Genome-wide investigation and analysis of microsatellites and compound microsatellites in Leptolyngbya species, Cyanobacteria

Dan Yao
Chengdu University
Lian-Ming Du
Chengdu University
Meijin Li
Peking University Shenzhen Graduate School
Maurycy Daroch
Peking University Shenzhen Graduate School
Jie Tang (tangjie@cdu.edu.cn)
Chengdu University https://orcid.org/0000-0002-8973-5200

Short Report

Keywords: Leptolyngbya, microsatellites, compound microsatellites, motif

Posted Date: May 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-211629/v1

Genome-wide investigation and analysis of microsatellites and compound microsatellites in *Leptolyngbya* species, Cyanobacteria

Dan Yao [1], Lianming Du [1], Meijin Li [2], Maurycy Daroch [2], Jie Tang [1,*]

[1] Key Laboratory of Coarse Cereal Processing, Ministry of Agriculture and Rural Affairs, Chengdu University, Chengdu 610106, China

[2] School of Environment and Energy, Peking University Shenzhen Graduate School, Shenzhen 518055, China

* Correspondence: Jie Tang, tangjie@cdu.edu.cn; Tel: 028-84616063; Fax: 028-84616063.

Abstract: Microsatellites (simple sequence repeats, SSRs) are ubiquitously distributed in almost all known genomes. Here, the first investigation was designed to examine the SSRs and compound microsatellites (CSSRs) in 36 genomes of *Leptolyngbya*. The results disclosed diversified patterns of distribution, abundance, density and diversity of SSRs and CSSRs in *Leptolyngbya* genomes. The numbers of SSRs and CSSRs were extremely uneven distributed among genomes, ranging from 11,086 to 27,292 and from 286 to 1,102, respectively. Mononucleotide SSRs were the most abundant category in 14 genomes, while the other 22 genomes followed the pattern: di- > mono- > trinucleotide SSRs. Both SSRs and CSSRs were overwhelmingly distributed in coding regions. The numbers of SSRs and CSSRs were significantly correlated with genome size ($P < 0.01$) and but not closely correlated with GC content ($P > 0.05$). Moreover, the motif $(A/T)_n$ and $(AG)_n$ was predominant in mononucleotide and dinucleotide SSRs, and unique motifs of CSSRs were identified in 33 genomes. This study provides the first insight into SSRs and CSSRs in *Leptolyngbya* genomes and will be useful to contribute to future use as molecular markers in closely-related species.

Keywords: *Leptolyngbya*; microsatellites; compound microsatellites; motif

Introduction

Leptolyngbya that are often found to be prosperous in thermal environments are ecologically important cyanobacteria in light of a crucial role in energy metabolism and matter cycling in ecosystems (Amin et al. 2017). *Leptolyngbya* strains have shown strong biotechnological potential in pharmaceutical (Vijayakumar and Menakha 2015) and biodegradation applications (Ibrahim et al. 2014). Although an increasing number of *Leptolyngbya* strains were proposed, identification of *Leptolyngbya*-like strains has been controversial due to their simple morphology (Bruno et al. 2009). Molecular markers, primarily 16S rRNA and/or 16S-23S intergenic spacer (ITS), alleviate the taxonomic recognition to some extent. However, it was ineffective in dealing with closely related *Leptolyngbya* species (Tang et al. 2018). Therefore, researches beneficial to developing more genetic markers are essential to address such biological question.

Microsatellites, also called simple sequence repeats (SSRs), are tandem repeats with a length of 1 - 6 bp in genomes (Ellegren and Hans 2004). SSRs have been developed as excellent genetic markers because of high abundance, reproducibility and wide genome coverage (Du et al. 2020), and have been employed in many disciplines, such as phylogenetics, evolution and population genetics (Oliveira et al. 2006). SSRs have been found to scatter in both coding and non-coding regions and showed a high mutation rate (10^{-6} to 10^{-2} events per locus per generation) (Bhargava and Fuentes 2010). Additionally, compound microsatellites (CSSRs) consist of two or more SSRs, e.g. $(GCA)_n$-$(C)_n$-$(CA)_n$, and are supposed to possess higher polymorphism than SSRs.

With the development of sequencing technology and *in silico* methodologies, conventional SSR mining based on genomic libraries is being replaced by computational mining from tremendous genome sequences (Evirgh et al. 2019; Wu et al. 2014). These new approaches significantly accelerate

the characterization of SSRs, and understanding of their origin and functions.

To date, there were 36 *Leptolyngbya* genomes available according to the genomic resources of the National Center for Biotechnology Information (NCBI), offering an opportunity of SSR discovery at the genomic level. To our knowledge, a genome-wide survey of SSRs and CSSRs is unavailable for *Leptolyngbya* genomes. The present study was designed to mine and analyze SSRs and CSSRs, and to further reveal the patterns of distribution, abundance, density and diversity of SSRs and CSSRs in *Leptolyngbya* genomes. This study provides the first insight into SSRs and CSSRs in *Leptolyngbya* genomes and may be useful for the future development of molecular markers.

Materials and methods

Genome sequences

According to the genomic resources of NCBI at the time of this study, a total of 36 genomes of *Leptolyngbya* strains were retrieved as data for SSR and CSSR analysis. Information regarding these genomes was summarized in Table 1 and Table S1. In addition, genomic annotations of the 36 *Leptolyngbya* genomes were also downloaded for corresponding analysis.

To illustrate the relationship among the strains studied, multi-locus sequence analysis (MLSA) was performed using concatenated sequences of 15 genes from each genome. These genes were *frr*, *pgk*, *rplA*, *rplB*, *rplC*, *rplE*, *rplK*, *rplL*, *rplM*, *rplN*, *rplP*, *rplT*, *rpmA*, *rpsC*, and *rpsS*. Genes were recommended locus for MLSA by reference (Shih et al. 2013) and selected based on a larger dataset with more genes given to the availability and completeness in genomes. Strains with much less common genes to other genomes were filtered for phylogenetic analysis. Sequences of each gene were aligned, edited and trimmed in Mega7 (Kumar et al. 2016). Sequences were concatenated using BioEdit 7 (http://www.mbio.ncsu.edu/bioedit/bioedit.html). Maximum-Likelihood (ML) phylogenetic

analyses were carried out using PhyML v3.0 (Guindon et al. 2010), and the substitution models were selected by Model Selection function implemented in PhyML (Vincent et al. 2017) under Akaike information criterion (AIC). Nonparametric bootstrap test (1000 replications) was performed to assess the robustness of tree topologies.

Identification and analysis of SSRs and CSSRs

The perfect SSR and CSSRs were identified in each genome using repeat search engine Krait v1.2.2 (Du et al. 2017). In light of small genomes in *Leptolyngbya* strains, the minimum repeats were customized to 6, 3, 3, 3, 3 and 3 for mono-, di-, tri-, tetra-, penta- and hexanucleotide SSRs, respectively (Ledenyova et al. 2019). The maximum distance allowed between any two adjacent SSRs (*d*max) was set to 10 bp for the CSSRs analysis. The other parameters in Krait were maintained as default. All identified perfect SSRs and CSSRs were mapped into coding and non-coding regions to feature coordinates using Krait. The complexity and motifs of CSSRs were investigated as well.

To mitigate the effect of genome size on the comparative analysis, the numbers of SSRs and CSSRs were normalized as relative abundance (RA), the number of SSRs and CSSRs per kb of the genome sequence studied, and relative density (RD), the total length contributed by each SSRs and CSSRs per kb of the genome sequence studied.

Statistical Analysis

To facilitate interpretation, statistical terms used in this study were abbreviated as follows. nSSR: number of SSRs in each genome; nCSSR: number of CSSRs in each genome; cSSR: individual SSR being part of such a CSSR; C: complexity defined by the number of cSSRs in a CSSR; ncSSR: number of cSSR in each genome; and cSSR%: percentage of ncSSR account for nSSR in each genome (cSSR% = ncSSR/ nSSR).

The Pearson correlation coefficient (ρ) was calculated using a custom R script to uncover the associations between variables, including genome size, GC content, nSSR, nCSSR and ncSSR. Significance levels of 0.05 and 0.01 were applied. Significance of CSSR representation in each genome was statistically evaluated by an index, *Z* (Jan 2006). *Z* scores were computed using the following equations:

$$\overline{C} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{ncSSR_i}{nCSSR_i}\right) \quad (1)$$

$$nCSSR_{exp} = \frac{ncSSR_i}{\overline{C}} \quad (2)$$

$$Z = \frac{(nCSSR_{obs} - nCSSR_{exp})}{\sqrt{nCSSR_{exp}}} \quad (3)$$

where n, number of genomes studied (n = 36); i, genome order; $ncSSR_i$, number of cSSR in genome; $nCSSR_i$ (also called $nCSSR_{obs}$), observed number of CSSRs in genome; $\overline{C}$, average of complexity of 36 genomes ($\overline{C}$ = 2.086 in this study); $nCSSR_{exp}$, expected number of CSSRs in genome.

Results

Phylogenetic relationship of *Leptolyngbya* strains

Based on the availability and quality of a single locus from each genome, the concatenated sequences of 15 genes representing 32 *Leptolyngbya* strains and three reference strains were constructed to infer the phylogenetic relationship. The ML tree (Fig. 1) and distance matrix of sequences (Table S2) suggested that considerably genetic divergences existed among those *Leptolyngbya* strains. The *Leptolyngbya* strains did not form a monophyletic clade but mixed with the reference strains, suggesting the interspecific heterogeneity within this genus. This result is not unexpected since it is well-known that *Leptolyngbya* is polyphyletic (Johansen et al. 2011). The tree was rooted by the most distant strains, *L. frigida* UCL18 and *Leptolyngbya* sp. PCC7376. Notably, *Leptolyngbya* sp. PCC7376 was proposed to a new genus as *Enugrolinea bermudensis* (Walter et al. 2017), while *L. frigida* UCL18

appeared to be affiliated with a newly reported genus *Stenomitos* (Shalygin et al. 2020).

Number, relative abundance and density of SSRs and CSSRs

Across the 36 *Leptolyngbya* genomes, a total of 599,033 perfect SSRs were identified (Table 1). Extremely uneven distribution of SSRs number was observed among genomes, ranging from 11,086 to 27,292. The relative abundance (RA) and relative density (RD) both showed significant dissimilarity among *Leptolyngbya* genomes (Table 1), shifting from 2.00 to 3.64/kb and from 13.20 to 24.21 bp/kb, respectively. However, great consistency of RA and RD was noticed within the subgroups (Table 1), e.g. S18-S19, S21-S22, and S27-S30.

There were 21,662 CSSRs identified in the 36 *Leptolyngbya* genomes (Table 1). Similar to SSRs, the number of CSSRs tremendously varied among genomes, from 286 to 1,102. Massive variations were also exhibited by RA and RD (Table 1). The maximum RA (0.18/kb) and RD (2.42 bp/kb) of CSSRs were both identified in *Leptolyngbya* sp. PCC 7376, while the lowest RA (0.05/kb) and RD (0.61 bp/kb) by *Leptolyngbya* sp. NIES-3755. Analogously, strains within each subgroup showed accordant RA and RD of CSSRs.

The number of cSSR in each genome (ncSSR) ranged from 580 to 2,303 (Table 1). And the results suggested that only a small part of all SSRs (less than 11%) in each genome consisted of a compound motif as reveal by cSSR% (Table 1). Within several subgroups, strains exhibited different cSSR%, e.g. S7 (8.85%) and S8 (7.19%), S21 (10.64%) and S22 (8.66%). This result indicated that the proportion of SSRs participating CSSR was inconsistent among strains though similar RA and RD of SSR and CSSR were shared by strains. The significance of CSSR representation, the Z scores, indicated that the $nCSSR_{obs}$ was less than $nCSSR_{exp}$ in 10 genomes, while the opposite results in the remaining 26 genomes. The greatest statistical significance was represented by the genome of *Leptolyngbya* sp.

131 BC1307.

132 **Distribution and diversity of SSRs**

133 As shown in Fig. 2a, mononucleotide, dinucleotide and trinucleotide SSRs accounted for the vast
134 majority of SSRs in each genome, from 98.58% to 99.44%. However, the most abundant category was
135 different among genomes. Mononucleotide SSRs were the most abundant in 14 genomes, accounting
136 for 35.38 to 52.85% of all SSRs, followed by dinucleotide and trinucleotide SSRs, while the other 22
137 genomes followed the pattern: di- > mono- > trinucleotide SSRs. The proportion of tetranucleotide
138 SSRs was more than that of pentanucleotide and hexanucleotide SSRs in each genome. Strains within
139 each subgroup followed the same distribution pattern of SSR type. Overwhelmingly, SSRs were found
140 to be distributed in coding regions of all 36 genomes analyzed (Fig. 2b). And only low percentages of
141 SSRs (15.55 - 28.79%) were located in non-coding regions.

142 A Heatmap (Fig. S1) was constructed to show the relative abundance of 335 standard motifs
143 identified in each genome. There were evident distinctions among genomes regarding motifs in
144 mononucleotide (0.07 - 1.75), dinucleotide (0.01- 1.04) and trinucleotide (0.002 – 0.416) repeat type.
145 However, consistency of relative abundance of standard motifs was observed among phylogenetically
146 closely-related strains, e.g. NIES-2104 and NIES-3755, and JSC-1 and IPPAS B-1204. The motif
147 $(A/T)_n$ was the predominant mononucleotide repeat type in genomes. $(AG)_n$, $(AC)_n$ and $(CG)_n$ were the
148 three most abundant dinucleotide SSRs motifs, among which $(AG)_n$ was particularly dominant. Among
149 the trinucleotide repeat type, (ACG)n and (CCG)n were the most abundant motifs. The relative
150 abundances of motifs in tetranucleotide, pentanucleotide and hexanucleotide repeat type were similar
151 among genomes.

152 **Complexity, motifs and distribution of CSSRs**

153 The complexity of CSSRs in 36 genomes ranged from 2 to 8 (Table S3), except for one CSSR with an
154 extremely high complexity of 28. A vast majority of complexity was 2, accounting for 92.66% of all
155 the CSSRs (Table S3). And the count of CSSRs decreases with the increase of complexity. The highest
156 complexity (C = 28) of single CSSR was observed in the genome of *Leptolyngbya* sp. O-77, with a
157 structure comprising multiple (AG)n, (GAA)n and (T)n. The complexity of CSSRs was different
158 among genomes. For example, the complexity of CSSRs in *Leptolyngbya* sp. BC1307 genome was up
159 to 3, while that in *Leptolyngbya* sp. PCC 7376 and *L. frigida* ULC18 genomes reached to 8. These
160 results suggested the great diversity of motifs among genomes. Moreover, unique motifs were
161 identified in 33 genomes (Table S4), and the number of unique motifs sharply varied among genomes,
162 from 8 (*L. boryana* PCC 6306) to 167 (*L. valderiana* BDU 20041).

163 The distribution of CSSRs, the same as SSRs, were also dominantly in coding regions of all 36
164 genomes analyzed (Fig. 3). The distribution pattern of SSRs and CSSRs obtained in the present study
165 was in accordance with the prevailing results that SSRs and CSSRs in prokaryotic genomes were
166 located more frequently in coding regions than in non-coding regions (Mrázek et al. 2007; Sreenu et al.
167 2003). Interestingly, it was noticed that the percentage of CSSRs in non-coding regions increased with
168 the increase in complexity.

169 **Correlation Analysis**

170 The Pearson correlation analysis (Table 2) showed that the count of SSRs was significantly positively
171 correlated with genome size ($\rho = 0.79$, $P < 0.01$) and but not closely correlated with GC content ($\rho =$
172 -0.22, $P > 0.05$). The number of CSSRs correlated positively with genome size ($\rho = 0.49$, $P < 0.01$),
173 number of SSRs ($\rho = 0.89$, $P < 0.01$) and number of cSSR ($\rho = 1$, $P < 0.01$), but had not significantly
174 correlation with GC content ($\rho = 0.08$, $P > 0.05$). Therefore, the degree of correlation with nCSSR was

ncSSR > nSSR > genome size > GC content.

Discussion

In this study, bioinformatics tools were employed to provide patterns of distribution, abundance, density and diversity of SSRs and CSSRs in 36 *Leptolyngbya* genomes. The results indicated the dissimilarity patterns of SSRs distribution among *Leptolyngbya* genomes (Table 1, Fig. 2), suggesting that SSRs might contribute to the genetic diversity of *Leptolyngbya* genomes. And the highly consistent patterns of SSRs distribution observed in subgroups implied that the dissimilarity patterns of SSRs distribution were probably ascribed to different species. The *Leptolyngbya* genomes differed in the most abundant repeat type, either mononucleotides or dinucleotides. This was in accordance with the prevalence of mononucleotide or dinucleotide repeats in prokaryotic genomes (Du et al. 2020), although sometimes trinucleotide SSRs (e.g. *Cyanobium gracile* PCC 6307) were the most abundant category of SSRs. Mononucleotide repeats were normally characterized as dominant SSRs in eukaryotic genomes, like all human chromosomes (Subbaya et al. 2003). Future studies are required to test the diversification of SSRs distribution in *Leptolyngbya* genomes through comparative genomics.

The smaller motifs were predominant in *Leptolyngbya* genomes (Fig. S1), and the occurrence decreased with the increase of motif length. This trend was shared in a wide range of organisms (Siqueira et al. 2018; Xue et al. 2018). The motif $(A/T)_n$ were the predominant mononucleotide repeat type in *Leptolyngbya* genomes, which was in agreement with the pattern in other cyanobacteria (Du et al. 2020). Among the dinucleotide, SSRs in *Leptolyngbya* genomes, $(AG)_n$ was the predominant motif, while other motifs, like $(AT)_n$, were also predominant in cyanobacteria, e.g. *Calothrix*.

The 36 *Leptolyngbya* genome sizes ranged from 3.9 Mb to 9.4 Mb (Table 1). The correlation analysis suggested a positively correlation between nSSR/nCSSR and genome size (ρ = 0.79/0.49, P <

197 0.01) (Table 2), although in several cases smaller genomes contained more SSRs or CSSRs (Table 1).
198 The GC content of all the *Leptolyngbya* genomes varied from 43.87% to 59.77% (Table 1).
199 Interestingly, GC content had no significant correlation with both nSSR and nCSSR (ρ = -0.22/0.08, P >
200 0.05). Nevertheless, the GC content might have influence on the GC content of SSRs, further affecting
201 the marker developments due to difficult amplification of GC-rich SSRs by PCR. In this study, SSRs of
202 *Leptolyngbya* genomes appeared to be AT-rich (Fig. S1), which might be valuable in the development
203 of SSRs markers.

204 The complexity analysis of CSSRs in the *Leptolyngbya* genomes showed that these CSSRs
205 primarily comprised two SSRs (complexity = 2) (Table S3). As the increase in complexity, the number
206 of CSSRs rapidly decreased. An extraordinary outlier with a complexity of 28 was found in the coding
207 region of *Leptolyngbya* sp. O-77 genome. However, this coding region was annotated as a hypothetical
208 protein. The motifs in CSSRs were quite diverse in each surveyed genome (Table S3), except for those
209 in the genomes of *L. boryana*. In addition, a large number of unique motifs were identified in 33
210 genomes (Table S4). These unique motifs were possibly shaped by two reasons. First, the diverse SSR
211 types in each genome generated various motifs (SSR-couple). Second, mutations within SSRs are
212 reported to be frequent (Xu and Peng 2000). The surveyed *Leptolyngbya* genomes were from diverse
213 niches (Table S1) and easily possessed diversified mutations during evolutionary processes. This
214 hypothesis was verified by the unique motifs obtained in this study that were differentiated from each
215 other by just one or two single mutations.

216 The SSRs and CSSRs identified in this study were predominantly distributed in coding regions of
217 each genome (Fig. 2b, Fig. 3). This result indicated a potential functional role of SSRs and CSSRs in
218 influencing transcription, protein function, gene regulation, and genome organization (Ellegren and

Hans 2004; Oliveira et al. 2006). The different percentages of distribution in coding regions among subgroups or phylotypes may suggest a different level of involvement in functions. However, the possible functions, as well as the mutational mechanism, remain mostly unknown (Bhargava and Fuentes 2010; Li et al. 2010). Moreover, overlapping genes extensively existed in prokaryotic genomes, possibly resulting in more influences caused by SSRs or CSSRs. The exact roles of SSRs or CSSRs need to be carefully investigated in future work.

Variations about genome sizes and distribution patterns of SSRs and CSSRs were evident in the surveyed *Leptolyngbya* genomes. This might be attributed to the fact that *Leptolyngbya* has been recognized as polyphyletic (Johansen et al. 2011), and distinct phylotypes existed in the current datasets (Fig.1 and Table S2). Consistent distribution patterns of SSRs and CSSRs were achieved at the species level, e.g. *L. boryana*, and within subgroups, e.g. *Leptolyngbya antarctica* ULC041bin1 and *Leptolyngbya* sp. ULC073bin1, and *Leptolyngbya* sp. O-77 and PKUAC-SCTA183 that were proposed to classify into one subgroup (Sciuto and Moro 2016; Tang et al. 2018).

According to the public microsatellite database (http://big.cdu.edu.cn/psmd/, updated on July, 2020), SSR number in *Leptolyngbya* genomes (11,086 to 27,292) is comparable to that of other cyanobacterial genomes (2,283 to 53,041). But evident variations of SSR number were observed at the genus level, such as *Thermosynechococcus* (7,490 to 7,724) and *Tolypothrix* (32,706 to 37,800). A similar situation was also noticed in CSSR and cSSR% between *Leptolyngbya* genomes and other cyanobacterial genomes.

Conclusively, a thorough survey was completed to disclose the patterns of distribution, abundance, density and diversity of SSRs and CSSRs in *Leptolyngbya* genomes. This study provides the first insight into SSRs and CSSRs in *Leptolyngbya* genomes and will be useful for the future development

241 of molecular markers in closely-related *Leptolyngbya* species. The mining of SSRs and CSSRs in
242 *Leptolyngbya* group will promote the research on the genomic distribution of SSRs and CSSRs in the
243 cyanobacterial genomes, and further facilitate understanding the origin, evolution and functions of
244 these repeats.

245 **Funding**

246 This research was supported by the National Natural Science Foundation of China (31970092,
247 32071480), and Key Laboratory of Coarse Cereal Processing (Ministry of Agriculture and Rural Affairs,
248 China) (2019CC12).

249 **Conflict of interest**

250 The authors declare that they have no conflict of interest.

251 **Ethical approval**

252 This article does not contain any studies with human participants or animals performed by any of the
253 authors.

254 **Informed consent**

255 All the authors have consent for publication.

256 **Figure legends**

257 **Fig. 1** Maximum likelihood phylogenetic tree of concatenated sequences of 15 genes representing 32
258 *Leptolyngbya* strains

259 **Fig. 2** The SSR distribution patterns in 36 *Leptolyngbya* genomes. **a** distribution of SSR repeat type. **b**
260 SSR distribution in coding and non-coding regions

261 **Fig. 3** The distribution of CSSR in coding and non-coding regions of 36 *Leptolyngbya* genomes

262

References

Amin A et al. (2017) Diversity and distribution of thermophilic Bacteria in hot springs of Pakistan Microb Ecol 74:116-127

Bhargava A, Fuentes FF (2010) Mutational dynamics of microsatellites Mol Biotechnol 44:250-266

Bruno L, Billi D, Bellezza S, Albertano P (2009) Cytomorphological and genetic characterization of troglobitic *Leptolyngbya* strains isolated from Roman hypogea Appl Environ Microbiol 75:608-617

Du L-M, Liu Q, Zhao K-L, Tang J, Fan Z-X (2020) PSMD: An extensive database for pan-species microsatellite investigation and marker development Mol Ecol Resour 20:283-291

Du L, Chi Z, Qin L, Zhang X, Yue B (2017) Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design Bioinformatics 34:681-683

Ellegren, Hans (2004) Microsatellites: simple sequences with complex evolution Nat Rev Genet 5:435-445

Evirgh RK, Hedayat N, Hafezian SH, Farhadi A, Bakhtiarizadeh MR (2019) Genome-wide identification of microsatellites and transposable elements in the dromedary camel genome using whole-genome sequencing data Frontiers in Genetics 10:692

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate Maximum-Likelihood phylogenies: assessing the performance of PhyML 3.0 Syst Biol 59:307-321 doi:10.1093/sysbio/syq010

Ibrahim WM, Karam MA, El-Shahat RM, Adway AA (2014) Biodegradation and utilization of organophosphorus pesticide malathion by cyanobacteria Biomed Res Int 2014:392682

Jan Mz (2006) Analysis of distribution indicates diverse functions of simple sequence repeats in *Mycoplasma* Genomes Mol Biol Evol 23:1370-1385

Johansen JR, Kovacik L, Casamatta DA, Iková KF, Kaštovský J (2011) Utility of 16S-23S ITS sequence and secondary structure for recognition of intrageneric and intergeneric limits within cyanobacterial taxa: *Leptolyngbya corticola* sp. nov. (Pseudanabaenaceae, Cyanobacteria) Nova Hedwigia 92:283-302

290 Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for
291 bigger datasets Mol Biol Evol 33:1870
292 Ledenyova ML, Tkachenko GA, Shpak IM (2019) Imperfect and compound microsatellites in the
293 genomes of *Burkholderia pseudomallei* strains Mol Biol 53:127-137
294 Li Y-C, Korol AB, Fahima T, Beiles A, Nevo E (2010) Microsatellites: genomic distribution, putative
295 functions and mutational mechanisms: a review Mol Ecol 11:2453-2465
296 Mrázek J, Guo X, Shah A (2007) Simple sequence repeats in prokaryotic genomes P Natl Acad Sci USA
297 104:8472-8477
298 Oliveira EJ, Pádua JG, Zucchi MI, Vencovsky R, Vieira MLC (2006) Origin, evolution and genome
299 distribution of microsatellites Genet Mol Biol 29:294-307
300 Sciuto K, Moro I (2016) Detection of the new cosmopolitan genus *Thermoleptolyngbya* (Cyanobacteria,
301 Leptolyngbyaceae) using the 16S rRNA gene and 16S–23S ITS region Mol Phylogenet Evol
302 105:15-35
303 Shalygin S, Shalygina R, Redkina V, Gargas C, Johansen J (2020) Description of *Stenomitos kolaenensis*
304 and *S. hiloensis* sp. nov. (Leptolyngbyaceae, Cyanobacteria) with an emendation of the genus
305 Phytotaxa 440:108-128
306 Shih PM et al. (2013) Improving the coverage of the cyanobacterial phylum using diversity-driven
307 genome sequencing P Natl Acad Sci USA 110:1053-1058
308 Siqueira MVBM, Queiroz-Silva JR, Bressan EA, Aline B, Pereira KJC, Pinto JG, Ann VE (2018)
309 Genetic characterization of cassava (Manihot esculenta) landraces in Brazil assessed with
310 simple sequence repeats Genet Mol Biol 32:104-110
311 Sreenu VB, Vishwanath A, Javaregowda N, Nagarajaram HA (2003) MICdb: database of prokaryotic
312 microsatellites Nucleic Acids Res 31:106-108
313 Subbaya, Subramanian, Rakesh, MishraLalji, Singh (2003) Genome-wide analysis of microsatellite
314 repeats in humans: their abundance and density in specific genomic regions Genome Biol 4:R13
315 Tang J, Jiang D, Luo Y, Liang Y, Li L, Shah MMR, Daroch M (2018) Potential new genera of
316 cyanobacterial strains isolated from thermal springs of western Sichuan, China Algal Res
317 31:14-20

318 Vijayakumar S, Menakha M (2015) Pharmaceutical applications of cyanobacteria - A review Journal of
319 Acute Medicine 5:15-23
320 Vincent L, Jean-Emmanuel L, Olivier G (2017) SMS: Smart Model Selection in PhyML Mol Biol Evol:9
321 Walter JM, Coutinho FH, Dutilh BE, Swings J, Thompson FL, Thompson CC (2017) Ecogenomics and
322 Taxonomy of Cyanobacteria Phylum Front Microbiol 8
323 Wu X, Zhou L, Zhao X, Tan Z (2014) The analysis of microsatellites and compound microsatellites in 56
324 complete genomes of *Herpesvirales* Gene 551:103-109
325 Xu X, Peng M, Z (2000) The direction of microsatellite mutations is dependent upon allele length Nat
326 Genet 24:396-399
327 Xue H et al. (2018) Genome-wide characterization of simple sequence repeats in Pyrus bretschneideri
328 and their application in an analysis of genetic diversity in pear BMC Genomics 19:473
329

Figures

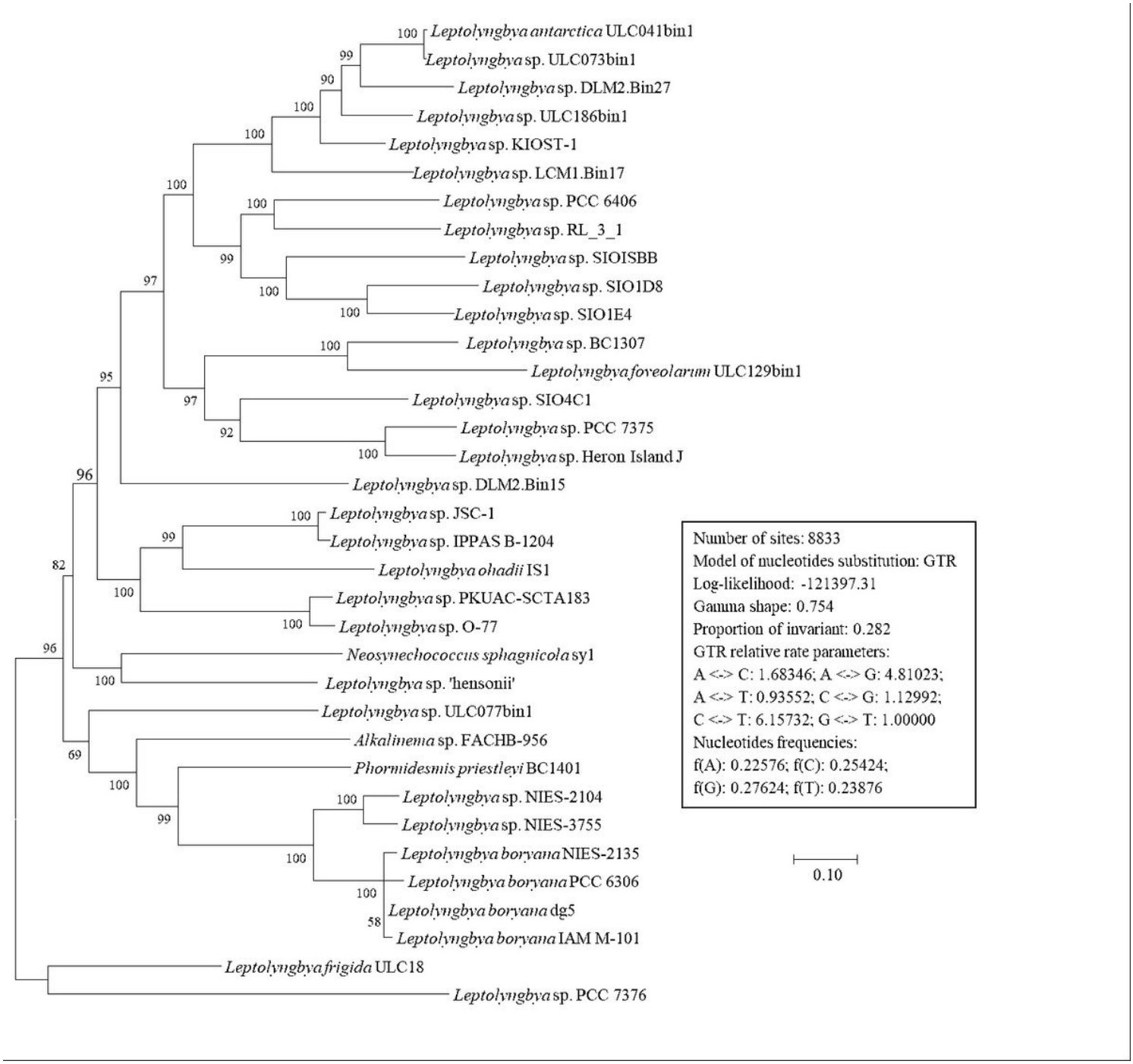

Figure 1

Maximum likelihood phylogenetic tree of concatenated sequences of 15 genes representing 32 Leptolyngbya strains

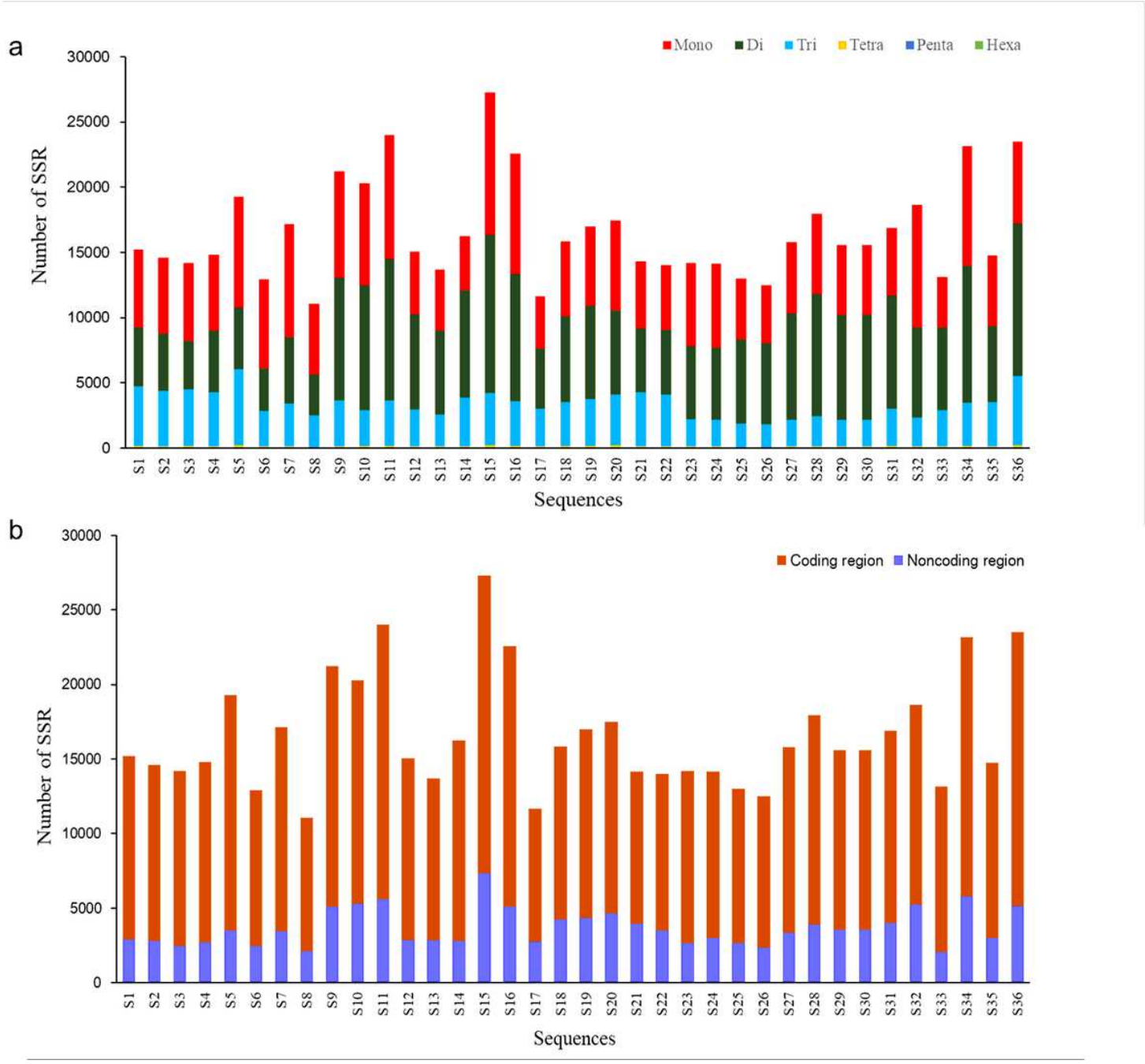

Figure 2

The SSR distribution patterns in 36 Leptolyngbya genomes. a distribution of SSR repeat type. b SSR distribution in coding and non-coding regions

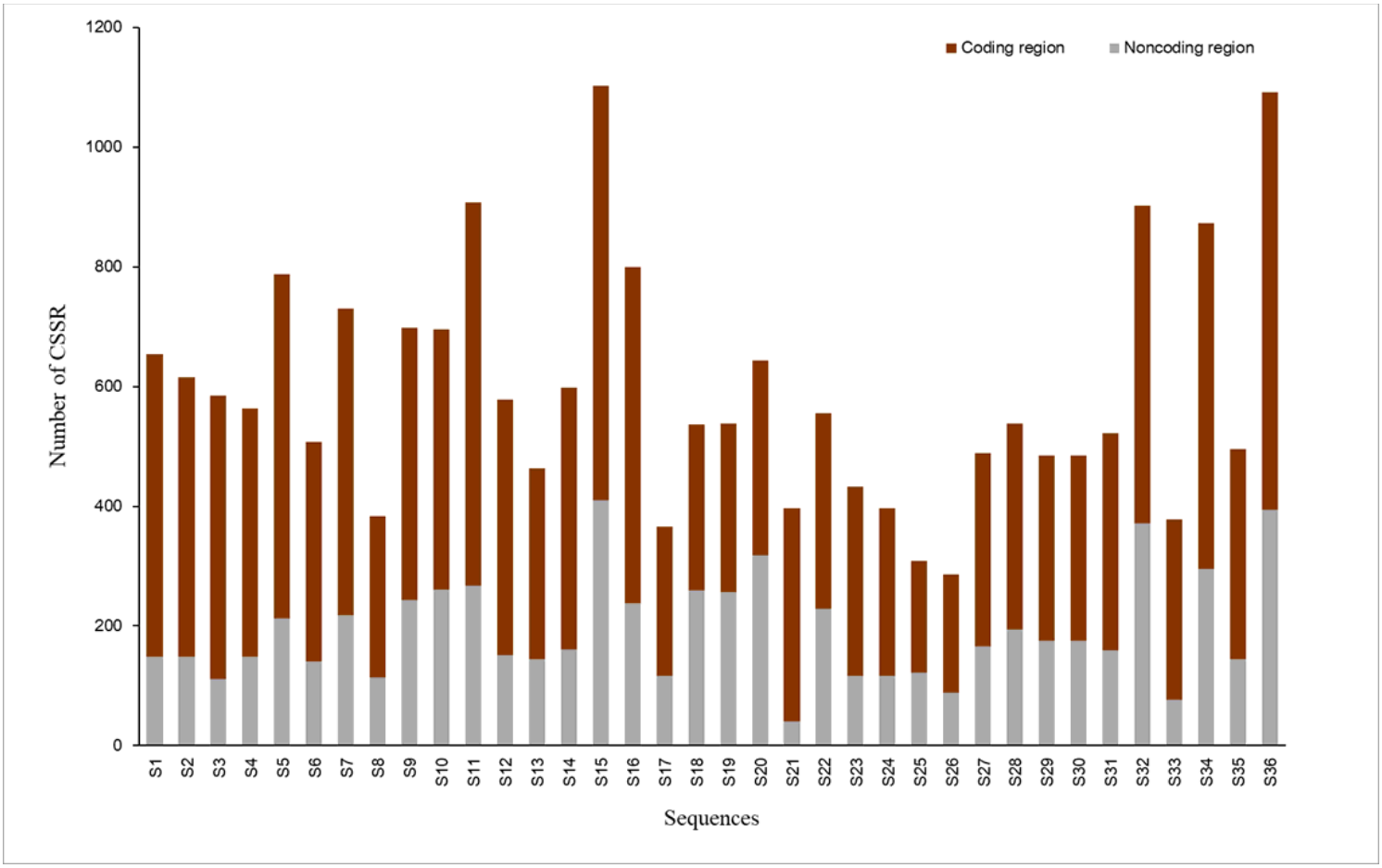

Figure 3

The distribution of CSSR in coding and non-coding regions of 36 Leptolyngbya genomes

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigS1.tiff
- TableS1.xlsx
- TableS2.xls
- TableS3.xlsx
- TableS4.xlsx