Photosynth Res (2014) 122:171–185
DOI 10.1007/s11120-014-0020-x

REGULAR PAPER

Phylogenetic analysis and molecular signatures defining a monophyletic clade of heterocystous cyanobacteria and identifying its closest relatives

Mohammad Howard-Azzeh · Larissa Shamseer · Herb E. Schellhorn · Radhey S. Gupta

Received: 30 January 2014 / Accepted: 22 May 2014 / Published online: 11 June 2014

Abstract Detailed phylogenetic and comparative genomic analyses are reported on 140 genome sequenced cyanobacteria with the main focus on the heterocyst-differentiating cyanobacteria. In a phylogenetic tree for cyanobacteria based upon concatenated sequences for 32 conserved proteins, the available cyanobacteria formed 8–9 strongly supported clades at the highest level, which may correspond to the higher taxonomic clades of this phylum. One of these clades contained all heterocystous cyanobacteria; within this clade, the members exhibiting either true (*Nostocales*) or false (*Stigonematales*) branching of filaments were intermixed indicating that the division of the heterocysts-forming cyanobacteria into these two groups is not supported by phylogenetic considerations. However, in both the protein tree as well as in the 16S rRNA gene tree, the akinete-forming heterocystous cyanobacteria formed a distinct clade. Within this clade, the members which differentiate into hormogonia or those which lack this ability were also separated into distinct groups. A novel molecular signature identified in this work that is uniquely shared by the akinete-forming heterocystous cyanobacteria provides further evidence that the members of this group are specifically related and they shared a common ancestor exclusive of the other cyanobacteria. Detailed comparative analyses on protein sequences from the genomes of heterocystous cyanobacteria reported here have also identified eight conserved signature indels (CSIs) in proteins involved in a broad range of functions, and three conserved signature proteins, that are either uniquely or mainly found in all heterocysts-forming cyanobacteria, but generally not found in other cyanobacteria. These molecular markers provide novel means for the identification of heterocystous cyanobacteria, and they provide evidence of their monophyletic origin. Additionally, this work has also identified seven CSIs in other proteins which in addition to the heterocystous cyanobacteria are uniquely shared by two smaller clades of cyanobacteria, which form the successive outgroups of the clade comprising of the heterocystous cyanobacteria in the protein trees. Based upon their close relationship to the heterocystous cyanobacteria, the members of these clades are indicated to be the closest relatives of the heterocysts-forming cyanobacteria.

Keywords Nostocales · Molecular signatures · Heterocyst-forming cyanobacteria · Akinete-forming cyanobacteria · Molecular phylogeny · Comparative genomics

Introduction

The Cyanobacteria constitute a large phylum of oxygenic phototrophic bacteria, which exhibit enormous diversity in terms of their morphological, physiological and developmental characteristics (Rippka et al. 1979; Castenholz 2001). The evolutionary relationships and the classification of bacteria within this phylum are, at present, poorly understood (Wilmotte and Herdman 2001; Komárek 2002).

Electronic supplementary material The online version of this article (doi:10.1007/s11120-014-0020-x) contains supplementary material, which is available to authorized users.

M. Howard-Azzeh · H. E. Schellhorn
Department of Biology, McMaster University, Hamilton L8N 3Z5, Canada

M. Howard-Azzeh · L. Shamseer · R. S. Gupta (✉)
Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton L8N 3Z5, Canada
e-mail: gupta@mcmaster.ca

This is largely due to the fact that for a long-time cyanobacterial taxonomy was governed by botanical criteria, which are based on morphological and developmental characteristics (Geitler 1932; Desikachary 1959; Rippka et al. 1979; Castenholz 2001; Hoffmann 2005). These characteristics are generally plastic in nature and most of them are not well suited for reliable classification (Stanier et al. 1978; Woese 1992; Gupta 1998; Wilmotte and Herdman 2001). Although the nomenclature and taxonomy of cyanobacteria was later placed under the International Code of Nomenclature of Bacteria (ICNB) (Stanier et al. 1978; De Vos and Truper 2000; Labeda 2000; Hoffmann 2005), most cyanobacterial names are still based on the original botanical criteria and very few taxa are validly described in terms of the bacteriological code (Oren 2004; Hoffmann 2005; Oren et al. 2009; Parte 2014). As a result, species/strains from large numbers of cyanobacterial genera (e.g., *Synechococcus, Nostoc, Calothrix, Cyanothece, Oscillatoria, Anabaena, Prochlorococcus, Leptolyngbya*, etc.) exhibit extensive polyphyletic branching in the 16S rRNA or other gene/protein trees (Honda et al. 1999; Turner et al. 1999; Wilmotte and Herdman 2001; Hoffmann et al. 2005; Rajaniemi et al. 2005; Gupta and Mathews 2010; Shih et al. 2013). Thus, it has proven very difficult to develop any reliable classification for cyanobacteria (Rippka et al. 1979; Castenholz 2001; Cavalier-Smith 2002; Hoffmann 2005; Sayers et al. 2010; Oren and Garrity 2014). In view of the enormous challenges that are faced in developing a meaningful classification of cyanobacteria under the Bacteriological Code, it has been recently proposed that the cyanobacteria be excluded from the groups of organisms that are covered by the Bacteriological Code (Oren and Garrity 2014).

Within cyanobacteria, one important group consists of the bacteria that are able to differentiate into heterocysts (Rippka et al. 1979; Castenholz 2001). Heterocysts are specialized cells that enable them to separate nitrogen fixation from photosynthesis. The heterocystous cyanobacteria have filamentous morphology and most of them use specialized cells, called hormogonia, for replication (Rippka et al. 1979; Castenholz 2001). The heterocyst-producing cyanobacteria are further divided into two groups based on their development of unbranched (false branching) or branched (true branching) filament colonies (Rippka et al. 1979; Cavalier-Smith 2002). The two kinds of heterocysts are placed into two different orders, *Nostocales* and *Stigonematales* by Cavalier-Smith (2002), and in two separate sections (IV and V, respectively) in a classification scheme proposed by Rippka et al. (1979). In the 16S rRNA tree, heterocystous cyanobacteria, which are comprised of >40 genera, form a monophyletic cluster (Giovannoni et al. 1988; Honda et al. 1999; Turner et al. 1999; Wilmotte and Herdman 2001; Henson et al. 2002; Rajaniemi et al. 2005). However, within this clade, the members of both orders exhibit extensive intermixing (Wilmotte and Herdman 2001; Gugger and Hoffmann 2004). Similar polyphyletic branching of the members from these two orders is observed in phylogenetic trees based upon *nifD* and *nifH* sequences (Zehr et al. 1997; Henson et al. 2002, 2004; Singh et al. 2013). These observations indicate that the division of heterocystous cyanobacteria into the two distinct groups, viz. *Nostocales* and *Stigonematales*, is not supported by the available evidence. Additionally, although the formation of hormogonia is regarded as a defining characteristic of heterocystous cyanobacteria (Cavalier-Smith 2002), these structures are not found in many members of this group (Castenholz 2001; Shih et al. 2013). Thus, it is important to identify other biochemical/molecular characteristics which are uniquely shared by either all or different subgroups of heterocystous cyanobacteria and can prove useful in clarifying their evolutionary relationships.

In recent years, genome sequences for large number of cyanobacteria have become available; these sequences provide a valuable resource for understanding the evolutionary relationships among cyanobacteria and for discovering molecular markers that are specific for the different main clades that are present within this phylum (Gupta 2000; Gao et al. 2009; Gupta 2009; Wu et al. 2009). Our earlier work on a limited number of cyanobacterial genomes identified many molecular signatures in the forms of conserved signature inserts or deletions (i.e., Indels) (CSIs) in protein sequences and conserved signature proteins (CSPs) that are specific for different clades of cyanobacteria (Gupta 2009, 2013; Gupta and Mathews 2010). However, the overall coverage of cyanobacterial diversity in these analyses was very limited and they included either no genomes, or only a few genomes, from the *Stigonematales* and *Nostocales* species. Hence, an examination of the relationships among the heterocystous cyanobacteria or the discovery of new molecular signatures for this group was not feasible at that time. Recently, as a result of the phylogenetically driven Genomic Encyclopedia of Bacteria and Archaea (GEBA) project (Wu et al. 2009; Shih et al. 2013), genome sequences for large numbers of cyanobacteria have become available (CyanoGEBA database) providing extensive coverage of the phylogenetic diversity within this phylum. These sequences, which more than triple the number of available cyanobacterial genomes, include multiple representatives from the orders *Nostocales* and *Stigonematales* (Shih et al. 2013), thus enabling detailed phylogenetic and comparative genomic studies on these bacteria. In this study, we have used these genome sequences to construct phylogenetic trees for 140 cyanobacterial species/strains based upon concatenated sequences for 32 universally distributed proteins. These studies

provide strong evidence that the heterocystous cyanobacteria form a monophyletic clade within the phylum cyanobacteria. In addition, comparative analyses of these genome sequences have identified 15 CSIs and 3 CSPs that are specific for either all sequenced heterocystous cyanobacteria or a number of their ancestral lineages, providing novel molecular markers and insights into the evolution of this important group of cyanobacteria.

Methods

Phylogenetic analysis

Phylogenetic analysis was performed on a concatenated sequence alignment of 32 highly conserved proteins that are found in most bacteria (Harris et al. 2003; Gupta 2009). The names and sequence characteristics of the proteins used for phylogenetic analysis is provided in Supplementary Table 1. Amino acid sequences for these proteins were obtained for 140 cyanobacterial species/strains, whose complete or draft genomes are now available in at least one of the following databases viz. Joint Genome Institute's Integrated Microbial Genomes Database (http://img.jgi.doe.gov/cgi-bin/w/main.cgi), NCBI genome database (http://www.ncbi.nlm.nih.gov/) and the EzGenome database (http://www.ezbiocloud.net/). Characteristics of these genomes are listed in Supplementary Table 2. The sequences for *Bacillus subtilis* subsp. *subtilis* 168 were included in our dataset to root the tree. Multiple sequence alignments of the proteins were created using Clustal_X 2.1 (Larkin et al. 2007) and after arranging them in the same species order, they were concatenated into a single file. Poorly aligned regions from the alignment were removed using Gblocks 0.91b (Castresana 2000). The resulting sequence alignment, which contains 15,279 aligned positions, was used for phylogenetic analysis. Maximum-likelihood (ML) and neighbor-joining (NJ) trees based on 100 bootstrap replicates of this sequence alignment were constructed using Mega 5.05 (Tamura et al. 2011) employing Jones-Taylor-Thornton substitution models (Jones et al. 1992). A phylogenetic tree was also constructed for 68 16S rRNA gene sequences (>1,100 bp in length) for members of the orders *Nostocales* and *Stigonematales* that are present in the SILVA ribosomal RNA gene database (Quast et al. 2013). This alignment included at least one representative members of different genera from the above two orders. A maximum-likelihood tree with 100 bootstrap replicates was constructed from this dataset based on the maximum composite-likelihood model of evolutionary gene change using Mega 5.05 (Nei and Kumar 2000; Tamura et al. 2011). The 16S rRNA gene of *Microcystis aeruginosa* NIES-843 was used to root this tree.

Identification of conserved signature indels

The identification of CSIs was carried out in a similar manner as described in earlier work (Gupta et al. 2003; Gupta 2009). Blastp searches were initially performed on proteins from the genome of *Nostoc* sp. PCC 7120. These searches were individually performed on all proteins from GI numbers 17227498 to 177228576 and 17231359 to 17232863 using the NCBI nr database. For each protein, sequences of 10–15 high-scoring homologues were obtained from the available *Nostocales* and other cyanobacteria as well as several outgroup species. Multiple sequence alignments of these proteins were constructed using Clustal_X 2.1 (Larkin et al. 2007). The resulting alignments were visually examined for the presence of indels that are flanked on both sides by at least 5–6 identically conserved amino acids in the neighboring 30–40 residues. Indels that were not flanked on either side by conserved regions were not further considered, as they do not provide useful molecular markers and could arise from alignment artefacts (Gupta et al. 2003; Gupta 2009). Species distribution patterns of all potentially useful indels were examined further by performing more detailed Blastp searches on short sequence regions (approximately 60–80 aa long) containing the indel and its flanking conserved regions. The top 250 blast hits were examined for the presence or absence of similar indels to determine the specificities of different indels (Gupta 2009). Protein sequence information from available draft genomes was obtained by tBlastn searches. In this work, we report the results of those CSIs that in most cases are specifically present in species from the orders *Nostocales* and *Stigonematales* and not present in other cyanobacteria or other bacteria.

Identification of conserved signature proteins specific for heterocystous cyanobacteria

Blastp searches were conducted on all previously identified proteins that were only found in a limited number of *Nostocales* species (from Table 4 and supplementary Table 5 of Gupta and Mathews 2010) to determine their species distribution. A protein was considered to be specific for heterocystous cyanobacteria, if all significant Blast hits (E value $<1 \times 10^{-4}$) were from this group and the protein was present in all (or most) heterocystous cyanobacteria, with very few exceptions (Gupta and Mathews 2010).

Results

Phylogenetic analysis of cyanobacteria based on a large dataset of protein sequences

The evolutionary relationships among different cyanobacteria until recently were only determined on the basis of 16S rRNA gene sequences (Honda et al. 1999; Turner et al. 1999; Wilmotte and Herdman 2001; Yarza et al. 2010). With the availability of genome sequences, it is now possible to construct phylogenetic trees based upon concatenated sequences for many conserved proteins, which provide a more reliable portrayal of the species relationships than those based on single genes or proteins (Rokas et al. 2003; Ciccarelli et al. 2006; Wu et al. 2009; Gupta and Mathews 2010). Phylogenetic trees for a limited number of cyanobacteria have been previously constructed based upon different datasets of protein sequences (Sanchez-Baracaldo et al. 2005; Zhaxybayeva et al. 2006; Shi and Falkowski 2008; Swingley et al. 2008; Gupta and Mathews 2010). Although, in these studies, a number of distinct clades of cyanobacteria were identified, due to the very limited coverage of cyanobacterial diversity in these analyses, inferences based upon them required further confirmation. Recently, Shih et al. (2013) reported sequencing of genomes from 54 phylogenetically and phenotypically diverse cyanobacteria (CyanoGEBA database), which greatly increases the information available for this phylum. In the phylogenetic tree, that they constructed for 123 cyanobacterial species/strains, based upon 31 conserved proteins, the sequenced cyanobacteria formed several distinct clusters and all heterocystous cyanobacteria grouped into one clade (Shih et al. 2013). More recently, draft genomes for a number of additional cyanobacterial species/strains, which include 11 new members from the heterocyst group (viz. *Cylindrospermopsis raciborskii* CS-509, *Anabaena* sp. 90, *Anabaena circinalis* AWQC131C, *Anabaena circinalis* AWQC310F, *Chlorogloeopsis fritschii* PCC 9212, *Chlorogloeopsis fritschii* PCC 6912, *Fischerella thermalis* PCC 7521, *Fischerella muscicola* PCC 73103, *Calothrix desertica* PCC 7102, *Mastigocoleus testarum* BC008, *Richelia intracellularis* HH01) have become available (see Methods). This has afforded a more detailed examination of the relationships among heterocystous cyanobacteria.

In this study, we have constructed phylogenetic trees for 140 cyanobacteria (listed in Supplementary Table 2), including 35 members of the heterocystous group, based upon concatenated sequences of 32 conserved proteins. The trees based upon this large dataset of protein sequences were constructed using both ML and NJ algorithms. The tree obtained using the ML method is shown in Fig. 1 and information for the NJ tree is provided in Supplementary Figure 1. The branching patterns of the cyanobacterial species/strains in the two trees are very similar and most nodes are supported by bootstrap values between 70 and 100 %. As noted earlier, and known from numerous other studies (Honda et al. 1999; Turner et al. 1999; Wilmotte and Herdman 2001; Rajaniemi et al. 2005; Gupta and Mathews 2010; Yarza et al. 2010; Shih et al. 2013), many cyanobacterial genera exhibited extensive polyphyletic branching indicating that the current names of many cyanobacteria are not informative and may, in fact, be misleading. Aside from this problem, the examined cyanobacterial species formed a number of strongly supported clades in these trees. These clades, arbitrarily marked 1–9, are generally similar to those observed by Shih et al. (2013), except for some differences in the grouping of species in the smaller clades.

Of these clades, Clade 1A includes all of the heterocystous cyanobacteria confirming their monophyletic origin. The closest relative of the heterocystous cyanobacteria in this tree is Clade 1B, which is comprised of three poorly characterized cyanobacteria (viz. *Synechocystis* PCC 7509, *Chroococcidiopsis* PCC 7203 and *Gloeocapsa* PCC 7428). A specific relationship between the Clade 1A and 1B cyanobacteria is supported by both ML and NJ methods and it was also observed in the tree by Shih et al. (2013). In addition to this relationship between Clade 1A and Clade 1B cyanobacteria, another small clade consisting of two species viz. *Crinalium epipsammum* PCC 9333 and *Chamaesiphon minutus* PCC 6605 (Clade 1C) also exhibited a close relationship to these two groups of cyanobacteria. A close relationship of the heterocystous cyanobacteria (Clade 1A) to the species which are part of the Clades 1B and 1C is also observed in the 16S rRNA tree (Wilmotte and Herdman 2001). Of the other main clades marked in the tree, the Clades 2 and 3 are mainly comprised of the species from the orders *Chroococcales* and *Oscillatoriales.* Although species/strains belonging to these orders of cyanobacteria are also present in a number of other clades, these other species are evolutionary unrelated to the members of these two clades. The phylogenetic tree also supports a grouping of the species/strains from Clades 1, 2 and 3 into a larger clade, which corresponds to Clade B in our earlier work (Gupta 2009; Gupta and Mathews 2010) and by Shih et al. (2013). In addition to these clades, another large and strongly supported clade in the tree, i.e., Clade 7, is comprised of marine unicellular cyanobacteria belonging to the genera *Prochlorococcus, Synechococcus* and *Cyanobium*. This clade, which is separated from all other cyanobacteria by a long branch, was referred to as Clade C in our earlier work (Gupta 2009; Gupta and Mathews 2010) or Syn/Pro clade by Sanchez-Baracaldo et al. (2005). Previously, many CSIs and CSPs, which are specific for members of this clade, have been identified

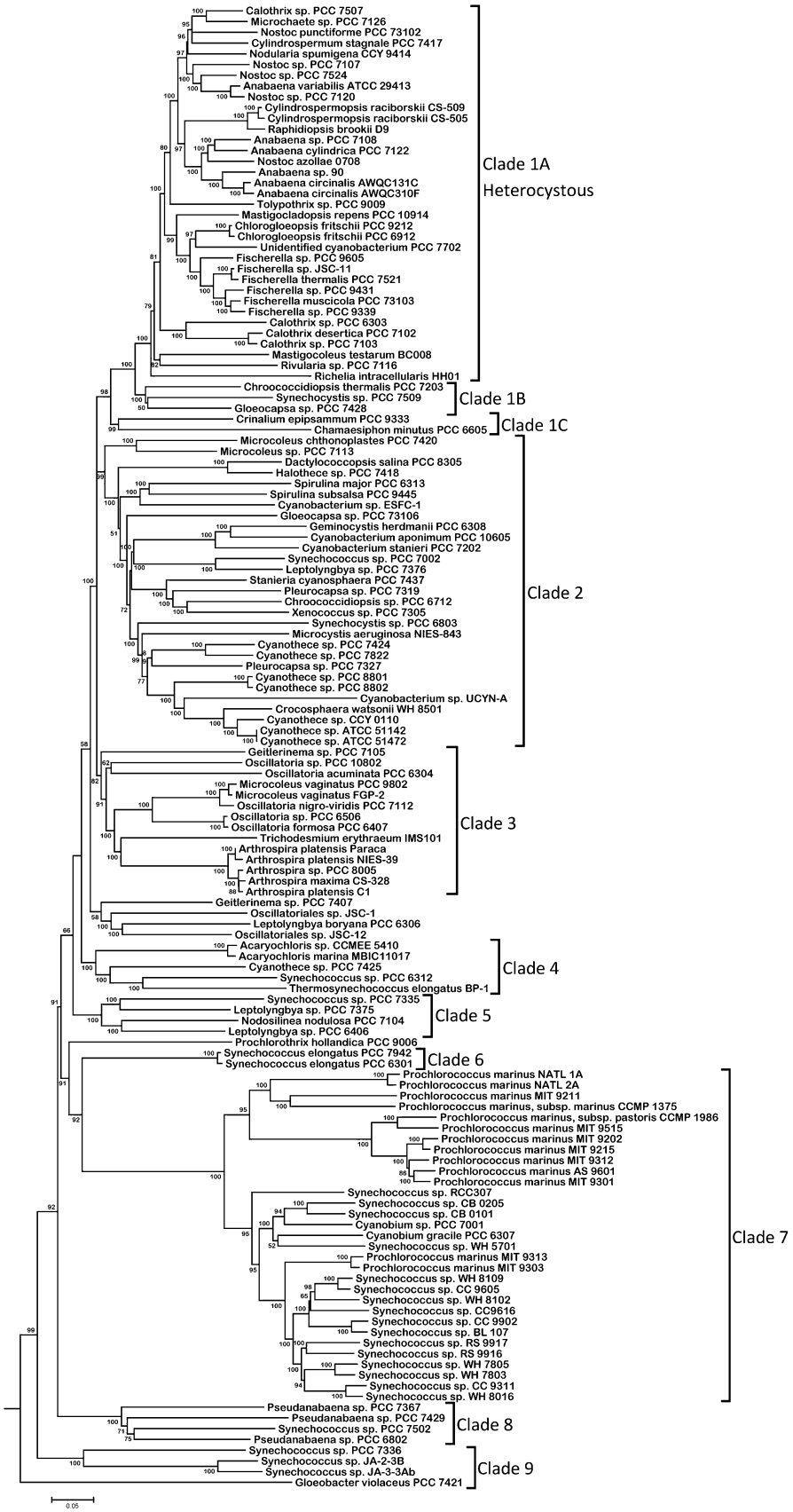


Fig. 1 A maximum-likelihood consensus tree based on 32 concatenated sequences for 140 species of cyanobacteria. The tree was rooted using the sequences for *Bacillus subtilis* subsp. *subtilis 168*. Numbers located at nodes indicate the bootstrap values out of 100. Major clades and subclades of cyanobacteria resolved in the tree are indicated

(Gupta 2009; Gupta and Mathews 2010). Besides these main clades, a number of smaller clades of cyanobacteria, which are comprised of between 3 and 5 species/strains, are also observed in the tree shown in Fig. 1 and in the work of Shih et al. (2013).

Since the focus of this work is on heterocystous cyanobacteria, a subtree for these bacteria and the two related clades excerpted from Fig. 1 is shown in Fig. 2. Within the Clade 1A, comprising of heterocystous cyanobacteria, a number of distinct subclades are also phylogenetically resolved. The members of these subclades differ from each other based upon morphological/developmental characteristic. One of these large subclades is comprised of the akinetes-forming *Nostocales* species/strains. This subclade can be further divided into two smaller subclades depending upon whether the members of these clades differentiate into hormogonia, or they lack such ability. However, it should be mentioned that the ability to differentiate into hormogonia is more broadly distributed among the heterocysts-forming cyanobacteria (Rippka et al. 1979; Cavalier-Smith 2002).

A phylogenetic tree was also constructed for the heterocystous cyanobacteria based on 16S rRNA gene

Fig. 3 A maximum-likelihood consensus tree based on 16S rRNA ▶ gene sequences, representing at least one member from every genera from the orders *Nostocales* and *Stigonematales* (except *Riveria*) and rooted with *Microcystis aeruginosa NIES-843*.[N] and [S] denote members belonging to the *Nostocales* and *Stigonematales* groups, respectively. The *boxed* species indicate those which were also used in the creation of concatenated protein trees and for comparative genomic analyses

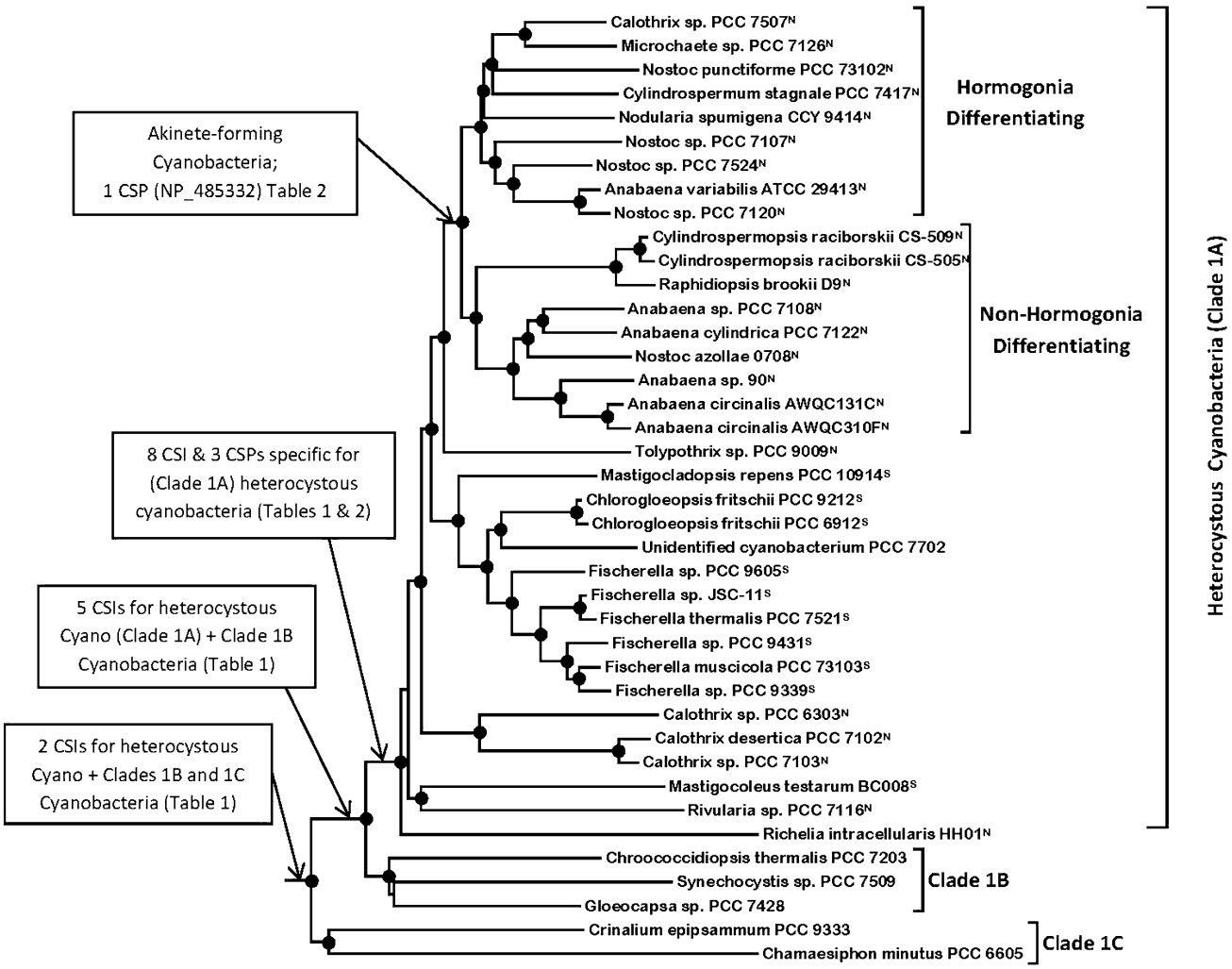
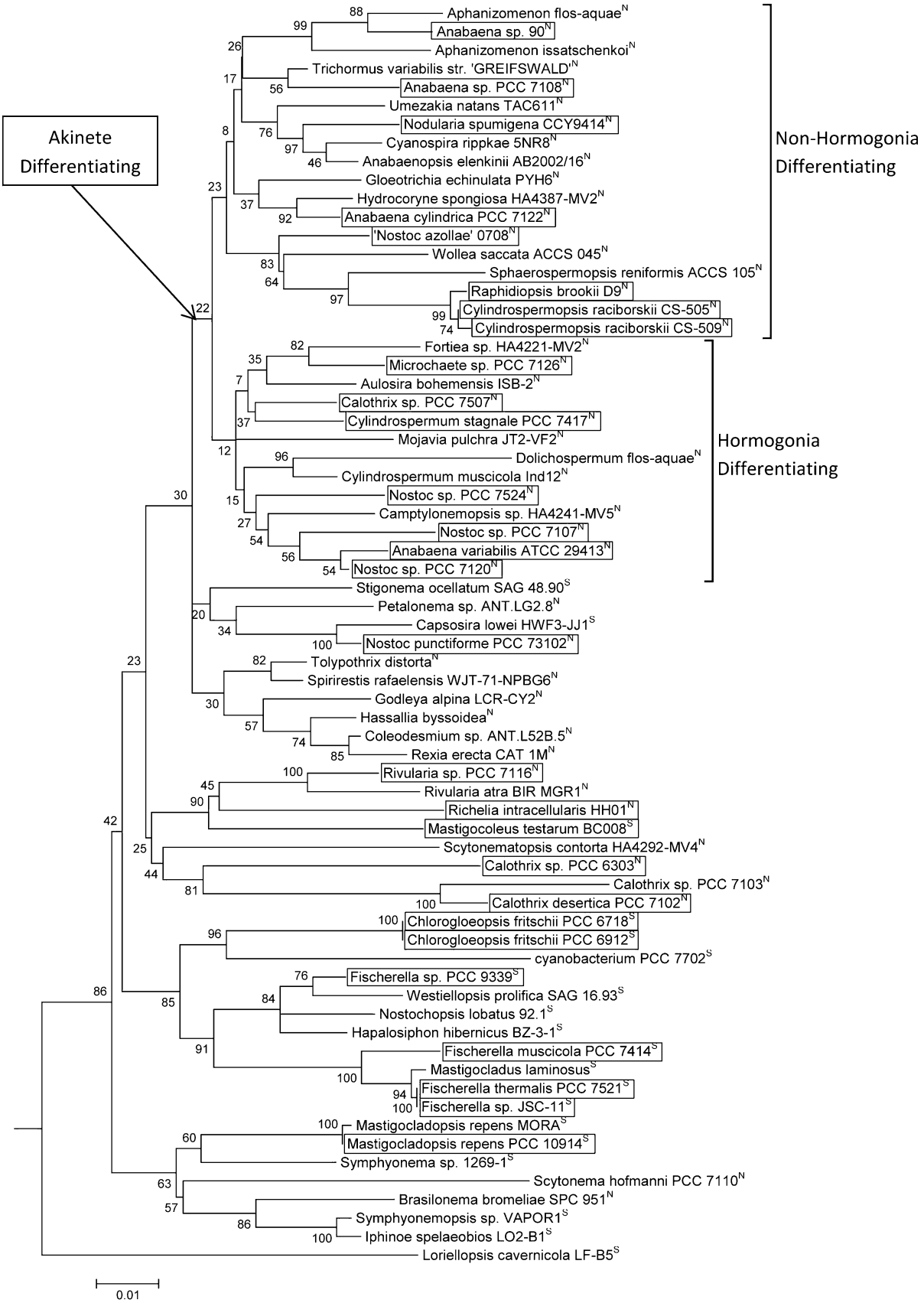


Fig. 2 A summary diagram showing the distribution of identified CSIs and CSPs specific to heterocystous cyanobacteria and its immediate relatives. The identified clades supported by both phylogenetic studies as well as various molecular markers are indicated. *Dots* indicate nodes that are supported by bootstrap scores of at least 80. The *superscripts* [N] and [S] denote members belonging to the suggested orders *Nostocales* and *Stigonematales*, respectively


Akinete Differentiating
Non-Hormogonia Differentiating
Hormogonia Differentiating
Aphanizomenon flos-aquae^N
Anabaena sp. 90^N
Aphanizomenon issatschenkoi^N
Trichormus variabilis str. 'GREIFSWALD'^N
Anabaena sp. PCC 7108^N
Umezakia natans TAC611^N
Nodularia spumigena CCY9414^N
Cyanospira rippkae 5NR8^N
Anabaenopsis elenkinii AB2002/16^N
Gloeotrichia echinulata PYH6^N
Hydrocoryne spongiosa HA4387-MV2^N
Anabaena cylindrica PCC 7122^N
'Nostoc azollae' 0708^N
Wollea saccata ACCS 045^N
Sphaerospermopsis reniformis ACCS 105^N
Raphidiopsis brookii D9^N
Cylindrospermopsis raciborskii CS-505^N
Cylindrospermopsis raciborskii CS-509^N
Fortiea sp. HA4221-MV2^N
Microchaete sp. PCC 7126^N
Aulosira bohemensis ISB-2^N
Calothrix sp. PCC 7507^N
Cylindrospermum stagnale PCC 7417^N
Mojavia pulchra JT2-VF2^N
Dolichospermum flos-aquae^N
Cylindrospermum muscicola Ind12^N
Nostoc sp. PCC 7524^N
Camptylonemopsis sp. HA4241-MV5^N
Nostoc sp. PCC 7107^N
Anabaena variabilis ATCC 29413^N
Nostoc sp. PCC 7120^N
Stigonema ocellatum SAG 48.90^S
Petalonema sp. ANT.LG2.8^N
Capsosira lowei HWF3-JJ1^S
Nostoc punctiforme PCC 73102^N
Tolypothrix distorta^N
Spirirestis rafaelensis WJT-71-NPBG6^N
Godleya alpina LCR-CY2^N
Hassallia byssoidea^N
Coleodesmium sp. ANT.L52B.5^N
Rexia erecta CAT 1M^N
Rivularia sp. PCC 7116^N
Rivularia atra BIR MGR1^N
Richelia intracellularis HH01^N
Mastigocoleus testarum BC008^S
Scytonematopsis contorta HA4292-MV4^N
Calothrix sp. PCC 6303^N
Calothrix sp. PCC 7103^N
Calothrix desertica PCC 7102^N
Chlorogloeopsis fritschii PCC 6718^S
Chlorogloeopsis fritschii PCC 6912^S
cyanobacterium PCC 7702^S
Fischerella sp. PCC 9339^S
Westiellopsis prolifica SAG 16.93^S
Nostochopsis lobatus 92.1^S
Hapalosiphon hibernicus BZ-3-1^S
Fischerella muscicola PCC 7414^S
Mastigocladus laminosus^S
Fischerella thermalis PCC 7521^S
Fischerella sp. JSC-11^S
Mastigocladopsis repens MORA^S
Mastigocladopsis repens PCC 10914^S
Symphyonema sp. 1269-1^S
Scytonema hofmanni PCC 7110^N
Brasilonema bromeliae SPC 951^N
Symphyonemopsis sp. VAPOR1^S
Iphinoe spelaeobios LO2-B1^S
Loriellopsis cavernicola LF-B5^S
0.01

sequences (Fig. 3). This tree includes representatives from different genera of heterocystous cyanobacteria. The branching of different species/strain in this tree is similar to those observed in earlier studies (Wilmotte and Herdman 2001; Gugger and Hoffmann 2004), with members from the orders *Nostocales* or *Stigonematales* (marked with the superscripts ^{N} and ^{S}, respectively) showed extensive intermixing. The cyanobacterial species/strains that are part of our protein tree are distributed throughout the rRNA tree indicating that our dataset provides a good representation of the known heterocystous cyanobacteria. Importantly, the different subclades of heterocystous cyanobacteria, which are seen in the protein tree, are also observed in the rRNA tree. Thus, a subclade consisting of the akinete-forming cyanobacteria and the two groups within it, which differentiate into hormogonia or lack such ability, are also resolved, with only isolated exceptions.

Conserved signature indels specific for *Nostocales*/*Stigonematales*

An important objective of this study was to identify molecular markers that can distinguish the heterocystous cyanobacteria from all other bacteria, or those which can provide insights into the origin and evolutionary relationships of these bacteria to other cyanobacteria. As noted earlier, CSIs and CSPs, which are restricted to a given group of related species, provide important molecular characteristics for evolutionary and taxonomic purposes (Baldauf and Palmer 1993; Delwiche et al. 1995; Gupta 1998, 2003; 2009; Rokas and Holland 2000). Recently, these markers have been used to define, in molecular terms, and to propose important taxonomic changes in a number of phyla of bacteria (viz. Spirochetes, Aquificae, Thermotogae, Chloroflexi) at multiple phylogenetic levels (Adeolu and Gupta 2013; Bhandari and Gupta 2014; Gupta and Lali 2013; Gupta et al. 2013; Naushad et al. 2014). We have previously identified several CSIs and CSPs that appeared specific, at that time, for either all cyanobacteria, or a number of their larger clades (Gupta 2009, 2010; Gupta and Mathews 2010). However, due to the paucity of sequence information for heterocystous cyanobacteria, no markers specific for these bacteria were identified at that time. As genome sequences are now available for large numbers of heterocystous cyanobacteria, comparative genomic analyses were undertaken to identify conserved signature indels (CSIs) that might be specifically present in this group of bacteria or provide information regarding their relationships to other cyanobacteria.

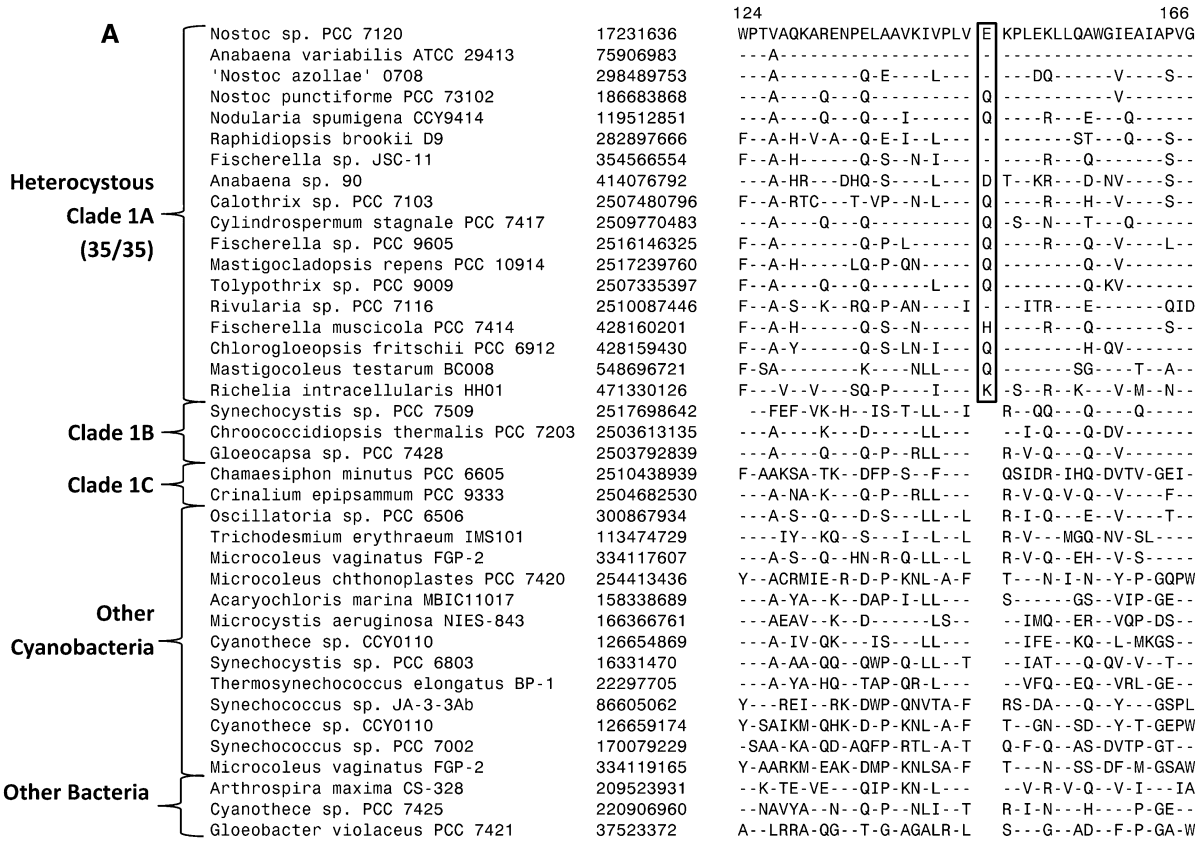
The present study has identified 15 CSIs in different proteins that are useful in this respect. Eight of these CSIs, which are present in proteins involved in diverse functions, are specific for all of the sequenced species/strains of heterocystous cyanobacteria (i.e., Clade 1A) and they are not found in the homologous proteins from any other bacteria. Two examples of the CSIs that are specific for the Clade 1A species are shown in Fig. 4. In the first example, one amino acid insert is present in a highly conserved region of a XRE family transcription regulator (Fig. 4a), which play an important role in the metabolism of toxic compounds (Saatcioglu et al. 1990). Another CSI showing similar specificity is shown in Fig. 4b, where a 7 aa insert in the protein all0200 (DUF111) is uniquely found in all heterocystous cyanobacteria. Sequence information for 6 other CSIs showing similar specificity is provided in Supplementary Figures S2–S7 and some of their characteristics are summarized in Table 1.

Fig. 4 Partial sequence alignments for the proteins **a** XRE family protein, showing a 1 aa insertion and **b** all0200, showing a 7 aa insertion that are both specifically present in all heterocystous cyanobacterial members and are flanked by conserved regions. *Dashes* in the sequence alignments show amino acid identity with the amino acid indicated on the top line. Sequence information for a limited number of species from other cyanobacterial taxa are shown here, but the indicated CSIs were not found in any other cyanobacteria ▶

In addition to the CSIs that are specific for all heterocystous cyanobacteria (Clade 1A), our analyses have also identified 5 CSIs, which are found in the Clade 1A as well as in members of the Clade 1B cyanobacteria, which forms the immediate out group of the Clade 1A in the phylogenetic tree (Fig. 1). Partial sequence alignment for one of these CSIs, containing a 5 amino acids insert in a highly conserved region of the 30S ribosomal protein S3, is shown in Fig. 5a. Sequence alignments for four other CSIs in different proteins, which also exhibit similar specificity, are provided in Supplementary Figures S8–S11 and some of their characteristics are summarized in Table 1. Two other CSIs, identified by our analysis, in addition to being commonly shared by Clade 1A and Clade 1B cyanobacteria, are also present in some or all members of the Clade 1C cyanobacteria (comprising of *Crinalium epipsammum* PCC 9333 and *Chamaesiphon minutus* PCC 6605), which forms the out group of the Clades 1A and 1B in the phylogenetic tree (Fig. 1). Sequence information for one of these CSIs consisting of a 7 aa insert in a conserved region of the pentapeptide repeat protein is shown in Fig. 5b and some characteristics of these CSI are also summarized in Table 1.

Signature proteins (CSPs) specific for the heterocystous cyanobacteria

Our earlier work on *Nostoc sp.* PCC 7120 identified a number of proteins which were specifically found in a


A
124
166
Nostoc sp. PCC 7120 17231636 WPTVAQKARENPELAAVKIVPLV E KPLEKLLQAWGIEAIAPVG
Anabaena variabilis ATCC 29413 75906983 ---A---------- - ----------
'Nostoc azollae' 0708 298489753 ---A--------Q-E----L--- - ---DQ------V----S--
Nostoc punctiforme PCC 73102 186683868 ---A----Q---Q---------- Q ----------V-------
Nodularia spumigena CCY9414 119512851 ---A----Q---Q---I------ Q ----R---E---Q------
Raphidiopsis brookii D9 282897666 F--A-H-V-A--Q-E-I--L--- - -------ST---Q---S--
Fischerella sp. JSC-11 354566554 F--A-H------Q-S--N-I--- - ----R---Q-------S--
Heterocystous Anabaena sp. 90 414076792 ---A-HR---DHQ-S----L--- D T--KR---D-NV----S--
Clade 1A Calothrix sp. PCC 7103 2507480796 F--A-RTC---T-VP--N-L--- Q ----R---H--V----S--
(35/35) Cylindrospermum stagnale PCC 7417 2509770483 ---A----Q---Q---------- Q -S--N---T---Q------
Fischerella sp. PCC 9605 2516146325 F--A--------Q-P-L------ Q ----R---Q--V----L--
Mastigocladopsis repens PCC 10914 2517239760 F--A-H-----LQ-P-QN----- Q --------Q--V-------
Tolypothrix sp. PCC 9009 2507335397 F--A----Q---Q------L--- Q --------Q-KV-------
Rivularia sp. PCC 7116 2510087446 F--A-S--K--RQ-P-AN----I - --ITR---E-------QID
Fischerella muscicola PCC 7414 428160201 F--A-H------Q-S--N----- H ----R---Q-------S--
Chlorogloeopsis fritschii PCC 6912 428159430 F--A-Y------Q-S-LN-I--- Q --------H-QV-------
Mastigocoleus testarum BC008 548696721 F-SA--------K----NLL--- Q -------SG----T--A--
Richelia intracellularis HH01 471330126 F---V--V---SQ-P----I--- K -S--R--K---V-M--N--
Synechocystis sp. PCC 7509 2517698642 --FEF-VK-H--IS-T-LL--I R--QQ---Q----Q-----
Clade 1B Chroococcidiopsis thermalis PCC 7203 2503613135 ---A----K---D-----LL--- --I-Q---Q-DV-------
Gloeocapsa sp. PCC 7428 2503792839 ---A----Q---Q-P--RLL--- R-V-Q---Q--V-------
Clade 1C Chamaesiphon minutus PCC 6605 2510438939 F-AAKSA-TK--DFP-S--F--- QSIDR-IHQ-DVTV-GEI-
Crinalium epipsammum PCC 9333 2504682530 ---A-NA-K---Q-P--RLL--- R-V-Q-V-Q--V----F--
Oscillatoria sp. PCC 6506 300867934 ---A-S--Q---D-S---LL--L R-I-Q---E--V----T--
Trichodesmium erythraeum IMS101 113474729 ----IY--KQ--S---I--L--L R-V---MGQ-NV-SL----
Microcoleus vaginatus FGP-2 334117607 ---A-S--Q--HN-R-Q-LL--L R-V-Q--EH--V-S-----
Microcoleus chthonoplastes PCC 7420 254413436 Y--ACRMIE-R-D-P-KNL-A-F T---N-I-N--Y-P-GQPW
Acaryochloris marina MBIC11017 158338689 ---A-YA--K--DAP-I-LL--- S------GS--VIP-GE--
Other Microcystis aeruginosa NIES-843 166366761 ---AEAV--K--D------LS-- --IMQ--ER--VQP-DS--
Cyanobacteria Cyanothece sp. CCY0110 126654869 ---A-IV-QK---IS---LL--- --IFE--KQ--L-MKGS--
Synechocystis sp. PCC 6803 16331470 ---A-AA-QQ--QWP-Q-LL--T --IAT---Q-QV-V--T--
Thermosynechococcus elongatus BP-1 22297705 ---A-YA-HQ--TAP-QR-L--- --VFQ--EQ--VRL-GE--
Synechococcus sp. JA-3-3Ab 86605062 Y---REI--RK-DWP-QNVTA-F RS-DA---Q--Y---GSPL
Cyanothece sp. CCY0110 126659174 Y-SAIKM-QHK-D-P-KNL-A-F T--GN--SD--Y-T-GEPW
Synechococcus sp. PCC 7002 170079229 -SAA-KA-QD-AQFP-RTL-A-T Q-F-Q--AS-DVTP-GT--
Microcoleus vaginatus FGP-2 334119165 Y-AARKM-EAK-DMP-KNLSA-F T---N--SS-DF-M-GSAW
Other Bacteria Arthrospira maxima CS-328 209523931 --K-TE-VE---QIP-KN-L--- --V-R-V-Q--V-I---IA
Cyanothece sp. PCC 7425 220906960 --NAVYA--N--Q-P--NLI--T R-I-N---H----P-GE--
Gloeobacter violaceus PCC 7421 37523372 A--LRRA-QG--T-G-AGALR-L S---G--AD--F-P-GA-W
B
131
174
Cylindrospermopsis raciborskii CS-505 282899760 AIVDIVGTCLSLDWLGIDS DETGLPL VYCSAFPTGGGTVRAAHG
Raphidiopsis brookii D9 282895922 ---------- -Q----- ----------
'Nostoc azollae' 0708 298493015 ----------G-------- NQE---- I---P----------
Nostoc punctiforme PCC 73102 186681196 ----------G------E- NDE-W-- L----------
Anabaena variabilis ATCC 29413 75908904 ----------G------A- -NA-F-- L----------
Nostoc sp. PCC 7120 17227696 ----------G------A- -NA-F-- L----------
Nodularia spumigena CCY9414 119509690 ----------G------A- NDQ---- L----L-----
Heterocystous Fischerella sp. JSC-11 354565251 ----------G-------- ------- L----------
Clade 1A Anabaena sp. PCC 7108 2506494187 ----------G-------- NQE---- I---P----------
(35/35) Calothrix sp. PCC 7507 2505799787 ----------G------E- NNA---- L----------
Fischerella sp. PCC 9605 2516146941 ----------G-----LE- N-Q---Q L----------
Mastigocladopsis repens PCC 10914 2517243987 ----------G------E- NHN---- L---PL-----
Rivularia sp. PCC 7116 2510089813 ----V-----G-------- NSE---- L---KL-------G----
Tolypothrix sp. PCC 9009 2507335463 --A-------G-------- N-K-M-- L----L-----
Fischerella thermalis PCC 7521 428159598 ----------G-------- ------- L----------
Fischerella muscicola PCC 7414 428160321 ----------G-------- ------- L----------
Chlorogloeopsis fritschii PCC 6912 428159566 ----------G-------- N-Q---- L----------
Mastigocoleus testarum BC008 548700893 ----------G-------N N-REF-- LF--EQ-------H----
Synechocystis sp. PCC 7509 2517698198 ----------G------EK L---P------N------
Clade 1B Chroococcidiopsis thermalis PCC 7203 2503615130 ----------G-------K I----L------IK----
Gloeocapsa sp. PCC 7428 2503796660 ----------G----E--K L----L-------K----
Clade 1C Chamaesiphon minutus PCC 6605 2510436581 ----------G--Y-DVAK I---PL-I----------
Crinalium epipsammum PCC 9333 2504686140 ----------G-----VEQ L----L-------K----
Cyanothece sp. PCC 7425 220909106 -L--------G------EE L----L-------K----
Microcystis aeruginosa NIES-843 166363888 ----------G----SVSE L---SL-------Q----
Synechococcus sp. PCC 7002 170078317 -L--------G----E--E L--APM-------K----
Acaryochloris marina MBIC11017 158338240 ----------G-H---VEE MF---L-------K----
Thermosynechococcus elongatus BP-1 22299839 -L--------G--Y-DVRQ -----L-------K----
Other Synechococcus sp. JA-3-3Ab 86605652 -L-------VG----QVER -IS-PH-I---W-DTE--
Cyanobacteria Synechococcus sp. RCC307 148242740 SL--VI-V-AG-LH--VQ- LWATPP-A-H-Q--T---
Trichodesmium erythraeum IMS101 113474503 --I-------G----D--L L---PL-I-----K----
Microcoleus vaginatus FGP-2 334121011 --I-------G----D-QE F-F--M-------K----
Prochlorococcus marinus AS9601 123968080 SL---I-V-AA-NY-NPKR ---NQPML-K-F-QTE--
Gloeobacter violaceus PCC 7421 37520111 -L--V-----G--Y-N-EA LH---L-V---L-H----
Staphylococcus epidermidis W23144 242234575 S-I--I-G-IA-EQ---NT L----I---H-KIHI---
Other Bacteria Geobacter sp. M21 253702319 S-------AIG--Y--VEA --A-GL-Y-R-F-QT---
Cohnella laeviribosi 517835598 S-I--I-V--A-EN--V-D IVV-PI---M-K--I---
Desulfotomaculum reducens MI-1 134298028 --------LVG-KK---EA ----PLNV-R---KC---
Eubacterium ramulus ATCC 29099 544229912 -LI-V--S--A-EL--V-E -QA-PVCV-N----C---

small number of the sequenced *Nostocales* species/strains (Gupta and Mathews 2010; Gupta 2010). In view of the large numbers of genome sequences that are now available for heterocystous cyanobacteria, Blastp searches on the sequences of these proteins were repeated to determine whether any of them are specifically found in all heterocystous cyanobacteria. The results of these studies show that for three of the proteins, of unknown functions, all significant Blast hits (except a few isolated exceptions as noted in Table 2) are limited to members of the Clade 1A comprising of heterocystous cyanobacteria. Due to the specific presence of these proteins in different sequenced members of the Clade 1A cyanobacteria, these CSPs appear to be distinctive characteristics of the heterocystous cyanobacteria. In addition to these three CSPs, our analysis has also identified one protein (accession number NP_485332), whose homologs are mainly limited to the akinete-forming cyanobacteria (Table 2), providing further evidence that these cyanobacteria form a distinct group within the heterocystous cyanobacteria.

Discussion

In this work, we have examined the evolutionary relationships among 140 genome sequence cyanobacteria, covering its diverse lineages, based upon concatenated sequences for 32 conserved proteins. The tree constructed in this work is the most comprehensive tree for cyanobacteria that has been made to date based upon genomic sequence data. The evolutionary relationships among different cyanobacterial taxa seen in this work are similar to those observed by Shih et al. (2013) in another detailed study based upon a different dataset of protein sequences, and they also show good concordance with the relationships seen in the 16S rRNA trees (Wilmotte and Herdman 2001; Yarza et al. 2010; Shih et al. 2013). The cyanobacterial species form at least 8–9 distinct clades in these trees (Fig. 1), which could correspond to the higher taxonomic groupings (e.g., classes or orders) within the phylum Cyanobacteria. Some of these clades were also identified in our earlier work based upon a limited numbers of cyanobacteria (Gupta 2009; Gupta and Mathews 2010).

The main focus of this work was on heterocystous cyanobacteria. In the trees based on large datasets of concatenated proteins, these bacteria formed a strongly

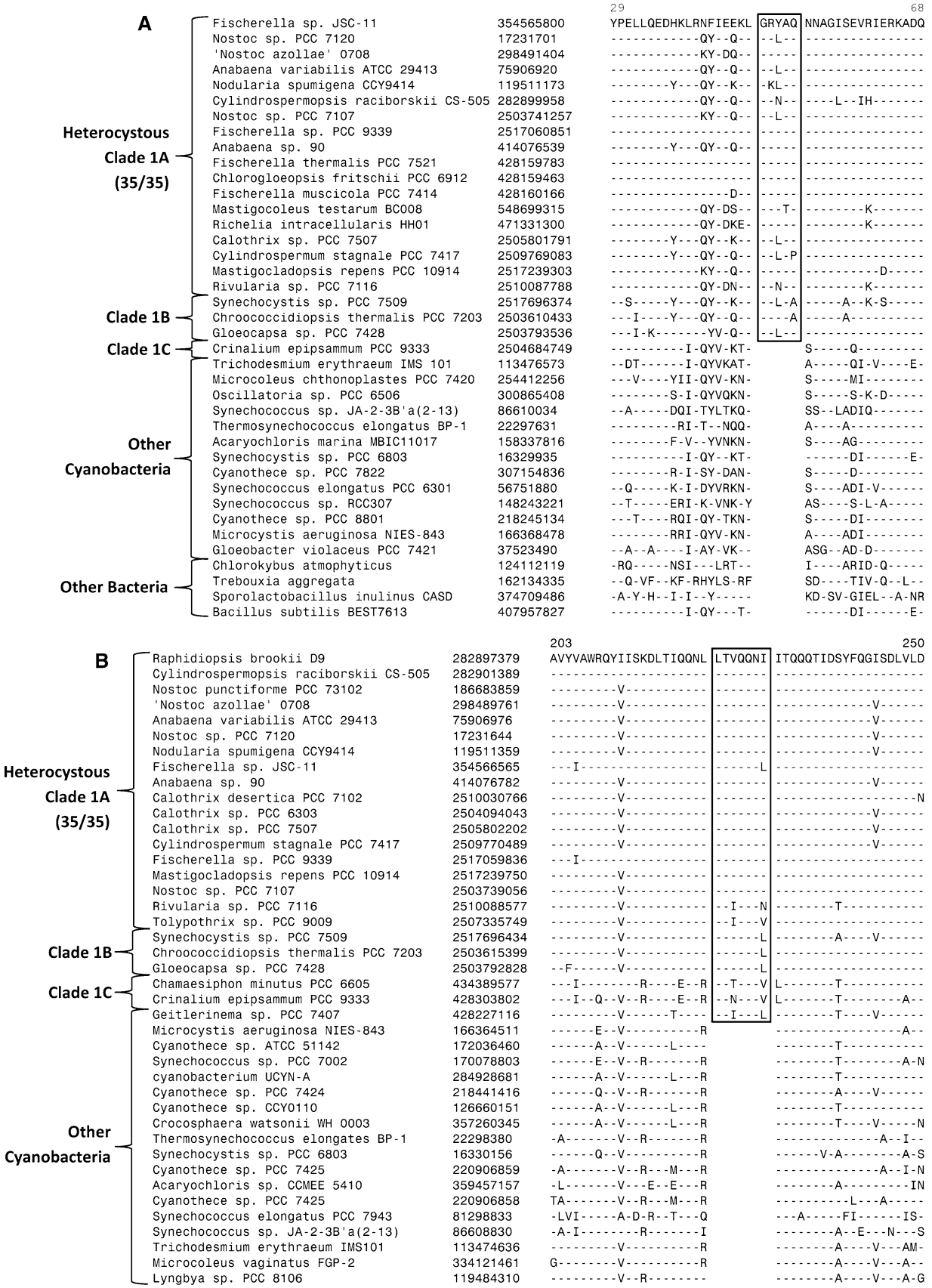
Fig. 5 **a** Partial sequence alignment of the 30S ribosomal protein S3 ▶ showing a 5 aa insert that is specifically present in all heterocystous cyanobacteria and Clade 1B cyanobacteria; **b** Sequence alignment of a pentapeptide repeat protein, showing a 7 aa insert in a conserved region that is commonly shared by all heterocystous cyanobacteria as well as Clade 1B and Clade 1C cyanobacteria. Additionally, this latter insert is also present in *Geitlerinema* sp. PCC 7407. Dashes in the sequence alignments indicate the presence of the same amino acid as shown on the top line. Sequence information for a limited number of species from other cyanobacterial groups is presented here

Table 1 Conserved Signature Indels that are specific for the heterocystous cyanobacteria

Protein Name	GenBank Identifier	Figure #	Indel Size	Indel Position	Specificity (Clade)
XRE family transcriptional regulator	17231636	Fig. 3a	1 aa ins	124–166	1A
Hypothetical protein DUF111	282899760	Fig. 3b	7 aa ins	131–174	1A
30S ribosomal protein S3	354565800	Fig. 4A	5 aa ins	29–68	1A + 1B
Pentapeptide repeat protein[a]	282897379	Fig. 4b	7 aa ins	203–250	1A + 1B + 1C
Outer membrane adhesin (OpcA)[a]	17231510	Sup. Fig. 2	1 aa ins	393–437	1A
Hypothetical protein Npun_R5589	17227740	Sup. Fig. 3	3 aa ins	12–58	1A
Hypothetical protein Aazo_3898	282896270	Sup. Fig. 4	2 aa ins	15–49	1A
PilT domain-containing protein	282900401	Sup. Fig. 5	1 aa del	138–166	1A
Arsenite efflux ATP-binding protein[a]	75909548	Sup. Fig. 6	2 aa ins	150–188	1A
TPR repeat protein[b]	282901375	Sup. Fig. 7	5 aa ins	271–308	1A
2-hydroxy-6-oxohepta-2,4-dienoate hydrolase[b]	17227812	Sup. Fig. 8	1 aa ins	215–246	1A + 1B
Ribonuclease II	282901218	Sup. Fig. 9	14 aa ins	177–237	1A + 1B
ATP synthase epsilon subunit	17232530	Sup. Fig. 10	2 aa ins	36–80	1A + 1B
Hypothetical protein Npun_R4490[b]	17232408	Sup. Fig. 11	1 aa ins	150–191	1A + 1B
Hypothetical protein Npun_R4929[b]	17231417	Sup. Fig. 12	5 aa ins	484–531	1A + 1B + 1C

[a] CSIs are also present in 1-2 homologous proteins from outgroup Cyanobacteria

[b] CSIs are missing from 1-2 members in-group Cyanobacteria


A
Heterocystous Clade 1A (35/35)
Clade 1B
Clade 1C
Other Cyanobacteria
Other Bacteria
B
Heterocystous Clade 1A (35/35)
Clade 1B
Clade 1C
Other Cyanobacteria

supported monophyletic clade, which is in accordance with their distinct branching in the 16S rRNA trees (Wilmotte and Herdman 2001; Gugger and Hoffmann 2004; Yarza et al. 2010). The monophyletic origin of the heterocystous cyanobacteria is also independently strongly supported by eight novel CSIs described here in different proteins, which are uniquely present in all of the heterocystous cyanobacteria (Table 1). The most parsimonious explanation to account for the presence of these CSIs is that the rare genetic changes responsible for them first occurred in a common ancestor of the Clade 1A and these changes were then vertically inherited by its different descendants (Gupta 2009). These CSIs provide novel genetic markers (synapomorphies) for the identification of heterocystous cyanobacteria in molecular terms.

The phylogenetic trees based on both protein sequences and the 16S rRNA gene sequences created in this study also show that the two previously suggested main groups within the heterocystous cyanobacteria (viz. sections IV and V or the orders *Nostocales* and *Stigonematales*) are not monophyletic and that the members of these groups exhibit extensive intermixing. The intermixing of the members of these two orders is also observed in phylogenetic trees based upon 16S rRNA as well as *nifH* and *nifD* gene sequences (Zehr et al. 1997; Wilmotte and Herdman 2001; Gugger and Hoffmann 2004; Henson et al. 2004; Singh et al. 2013). While the division of the heterocystous cyanobacteria into the two main groups, viz. *Nostocales* and *Stigonematales*, is not supported by available evidence, this work has identified a number of novel subclades within the heterocystous cyanobacteria (see Fig. 2). One of these subclades consists of the akinete-forming cyanobacteria (Figs. 2, 3). The distinctness of this subclade of heterocystous cyanobacteria is supported not only by phylogenetic analyses but also by a CSP identified in this work that is mainly limited to this group of cyanobacteria. Within the akinete-forming cyanobacteria, two smaller subclades are also distinguished by both phylogenetic means and by their shared morphological and developmental characteristics (Figs. 2, 3). The members of one of these subclades are capable of differentiating into distinct hormogonia, while the members of the other subclade lack this ability. Apart from these subclades, several deeper branching subclades that grouped the remainder of the heterocystous cyanobacteria are also observed.

The ability of the cells to differentiate into specialized cell types (viz. heterocysts) is found only within a define lineage of cyanobacteria (Rippka et al. 1979; Castenholz 2001). To understand the origin of heterocysts-forming bacteria, it is of much interest to identify their closest relatives within the cyanobacteria. The results of our studies provide important insights in this respect. In the phylogenetic trees based upon concatenated protein sequences, the members from the Clade-1B cyanobacteria are found to be the immediate out group of the heterocysts-forming cyanobacteria (Clade-1A). A close relationship between these two groups was also observed in the tree by Shih et al. (2013). Additionally, another deeper branching clade (Clade 1C) consisting of the two species/strains (viz. *Crinalium epipsammum* PCC 9333 and *Chamaesiphon minutus* PCC 6605) also exhibits a close relationship to the above two clades of cyanobacteria. Strong and independent evidence that the species from the Clade-1B are the closest relatives of the heterocysts-forming cyanobacteria is provided by our identification of five CSIs in proteins involved in different functions that are uniquely shared by the members of these two subclades of (Clade-1A and 1B) cyanobacteria. These results provide strong evidence that the members of the Clade-1B cyanobacteria are the immediate ancestor of the heterocysts-forming cyanobacteria. Additionally, a specific relationship of these two clades of cyanobacteria to the Clade-1C species/strains is also supported by phylogenetic analysis and by two of the identified CSIs. The members of the Clades-1B and 1C are presently very poorly characterized. However, as these are indicated to be the closest relatives of the heterocysts-forming cyanobacteria, it is important to determine what other genetic, physiological or morphological characteristics are commonly shared by the members of these clades and the other heterocystous cyanobacteria.

This work has identified for the first time multiple molecular markers in the forms of CSIs and CSPs that are uniquely shared by the heterocystous cyanobacteria or their closest relatives. The cellular functions of these molecular signatures are presently not known. However, due to the unique presence of these molecular characteristics in these specific lineages of cyanobacteria, it is likely that these genetic changes (or genes) are in some way linked to the unique morphological (viz. heterocysts formation) and associated biochemical characteristics (nitrogen fixation) of these cyanobacteria. Earlier work on a number of CSIs and CSPs, which were specific for other groups, has shown that these molecular characteristics are essential for the groups where they are found and hence serve important functions in the particular groups of bacteria (Fang et al. 2005; Singh and Gupta 2009; Schoeffler et al. 2010). Therefore, studies on understanding the cellular functions of the CSPs and CSIs that are specific for the heterocystous cyanobacteria could provide important insights into the novel biochemical or structural aspects of these bacteria.

Table 2 Conserved signature proteins specific for all *Nostocales* and *Stigonematales* members

Species/Strain	NP_485191 (364 aa)	NP_488605 (92 aa)	NP_485189 (847 aa)	NP_485332[a] (70 aa)
	Expect values (*E* value) of the observed hits			
Nostoc sp. PCC 7120	0	$7e^{-61}$	0	$3e^{-42}$
Anabaena variabilis ATCC 29413	0	$2e^{-57}$	0	$3e^{-41}$
Nostoc punctiforme PCC 73102	0	$2e^{-43}$	0	$1e^{-35}$
Nostoc sp. PCC 7524	0	$2e^{-43}$	0	$5e^{-39}$
Nostoc sp. PCC 7107	0	$2e^{-42}$	0	$7e^{-34}$
Calothrix sp. PCC 7507	0	$2e^{-40}$	0	$4e^{-34}$
Cylindrospermum stagnale PCC 7417	0	$2e^{-45}$	0	$8e^{-38}$
Anabaena cylindrica PCC 7122	$1e^{-178}$	$7e^{-37}$	0	$6e^{-35}$
Microchaete sp. PCC 7126	$3e^{-172}$	$2e^{-36}$	0	$1e^{-37}$
Nostoc azollae' 0708	$6e^{-168}$	$1e^{-37}$	0	$1e^{-35}$
Nodularia spumigena CCY9414	$2e^{-167}$	$8e^{-42}$	0	$2e^{-34}$
Anabaena sp. 90	$1e^{-163}$	$2e^{-34}$	0	$2e^{-33}$
Anabaena sp. PCC 7108	$1e^{-161}$	$1e^{-38}$	0	$1e^{-36}$
Anabaena circinalis AWQC131C	$8e^{-148}$	$7e^{-38}$	0	$5e^{-33}$
Anabaena circinalis AWQC310F	$2e^{-147}$	$6e^{-39}$	0	$5e^{-33}$
Raphidiopsis brookii D9	$2e^{-142}$	$8e^{-33}$	0	$7e^{-34}$
Cylindrospermopsis raciborskii CS-509	$4e^{-141}$	$7e^{-30}$	0	$2e^{-34}$
Cylindrospermopsis raciborskii CS-505	$4e^{-141}$	$6e^{-30}$	0	$1e^{-34}$
Tolypothrix sp. PCC 9009	0	$5e^{-39}$	0	–
Calothrix sp. PCC 7103	$2e^{-139}$	$1e^{-29}$	0	–
Calothrix desertica PCC 7102	$6e^{-137}$	$3e^{-31}$	0	–
Mastigocladopsis repens PCC 10914	$1e^{-108}$	$4e^{-38}$	0	–
Fischerella muscicola PCC 7414	$3e^{-108}$	$2e^{-44}$	0	–
Fischerella sp. PCC 9339	$7e^{-108}$	$2e^{-43}$	0	–
Fischerella sp. JSC-11	$3e^{-106}$	$8e^{-44}$	0	–
Calothrix sp. PCC 6303	$9e^{-101}$	$3e^{-31}$	0	–
Chlorogloeopsis fritschii PCC 6912	$9e^{-99}$	$3e^{-43}$	0	–
Fischerella sp. PCC 9605	$6e^{-98}$	$9e^{-42}$	0	–
cyanobacterium PCC 7702	$2e^{-93}$	$6e^{-37}$	$2e^{-85}$	–
Mastigocoleus testarum BC008	$4e^{-92}$	$3e^{-39}$	0	–
Fischerella sp. PCC 9431	$1e^{-86}$	$4e^{-44}$	0	–
Rivularia sp. PCC 7116	$1e^{-74}$	$3e^{-27}$	0	–
Richelia intracellularis HH01	–	$2e^{-27}$	–	–
Other cyanobacteria with significant hits				
Oscillatoria sp. PCC 10802	$4e^{-125}$	–	0	$2e^{-17}$
Moorea producens 3L	$2e^{-67}$	–	$4e^{-135}$	–
Leptolyngbya boryana	–	$1e^{-15}$	–	–
Microcoleus vaginatus FGP-2	–	–	–	$6e^{-26}$
Oscillatoria nigro-viridis PCC 7112	–	–	–	$8e^{-26}$
Next Best BLAST Hits				
Streptomyces aurantiacus	$5e^{-19}$		$8e^{-47}$	–
Zavarzinella formosa	–	–	$3e^{-19}$	–
Micromonospora sp. ATCC 39149	$5e^{-03}$	–	–	–
Streptomyces sp. FxanaD5	–	$9e^{-01}$	–	–
Frankia sp. EUN1f	–	–	$3e^{-05}$	–
Glycine max	–	–	–	$2e^{00}$

[a] This protein is primarily found in the akinete-forming heterocystous cyanobacteria

Expect values of 0 in the table correspond to *E* value $<1e^{-200}$

Acknowledgments This work was supported by an MRI-ORF Water Round research grant. We thank Mobolaji Adeolu for helpful comments on the manuscript.

References

Adeolu M, Gupta RS (2013) Phylogenomics and molecular signatures for the order Neisseriales: proposal for division of the order Neisseriales into the emended family Neisseriaceae and Chromobacteriaceae fam. nov. Antonie Van Leeuwenhoek 104:1–24

Baldauf SL, Palmer JD (1993) Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. Proc Natl Acad Sci USA 90:11558–11562

Bhandari V, Gupta RS (2014) Molecular signatures for the phylum (class) Thermotogae and a proposal for its division into three orders (*Thermotogales*, *Kosmotogales* ord. nov. and *Petrotogales* ord. nov.) containing four families (*Thermotogaceae*, *Fervidobacteriaceae* fam. nov., *Kosmotogaceae* fam. nov. and *Petrotogaceae* fam. nov.) and a new genus *Pseudothermotoga* gen. nov. with five new combinations. Antonie Van Leeuwenhoek 105:143–168

Castenholz RW (2001) Phylum BX cyanobacteria oxygenic photosynthetic bacteria. In Bergey's manual of systematic bacteriology. Springer, New York, pp 474–487

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

Cavalier-Smith T (2002) The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. Int J Syst Evol Microbiol 52:7–76

Ciccarelli FD, Doerks T, Von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287

De Vos P, Truper HG (2000) Judicial Commission of the International Committee on Systematic Bacteriology; IXth International (IUMS) Congress of Bacteriology and Applied Microbiology. Int J Syst Evol Microbiol 50:2239–2244

Delwiche CF, Kuhsel M, Palmer JD (1995) Phylogenetic analysis of tufA sequences indicates a cyanobacterial origin of all plastids. Mol Phylogenet Evol 4:110–128

Desikachary TV (1959) Cyanophyta, Indian Council of Agricultural Research, monographs on Algae. New Delhi, India

Fang G, Rocha E, Danchin A (2005) How essential are nonessential genes? Mol Biol Evol 22:2147–2156

Gao B, Mohan R, Gupta RS (2009) Phylogenomics and protein signatures elucidating the evolutionary relationships among the Gammaproteobacteria. Int J Syst Evol Microbiol 59:234–247

Geitler L (1932) Cyanophycea. Rabenhorst's Kryptogamen-Flora von Deutschland, Österreich und der Schweiz, Reprint 1985:14

Giovannoni SJ, Turner S, Olsen GJ, Barns S, Lane DJ, Pace NR (1988) Evolutionary relationships among cyanobacteria and green chloroplasts. J Bacteriol 170:3584–3592

Gugger MF, Hoffmann L (2004) Polyphyly of true branching cyanobacteria (Stigonematales). Int J Syst Evol Microbiol 54:349–357

Gupta RS (1998) Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol Mol Biol Rev 62:1435–1491

Gupta RS (2000) The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. FEMS Microbiol Rev 24:367–402

Gupta RS (2003) Evolutionary relationships among photosynthetic bacteria. Photosynth Res 76:173–183

Gupta RS (2009) Protein signatures (molecular synapomorphies) that are distinctive characteristics of the major cyanobacterial clades. Int J Syst Evol Microbiol 59:2510–2526

Gupta RS (2010) Molecular signatures for the main phyla of photosynthetic bacteria and their subgroups. Photosynth Res 104:357–372

Gupta RS (2013) Molecular markers for photosynthetic bacteria and insights into the origin and spread of photosynthesis. Adv Bot Res 66:37–66

Gupta RS, Lali R (2013) Molecular signatures for the phylum Aquificae and its different clades: proposal for division of the phylum Aquificae into the emended order Aquificales, containing the families Aquificaceae and Hydrogenothermaceae, and a new order Desulfurobacteriales ord. nov., containing the family Desulfurobacteriaceae. Antonie Van Leeuwenhoek 104:349–368

Gupta RS, Mathews DW (2010) Signature proteins for the major clades of Cyanobacteria. BMC Evol Biol 10:24

Gupta RS, Pereira M, Chandrasekera C, Johari V (2003) Molecular signatures in protein sequences that are characteristic of cyanobacteria and plastid homologues. Int J Syst Evol Microbiol 53:1833–1842

Gupta RS, Chander P, George S (2013) Phylogenetic framework and molecular signatures for the class Chloroflexi and its different clades; proposal for division of the class Chloroflexi class. nov. into the suborder Chloroflexineae subord. nov., consisting of the emended family Oscillochloridaceae and the family Chloroflexaceae fam. nov., and the suborder Roseiflexineae subord. nov., containing the family Roseiflexaceae fam. nov. Antonie Van Leeuwenhoek 103:99–119

Harris JK, Kelley ST, Spiegelman GB, Pace NR (2003) The genetic core of the universal ancestor. Genome Res 13:407–412

Henson BJ, Watson LE, Barnum SR (2002) Molecular differentiation of the heterocystous cyanobacteria, Nostoc and Anabaena, based on complete NifD sequences. Curr Microbiol 45:161–164

Henson BJ, Hesselbrock SM, Watson LE, Barnum SR (2004) Molecular phylogeny of the heterocystous cyanobacteria (subsections IV and V) based on nifD. Int J Syst Evol Microbiol 54:493–497

Hoffmann L (2005) Nomenclature of Cyanophyta/Cyanobacteria: roundtable on the unification of the nomenclature under the Botanical and Bacteriological Codes. Algol Stud 117:13–29

Hoffmann L, Komarek J, Kastovsky J (2005) System of cyanoprokaryotes (cyanobacteria) state in 2004. Algol Stud 117:95–115

Honda D, Yokota A, Sugiyama J (1999) Detection of seven major evolutionary lineages in cyanobacteria based on the 16S rRNA gene sequence analysis with new sequences of five marine Synechococcus strains. J Mol Evol 48:723–739

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comp Appl Biosci CABIOS 8:275–282

Komárek J (2002) Problems in cyanobacterial taxonomy: implication for most common toxin producing species. Rapporti Istisan 9:6–43

Labeda DP (2000) International committee on systematic bacteriology; IXth international (IUMS) congress of bacteriology and applied microbiology. Int J Syst Evol Microbiol 50:2245–2247

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948

Naushad HS, Lee B, Gupta RS (2014) Conserved signature indels and signature proteins as novel tools for understanding microbial phylogeny and systematics: identification of molecular signatures that are specific for the phytopathogenic genera Dickeya, Pectobacterium and Brenneria. Int J Syst Evol Microbiol 64:366–383

Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press, USA

Oren A (2004) A proposal for further integration of the cyanobacteria under the Bacteriological Code. Int J Syst Evol Microbiol 54:1895–1902

Oren A, Garrity GM (2014) Proposal to change general consideration 5 and principle 2 of the International Code of Nomenclature of Prokaryotes. Int J Syst Evol Microbiol 64:309–310

Oren A, Komarek J, Hoffmann L (2009) Nomenclature of the Cyanophyta/Cyanobacteria/Cyanoprokaryotes: What has happened since IAC Luxembourg? Algol Stud 130:17–26

Parte AC (2014) LPSN-list of prokaryotic names with standing in nomenclature. Nucleic Acids Res 42:D613–D616

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596

Rajaniemi P, Hrouzek P, Kastovska K, Willame R, Rantala A, Hoffmann L, Komarek J, Sivonen K (2005) Phylogenetic and morphological evaluation of the genera Anabaena, Aphanizomenon, Trichormus and Nostoc (Nostocales, Cyanobacteria). Int J Syst Evol Microbiol 55:11–26

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

Rokas A, Holland PW (2000) Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol 15:454–459

Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804

Saatcioglu FAHR, Perry DJ, Pasco DS, Fagan JB (1990) Multiple DNA-binding factors interact with overlapping specificities at the aryl hydrocarbon response element of the cytochrome P450IA1 gene. Mol Cell Biol 10:6408–6416

Sanchez-Baracaldo P, Hayes PK, Blank CE (2005) Morphological and habitat evolution in the Cyanobacteria using a compartmentalization approach. Geobiology 3:145–165

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S (2010) Database resources of the national center for biotechnology information. Nucleic Acids Res 38:D5–D16

Schoeffler AJ, May AP, Berger JM (2010) A domain insertion in *Escherichia coli* GyrB adopts a novel fold that plays a critical role in gyrase function. Nucleic Acids Res 38:7830–7844

Shi T, Falkowski PG (2008) Genome evolution in cyanobacteria: the stable core and the variable shell. Proc Natl Acad Sci 105:2510–2515

Shih PM, Wu D, Latifi A, Axen SD, Fewer DP, Talla E, Calteau A, Cai F, de Marsac NT, Rippka R (2013) Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. Proc Natl Acad Sci 110:1053–1058

Singh B, Gupta RS (2009) Conserved inserts in the Hsp60 (GroEL) and Hsp70 (DnaK) proteins are essential for cellular growth. Mol Genet Genomics 281:361–373

Singh P, Singh SS, Elster J, Mishra AK (2013) Molecular phylogeny, population genetics, and evolution of heterocystous cyanobacteria using nifH gene sequences. Protoplasma 250:751–764

Stanier RY, Sistrom WR, Hansen TA, Whitton BA, Castenholz RW, Pfennig N, Gorlenko VN, Kondratieva EN, Eimhjellen KE, Whittenbury R (1978) Proposal to place the nomenclature of the cyanobacteria (blue-green algae) under the rules of the International Code of Nomenclature of Bacteria. Int J Syst Bacteriol 28:335–336

Swingley WD, Blankenship RE, Raymond J (2008) Integrating Markov clustering and molecular phylogenetics to reconstruct the cyanobacterial species tree from conserved protein families. Mol Biol Evol 25:643–654

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739

Turner SE, Pryer KM, Miao VPW, Palmer JD (1999) Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis 1. J Eukaryot Microbiol 46:327–338

Wilmotte A, Herdman M (2001) Phylogenetic relationships among the cyanobacteria based on 16S rRNA sequences. In: Boone DR, Castenholz RW, Garrity GM (eds) Bergey's manual of systematic bacteriology. Springer, New York, pp 487–493

Woese CR (1992) Prokaryote systematics: the evolution of a science. In: Balows A, Trüper HG, Dworkin M, Harder W, Schleifer KH (eds) The prokaryotes. Springer, New York, pp 3–18

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462:1056–1060

Yarza P, Ludwig W, Euzeby J, Amann R, Schleifer KH, Glockner FO, Rossello-Mora R (2010) Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. Syst Appl Microbiol 33:291–299

Zehr JP, Mellon MT, Hiorns WD (1997) Phylogeny of cyanobacterial nifH genes: evolutionary implications and potential applications to natural assemblages. Microbiology 143:1443–1450

Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT (2006) Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res 16:1099–1108