3 Biotech (2019) 9:87
https://doi.org/10.1007/s13205-019-1635-6

ORIGINAL ARTICLE

Cyanobacterial phylogenetic analysis based on phylogenomics approaches render evolutionary diversification and adaptation: an overview of representative orders

Ratna Prabha[1,2] · **Dhananjaya P. Singh**[1]

Received: 1 August 2018 / Accepted: 11 February 2019 / Published online: 15 February 2019

Abstract

Phylogenetic studies based on a definite set of marker genes usually reconstruct evolutionary relationships among the prokaryotic species. Based on specific target sequences, such studies represent variations and allow identification of similarities or dissimilarities in organisms. With the advent of completely sequenced genomes and accumulation of information on whole prokaryotic genomes, phylogenetic reconstructions should be considered more reliable if they are ideally based on entire genomes to resolve phylogenetic interest. We applied phylogenomics approaches taking into account completely sequenced cyanobacterial genomes to reconstruct underlying species that represented major taxonomic classes and belonged to distinctly different habitats (freshwater, marine, soils, and rocks). We did not rely on describing phylogeny of all representative class of cyanobacterial species on the basis of only ribosomal gene, 16S rDNA gene. In contrast, we analyzed combined molecular marker and phylogenomics approaches (genome alignment, gene content and gene order, composition vector and protein domain content) for accurately inferring phylogenetic relationship of species. We have shown that this approach reflects the impact of evolution on the organisms and considers connects with the ecological adaptation in cyanobacteria in different habitats. Analysis revealed that the members from marine habitat occupy different profile than those from freshwater. Impact of GC content and genomic repetitiveness over the diversification of cyanobacterial species and their possible role in adaptation was also reflected. Members occupying similar habitats cover more evolutionary distance together and also evolve various strategies for adaptation and survival either through genomic repetitiveness or preferences for genes of particular functions or modified GC content. Genomes undergo different changes for their adaptation in diverse habitats.

Keywords Cyanobacterial evolution · Phylogeny · Ecological adaptation · Genomic repetitiveness · Functional profile · Phylogenomics · Cyanobacteria

Introduction

The universal 'tree of life' constructed on the basis of molecular analysis of ribosomal RNA (rRNA) genes has led to molecular classification of microorganisms (Woese and Fox 1977; Woese 1987). Although the rRNA-based tree of life is the most common and widely accepted approach for microbial identification, it is not quite sufficient to resolve accurate phylogeny of interest (Eisen 2000; Capella-Gutierrez et al. 2014; Adato et al. 2015). The most prominent objection disfavouring this approach was whether a single-gene tree solely represents the evolutionary history of the organisms (Eisen 2000; Sleator 2013). In molecular taxonomy, a single tree of life generally reflects species relatedness through vertical descent. However, not all genes follow similar tendency. Many genes are transferred between lineages horizontally or

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13205-019-1635-6) contains supplementary material, which is available to authorized users.

✉ Dhananjaya P. Singh
dpsfarm@rediffmail.com

[1] ICAR-National Bureau of Agriculturally Important Microorganisms, Kushmaur, Maunath Bhanjan 275101, India

[2] Department of Biotechnology, Mewar University, Gangrar, Chittorgarh, Rajasthan, India

laterally via 'Horizontal Gene Transfer' (HGT), a phenomenon quite frequent in prokaryotic organisms. Such gene transfers within species usually complicate the evolutionary reconstruction based on single gene. They represent that some species are chimeric in nature having diverse histories for different parts of their genome (Eisen 2000; Prasanna and Mehra 2013; Rinke et al. 2013). Multiple approaches were derived to explore phylogeny among organisms on the basis of various structural and functional genes and protein sequences (Sawa et al. 2003; Auch et al. 2010; Lang et al. 2013). Such methods end with troubles in deciphering evolutionary history of organisms due to duplication, deletion or horizontal gene transfer. Another problem in gene(s) based approach lies in the identification of gene to be used for phylogenetic analysis as it should be significantly conserved across the diverse genomes (Sawa et al. 2003). There existed some reports on the phylogenetic analysis of archaea and bacteria using universal genes (Nelson et al. 1999; Makarova et al. 1999). It was, therefore, suggested that the tree generated from universal genes is not necessarily accurate due to factors such as convergence, misidentification of orthologs, gene conversion and ambiguities in sequence alignment (Eisen 2000). The molecular phylogenies that have led to the classification of three primary kingdoms or domains were based on single characteristic gene sequence such as 16S rRNA gene (Weisburg et al. 1991; Li et al 2002; Větrovský and Baldrian 2013). However, using various characteristic gene sequences for phylogeny reconstruction leads to contrast conclusions reflecting problems owing to horizontal gene transfer, unrecognized paralogy and highly variable rates of evolution (Woese 1998; Rudi and Sekelja 2013). Therefore, equating the history of the whole genome with that of a single although highly conserved gene or only a part of the genome or a major region is not conclusive.

Theoretically, the evolutionary history of different genomes can be assessed by comparing phylogenetic trees for each gene in every genome (Eisen 2000). Practically, such approaches have their own limitations because phylogenetic tree reconstruction is cumbersome when every gene in the genome is considered. Ideally, for making the genetic tree reconstruction reliable, the set of genes analyzed should be present in all the species considered, sequence alignments need to be carefully examined and ambiguous or hypervariable alignment regions should be excluded from phylogenetic analysis. Phylogenetic reconstructions based on whole genome sequences are emerging as the most reliable alternatives (Takahashi et al. 2009). It allows comparison of entire gene families or overall gene content, insertions and deletions or HGT events (Sawa et al. 2003). Evaluation of closely related genome sequences can endow with an evidence of macroscopic genome polymorphism which occurs during the phase of recombination processes (Kawai et al. 2006).

Conventional phylogenetics and phylogenomics

Evolutionary biology has witnessed revolution with the disclosure of fact that comparison of gene sequences can reveal the evolutionary history of species (Losos et al. 2013) Ribosomal RNA was analyzed for the interpretation of evolutionary classification of microbes as universal tree of life (Zuckerkandl and Pauling 1965; Woese 1987; Woese and Fox 1977; Eisen 2000). Archaea was recognized as the third domain of life along with well-accepted domains Bacteria and Eukarya on the basis of ssu-rRNA (Rajendhran and Gunasekaran 2011). For evolutionary studies, ssu-rRNA is broadly accepted as 'phylogenetic marker' that possesses features such as presence across all three domains of life, highly conserved regions, sequence and structure level conservation and sequence variations to facilitate studies on evolutionary events (Wu et al. 2013). The gene also serves the purpose of 'ecological marker' along with 'phylogenetic marker' for the analysis of microbes (Wu et al. 2013; Moreno-Hagelsieb et al. 2013). Along with microbes, these gene sequences were also used as phylogenetic marker for other organisms and deciphered valuable information about their phylogenetic relationships. Ribosomal rRNA genes are present in many copies per genome and repeated copies have the similar prototype of concerted evolution which facilitates its analysis either by direct RNA and DNA sequencing or by restriction enzyme methodologies (Hillis and Dixon 1991; Bryant et al. 2013).

Though, there exists many controversies with rRNA gene based tree of life, major concern relates to the use of single gene for defining evolutionary history (Eisen 2000; Daubin and Szöllősi 2016). ssu-rRNA genes also have wide deviation in copy number among different organisms which poses limitation on estimation of their relative abundance across different groups. Primers which are universally used for amplification of ssu-rRNA genes also tend to have slight biasness towards certain taxonomic groups. This gene is reported to influenced by the processes such as HGT, convergent evolution or variations in evolutionary rates in due course of time (Wu et al. 2013). Among phylogeny based on this gene, limited mutable sites and restricted length generates the problem of saturation (Henz et al. 2005). In addition, 16S rRNA gene sequence identity does not correlate with DNA–DNA hybridization (DDH) values which are critical conclusive factor for establishment of new species (Klenk and Göker 2010; Kim et al. 2014). In addition to 16S rRNA genes, 23S rRNA, the β-subunit of F1F0 ATPase, DNA gyrase b gene and elongation factor Tu is also used as phylogenetic marker genes (Ludwig and Schleifer 1999). Though again, the major problem associated with the use of these genes

as phylogenetic marker lies in the fact that whether a single gene can resolve evolutionary history of any species. There are experimental facts that rRNA gene is also subjected to horizontal gene transfer (Li et al. 2002; Yabuki et al. 2014) and single-gene-based phylogeny is unable to cope with horizontal gene transfer, unrecognizable paralogy and extremely inconsistent rates of evolution (Wu et al. 2013).

Evolution of gene content in a genome is a complicated and unresolved process (Snel et al. 2002; Comas et al. 2006). Studies revealed that presence or absence of genes between genomes is dependents on the genome size (Snel et al. 1999) and evolutionary distance (Snel et al. 2002). Genome size is ruled by many different factors involving addition/deletion of genetic information, amplifications, horizontal transfer and selection. Prokaryotic genomes tend to have strong deletion biases where DNA lacking adaptive value is rapidly deleted (Mira et al. 2001). This has indicated that the recombination leads to deletions more often than amplifications (Treangen et al. 2009). Larger genomes are reported to preferentially obtain genes involved in regulation, secondary metabolism and energy conversion. Bacteria with larger genomes are ecologically more successful in the environment where resources are varied and poor (Prabha et al. 2016). Complete understanding of these facts will provide detail insight into interaction between ecology and genome evolution (Konstantinidis and Tiedje 2004).

Genomics technologies opened array of scope for microbial biologists to understand factors leading to complex evolutionary patterns of the genomes (Eisen 2000; Wolf et al. 2001; Mirkin et al. 2003; Henz et al. 2005; Zhao et al. 2012). Microbial genomes acquire foreign DNA from surroundings more frequently as compared to higher organisms through HGT events (Capella-Gutierrez et al. 2014). HGT in combination with intra-genomic rearrangements and duplication events leads to bacterial adaptations to different environmental niches and variance in closely related species (Sicheritz-Pontén and Anderssona 2001; Oliveira et al. 2017). Genome size, geometry, GC content and gene number are the important parameters and deviation in any of these parameters reflects typical change in the bacterial genomes (Bentley and Parkhill 2004). Many genomes reflects changes in genome size and gene content as directed by the processes such as deletion, duplication and lateral gene transfer events (Nilsson et al. 2005; Cordero and Hogeweg 2009). Overall, the genome size of bacteria is maintained in an equilibrium between the duplication or HGT and mutations leading to elimination of function followed by deletions (the loss of genes) (Wernegreen et al. 2000). For better understanding of this phenomenon, in-depth phylogenetic analysis is required at genome level (Sicheritz-Pontén and Anderssona 2001). Increasing number of complete genome sequences makes it possible to use wealth of genomic information for phylogenetic reconstruction, focus on entire genome and its genes rather than a single gene or group of genes (Wolf et al. 2001; Mirkin et al. 2003; Henz et al. 2005). Availability of large number of bacterial whole genome sequences facilitated biologists to explore and examine evolutionary hypotheses on a larger scale than ever (Zhao et al. 2012; Land et al. 2015).

Phylogenomics refer to such studies that involve large-scale genomic comparisons for phylogenetic reconstruction. This reflects complex evolutionary pattern for microbes that involve not only vertical descent or lateral gene transfer but also include a mix of recombination, duplication, invention, loss, degradation and convergence of genes during selection processes (Eisen 2000; Meier-Kolthoff et al. 2014). Phylogenomics approaches hold promise for better interpretation of genome organization and function. Recently sequenced genomes provided a clear picture of genome evolution and phylogeny (Medina 2005). Enormous information lead to decipher salient genomic features in terms of gene content and gene order and create reliable phylogenetic reconstructions. Based on the information of gene and gene family content, gene order, protein domain content, protein orthologs and fold information in the genomes, whole genome phylogenetic trees have become more reliable (Klenk and Göker 2010). Though in prokaryotic genomes, gene content and gene order phylogenies are influenced by gene loss and horizontal gene transfer (Klenk and Göker 2010).

Phylogenetic trees are, therefore, valuable means for analyses such as taxonomy assignment, metagenomics studies, inference of co-speciation, identification of ecological trends, epidemiological and biogeographical events, phylogenetic profiling analysis and genomes selection for sequencing (Lang et al. 2013). Currently, reliability in methods, tools, and approaches raises expectations from phylogenomics analysis to infer identification of taxon-specific gene families as a source of explicit physiological features, taxonomic and evolutionary purposes and lacuna in single-gene phylogenies (Klenk and Göker 2010; Lang et al. 2013).

Approaches for phylogenomics analysis

Genome trees were suggested to capture overwhelming information of the phylogeny. Approaches for whole genome-based phylogenies can be broadly classified into three categories as based on (1) sequence alignment, (2) gene content and gene order, and (3) sequence statistics. Certain parameter-free and whole-genome-based composition vector approaches were also developed for phylogenetic reconstruction (Qi et al. 2004; Hao and Qi 2004; Qiang et al. 2010; Bromberg et al. 2016). Thus, phylogenetic trees considering complete genome sequences can be reconstructed into five different ways: (1) alignment-free trees in which statistic properties of genome are considered; (2) gene

content trees in which presence or absence of genes are considered; (3) trees based on chromosomal gene order; (4) trees based on average sequence similarity; and (5) phylogenomics based genome trees (Snel et al. 2005). In another study, four major approaches considering (1) gene and/or protein content, (2) gene order, (3) shared gene content, and (4) information theory and genome compression were mentioned for phylogenetic analysis (Khiripet 2005).

Alignment-based methods were in practice from the time of rRNA sequence rooted phylogenetic trees. However, owing to sequence size limitations, alignment-based methods are complicated for application at the entire genome level. Furthermore, it is also not feasible to carry out multiple sequence alignment at whole genome level because of huge sequence size (from millions to billions of base pairs) (Klenk and Göker 2010). Gene duplication affects gene content of genomes and thereby generates inconsistency in phylogeny of closely and distantly related genomes. Gene content tends to have strong phylogenetic signal and assist in removing numerous taxonomic uncertainties (Tekaia et al. 1999; Snel et al. 2005; Anselmetti et al. 2018). It, however, depends up on the availability of complete genomes. In case of incomplete genomes, gene content approach can be used via signature genes, where for each clade, core genes ubiquitous in every genome in a phylogenetically coherent group is identified (Charlebois and Doolittle 2004; Dutilh et al. 2008). Gene order is also used for the assessment of phylogenetic relationship in closely related genomes. However, this parameter lacks resolution as genome rearrangement is not a frequent event in nature (Vishnoi et al. 2010; Gu et al. 2005; Zhou et al. 2017). Alignment-free approaches such as *k*-string approach or composition vector approach are also available for the construction of genome trees (Xu and Hao 2009). Such approaches are computationally less costly and utilize utmost content of the genomes (Vishnoi et al. 2010).

Cyanobacterial phylogeny

Cyanobacteria are among the most primitive oxygenic photosynthetic organisms. They have been studied extensively for different biological processes including photosynthesis, bioenergetics, nitrogen fixation, environmental stress adaptation and molecular evolution (Koksharova and Wolk 2002; Cassier-Chauvat and Chauvat 2018). Cyanobacterial genomes reveal a complex evolutionary history (Zhaxybayeva et al. 2006; Prabha et al. 2016). These organisms have found their origin in ancient group of photosynthetic prokaryotes. They have shown distinctions in their habitats, cellular differentiation strategies, physiological capacities and metabolic complexity (Beck et al. 2012). Diversity across various cyanobacteria in terms of their size, gene number and GC content is reflected in their whole genome sequences (Larsson et al. 2011; Prabha et al. 2016). This has facilitated studies on the factors governing variations among the organisms and mechanisms responsible for evolutionary diversification. We, therefore, determined phylogenetic relationship within different cyanobacterial species from diverse taxonomic groups by considering their entire genomic sequence and features (genome alignment, gene content and gene order, protein domain content). The study reflected changes in cyanobacterial genomes towards their adaptation in different ecological niches during the evolution. We also compared conventional and phylogenomics approaches reflecting evolutionary history of cyanobacteria and provides light over process of diversification of these organisms.

Materials and methods

Species and genome sequences

Forty-one cyanobacterial species representing five different taxonomic orders (Chroococcales, Prochlorales, Nostacales, Oscillatoriales and Gloeobacterales), for which complete genome sequences were available, were taken into the study. All the genome sequences were downloaded from NCBI Genome database (http://www.ncbi.nlm.nih.gov/genome/). Maximum number of species (22) represented Chroococcales followed by Prochlorales (12). Order Nostacales, Oscillatoriales and Gloeobacterales had 4, 2, and 1 species, respectively (Table 1).

Genomic features

Genomic features of all of these cyanobacteria were identified.

Genome size, GC content: Information about these parameters was obtained from NCBI Genome database (http://www.ncbi.nlm.nih.gov/genbank/genomes).

%-coding information: Information about how much percentage of genome is coding was calculated as

$$\%\text{-coding} = (\text{number of base pairs involved in coding for gene or other products/total number of base pairs}) \times 100.$$

Gene density: Gene density was calculated as genes/Mb.

Phylogenetic reconstruction on the basis of 16S rRNA gene

For alignment of 16S rRNA gene, MUSCLE (MUltiple Sequence Comparison by Log-Expectation) was used (Edgar 2004). 16S rRNA gene phylogeny was constructed using MEGA 5.0 (Tamura et al. 2011) employing Neighbor-Joining reconstruction with 100 bootstrap iterations. Tree

Table 1 Details of cyanobacterial genomes under consideration [genome size (Mb), GC content (percentage), coding genome (percentage), gene density (genes/Mb), microsatellites (number)]

S. no.	Taxonomy	Organism name	Abbreviation used	Habitat	Genome Size	GC content	Coding genome	Gene density	Micro-satellites
1	Chroococcales	*Acaryochloris marina* MBIC11017	Am_ MBIC11017	Marine	8.36	47	83.26	1025	6826
2		*Cyanothece* sp. ATCC 51142	Cs_ ATCC51142	Marine	5.46	37.9	86.8	983	6540
3		*Cyanothece* sp. PCC 7424	Cs_PCC7424	Fresh water	6.55	38.5	81.46	907	7975
4		*Cyanothece* sp. PCC 7425	Cs_PCC7425	Fresh water	5.79	50.6	85.28	951	7023
5		*Cyanothece* sp. PCC 7822	Cs_PCC7822	Fresh water	7.84	40.1	82.83	898	7512
6		*Cyanothece* sp. PCC 8801	Cs_PCC8801	Fresh water	4.79	39.8	84.85	964	5587
7		*Cyanothece* sp. PCC 8802	Cs_PCC8802	Fresh water	4.8	39.8	85.1	979	5567
8		*Microcystis aeruginosa* NIES-843	Ma_NIES_843	Fresh water	5.84	42.3	81.36	1090	6639
9		*Synechococcus* sp. CC9311	Ss_CC9311	Fresh water	2.61	55.5	87.21	1128	3317
10		*Synechococcus* sp. CC9605	Ss_CC9605	Fresh water	2.51	55.4	86.94	1098	3782
11		*Synechococcus* sp. CC9902	Ss_CC9902	Marine	2.23	52.4	90	1056	2837
12		*Synechococcus* sp. JA-2-3B'a(2–13)	Ss_JA_2_3Ba	Marine	3.05	59.2	85.48	965	4386
13		*Synechococcus* sp. JA-3-3Ab	Ss_JA-3-3Ab	Marine	2.93	54.2	84.86	989	4509
14		*Synechococcus* sp. PCC 7002	Ss_PCC7002	Hot spring	3.41	58.5	87.64	950	3426
15		*Synechococcus* sp. RCC307	SsRCC307	Hot spring	2.22	60.2	94.16	1163	3586
16		*Synechococcus* sp. WH 7803	Ss_WH7803	Marine	2.37	49.2	93.39	1091	3425
17		*Synechococcus* sp. WH 8102	Ss_WH8102	Marine	2.43	60.8	90.3	1062	3580
18		*Synechococcus elongatus* PCC 6301	Se_PCC6301	Marine	2.7	60.2	88.04	956	3289
19		*Synechococcus elongatus* PCC 7942	Se_PCC7942	Marine	2.74	59.4	89.21	991	3324
20		*Synechocystis* sp. PCC 6803	Sy_PCC6803	Fresh water	3.95	47.4	87.14	918	4410
21		*Thermosynechococcus elongatus* BP-1	Te_BP_1	Hot spring	2.59	53.9	89.99	975	2956
22		cyanobacterium UCYN-A	C_UCYNA	Marine	1.44	31.1	81.41	862	2349
23	Gloeobacterales	*Gloeobacter violaceus* PCC 7421	Gv_PCC7421	Rock	4.66	62	89.36	962	7078

Table 1 (continued)

S. no.	Taxonomy	Organism name	Abbreviation used	Habitat	Genome Size	GC content	Coding genome	Gene density	Micro-satellites
24	Nostacales	*Anabaena variabilis* ATCC 29413	Av_ ATCC29413	Multiple	7.11	41.4	82.33	818	7666
25		*Nostoc* sp. PCC 7120	Ns_PCC7120	Multiple	7.21	38.3	82.5	862	7627
26		*Nostoc punctiforme* PCC 73102	Np_PCC73102	Fresh water	9.06	41.4	77.43	791	9646
27		*'Nostoc azollae'* 0708	Na_0708	Multiple	5.49	41.3	52.13	980	7175
28	Oscillatoriales	*Arthrospira platensis* NIES-39	Ap_NIES-39	Fresh water	6.79	44.3	81.24	983	7739
29		*Trichodesmium erythraeum* IMS101	Te_IMS101	Marine	7.75	34.1	60.11	661	12,530
30	Prochlorales	*Prochlorococcus marinus str*. AS9601	Pm_AS9601	Marine	1.67	31.3	91.15	1177	2967
31		*Prochlorococcus marinus str*. MIT 9211	Pm_MIT9211	Marine	1.69	38	90.12	1124	2269
32		Prochlorococcus marinus str. MIT 9215	Pm_MIT9215	Marine	1.74	31.1	89.62	1180	3094
33		*Prochlorococcus marinus str*. MIT 9301	Pm_MIT9301	Marine	1.64	31.3	91.21	1196	2798
34		*Prochlorococcus marinus str*. MIT 9303	Pm_MIT9303	Marine	2.68	50	84.52	1170	3650
35		*Prochlorococcus marinus str*. MIT 9312	Pm_MIT9312	Marine	1.71	31.2	89.59	1085	3042
36		Prochlorococcus marinus str. MIT 9313	Pm_MIT9313	Marine	2.41	50.7	82.23	967	3234
37		Prochlorococcus marinus str. MIT 9515	Pm_MIT9515	Marine	1.7	30.8	88.92	1155	3146
38		*Prochlorococcus marinus str*. NATL1A	Pm_NATL1A	Marine	1.86	35	87.29	1210	2913
39		*Prochlorococcus marinus str*. NATL2A	Pm_NATL2A	Marine	1.84	35.1	85.62	1211	2798
40		*Prochlorococcus marinus* subsp. *marinus str*. CCMP1375	Pm_ CCMP1375	Marine	1.75	36.4	89.22	1103	2421
41		*Prochlorococcus marinus* subsp. *pastoris str*. CCMP1986	Pm_ CCMP1986	Marine	1.66	30.8	88.42	1061	3112

visualization was done via TreeDyn program (Chevenet et al. 2006).

Reconstruction of complete genome-based phylogenies

Distance estimation

On the basis of genome alignment: Alignment of complete genomes was carried out through MUMMER (Delcher et al. 2002) from GGDC web server (Auch et al. 2010) (parameters: coverage algorithm with distance function and 100% identity). DNA–DNA hybridization (DDH) is an extensively used technique for estimation of the overall similarity among the genomes of two organisms (Auch et al. 2010). In fact, the "gold standard" of bacterial species delineation is in general genome similarity identified through DDH, which is a strictly rigorous technique, however, at some instances yields inconsistent results (Colston et al. 2014). In silico approaches for the genome sequence comparison are also available as an alternative of DDH. GGDC web server is based on this only and implies distance methods based as high-scoring segment pairs (HSPs) or maximally unique matches (MUMs) for estimation of similarity among different genomes (Auch et al. 2010).

On the basis of alignment-free composition vector approach: Alignment-free composition vector (CV) approach was used for intergenomic distance estimation from CVTree platform (k-tuple length of 6) (Xu and Hao 2009). Composition vector (CV) approach is an alignment-free method based on K-tuple counting and background subtraction. This approach infers trees from whole-genome data through an alignment-free and parameter-free way (Zuo et al. 2010).

On the basis of overlapping gene content and gene order: For estimation of intergenomic distance on the basis of overlapping genes (OG), OGtree algorithm was used (parameters: 1 for weight of overlapping-gene order and gene content, 1e−9 as threshold of *E* value and 80% for threshold of alignment coverage in each sequence). OGtree constructs OG distance between the genomes by the analysis of both the OG content and OG order (Jiang et al. 2008). The overlapping genes (OGs) are adjoining genes whose coding sequences are partially or completely overlapped (Jiang et al. 2008). Overlapping genes (OGs) correspond to widely available genomic feature of bacterial genomes and are also used as rare genomic markers for phylogenetic analysis of closely related bacterial species (Zhang and Lin 2015). OGs are potentially involved in different important processes including regulation of gene expression or development of genome compaction (Zhang and Lin 2015). In fact, OGs are much conserved among species rather than non-OGs (Jiang et al. 2008).

On the basis of whole-genome protein domain content: Whole-genome protein domain content approach considers protein domains of the entire genome for the estimation of distances. ProdocTree (http://ibi.cqupt.edu.cn/prodoctree/index.php) was used and it uses the bit score of hmmscan result of each Pfam protein domain for calculation of the coordinate of a dimension of a multi-dimensional space and provides euclidean distances between two points in the space where every point is representative of a species. Protein structural domains represent evolutionary units, the relationships of whom can be assessed along long evolutionary distances (Yang and Bourne 2009). Domain architectures are shown to be conserved at large phylogenetic distances, in prokaryotes and eukaryotes, both (Koehorst et al. 2016).

Tree visualization

Neighbor-Net algorithm (Bryant and Moulton 2004) from SplitsTree software (Huson 1998) was used for phylogenetic network construction for complete genomes.

Functional characterization and COG assignment

Functional characterization of all the cyanobacterial genomes was done using the Clusters of Orthologous Groups (COG) database (Tatusov et al. 2003). For each of the genome, all the genes were subjected to COG assignment using Function Profile tool from IMG database (Markowitz et al. 2012). Function Profile tool assists in identification of the genes associated with a particular function in query genome and thus, genes are expected to share at least the same general functions with their COG matches. Once the genes were assigned COGs, they were clustered in 22 functional categories, which were further grouped in four major classes (Table 2). Self-made Perl scripts were used for the grouping of different COG categories.

Repeat identification and analysis

Information about various kinds of microsatellites (Simple Sequence Repeats or SSRs) present in cyanobacterial species was extracted through Imperfect Microsatellite Extractor (IMEx) tool (Mudunuri and Nagarajaram 2007). Parameters taken were, Repeat Type: Perfect; Min. Repeat Number: mono:6, di: 3, tri-hexa:2.

Results and discussion

Habitat and genomic features

Genome size of the cyanobacterial species under study varied from 1.44 Mb (C_UCYN-A) to 9.06 Mb

Table 2 Functional distribution of cyanobacterial genomes under consideration (values represents percentage of the particular COG category identified in the cyanobacteria)

COG Function Class	Metabolism								Cellular Processes and Signalling							Information Storage & Processing					Poorly Categorized	
COG Category	E	G	F	C	H	I	P	Q	D	M	N	O	T	U	V	A	B	J	K	L	S	R
Te_BP_1	9.41	5.34	3.12	6.99	7.25	2.67	6.87	0.45	1.27	5.79	1.21	5.28	2.42	0.89	1.97	0.00	0.06	8.33	4.07	7.76	8.07	10.81
Am_MBIC11017	6.20	4.49	2.17	5.91	4.37	2.84	5.94	1.51	1.45	5.65	1.13	4.69	4.20	1.39	2.55	0.00	0.03	4.63	7.62	9.01	10.66	13.58
Se_PCC6301	9.28	5.43	3.66	7.26	7.08	2.44	7.02	0.61	1.10	5.92	1.16	5.19	2.63	0.98	2.14	0.00	0.12	8.00	4.15	4.52	8.55	12.76
Se_PCC7942	9.39	5.32	3.59	7.18	6.94	2.45	7.36	0.60	1.08	5.92	1.14	5.08	2.57	0.96	2.09	0.00	0.12	7.83	4.25	4.49	8.91	12.74
Sy_PCC6803	7.62	4.84	2.69	6.36	5.15	2.46	6.77	0.85	1.39	6.23	1.43	4.84	3.85	0.94	2.24	0.00	0.04	6.27	4.57	7.39	11.83	12.23
Ma_NIES_843	8.00	4.08	2.30	5.70	4.67	2.42	5.98	1.58	1.39	5.66	1.07	4.48	2.34	0.71	1.74	0.00	0.04	5.50	4.12	11.33	14.57	12.32
Cs_ATCC51142	7.16	5.40	2.36	7.35	4.91	2.92	7.05	1.80	1.20	6.11	1.42	4.05	3.75	1.09	1.87	0.00	0.04	5.47	4.09	5.51	13.08	13.38
Cs_PCC7424	8.02	5.25	2.15	6.11	4.81	2.32	6.28	1.77	0.92	5.42	1.36	4.40	3.96	0.99	1.77	0.03	0.10	5.12	5.25	7.71	12.76	13.51
Cs_PCC7425	7.83	5.37	2.25	5.78	4.71	2.98	6.96	0.90	1.25	6.69	1.11	4.78	5.37	0.66	2.46	0.00	0.07	4.95	5.40	5.78	11.67	13.02
Cs_PCC7822	7.48	5.37	2.21	6.36	4.47	2.84	6.65	1.82	1.34	5.94	1.41	4.19	4.51	0.99	2.11	0.03	0.10	4.76	5.02	6.65	13.20	12.56
Cs_PCC8801	7.33	4.99	2.50	6.65	5.15	2.38	6.50	2.06	1.23	5.94	1.86	4.40	4.00	0.87	1.94	0.00	0.08	5.66	4.44	5.94	14.46	11.64
Cs_PCC8802	7.21	4.96	2.42	6.61	5.20	2.26	6.37	2.05	1.17	5.88	2.05	4.43	3.51	0.89	2.10	0.00	0.08	5.76	4.59	5.48	15.07	11.93
Av_ATCC29413	7.37	4.71	2.10	7.06	4.55	2.63	7.50	2.32	0.97	6.49	0.82	4.64	4.14	0.85	2.76	0.03	0.06	4.92	5.65	5.33	12.42	12.67
Na_0708	7.36	5.25	2.40	7.11	5.49	2.21	5.05	1.18	1.18	6.72	1.03	5.05	3.48	1.08	1.57	0.00	0.05	6.82	4.46	11.53	10.20	10.79
Np_PCC73102	7.39	4.96	1.97	6.40	4.15	3.61	6.01	3.22	1.27	6.38	1.27	4.57	5.05	0.87	2.99	0.03	0.06	4.40	5.61	4.06	12.58	13.15
Ns_PCC7120	6.78	4.41	2.01	5.68	4.29	2.25	7.43	1.86	1.15	6.60	0.89	4.47	5.09	1.01	3.26	0.03	0.06	4.68	6.48	6.19	13.14	12.25
Ap_NIES-39	7.84	5.13	2.42	5.83	4.76	2.42	4.68	0.70	0.82	6.32	1.35	4.97	4.72	0.74	2.09	0.00	0.08	5.79	4.72	7.72	13.96	12.93
Te_IMS101	9.22	5.15	2.58	6.96	5.65	2.49	4.84	1.81	0.86	5.83	1.31	5.33	3.44	1.22	1.81	0.00	0.09	6.19	4.52	7.41	10.76	12.52
Ss_JA_2_3Ba	10.11	4.68	3.00	7.16	6.93	2.02	7.51	0.69	1.33	5.03	0.81	5.66	2.60	0.69	1.27	0.00	0.12	7.34	4.45	9.13	8.26	11.21
Ss_JA-3-3Ab	9.79	4.62	3.00	6.61	7.21	1.92	6.73	0.48	1.14	4.68	0.90	5.59	2.58	0.78	1.32	0.00	0.12	7.81	4.32	11.05	8.17	11.17
Gv_PCC7421	7.80	5.33	2.43	6.44	5.45	3.22	6.84	1.67	1.15	6.88	0.64	4.22	3.42	0.84	2.67	0.00	0.08	5.85	6.48	4.22	12.73	11.65
C_UCYNA	7.36	5.70	2.61	7.48	8.67	3.44	4.63	0.83	1.43	8.08	0.00	6.53	1.31	1.66	1.19	0.00	0.00	13.54	3.92	5.58	6.77	9.26
Pm_AS9601	11.76	5.79	4.27	7.02	8.63	3.51	4.55	0.76	1.42	5.98	0.00	6.07	0.95	1.04	1.52	0.00	0.00	11.57	2.85	5.22	6.45	10.63
Pm_MIT9211	11.51	5.66	4.43	7.08	8.58	3.30	4.62	0.66	1.51	5.85	0.85	6.13	1.04	1.04	1.42	0.00	0.00	11.32	2.92	5.09	6.89	10.09
Pm_MIT9215	12.15	5.23	4.30	7.20	8.60	3.36	4.49	0.84	1.50	6.73	0.00	5.98	0.93	0.93	1.31	0.00	0.00	11.21	2.62	5.23	6.54	10.84
Pm_MIT9301	11.85	5.50	4.45	7.01	8.82	3.41	5.31	0.85	1.33	5.59	0.47	6.45	0.85	0.95	1.14	0.00	0.00	11.37	2.94	4.93	5.97	10.81
Pm_MIT9303	11.25	6.48	3.50	6.78	7.15	3.06	4.84	1.27	1.12	6.63	0.60	5.51	1.42	1.19	1.71	0.00	0.00	9.69	3.58	5.29	8.27	10.66
Pm_MIT9312	11.99	5.52	4.21	7.02	8.71	3.65	4.40	0.84	1.40	6.09	0.37	6.09	0.94	1.12	1.31	0.00	0.00	11.24	3.00	5.06	6.55	10.49
Pm_MIT9313	11.22	6.27	3.64	6.89	7.59	3.10	4.33	1.08	1.24	6.81	0.46	5.34	1.32	0.77	2.24	0.00	0.00	10.22	3.64	5.11	8.36	10.37
Pm_MIT9515	11.68	5.70	4.46	7.12	8.74	3.61	4.75	0.66	1.52	6.65	0.00	5.79	1.04	0.95	1.23	0.00	0.00	11.59	2.37	5.22	6.55	10.35
Pm_NATL1A	11.85	5.46	4.26	6.76	8.80	3.33	5.00	0.65	1.39	6.39	0.19	6.48	1.20	0.93	1.39	0.00	0.00	11.02	3.06	5.28	6.67	9.91
Pm_NATL2A	11.86	5.23	4.30	7.00	8.78	3.27	5.04	0.75	1.49	5.79	0.19	6.54	1.03	0.93	1.49	0.00	0.00	10.92	3.08	5.14	7.00	10.18
Pm_CCMP1375	11.52	6.27	4.21	7.40	8.80	3.09	4.40	0.66	1.40	6.18	0.09	5.71	1.12	1.03	1.69	0.00	0.00	11.05	3.09	5.43	6.84	10.02
Pm_CCMP1986	11.75	5.73	4.32	7.05	8.83	3.38	4.79	0.75	1.41	5.92	0.00	6.39	0.94	0.85	1.22	0.00	0.00	11.09	3.10	5.08	6.77	10.62
Ss_CC9311	10.58	6.61	3.55	7.59	6.96	3.13	5.57	1.25	1.25	6.05	0.56	5.57	1.81	0.63	1.74	0.00	0.07	8.91	3.90	4.45	8.77	11.06
Ss_CC9605	11.08	6.49	3.79	7.22	7.58	2.77	5.76	0.66	1.24	6.56	0.29	5.39	1.31	0.80	1.68	0.00	0.07	9.40	3.13	4.66	8.97	11.15
Ss_CC9902	10.55	6.61	3.72	7.37	7.90	3.04	4.71	0.76	1.06	7.67	0.15	5.85	0.99	0.76	1.59	0.00	0.08	9.79	3.34	5.01	8.43	10.63
Ss_PCC7002	8.51	4.70	3.04	7.07	5.83	2.53	6.91	1.03	1.34	5.68	1.14	5.21	3.87	0.88	2.27	0.00	0.10	7.22	4.80	4.39	11.09	12.38
Ss_RCC307	10.40	6.50	3.47	6.79	7.73	2.96	5.85	1.08	1.30	6.93	0.72	5.70	1.30	0.94	1.66	0.00	0.07	9.17	3.90	4.69	8.16	10.69
Ss_WH7803	10.10	6.71	3.53	6.85	7.70	2.82	5.65	0.92	1.20	6.78	0.42	5.37	1.55	0.92	2.12	0.00	0.07	8.97	3.60	4.73	9.11	10.88
Ss_WH8102	10.34	6.29	3.69	6.94	7.45	3.11	5.57	0.94	1.23	6.80	0.22	5.57	1.52	0.65	1.88	0.00	0.07	9.26	3.40	4.92	8.82	11.35

COG categories: *E* amino acid transport and metabolism, *G* carbohydrate transport and metabolism, *F* nucleotide transport and metabolism, *C* energy production and conversion, *H* coenzyme transport and metabolism, *I* lipid transport and metabolism, *P* inorganic ion transport and metabolism, *Q* secondary metabolites biosynthesis, transport and catabolism, *D* cell cycle control, cell division, chromosome partitioning, *M* cell wall/membrane/envelope biogenesis, *N* cell motility, *O* posttranslational modification, protein turnover, chaperones, *T* signal transduction mechanisms, *U* intracellular trafficking, secretion, and vesicular transport, *V* defense mechanisms, *A* RNA processing and modification, *B* chromatin structure and dynamics, *J* translation, ribosomal structure and biogenesis, *K* Transcription, *L* replication, recombination and repair, *S* function unknown, *R* general function prediction only

(Np_PCC73102) (Table 1). All the organisms contain single circular chromosome as their major genetic material while Cs_ATCC51142 and Av_ATCC29413 have additional chromosome, i.e., linear chromosome and incision element, respectively. For all cyanobacteria, GC content ranged from 30.8% (Pm_MIT9515) to 62% (Gv_PCC7421). Cyanobacteria from marine habitat has small genome size as compared to the members inhabiting freshwater, soil or multiple habitats (Table 1). Larger genomes showed low gene density as compared to smaller ones (Table 1).

16S Rrna-based phylogenetic analysis

We applied 16S rRNA phylogenetic approach to have an insight on the phylogeny of cyanobacterial species. Phylogenetic tree on the basis of 16S rRNA genes divided all cyanobacteria in two branches (Fig S1). The first branch included four species (all the three thermophilic strains (Ss_JA23Ba, Ss_JA33Ab and Te_BP_1) and the only member of Gloeobacterales (Gv_PCC7421 from rock habitat)), second branch comprised rest of the species. In the second branch, all the marine pico-cyanobacteria of the order Prochlorales were grouped with marine and freshwater species of *Synechococcus*. All the members of the order Nostacales shared same branch, joint earlier with a branch having two members of Oscillatoriales. Rest members of the order Chroococcales (*Cyanothece* sps.) clustered on a single branch with members of Nostacales and Oscillatoriales as their nearest neighbor (Fig S1). The tree reflected that cyanobacterial species covering long evolutionary distance together occupy similar habitats and generally possess similar genomic features such as genome size and GC composition as evident from the 16S rRNA-based phylogenetic tree.

Complete genome-based phylogenetic analyses

Genome comparisons suggest that horizontal gene transfer and differential gene loss constitute major evolutionary phenomenon in prokaryotes (Koonin et al. 2001). Whole genome approaches of reconstructing phylogenetic tree have become more apparent due to increasing rate of sequencing projects (Wolf et al. 2002; Delsuc et al. 2005). Considering the entire genome sequences, phylogenetic reconstruction of cyanobacterial species were created using different approaches i.e. genome alignment (Fig S2), alignment-free composition vector approach (Fig S3), overlapping gene content and gene order (Fig S4) and whole-genome protein domain content (Fig S5). Different phylogenomics analyses yielded following results:

Diverse clades for the Order Chroococcales

From the different phylogenomics reconstructions (Fig S1–Fig S5), it is clear that among the order Chroococcales four different clades are identified:

1. Clade of *Cyanothece* sps. with members of Nostacales and Oscillatoriales.
2. Clade of marine *Synechococcus* sps. with *Prochlorococcus* sps.
3. Both the thermophilic *Synechococcus* sps. occupied same lineage and showed similarity with Gv_PCC7421 (Gloeobacterales).
4. Clade occupied rest of the *Chroococcales* species, i.e. Am_MBIC11017, Te_BP_1, C_UCYNA, Se_PCC6301 and Se_PCC7942.

Monophyletic clade for Order Nostacales

In all the phylogenomics reconstructions, members of the Order Nostacales occupied different branches of a single clade along with members of Chroococcales (strains of Cyanothece sps.) and Oscillatoriales as nearest neighbor (Fig S2-S5). In the order Nostacales, Av_ATCC29413 and Ns_PCC7120 cover maximum distance together compared to rest of the two species (Np_PCC73102 and Na_0708) (Fig S2, Fig S4-S5). This is also evident in the phylogenetic reconstruction based on 16S rRNA gene approach (Fig S1). Important to note here is that all the four members of order Nostacales occupied almost similar kind of habitats.

Common clade of marine cyanobacterial species

In all the phylogenetic reconstructions, marine cyanobacterial strains of *Synechococcus* sps. and *Prochlorococcus* sps. occupy the same clade, though they represent different taxonomic orders i.e. Chroococcales and Prochlorales, respectively (Fig S1-S5). Both these groups of marine cyanobacteria show similarity in their genomic features and habitat. Furthermore, it was observed that, among the order Prochlorales, all the high light-adapted strains (Pm_CCMP1986, Pm_MIT9515, Pm_MIT9312, Pm_MIT9301, Pm_AS9601, Pm_MIT9215) formed a single clade thereby, supporting their monophyly origin, whereas six low light-adapted strains (Pm_MIT9211, Pm_MIT9303, Pm_MIT9313, Pm_NATL1A, Pm_NATL2A, Pm_CCMP1375) occupied different branches suggesting parallel evolution (Fig S3-S5). It was also observed that Pm_MIT9303 and Pm_MIT9313 shared branching with Synechococcus sps. rather than Prochlorococcus sps. (Fig S5). Marine cyanobacteria

C_UCYNA also occupied a branch closer to the *Prochlorococcus* strains (Fig S3).

Functional genome profile and its role in adaptation and diversification

We identified functional profile of all 41 cyanobacterial species which provided insights over the functional composition of each genome. 2406 COGs from 22 different categories of four major functional classes ('Metabolism', 'Cellular

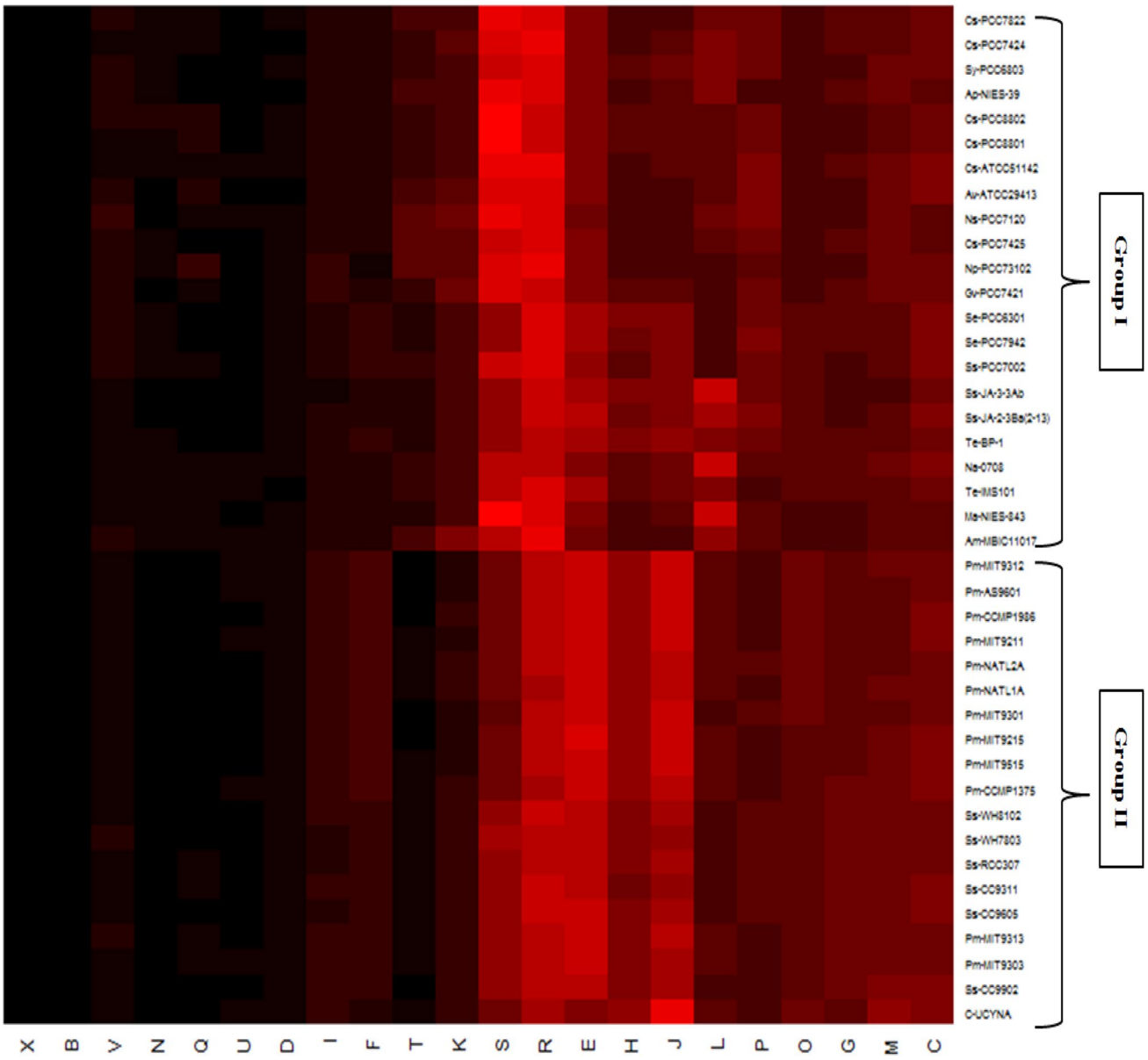

Fig. 1 Heatmap based on the percentage distribution of genes in each functional category for all the cyanobacteria under consideration. Color coding varies from black to red, where black represents lowest value and red represents highest. COG categories: *E* amino acid transport and metabolism, *G* carbohydrate transport and metabolism, *F* nucleotide transport and metabolism, *C* energy production and conversion, *H* coenzyme transport and metabolism, *I* lipid transport and metabolism, *P* inorganic ion transport and metabolism, *Q* secondary metabolites biosynthesis, transport and catabolism, *D* cell cycle control, cell division, chromosome partitioning, *M* cell wall/membrane/envelope biogenesis, *N* cell motility, *O* Posttranslational modification, protein turnover, chaperones, *T* signal transduction mechanisms, *U* intracellular trafficking, secretion, and vesicular transport, *V* defense mechanisms, *A* RNA processing and modification, *B* chromatin structure and dynamics, *J* translation, ribosomal structure and biogenesis, *K* transcription, *L* replication, recombination and repair, *F* function unknown, *R* general function prediction only

processes and signalling', 'Information storage & processing' and 'Poorly categorized') were assigned to these species (Table 2). While analyzing distribution of each functional category, it was observed that across all cyanobacteria, genes with metabolic functions gained maximum share (Fig. 1). Next most abundant functional category in most of the cyanobacteria (specifically those inhabiting freshwater and multiple habitat) was that of poorly categorized genes [Function unknown (S) and General function prediction only (R)]. While marine cyanobacteria preferred genes for 'Information Storage and Processing' over 'Cellular Processes and Signalling', later one is preferred by cyanobacteria from other habitats (Fig. 1). Earlier reports suggested conservation of genes involved in Information processing and signalling in large evolutionary distances (Makarova et al. 1999; Mushegian and Koonin 1996; Azuma and Ota 2009). This may be because they encode for the basic functionalities of the cells (e.g., transcription, translation, repair etc.) and any change in them leads to disruption in normal cellular machinery (Caffrey et al. 2012). In general, habitat seems to influence the functional profile as members from similar habitats possess similar functional profile (Fig. 1). Bacterial genomes contain specific functional gene inventories which are in concurrence with their survival in the particular ecological niche (Allen and Banfield 2005). In the heatmap deduced from the functional profile of cyanobacterial species, cyanobacteria forms two different groups i.e. group I and group II (Fig. 1). Group I majorly includes cyanobacteria from freshwater and other habitats while Group II includes cyanobacteria exclusively from marine habitats. Marine species showed almost different functional genome profile as compared to rest of the cyanobacteria. Most abundant functional category in the cyanobacteria from freshwater and other habitats was Function unknown (S) and General function prediction only (R) (Fig. 1), which possibly reflects that organisms have gained a number of genes, maximum of which remained unknown though they definitely have some important role in their survival.

Influence of GC content over adaptation and diversification

GC content is a well-defined compositional feature of organisms and genomic signature considered to be biased across the tree of life (Nakabachi et al. 2006; Nalbantoglu 2011). GC content is the simplest compositional parameter likely to be affected by the environment or lifestyle of any microbial species and is related to phylogenetic variation (Lawrence and Ochman 1997; Nalbantoglu 2011; Dutta and Paul 2012; Bossert et al. 2017). This feature generally remains constant within a microbial species but becomes variable when it comes across the organisms (Dutta and Paul 2012). In the molecular marker based phylogenetic reconstruction, it is evident that members occupying the same clade possess similar genome size, GC composition and also occupy similar kind of habitat (Fig S1). This fact was also identified in the phylogenomics reconstructions (composition vector approach (Fig S3) and gene content and gene order approach (Fig S4)), where it was reflected that members formed clade with other members of same or different taxonomic orders having similar GC content.

Relation between habitat and different phylogenetic analyses

Habitat of cyanobacteria emerge as a major factor behind their grouping in different phylogenetic reconstructions whether it was phylogenetic analysis or phylogenomics approaches. Even though genome alignment-based phylogenomics analysis showed complicated and varied pattern, habitat was a major influencing factor behind the clustering of cyanobacterial species. Cyanobacteria from similar habitats possess similar kind of genomic features such as genome size, GC content (Table 1), genomic repetitiveness and functional profile (Table 2). Among different phylogenomics reconstructions they occupied the same lineages (Fig S1–S5).

Identification of repeats and their role in ecological adaptation and diversification

Microsatellite are widely used for different studies related to strain typing, genetic mapping, population genetics, phylogenetics, and microevolution analysis (Lim et al. 2004). Microsatellite mining has been carried out in bacterial species *Escherichia coli* (Gur-Arie et al. 2000), *Lactobacillus* (Basharat and Yasmin 2015), *Haemophilus influenzae* (Hood et al. 1996) and others (Mrázek et al. 2007) and in fungal species *S. cerevisiae* (Field and Wills 1998a, b; Kruglyak et al. 2000; Pupko and Graur 1999), *Sphaeropsis sapinea* (Burgess et al. 2001), *Fusarium pseudograminearum* (Scott and Chakraborty 2008) and *Magnaporthe grisea* (Kaye et al. 2003; Lim et al. 2004; Li et al. 2009) etc.

Mono- to hexa-nucleotide repeats were identified for every cyanobacterial species under study. Total 197,750 repeats were identified, ranging from 2269 (Pm_MIT9211) to 12,530 (Te_IMS101) in individual genomes. It was observed that repetitiveness of genome increases with size and smaller genomes tend to have low number of repeats in comparison to larger genome. Furthermore, rather than small-motif repeats (mono- to tetra-nucleotide), penta- and hexa-nucleotide repeats occupied a major proportion of the entire distribution (95%) (Fig. 2). Across all the cyanobacteria, penta-nucleotide repeats occupied 59% of total distribution while hexa-nucleotide repeats represent 36%. Mono- to tetra-nucleotide repeats were present in a very small amount

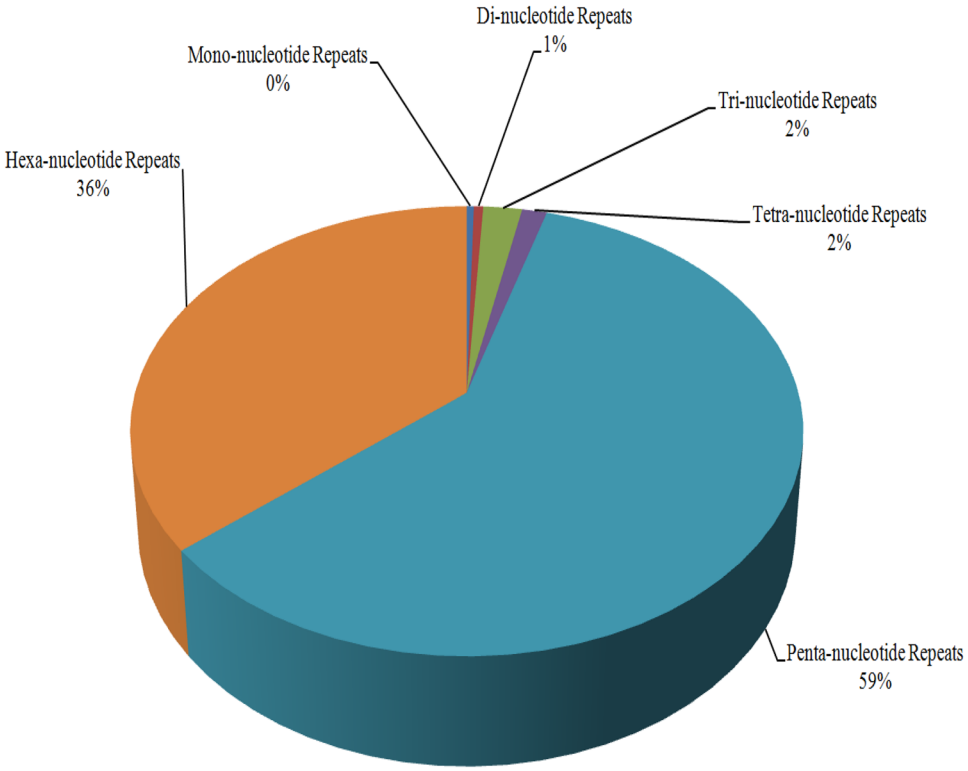

Fig. 2 Distribution of different kind of repeats in cyanobacterial species

in each genome (Fig. 2). Across all the species, strains of *Prochlorococcus marinus* (marine pico-cyanobacteria) possesses highest percentage of mono-, tetra- and penta-nucleotide motifs SSRs on an average but when it comes to hexa-nucleotides, they occupy lowest position. Cyanobacteria with large genome size possess larger motif repeats as compared to those with smaller genome size.

Organisms evolve many mechanisms to cope with environmental situations. Presence of repeats is considered as one possible mechanism towards this adaptation (Treangen et al. 2009; Qin et al. 2014). Repeats affect phenotypic variation either through involvement in the gene expression at the transcriptional level (van Ham et al. 1993; Weiser et al. 1989) or by inducing reversible premature ending of translation when present within coding regions (Bayliss et al. 2001; Henderson et al. 1999; Wang et al. 2000). Our analysis has shown that that marine pico-cyanobacteria with small genome size and inhabiting ecological niche where nutrients are available in plenty have repeats of smaller motifs (mono- to tetra-nucleotide). In contrast, cyanobacteria with large genome size and occupying diverse habitats from freshwater to soil or even rock where nutrition resources are scarce and diverse, possess a large repeat numbers and larger motifs (penta- and hexa-nucleotides). Rather than SSRs with larger motifs, shorter motifs SSRs are reported to be less stable, though the phenomenon behind it is still unclear (Mrazek et al. 2007). This could be the consequence of either diverse evolutionary strategies, recombination approaches or both (or none) (Mrazek et al. 2007). One of the major reasons behind the under-representation of SSR of motifs that are not multiples of three nucleotides in coding sequences lies in the fact that recombination may cause frameshifts leading to gene inactivation (Treangen et al. 2009; Field and Wills 1998a, b; Ackermann and Chao 2006; Qi et al. 2015).

Variation of simple sequence repeats within genes should be very significant for regular gene activity. Expansion or contraction of encoding SSR will straightforwardly influence the corresponding gene products and can even lead to phenotypic changes (Li et al. 2004). In short, repeats facilitates invention of novel functions from pre-existing ones through evolutionary tinkering, though they pose problem in chromosome integrity and organization. Owing to these factors, understanding of the role of repeats in any genomes requires in-depth study about their rate of creation and outcomes from a functional and evolutionary perspective (Treangen et al. 2009). Thus, it can be hypothesized that repeats have played an important role in adaptation of cyanobacteria. They help and assist them in survival in diverse ecological conditions while some cyanobacteria have also possibly evolved them for resistance activities.

Conclusion

Genome comparisons suggest that horizontal gene transfer and differential gene loss constitute major evolutionary phenomenon in prokaryotes. The extant of such events

makes feasibility of 'Tree of Life' reconstruction doubtful as the trees prepared from different genes often state different evolutionary histories. Therefore, alternative approaches to construct tree on the basis of comparisons of complete gene sets or whole genomes can reveal a phylogenetic signal that can support evolutionary history of genomes and suggests the possibility of delineation of undetected clades of organisms. With the increasing rate of sequencing of prokaryotic genomes and the whole genome approaches of reconstructing phylogenetic tree, the concept of universal species tree might be established.

In this study, phylogenetic reconstruction based on the entire genome of different cyanobacteria clearly indicated that clustering of the organisms varied in accordance with their habitats and genome size. Cyanobacteria inhabiting similar habitats tend to have almost similar genome size (and GC-content) and occupied similar lineage during the course of evolution. The study on the evolutionary history of cyanobacterial genomes, even though having several complications, clearly suggested that ecological conditions and the modifications caused within the genomes due to them had great impact on cyanobacterial evolutionary relationships. Habitat also plays an important role in genomic repetitiveness, though, rather than having a direct influence, it majorly affects genome size which eventually is correlated with repeats. Thus, we inferred that maybe large genomes residing in different ecological conditions with scarce and diverse nutritional sources has generated larger repeats (even with larger motifs) which can facilitate development of certain novel function or will play a role in their adaptation. Evolutionary speaking, repeat distribution is a result of selection among different cyanobacterial species and it can be stated that complicated mechanisms are involved in evolution and functioning of repeats. In our study, it was observed that members with different habitats (freshwater, terrestrial or rocks) preferentially accumulate genes for regulation, motility and secondary metabolism in contrast to the genes responsible for informational consequences that are abundant in marine members. A broad metabolic diversity is visible in the large sized cyanobacteria. Furthermore, a large fraction of genes are present in freshwater and terrestrial cyanobacteria, for whom no function is identified till now. The characteristics of gene gain within the genomes can help in understanding the interaction between ecological conditions and genomic evolution. Though, it's clear that micro-evolutionary processes (functional divergence) couples with macro-evolutionary processes (HGT or genome shrinkage) supports for survival and adaptation of cyanobacterial population to diverse ecological niches.

Acknowledgements Authors are grateful to Indian Council of Agricultural Research, New Delhi, India for financial support in the form of "National Agricultural Bioinformatics Grid" (NABG), NAIP. The work is a part of Ph.D. degree program of RP.

Compliance with ethical standards

Conflict of interest Authors declared no conflict of interest.

References

Ackermann M, Chao L (2006) DNA sequences shaped by selection for stability. PLoS Genet 2(2):e22

Adato O, Ninyo N, Gophna U, Snir S (2015) Detecting horizontal gene transfer between closely related taxa. PLoS Comput Biol 11(10):e1004408. https://doi.org/10.1371/journal.pcbi.1004408

Allen EE, Banfield JF (2005) Community genomics in microbial ecology and evolution. Nat Rev Microbiol 3(6):489–498

Anselmetti Y, Duchemin W, Tannier E, Chauve C, Bérard S (2018) Phylogenetic signal from rearrangements in 18 Anopheles species by joint scaffolding extant and ancestral genomes. BMC Genom 19(Suppl 2):96

Auch AF, von Jan M, Klenk H, Göker M (2010) Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genom Sci 2:117–134

Azuma Y, Ota M (2009) An evaluation of minimal cellular functions to sustain a bacterial cell. BMC Syst Biol 3:111

Basharat Z, Yasmin A (2015) Survey of compound microsatellites in multiple Lactobacillus genomes. Can J Microbiol 61(12):898–902. https://doi.org/10.1139/cjm-2015-0136

Bayliss CD, Field D, Moxon ER (2001) The simple sequence contingency loci of *Haemophilus influenzae* and *Neisseria meningitidis*. J Clin Investig 107:657–666

Beck C, Knoop H, Axmann IM, Steuer R (2012) The diversity of cyanobacterial metabolism: genome analysis of multiple phototrophic microorganisms. BMC Genom 13:56

Bentley SD, Parkhill J (2004) Comparative genomic structure of prokaryotes. Annu Rev Genet 38:771–792

Bossert S, Murray EA, Blaimer BB, Danforth BN (2017) The impact of GC bias on phylogenetic accuracy using targeted enrichment phylogenomic data. Mol Phylogenet Evol 111:149–157

Bromberg R, Grishin NV, Otwinowski Z (2016) Phylogeny reconstruction with alignment-free method that corrects for horizontal gene transfer. PLoS Comput Biol 12(6):e1004985. https://doi.org/10.1371/journal.pcbi.1004985

Bryant D, Moulton V (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 21:255–265

Bryant WA, Faruqi AA, Pinney JW (2013) Analysis of metabolic evolution in bacteria using whole-genome metabolic models. J Comput Biol 20:755–764

Burgess T, Wingfield BD, Wingfield MJ (2001) Comparison of genotypic diversity in native and introduced populations of *Sphaeropsis sapinea* isolated from *Pinus radiata*. Mycol Res 105(11):1331–1339

Caffrey BE, Williams TA, Jiang X et al (2012) Proteome-wide analysis of functional divergence in bacteria: exploring a host of ecological adaptations. PLoS One 7(4):e35659

Capella-Gutierrez S, Kauff F, Gabaldón T (2014) A phylogenomics approach for selecting robust sets of phylogenetic markers. Nucleic Acids Res 42:e54

Cassier-Chauvat C, Chauvat F (2018) Cyanobacteria: wonderful microorganisms for basic and applied research. eLS. https://doi.org/10.1002/9780470015902.a0027884

Charlebois RL, Doolittle WF (2004) Computing prokaryotic gene ubiquity: rescuing the core from extinction. Genome Res 14:2469–2477

Chevenet F, Brun C, Bañuls AL, Jacq B, Christen R (2006) TreeDyn: towards dynamic graphics and annotations for analyses of trees. BMC Bioinform 7:439

Colston SM, Fullmer MS, Beka L, Lamy B, Gogarten JP, Graf J (2014) Bioinformatic genome comparisons for taxonomic and phylogenetic assignments using Aeromonas as a test case. MBio 5(6):e02136

Comas I, Moya A, Azad RK, Lawrence JG, Gonzalez-Candelas F (2006) The evolutionary origin of xanthomonadales genomes and the nature of the horizontal gene transfer process. Mol Biol Evol 23(11):2049–2057

Cordero OX, Hogeweg P (2009) The impact of long-distance horizontal gene transfer on prokaryotic genome size. Proc Natl Acad Sci USA 106(51):21748–21753

Daubin V, Szöllősi GJ (2016) Horizontal gene transfer and the history of life. Cold Spring Harb Perspect Biol 8(4):a018036. https://doi.org/10.1101/cshperspect.a018036

Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) 'Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res 30:2478–2483

Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6(5):361–375

Dutilh BE, Snel B, Ettema TJG, Huynen MA (2008) Signature genes as a phylogenomic tool. Mol Biol Evol 25:1659–1667

Dutta C, Paul S (2012) Microbial lifestyle and genome signatures. Curr Genom 13(2):153–162

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

Eisen JA (2000) Assessing evolutionary relationships among microbes from whole-genome analysis. Curr Opin Microbiol 3:475–480

Field D, Wills C (1998a) Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distributions of microsatellites in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces. Proc Natl Acad Sci USA 95(4):1647–1652

Field D, Wills C (1998b) Abundant microsatellite polymorphism in *Saccharomyces cerevisiae* and the different distributions of microsatellites in 8 prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces. P Natl Acad Sci USA 95:1647–1652

Gu X, Huang W, Xu D, Zhang H (2005) GeneContent: software for whole-genome phylogenetic analysis. Bioinformatics 21:1713–1714

Gur-Arie R, Cohen CJ, Eitan Y, Shelef L, Hallerman EM, Kashi Y (2000) Simple sequence repeats in Escherichia coli: abundance, distribution, composition, and polymorphism. Genome Res 10(1):62–71

Hao B, Qi J (2004) Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. J Bioinform Comput Biol 2:1–19

Henderson IR, Owen P, Nataro JP (1999) Molecular switches—the ON and OFF of bacterial phase variation. Mol Microbiol 33(9):19–32

Henz SR, Huson DH, Auch AF, Nieselt-Struwe K, Schuster SC (2005) Whole-genome prokaryotic phylogeny. Bioinformatics 21:2329–2335

Hillis DM, Dixon MT (1991) Ribosomal DNA: molecular evolution and phylogenetic inference. Q Rev Biol 66:411–453

Hood DW, Deadman ME, Jennings MP, Bisercic M, Fleischmann RD, Venter JC et al (1996) DNA repeats identify novel virulence genes in Haemophilus influenzae. Proc Natl Acad Sci USA 93:11121–11125. https://doi.org/10.1073/pnas.93.20.11121

Huson DH (1998) SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 14:68–73

Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci USA 96:3801–3806

Jiang LW, Lin KL, Lu CL (2008) OGtree: a tool for creating genome trees of prokaryotes based on overlapping genes. Nucleic Acids Res 36:W475–W480

Kawai M, Nakao K, Uchiyama I, Kobayashi I (2006) How genomes rearrange: Genome comparison within bacteria Neisseria suggests roles for mobile elements in formation of complex genome polymorphisms. Gene 383:52–63

Kaye C, Milazzo J, Rozenfeld S, Lebrun MH, Tharreau D (2003) The development of simple sequence repeat markers for *Magnaporthe grisea* and their integration into an established genetic linkage map. Fungal Genet Biol 40:207–214

Khiripet N (2005) Bacterial whole genome phylogeny using proteome comparison and optimal reversal distance. In: Computational systems bioinformatics conference, 2005, workshops and poster abstracts, IEEE, pp 63–64

Kim M, Oh HS, Park SC, Chun J (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol 64(Pt 2):346–351

Klenk HP, Göker M (2010) En route to a genome-based classification of Archaea and bacteria? Syst Appl Microbiol 33:175–182

Koehorst JJ, Saccenti E, Schaap PJ, Martins Dos Santos VAP, Suarez-Diez M (2016) Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics. Version 3. F1000Res 5:1987 **(2016 [revised 2017 Jun 27])**

Koksharova O, Wolk C (2002) Genetic tools for cyanobacteria. Appl Microbiol Biotechnol 58(2):123–137

Konstantinidis KT, Tiedje JM (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. Proc Natl Acad Sci USA 101(9):3160–3165

Koonin EV, Makarova KS, Aravind L (2001) Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol 55:709–742

Kruglyak S, Durret R, Shug MC, Aquadro CF (2000) Distribution and abundance of microsatellites in the yeast genome can be explained by a balance slippage between events and point mutations. Mol Biol Evol 17:1210–1219

Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW (2015) Insights from 20 years of bacterial genome sequencing. Funct Integr Genom 15(2):141–161

Lang JM, Darling AE, Eisen JA (2013) phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. PLoS One 8:e62510

Larsson J, Nylander JA, Bergman B (2011) Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. BMC Evol Biol 11:187

Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 44:383–397

Li W, Fang W, Ling L, Wang J, Xuan Z, Chen R (2002) Phylogeny based on whole genome as inferred from complete information set analysis. J Biol Phys 28:439–447

Li Y, Korol AB, Fahima T, Nevo E (2004) Microsatellites within genes: structure, function, and evolution. Mol Biol Evol 21(6):991–1007

Li C, Liu L, Yang J, Li J, Su Y, Zhang Y, Wang Y, Zhu Y (2009) Genome wide analysis of microsatellite sequence in seven filamentous fungi. Interdiscip Sci Comput Life Sci 1:141–150

Lim S, Notley-McRobb L, Lim M, Carter DA (2004) A comparison of the nature and abundance of microsatellites in 14 fungal genomes. Fungal Genet Biol 41:1025–1036

Losos JB, Arnold SJ, Bejerano G, Brodie ED III, Hibbett D, Hoekstra HE, Mindell DP, Monteiro A, Moritz C, Orr HA, Petrov

DA, Renner SS, Ricklefs RE, Soltis PS, Turner TL (2013) Evolutionary biology for the 21st century. PLoS Biol 11:e1001466. https://doi.org/10.1371/journal.pbio.1001466

Ludwig W, Schleifer KH (1999) Phylogeny of bacteria beyond the 16S rRNA standard. ASM News 65:752–757

Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV (1999) Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. Genome Res 9:608–628

Markowitz VM, Chen IM, Palaniappan K et al (2012) IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res 40(Database issue):D115–D122

Medina M (2005) Genomes, phylogeny, and evolutionary systems biology. Proc Natl Acad Sci USA 102:6630–6635

Meier-Kolthoff JP, Auch AF, Klenk H, Göker M (2014) Highly parallelized inference of large genome-based phylogenies. Concurr Comput Pract Exp 26:1715–1729

Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. Trends Genet 17(10):589–596

Mirkin BG, Fenner TI, Galperin MY, Koonin EV (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol 3(2)

Moreno-Hagelsieb G, Wang Z, Walsh S, ElSherbiny A (2013) Phylogenomic clustering for selecting non-redundant genomes for comparative genomics. Bioinformatics 29:947–949

Mrazek J, Guo X, Shah A (2007) Simple sequence repeats in prokaryotic genomes. Proc Natl Acad Sci USA 104:8472–8477

Mrázek J, Guo X, Shah A (2007) Simple sequence repeats in prokaryotic genomes. PNAS 104(20):8472–8477. https://doi.org/10.1073/pnas.0702412104

Mudunuri SB, Nagarajaram HA (2007) IMEx: imperfect microsatellite extractor. Bioinformatics 23(10):1181–1187

Mushegian AR, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci USA 93(19):10268–10273

Nakabachi A, Yamashita A, Toh H et al (2006) The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. Science 314:267

Nalbantoglu OU (2011) Computational genomic signatures and metagenomics. Electrical engineering. Theses and dissertations, p 19

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, White O, Salzberg SL, Smith HO, Venter JC, Fraser CM (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 399:323–329

Nilsson AI, Koskiniemi S, Eriksson S et al (2005) Bacterial genome size reduction by experimental evolution. Proc Natl Acad Sci USA 102(34):12112–12116

Oliveira PH, Touchon M, Cury J, Rocha EPC (2017) The chromosomal organization of horizontal gene transfer in bacteria. Nat Commun 8(1):841

Prabha R, Singh DP, Somvanshi P, Rai A (2016) Functional profiling of cyanobacterial genomes and its role in ecological adaptations. Genom Data 9:89–94

Prasanna AN, Mehra S (2013) Comparative phylogenomics of pathogenic and non-pathogenic *Mycobacterium*. PLoS One 8:e71248

Pupko T, Graur D (1999) Evolution of microsatellites in the yeast *Saccharomyces cerevisiae*: role of length and number of repeated units. J Mol Evol 48:313–316

Qi J, Luo H, Hao B-L (2004) CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res 32:W45–W47

Qi W-H, Jiang X-M, Du L-M, Xiao G-S, Hu T-Z, Yue B-S et al (2015) Genome-wide survey and analysis of microsatellite sequences in bovid species. PLoS One 10(7):e0133667

Qiang, Li, Zhao, Xu, Bailin, Hao (2010) Composition vector approach to whole-genome-based prokaryotic phylogeny: success and foundations. J Biotechnol 149:115–119

Qin L, Zhang Z, Zhao X, Wu X, Chen Y, Tan Z, Li S (2014) Survey and analysis of simple sequence repeats (SSRs) present in the genomes of plant viroids. FEBS Open Bio 4:185–189

Rajendhran J, Gunasekaran P (2011) Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. Microbiol Res 166(2):99–110

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu WT, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T (2013) Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431–437

Rudi K, Sekelja M (2013) High or low correlation between co-occuring gene clusters and 16S rRNA gene phylogeny. FEMS Microbiol Lett 339:23–29

Sawa G, Dicks J, Roberts IN (2003) Current approaches to whole genome phylogenetic analysis. Brief Bioinform 4:63–74

Scott JB, Chakraborty S (2008) Identification of 11 polymorphic simple sequence repeat loci in the phytopathogenic fungus *Fusarium pseudograminearum* as a tool for genetic studies. Mol Ecol Resour 8(3):628–630

Sicheritz-Pontén T, Anderssona SGE (2001) A phylogenomic approach to microbial evolution. Nucleic Acids Res 29:545–552

Sleator RD (2013) A beginner's guide to phylogenetics. Microb Ecol 66:1–4

Snel B, Bork P, Huynen MA (1999) Genome phylogeny based on gene content. Nat Genet 21(1):108–110

Snel B, Bork P, Huynen MA (2002) The identification of functional modules from the genomic association of genes. Proc Natl Acad Sci USA 99(9):5890–5895

Snel B, Huynen MA, Dutilh BE (2005) Genome trees and the nature of genome evolution. Ann Rev Microbiol 59:191–209

Takahashi M, Kryukov K, Saitou N (2009) Estimation of bacterial species phylogeny through oligonucleotide frequency distances. Genomics 93:525–533

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

Tatusov RL, Fedorova ND, Jackson JD et al (2003) The COG database: an updated version includes eukaryotes. BMC Bioinform 4:41

Tekaia F, Lazcano A, Dujon B (1999) The genomic tree as revealed from whole proteome comparisons. Genome Res 9:550–557

Treangen TJ, Abraham AL, Touchon M, Rocha EP (2009) Genesis, effects and fates of repeats in prokaryotic genomes. FEMS Microbiol Rev 33(3):539–571

van Ham SM, van Alphen L, Mooi FR, van Putten JP (1993) Phase variation of *H. influenzae fimbriae*: transcriptional control of two divergent genes through a variable combined promoter region. Cell 73:1187–1196

Větrovský T, Baldrian P (2013) The Variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PLoS One 8:e57923

Vishnoi A, Roy R, Prasad HK, Bhattacharya A (2010) Anchor-based whole genome phylogeny (ABWGP): a tool for inferring evolutionary relationship among closely related microorganims. PLoS One 5:e14159

Wang L, Jiang T, Gusfield D (2000) A more efficient approximation scheme for tree alignment. SIAM J Comput 30(1):283–299

Weisburg WG, Barns SM, Pelletier DA, Lane DJ (1991) 16S ribosomal DNA amplification for phylogenetic study. J Bacteriol 173:697–703

Weiser JN, Love JM, Moxon ER (1989) The molecular mechanism of phase variation of *H. influenzae* lipopolysaccharide. Cell 59:657–665

Wernegreen JJ, Ochman H, Jones IB, Moran NA (2000) 'The decoupling of genome size and sequence divergence in a symbiotic bacterium. J Bacteriol 182:3867–3869

Woese C (1987) Bacterial evolution. Microbiol Rev 51:221–271

Woese C (1998) The universal ancestor. Proc Natl Acad Sci USA 95:6854–6859

Woese C, Fox G (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090

Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol Biol 1:8

Wolf YI, Rogozin IB, Grishin NV, Koonin EV (2002) Genome trees and the tree of life. Trends Genet 18:472–479

Wu D, Jospin G, Eisen JA (2013) Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. PLoS One 8:e77033

Xu Z, Hao B (2009) CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Res 37:W174–W178

Yabuki A, Toyofuku T, Takishita K (2014) Lateral transfer of eukaryotic ribosomal RNA genes: an emerging concern for molecular ecology of microbial eukaryotes. ISME J 8(7):1544–1547

Yang S, Bourne PE (2009) The Evolutionary History of Protein Domains Viewed by Species Phylogeny. PLoS ONE 4(12):e8378. https://doi.org/10.1371/journal.pone.0008378

Zhang YC, Lin K (2015) Phylogeny inference of closely related bacterial genomes: combining the features of both overlapping genes and collinear genomic regions. Evol Bioinform Online 11(Suppl 2):1–9

Zhao Y, Wu J, Yang J et al (2012) PGAP: Pan-genomes analysis pipeline. Bioinformatics 28:416–418

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

Zhou L, Lin Y, Feng B, Zhao J, Tang J (2017) Phylogeny analysis from gene-order data with massive duplications. BMC Genom 18(Suppl 7):760

Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. J Theor Biol 8:357–366

Zuo G, Xu Z, Yu H, Hao B.. Genomics (2010) Jackknife and bootstrap tests of the composition vector trees. Proteom Bioinform 8(4):262–267