Intra–genomic 16S rRNA gene heterogeneity in cyanobacterial genomes

Niclas Engene[1] & William H. Gerwick*[1]

[1]*Scripps Institution of Oceanography, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093–0212, USA; *corresponding author e–mail: wgerwick@ucsd.edu*

Abstract: The ability of the small ribosomal subunit (16S) rRNA to infer fine–scale phylogenetic relationships is, in principal, impaired by the presence of multiple and variable gene copies within the same genome. This study investigated the extent of intra–genomic 16S rRNA gene heterogeneity in cyanobacteria. Using bioinformatics, all available sequenced cyanobacterial genomes were screened for microheterogeneity between their paralogous ribosomal genes. As a result, cyanobacteria were found to commonly contain multiple ribosomal operons and the numbers of copies were relatively proportional to genome size. Moreover, intra–genomic paralogous 16S rRNA gene copies often contain point–mutations that were validated by secondary structure modeling to be true point mutations rather than sequencing errors. Although microheterogeneity between paralogous 16S rRNA genes is relatively common in cyanobacterial genomes, the degree of sequence divergence is relatively low. We conclude that cyanobacterial intra–genomic 16S rRNA gene heterogeneity generally has a relatively small impact on species delineation and inference of evolutionary histories of cyanobacteria.

Keywords: 16S rRNA, cyanobacteria, genome, intra–genomic gene heterogeneity, microheterogeneity

Introduction

In consistency with the unifying biological theory of evolution, a framework of phylogenetic relationships is necessary for establishing informative and reliable taxonomic systems (Pleijel & Rouse 2000). In the transition from morphology to phylogeny based classification systems, it is crucial to carefully determine optimum evolutionary metrics as well as their capacities to infer evolutionary relationships. In this regard, the small ribosomal subunit (SSU) rRNA gene has long been considered the "gold standard" in interpreting evolutionary relationships, and the 16S rRNA gene has consequently been embraced in cyanobacterial classification systems (Woese et al. 1990; Wilmotte & Herdman 2001; Hoffman et al. 2005).

A recognized limitation in using the 16S rRNA genes has been that ribosomal genes often occur in multiple copies in microbial genomes and, importantly, that paralogous gene copies can vary in sequence between the different operons (Acinas et al. 2004; Case et al. 2007). Recent studies have detected intra–genomic gene heterogeneity of the 16S rRNA gene in cyanobacteria of the genus *Lyngbya* (Engene et al. 2010). In this cyanobacterial genus the genetic variation between the paralogous 16S rRNA gene copies was of equivalent magnitude to sequence variation between different morphological species. This degree of intra–genomic gene heterogeneity challenges the ability of this gene to reliably distinguish between different species, let alone between different sub–species. However, the frequency and amount of general intra–genomic gene heterogeneity in cyanobacteria is, to date, unclear. *In silico* investigations of sequenced bacterial genomes have been conducted, but these studies included a limited number of cyanobacterial genomes (Coenye & Vandamme 2003; Acinas et al. 2004; Case et al. 2007). This study was aimed at assessing general heterogeneity in cyanobacterial 16S rRNA genes through a bioinformatic screening of sequenced cyanobacterial genomes. Moreover, the implication of potential heterogeneity on fine–scale phylogenetic resolution was considered.

Materials and Methods

Cyanobacterial genomes were obtained from the National Center for Biotechnology information (NCBI) Microbial Genome Database web pages (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html). The BLASTn 2.2.18 alignment algorithm available at the NCBI web page was used to screen and obtain gene sequences as well as to determine percent sequence divergence among them. Multiple sequence alignments were performed using the ClustalW algorithm (Thompson et al. 1994). Secondary RNA structures were predicted by the CLC Combined Workbench 3.5.2 (CLC bio, Cambridge, MA, USA). Mutation types and domains of the 16S rRNA genes were determined by superimposing their secondary structures on the SSU model for *Escherichia coli* strain J01695 (Cannone et al. 2002). A total of twelve house–keeping genes (DNA–G, FRR, rpsB, NusA, PGK, PyrG, rpoB, rpsC, rpl2, rpl3, rpl4, and TSF) were downloaded and concatenated for phylogenetic inference of the cyanobacterial strains. All gene sequences were aligned using the L–INS–I algorithm in MAFFT 6.717 and manually refined. Appropriate nucleotide substitution model (WAG+I+G) was selected using Akaike information criterion (AIC/AIC$_C$) and Bayesian information criterion (BIC) in jModeltest. The Maximum likelihood (ML) inference was performed using RaxML. The analysis was run using the WAG+I+G model (selected by AIC and BIC criteria) assuming a heterogeneous substitution rates and gamma substitution of variable sites (proportion of invariable sites (pINV) = 0.265, shape parameter (α) = 0868, number of rate categories = 4). Bootstrap resampling was performed on 500 replicates.

Results and Discussion

Redundancy of ribosomal genes

To gain insight into the redundancy of cyanobacterial ribosomal genes, all 59 publicly accessible cyanobacterial genomes were assessed for the number of ribosomal genes (rrs) and their clustering into ribosomal operons (rrn) (Table 1). In the vast majority (>97%) of the cyanobacterial rrn's, the ribosomal genes were organized in the typical bacterial 16S–23S–5S gene configuration. The number of rrn's varied between one to four copies with an average of 1.8 ± 0.8 copies per genome. This is relatively low compared with the 4.2 rrn's present in an average bacterial genome (Case et al. 2007). A plausible explanation for the average low copy number of ribosomal operons in cyanobacterial genomes was the fact that 45 out of the 59 sequenced genomes belonged to the unicellular order Chroococcales and among the genomes that possessed one rrn per genome, more than 77% belonged to these unicellular forms (Fig. 1). Unicellular cyanobacteria are mainly pelagic, and similar trends of smaller rrn copy numbers are also observed in the pelagic filamentous *Arthrospira maxima* CS–328, *Arthrospira* sp. PCC 8005 and *Arthrospira platensis* Paraca, which live predominately in the open ocean environment. The open ocean represents a relatively constant habitat that allows organisms to specialize in their environment. Specialized organisms normally have less need for rapid adjustments and consequently possess a lower number of rrn copies compared with more generalized organisms (Klappenbach et al. 2000). Moreover, Chroococcales have relatively small size genomes (mean: 3.2 ± 1.5 Mbp), ranging from 1.6 Mbp in *Prochlorococcus marinus* MIT 9301 to *Acaryochloris marina* MBIC11017 with an abnormally large genome size of 6.5 Mbp. Thus, their lower rrn copy numbers may be a consequence of the smaller genome sizes of these unicellular cyanobacteria.

In filamentous cyanobacteria, the genomes increase in size and were found to typically contain multiple rrn copies. For example, the average genome size for filamentous forms of cyanobacteria was 6.1 ± 1.6 Mbp and contained an average of 2.4 ± 1.3 rrn's. The order Oscillatoriales (n = 7 strains) has an average genome size of 6.8 ± 1.2 Mbp and contained an average of 1.7 ± 1.2 rrn's, while the heterocystous order Nostocales (n = 7 strains) has an average genome size of 5.5 ± 1.7 Mbp and contained an average of 3.0 ± 1.2 rrn's. Smaller genomes with few ribosomal operons may, therefore, be a consequence of specialized organisms living in a uniform environment.

16S rRNA gene heterogeneity

In total, 62.7% of all cyanobacterial genomes and 64.3% of filamentous forms contained more than one ribosomal operon (Table 1). Among these 37 cyanobacterial genomes with multiple rrn's, more than one third (35.1%) of the genomes displayed sequence divergence between at least two of their 16S rRNA gene copies. As these assessments are based on bioinformatic analyses of sequenced genomes, it is important to note that assembly programs typically assemble closely related (<6%) sequence reads together and form consensus sequences. The formation of consensus sequences has been found to overlook sequence variations between paralogous 16S rRNA gene copies, as observed in the unpublished genome assembly of *Lyngbya majuscula* 3L, and thus, genome sequences likely underestimate their microheterogeneity (E. Monroe, personal communication).

The intra–genomic sequence divergence between the 16S rRNA genes of different rrn's ranged from 0% to 0.6% with an average divergence of 0.2% in all cyanobacteria. This value is relatively low compared with other gram–negative bacteria (Case et al. 2007; Rastogi et al. 2009). The intra–genomic gene sequence divergence in filamentous cyanobacteria was slightly higher with 0.3% divergence, which could correlate to variety between the higher numbers of copies. However, there also appears to be a trend between genomes with higher gene sequence divergence and secondary metabolite diversity. For example, the microcystin–producing *Microcystis aeruginosa* NIES–843 and the hepatotoxic *Nodularia spumigena* CCY 9414 both have relatively high rates of divergence between their 16S rRNA genes (0.3% and 0.5% sequence divergence, respectively). A potential reason for this large divergence may be that the genomes of such cyanobacteria undergo higher rates of recombination, which ultimately results in a richer diversity of secondary metabolites as well as a higher rate of intra–genomic gene heterogeneity in house–keeping genes.

Mutation types in intra–genomic 16S rRNA gene variations

The mutation types and locations between paralogous 16S rRNA genes were examined by superimposing the sequenced genes on established secondary structure models. This was undertaken to examine the likelihood that the nucleotide substitutions in sequenced genomes were true mutations and not artifacts of genome sequencing. Summarizing the 41 mutations responsible for intra–genomic 16S rRNA gene heterogeneity in sequenced cyanobacterial genomes, 61% were nucleotide substitutions, 37% were nucleotide deletions and only 2% were nucleotide insertions. Among the 24 nucleotide substitutions, 79% were transitional mutations and 21% were transversional mutations (Table 1). Transitional mutations are usually point–mutations which result from errors caused by DNA–polymerases during replication and are estimated to constitute approximately two thirds of all nucleotide substitutions (Collins & Jukes 1994).

The majority of the nucleotide substitutions were located in the ribosomal loop–regions, either in hairpin–loops (44%) or in interior–loops (31%) of the gene helices. The mutations occuring in stem–regions (25%) were typically limited to cytosine to thymine substitutions and only occurred if the nucleotide on the complementary DNA strand was a guanine. The resulting uracil and guanine base pairing is energetically allowed, even though it is slightly more constraining than the cytosine and guanine Watson–Crick base–pairing (Gautheret et al. 1995). Thus, the observed nucleotide substitutions in these genomic 16S rRNA genes were typically the result of thermodynamically allowed gene mutations. The secondary structures of ribosomal genes are essential for ribosome assemblage and ribosomal protein interactions (Van de Peer et al. 1996). Therefore, the conserved secondary structures observed in these cyanobacterial 16S rRNA gene sequences are unlikely to affect ribosomal function, and are thus tolerated without strong selection pressure. On the basis of the types of mutations and their locations, we predict that all or a large proportion of the observed nucleotide variations are true mutations rather than artifacts from genome sequencing.

Conclusions

Cyanobacteria, like most bacteria, frequently have multiple copies of their ribosomal operons, and these appear to be reasonably correlated with genome size of the organism. Among cyanobacteria with multiple ribosomal gene copies, variations between gene copies are relatively frequent and may be underestimated due to limitations of genome assembly programs. Despite the relatively frequent occurrence of intra–genomic 16S rRNA gene heterogeneity, the extent of sequence divergence is typically quite small (mean = 0.2%) and generally much less than in other groups of bacteria (Case et al. 2007). Furthermore, the average cyanobacterial intra–genomic 16S rRNA gene heterogeneity (mean = 0.2%) is of an order less than the 3% gene sequence divergence typically used to delineate microbial species (Tindall et al. 2010). It is, however, important to note that much higher degrees of intra–genomic gene heterogeneity (>1%) have been reported in filamentous marine cyanobacteria (Engene et al. 2010). Moreover, single nucleotide substitutions can, in principal, have large impacts on the

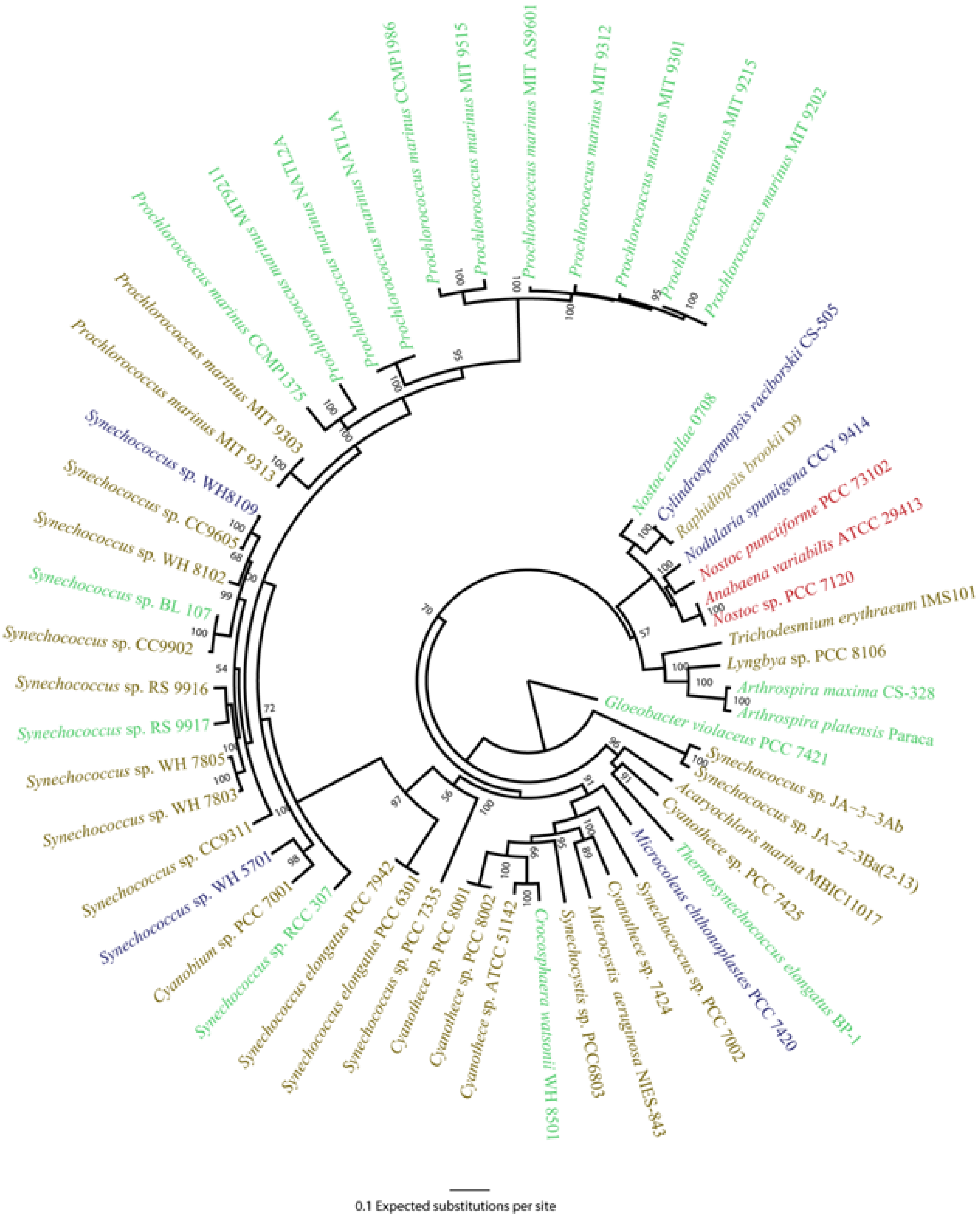

Fig. 1. Phylogenetic relationships of the 59 cyanobacterial strains with sequenced genomes. The number of paralogous 16S rRNA gene copies within each genome is displayed in the following colors: (green) 1 copy; (brown) 2 copies; (blue) 3 copies; (red) 4 copies. The phylogenetic inference was performed by concatenation of twelve house–keeping genes using RaxML. The scale bar is equivalent to 0.1 expected substitutions per nucleotide position.

Table 1. Ribosomal operon redundancy and intra–genomic SSU rRNA gene heterogeneity in cyanobacterial genomes.

Species (strain)	Acc. Nr.	Genome size (Mbp)	SSU Nr.[a]	Variable SSU	Gene length (bp)	P–distance (%)	Hetero. Positions	Mutation types[c]	Structural location
Prochlorococcus (n = 13)									
P. marinus MIT 9301	CP000576	1.6	1	–	1498	–	–	–	–
P. marinus MIT 9202	ACDW0000	1.7	1	–	1441	–	–	–	–
P. marinus AS 9601	CP000551	1.7	1	–	1465	–	–	–	–
P. marinus CCMP1986	CP000951	1.7	1	–	1465	–	–	–	–
P. marinus MIT 9211	CP000878	1.7	1	–	1465	–	–	–	–
P. marinus MIT 9312	CP000111	1.7	1	–	1465	–	–	–	–
P. marinus MIT 9215	CP000825	1.7	1	–	1465	–	–	–	–
P. marinus MIT 9515	CP000552	1.7	1	–	1465	–	–	–	–
P. marinus CCMP1375	AE017126	1.8	1	–	1465	–	–	–	–
P. marinus NATL1A	CP000553	1.9	1	–	1465	–	–	–	–
P. marinus NATL2A	CP000095	1.9	1	–	1451	–	–	–	–
P. marinus MIT9313	BX548175	2.4	2	0	1465	–	–	–	–
P. marinus MIT9303	CP000554	2.7	2	0	1401	–	–	–	–
Synechococcus (n = 18)									
Synechococcus sp. WH 8109	ACNY0000	2.1	3	0	1456	–	–	–	–
Synechococcus sp. RCC 307	CT978603	2.2	1	–	1498	–	–	–	–
Synechococcus sp. CC9902	CP000097	2.2	2	0	1479	–	–	–	–
Synechococcus sp. BL107	AATZ0000	2.3	1	–	1479	–	–	–	–
Synechococcus sp. WH 8102	BX548020	2.4	2	2	1462/1464	0.1	2	2d	HL, IL
Synechococcus sp. WH 7803	CT971583	2.4	2	0	1497	–	–	–	–
Synechococcus sp. CC 9605	CP000110	2.5	2	0	1440	–	–	–	–
Synechococcus sp. RS9917	AANP0000	2.6	1	–	1448	–	–	–	–
Synechococcus sp. WH 7805	AAOK0000	2.6	2	2	1498/1497	0.1	1	d	IL
Synechococcus sp. CC 9311	CP000435	2.6	2	2	1477	0.1	1	ti	HL
Synechococcus sp. RS 9916	AAUA0000	2.7	2	2	1498	0.1	1	d	IL
S. elongatus PCC 6301	AP008231	2.7	2	0	1489	–	–	–	–

Table 1 Cont.

Species (strain)	Acc. Nr.	Genome size (Mbp)	SSU Nr.[a]	Variable SSU	Gene length (bp)	P–distance (%)	Hetero. Positions	Mutation types[c]	Structural location
S. elongatus PCC 7942	CP000100	2.7	2	0	1490	–	–	–	–
Synechococcus sp. JA–3–3Ab	CP000239	2.9	2	0	1480	–	–	–	–
Synechococcus sp. PCC 7002	AB015058	3.0	2	0	1452	–	–	–	–
Synechococcus sp. WH 5701	AANO0000	3.0	3	0	1447	–	–	–	–
Synechococcus sp. JA–2–3B'a(2–13)	CP000240	3.0	2	0	1479	–	–	–	–
Synechococcus sp. PCC 7335	AB015062	6.0	2	0	1448	–	–	–	–
Cyanothece (n = 7)									
Cyanothece sp. PCC 8801	CP001287	4.7	2	0	1479	–	–	–	–
Cyanothece sp. PCC 8802	CP001701	4.7	2	2	1479	0.1	1	ti	HL
Cyanothece sp. ATCC 51142	CP000806	4.9	2	2	1489	0.1	1	tv	S
Cyanothece sp. PCC 7822	ABVE0000	5.7	2	0	1492	–	–	–	–
Cyanothece sp. PCC 7425	CP001344	5.4	2	0	1474	–	–	–	–
Cyanothece sp. CCY 0110	AAXW0000	5.9	2	0	1488	–	–	–	–
Cyanothece sp. PCC 7424	CP001291	5.9	2	0	1483	–	–	–	–
Other Chroococcales (n = 7)									
Thermosynechococcus elongatus BP–1	BA000039	2.6	1	–	1491	–	–	–	–
Cyanobium sp. PCC 7001	ABSE0000	2.8	2	0	1448	–	–	–	–
Synechocystis sp. PCC 6803	BA000022	3.6	2	0	1489	–	–	–	–
Gloeobacter violaceus PCC 7421	BA000045	4.7	1	–	1484	–	–	–	–
Microcystis aeruginosa NIES–843	AP009552	5.8	2	2	1477	0.3	4	4ti	2IL, 2HL
Crocosphaera watsonii WH 8501	AADV0000	6.2	1	–	1408	–	–	–	–
Acaryochloris marina MBIC11017	CP000828	6.5	2	0	1500	–	–	–	–
Oscillatoriales (n = 7)									
Arthrospira platensis Paraca	ACKSK0000	5.0	1	–	1483	–	–	–	–
Arthrospira maxima CS–328	ABYK0000	6.0	1	–	1482	–	–	–	–
Arthrospira sp. PCC 8005	ADDH0000	6.1	1	–	1482	–	–	–	–
Oscillatoria sp. PCC 6506	CACA00000000	6.7	1	–	1480	–	–	–	
Lyngbya aestoarii PCC 8106	AAVU0000	7.0	2	2	1493/1494	0.3	4	3ti, i	2HL, 2IL

Table 1 Cont.

Species (strain)	Acc. Nr.	Genome size (Mbp)	SSU Nr.[a]	Variable SSU	Gene length (bp)	P–distance (%)	Hetero. Positions	Mutation types[c]	Structural location
Microcoleus chthonoplastes PCC7420	ABRS0000	8.7	4	3	1482	0.6	9	2ti, tv, 6d	3HL, S*, 4IL
Trichodesmium erythraeum IMS101	CP000393	7.8	2	0	1482	–	–	–	–
Nostocales (n = 7)									
Raphidiopsis brookii D9	ACYB0000	3.2	2	2	1487	0.1	1	ti	S*
Cylindrospermopsis raciborskii CS–505	ACYA0000	3.9	3	0	1487	–	–	–	–
Nostoc azollae 0708	ACIR0000	5.0	1	–	1492	–	–	–	–
Anabaena variabilis ATCC 29413	CP000117	6.4	4	0	1488	–	–	–	–
Nostoc punctiforme PCC 7120	BA000019	6.4	4	2	1489	0.1	1	ti	5'
Nostoc punctiforme PCC 73102	CP001037	8.2	4	2	1489	0.1	2	2ti	2S*
Nodularia spumigena CCY 9414	AAVW0000	5.3	3	3	1563/1566/1485	0.5	13	5ti, 3tv, 5d	3IL, 2S*, HL, 4C, 3S

interpretations of DNA fingerprinting methods if these mutations are located in restriction sites (Roudière et al. 2007). Intra–genomic gene heterogeneity should, therefore, not be neglected when inferring phylogenetic relationships of cyanobacteria, especially on an infra–species level. However, we argue that the general degree of microheterogeneity in paralogous 16S rRNA gene copies of cyanobacteria is relatively small and, thus, has only minor impact on the inference of phylogenetic relationships and evolutionary histories.

Acknowledgements

We gratefully thank S. Podel and E. Monroe for bioinformatics advice and discussions. Financial support for this study came from the Swedish–American Foundation (NE), NIH grants (CA–108874 and NS053398).

References

Acinas, S.G., Marcelino, L.A., Klepac–Ceraj, V. & Polzm M. F.(2004): Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. – J. Bacter. 186: 2629–2635.

Cannone, J.J., Subramanin, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V., Muller, K.M., Pnde, N., Schang, Z., Yu, N. & Gutell R. R.(2002): The Comparative RNA Web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. – BMC Bioinformatics 3: 1471–2105.

Case, R.J., Boucher, Y., Dahllof, I., Holmstrom, C., Doolittle, W.F. & Kjelleberg, S.(2007): Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. – Appl. Env. Mic. Biol. 73: 278–288.

Coenye, T. & Vandamme, P.(2003): Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. – FEMS Microbiol. Lett. 228: 45–49.

Collins, D. W. & Jukes, T.H.(1994): Rates of transition and transversion in coding sequences since the human–rodent divergence. – Genomics 20: 386–96.

Engene, N., Coates, R.C. & Gerwick, W.H. (2010): 16S rRNA gene heterogeneity in the filamentous marine cyanobacterial genus *Lyngbya*. – J. Phycol. 46: 591–601.

Gautheret, D., Konings, D. & Gutell, R.R.(1995): G.U base pairing motifs in ribosomal RNA. –

RNA 1: 807–814.

Hoffman, L., Komárek, J. & Kaštovský, J.(2005): System of cyanoprokaryotes (cyanobacteria) – state in 2004. –Algological Studies 117: 95–115.

Klappenbach, J.A., Dunbar, J.M. & Schmidt, T.M.(2000): rRNA Operon Copy Number Reflects Ecological Strategies of Bacteria. – Appl. Env. Mic. Biol. 66: 1328–1333.

Pleijel, F. & Rouse, G. W.(2000): Least–inclusive taxonomic unit: a new taxonomic concept for biology. – Proc. R. Soc. Lond. 267: 627–630.

Rappe, M.S., Connon, S.A., Vergin, K.L. & Giovannoni, S.J.(2002): Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. – Nature 418: 630–633.

Rastogi, R., Wu, M., DasGupta, I. & Fox, G.E.(2009): Visualization of ribosomal RNA operon copy number distribution. – BMC Microbiol. 25: 208.

Roudière, L., Lorto, S., Tallagrand, E., Marchandin, H., Jeannot, J.–L. & Jumas–Bilak, E.(2007): Molecular fingerprint of bacterial communities and 16S rDNA intra–species heterogeneity: a pitfall that should be considered. – Pathologie Biologie 55: 434–440.

Tindall, B.J., Rosselló–Móra, R., Busse, H.–J., Ludwig, W. & Kämpfer, P. (2010): Notes on the characterization of prokaryote strains for taxonomic purposes. – Int. J. Syst. Evol. Microbiol. 60: 249–266.

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994): CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position–specific gap penalties and weight matrix choice. – Nucleic Acids Res. 22: 4673–4680.

Van de Peer, Y., Chapelle, S. & De Wachter, R.(1996): The European database on small subunit ribosomal RNA. – Nucleic Acids Res. 24: 3381–3391.

Wilmotte, A. & Herdman, M.(2001): Bergey's Manual of Systematic Bacteriology, 2nd ed. – pp. 487–493, Springer, New York.

Woese, C.R., Kandler, O. & Wheelis, M.L.(1990): Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. – Proc. Nati. Acad. Sci. USA. 87: 4576–4579.

Recieved Sept 2010
Accepted Dec 2010