 73102
Nostoc azollae 0708
Anabaena cylindrica PCC 7122 c
Anabaena sp. 90 Chromosome chANA01
Anabaena sp. WA102
Cylindrospermum stagnale PCC 7417
Nodularia spumigena CCY9414
Nostoc sp. PCC 7120
Trichormus variabilis ATCC 29413
Nostoc sp. PCC 7524
Nostoc sp. PCC 7107
Nostoc piscinale CENA21
Rivularia sp. PCC 7116
Arthrospira platensis NIES 39
Arthrospira platensis YZ
Arthrospira sp PCC 8005
Trichodesmium erythraeum IMS101
Oscillatoria nigro viridis PCC 7112
Oscillatoria acuminata PCC 6304
Leptolyngbya sp. O 77
Leptolyngbya boryana dg5
Geitlerinema sp. PCC 7407
Thermosynechococcus sp. NK55a
Thermosynechococcus elongatus BP 1
Synechococcus lividus PCC 6715
Synechococcus sp PCC 6312
Cyanothece sp. PCC 7425
Acaryochloris marina MBIC11017
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301
Synechococcus sp. RCC307
Cyanobium gracile PCC 6307
Synechococcus sp. WH 8109
Synechococcus sp. CC9605
Synechococcus sp. CC9902
Synechococcus sp. WH 8102
Synechococcus sp. WH 8103
Synechococcus sp. CC9311
Synechococcus sp. WH 8020
Prochlorococcus marinus str MIT 9313
Prochlorococcus marinus subsp marinus str CCMP1375
Prochlorococcus marinus str. NATL2A
Prochlorococcus marinus str. NATL1A
Prochlorococcus marinus subsp. pastoris str. CCMP1986
Prochlorococcus marinus str. MIT 9515
Prochlorococcus marinus str MIT 9312
Prochlorococcus sp RS01
Prochlorococcus marinus str. AS9601
0.005
log (SCC)
0.026
length=1.474

Figure 4. Mapping of the *SCC* complexity metric on the Cyanobacteria tree. We used two functions (*contMap* and *fastAnc*) from the *phytools* R package (Revell, 2012). The *contMap* R function allows plotting a tree with a mapped continuous character, such as any of our complexity measures. Mapping is accomplished by estimating states at internal nodes using maximum likelihood with the function *fastAnc* and interpolating the states along each edge using equation 2 of Felsenstein (1985).

Table 1. Phylogenetic signals (*K*) of the six metrics of genome complexity and the three genome parameters.

Genome parameters and metrics of genome complexity	*K*	Probability, *p*
SSC	0.34	0.001
SCC_{RY}	0.66	0.001
SCC_{SW}	0.32	0.001
SCC_{KM}	0.26	0.001
BB	0.15	0.001
GS	1.00	0.001
Genome size	0.46	0.001
%GC	3.96	0.001
No. of genes	0.31	0.001

Table 2. Phylogenetic Pearson correlations (*r*) among metrics of genome complexity and genome parameters. Statistical significance was corrected by false discovery rates to control for multiple testing.

	SCC	SCC_{SW}	SCC_{RY}	SCC_{KM}	*BB*	*GS*	Genome Size	%GC
SCC_{SW}	0.66^{***}							
SCC_{RY}	0.52^{***}	0.09^{ns}						
SCC_{KM}	0.30^{**}	0.09^{ns}	0.03^{ns}					
BB	0.38^{***}	-0.02^{ns}	0.53^{***}	-0.04^{ns}				
GS	0.34^{***}	0.20^{ns}	0.41^{***}	-0.20^{ns}	0.19^{ns}			
Genome Size	0.22^{*}	0.10^{ns}	0.31^{**}	-0.05^{ns}	0.32^{**}	0.001^{ns}		
%GC	-0.06^{ns}	0.26^{*}	-0.38^{***}	-0.11^{ns}	-0.09^{ns}	-0.1^{ns}	-0.11^{ns}	
No. of genes	0.12^{ns}	0.07^{ns}	0.26^{*}	-0.09^{ns}	0.26^{*}	-0.09^{ns}	0.86^{***}	-0.09^{ns}

*** $p < 0.001$; ** $0.001 < p < 0.01$; *$0.01 < p < 0.05$; ^{ns} $p > 0.05$

Table 3. Proofs of the minimum. D'Agostino-Pearson test of skewness for the entire Phylum (n = 91) and number (*n*) of times that the metric/parameter of a given derived internal or terminal node (x_d) is higher or lower than the basal node (x_b) (Chi-square test), as well as the average absolute difference ($|x_d$-$x_b|$, Student's t-test) between nodes that are higher or lower than x_b.

Complexity measure			Higher than x_b		Lower than x_b		Chi-square test	Student's t-test
	Skewness	p-value	*n*	$\|x_d$-$x_b\|$	*n*	$\|x_d$-$x_b\|$	p-value	p-value
SCC	0.3470	2.78E-01	139	0.00265	40	0.00207	1.3659E-13	0.0848
SCC_{SW}*	0.9455	5.31E-04	48	0.00215	131	0.00108	5.5147E-10	5.8066E-07
SCC_{RY}	2.1530	9.49E-12	130	0.00115	49	0.00051	1.1410E-09	0.0031
SCC_{KM}*	1.9214	6.76E-11	138	0.00079	43	0.00035	1.6496E-12	0.0005
*BB**	0.7018	2.31E-02	116	0.07421	63	0.07290	7.4510E-05	0.4452
GS	0.6050	4.30E-02	30	0.05647	149	0.07073	5.8695E-19	0.0460
Genome Size*	0.3805	2.31E-01	112	1117595	67	1959615	0.0008	3.3185E-07
%GC*	0.6496	4.53E-02	20	6.245	159	8.705	2.7724E-25	0.0053
No. of genes*	0.3367	3.78E-01	105	796.8	74	1488.6	0.0205	1.1460E-07

*these metrics/parameters showed twice $x_d = x_b$

Table 4. Ancestor-descendent proof. For any complexity measure/genome parameter we test whether the derived clades present higher or lower values than the corresponding ancestral clade for any node.

Complexity measure	Derived nodes with a higher value than the ancestor of a given clade	Derived nodes with a lower value than the ancestor of a given clade	Fisher exact text *p*-value
SCC	36	2	0.0001
SCC_{SW}	19	9	0.2772
SCC_{RY}	15	5	0.1908
SCC_{KM}	15	15	1.0000
BB	58	32	0.0703
GS	33	5	0.0011
Genome Size	68	36	0.0018
%GC	68	8	0.0350
No. of genes	38	32	0.0143

Table 5. Summary of the results for each complexity metric and genome parameter. *K* is the phylogenetic signal. The signs "+", "-" or "0" indicate the existence of a positive, negative or no statistical evidence, respectively, for the corresponding test: the general trend, the driven trend, the three types of test of the minimum (i.e., skewness, Chi-square test and Student's t-test), the ancestor-descendant test and the trend in the case of the four sub-clades.

Complexity metric / parameter	*K*	Number of significant correlations	General trend	Driven trend						
				Minimum test			Ancestor-descendant test	Trend in the four sub-clades		
				Skewness	Chi-square test	Student's t-test		+	-	0
SCC	+	6	+	0	+	0	+	2	0	2
SCC_{SW}	+	2	+	+	-	+	0	1	2	1
SCC_{RY}	+	6	+	+	+	+	0	2	0	2
SCC_{KM}	+	1	0	+	+	+	0	0	2	2
BB	+	4	-	+	+	0	0	0	1	3
GS	+	2	+	+	-	-	+	3	0	1
Genome Size	+	4	-	+	+	-	+	0	2	2

%GC	+	2	-	0	-	-	+	0	2	2
No. of genes	+	3	-	+	+	-	+	0	2	2

Additional files

Supplementary files

Supplementary file 1. Table S1. General genome features, genome parameters, and metrics of genome complexity of Cyanobacteria genomes

Supplementary file 2. Figure S1. Cyanobacteria tree mapping the rest of the metrics and parameters. We used two functions (*contMap* and *fastAnc*) from the *phytools* R package (Revell, 2012). The *contMap* R function allows plotting a tree with a mapped continuous character, such as any of our complexity measures. The mapping is accomplished by estimating states at internal nodes using maximum likelihood with the function *fastAnc* and interpolating the states along each edge using equation 2 of Felsenstein (1985).

Supplementary file 3. Figure S2. Indication on the phylogenetic tree of the Cyanobacteria the location of the four monophyletic clades (97, 132, 162 and 172) where evolutionary trends of metrics and genome parameters were evaluated.

Supplementary file 4. Table S2. Ridge regression of genome complexity metrics and genome parameters versus age (distance from the root) in the phylum and the four selected monophyletic clades.

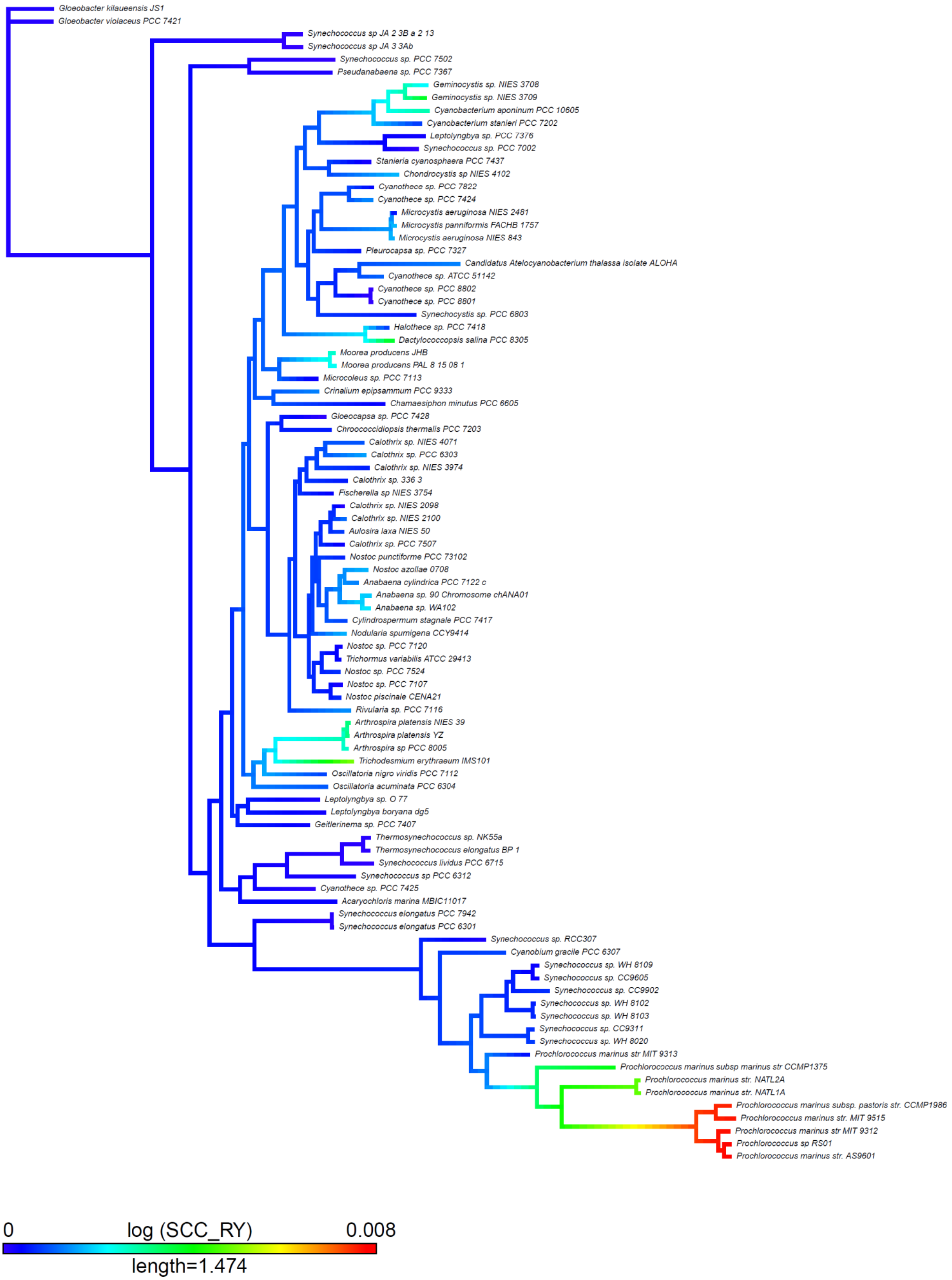
Supplementary file 5. Figure S3. Phylogenetic trends of genomic complexity metrics and standard genome parameters in the clades 97, 132, 162 and 172. The estimated genomic value for each tip (red dots) or node (black dots) in the phylogenetic tree is regressed against its evolutionary age (i.e. distance from the root). The shaded area of the plot shows a 95% confidence interval of the estimated genomic values. The statistical significance of the regression is tested against 10,000 slopes obtained under simulated Brownian evolution.

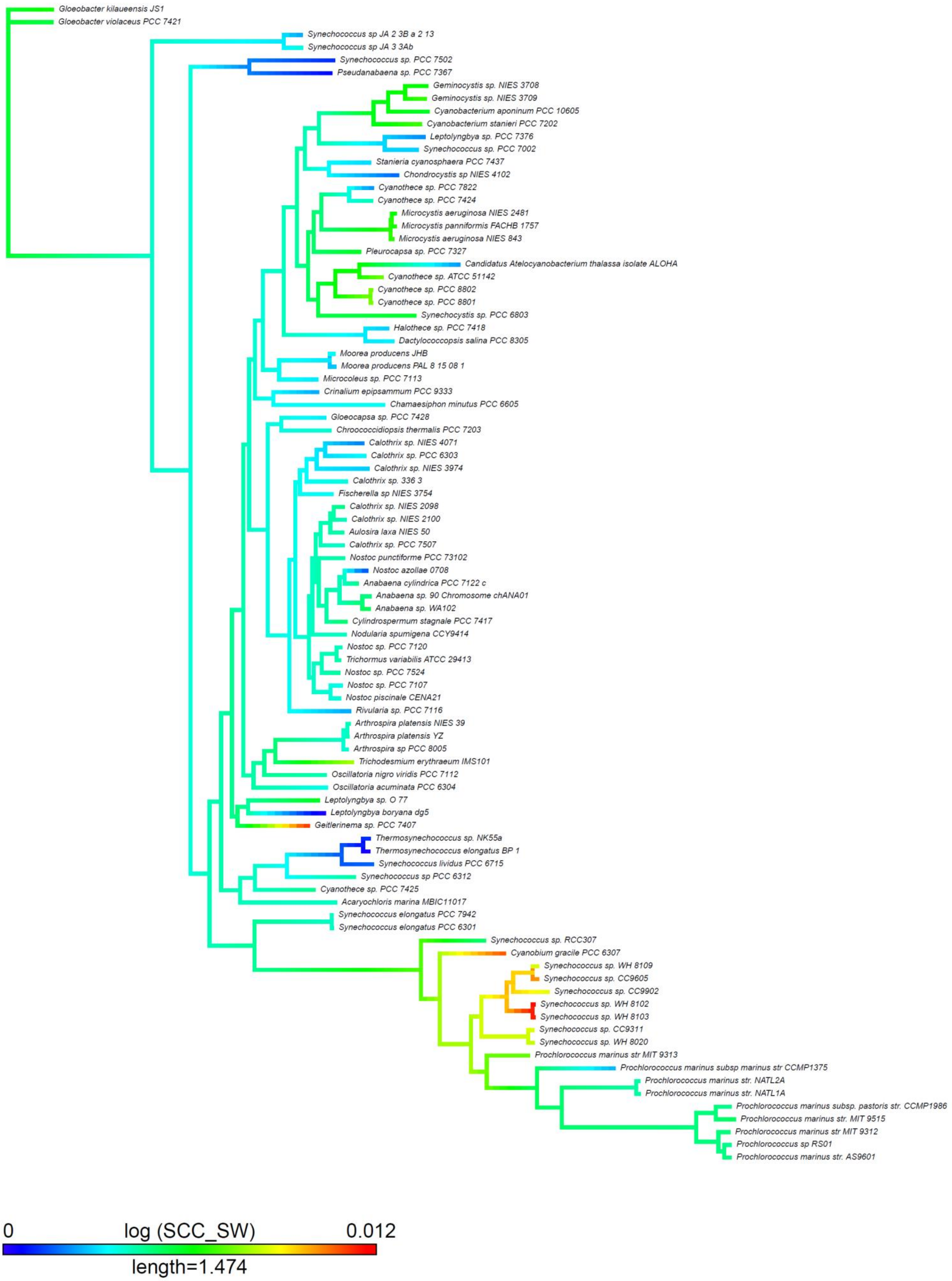
Supplementary file 6. Table S3. Sub-clade proof, second sub-clade test. Median values in the entire phylum and the four sub-clades. In the protocol adopted here, a subclade drawn from the tail is defined as a monophyletic subset chosen such that the mean fits distribution is greater than the mean of the parent distribution (McShea, 1994).

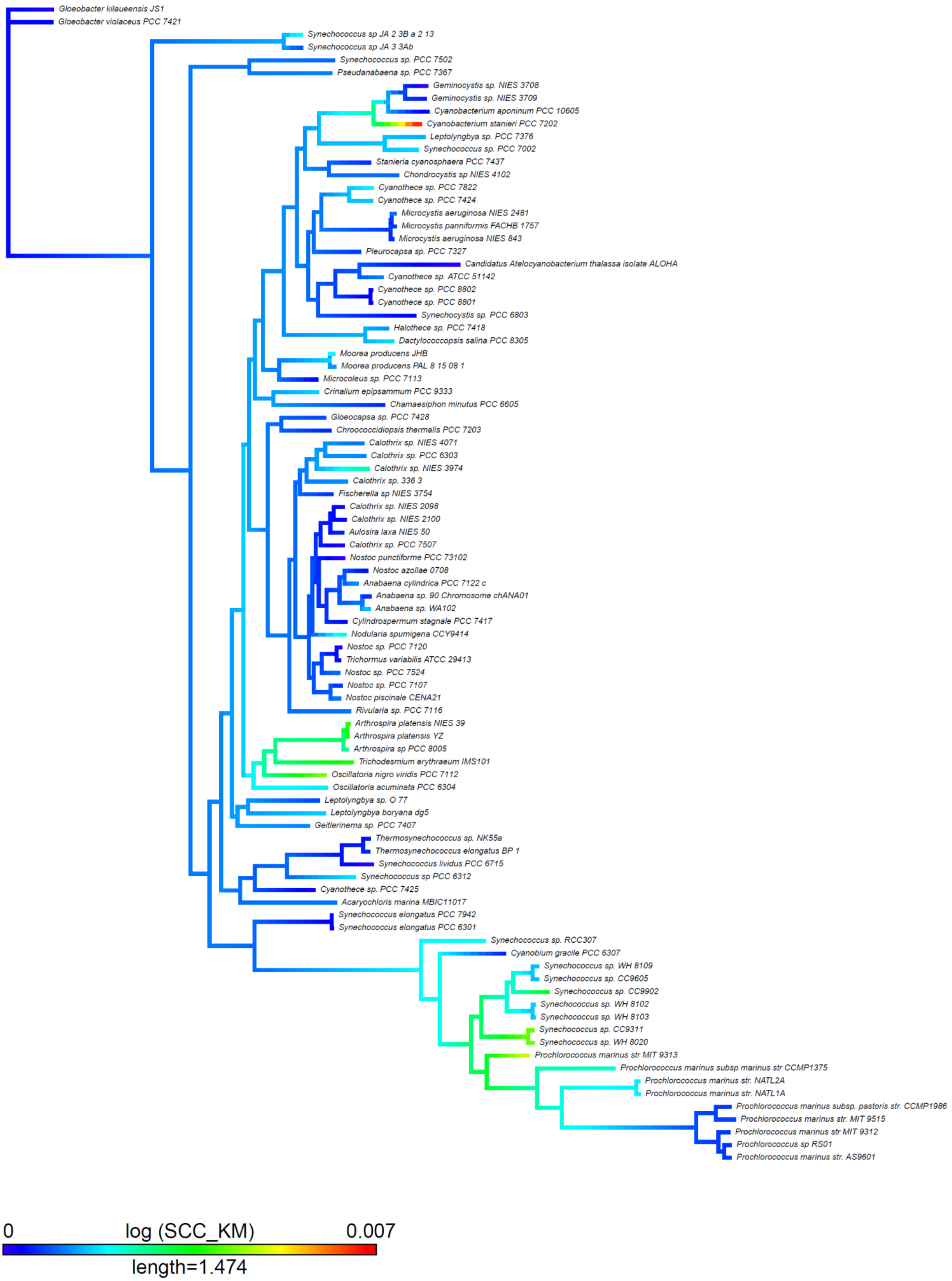
Supplementary File 1. Table S1.

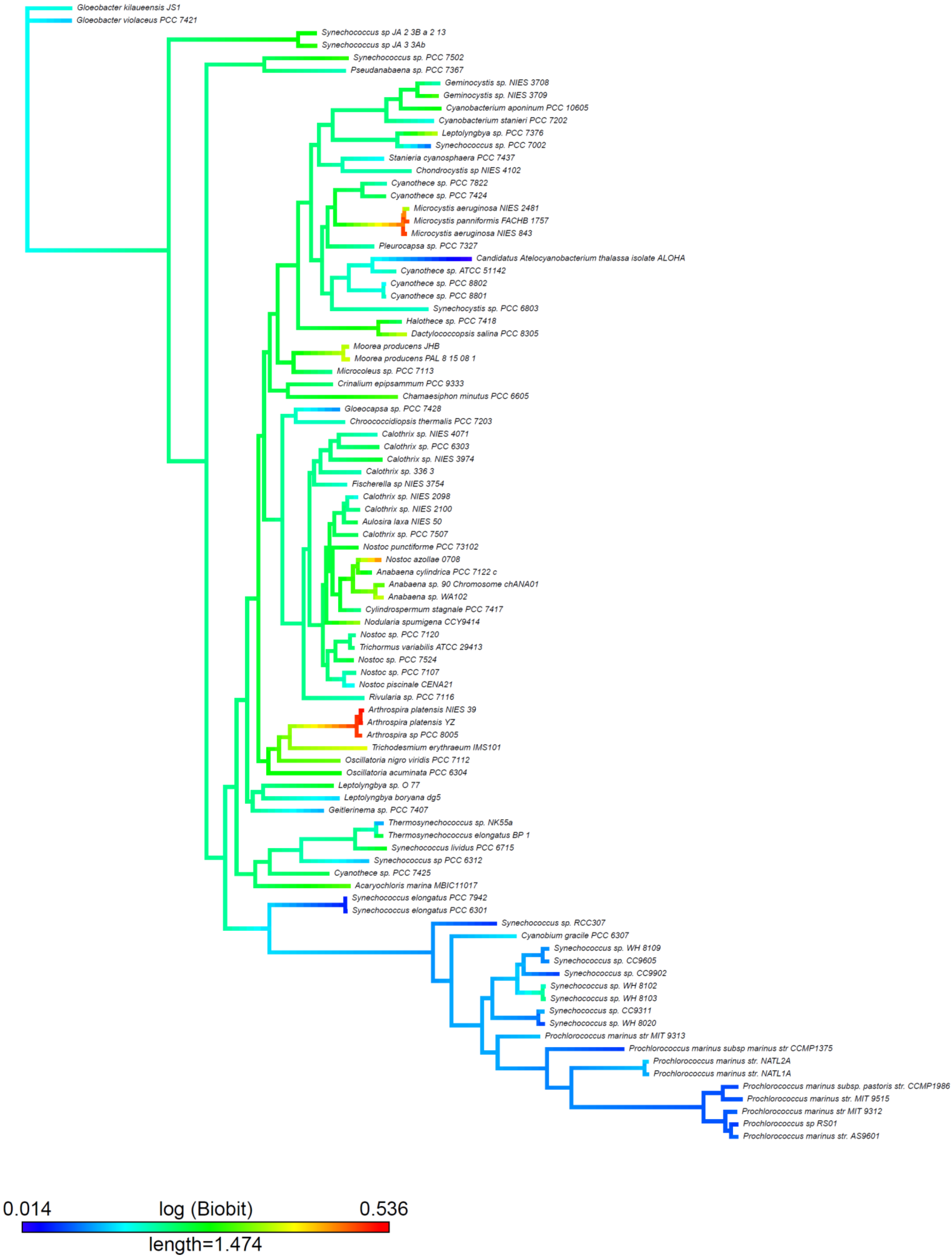
		Assembly							Genome		No. of
Species	NCBIRef	accession	SCC	SCC_SW	SCC_RY	SCC_KM	GS	BB	size	%GC	genes
Leptolyngbya_boryana_dg5	AP014642.1	GCF_002142495.1	0.0055	0.0003	0.0003	0.0016	0.5944	0.1451	6176365	46.99	6144
Calothrix_sp._NIES_2098	AP018172.1	GCF_002368175.1	0.0094	0.0051	0.0001	0.0002	0.6425	0.1706	8656060	41.20	7082
Chondrocystis_sp_NIES_4102	AP018281.1	GCF_002368355.1	0.0111	0.0019	0.0019	0.0010	0.7459	0.2086	4516676	36.53	4198
Thermosynechococcus_elongatus_BP_1	BA000039.2	GCF_000011345.1	0.0047	0.0001	0.0001	0.0006	0.6483	0.3032	2593857	53.92	2476
Gloeobacter_violaceus_PCC_7421	BA000045.2	GCF_000011385.1	0.0095	0.0058	0.0001	0.0003	0.8216	0.1380	4659019	62.00	4430
Prochlorococcus_marinus_subsp._pastoris_str._CCMP1986	BX548174.1	GCF_000011465.1	0.0173	0.0049	0.0082	0.0006	0.8197	0.0726	1657990	30.80	1790
Anabaena_sp._90_Chromosome_chANA01	CP003284.1	GCF_000312705.1	0.0149	0.0049	0.0020	0.0003	0.7406	0.3151	4329264	38.10	4531
Rivularia_sp._PCC_7116	CP003549.1	GCF_000316665.1	0.0128	0.0026	0.0015	0.0011	0.7498	0.1990	8698463	37.54	6526
Pseudanabaena_sp._PCC_7367	CP003592.1	GCF_000317065.1	0.0079	0.0007	0.0002	0.0011	0.6684	0.1972	4557046	46.31	3877
Oscillatoria_nigro_viridis_PCC_7112	CP003614.1	GCF_000317475.1	0.0163	0.0045	0.0009	0.0043	0.5948	0.3362	7479014	45.87	6408
Gloeocapsa_sp._PCC_7428	CP003646.1	GCF_000317555.1	0.0085	0.0034	0.0000	0.0008	0.5904	0.1100	5431448	43.27	4996
Cyanobacterium_stanieri_PCC_7202	CP003940.1	GCF_000317655.1	0.0257	0.0072	0.0008	0.0067	0.7436	0.1744	3163381	38.66	2837
Leptolyngbya_sp._PCC_7376	CP003946.1	GCF_000316605.1	0.0078	0.0023	0.0003	0.0014	0.6078	0.3789	5125950	43.87	4167
Microcystis_panniformis_FACHB_1757	CP011339.1	GCF_001264245.1	0.0173	0.0067	0.0021	0.0006	0.6771	0.5209	5686839	42.35	5974
Arthrospira_sp_PCC_8005	FO818640.1	GCF_000973065.1	0.0136	0.0040	0.0027	0.0023	0.6321	0.5083	6228153	44.73	5293
Synechocystis_sp._PCC_6803	NC_000911.1	GCF_001318385.1	0.0109	0.0060	0.0005	0.0003	0.6692	0.1887	3573470	47.72	3204
Nostoc_sp._PCC_7120	NC_003272.1	GCF_000009705.1	0.0107	0.0043	0.0003	0.0001	0.6197	0.2037	6413771	41.35	5842
Prochlorococcus_marinus_subsp_marinus_str_CCMP1375	NC_005042.1	GCF_000007925.1	0.0116	0.0028	0.0043	0.0023	0.7325	0.0679	1751080	36.44	1882
Synechococcus_sp._WH_8102	NC_005070.1	GCF_000195975.1	0.0176	0.0124	0.0006	0.0017	0.7361	0.1826	2434428	59.41	2513
Prochlorococcus_marinus_str_MIT_9313	NC_005071.1	GCF_000011485.1	0.0156	0.0077	0.0005	0.0049	0.5902	0.1334	2410873	50.74	2369
Synechococcus_elongatus_PCC_6301	NC_006576.1	GCF_000010065.1	0.0067	0.0043	0.0000	0.0003	0.6473	0.0564	2696255	55.48	2602
Prochlorococcus_marinus_str._NATL2A	NC_007335.2	GCF_000012465.1	0.0153	0.0046	0.0051	0.0018	0.7571	0.1461	1842899	35.12	1953
Trichormus_variabilis_ATCC_29413	NC_007413.1	GCF_000204075.1	0.0105	0.0046	0.0004	0.0004	0.6206	0.2421	6365727	41.42	5676
Synechococcus_sp._CC9902	NC_007513.1	GCF_000012505.1	0.0156	0.0092	0.0008	0.0034	0.6111	0.0690	2234828	54.16	2337
Synechococcus_sp._CC9605	NC_007516.1	GCF_000012625.1	0.0167	0.0116	0.0003	0.0015	0.7302	0.1029	2510659	59.22	2665
Prochlorococcus_marinus_str_MIT_9312	NC_007577.1	GCF_000012645.1	0.0176	0.0052	0.0084	0.0007	0.8165	0.0852	1709204	31.21	1826
Synechococcus_elongatus_PCC_7942	NC_007604.1	GCF_000012525.1	0.0067	0.0043	0.0001	0.0003	0.6474	0.0523	2695903	55.47	2685
Synechococcus_sp_JA_3_3Ab	NC_007775.1	GCF_000013205.1	0.0105	0.0040	0.0000	0.0008	0.7850	0.3043	2932766	60.24	2611
Synechococcus_sp_JA_2_3B_a_2_13	NC_007776.1	GCF_000013225.1	0.0086	0.0026	0.0001	0.0020	0.7515	0.2847	3046682	58.45	2692
Trichodesmium_erythraeum_IMS101	NC_008312.1	GCF_000014265.1	0.0228	0.0085	0.0054	0.0037	0.8707	0.4000	7750108	34.14	4549
Synechococcus_sp._CC9311	NC_008319.1	GCF_000014585.1	0.0166	0.0091	0.0008	0.0042	0.5892	0.1273	2606748	52.45	2663
Prochlorococcus_marinus_str._AS9601	NC_008816.1	GCF_000015645.1	0.0178	0.0051	0.0085	0.0007	0.8132	0.0896	1669886	31.32	1784
Prochlorococcus_marinus_str._MIT_9515	NC_008817.1	GCF_000015665.1	0.0178	0.0055	0.0084	0.0007	0.8245	0.0798	1704176	30.79	1794
Prochlorococcus_marinus_str._NATL1A	NC_008819.1	GCF_000015685.1	0.0157	0.0047	0.0052	0.0019	0.7612	0.1136	1864731	34.98	1976
Synechococcus_sp._RCC307	NC_009482.1	GCF_000063525.1	0.0100	0.0053	0.0002	0.0020	0.7824	0.0620	2224914	60.84	2388
Acaryochloris_marina_MBIC11017	NC_009925.1	GCF_000018105.1	0.0087	0.0043	0.0004	0.0011	0.4988	0.3259	6503724	47.25	7163
Microcystis_aeruginosa_NIES_843	NC_010296.1	GCF_000010625.1	0.0184	0.0072	0.0020	0.0005	0.6789	0.5272	5842795	42.33	5190
Synechococcus_sp._PCC_7002	NC_010475.1	GCF_000019485.1	0.0105	0.0031	0.0000	0.0017	0.6137	0.0924	3008047	49.63	3148
Cyanothece_sp._ATCC_51142	NC_010546.1	GCF_000017845.1	0.0159	0.0078	0.0012	0.0012	0.7318	0.1957	4934271	37.88	4942
Nostoc_punctiforme_PCC_73102	NC_010628.1	GCF_000020025.1	0.0095	0.0046	0.0007	0.0000	0.6249	0.2768	8234322	41.41	6984
Cyanothece_sp._PCC_8801	NC_011726.1	GCF_000021805.1	0.0149	0.0079	0.0000	0.0004	0.6980	0.2034	4679413	39.76	4326
Cyanothece_sp._PCC_7424	NC_011729.1	GCF_000021825.1	0.0148	0.0045	0.0015	0.0016	0.7157	0.2530	5942652	38.61	5603
Cyanothece_sp._PCC_7425	NC_011884.1	GCF_000022045.1	0.0086	0.0049	0.0001	0.0002	0.6005	0.2504	5374574	50.79	5202
Cyanothece_sp._PCC_8802	NC_013161.1	GCF_000024045.1	0.0146	0.0078	0.0000	0.0002	0.6958	0.1516	4669813	39.82	4371
Candidatus_Atelocyanobacterium_thalassa_isolate_ALOHA	NC_013771.1	GCF_000025125.1	0.0106	0.0025	0.0014	0.0000	0.7835	0.0141	1443806	31.12	1148
Nostoc_azollae_0708	NC_014248.1	GCF_000196515.1	0.0123	0.0016	0.0018	0.0004	0.7431	0.4708	5354700	38.45	3669
Cyanothece_sp._PCC_7822	NC_014501.1	GCF_000147335.1	0.0131	0.0024	0.0005	0.0019	0.6672	0.2010	6091620	40.22	6683
Arthrospira_platensis_NIES_39	NC_016640.1	GCF_000210375.1	0.0150	0.0044	0.0035	0.0038	0.6427	0.5361	6788435	43.65	5872
Cyanobium_gracile_PCC_6307	NC_019675.1	GCF_000316515.1	0.0161	0.0114	0.0009	0.0005	0.9722	0.1545	3342364	68.71	3191
Nostoc_sp._PCC_7107	NC_019676.1	GCF_000316625.1	0.0095	0.0038	0.0002	0.0006	0.6642	0.2174	6329823	40.36	5192
Synechococcus_sp_PCC_6312	NC_019680.1	GCF_000316685.1	0.0107	0.0046	0.0002	0.0018	0.6020	0.1331	3697276	48.52	3528
Calothrix_sp._PCC_7507	NC_019682.1	GCF_000316575.1	0.0094	0.0042	0.0003	0.0001	0.6144	0.2470	7023215	42.25	5836
Nostoc_sp._PCC_7524	NC_019684.1	GCF_000316645.1	0.0113	0.0047	0.0008	0.0010	0.6332	0.2669	6635030	41.53	5405
Pleurocapsa_sp._PCC_7327	NC_019689.1	GCF_000317025.1	0.0118	0.0061	0.0005	0.0007	0.6080	0.2026	4986817	45.19	4271
Oscillatoria_acuminata_PCC_6304	NC_019693.1	GCF_000317105.1	0.0128	0.0038	0.0011	0.0018	0.5971	0.2879	7689443	47.60	5879
Chroococcidiopsis_thermalis_PCC_7203	NC_019695.1	GCF_000317125.1	0.0100	0.0042	0.0006	0.0005	0.5662	0.1985	6315792	44.44	5716
Chamaesiphon_minutus_PCC_6605	NC_019697.1	GCF_000317145.1	0.0095	0.0037	0.0003	0.0005	0.5962	0.3231	6284095	45.73	6013
Synechococcus_sp._PCC_7502	NC_019702.1	GCF_000317085.1	0.0058	0.0012	0.0000	0.0010	0.6733	0.3091	3510253	40.62	3442
Geitlerinema_sp._PCC_7407	NC_019703.1	GCF_000317045.1	0.0187	0.0118	0.0003	0.0012	0.7190	0.1261	4681111	58.46	3789
Microcoleus_sp._PCC_7113	NC_019738.1	GCF_000317515.1	0.0092	0.0034	0.0003	0.0004	0.4995	0.2215	7470429	46.21	6350
Stanieria_cyanosphaera_PCC_7437	NC_019748.1	GCF_000317575.1	0.0092	0.0032	0.0002	0.0006	0.7960	0.1622	5041209	35.95	4758
Calothrix_sp._PCC_6303	NC_019751.1	GCF_000317435.1	0.0123	0.0036	0.0017	0.0012	0.6702	0.2512	6767834	39.80	5465
Crinalium_epipsammum_PCC_9333	NC_019753.1	GCF_000317495.1	0.0099	0.0026	0.0014	0.0016	0.6482	0.2338	5315554	40.16	4728
Cylindrospermum_stagnale_PCC_7417	NC_019757.1	GCF_000317535.1	0.0118	0.0053	0.0009	0.0004	0.6193	0.2233	7003560	42.30	6158
Anabaena_cylindrica_PCC_7122_c	NC_019771.1	GCF_000317695.1	0.0127	0.0049	0.0014	0.0012	0.7081	0.2648	6395836	38.80	5834
Cyanobacterium_aponinum_PCC_10605	NC_019776.1	GCF_000317675.1	0.0216	0.0061	0.0031	0.0002	0.8160	0.2883	4114099	34.96	3426
Halothece_sp._PCC_7418	NC_019779.1	GCF_000317635.1	0.0102	0.0030	0.0009	0.0013	0.6501	0.2452	4179170	42.92	3710
Dactylococcopsis_salina_PCC_8305	NC_019780.1	GCF_000317615.1	0.0165	0.0035	0.0043	0.0018	0.6906	0.3888	3781008	42.44	3429
Gloeobacter_kilaueensis_JS1	NC_022600.1	GCF_000484535.1	0.0116	0.0067	0.0002	0.0002	0.8134	0.1802	4724791	60.54	4336
Thermosynechococcus_sp._NK55a	NC_023033.1	GCF_000505665.1	0.0045	0.0014	0.0001	0.0004	0.6403	0.0708	2520064	53.81	2287
Geminocystis_sp._NIES_3708	NZ_AP014815.1	GCF_001548095.1	0.0159	0.0066	0.0023	0.0000	0.8533	0.1995	3883409	32.28	3376
Geminocystis_sp._NIES_3709	NZ_AP014821.1	GCF_001548115.1	0.0174	0.0071	0.0043	0.0005	0.8448	0.3198	4150181	33.34	3587
Fischerella_sp_NIES_3754	NZ_AP017305.1	GCF_001548455.1	0.0127	0.0037	0.0004	0.0006	0.6563	0.1869	5821603	40.99	4584
Leptolyngbya_sp._O_77	NZ_AP017367.1	GCF_001548395.1	0.0143	0.0061	0.0002	0.0008	0.6679	0.2720	5480261	55.93	4291
Calothrix_sp._NIES_2100	NZ_AP018178.1	GCF_002368195.1	0.0105	0.0043	0.0013	0.0004	0.6439	0.2308	9909373	41.41	7574
Calothrix_sp._NIES_3974	NZ_AP018254.1	GCF_002368395.1	0.0138	0.0030	0.0006	0.0025	0.6759	0.2685	5985875	41.53	4529
Calothrix_sp._NIES_4071	NZ_AP018255.1	GCF_002368455.1	0.0100	0.0022	0.0011	0.0010	0.6799	0.1937	11064963	39.05	9858
Aulosira_laxa_NIES_50	NZ_AP018307.1	GCF_002368055.1	0.0105	0.0045	0.0009	0.0005	0.6713	0.2562	8460156	40.68	7124
Synechococcus_sp._WH_8109	NZ_CP006882.1	GCF_000161795.2	0.0124	0.0081	0.0003	0.0015	0.7262	0.1098	2111515	60.09	2232
Nodularia_spumigena_CCY9414	NZ_CP007203.1	GCF_000340565.2	0.0126	0.0043	0.0018	0.0023	0.6469	0.3625	5462271	41.19	4566
Calothrix_sp._336_3	NZ_CP011382.1	GCF_000734895.2	0.0114	0.0040	0.0006	0.0012	0.6534	0.2009	6283267	41.04	5232
Anabaena_sp._WA102	NZ_CP011456.1	GCF_001277295.1	0.0173	0.0057	0.0023	0.0018	0.7350	0.3991	5705437	38.39	4927
Synechococcus_sp._WH_8020	NZ_CP011941.1	GCF_001040845.1	0.0162	0.0088	0.0009	0.0042	0.6027	0.0675	2661166	53.12	2725
Nostoc_piscinale_CENA21	NZ_CP012036.1	GCF_001298445.1	0.0112	0.0041	0.0007	0.0012	0.6662	0.1581	7094556	40.54	5232
Microcystis_aeruginosa_NIES_2481	NZ_CP012375.1	GCF_001704955.2	0.0146	0.0065	0.0001	0.0012	0.6534	0.3254	4293006	42.91	3966
Arthrospira_platensis_YZ	NZ_CP013008.1	GCF_001611905.1	0.0146	0.0043	0.0034	0.0037	0.6396	0.5223	6520772	44.19	5551
Moorea_producens_PAL_8_15_08_1	NZ_CP017599.1	GCF_001767235.1	0.0132	0.0025	0.0023	0.0009	0.5947	0.3972	9673108	43.52	6792
Moorea_producens_JHB	NZ_CP017708.1	GCF_001854205.1	0.0136	0.0040	0.0030	0.0022	0.5959	0.3742	9373345	43.55	6690
Synechococcus_lividus_PCC_6715	NZ_CP018092.1	GCF_002754935.1	0.0049	0.0018	0.0000	0.0000	0.6114	0.2759	2659739	53.51	2227
Prochlorococcus_sp_RS01	NZ_CP018345.1	GCF_001989435.1	0.0177	0.0049	0.0085	0.0006	0.8123	0.0646	1657699	31.38	1793
Synechococcus_sp._WH_8103	NZ_LN847356.1	GCF_001182765.1	0.0172	0.0120	0.0007	0.0015	0.7394	0.2345	2429688	59.47	2492

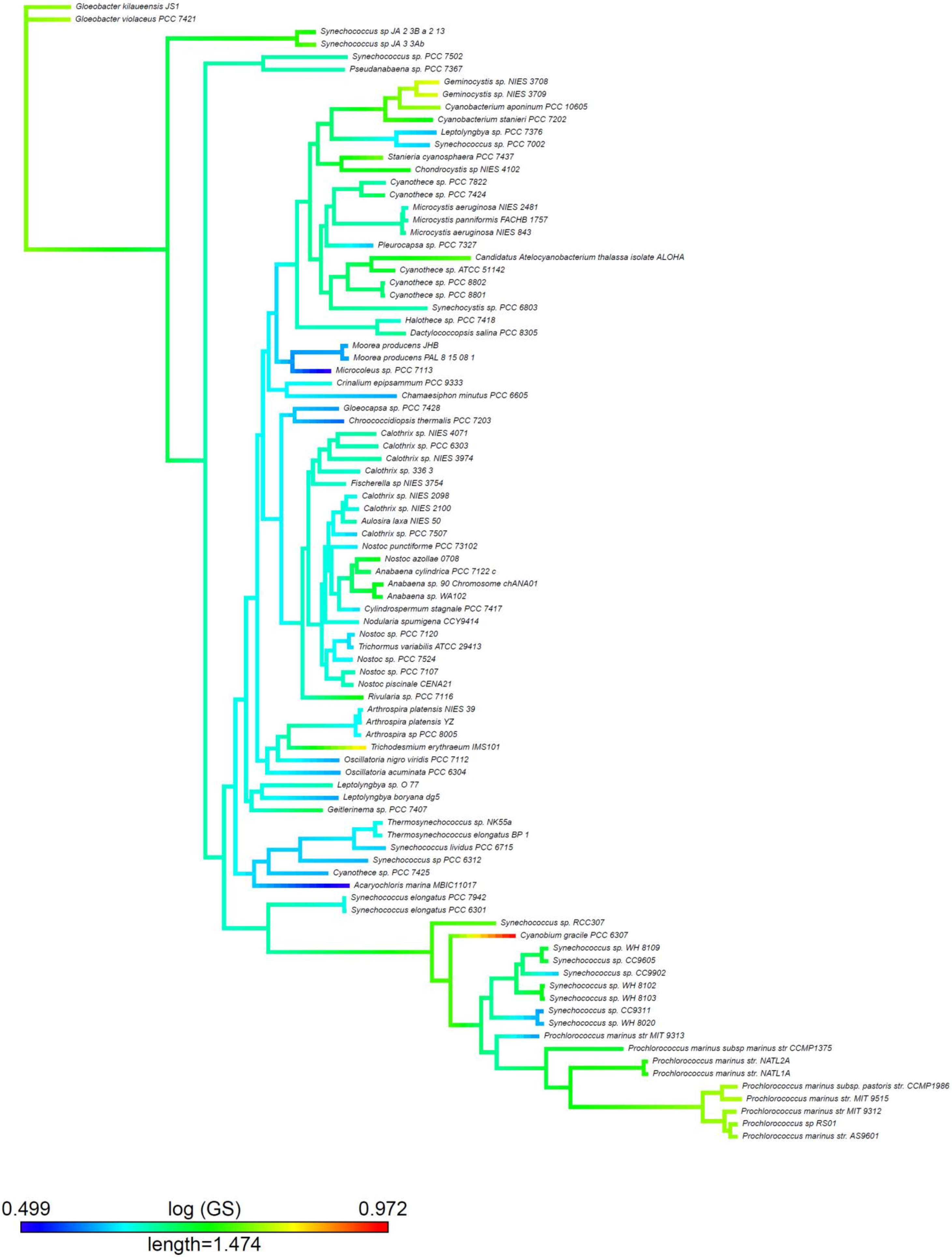
Supplementary File 2. Figure S1.

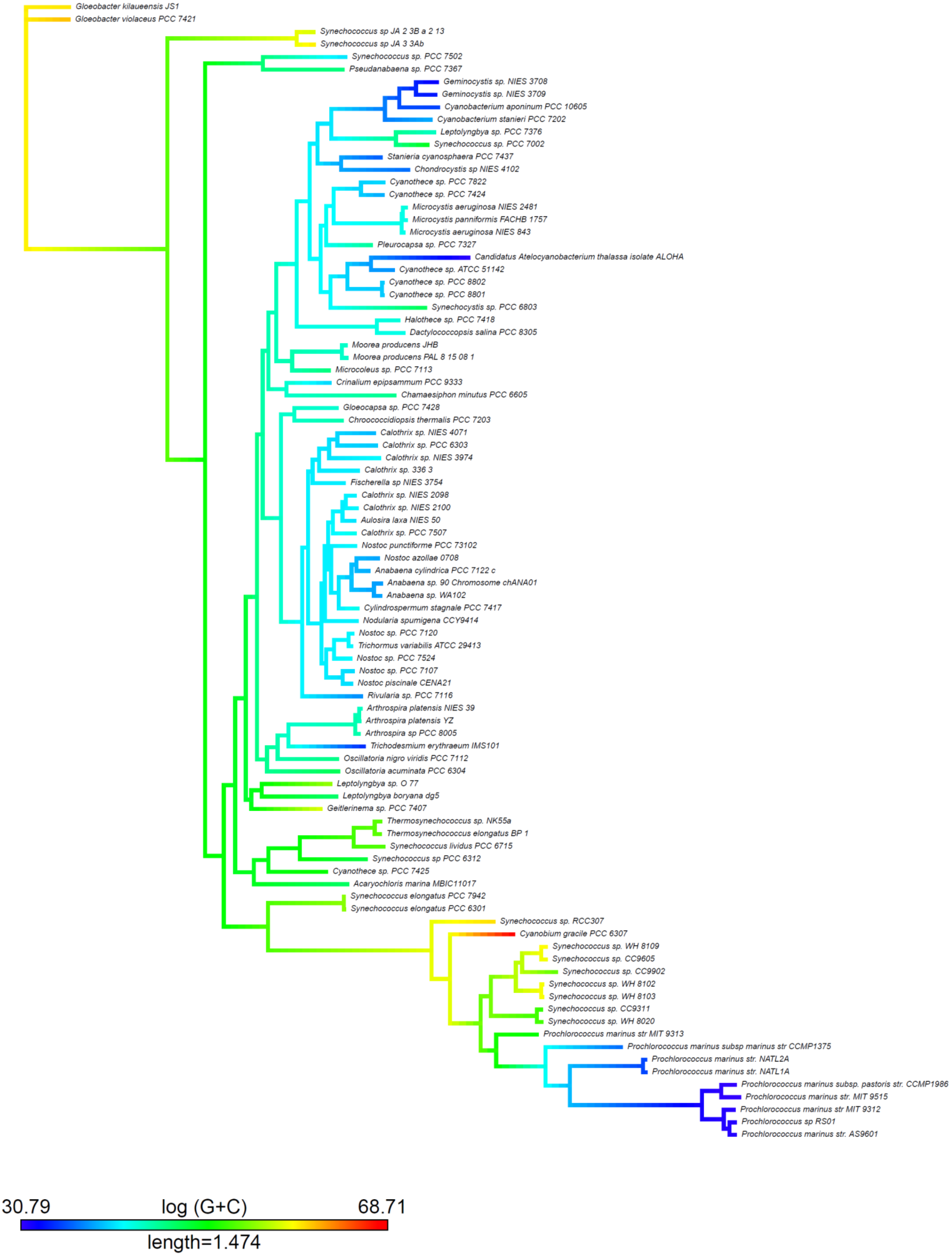

Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Synechococcus sp JA 2 3B a 2 13
Synechococcus sp JA 3 3Ab
Synechococcus sp. PCC 7502
Pseudanabaena sp. PCC 7367
Geminocystis sp. NIES 3708
Geminocystis sp. NIES 3709
Cyanobacterium aponinum PCC 10605
Cyanobacterium stanieri PCC 7202
Leptolyngbya sp. PCC 7376
Synechococcus sp. PCC 7002
Stanieria cyanosphaera PCC 7437
Chondrocystis sp NIES 4102
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Microcystis aeruginosa NIES 2481
Microcystis panniformis FACHB 1757
Microcystis aeruginosa NIES 843
Pleurocapsa sp. PCC 7327
Candidatus Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp. ATCC 51142
Cyanothece sp. PCC 8802
Cyanothece sp. PCC 8801
Synechocystis sp. PCC 6803
Halothece sp. PCC 7418
Dactylococcopsis salina PCC 8305
Moorea producens JHB
Moorea producens PAL 8 15 08 1
Microcoleus sp. PCC 7113
Crinalium epipsammum PCC 9333
Chamaesiphon minutus PCC 6605
Gloeocapsa sp. PCC 7428
Chroococcidiopsis thermalis PCC 7203
Calothrix sp. NIES 4071
Calothrix sp. PCC 6303
Calothrix sp. NIES 3974
Calothrix sp. 336 3
Fischerella sp NIES 3754
Calothrix sp. NIES 2098
Calothrix sp. NIES 2100
Aulosira laxa NIES 50
Calothrix sp. PCC 7507
Nostoc punctiforme PCC 73102
Nostoc azollae 0708
Anabaena cylindrica PCC 7122 c
Anabaena sp. 90 Chromosome chANA01
Anabaena sp. WA102
Cylindrospermum stagnale PCC 7417
Nodularia spumigena CCY9414
Nostoc sp. PCC 7120
Trichormus variabilis ATCC 29413
Nostoc sp. PCC 7524
Nostoc sp. PCC 7107
Nostoc piscinale CENA21
Rivularia sp. PCC 7116
Arthrospira platensis NIES 39
Arthrospira platensis YZ
Arthrospira sp PCC 8005
Trichodesmium erythraeum IMS101
Oscillatoria nigro viridis PCC 7112
Oscillatoria acuminata PCC 6304
Leptolyngbya sp. O 77
Leptolyngbya boryana dg5
Geitlerinema sp. PCC 7407
Thermosynechococcus sp. NK55a
Thermosynechococcus elongatus BP 1
Synechococcus lividus PCC 6715
Synechococcus sp PCC 6312
Cyanothece sp. PCC 7425
Acaryochloris marina MBIC11017
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301
Synechococcus sp. RCC307
Cyanobium gracile PCC 6307
Synechococcus sp. WH 8109
Synechococcus sp. CC9605
Synechococcus sp. CC9902
Synechococcus sp. WH 8102
Synechococcus sp. WH 8103
Synechococcus sp. CC9311
Synechococcus sp. WH 8020
Prochlorococcus marinus str MIT 9313
Prochlorococcus marinus subsp marinus str CCMP1375
Prochlorococcus marinus str. NATL2A
Prochlorococcus marinus str. NATL1A
Prochlorococcus marinus subsp. pastoris str. CCMP1986
Prochlorococcus marinus str. MIT 9515
Prochlorococcus marinus str MIT 9312
Prochlorococcus sp RS01
Prochlorococcus marinus str. AS9601
0
log (SCC_RY)
0.008
length=1.474

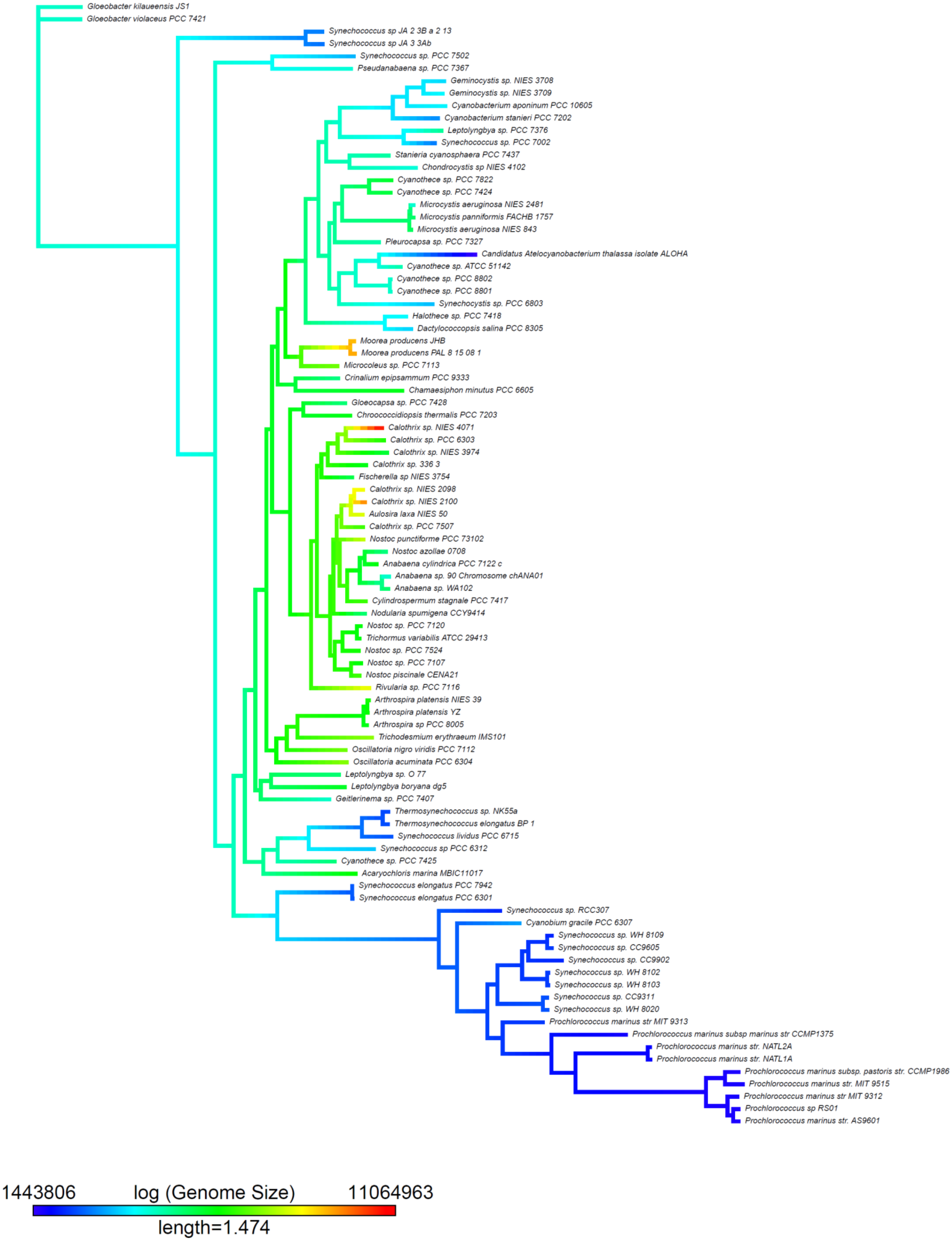

Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Synechococcus sp JA 2 3B a 2 13
Synechococcus sp JA 3 3Ab
Synechococcus sp. PCC 7502
Pseudanabaena sp. PCC 7367
Geminocystis sp. NIES 3708
Geminocystis sp. NIES 3709
Cyanobacterium aponinum PCC 10605
Cyanobacterium stanieri PCC 7202
Leptolyngbya sp. PCC 7376
Synechococcus sp. PCC 7002
Stanieria cyanosphaera PCC 7437
Chondrocystis sp NIES 4102
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Microcystis aeruginosa NIES 2481
Microcystis panniformis FACHB 1757
Microcystis aeruginosa NIES 843
Pleurocapsa sp. PCC 7327
Candidatus Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp. ATCC 51142
Cyanothece sp. PCC 8802
Cyanothece sp. PCC 8801
Synechocystis sp. PCC 6803
Halothece sp. PCC 7418
Dactylococcopsis salina PCC 8305
Moorea producens JHB
Moorea producens PAL 8 15 08 1
Microcoleus sp. PCC 7113
Crinalium epipsammum PCC 9333
Chamaesiphon minutus PCC 6605
Gloeocapsa sp. PCC 7428
Chroococcidiopsis thermalis PCC 7203
Calothrix sp. NIES 4071
Calothrix sp. PCC 6303
Calothrix sp. NIES 3974
Calothrix sp. 336 3
Fischerella sp NIES 3754
Calothrix sp. NIES 2098
Calothrix sp. NIES 2100
Aulosira laxa NIES 50
Calothrix sp. PCC 7507
Nostoc punctiforme PCC 73102
Nostoc azollae 0708
Anabaena cylindrica PCC 7122 c
Anabaena sp. 90 Chromosome chANA01
Anabaena sp. WA102
Cylindrospermum stagnale PCC 7417
Nodularia spumigena CCY9414
Nostoc sp. PCC 7120
Trichormus variabilis ATCC 29413
Nostoc sp. PCC 7524
Nostoc sp. PCC 7107
Nostoc piscinale CENA21
Rivularia sp. PCC 7116
Arthrospira platensis NIES 39
Arthrospira platensis YZ
Arthrospira sp PCC 8005
Trichodesmium erythraeum IMS101
Oscillatoria nigro viridis PCC 7112
Oscillatoria acuminata PCC 6304
Leptolyngbya sp. O 77
Leptolyngbya boryana dg5
Geitlerinema sp. PCC 7407
Thermosynechococcus sp. NK55a
Thermosynechococcus elongatus BP 1
Synechococcus lividus PCC 6715
Synechococcus sp PCC 6312
Cyanothece sp. PCC 7425
Acaryochloris marina MBIC11017
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301
Synechococcus sp. RCC307
Cyanobium gracile PCC 6307
Synechococcus sp. WH 8109
Synechococcus sp. CC9605
Synechococcus sp. CC9902
Synechococcus sp. WH 8102
Synechococcus sp. WH 8103
Synechococcus sp. CC9311
Synechococcus sp. WH 8020
Prochlorococcus marinus str MIT 9313
Prochlorococcus marinus subsp marinus str CCMP1375
Prochlorococcus marinus str. NATL2A
Prochlorococcus marinus str. NATL1A
Prochlorococcus marinus subsp. pastoris str. CCMP1986
Prochlorococcus marinus str. MIT 9515
Prochlorococcus marinus str MIT 9312
Prochlorococcus sp RS01
Prochlorococcus marinus str. AS9601
0
log (SCC_SW)
0.012
length=1.474

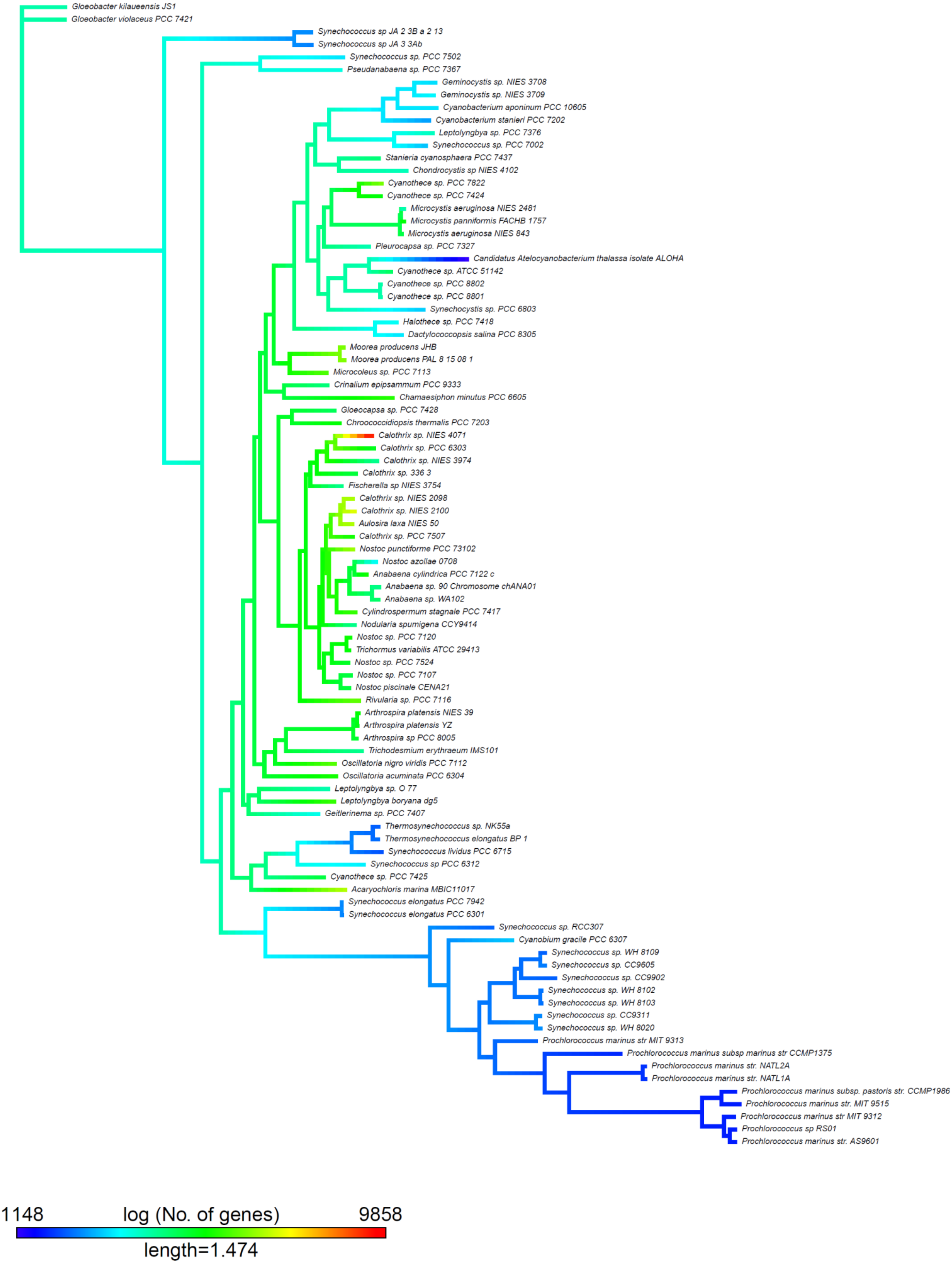

Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Synechococcus sp JA 2 3B a 2 13
Synechococcus sp JA 3 3Ab
Synechococcus sp. PCC 7502
Pseudanabaena sp. PCC 7367
Geminocystis sp. NIES 3708
Geminocystis sp. NIES 3709
Cyanobacterium aponinum PCC 10605
Cyanobacterium stanieri PCC 7202
Leptolyngbya sp. PCC 7376
Synechococcus sp. PCC 7002
Stanieria cyanosphaera PCC 7437
Chondrocystis sp NIES 4102
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Microcystis aeruginosa NIES 2481
Microcystis panniformis FACHB 1757
Microcystis aeruginosa NIES 843
Pleurocapsa sp. PCC 7327
Candidatus Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp. ATCC 51142
Cyanothece sp. PCC 8802
Cyanothece sp. PCC 8801
Synechocystis sp. PCC 6803
Halothece sp. PCC 7418
Dactylococcopsis salina PCC 8305
Moorea producens JHB
Moorea producens PAL 8 15 08 1
Microcoleus sp. PCC 7113
Crinalium epipsammum PCC 9333
Chamaesiphon minutus PCC 6605
Gloeocapsa sp. PCC 7428
Chroococcidiopsis thermalis PCC 7203
Calothrix sp. NIES 4071
Calothrix sp. PCC 6303
Calothrix sp. NIES 3974
Calothrix sp. 336 3
Fischerella sp NIES 3754
Calothrix sp. NIES 2098
Calothrix sp. NIES 2100
Aulosira laxa NIES 50
Calothrix sp. PCC 7507
Nostoc punctiforme PCC 73102
Nostoc azollae 0708
Anabaena cylindrica PCC 7122 c
Anabaena sp. 90 Chromosome chANA01
Anabaena sp. WA102
Cylindrospermum stagnale PCC 7417
Nodularia spumigena CCY9414
Nostoc sp. PCC 7120
Trichormus variabilis ATCC 29413
Nostoc sp. PCC 7524
Nostoc sp. PCC 7107
Nostoc piscinale CENA21
Rivularia sp. PCC 7116
Arthrospira platensis NIES 39
Arthrospira platensis YZ
Arthrospira sp PCC 8005
Trichodesmium erythraeum IMS101
Oscillatoria nigro viridis PCC 7112
Oscillatoria acuminata PCC 6304
Leptolyngbya sp. O 77
Leptolyngbya boryana dg5
Geitlerinema sp. PCC 7407
Thermosynechococcus sp. NK55a
Thermosynechococcus elongatus BP 1
Synechococcus lividus PCC 6715
Synechococcus sp PCC 6312
Cyanothece sp. PCC 7425
Acaryochloris marina MBIC11017
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301
Synechococcus sp. RCC307
Cyanobium gracile PCC 6307
Synechococcus sp. WH 8109
Synechococcus sp. CC9605
Synechococcus sp. CC9902
Synechococcus sp. WH 8102
Synechococcus sp. WH 8103
Synechococcus sp. CC9311
Synechococcus sp. WH 8020
Prochlorococcus marinus str MIT 9313
Prochlorococcus marinus subsp marinus str CCMP1375
Prochlorococcus marinus str. NATL2A
Prochlorococcus marinus str. NATL1A
Prochlorococcus marinus subsp. pastoris str. CCMP1986
Prochlorococcus marinus str. MIT 9515
Prochlorococcus marinus str MIT 9312
Prochlorococcus sp RS01
Prochlorococcus marinus str. AS9601
0
log (SCC_KM)
0.007
length=1.474


Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Synechococcus sp JA 2 3B a 2 13
Synechococcus sp JA 3 3Ab
Synechococcus sp. PCC 7502
Pseudanabaena sp. PCC 7367
Geminocystis sp. NIES 3708
Geminocystis sp. NIES 3709
Cyanobacterium aponinum PCC 10605
Cyanobacterium stanieri PCC 7202
Leptolyngbya sp. PCC 7376
Synechococcus sp. PCC 7002
Stanieria cyanosphaera PCC 7437
Chondrocystis sp NIES 4102
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Microcystis aeruginosa NIES 2481
Microcystis panniformis FACHB 1757
Microcystis aeruginosa NIES 843
Pleurocapsa sp. PCC 7327
Candidatus Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp. ATCC 51142
Cyanothece sp. PCC 8802
Cyanothece sp. PCC 8801
Synechocystis sp. PCC 6803
Halothece sp. PCC 7418
Dactylococcopsis salina PCC 8305
Moorea producens JHB
Moorea producens PAL 8 15 08 1
Microcoleus sp. PCC 7113
Crinalium epipsammum PCC 9333
Chamaesiphon minutus PCC 6605
Gloeocapsa sp. PCC 7428
Chroococcidiopsis thermalis PCC 7203
Calothrix sp. NIES 4071
Calothrix sp. PCC 6303
Calothrix sp. NIES 3974
Calothrix sp. 336 3
Fischerella sp NIES 3754
Calothrix sp. NIES 2098
Calothrix sp. NIES 2100
Aulosira laxa NIES 50
Calothrix sp. PCC 7507
Nostoc punctiforme PCC 73102
Nostoc azollae 0708
Anabaena cylindrica PCC 7122 c
Anabaena sp. 90 Chromosome chANA01
Anabaena sp. WA102
Cylindrospermum stagnale PCC 7417
Nodularia spumigena CCY9414
Nostoc sp. PCC 7120
Trichormus variabilis ATCC 29413
Nostoc sp. PCC 7524
Nostoc sp. PCC 7107
Nostoc piscinale CENA21
Rivularia sp. PCC 7116
Arthrospira platensis NIES 39
Arthrospira platensis YZ
Arthrospira sp PCC 8005
Trichodesmium erythraeum IMS101
Oscillatoria nigro viridis PCC 7112
Oscillatoria acuminata PCC 6304
Leptolyngbya sp. O 77
Leptolyngbya boryana dg5
Geitlerinema sp. PCC 7407
Thermosynechococcus sp. NK55a
Thermosynechococcus elongatus BP 1
Synechococcus lividus PCC 6715
Synechococcus sp PCC 6312
Cyanothece sp. PCC 7425
Acaryochloris marina MBIC11017
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301
Synechococcus sp. RCC307
Cyanobium gracile PCC 6307
Synechococcus sp. WH 8109
Synechococcus sp. CC9605
Synechococcus sp. CC9902
Synechococcus sp. WH 8102
Synechococcus sp. WH 8103
Synechococcus sp. CC9311
Synechococcus sp. WH 8020
Prochlorococcus marinus str MIT 9313
Prochlorococcus marinus subsp marinus str CCMP1375
Prochlorococcus marinus str. NATL2A
Prochlorococcus marinus str. NATL1A
Prochlorococcus marinus subsp. pastoris str. CCMP1986
Prochlorococcus marinus str. MIT 9515
Prochlorococcus marinus str MIT 9312
Prochlorococcus sp RS01
Prochlorococcus marinus str. AS9601
0.014
log (Biobit)
0.536
length=1.474

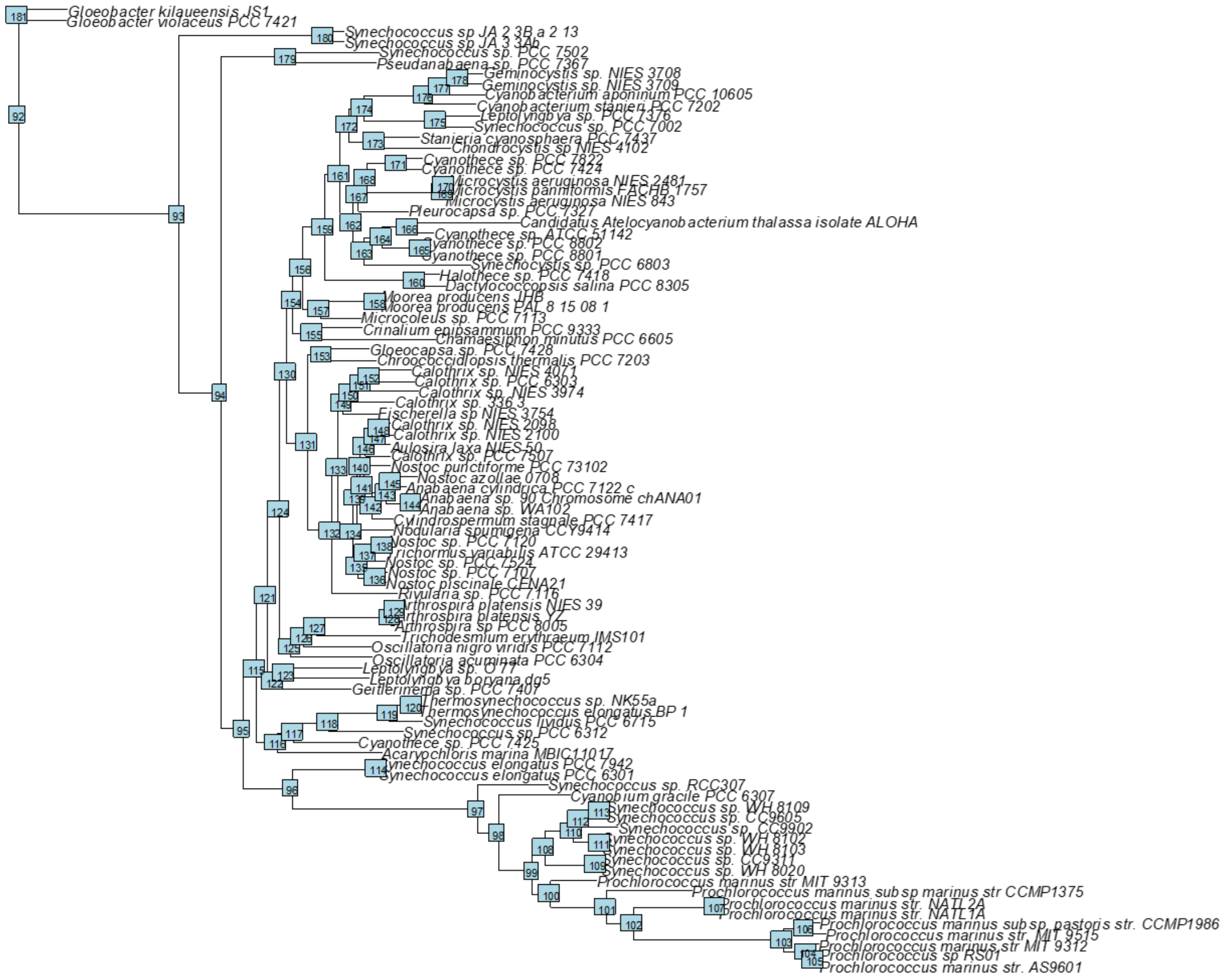

Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Synechococcus sp JA 2 3B a 2 13
Synechococcus sp JA 3 3Ab
Synechococcus sp. PCC 7502
Pseudanabaena sp. PCC 7367
Geminocystis sp. NIES 3708
Geminocystis sp. NIES 3709
Cyanobacterium aponinum PCC 10605
Cyanobacterium stanieri PCC 7202
Leptolyngbya sp. PCC 7376
Synechococcus sp. PCC 7002
Stanieria cyanosphaera PCC 7437
Chondrocystis sp NIES 4102
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Microcystis aeruginosa NIES 2481
Microcystis panniformis FACHB 1757
Microcystis aeruginosa NIES 843
Pleurocapsa sp. PCC 7327
Candidatus Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp. ATCC 51142
Cyanothece sp. PCC 8802
Cyanothece sp. PCC 8801
Synechocystis sp. PCC 6803
Halothece sp. PCC 7418
Dactylococcopsis salina PCC 8305
Moorea producens JHB
Moorea producens PAL 8 15 08 1
Microcoleus sp. PCC 7113
Crinalium epipsammum PCC 9333
Chamaesiphon minutus PCC 6605
Gloeocapsa sp. PCC 7428
Chroococcidiopsis thermalis PCC 7203
Calothrix sp. NIES 4071
Calothrix sp. PCC 6303
Calothrix sp. NIES 3974
Calothrix sp. 336 3
Fischerella sp NIES 3754
Calothrix sp. NIES 2098
Calothrix sp. NIES 2100
Aulosira laxa NIES 50
Calothrix sp. PCC 7507
Nostoc punctiforme PCC 73102
Nostoc azollae 0708
Anabaena cylindrica PCC 7122 c
Anabaena sp. 90 Chromosome chANA01
Anabaena sp. WA102
Cylindrospermum stagnale PCC 7417
Nodularia spumigena CCY9414
Nostoc sp. PCC 7120
Trichormus variabilis ATCC 29413
Nostoc sp. PCC 7524
Nostoc sp. PCC 7107
Nostoc piscinale CENA21
Rivularia sp. PCC 7116
Arthrospira platensis NIES 39
Arthrospira platensis YZ
Arthrospira sp PCC 8005
Trichodesmium erythraeum IMS101
Oscillatoria nigro viridis PCC 7112
Oscillatoria acuminata PCC 6304
Leptolyngbya sp. O 77
Leptolyngbya boryana dg5
Geitlerinema sp. PCC 7407
Thermosynechococcus sp. NK55a
Thermosynechococcus elongatus BP 1
Synechococcus lividus PCC 6715
Synechococcus sp PCC 6312
Cyanothece sp. PCC 7425
Acaryochloris marina MBIC11017
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301
Synechococcus sp. RCC307
Cyanobium gracile PCC 6307
Synechococcus sp. WH 8109
Synechococcus sp. CC9605
Synechococcus sp. CC9902
Synechococcus sp. WH 8102
Synechococcus sp. WH 8103
Synechococcus sp. CC9311
Synechococcus sp. WH 8020
Prochlorococcus marinus str MIT 9313
Prochlorococcus marinus subsp marinus str CCMP1375
Prochlorococcus marinus str. NATL2A
Prochlorococcus marinus str. NATL1A
Prochlorococcus marinus subsp. pastoris str. CCMP1986
Prochlorococcus marinus str. MIT 9515
Prochlorococcus marinus str MIT 9312
Prochlorococcus sp RS01
Prochlorococcus marinus str. AS9601
0.499
log (GS)
0.972
length=1.474


Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Synechococcus sp JA 2 3B a 2 13
Synechococcus sp JA 3 3Ab
Synechococcus sp. PCC 7502
Pseudanabaena sp. PCC 7367
Geminocystis sp. NIES 3708
Geminocystis sp. NIES 3709
Cyanobacterium aponinum PCC 10605
Cyanobacterium stanieri PCC 7202
Leptolyngbya sp. PCC 7376
Synechococcus sp. PCC 7002
Stanieria cyanosphaera PCC 7437
Chondrocystis sp NIES 4102
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Microcystis aeruginosa NIES 2481
Microcystis panniformis FACHB 1757
Microcystis aeruginosa NIES 843
Pleurocapsa sp. PCC 7327
Candidatus Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp. ATCC 51142
Cyanothece sp. PCC 8802
Cyanothece sp. PCC 8801
Synechocystis sp. PCC 6803
Halothece sp. PCC 7418
Dactylococcopsis salina PCC 8305
Moorea producens JHB
Moorea producens PAL 8 15 08 1
Microcoleus sp. PCC 7113
Crinalium epipsammum PCC 9333
Chamaesiphon minutus PCC 6605
Gloeocapsa sp. PCC 7428
Chroococcidiopsis thermalis PCC 7203
Calothrix sp. NIES 4071
Calothrix sp. PCC 6303
Calothrix sp. NIES 3974
Calothrix sp. 336 3
Fischerella sp NIES 3754
Calothrix sp. NIES 2098
Calothrix sp. NIES 2100
Aulosira laxa NIES 50
Calothrix sp. PCC 7507
Nostoc punctiforme PCC 73102
Nostoc azollae 0708
Anabaena cylindrica PCC 7122 c
Anabaena sp. 90 Chromosome chANA01
Anabaena sp. WA102
Cylindrospermum stagnale PCC 7417
Nodularia spumigena CCY9414
Nostoc sp. PCC 7120
Trichormus variabilis ATCC 29413
Nostoc sp. PCC 7524
Nostoc sp. PCC 7107
Nostoc piscinale CENA21
Rivularia sp. PCC 7116
Arthrospira platensis NIES 39
Arthrospira platensis YZ
Arthrospira sp PCC 8005
Trichodesmium erythraeum IMS101
Oscillatoria nigro viridis PCC 7112
Oscillatoria acuminata PCC 6304
Leptolyngbya sp. O 77
Leptolyngbya boryana dg5
Geitlerinema sp. PCC 7407
Thermosynechococcus sp. NK55a
Thermosynechococcus elongatus BP 1
Synechococcus lividus PCC 6715
Synechococcus sp PCC 6312
Cyanothece sp. PCC 7425
Acaryochloris marina MBIC11017
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301
Synechococcus sp. RCC307
Cyanobium gracile PCC 6307
Synechococcus sp. WH 8109
Synechococcus sp. CC9605
Synechococcus sp. CC9902
Synechococcus sp. WH 8102
Synechococcus sp. WH 8103
Synechococcus sp. CC9311
Synechococcus sp. WH 8020
Prochlorococcus marinus str MIT 9313
Prochlorococcus marinus subsp marinus str CCMP1375
Prochlorococcus marinus str. NATL2A
Prochlorococcus marinus str. NATL1A
Prochlorococcus marinus subsp. pastoris str. CCMP1986
Prochlorococcus marinus str. MIT 9515
Prochlorococcus marinus str MIT 9312
Prochlorococcus sp RS01
Prochlorococcus marinus str. AS9601
30.79
log (G+C)
68.71
length=1.474


Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Synechococcus sp JA 2 3B a 2 13
Synechococcus sp JA 3 3Ab
Synechococcus sp. PCC 7502
Pseudanabaena sp. PCC 7367
Geminocystis sp. NIES 3708
Geminocystis sp. NIES 3709
Cyanobacterium aponinum PCC 10605
Cyanobacterium stanieri PCC 7202
Leptolyngbya sp. PCC 7376
Synechococcus sp. PCC 7002
Stanieria cyanosphaera PCC 7437
Chondrocystis sp NIES 4102
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Microcystis aeruginosa NIES 2481
Microcystis panniformis FACHB 1757
Microcystis aeruginosa NIES 843
Pleurocapsa sp. PCC 7327
Candidatus Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp. ATCC 51142
Cyanothece sp. PCC 8802
Cyanothece sp. PCC 8801
Synechocystis sp. PCC 6803
Halothece sp. PCC 7418
Dactylococcopsis salina PCC 8305
Moorea producens JHB
Moorea producens PAL 8 15 08 1
Microcoleus sp. PCC 7113
Crinalium epipsammum PCC 9333
Chamaesiphon minutus PCC 6605
Gloeocapsa sp. PCC 7428
Chroococcidiopsis thermalis PCC 7203
Calothrix sp. NIES 4071
Calothrix sp. PCC 6303
Calothrix sp. NIES 3974
Calothrix sp. 336 3
Fischerella sp NIES 3754
Calothrix sp. NIES 2098
Calothrix sp. NIES 2100
Aulosira laxa NIES 50
Calothrix sp. PCC 7507
Nostoc punctiforme PCC 73102
Nostoc azollae 0708
Anabaena cylindrica PCC 7122 c
Anabaena sp. 90 Chromosome chANA01
Anabaena sp. WA102
Cylindrospermum stagnale PCC 7417
Nodularia spumigena CCY9414
Nostoc sp. PCC 7120
Trichormus variabilis ATCC 29413
Nostoc sp. PCC 7524
Nostoc sp. PCC 7107
Nostoc piscinale CENA21
Rivularia sp. PCC 7116
Arthrospira platensis NIES 39
Arthrospira platensis YZ
Arthrospira sp PCC 8005
Trichodesmium erythraeum IMS101
Oscillatoria nigro viridis PCC 7112
Oscillatoria acuminata PCC 6304
Leptolyngbya sp. O 77
Leptolyngbya boryana dg5
Geitlerinema sp. PCC 7407
Thermosynechococcus sp. NK55a
Thermosynechococcus elongatus BP 1
Synechococcus lividus PCC 6715
Synechococcus sp PCC 6312
Cyanothece sp. PCC 7425
Acaryochloris marina MBIC11017
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301
Synechococcus sp. RCC307
Cyanobium gracile PCC 6307
Synechococcus sp. WH 8109
Synechococcus sp. CC9605
Synechococcus sp. CC9902
Synechococcus sp. WH 8102
Synechococcus sp. WH 8103
Synechococcus sp. CC9311
Synechococcus sp. WH 8020
Prochlorococcus marinus str MIT 9313
Prochlorococcus marinus subsp marinus str CCMP1375
Prochlorococcus marinus str. NATL2A
Prochlorococcus marinus str. NATL1A
Prochlorococcus marinus subsp. pastoris str. CCMP1986
Prochlorococcus marinus str. MIT 9515
Prochlorococcus marinus str MIT 9312
Prochlorococcus sp RS01
Prochlorococcus marinus str. AS9601
1443806
log (Genome Size)
11064963
length=1.474


Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Synechococcus sp JA 2 3B a 2 13
Synechococcus sp JA 3 3Ab
Synechococcus sp. PCC 7502
Pseudanabaena sp. PCC 7367
Geminocystis sp. NIES 3708
Geminocystis sp. NIES 3709
Cyanobacterium aponinum PCC 10605
Cyanobacterium stanieri PCC 7202
Leptolyngbya sp. PCC 7376
Synechococcus sp. PCC 7002
Stanieria cyanosphaera PCC 7437
Chondrocystis sp NIES 4102
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Microcystis aeruginosa NIES 2481
Microcystis panniformis FACHB 1757
Microcystis aeruginosa NIES 843
Pleurocapsa sp. PCC 7327
Candidatus Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp. ATCC 51142
Cyanothece sp. PCC 8802
Cyanothece sp. PCC 8801
Synechocystis sp. PCC 6803
Halothece sp. PCC 7418
Dactylococcopsis salina PCC 8305
Moorea producens JHB
Moorea producens PAL 8 15 08 1
Microcoleus sp. PCC 7113
Crinalium epipsammum PCC 9333
Chamaesiphon minutus PCC 6605
Gloeocapsa sp. PCC 7428
Chroococcidiopsis thermalis PCC 7203
Calothrix sp. NIES 4071
Calothrix sp. PCC 6303
Calothrix sp. NIES 3974
Calothrix sp. 336 3
Fischerella sp NIES 3754
Calothrix sp. NIES 2098
Calothrix sp. NIES 2100
Aulosira laxa NIES 50
Calothrix sp. PCC 7507
Nostoc punctiforme PCC 73102
Nostoc azollae 0708
Anabaena cylindrica PCC 7122 c
Anabaena sp. 90 Chromosome chANA01
Anabaena sp. WA102
Cylindrospermum stagnale PCC 7417
Nodularia spumigena CCY9414
Nostoc sp. PCC 7120
Trichormus variabilis ATCC 29413
Nostoc sp. PCC 7524
Nostoc sp. PCC 7107
Nostoc piscinale CENA21
Rivularia sp. PCC 7116
Arthrospira platensis NIES 39
Arthrospira platensis YZ
Arthrospira sp PCC 8005
Trichodesmium erythraeum IMS101
Oscillatoria nigro viridis PCC 7112
Oscillatoria acuminata PCC 6304
Leptolyngbya sp. O 77
Leptolyngbya boryana dg5
Geitlerinema sp. PCC 7407
Thermosynechococcus sp. NK55a
Thermosynechococcus elongatus BP 1
Synechococcus lividus PCC 6715
Synechococcus sp PCC 6312
Cyanothece sp. PCC 7425
Acaryochloris marina MBIC11017
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301
Synechococcus sp. RCC307
Cyanobium gracile PCC 6307
Synechococcus sp. WH 8109
Synechococcus sp. CC9605
Synechococcus sp. CC9902
Synechococcus sp. WH 8102
Synechococcus sp. WH 8103
Synechococcus sp. CC9311
Synechococcus sp. WH 8020
Prochlorococcus marinus str MIT 9313
Prochlorococcus marinus subsp marinus str CCMP1375
Prochlorococcus marinus str. NATL2A
Prochlorococcus marinus str. NATL1A
Prochlorococcus marinus subsp. pastoris str. CCMP1986
Prochlorococcus marinus str. MIT 9515
Prochlorococcus marinus str MIT 9312
Prochlorococcus sp RS01
Prochlorococcus marinus str. AS9601
1148
log (No. of genes)
9858
length=1.474

Supplementary File 2. Figure S3.


Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Synechococcus sp JA 2 3B a 2 13
Synechococcus sp JA 3 3Ab
Synechococcus sp. PCC 7502
Pseudanabaena sp. PCC 7367
Geminocystis sp. NIES 3708
Geminocystis sp. NIES 3709
Cyanobacterium aponinum PCC 10605
Cyanobacterium stanieri PCC 7202
Leptolyngbya sp. PCC 7376
Synechococcus sp. PCC 7002
Stanieria cyanosphaera PCC 7437
Chondrocystis sp NIES 4102
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Microcystis aeruginosa NIES 2481
Microcystis panniformis FACHB 1757
Microcystis aeruginosa NIES 843
Pleurocapsa sp. PCC 7327
Candidatus Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp. ATCC 51142
Cyanothece sp. PCC 8802
Cyanothece sp. PCC 8801
Synechocystis sp. PCC 6803
Halothece sp. PCC 7418
Dactylococcopsis salina PCC 8305
Moorea producens JHB
Moorea producens PAL 8 15 08 1
Microcoleus sp. PCC 7113
Crinalium epipsammum PCC 9333
Chamaesiphon minutus PCC 6605
Gloeocapsa sp. PCC 7428
Chroococcidiopsis thermalis PCC 7203
Calothrix sp. NIES 4071
Calothrix sp. PCC 6303
Calothrix sp. NIES 3974
Calothrix sp. 336 3
Fischerella sp NIES 3754
Calothrix sp. NIES 2098
Calothrix sp. NIES 2100
Aulosira laxa NIES 50
Calothrix sp. PCC 7507
Nostoc punctiforme PCC 73102
Nostoc azollae 0708
Anabaena cylindrica PCC 7122 c
Anabaena sp. 90 Chromosome chANA01
Anabaena sp. WA102
Cylindrospermum stagnale PCC 7417
Nodularia spumigena CCY9414
Nostoc sp. PCC 7120
Trichormus variabilis ATCC 29413
Nostoc sp. PCC 7524
Nostoc sp. PCC 7107
Nostoc piscinale CENA21
Rivularia sp. PCC 7116
Arthrospira platensis NIES 39
Arthrospira platensis YZ
Arthrospira sp PCC 8005
Trichodesmium erythraeum IMS101
Oscillatoria nigro viridis PCC 7112
Oscillatoria acuminata PCC 6304
Leptolyngbya sp. O 77
Leptolyngbya boryana dg5
Geitlerinema sp. PCC 7407
Thermosynechococcus sp. NK55a
Thermosynechococcus elongatus BP 1
Synechococcus lividus PCC 6715
Synechococcus sp PCC 6312
Cyanothece sp. PCC 7425
Acaryochloris marina MBIC11017
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301
Synechococcus sp. RCC307
Cyanobium gracile PCC 6307
Synechococcus sp. WH 8109
Synechococcus sp. CC9605
Synechococcus sp. CC9902
Synechococcus sp. WH 8102
Synechococcus sp. WH 8103
Synechococcus sp. CC9311
Synechococcus sp. WH 8020
Prochlorococcus marinus str MIT 9313
Prochlorococcus marinus subsp marinus str CCMP1375
Prochlorococcus marinus str. NATL2A
Prochlorococcus marinus str. NATL1A
Prochlorococcus marinus subsp. pastoris str. CCMP1986
Prochlorococcus marinus str. MIT 9515
Prochlorococcus marinus str MIT 9312
Prochlorococcus sp RS01
Prochlorococcus marinus str. AS9601
181
180
179
178
177
176
175
174
173
172
171
170
169
168
167
166
165
164
163
162
161
160
159
158
157
156
155
154
153
152
151
150
149
148
147
146
145
144
143
142
141
140
139
138
137
136
135
134
133
132
131
130
129
128
127
126
125
124
123
122
121
120
119
118
117
116
115
114
113
112
111
110
109
108
107
106
105
104
103
102
101
100
99
98
97
96
95
94
93
92

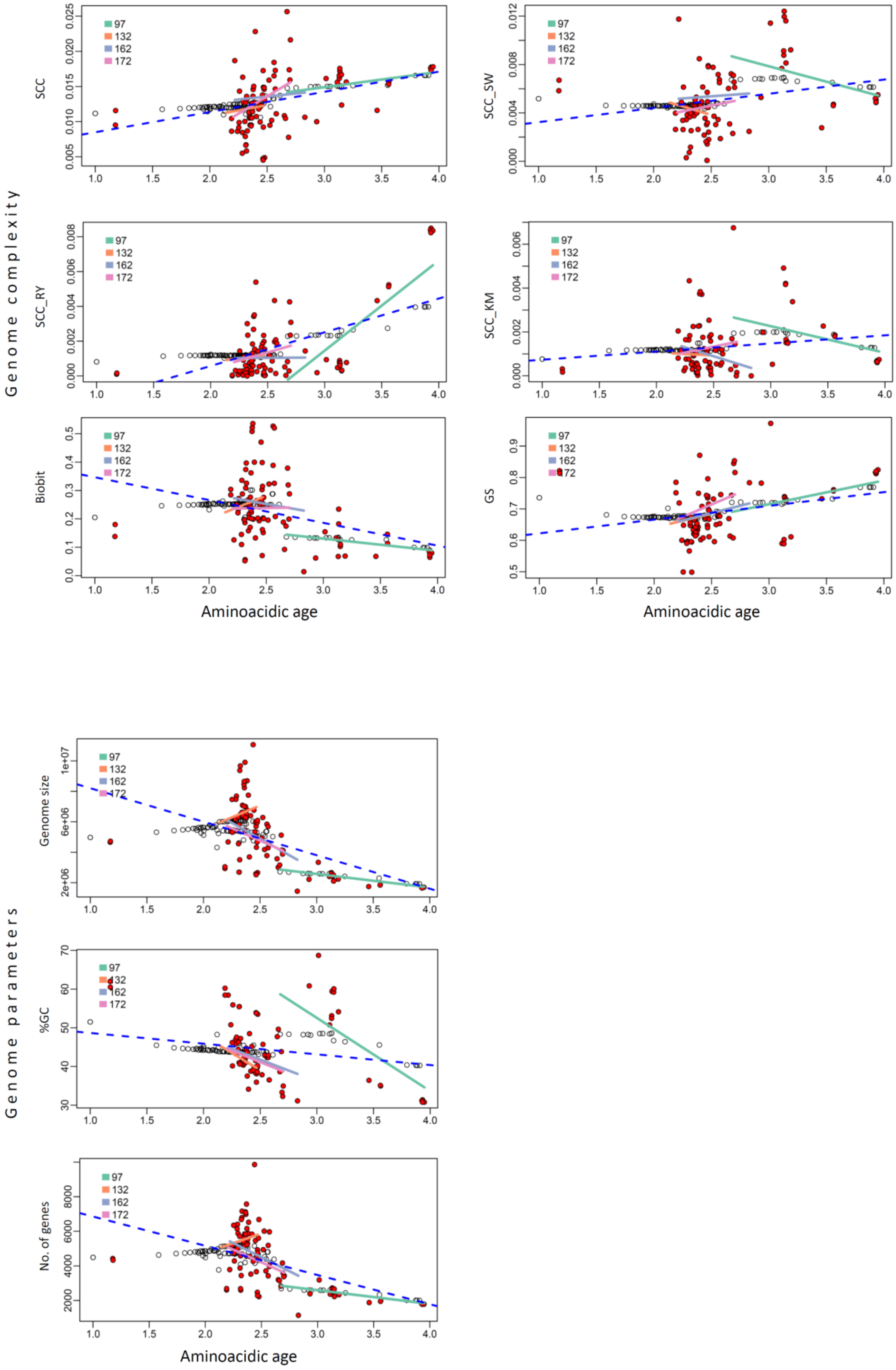
Supplementary File 4. Table S2

	Phylum (n = 91)		Clade 97 (n = 18)		Clade 132 (n = 22)		Clade 162 (n = 11)		Clade 172 (n = 8)	
Metrics:	Slope	P-value	Regression	P-value	Regression	P-value	Regression	P-value	Regression	P-value
SCC	0.14	0.003	0.28	0.023	0.38	0.130	0.22	0.283	0.55	0.062
SCC_SW	0.10	0.013	-0.26	0.008	-0.70	0.016	0.13	0.359	0.36	0.175
SCC_RY	0.23	0.000	0.62	0.000	0.38	0.206	0.02	0.505	0.45	0.121
SCC_KM	0.05	0.158	-0.28	0.007	-0.11	0.337	-0.87	0.006	0.16	0.330
GS	0.09	0.042	0.20	0.096	0.64	0.028	0.53	0.094	0.64	0.038
BB	-0.15	0.004	-0.25	0.053	0.52	0.067	-0.14	0.640	0.02	0.477
Genome parameters:										
Genome size	-0.23	0.081	-0.53	0.044	0.46	0.304	-0.90	0.090	-1.36	0.036
%GC	-0.07	0.362	-0.50	0.026	-2.82	0.001	-0.68	0.180	-0.68	0.185
No. of genes	-0.19	0.148	-0.56	0.027	0.36	0.337	-0.59	0.195	-1.52	0.018

Positive trend

Negative trend

Supplementary File 5. Figure S3



Supplementary File 6. Table S3

	Phylum		Subclade 97		Subclade 132		Subclade 162		Subclade 172			Mood's Median test p-value (Phylum vs. Subclade)			
Metric / Parameter	n	Median	n	Median	n	Median	n	Median	n	Median		Subclade 97	Subclade 132	Subclade 162	Subclade 172
SCC	91	1.24E-02	18	1.64E-02	22	1.14E-02	11	1.47E-02	8	1.35E-02		3.9E-01	🟥	3.6E-01	3.3E-01
SCC_{SW}	91	4.48E-03	18	6.59E-03	22	4.29E-03	11	6.51E-03	8	4.65E-03		2.6E-01	🟥	3.4E-01	3.1E-01
SCC_{RY}	91	7.35E-04	18	9.12E-04	22	8.19E-04	11	5.46E-04	8	1.35E-03		3.9E-01	4.6E-01	🟥	3.3E-01
SCC_{KM}	91	1.00E-03	18	1.59E-03	22	5.96E-04	11	5.93E-04	8	8.04E-04		3.9E-01	🟥	🟥	🟥
GS	91	6.68E-01	18	7.48E-01	22	6.60E-01	11	6.79E-01	8	7.71E-01		3.9E-01	🟥	3.6E-01	3.3E-01
Biobit	91	2.04E-01	18	9.63E-02	22	2.45E-01	11	2.03E-01	8	2.04E-01		🟥	4.0E-01		3.3E-01
%GC	91	4.23E+01	18	5.16E+01	22	4.10E+01	11	4.02E+01	8	3.62E+01		3.9E-01	🟥	🟥	🟥
No. of genes	91	4336	18	2284.5	22	5570.5	11	4371	8	3506.5		🟥	4.0E-01	3.6E-01	🟥
Genome Size	91	4934270	18	2168210	22	6524400	11	4934270	8	4132140		🟥	4.0E-01	3.6E-01	🟥
🟥 Not necessary, as the subclade median is **lower** than the phylum median															