Protein signatures (molecular synapomorphies) that are distinctive characteristics of the major cyanobacterial clades

Radhey S. Gupta

Correspondence
Radhey S. Gupta
gupta@mcmaster.ca

Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario L8N 3Z5, Canada

A combination of phylogenomic and signature sequence-based (or phenetic) approaches was used to understand the evolutionary relationships among cyanobacteria. Phylogenetic trees were constructed for 34 cyanobacteria whose genomes have been sequenced, based on concatenated sequences for 45 conserved proteins and also the 16S rRNA gene. In parallel, sequence alignments of various proteins were examined to identify conserved indels (i.e. molecular signatures or synapomorphies) that are specific for either all cyanobacteria or their various clades in the phylogenetic trees. Of the >40 molecular signatures described in this work, 15 are specific for all cyanobacteria. The other cyanobacterial clades that can now be identified and circumscribed in molecular terms by using these signatures include a deep-branching clade (clade A, corresponding to the subclass Gloeobacterophycidae), consisting of *Gloeobacter violaceus* and two diazotrophic *Synechococcus* strains (JA-3-3Ab and JA2-3-B'a) (15 aa insert in EF-G); a clade comprising all other cyanobacteria except those from clade A [18 aa insert in DNA polymerase I (Pol I), 2 aa insert in the DnaX protein, 4 aa insert in TrpRS and 4–5 aa insert in tryptophan synthase beta subunit]; a clade (clade C, corresponding to the subclass Synechococcophycidae) of various marine unicellular *Synechococcus* and *Prochlorococcus* cyanobacteria (12 aa insert in Pol I, 3 aa insert in RpoB, 2 aa insert in KgsA, 6 aa insert in TyrRS, 2 aa insert in tRNA-mG1 transferase and 1 aa deletion in the RpoC protein); a clade of the low-B/A ecotype *Prochlorococcus* strains (5 aa deletion in LeuRS and 1 aa insert in the Ffh protein); a clade consisting of the *Nostocales* species/strains (subclass Nostocophycidae; 4 aa insert in the PetA protein and 5 aa insert in the ribosomal protein S3); a clade of the order *Chroococcales* (1 aa insert in RecA); a clade comprising the orders *Nostocales*, *Oscillatoriales* and *Chroococcales* [19 aa insert in DnaE, 13 aa insert in GDP-mannose pyrophosphorylase and 22–27 aa insert in NADP(H)-quinone oxidoreductase subunit D]. Two additional conserved indels in the translation-initiation factor IF-2 and riboflavin synthase alpha subunit suggest an intermediate placement of the *Oscillatoriales* in between the orders *Nostocales* and *Chroococcales*. The unique presence of these molecular signatures in all available sequences from the indicated groups of cyanobacteria, but not in any other cyanobacteria (or bacteria), indicates that these synapomorphies provide novel and potentially useful means for circumscription of several important taxonomic clades of cyanobacteria in more definitive terms. The species-distribution patterns of these synapomorphies also indicate that the plant/plastid homologues are not derived from the clade A or C cyanobacteria.

INTRODUCTION

Cyanobacteria form one of the main phyla within *Bacteria* and they are the sole prokaryotic group capable of carrying out oxygenic photosynthesis (Kondratieva *et al.*, 1992;

Abbreviations: ML, maximum-likelihood; NJ, neighbour-joining; Pol I, DNA polymerase I.

A list of proteins used in phylogenetic analyses and 22 supplementary figures are available with the online version of this paper.

Castenholz, 2001). They exhibit enormous diversity in terms of their morphology (cell size, shape and arrangement, e.g. filamentous, unicellular or colonial), physiology (e.g. ability to fix nitrogen, use of sulfide as an electron donor) and other characteristics (e.g. motility, thermophily) (Rippka *et al.*, 1979; Anagnostidis & Komarek, 1985; Wilmotte & Golubic, 1991; Kondratieva *et al.*, 1992; Castenholz, 2001; Sánchez-Baracaldo *et al.*, 2005). However, most of these characteristics in the past have

not shown good correlation with cyanobacterial phylogeny based on the 16S rRNA gene or other gene/protein sequences (Honda *et al.*, 1999; Turner *et al.*, 1999; Ishida *et al.*, 2001; Robertson *et al.*, 2001; Wilmotte & Herdman, 2001; Gugger & Hoffmann, 2004; Sánchez-Baracaldo *et al.*, 2005; Swingley *et al.*, 2008a). Thus, it remains to be determined whether some of them will provide reliable markers for understanding cyanobacterial taxonomy or evolution. In 16S rRNA gene-based trees, which provide the primary means for taxonomic and evolutionary studies (Olsen & Woese, 1993; Honda *et al.*, 1999; Turner *et al.*, 1999; Wilmotte & Herdman, 2001), cyanobacteria group into 14 clusters (Wilmotte & Herdman, 2001). However, the taxonomic significances of these clusters or how they are related evolutionarily is not clear (Honda *et al.*, 1999; Turner *et al.*, 1999; Wilmotte & Herdman, 2001; Hoffmann, 2005). Apart from the 16S rRNA gene, sequence information for cyanobacteria for other genes/proteins was, until recently, quite limited; however, phylogenetic analyses based on them generally indicated results very similar to those seen with the 16S rRNA gene (Giovannoni *et al.*, 1988; Delwiche *et al.*, 1995; Honda *et al.*, 1999; Henson *et al.*, 2002; Seo & Yokota, 2003; Zhaxybayeva *et al.*, 2006; Shi & Falkowski, 2008). Many of the problems that are currently encountered in understanding cyanobacterial phylogeny can be attributed to their outdated taxonomy and nomenclature, which have been governed in the past by both the Botanical and Bacteriological Codes (Rippka *et al.*, 1979; Anagnostidis & Komarek, 1985; Castenholz, 2001; Oren, 2004; Garrity *et al.*, 2005). Thus, despite the fact that cyanobacteria constitute a large phylum containing >4000 isolates (Maidak *et al.*, 2001), only a small number of species and higher taxonomic groups within this phylum have validly published names under the Bacteriological Code (Oren, 2004; Garrity *et al.*, 2005; Hoffmann, 2005). At the 16th Symposium of the International Association for Cyanophyte Research held in 2004, which considered some of these problems, a number of recommendations for improvement of cyanobacterial nomenclature and for their integration under the Bacteriological Code were proposed (Oren, 2004; Hoffmann, 2005; Oren & Tindall, 2005). At this meeting, a new proposal for the classification of cyanobacteria into four subclasses, viz. Gloeobacteriophycidae, Synechococcophycidae, Oscillatoriophycidae and Nostocophycidae, was also made (Hoffmann *et al.*, 2005).

The availability of genome sequences in recent years has provided new opportunities for understanding cyanobacterial taxonomy and evolution. Based upon genomic sequences, an important and frequently used approach is to assemble phylogenies based upon combined sequences for a large number of proteins (Eisen, 1998; Ciccarelli *et al.*, 2006). Phylogenies based upon a large number of characters derived from multiple conserved (or slow-evolving) genes/proteins are better able to resolve deeper-branching evolutionary relationships than those based on

any single gene or protein (Hansmann & Martin, 2000; Brown *et al.*, 2001; Rokas *et al.*, 2003; Delsuc *et al.*, 2005; Ciccarelli *et al.*, 2006). Recently, a number of studies have reported phylogenomic analysis for limited numbers of cyanobacteria (between 11 and 24) based upon concatenated sequences for different large sets of proteins present in various cyanobacteria (Giovannoni *et al.*, 1988; Honda *et al.*, 1999; Wilmotte & Herdman, 2001; Seo & Yokota, 2003; Sánchez-Baracaldo *et al.*, 2005; Shi & Falkowski, 2008; Swingley *et al.*, 2008a). These studies have been very useful in confirming the existence of certain important clades within cyanobacteria and also in clarifying their relative branching positions. However, based upon these studies, one cannot circumscribe various cyanobacterial clades in definitive terms, which is a prerequisite for developing reliable or stable taxonomy (Oren & Stackebrandt, 2002; Hoffmann *et al.*, 2005). This approach is also limited to species whose genomes have been sequenced.

Hence, it is important to identify other reliable molecular markers that are consistent with the results of 16S rRNA gene-based trees as well as phylogenomic and other approaches, but which also enable circumscription of the major cyanobacterial clades in more definitive (molecular) terms. One approach that has been very useful in this regard involves identifying molecular synapomorphies consisting of conserved inserts or deletions (i.e. indels) in widely distributed proteins that are distinctive characteristics of various main clades within a given phylum (Rivera & Lake, 1992; Delwiche *et al.*, 1995; Gupta, 1998, 2000; Griffiths & Gupta, 2007). This approach has been used successfully in our work to elucidate the branching order and inter-relationships within different main phyla of bacteria (Gupta *et al.*, 1999; Griffiths & Gupta, 2001; Gupta, 2001, 2003), as well as a number of main groups (or phyla) of bacteria including *Alphaproteobacteria* (Gupta & Mok, 2007), *Chlamydiae* (Gupta & Griffiths, 2006) and the *Bacteroidetes–Chlorobi* group (Gupta, 2004). By using this approach, we have previously described 14 conserved indels in 10 widely distributed proteins that were distinctive characteristics of various available cyanobacteria, but not found in any other bacteria (Gupta *et al.*, 2003). The shared presence of several of these indels by the plastid/plant homologues (Gupta *et al.*, 2003) also provided evidence for their cyanobacterial ancestry (Margulis, 1970; Gray & Doolittle, 1982; Morden *et al.*, 1992; Whatley, 1993; Palmer & Delwiche, 1998). In this work, we describe 28 novel conserved indels in many important and widely distributed proteins that are distinctive characteristics of various main clades of cyanobacteria that can be identified by means of phylogenetic analyses. Because of their specific presence in these clades, these conserved indels or synapomorphies also provide new means for potential circumscription of a number of important taxonomic groups of cyanobacteria in molecular terms and for understanding cyanobacterial evolution.

METHODS

Phylogenetic analyses. Phylogenetic analyses were carried out on a set of proteins involved in important housekeeping functions that are present in most organisms (Harris *et al.*, 2003). BLAST searches were conducted on each of these proteins to determine whether their homologues were present in all 34 cyanobacteria listed in Table 1 and the two outgroup species (*Bacillus subtilis* and *Staphylococcus aureus*) used in this work. Except for *Crocospaera watsonii* WH8501, the genomes of all other cyanobacteria are now available (Table 1). For the 45 proteins for which information is presented in Supplementary Table S1 (available in IJSEM Online), orthologues were available from

all 34 cyanobacteria and the two outgroup species. The multiple sequence alignments for these proteins were created by using the CLUSTAL_X 1.83 program (Jeanmougin *et al.*, 1998) and they were concatenated into a single large file. This unedited sequence alignment was imported into the Gblocks 0.91b program (Castresana, 2000) to remove poorly aligned regions. This program was used with default settings except the allowed gap position parameter was changed to 0.5. The resulting final alignment of 16 822 amino acid sites was used for phylogenetic analyses. A neighbour-joining (NJ) tree based on 1000 bootstrap replicates was constructed by the Kimura model (Kimura, 1983) using the TREECON 1.3b program (Van de Peer & De Wachter, 1994). The maximum-

Table 1. Cyanobacterial genomes studied in this work

Abbreviations: DOE JGI, Department of Energy Joint Genome Institute; GBM, Gordon and Betty Moore; TIGR, The Institute of Genome Research. The genome of *Crocospaera watsonii* WH8501 was not sequenced fully.

Species/strain	Genome size (Mb)	DNA G+C content (mol%)	No. proteins*	GenBank accession no.	Centre/reference
<i>Acaryochloris marina</i> MBIC11017	8.36	47.0	6254	NC_009925.1	Swingley <i>et al.</i> (2008b)
<i>Anabaena variabilis</i> ATCC 29413	7.07	41.4	5043	NC_007413.1	DOE JGI
<i>Gloeobacter violaceus</i> PCC7421	4.66	62.0	4430	NC_005125.1	Nakamura <i>et al.</i> (2003)
<i>Nostoc</i> sp. PCC7120	7.21	41.3	5366	NC_003272.1	Kaneko <i>et al.</i> (2001)
<i>Nostoc punctiforme</i> PCC73102	8.20	41.4	6087	NC_010628	DOE JGI
<i>Prochlorococcus marinus</i>					
AS9601	1.70	31.3	1921	NC_008816.1	J. Craig Venter Institute
MIT9211	1.70	39.7	1855	NC_009976.1	Rocap <i>et al.</i> (2003)
MIT9215	1.70	31.0	1983	NC_009840.1	DOE JGI
MIT9301	1.60	31.3	1907	NC_009091.1	GBM Foundation
MIT9303	2.70	50.0	2997	NC_008820.1	J. Craig Venter Institute
MIT9312	1.71	31.2	1810	NC_007577.1	DOE JGI
MIT9313	2.41	50.7	2269	NC_005071.1	Rocap <i>et al.</i> (2003)
MIT9515	1.70	30.8	1906	NC_008817.1	J. Craig Venter Institute
NATL1A	1.90	35.0	2193	NC_008819.1	J. Craig Venter Institute
NATL2A	1.80	35.1	2163	NC_007335.2	DOE JGI
<i>Prochlorococcus marinus</i> subsp. <i>marinus</i>					
CCMP1375	1.75	36.4	1883	NC_005042.1	Dufresne <i>et al.</i> (2003)
CCMP1986	1.70	30.8	1717	NC_005072.1	Rocap <i>et al.</i> (2003)
<i>Synechococcus elongatus</i>					
PCC6301	2.70	55.5	2527	NC_006576.1	Sugita <i>et al.</i> (2007)
PCC7942	2.75	55.4	2612	NC_007604.1	DOE JGI
<i>Synechococcus</i> sp.					
CC9311	2.61	52.4	2892	NC_008319.1	Palenik <i>et al.</i> (2006)
CC9605	2.51	59.2	2645	NC_007516.1	Dufresne <i>et al.</i> (2008)
CC9902	2.23	54.2	2307	NC_007513.1	Dufresne <i>et al.</i> (2008)
JA-2-3B'a(2-13)	3.05	58.5	2862	NC_007776.1	TIGR
JA-3-3Ab	2.93	60.2	2760	NC_007775.1	TIGR
RCC307	2.20	60.8	2535	NC_009482.1	Dufresne <i>et al.</i> (2008)
RS9917	2.58	64.8	2820	–	Dufresne <i>et al.</i> (2008)
WH5701	2.86	66.0	3129	–	Dufresne <i>et al.</i> (2008)
WH7803	2.40	60.2	2533	NC_009481.1	Dufresne <i>et al.</i> (2008)
WH7805	2.62	57.5	2934	–	Dufresne <i>et al.</i> (2008)
WH8102	2.43	59.4	2519	NC_005070.1	Palenik <i>et al.</i> (2003)
<i>Synechocystis</i> sp. PCC6803	3.95	47.4	3172	NC_000911.1	Kaneko <i>et al.</i> (1996)
<i>Thermosynechococcus elongatus</i> BP-1	2.59	53.9	2476	NC_004113.1	Nakamura <i>et al.</i> (2002)
<i>Trichodesmium erythraeum</i> IMS101	7.80	34.1	4451	NC_008312.1	DOE JGI

*Total number of proteins or open reading frames in the genome.

likelihood (ML) analysis was carried out by using the WAG + F model with gamma distribution of evolutionary rates with four categories by using the TREE-PUZZLE program with 10 000 puzzling steps (Schmidt *et al.*, 2002). The 16S rRNA gene sequences for the same cyanobacteria were obtained from their genomic sequences. A multiple sequence alignment of them was created by using the CLUSTAL_X 1.83 program. An NJ tree based on 1000 bootstrap replicates was constructed based on distances calculated using Kimura's two-parameter model (Kimura, 1980) using the TREECON 1.3b program. This tree was also rooted by using *B. subtilis* and *S. aureus* sequences.

Identification of conserved indels that are specific for various cyanobacterial clades. To identify conserved indels that are specific for cyanobacteria, multiple sequence alignments of different proteins made in our earlier work (Gupta, 1998, 2004; Gupta & Griffiths, 2006; Gupta & Mok, 2007) were examined. BLAST searches were also performed on many other conserved proteins from the genome of *Nostoc* sp. PCC7120 to retrieve cyanobacterial homologues, as well as high-scoring homologues from other bacteria, to generate multiple sequence alignments. All of these sequence alignments were inspected visually to identify conserved indels that were restricted to cyanobacteria (Gupta *et al.*, 2003). The indels that were not flanked by conserved regions were excluded from further consideration, as they do not provide reliable molecular markers (Gupta, 1998). The species distribution of all conserved indels was evaluated by carrying out detailed BLASTP searches on short sequence segments (between 60 and 150 aa depending upon the lengths of the indels) containing the indels and their flanking conserved regions. In cases where the indels were large, two separate BLASTP searches were performed, one with a sequence containing the indel and the other with a sequence lacking the indel, to determine the presence or absence of the indel in various species/strains. The sequence information for various conserved indels from representative cyanobacteria and some other bacteria was compiled into signature files. Due to space consideration, sequences for some closely related *Prochlorococcus* and *Synechococcus* strains are not shown in several of the alignment figures.

To infer whether a given indel is an insert or a deletion in cyanobacteria or their particular groups, the presence of this indel in other cyanobacteria, as well as other phyla of bacteria, was determined. If this indel was lacking in these groups, thereby indicating that the absence of indel is the ancestral character state, then the indel in question was interpreted as an insert in the given group of cyanobacteria. On the other hand, if all of these bacteria contained the particular sequence region that was absent in a given clade of cyanobacteria, then the indel was considered a deletion.

RESULTS

Phylogenomic/phylogenetic analyses on cyanobacteria

Genome sequences for a large number of cyanobacteria have become available in recent years (Table 1). Phylogenetic analysis on these cyanobacteria was carried out based on a concatenated sequence alignment for 45 widely distributed proteins (see Methods). A rooted NJ tree based on this dataset is shown in Fig. 1 and an ML tree based on these sequences is provided as Supplementary Fig. S1, available in IJSEM Online. The branching patterns of cyanobacteria in these trees are very similar and these results are comparable to those observed in other recent studies based on different large datasets of protein

sequences (Sánchez-Baracaldo *et al.*, 2005; Shi & Falkowski, 2008; Swingley *et al.*, 2008a). In both NJ and ML trees, a clade consisting of *Gloeobacter violaceus* and *Synechococcus* spp. (JA-3-3Ab and JA2-3-B'a) (referred to here as clade A) showed the deepest branching within the cyanobacteria. The deep branching of clade A species has also been observed in a number of earlier studies (Giovannoni *et al.*, 1988; Honda *et al.*, 1999; Wilmotte & Herdman, 2001; Seo & Yokota, 2003; Sánchez-Baracaldo *et al.*, 2005; Shi & Falkowski, 2008; Swingley *et al.*, 2008a). According to a recent proposal for the classification of cyanobacteria (Hoffmann *et al.*, 2005), the *Gloeobacterales* are placed into a separate subclass (Gloeobacterophycidae). Most other cyanobacteria in our dataset can be grouped into two major clades. One of these clades (designated clade B) comprises diverse cyanobacteria such as *Thermosynechococcus*, *Synechocystis*, *Crocospaera*, *Acaryochloris*, *Trichodesmium*, *Nostoc* and *Anabaena* spp., and it includes both the subclasses Oscillatoriophycidae and Nostocophycidae of cyanobacteria (Hoffmann *et al.*, 2005). In contrast, the other main clade (clade C) is composed entirely of different strains/isolates of *Prochlorococcus* and *Synechococcus*. This latter clade corresponds to the subclass Synechococcophycidae (Hoffmann *et al.*, 2005), except that our analysis indicates that *Acaryochloris* is not part of this clade or subclass. Based on their genetic distances (Figs 1 and 2), the clade C isolates are genetically not as diverse as the taxa in clade B. Within clade C, two subclades can be distinguished. One of these subclades consisted entirely of the *Prochlorococcus* strains/ecotypes, whereas the other subclade mainly comprised various marine *Synechococcus* strains, except for the branching of *Prochlorococcus marinus* MIT9303 and MIT9313 within them. In clade B, a number of subclades including those corresponding to *Nostoc/Anabaena* (order Nostocales) and *Synechocystis/Crocospaera* (order Chroococcales) were resolved with high statistical support (i.e. supported by >90% puzzled quartets or bootstrap samples). The branching position of *Synechococcus elongatus* was uncertain in these trees; in the ML tree (Supplementary Fig. S1), it branched in between clades B and C, whereas in the NJ tree (Fig. 1), it branched with clade B, although this association lacked statistical support.

An NJ tree for these cyanobacteria was also constructed based on 16S rRNA gene sequences (Fig. 2). This tree consisted of only two main clades, one consisting of the clade A and B cyanobacteria as well as *Synechococcus elongatus*, whilst the clade C species/strains/isolates comprised the other clade. In contrast to the tree based on the concatenated protein sequences (Fig. 1), the deep branches in the 16S rRNA gene-based tree, including that indicating the branching of *Thermosynechococcus elongatus* with clade A, were not supported statistically. These trees provide us with a framework for understanding the evolutionary significances of various conserved indels identified in this work.

Fig. 1. NJ distance tree for various cyanobacteria whose genomes are now available, based on concatenated sequences for 45 conserved proteins. The tree was rooted by using *Bacillus subtilis* and *Staphylococcus aureus* sequences. Numbers at nodes indicate bootstrap scores out of 1000. Bar, 0.1 substitutions per site.

Fig. 2. NJ distance tree for the same group of cyanobacteria as shown in Fig. 1, based on 16S rRNA gene sequences. The tree was bootstrapped 1000 times and numbers at nodes indicate percentage bootstrap scores. Bar, 0.1 substitutions per site.

Usefulness of signature indels (synapomorphies) for taxonomic/evolutionary studies

The shared derived characters that are unique to particular groups or clades of organisms (i.e. synapomorphies) provide an important means of identifying various monophyletic clades and also for understanding how these clades are related to each other. In the past, this approach has largely been employed by using morphology and other observable traits (Sneath & Sokal, 1973; Sneath, 2001). However, such traits are often either plesiomorphic (i.e. a particular character is not limited to a given group) or exhibit homoplasy (the derived character state has evolved independently in the given group of organisms), limiting their utility as phylogenetic or taxonomic markers. In recent years, the availability of genome sequences has led to the discovery of important molecular characteristics that are uniquely shared by different groups of organisms and provide important means for their grouping and for evolutionary studies. The conserved indels of defined lengths that are present in gene/protein sequences at specific positions, and which are uniquely shared by particular groups of organisms, form an important class of such characters (synapomorphies) and have proven very useful in clarifying many important evolutionary relationships (Rivera & Lake, 1992; Baldauf & Palmer, 1993; Gupta, 1998; Rokas & Holland, 2000). When a conserved indel of defined length and sequence is found specifically in a particular group of species, then the simplest and most parsimonious explanation for this observation is that these species shared a common ancestor, in which the rare genetic change responsible for this indel first occurred and was then passed on to various descendants (Rivera & Lake, 1992; Gupta, 1998; Rokas & Holland, 2000). However, such changes can also occur due to lateral gene transfers between or among species (Boucher *et al.*, 2003; Zhaxybayeva *et al.*, 2006); hence, it is important to exclude this possibility by comparing the inference obtained by the signature approach with that derived from other independent means, such as phylogenetic or phylogenomic methods. I describe below many conserved indels or molecular synapomorphies that are distinctive characteristics of various clades of cyanobacteria.

Conserved indels that are specific for cyanobacteria and their deepest-branching clades

Previously, we described 14 conserved indels in 10 proteins that are specific for all cyanobacterial homologues (Gupta *et al.*, 2003). Since that time, sequence information for many cyanobacteria and a large number of other bacteria has become available. Hence, it was of importance to check the species specificities of these signatures. Of these 14 indels, 13 (including the 6, 7 and 28 aa inserts in the DNA helicase II protein, 14 aa insert in ADP-glucose pyrophosphorylase, 3 aa insert in FtsH protein, 11–13 aa insert in phytoene synthase, 5 aa insert in EF-Tu, 2 and 7 aa deletions in the ribosomal S1 protein, 2 aa insert in SecA

protein, 6 aa insert and 1 aa deletion in inosine-5'-monophosphate dehydrogenase and 1 aa insert in the major sigma factor) were present in all sequenced (available) cyanobacteria, but not found in any other bacteria (results not shown). In this work, I have identified two additional conserved indels, viz. a 3 aa deletion in carbamoylphosphate synthase (Supplementary Fig. S2) and a 3 aa insert in prolyl-tRNA synthetase (Supplementary Fig. S3), that are also specific for cyanobacteria and are not found in any other bacteria. In prolyl-tRNA synthetase, a 2 aa insert is also found in various chlamydial species in the same position where the 3 aa insert is present in cyanobacteria. Earlier studies have indicated the presence of chlamydia-like genes in green plants and algae, but not in cyanobacteria (Horn *et al.*, 2004; Huang & Gogarten, 2007; Moustafa *et al.*, 2008). In a phylogenetic tree for ProRS sequences, the chlamydiae and cyanobacteria do not group together (results not shown), indicating that the inserts in these two groups, which are of different lengths, have originated independently. Because of their cyanobacterial specificities, the genetic changes responsible for these molecular signatures probably occurred in a common ancestor of cyanobacteria, as indicated in Fig. 3.

In contrast to the above indels, the 18 aa insert in DNA polymerase I (Pol I), identified previously (Gupta *et al.*, 2003), was present in all other cyanobacteria except those corresponding to clade A (Fig. 4). This indel was also absent in various other bacteria, indicating that this synapomorphy is an insert in other cyanobacteria that was introduced after the divergence of clade A (see Fig. 3). In addition to the Pol I signature, we have identified several other conserved indels that support a similar relationship. These include a 4–5 aa insert in the tryptophan synthase beta chain (Supplementary Fig. S4), a 4 aa insert in tryptophanyl-tRNA synthetase (TrpRS) (Supplementary Fig. S5) and a 2 aa insert in the DNA polymerase III gamma/tau subunits (encoded by the *dnaX* gene) (Supplementary Fig. S6). The conserved inserts in these proteins are commonly shared by all other cyanobacteria, but they are absent in the clade A species/strains as well as other bacteria. These signatures/synapomorphies support the view that the clade A species/strains constitute the deepest-branching lineage within cyanobacteria (Fig. 3), which is in accordance with the phylogenetic tree based upon our protein dataset (Fig. 1; Supplementary Fig. S1) and also with earlier phylogenomic studies based on other datasets of proteins (Honda *et al.*, 1999; Turner *et al.*, 1999; Sánchez-Baracaldo *et al.*, 2005; Shi & Falkowski, 2008; Swingley *et al.*, 2008a). Interestingly, the insert in the tryptophan synthase beta chain (Supplementary Fig. S4), which is absent in the clade A cyanobacteria but found in all other cyanobacteria, is also found in various plant and algal homologues. The shared presence of this synapomorphy in the plant/algal homologues and various cyanobacteria, except those from clade A, provides evidence that the plant/algal homologues of this protein are derived from a cyanobacterium not belonging to clade A.

Fig. 3. Interpretive diagram indicating the evolutionary stages where different conserved indels specific for either all cyanobacteria or those distinguishing the clade A cyanobacteria from others were probably introduced. These signatures provide evidence for the deepest branching of the clade A species/strains within cyanobacteria. The signatures marked * were described in our earlier work (Gupta *et al.*, 2003) and the present work confirms that they are specifically present in all sequenced cyanobacteria.

We have also identified a 15 aa insert in a highly conserved region of the protein synthesis elongation factor-G (EF-G) that is found only in the clade A cyanobacteria (Fig. 5). This synapomorphy provides a molecular marker for the clade A species/strains (subclass *Gloeobacterophycidae*) and, in conjunction with other signatures described above (Fig. 3), provides evidence that this group is distinct from all other cyanobacteria. A number of conserved inserts in other proteins are specifically present in the two *Synechococcus* strains JA-3-3Ab and JA2-3-B'a that are part of clade A. These include a 12 aa insert in the protein serine hydroxymethyltransferase (GlyA), an 8 aa insert in the RNA polymerase beta subunit (RpoB), a 6 aa insert in the ribosomal L2 protein and a 2 aa insert in the prolyl-tRNA synthetase (Supplementary Figs S7-S10).

Signature indels for the clade C cyanobacteria

As indicated earlier, clade C is composed entirely of different strains/isolates and ecotypes of *Prochlorococcus* and *Synechococcus*. Of these, *Synechococcus* spp. are ubiquitous in different aquatic environments, including estuarine, coastal and offshore waters (Palenik *et al.*, 2006).

In contrast, *Prochlorococcus* spp. are found mainly in warm, oligotrophic oceanic settings (Rocap *et al.*, 2003). We have identified a number of conserved indels in widely distributed proteins that are specific for the clade C species/strains (subclass *Synechococcophycidae*) and also for one of its subgroups. The signatures that are specific for clade C include a 3 aa insert in the RNA polymerase β subunit (RpoB; Fig. 6), a 2 aa insert in the protein KsgA (Supplementary Fig. S11) that carries out essential dimethylation of the 16S rRNA, a 6 aa insert in the tyrosyl-tRNA synthetase (Supplementary Fig. S12), a 2 aa insert in the tRNA (guanine-N1-)-methyltransferase (Supplementary Fig. S13) and a 1 aa insert in the RNA polymerase β' subunit (RpoC; Supplementary Fig. S14). The inserts in all of these proteins are present in all of the clade C species/strains (sequences for some closely related strains of *Prochlorococcus* and *Synechococcus* strains are not shown in these alignments), but they are not found in other cyanobacteria, including *Acaryochloris marina*, which has also been suggested to belong to the subclass *Synechococcophycidae* (Hoffmann *et al.*, 2005). Another prominent signature for clade C (a 12 aa insert) is found in Pol I (Supplementary Fig. S15), which, as indicated

			845		910
Other cyano	<i>Nostoc</i> PCC7120	NP_485297	GYVETILGARRYFDFTNNSLRKL	KGSKPEDIDL SKLNLGPF	DAGLLRSAANAPIQSSADIITKA
	<i>Nostoc</i> PCC73102	ZP_00107254	-----N-----R-----	-----N-----	-----N-----
	<i>Anabaena variabilis</i>	YP_321100	-----N-DST--LN-----	-----N-----AK-----	-----N-----
	<i>Lyngbya</i> PCC8106	ZP_01624605	---T-L-----H-ESP--KA--	-E-PL---F-DE--KT-QN	---QF--A-----
	<i>Nodularia spumigena</i>	ZP_01631596	-----V--K-D-----	-----N-----E-----	-----N-----
	<i>Synechocystis</i> PCC6803	NP_442677	---T-V-----N-VTEA--Q-	R-KTVTEI--VDV-MNYN	---Q-----
	<i>Crocospheara watsonii</i>	ZP_00516744	---T-----N-FDDN-YRF	R-HD-QN---N-ININYN	-SQ---A-----
	<i>Cyanothece</i> CCY0110	ZP_01729743	---T-----N-FDDN-YRF	R-HD-QT---NQ-NINYN	-SQ---A-----
	<i>Microcystis aeruginosa</i>	YP_001656825	-----N-D-PL-KR-	R-LSLH---DS--LNNE	E-Q-----V-
	<i>Acaryochloris marina</i>	YP_001516479	-----E-DSRG-K-Y	R-KAL-ELAEVD-SDIKM*	-R---A-----L---
	<i>Thermosynechococcus elongatus</i>	NP_682129	-----A-ESRE-QS-	R-KPLDVLADVDPK-KM*	ER---A-----C-
	<i>Trichodesmium erythraeum</i>	YP_720469	-----ELESKTI-D-	-DKE--E-N-QE--FQGN	---QI-----S-----
	<i>Synechococcus elongatus</i> PCC7942	YP_399213	---V-----EDTG-Q--	R-D--S--D-IRPSRF	E-Q--A-----V-
	<i>Synechococcus elongatus</i> PCC6301	YP_172027	---V-----EDTG-Q--	R-D--S--D-IRPSRF	E-Q--A-----V-
	<i>Synechococcus</i> WH8102	NP_896795	-----P-H-DR-G-GR-	L-KE-LE---DVARRG-M	E-QQ--A-----V-
	<i>Synechococcus</i> RS9916	ZP_01470390	-----P-H-DR-G-GR-	L-KD-LE---DVARRG-M	E-QQ--A-----L-
	<i>Synechococcus</i> WH5701	ZP_01085666	-----P-A-DPGG-GR-	R-KP--E-E-EVARRA-M	E-QQ--A-----L-
	<i>Synechococcus</i> RS9917	ZP_01080692	-----P-H-DR-G-GR-	L-KD-LE---EVARRG-M	E-QQ--A-----L-
	<i>Synechococcus</i> WH7803	YP_001225340	---M-----P-H-DR-G-GR-	L-KD-FE---DVARRG-M	E-QQ--A-----L-
	<i>Synechococcus</i> RCC307	YP_001227925	---LM-----P-H-DPQG-GR-	--TD-AE---DVARRG-M	E-QQ--A-----K-
	<i>Prochlorococcus marinus</i> CCMP1375	NP_875626	---Q-L-----N-DK-G-GR-	L--D-MN---KRAARRA-M	E-QQ--A-----
	<i>Prochlorococcus marinus</i> NATL1A	YP_001015410	-F---L-----H-NK-G-GR-	L-TP-NE---TTARRA-M	E-QQ--A-----L-
	<i>Prochlorococcus marinus</i> MIT9211	YP_001551099	-F-----H-H-DK-G-GR-	L-KN-LE---KLARRA-M	E-QQ--A-----V-
	<i>Prochlorococcus marinus</i> MIT9301	YP_001019153	---K-F--K-E-K-DK-G-GR-	I-KD-YE---QSARRA-M	E-QS--A-----
	<i>Prochlorococcus marinus</i> AS9601	YP_001009706	---K-F--K-E-K-DK-G-GR-	I-KD-YE---QAARRA-M	E-QS--A-----
	<i>Prochlorococcus marinus</i> MIT9312	YP_397731	---K-F--K-E-K-DK-G-GR-	I-KD-YE---QSARRA-M	E-QS--A-----
<i>Prochlorococcus marinus</i> CCMP1986	NP_893257	-----F--K-E-K-DK-G-GR-	I-KD-YE---QTARRA-M	E-QS--A-----	
<i>Prochlorococcus marinus</i> MIT9313	NP_894987	-----L-----P-H-DR-G-GR-	L-KD-M---DVARRG-M	E-QQ--A-----V-	
Cyano- clade A	<i>Synechococcus</i> JA-3-3A	YP_474011	-----L-----PNLSQLSGHR	KQAE--A-V-----AS--V-	
	<i>Synechococcus</i> JA-2-3B	YP_476390	-----L-----PNLSRLTGHR	KQAE--A-V-----AS--V-	
	<i>Gloeobacter violaceus</i>	NP_923582	-----L-----RGLGQLNQD	RE-A--A-F-----TA-----	
	<i>Escherichia coli</i>	16131704	-----LD---L-LPDIKS-NGAR	R-AAE-A-I--M-TA-----R	
	<i>Xylella fastidiosa</i>	15837705	-----VF---L-LNSIASGNQTC	R-AAE-A-I--M-TA-----R	
	<i>Nitrosomonas europaea</i>	22955357	-----V-----LQLSDIRSNQ-NR	QM-AAE-A-I--M-TA-----L	
	<i>Rhodospirillum rubrum</i>	22965813	---W-A-PF---IPIPGIQDKNEAA	R-FAE-Q-I-----GA-----R	
	<i>Chlorobium tepidum</i>	21674485	---T-LM-----VFDL-SANBNI	RKAAE-VTM-T-----TA-----P	
	<i>Borrelia burgdorferi</i>	15594893	--S---K---IKET-SNNVLE	RSAAE-T-I-SI-----A---M---	
	<i>Deinococcus radiodurans</i>	15006710	-----LY-----VFGLSRRN-VQ	REAAE-L-Y-M-----TA---M-L	
	<i>Thermotoga maritima</i>	15644367	---R-LF--K-DIFQLMARD-NT	Q-EGE-I-I-T-----TA-----L	
	<i>Clostridium perfringens</i>	18310976	---L-M-N-----IPEA-A-NKIV	K-LGE-L-M-----TA-----L	
	Other bacteria	<i>Streptococcus coelicolor</i>	21220485	---TA-LF-----LPDL-SDN-QR	REAAE-M-L-----TA---V---
<i>Staphylococcus aureus</i>		21283362	-----L-H-----IPDTISRNFN	RGFAE-T-M-T-----A-----L	
<i>Bacillus subtilis</i>		16079961	---T-LMH---IPELTISRNFNI	RSFAE-T-M-T-----A-----K-	

Fig. 4. Partial sequence alignment of DNA polymerase I (Pol I) showing a large conserved insert (boxed) that is uniquely present in various cyanobacteria except those from the clade A species/strains. Dashes in all alignments indicate identity with the amino acid on the top line. Sequence information for only a few bacteria from other phyla is presented; however, this insert is not found in any other bacteria. Due to space considerations, sequence information for a number of *Prochlorococcus* and *Synechococcus* species/strains (*Synechococcus* CC9311, WH7805, BL107, CC9605 and CC9902; *Prochlorococcus marinus* MIT9303, MIT9215, MIT9515 and NATL2A) that are closely related to the strains shown here is not presented in this and several other sequence alignments. However, the sequences from these strains behaved in the same manner as those shown here, unless indicated otherwise. At the position marked *, three extra amino acids are present in *Acaryochloris marina* and *Thermosynechococcus elongatus*.

previously, also contains a large insert that is specific for various cyanobacteria except those from clade A (Fig. 4). In contrast to the other clade C-specific signatures, the insert in Pol I is lacking in *Synechococcus* WH5701, which lies in the middle of the clade C species/strains. To account for the absence of the Pol I insert in *Synechococcus* WH5701, one must postulate that either the Pol I gene in this cyanobacterium was acquired laterally from some other bacterium lacking the insert, or that the insert in this case has somehow been deleted/lost. Based upon the available information, we are unable to distinguish between these possibilities. Of the inserts that are specific for the clade C cyanobacteria, the homologues of RpoB are also present in various plastids and green plants, and this insert is not found in any of them (Fig. 6). Likewise, the plant homologues of the protein tRNA (guanine-N1-)-methyltransferase also lack the insert in this protein (Supplementary Fig. S13). These observations suggest that

the plastid/plant homologues of these proteins have not originated from clade C cyanobacteria.

Within clade C, among the *Prochlorococcus* isolates, two phylogenetically and physiologically different subgroups or ecotypes (high-B/A and low-B/A) have been identified (Moore *et al.*, 1998; Rocap *et al.*, 2002). The strains from these two ecotypes differ in terms of the relative ratios of chlorophyll *b* and *a2* in their light-harvesting systems. Strains from these two subgroups also differ in their ability to grow at different light intensities and copper toxicities and to use nitrite or nitrate as nitrogen sources (Ferris & Palenik, 1998; Rocap *et al.*, 2002). In the phylogenetic trees shown in Figs 1 and 2, the low-B/A *Prochlorococcus* isolates formed a distinct subclade within clade C that was separated from all other clade C species/strains by a long branch length and a 100 % bootstrap score. We have identified two conserved inserts that are specific for the

		55	106
Cyano-clade A	<i>Gloeobacter violaceus</i>	NP_926873	QERERGITITAAAITTAWT
	<i>Synechococcus</i> JA-3-3A	YP_474745	-----R--
	<i>Synechococcus</i> JA-2-3B	YP_477162	-----R--
	<i>Prochlorococcus marinus</i> MIT9303	YP_001018361	-----S-T-N
	<i>Prochlorococcus marinus</i> AS9601	YP_001010102	-----S-S-Q
	<i>Prochlorococcus marinus</i> NATL1A	YP_001015770	-----S-T-D
	<i>Prochlorococcus marinus</i> MIT9515	YP_001012003	-----S-S-Q
	<i>Prochlorococcus marinus</i> CCMP1375	NP_876056	-----S-S-Q
	<i>Prochlorococcus marinus</i> MIT9211	YP_001551517	-----S-S-Q
	<i>Prochlorococcus marinus</i> MIT9301	YP_001091924	-----S-S-Q
	<i>Synechococcus</i> WH7805	ZP_01123454	-----S-S-N
	<i>Synechococcus</i> WH5701	ZP_01084699	-----S-S-K
	<i>Synechococcus</i> CC9311	YP_729614	-----S-S-N
	<i>Synechococcus</i> RCC307	YP_001228425	-----S--N
	<i>Synechococcus</i> WH8102	NP_898228	-----S-S-Q
	<i>Synechococcus</i> RS9917	ZP_01080868	-----S-S-K
	<i>Synechococcus</i> CC9902	YP_378022	-----S-S-K
Other cyano	<i>Thermosynechococcus elongatus</i>	NP_682539	-----S-S-R
	<i>Trichodesmium erythraeum</i>	YP_720404	-----S-K-L
	<i>Lyngbya</i> PCC8106	ZP_01619783	-----S-S-R
	<i>Crocospaera watsonii</i>	ZP_00515233	--Q-----S-N-L
	<i>Microcystis aeruginosa</i>	CA087589	-----S-S-Q
	<i>Cyanothece</i> CCY0110	ZP_01730031	--Q-----S-S-L
	<i>Nostoc punctiforme</i>	ZP_00107087	-----S-S-K
	<i>Synechocystis</i> PCC6803	NP_441642	-----S-D-L
	<i>Arthrospira platensis</i>	P13550	-----S-S-L
	<i>Anabaena variabilis</i>	YP_321807	-----S-S-K
	<i>Nostoc</i> sp. PCC7120	NP_488378	-----S-S-K
	<i>Acaryochloris marina</i>	YP_001516130	-----S-T-Q
	<i>Synechococcus elongatus</i> PCC6301	YP_171365	-----S-S-K
	<i>Helicobacter pylori</i>	NP_207986	--K-----S-T-CF-K
	<i>Clostridium difficile</i>	ZP_03126038	--K-----S-T-AS-K
	<i>Petrogla mobilis</i>	YP_001567840	--K-----S-TSAF-K
	<i>Micrococcus luteus</i>	ZP_02944573	--K-----S-V-CF-N
Other bacteria	<i>Corynebacterium jeikeium</i>	YP_251635	--K-----S-T-CF-K
	<i>Ehrlichia ruminantium</i>	YP_180032	--K-----S-T-CF-N
			D-R-----

Fig. 5. Partial sequence alignment of the protein synthesis elongation factor-G (EF-G) showing a large insert (boxed) that is unique to the clade A cyanobacteria, but not found in any other cyanobacteria or bacteria from other phyla.

low-B/A subgroup of *Prochlorococcus marinus* isolates. In the protein leucyl-tRNA synthetase, which is essential for protein synthesis, a 5 aa deletion is present that is specific for the various low-B/A subgroup *Prochlorococcus* strains (MIT9515, CCMP1986, MIT9312, MIT9215, MIT9301 and AS9601), but not found in any other cyanobacteria (Supplementary Fig. S16). Another 1 aa insert in a conserved region of the signal-recognition particle protein Ffh or SRP54 is also specific for the low-B/A subgroup of *Prochlorococcus* strains (Supplementary Fig. S17). Additionally, in the TrpRS signature shown in Supplementary Fig. S5, all of the low-B/A subgroup *Prochlorococcus marinus* strains contained a 3 aa insert, whilst a 4 aa insert was found in other cyanobacteria. These signature indels provide evidence that the low-B/A subgroup of *Prochlorococcus marinus* strains is phylogenetically and molecularly distinct from all other clade C species/strains. The evolutionary stages where genetic changes responsible for these signature sequences probably occurred are indicated in Fig. 7.

Conserved indels that are specific for the clade B cyanobacteria

A number of conserved indels have been identified that are specific for the clade B cyanobacteria, which encompass the

majority of known cyanobacteria (Honda *et al.*, 1999; Turner *et al.*, 1999; Castenholz, 2001; Wilmotte & Herdman, 2001; Sánchez-Baracaldo *et al.*, 2005; Swingley *et al.*, 2008a). Within clade B, heterocyst-forming cyanobacteria form a monophyletic group (subclass Nostocophycidae) (Turner *et al.*, 1999; Adams, 2000; Wilmotte & Herdman, 2001; Hoffmann, 2005; Rajaniemi *et al.*, 2005). These heterocystous cyanobacteria include both *Nostocales* (non-branching and dividing always in a plane at a right angle to the long axis of the trichome) and *Stigonematales* (showing true branching) (Rippka *et al.*, 1979; Wilmotte & Herdman, 2001; Rajaniemi *et al.*, 2005). We have identified two conserved indels that are specific for these cyanobacteria. In the PetA protein, which is a precursor of the apocytochrome *f*, a 4 aa insert in a conserved region is present in various *Nostocales* species/strains (*Nostoc*, *Anabaena* and *Nodularia*) as well as in *Mastigocladus laminosus*, which belongs to the order *Stigonematales* (Fig. 8). The unique shared presence of this insert in these cyanobacteria provides further evidence that they form a monophyletic group. Another conserved insert (5 aa) specific for the *Nostocales* is found in ribosomal protein S3 (Supplementary Fig. S18). The sequence information for this protein is not available for *Stigonematales*. In ribosomal protein S3, in the same position where the 5 aa insert is found in the *Nostocales*, an

		185	239
Cyanobacteria	<i>Synechococcus</i> RS9917	ZP_01081213	VLMRAMGL SDNDVVDKLRHPEYYRKSI EAA NDEGISSSDQALLELYKKLRPGEPF
	<i>Synechococcus</i> WH7803	YP_001225785	-----I-----K-----
	<i>Synechococcus</i> RS9916	ZP_01470995	-----I-----K-----
	<i>Synechococcus</i> CC9902	YP_376617	-----L-----F-K-----
	<i>Synechococcus</i> WH8102	NP_896706	-----L-----F-K-----
	<i>Synechococcus</i> WH5701	ZP_01084135	-----I-----L-----K-----
	<i>Synechococcus</i> RCC307	YP_001228215	-M-----I-----L-----K-----
	<i>Prochlorococcus marinus</i> MIT9313	NP_895334	-----I-----K-----
	<i>Prochlorococcus marinus</i> CCMP1375	NP_876031	-----I-----K-----
	<i>Prochlorococcus marinus</i> MIT9211	YP_001551492	-----I-----K-----
	<i>Prochlorococcus marinus</i> NATL1A	YP_001015706	-----I-----F-K-----
	<i>Prochlorococcus marinus</i> MIT9301	YP_001091900	-----F-QS-----
	<i>Prochlorococcus marinus</i> MIT9312	YP_398074	-----F-QN-----
	<i>Prochlorococcus marinus</i> CCMP1986	NP_893602	-----F-KQ-----
	<i>Prochlorococcus marinus</i> AS9601	YP_001010079	-----F-QN-----
	<i>Synechococcus elongatus</i> PCC6301	YP_173217	--LK-L-- --EIY-----Q-T-- DK-QF--E-M--R-----
	<i>Thermosynechococcus elongatus</i>	NP_681431	--LK-L-- T-SEILER-----Q-TV EK-KF--E-I-----
	<i>Nostoc punctiforme</i>	ZP_00111111	--LK-L-- --EIF-A-----FQ-T-- EK-QF--E-M--R-----
	<i>Nodularia spumigena</i>	ZP_01630667	--LK-L-- --EIF-A-----FQ-T-- EK-QF--E-M--R-----
	<i>Anabaena variabilis</i>	YP_324701	--LK-L-- --EIF-A-----FQ-T-- EK-QF--E-M--R-----
Other cyanobacteria	<i>Nostoc</i> sp. PCC7120	NP_485634	--LK-L-- --EIF-A-----FQ-T-- EK-QF--E-M--R-----
	<i>Lyngbya</i> PCC8106	ZP_01620022	--LK-L-- -N-EIF-A-----D-FQ-T-- DR-EF--E-M--R-----
	<i>Acaryochloris marina</i>	YP_001517898	--LK-L-- --GEIF-R-----Q-T-- DK-QFG--E-M--R-----
	<i>Microcystis aeruginosa</i>	CAC087527	--LK-L-- --EIF-S-----F-Q-TL DK-NP--E-M--D--R-----
	<i>Crocospira watsonii</i>	ZP_00516049	--LK-L-- --EIL-S-----F-QRTL DK-NPT--E-M--R-----
	<i>Synechocystis</i> PCC6803	NP_440685	--LK-L-- --EIL-S-S-----F-Q-TL DK-NPT--E-M--V-----
	<i>Trichodesmium erythraeum</i>	YP_722576	--LK-L-- G-SEIF-S-----D-FQ-T-- EK-QYG--E-M--R-----
	<i>Gloeobacter violaceus</i>	NP_925229	--LK-L-- --EIL-AF-----FQ-T-- EK-NY--E-M--R-----
	<i>Synechococcus</i> JA-3-3A	YP_473891	--LK-L-- --SEIY-S-----FFQ-TM EK-HY--E-M--R-----
	<i>Synechococcus</i> JA-2-3B	YP_478633	--LK-L-- --SEIY-N-----FFQ-TM EK-PYT--E-M--R-----
Plastids	<i>Rhodomonas salina</i>	YP_001293553	-FLK-----S-IYNG-----LK-F RV-NYTTEE--IQM-T-----
	<i>Guillardia theta</i>	NP_050742	-FLK-----T-IYNG-----LK-TF RF-NYTTEE--IQM-N-----
	<i>Cyanophora paradoxa</i>	NP_043230	IFLK-LN--ESELISA-T---FLQ-T- YEH-DCT-EE--I-FHRQ-----
	<i>Bacillus subtilis</i>	ABX64457	--L-L-FG--QEIL-LIGEN--L-NTL DKDNTENS-----I-ER-----
Other bacteria	<i>Staphylococcus aureus</i>	YP_039996	--L-L-FS--QEII-L-GDN--L-NTL EKD-TENTE-----I-ER-----
	<i>Geobacillus kaustophilus</i>	YP_145951	--L-L-FS--QEII-L-GDN--L-NTL EKDNTD-TEK--I-I-ER-----
	<i>Clostridium novyi</i>	YP_877190	I-I--L-L-G--AEI--WFGEE-RLKAT- EKDNTKTREE-----I-R-----
	<i>Listeria monocytogenes</i>	ZP_00229189	--L-L-FG--QEII-LIGDND--L-NTL EKDNTDNAEK-----I-ER-----
	<i>Mycobacterium tuberculosis</i>	ZP_00879929	--LK-L-W TSEQL-ERFSGS-IM-STL EKDNTVGT-E-----DI-R-----
	<i>Streptomyces coelicolor</i>	NP_628815	--LK-L-W TTEQLLEEFGEY-SM-ATL EKDHTQGQ-D-----DI-R-----

Fig. 6. Partial sequence alignment of the RNA polymerase β subunit (RpoB) showing a 3 aa insert (boxed) that is unique for the clade C cyanobacteria, but not found in any other cyanobacteria, plastid homologues or bacteria from other phyla.

8 aa insert is present in many green plants, but not in any of the plastid homologues (Supplementary Fig. S19). The different lengths and sequences of these inserts indicate that they are of independent origin.

Cyanobacteria such as *Synechocystis*, *Microcystis*, *Crocospira* and *Cyanothece*, etc., which belong to the order *Chroococcales*, form another well-defined clade in phylogenetic trees (Figs 1 and 2) (Honda *et al.*, 1999; Turner *et al.*, 1999; Sánchez-Baracaldo *et al.*, 2005; Shi & Falkowski, 2008; Swingley *et al.*, 2008a). This clade has been referred to as the SPM clade in earlier work (Turner *et al.*, 1999; Sánchez-Baracaldo *et al.*, 2005) and it forms a part of the subclass Oscillatoriophycidae in the proposal by Hoffmann *et al.* (2005). We have identified a 1 aa insert in a highly conserved region of the RecA protein that is commonly shared by all of these cyanobacteria (Fig. 9). This insert is also present in *Synechococcus* PCC7002, which branches with this clade in the phylogenetic trees (Fig. 2) (Turner *et al.*, 1999; Sánchez-Baracaldo *et al.*, 2005). In phylogenetic trees based on different datasets, *Trichodesmium erythraeum* generally branches in between the *Nostocales* and the *Chroococcales*, often as an outgroup of the *Nostocales* clade (Figs 1 and 2) (Honda *et al.*, 1999; Turner *et al.*, 1999; Sánchez-Baracaldo *et al.*, 2005; Shi & Falkowski, 2008; Swingley *et al.*, 2008a). We have identified

a 4 aa deletion in translation-initiation factor IF-2 that, within cyanobacteria, is uniquely present in various *Nostocales*, as well as in *Trichodesmium erythraeum* and *Lyngbya* PCC8106 (Supplementary Fig. S19), both of which belong to the order *Oscillatoriales*. The shared presence of this signature in the *Nostocales* and *Oscillatoriales* provides further evidence that these two orders of cyanobacteria are close relatives, as also seen in phylogenetic trees (Figs 1 and 2). Interestingly, another conserved signature consisting of a 7 aa insert in the α subunit of riboflavin synthase is uniquely shared by various *Chroococcales* and *Oscillatoriales* species/strains (Supplementary Fig. S20). A close relationship between these two cyanobacterial orders, which are grouped together in the subclass Oscillatoriophycidae (Hoffmann *et al.*, 2005), has also been reported in earlier work (Turner *et al.*, 1999). These molecular signatures support an intermediate placement of the *Oscillatoriales* in between the orders *Nostocales* and *Chroococcales*, as indicated in Fig. 7.

Three other large signature indels provide evidence that cyanobacteria from the orders *Nostocales*, *Oscillatoriales* and *Chroococcales* shared a common ancestor, exclusive of other cyanobacteria for which sequence information is presently available (Fig. 7). In the DnaE protein, a 19 aa insert is present specifically in all available sequences from

Fig. 7. Interpretive diagram showing various signatures that are specific for the clade B and C cyanobacteria and the evolutionary stages where they were probably introduced. The signatures for the nodes marked ● are described in Fig. 3.

these orders, but not in any other cyanobacteria or bacteria (Fig. 10). Likewise, in the protein GDP-mannose pyrophosphorylase, a 13 aa deletion is specifically present in these cyanobacteria, to the exclusion of all others (Supplementary Fig. S21). Another prominent indel of about 22–27 aa in the protein NAD(P)H-quinone oxidoreductase subunit D (Supplementary Fig. S22) is also specific for this clade of cyanobacteria. Most cyanobacteria contain at least two homologues of this protein (Melo *et al.*, 2004) and this insert is present in only one of the two homologues from these groups of cyanobacteria, but not in other cyanobacteria. The evolutionary stages where genetic changes responsible for these conserved indels have probably occurred are indicated in Fig. 7.

DISCUSSION

In this work, I have used a combination of phylogenomic and signature sequence-based (i.e. phenetic) approaches to

understand the evolutionary relationships among cyanobacteria. Phylogenetic trees were constructed based on concatenated sequences for 45 widely distributed proteins from the genomes of 34 sequenced cyanobacteria, as well as 16S rRNA gene sequences. In parallel, I have also identified 28 new conserved indels that are distinctive characteristics of various cyanobacterial clades that can be resolved in these phylogenetic trees. The identified conserved indels are found uniquely in all species/strains from the indicated groups of cyanobacteria, but they are not found in other cyanobacteria/bacteria. These results suggest strongly that the rare genetic changes responsible for these conserved indels occurred only once in a common ancestor of these clades and that their species-distribution patterns have not been affected by non-specific mechanisms such as lateral gene transfers (Boucher *et al.*, 2003; Zhaxybayeva *et al.*, 2006). These conserved indels thus provide potentially useful molecular synapomorphies for the identification and circumscription of a number of monophyletic clades within the cyanobacteria. The cyanobacterial clades that

Fig. 8. Partial sequence alignment of the PetA protein showing a 4 aa insert that is specific for the sequenced taxa from the orders *Nostocales/Stigonematales*. This insert is not found in any other cyanobacteria or the plastid homologues.

can now be distinguished and circumscribed on the basis of these synapomorphies include the *Nostocales* (plus *Stigonematales*), the *Chroococcales*, a deep-branching clade (clade A) comprising *Gloeobacter* and the two diazotrophic *Synechococcus* strains (subclass *Gloeobacterophycidae*), a clade comprising all other cyanobacteria except those from clade A, a clade consisting of various marine unicellular *Synechococcus* and *Prochlorococcus* strains/isolates (clade C) and a subclade of the low-B/A ecotype *Prochlorococcus* strains (Ferris & Palenik, 1998; Rocap *et al.*, 2002).

Hoffmann *et al.* (2005) proposed a new classification scheme for cyanobacteria that divides them into four subclasses. Of the clades that we can identify based upon the identified signature sequences, clade A corresponds to the subclass *Gloeobacterophycidae* (Hoffmann *et al.*, 2005). The two *Synechococcus* strains (JA-3-3A and JA-2-3B) that group reliably with this clade both in phylogenetic trees and based upon different signature sequences have presumably been misidentified as *Synechococcus* and they are clearly indicated to be part of this clade. The clade comprising the *Nostocales* (plus *Stigonematales*) corresponds to the subclass *Nostocophycidae*, whereas clade C corresponds to the subclass *Synechococcophycidae*, with the exception that *Acaryochloris* is indicated to not be a part of this clade. The fourth proposed subclass of cyanobacteria, *Oscillatoriophycidae*, comprises the orders *Oscillatoriales* and *Chroococcales*. We have identified a

conserved insert in the protein riboflavin synthase that is uniquely shared by the available sequences from these orders. Although this signature supports a grouping of these two orders in the subclass *Oscillatoriophycidae* (Hoffmann *et al.*, 2005), the results from various phylogenetic studies (Figs 1 and 2) (Sánchez-Baracaldo *et al.*, 2005; Swingley *et al.*, 2008a), as well as a signature in the IF-2 protein (Supplementary Fig. S19), indicate that the order *Oscillatoriales*, represented in this work by *Trichodesmium erythraeum* and *Lyngbya* PCC8106, is related more closely to the *Nostocales* than to the *Chroococcales*. However, it should be noted that the filamentous non-heterocystous cyanobacteria such as those belonging to the order *Oscillatoriales* are a heterogeneous and possibly polyphyletic assemblage (Hoffmann *et al.*, 2005). Although our analysis based on the genomes for *Trichodesmium erythraeum* and *Lyngbya* PCC8106 suggests an intermediate placement of the order *Oscillatoriales* between the orders *Nostocales* and *Chroococcales* (Fig. 7), it is possible that this placement could be a consequence of limited representation of taxa (only two genomes) from this order in our work. Hence, it is important to confirm this inference by using sequence data for various other members of the orders *Oscillatoriales*, *Nostocales* and *Chroococcales*.

In addition to these signatures, a number of other identified synapomorphies clarify certain evolutionary

		220	258
Chroococcales	<i>Synechocystis</i> PCC6803	NP_443045	YASVRLDIRRIQT LKKGTE G EYGIRAKVKVAKNKVAPPF
	<i>Microcystis aeruginosa</i>	CA090208	-----
	<i>Cyanothece</i> sp. CCY0110	ZP_01728899	-----S-----
	<i>Cyanothece</i> sp. ATCC 51142	YP_001806053	-----S-----
	<i>Cyanothece</i> sp. PCC8801	EBDT59347	-----S-----
	<i>Crocospheara watsonii</i>	ZP_00519282	-----S-----
	<i>Synechococcus</i> PCC7002	YP_001733692	-----S-----
	<i>Nostoc punctiforme</i>	ZP_00105721	-----D-----N-V-----
	<i>Nostoc</i> sp. PCC7120	NP_487312	-----D-----F-N-V-----
	<i>Anabaena variabilis</i>	NP_325417	-----D-----F-N-V-----
	<i>Nodularia spumigena</i>	ZP_01631726	-----F-N-V-----
	<i>Acaryochloris marina</i>	YP_001519757	-----M-----
	<i>Trichodesmium erythraeum</i>	YP_722357	-----T-----SS-----H-----
	<i>Lyngbya</i> sp. PCC8106	ZP_01624747	-C-----D-----
	<i>Synechococcus elongatus</i> PCC6301	YP_171875	-----T-----V-----
Other cyanobacteria	<i>Synechococcus elongatus</i> PCC7942	YP_399367	-----T-----V-----
	<i>Thermosynechococcus elongatus</i>	NP_682879	-----V-----F-----
	<i>Gloeobacter violaceus</i>	NP_926362	-----KAE-----QD-----N-VR-----V-----
	<i>Synechococcus</i> JA-3-3A	YP_478173	-----V-----A-----R-----V-----
	<i>Synechococcus</i> JA-2-3B	YP_474545	-----V-----A-----R-----V-----
	<i>Prochlorococcus marinus</i> MIT9301	YP_001091977	-----R-----
	<i>Prochlorococcus marinus</i> NATL1A	YP_001015827	-----R-----
	<i>Prochlorococcus marinus</i> AS9601	YP_001010159	-----R-----
	<i>Prochlorococcus marinus</i> MIT9312	YP_398151	-----R-----
	<i>Prochlorococcus marinus</i> CCMP1986	NP_893679	-----R-----
	<i>Prochlorococcus marinus</i> CCMP1375	NP_876107	-----R-----
	<i>Prochlorococcus marinus</i> MIT9313	NP_895554	-----R-----
	<i>Prochlorococcus marinus</i> MIT9211	YP_001551566	-----R-D-----
	<i>Synechococcus</i> WH5701	ZP_01083760	-----R-----
	<i>Synechococcus</i> WH7803	YP_001224162	-----R-----
	<i>Synechococcus</i> RS9916	ZP_01473032	-----R-----
	<i>Synechococcus</i> WH5701	ZP_01086496	-----F-----
	<i>Synechococcus</i> RS9917	ZP_01080804	-----R-----
	<i>Synechococcus</i> WH8102	NP_898153	-----F-----
	<i>Synechococcus</i> RCC307	YP_001228367	-----RA-----

Fig. 9. Partial sequence alignment of the RecA protein showing a 1 aa insert in a highly conserved region that is specific for the sequenced species/strains from the order *Chroococcales*. This insert is also present in *Synechococcus* PCC7002, which branches with this group in phylogenetic trees (see Fig. 2) (Turner *et al.*, 1999; Sánchez-Baracaldo *et al.*, 2005).

relationships that are not resolved in phylogenetic trees. Species belonging to the orders *Nostocales*, *Oscillatoriales* and *Chroococcales* are indicated as late-diverging lineages within clade B of cyanobacteria (Honda *et al.*, 1999; Turner *et al.*, 1999; Sánchez-Baracaldo *et al.*, 2005; Swingley *et al.*, 2008a). However, a clade comprising these cyanobacteria is only weakly supported in some phylogenetic trees (Sánchez-Baracaldo *et al.*, 2005; Swingley *et al.*, 2008a). In the present work, we have identified three prominent signatures that are uniquely shared by various sequenced taxa from the orders *Nostocales*, *Oscillatoriales* and *Chroococcales*. These results provide strong suggestive evidence that species from these orders shared a common ancestor exclusive of all other cyanobacteria and that these signatures are revealing a higher taxonomic clade within the cyanobacteria.

The evolutionary relationship among cyanobacteria was studied in this work based upon species/strains whose genomes have been sequenced (Table 1). About 60 % of these genomes are from marine *Prochlorococcus* and *Synechococcus* strains/isolates. Thus, the present dataset does not truly reflect the relative abundance or genetic diversity of this phylum. Nonetheless, the available genomes cover many important groups/orders within

cyanobacteria and they permit comparison of the results obtained by using the 16S rRNA gene with those obtained by other approaches. Some of the clades, such as those corresponding to the orders *Nostocales*, *Chroococcales* or clade C, are clearly distinguished in all of the trees (Figs 1 and 2) as well as by means of conserved indels in protein sequences. However, the deeper branching of the clade A cyanobacteria, whilst resolved with strong statistical support in the phylogenomic tree based on protein sequences (Fig. 1) and also based on conserved indels (Fig. 3), was not resolved in the 16S rRNA gene-based tree (Fig. 2). Importantly, on the basis of conserved indels, we were able to clearly distinguish not only all of the clades that can be resolved in both the 16S rRNA gene and the protein trees, but also a deeper-branching clade comprising the orders *Nostocales*, *Chroococcales* and *Oscillatoriales* (Fig. 7) that was not resolved by phylogenetic methods. These observations indicate that the results obtained by using the phylogenomic and signature-sequence approaches are, in general, in very good agreement with those suggested by 16S rRNA gene analysis, but these other methods are better suited for resolving deeper-branching relationships that are not resolved in 16S rRNA gene-based trees (Wilmotte & Herdman, 2001; Hoffmann *et al.*, 2005).

Fig. 10. Excerpt from a sequence alignment for the DnaE protein showing a 19 aa insert (boxed) that is specific for the sequenced species/strains from the orders *Nostocales*, *Oscillatoriales* and *Chroococcales*, but not found in any other cyanobacteria. This insert was probably introduced into a common ancestor of these orders, as indicated in Fig. 7.

16S rRNA gene-based trees currently provide the primary means for understanding microbial taxonomy and phylogeny. Currently, 16S rRNA gene sequences from >4500 cyanobacteria are available in the Ribosomal Database Project (RDP release 10; Maidak *et al.*, 2001). However, with the large increase in the numbers of these sequences, the ability of 16S rRNA gene-based trees to resolve the relative branching positions of different taxa (particularly the deeper branches) has diminished greatly and, in some cases, been almost entirely lost (Ludwig & Schleifer, 1999; Wilmotte & Herdman, 2001; Hoffmann *et al.*, 2005). This loss of phylogenetic clarity is in large part a consequence of many variables, such as differences in base compositions or evolutionary rates among various lineages, long branch-length effect, etc., that affect the branching of species in phylogenetic trees markedly, but are very difficult to control or correct for (Felsenstein, 2004; Delsuc *et al.*, 2005). Hence, other stable characteristics that are consistent with the 16S rRNA gene-based as well as other phylogenetic approaches, but are affected minimally by these variables, are of particular importance for taxonomic and evolutionary studies (Rivera & Lake, 1992; Delwiche *et al.*, 1995; Gupta, 1998, 2000; Rokas & Holland, 2000; Skophammer *et al.*, 2007; Griffiths & Gupta, 2007).

Conserved indels, because the genetic changes that give rise to them are of a highly specific nature, are less likely to occur independently in different taxa. Further, because these indels are present in highly conserved regions of the proteins, their presence or absence is not affected by differences in base composition or evolutionary rates among various lineages. Hence, such indels have provided generally reliable synapomorphic characteristics for identifying different monophyletic clades that are due to commonly shared ancestry (Rivera & Lake, 1992; Delwiche *et al.*, 1995; Gupta, 1998, 2000; Rokas & Holland, 2000; Gupta *et al.*, 2003; Skophammer *et al.*, 2007; Griffiths & Gupta, 2007).

In this work, we have also evaluated the reliability of many cyanobacterial signatures that were described in earlier work (Gupta *et al.*, 2003). Of the 14 cyanobacteria-specific indels described previously when sequence data were available for only eight cyanobacteria, 13 are present in all (>34) sequenced cyanobacteria. One of these signatures (the insert in Pol I; Fig. 4) is absent only in clade A cyanobacteria for which no sequence information was available at that time. Similar results have been obtained for numerous other indels that are distinctive characteristics of other groups/phyla of bacteria (Gao & Gupta,

2005; Gupta & Griffiths, 2006; Gupta & Mok, 2007). These observations indicate strongly that the vast majority of these conserved indels retain their specificity as sequence information from new genomes or other sources becomes available and that many of these signatures will prove useful as reliable molecular markers for taxonomic and evolutionary studies. It should be emphasized that the placement of any species/strain into a given clade by using this approach is based upon several independent signatures that provide complementary information. Some of these signatures serve to exclude a given species/strain from particular groups or clades, whereas others point to its inclusion in more and more specific clades. The information provided by all of the signatures is generally internally highly consistent and only in exceptional cases is contradictory placement of a species in alternative position(s) indicated. For placement of any species or strain into a particular clade, in most cases, partial sequence information for only a few genes containing the diagnostic signature indels is sufficient. Because all of these signatures are flanked on both sides by conserved sequence regions, degenerate PCR primers for obtaining the requisite sequence information from other cyanobacteria could be designed readily. Hence, it should be of much interest to obtain sequence information for many of these signatures from a broad range of cyanobacteria, to confirm the usefulness and reliability of these molecular markers for taxonomic studies.

The present work also provides useful insights into the evolutionary relationship between cyanobacteria and plastids. The shared presence of several cyanobacteria-specific indels in the plastid/plant homologues (Gupta *et al.*, 2003) has previously supported the view that plastids have originated from a cyanobacterial ancestor (Gray & Doolittle, 1982; Morden *et al.*, 1992; Delwiche *et al.*, 1995; Palmer & Delwiche, 1998; Gupta *et al.*, 2003; Archibald, 2006; Stiller, 2007). In this work, the shared presence of the 5 aa insert in the tryptophan synthase beta subunit (Supplementary Fig. S4) in different green plants and algae and all cyanobacteria except those from clade A provides evidence that the plant/plastid homologues have not originated from clade A. The signature indels in several other proteins (RpoB, Fig. 6; RpoC, Supplementary Fig. S14) that are uniquely found in the clade C homologues, but are absent in other cyanobacteria as well as plant/plastid homologues, provide evidence against the origin of plastids from clade C cyanobacteria. By exclusion, these results suggest that the plant/plastid homologues have probably originated from clade B cyanobacteria. A recent comparative genomic study has suggested that the plastid homologues are most similar to the order *Nostocales* (part of clade B) (Deutsch *et al.*, 2008). However, in our work, conserved inserts in the PetA (Fig. 8) and ribosomal S3 (Supplementary Fig. S18) proteins, which are specific for the order *Nostocales*, are not found in the corresponding plastid homologues, thereby questioning the validity of this inference. Hopefully, in future studies, additional signature

sequences will be identified that will allow us to further pinpoint the cyanobacterial ancestor of the plastids.

ACKNOWLEDGEMENTS

This work was supported by a research grant from the Natural Science and Engineering Research Council of Canada. I thankfully acknowledge the technical assistance provided by Larissa Shamseer, Adeel Mahmood and Divya Wilson in the sequence-alignment studies and in the creation of some of the signature files.

REFERENCES

- Adams, D. G. (2000). Heterocyst formation in cyanobacteria. *Curr Opin Microbiol* **3**, 618–624.
- Anagnostidis, K. & Komarek, J. (1985). Modern approaches to the classification system of *Cyanophytes*. 1 – Introduction. *Arch Hydrobiol* **38–39**, 291–302.
- Archibald, J. M. (2006). Algal genomics: exploring the imprint of endosymbiosis. *Curr Biol* **16**, R1033–R1035.
- Baldauf, S. L. & Palmer, J. D. (1993). Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc Natl Acad Sci U S A* **90**, 11558–11562.
- Boucher, Y., Douady, C. J., Papke, R. T., Walsh, D. A., Boudreau, M. E., Nesbø, C. L., Case, R. J. & Doolittle, W. F. (2003). Lateral gene transfer and the origins of prokaryotic groups. *Annu Rev Genet* **37**, 283–328.
- Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. & Stanhope, M. J. (2001). Universal trees based on large combined protein sequence data sets. *Nat Genet* **28**, 281–285.
- Castenholz, R. W. (2001). Phylum BX. Cyanobacteria: oxygenic photosynthetic bacteria. In *Bergey's Manual of Systematic Bacteriology*, 2nd edn, vol. 1, pp. 474–487. Edited by D. R. Boone & R. W. Castenholz. New York: Springer.
- Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540–552.
- Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B. & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287.
- Delsuc, F., Brinkmann, H. & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* **6**, 361–375.
- Delwiche, C. F., Kuhsel, M. & Palmer, J. D. (1995). Phylogenetic analysis of *tufA* sequences indicates a cyanobacterial origin of all plastids. *Mol Phylogenet Evol* **4**, 110–128.
- Deutsch, O., Landan, G., Roettger, M., Gruenheit, N., Kowallik, K. V., Allen, J. F., Martin, W. & Dagan, T. (2008). Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Mol Biol Evol* **25**, 748–761.
- Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I. M., Barbe, V., Duprat, S., Galperin, M. Y., Koonin, E. V. & other authors (2003). Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci U S A* **100**, 10020–10025.
- Dufresne, A., Ostrowski, M., Scanlan, D. J., Garczarek, L., Mazard, S., Palenik, B. P., Paulsen, I. T., de Marsac, N. T., Wincker, P. & other authors (2008). Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol* **9**, R90.
- Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res* **8**, 163–167.

- Felsenstein, J. (2004). *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates.
- Ferris, M. J. & Palenik, B. (1998). Niche adaptation in ocean cyanobacteria. *Nature* **396**, 226–228.
- Gao, B. & Gupta, R. S. (2005). Conserved indels in protein sequences that are characteristic of the phylum *Actinobacteria*. *Int J Syst Evol Microbiol* **55**, 2401–2412.
- Garrity, G. M., Bell, J. A. & Lilburn, T. G. (2005). The revised road map to the *Manual*. In *Bergey's Manual of Systematic Bacteriology*, 2nd edn, vol. 2, part A, pp. 159–220. Edited by D. J. Brenner, N. R. Krieg & J. T. Staley. New York: Springer.
- Giovannoni, S. J., Turner, S., Olsen, G. J., Barns, S., Lane, D. J. & Pace, N. R. (1988). Evolutionary relationships among cyanobacteria and green chloroplasts. *J Bacteriol* **170**, 3584–3592.
- Gray, M. W. & Doolittle, W. F. (1982). Has the endosymbiont hypothesis been proven? *Microbiol Rev* **46**, 1–42.
- Griffiths, E. & Gupta, R. S. (2001). The use of signature sequences in different proteins to determine the relative branching order of bacterial divisions: evidence that *Fibrobacter* diverged at a similar time to *Chlamydia* and the *Cytophaga*–*Flavobacterium*–*Bacteroides* division. *Microbiology* **147**, 2611–2622.
- Griffiths, E. & Gupta, R. S. (2007). Phylogeny and shared conserved inserts in proteins provide evidence that *Verrucomicrobia* are the closest known free-living relatives of chlamydiae. *Microbiology* **153**, 2648–2654.
- Gugger, M. F. & Hoffmann, L. (2004). Polyphyly of true branching cyanobacteria (*Stigonematales*). *Int J Syst Evol Microbiol* **54**, 349–357.
- Gupta, R. S. (1998). Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaeobacteria, eubacteria and eukaryotes. *Microbiol Mol Biol Rev* **62**, 1435–1491.
- Gupta, R. S. (2000). The phylogeny of *Proteobacteria*: relationships to other eubacterial phyla and eukaryotes. *FEMS Microbiol Rev* **24**, 367–402.
- Gupta, R. S. (2001). The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins. *Int Microbiol* **4**, 187–202.
- Gupta, R. S. (2003). Evolutionary relationships among photosynthetic bacteria. *Photosynth Res* **76**, 173–183.
- Gupta, R. S. (2004). The phylogeny and signature sequences characteristics of *Fibrobacteres*, *Chlorobi*, and *Bacteroidetes*. *Crit Rev Microbiol* **30**, 123–143.
- Gupta, R. S. & Griffiths, E. (2006). Chlamydiae-specific proteins and indels: novel tools for studies. *Trends Microbiol* **14**, 527–535.
- Gupta, R. S. & Mok, A. (2007). Phylogenomics and signature proteins for the alpha *Proteobacteria* and its main groups. *BMC Microbiol* **7**, 106.
- Gupta, R. S., Mukhtar, T. & Singh, B. (1999). Evolutionary relationships among photosynthetic prokaryotes (*Heliobacterium chlorum*, *Chloroflexus aurantiacus*, cyanobacteria, *Chlorobium tepidum* and proteobacteria): implications regarding the origin of photosynthesis. *Mol Microbiol* **32**, 893–906.
- Gupta, R. S., 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.
- Hansmann, S. & Martin, W. (2000). Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. *Int J Syst Evol Microbiol* **50**, 1655–1663.
- Harris, J. K., Kelley, S. T., Spiegelman, G. B. & Pace, N. R. (2003). The genetic core of the universal ancestor. *Genome Res* **13**, 407–412.
- Henson, B. J., Watson, L. E. & Barnum, S. R. (2002). Molecular differentiation of the heterocystous cyanobacteria, *Nostoc* and *Anabaena*, based on complete *NifD* sequences. *Curr Microbiol* **45**, 161–164.
- Hoffmann, L. (2005). Nomenclature of *Cyanophyta/Cyanobacteria*: roundtable on the unification of the nomenclature under the Botanical and Bacteriological Codes. *Arch Hydrobiol* **159**, 13–29.
- Hoffmann, L., Komárek, J. & Kaštovský, J. (2005). System of *Cyanoprokaryotes* (*Cyanobacteria*) – state in 2004. *Arch Hydrobiol* **159**, 95–155.
- 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.
- Horn, M., Collingro, A., Schmitz-Esser, S., Beier, C. L., Purkhold, U., Fartmann, B., Brandt, P., Nyakatura, G. J., Droege, M. & other authors (2004). Illuminating the evolutionary history of chlamydiae. *Science* **304**, 728–730.
- Huang, J. & Gogarten, J. P. (2007). Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? *Genome Biol* **8**, R99.
- Ishida, T., Watanabe, M. M., Sugiyama, J. & Yokota, A. (2001). Evidence for polyphyletic origin of the members of the orders of *Oscillatoriales* and *Pleurocapsales* as determined by 16S rDNA analysis. *FEMS Microbiol Lett* **201**, 79–82.
- Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. & Gibson, T. J. (1998). Multiple sequence alignment with CLUSTAL_X. *Trends Biochem Sci* **23**, 403–405.
- Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M. & other authors (1996). Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* **3**, 109–136.
- Kaneko, T., Nakamura, Y., Wolk, C. P., Kuritz, T., Sasamoto, S., Watanabe, A., Iriguchi, M., Ishikawa, A., Kawashima, K. & other authors (2001). Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res* **8**, 205–213.
- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**, 111–120.
- Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.
- Kondratieva, E. N., Pfennig, N. & Truper, H. G. (1992). The phototrophic prokaryotes. In *The Prokaryotes*, pp. 312–330. Edited by A. Balows, H. G. Trüper, M. Dworkin, W. Harder & K. H. Schleifer. New York: Springer.
- Ludwig, W. & Schleifer, K. H. (1999). Phylogeny of *Bacteria* beyond the 16S rRNA standard. *ASM News* **65**, 752–757.
- Maidak, B. L., Cole, J. R., Lilburn, T. G., Parker, C. T., Jr, Saxman, P. R., Farris, R. J., Garrity, G. M., Olsen, G. J., Schmidt, T. M. & Tiedje, J. M. (2001). The RDP-II (Ribosomal Database Project). *Nucleic Acids Res* **29**, 173–174.
- Margulis, L. (1970). *Origin of Eukaryotic Cells*. New Haven, CT: Yale University Press.
- Melo, A. M., Bandeiras, T. M. & Teixeira, M. (2004). New insights into type II NAD(P)H:quinone oxidoreductases. *Microbiol Mol Biol Rev* **68**, 603–616.
- Moore, L. R., Rocap, G. & Chisholm, S. W. (1998). Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393**, 464–467.

- Morden, C. W., Delwiche, C. F., Kuhse, M. & Palmer, J. D. (1992). Gene phylogenies and the endosymbiotic origin of plastids. *Biosystems* 28, 75–90.
- Moustafa, A., Reyes-Prieto, A. & Bhattacharya, D. (2008). Chlamydiae has contributed at least 55 genes to Plantae with predominantly plastid functions. *PLoS One* 3, e2205.
- Nakamura, Y., Kaneko, T., Sato, S., Ikeuchi, M., Katoh, H., Sasamoto, S., Watanabe, A., Iriguchi, M., Kawashima, K. & other authors (2002). Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1. *DNA Res* 9, 123–130.
- Nakamura, Y., Kaneko, T., Sato, S., Mimuro, M., Miyashita, H., Tsuchiya, T., Sasamoto, S., Watanabe, A., Kawashima, K. & other authors (2003). Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. *DNA Res* 10, 137–145.
- Olsen, G. J. & Woese, C. R. (1993). Ribosomal RNA: a key to phylogeny. *FASEB J* 7, 113–123.
- 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. & Stackebrandt, E. (2002). Prokaryote taxonomy online: challenges ahead. *Nature* 419, 15.
- Oren, A. & Tindall, B. J. (2005). Nomenclature of the cyanophyta/cyanobacteria/cyanoprokaryotes under the International Code of Nomenclature of Prokaryotes. *Arch Hydrobiol* 117, 39–52.
- Palenik, B., Brahamsha, B., Larimer, F. W., Land, M., Hauser, L., Chain, P., Lamerdin, J., Regala, W., Allen, E. E. & other authors (2003). The genome of a motile marine *Synechococcus*. *Nature* 424, 1037–1042.
- Palenik, B., Ren, Q., Dupont, C. L., Myers, G. S., Heidelberg, J. F., Badger, J. H., Madupu, R., Nelson, W. C., Brinkac, L. M. & other authors (2006). Genome sequence of *Synechococcus* CC9311: insights into adaptation to a coastal environment. *Proc Natl Acad Sci U S A* 103, 13555–13559.
- Palmer, J. D. & Delwiche, C. F. (1998). The origin and evolution of plastids and their genomes. In *Molecular Systematics of Plants II: DNA Sequencing*, pp. 375–409. Edited by D. E. Sotis, P. E. Soltis & J. J. Doyle. Norwell, MA: Kluwer Academic.
- Rajaniemi, P., Hrouzek, P., Kastovska, K., Willame, R., Rantala, A., Hoffmann, L., Komárek, 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, J. B., Herdman, M. & Stanier, R. Y. (1979). Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *J Gen Microbiol* 111, 1–61.
- Rivera, M. C. & Lake, J. A. (1992). Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science* 257, 74–76.
- Robertson, B. R., Tezuka, N. & Watanabe, M. M. (2001). Phylogenetic analyses of *Synechococcus* strains (cyanobacteria) using sequences of 16S rDNA and part of the phycocyanin operon reveal multiple evolutionary lines and reflect phycobilin content. *Int J Syst Evol Microbiol* 51, 861–871.
- Rocap, G., Distel, D. L., Waterbury, J. B. & Chisholm, S. W. (2002). Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S–23S ribosomal DNA internal transcribed spacer sequences. *Appl Environ Microbiol* 68, 1180–1191.
- Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L. & other authors (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* 424, 1042–1047.
- Rokas, A. & Holland, P. W. (2000). Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol* 15, 454–459.
- Rokas, A., Williams, B. L., King, N. & Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425, 798–804.
- Sánchez-Baracaldo, P., Hayes, P. K. & Blank, C. E. (2005). Morphological and habitat evolution in the cyanobacteria using a compartmentalization approach. *Geobiology* 3, 145–165.
- Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18, 502–504.
- Seo, P. S. & Yokota, A. (2003). The phylogenetic relationships of cyanobacteria inferred from 16S rRNA, *gyrB*, *rpoC1* and *rpoD1* gene sequences. *J Gen Appl Microbiol* 49, 191–203.
- Shi, T. & Falkowski, P. G. (2008). Genome evolution in cyanobacteria: the stable core and the variable shell. *Proc Natl Acad Sci U S A* 105, 2510–2515.
- Skophammer, R. G., Servin, J. A., Herbold, C. W. & Lake, J. A. (2007). Evidence for a Gram-positive, eubacterial root of the tree of life. *Mol Biol Evol* 24, 1761–1768.
- Sneath, P. H. A. (2001). Numerical taxonomy. In *Bergey's Manual of Systematic Bacteriology*, 2nd edn, vol. 1, pp. 39–42. Edited by D. R. Boone & R. W. Castenholz. New York: Springer.
- Sneath, P. H. A. & Sokal, R. R. (1973). *Numerical Taxonomy: the Principles and Practice of Numerical Classification*. San Francisco: W. H. Freeman.
- Stiller, J. W. (2007). Plastid endosymbiosis, genome evolution and the origin of green plants. *Trends Plant Sci* 12, 391–396.
- Sugita, C., Ogata, K., Shikata, M., Jikuya, H., Takano, J., Furumichi, M., Kanehisa, M., Omata, T., Sugiura, M. & Sugita, M. (2007). Complete nucleotide sequence of the freshwater unicellular cyanobacterium *Synechococcus elongatus* PCC 6301 chromosome: gene content and organization. *Photosynth Res* 93, 55–67.
- Swingle, W. D., Blankenship, R. E. & Raymond, J. (2008a). Integrating Markov clustering and molecular phylogenetics to reconstruct the cyanobacterial species tree from conserved protein families. *Mol Biol Evol* 25, 643–654.
- Swingle, W. D., Chen, M., Cheung, P. C., Conrad, A. L., Dejesa, L. C., Hao, J., Honchak, B. M., Karbach, L. E., Kurdoglu, A. & other authors (2008b). Niche adaptation and genome expansion in the chlorophyll *d*-producing cyanobacterium *Acaryochloris marina*. *Proc Natl Acad Sci U S A* 105, 2005–2010.
- Turner, S., Pryer, K. M., Miao, V. P. & Palmer, J. D. (1999). Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. *J Eukaryot Microbiol* 46, 327–338.
- Van de Peer, Y. & De Wachter, R. (1994). TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput Appl Biosci* 10, 569–570.
- Whatley, J. M. (1993). The endosymbiotic origin of chloroplasts. *Int Rev Cytol* 144, 259–299.
- Willemotte, A. & Golubic, S. (1991). Morphological and genetic criteria in the taxonomy of *Cyanophyta/Cyanobacteria*. *Arch Hydrobiol* 64, 1–24.
- Willemotte, A. & Herdman, M. (2001). Phylogenetic relationships among the cyanobacteria based on 16S rRNA sequences. In *Bergey's Manual of Systematic Bacteriology*, 2nd edn, vol. 1, pp. 487–493. Edited by D. R. Boone & R. W. Castenholz. New York: Springer.
- Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F. & Papke, R. T. (2006). Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res* 16, 1099–1108.