Role of Diversity-Generating Retroelements for Regulatory Pathway Tuning in Cyanobacteria

Alec Vallota-Eastman[1], Eleanor C. Arrington[1], Siobhan Meeken[2], Simon Roux[3], Krishna Dasari[4], Sydney Rosen[4], Jeff F. Miller[5,6], David L. Valentine[7,8], Blair G. Paul*[2]

[1]Interdepartmental Graduate Program for Marine Science, University of California Santa Barbara, CA, 93106, USA

[2]Josephine Bay Paul Center, Marine Biological Laboratory, 7 MBL St, Woods Hole, MA 02543, USA

[3] DOE Joint Genome Institute, Berkeley, CA, 94720, USA

[4]Research Mentorship Program (RMP), University of California, Santa Barbara, CA, 93106, USA

[5]Microbiology, Immunology and Molecular Genetics, University of California, Los Angeles, CA, 90095, USA

[6]California NanoSystems Institute, University of California, Los Angeles, CA, 90095, USA

[7]Marine Science Institute, University of California Santa Barbara, CA, 93106, USA

[8]Department of Earth Science, University of California Santa Barbara, CA, 93106, USA

*Corresponding author: Blair G. Paul, bgpaul@mbl.edu

Abstract

Background

Cyanobacteria maintain extensive repertoires of regulatory genes that are vital for adaptation to environmental stress. Some cyanobacterial genomes have been noted to encode diversity-generating retroelements (DGRs), which promote protein hypervariation through localized retrohoming and codon rewriting in target genes. Past research has shown DGRs to mainly diversify proteins involved in cell-cell attachment or viral-host attachment within viral, bacterial, and archaeal lineages. However, these elements may be critical in driving variation for proteins involved in other core cellular processes.

Results

Members of 31 cyanobacterial genera encode at least one DGR, and together, their retroelements form a monophyletic clade of closely-related reverse transcriptases. This class of retroelements diversifies target proteins with unique domain architectures: modular ligand-binding domains often paired with a second domain that is linked to signal response or regulation. Comparative analysis indicates recent intragenomic duplication of DGR targets as paralogs, but also apparent intergenomic exchange of DGR components. The prevalence of DGRs and the paralogs of their targets is disproportionately high among colonial and filamentous strains of cyanobacteria.

Conclusions

We find that colonial and filamentous cyanobacteria have recruited DGRs to optimize a ligand-binding module for apparent function in signal response or regulation. These represent a unique class of hypervariable proteins, which might offer cyanobacteria

a form of plasticity to adapt to environmental stress. This analysis supports the hypothesis that DGR-driven mutation modulates signaling and regulatory networks in cyanobacteria, suggestive of a new framework for the utility of localized genetic hypervariation.

Main Text

Background

Cyanobacteria are a remarkably diverse lineage, in terms of metabolisms, morphologies, and habitat distribution. Perhaps most notably, this phylum contains the only prokaryotic organisms known to have evolved the capability for oxygenic photosynthesis; this trait was later acquired by eukaryotes through endosymbiosis with cyanobacteria, resulting in the formation of chloroplasts [1,2], and driving the modern biosphere. Cyanobacteria have evolved an array of morphologies, including complex multicellular forms [3–6]. Representatives are typically classified into five subsections [7,8]. Species of subsections I and II consist of single coccoid cells. Subsections III-V represent multicellular species that form filaments of varying complexity. Members of subsection III form reversibly-differentiable filaments of vegetative cells. Among subsections IV and V, cells can carry out terminal cellular differentiation in response to environmental stimuli, forming spore-like cells that are resistant to desiccation (akinetes), micro-oxic cells specialized for N_2 fixation (heterocysts), and motile filaments (hormogonia) [9]. This morphological and metabolic complexity has allowed cyanobacteria to inhabit diverse environments.

Certain members of the cyanobacterial phylum possess an extensive capacity to adapt to various environmental pressures through tightly-controlled regulation of complex cellular programs for signal response. This is exemplified by abilities for metabolic switching (i.e. CO_2/N_2 fixation), maintaining photoreceptors of varying intensity for binary programs of circadian rhythm,

64 and forming specialized cells which can sometimes be terminally differentiated and lead to
65 multicellularity [9]. To regulate these complex programs, cyanobacteria have an extensive
66 repertoire of genes governing signal transduction including proteases, kinases, and nucleases.
67 Notably, paralogs of these regulatory proteins are more abundant among the more complex species
68 of cyanobacteria (i.e. those belonging to subsections III-V) [10,11]. However, the mechanisms to
69 diversify and adapt specific functionality in these duplicated genes remain largely unexplored. One
70 mechanism may be the application of diversity-generating retroelements, known to accelerate the
71 evolution of the proteins they target.

72 Diversity-generating retroelements (DGRs) have been noted within the genomes of several
73 genera of cyanobacteria [12–14]. In other bacterial and viral systems, DGRs drive site-specific
74 hypermutation of a subset of codons in target genes. These retroelements utilize a uniquely targeted
75 form of retrotransposition. To this end, DGRs insert variants into a flexible coding scaffold, while
76 avoiding non-specific variation in conserved portions of a gene. The essential features of a DGR
77 are most often found within a single genomic locus spanning ~ 5 – 10 kbp (Fig. 1a), though the
78 synteny and organization of DGR components can vary [13]. Diversification is mechanistically
79 carried out by a reverse transcriptase (RT), which acts upon an RNA transcript, encoded by a
80 template repeat (TR) region in the locus. This region is nearly identical to a variable region (VR)
81 that typically resides in a nearby gene, which encodes a DGR-variable protein (VP). The TR-RNA
82 intermediate is reverse transcribed into cDNA wherein A → N mutation is highly favored by the
83 error-prone RT. This cDNA then replaces VR, whose sequence commonly corresponds to flexible
84 residues in ligand binding structural domains belonging to the C-type lectin or Immunoglobulin
85 (Ig)-like protein families.

86 The first DGR variable protein was characterized from the bacteriophage, BPP-1. In these
87 phage, DGRs diversify tail fiber tip proteins that recognize and bind to *Bordetella* host receptors
88 [12,15]. Other cellular DGRs have been characterized in bacterial pathogens, including *Legionella*
89 *pneumophila* [16] and *Treponema denticola* [17], where DGR target genes encode for cellular
90 surface proteins, presumably involved in cell-cell attachment. The conserved function of cell-cell
91 or viral-cell attachment in these target genes lends to a perspective of DGRs for broad use in host
92 recognition for symbiosis or infection. Moreover, several genera of cyanobacteria were identified
93 in recent genomic and metagenomic surveys of DGRs [13,18]. The essential components of DGRs
94 can be found across most lineages of prokaryotic life [13,18–22], suggesting broad utility of this
95 form of localized mutation.

96 Whereas previously characterized DGR target proteins appear to share a functional role in
97 extracellular attachment to ligands displayed on foreign cells, these retroelements could potentially
98 diversify other cellular proteins with entirely distinct functions. The intermediate RNA, which
99 preserves a template for DGR mutagenesis, has been shown to be highly expressed in lab isolates
100 of *Trichodesmium erythraeum* IMS101 [14] and in *Nodularia spumigena* CCY9414 under light
101 and oxidative stress [23,24]. Here, a systematic analysis of DGRs and their variable proteins in
102 cyanobacterial genomes leads to a new perspective on the utility of diversification and
103 optimization of modular protein domains in paralogs that appear linked to signaling and
104 transcriptional control.

105 Results and discussion

106 A Conserved Subclass of Retroelements in Cyanobacteria

107 Our analysis identified 58 DGRs that include 90 target genes (i.e. encoding VPs) in 52
108 genomes of cyanobacteria spanning 31 different genera. These include filamentous, colonial, and

109 symbiotic organisms (Fig. 1b and Additional file 1: Table S1). Sequence clustering of the 58 DGRs
110 (RT sequences at 95% identity) suggests they represented 49 distinct DGRs, although the full set
111 of 58 were examined further. All DGRs were identified by presence of diagnostic and essential
112 components: an RT gene; one or more VP genes with VR regions; and a TR region. Among the
113 52 genomes analyzed, four contain duplicate DGR cassettes, based on clustering, while one
114 contains two unique DGR-RTs. Moreover, several individual DGRs have multiple target genes,
115 and some VP genes have VRs with homology to other genes dispersed throughout the genome
116 (paralogs) (Fig. 1c).

117 To evaluate the diversity of cyanobacteria-encoded DGRs, we first compared these
118 representatives to a recently developed, global metagenomic DGR dataset [18]. Cyanobacterial
119 DGR-RTs were clustered (i.e. at ≥ 50% AAI) with sequences in the global metagenomic dataset,
120 then linked to a corresponding DGR clade and target protein cluster. All DGRs from our dataset
121 were closely related to DGR Clade-5. The global dataset RTs in DGR Clade-5 are affiliated with
122 target proteins in protein cluster 1 (i.e. PC_00001), which primarily contains cellular proteins that
123 appear to be membrane-bound [18]. Given that the cyanobacterial DGRs appear to cluster tightly
124 together, we next sought to analyze phylogenetic relationships within this set.

125 Phylogenetic analysis of cyanobacterial DGR-RTs revealed a monophyletic clade, unique
126 from all other bacterial DGR-RTs (Fig. 2). The cyanobacterial DGR-RT clade comprises
127 sequences that span nearly all major cyanobacterial genera within morphological subclasses I, III,
128 IV, and V (Fig. 3a). None of the DGR-containing genomes correspond to genera within subclass
129 II. Strikingly, cyanobacterial reverse transcriptases within the monophyletic cyanobacterial DGR
130 clade share an average global sequence identity of 67% (minimum 55%; amino acid sequence).
131 Whereas members of this DGR-RT subgroup do not appear to be shared with other bacteria or

132 archaea, their phylogenetic relationships suggest a complex evolutionary history punctuated by
133 horizontal exchange within the cyanobacterial phylum (Fig. 3b). Although none of the
134 cyanobacterial DGRs could be definitively assigned to prophage elements, they were identified on
135 plasmids of *Anabaena* sp. 90 (CP003287) and *Fischerella* sp. NIES-4106 (AP018301), which may
136 indicate a vehicle for retroelement transfer between closely related populations. Among members
137 of this RT clade, each corresponding DGR-VP contains a ligand-binding C-type lectin-like domain
138 (CLec) with additional functional domains described in detail below.

139 **Intragenomic Dispersal of Hypervariable Binding Domains**

140 DGR variable proteins often contain multiple distinct structural domains [13,21,25]. To
141 investigate the specific functions of cyanobacterial DGR-targeted proteins (i.e. containing the VR
142 scaffold), we first separately analyzed the ligand-binding CLec domains in all DGR-VPs. This
143 approach identified a conserved module (i.e. a putative C-terminal domain) found in each of the
144 52 cyanobacterial VP representatives (Additional file 1: Table S1). The entire set of VR-containing
145 modules share sequence homology with 50.5% average identity and, moreover, all of these protein
146 sequences were clustered together with >30% pairwise amino acid identity. Structural prediction
147 of the representative C-terminal domain sequence (i.e. obtained from clustering) determined that
148 each module most closely resembles the C-type Lectin fold, which is represented by the CLec-like
149 superfamily (InterPro: IPR016187). A search for similar proteins in the Uniprot database identified
150 sequences from an array of other genomes among which 92% belong to cyanobacterial phyla
151 (Additional file 2: Table S2). The similarity between CLec domains found in diverse DGRs may
152 underlie a conserved utility for diversifying this module across different cyanobacterial taxa. The
153 CLec-like superfamily has been linked to a variety of molecular processes in cells and viruses
154 spanning the tree of life, with a common functional role in ligand binding generally predicted for

155 this fold [26–28]. Thus, the modular and dispersed nature of a highly conserved CLec subclass
156 may further point to multifaceted functional significance in cyanobacteria.

157 We next sought to address whether hypervariable CLec modules might arise from gene
158 duplication and intragenomic dispersal, resulting in recognizable sets of paralogs in cyanobacterial
159 genomes. This search was limited to 21 high-quality genomes of the 52-genome total, such that
160 draft genomes composed of >50 scaffolds were removed from the analysis. This approach
161 uncovered 21 genomes that have multiple genes encoding CLec domain-containing proteins, with
162 varying degrees of VR/TR homology (Fig. 4 and Additional file 3: Table S3). These paralogs occur
163 both within DGR loci and dispersed throughout the genome and most often consist of either a
164 single CLec domain or the C-terminal CLec grafted to an N-terminal putative serine kinase
165 domain. Taken together, the multi-genome set of 219 cyanobacterial orthologs across 21 genomes
166 share average pairwise identity of 50.5% within their CLec domains. The complete set of 219
167 orthologs comprises 121 genes that appear to be DGR-diversified based on VR/TR homology,
168 including 45 VP genes encoded within a DGR. The additional 76 remote targets were associated
169 with their respective genome's DGR(s) using a threshold of TR identity greater than 50%; these
170 matches were exclusively found near the 3' terminus of CLec-encoding genes. The proximity to
171 3'-termini suggests that conserved, cis-acting features - such as DNA cruciforms or initiation of
172 mutagenic homing sites required for cDNA integration - may play a role in activating remote
173 targets.

174 The genome of *Nostoc* sp. PCC 7120 (formerly *Anabaena*), contains two DGRs and several
175 dispersed VP paralogs (Fig. 4), providing the opportunity to examine the evolutionary history of
176 these genes in an extensively-studied model organism. Within this genome, we identified three
177 highly similar VP homologs (≥ 60% amino acid identity) in dispersed loci, wherein these genes

178 may have proliferated by duplication and transposition from a common ancestral gene. Notably,
179 one of these paralogs (*all3226*) contains remnant TR-VR homology, despite an absence of
180 proximal RT genes or pseudogenes. Taken together, this suggests a capacity for intragenomic
181 dispersal of DGR-targeted variable proteins, and perhaps removal of diversification components
182 once an optimal variant is selected. In addition to its tractability, the common constellation of DGR
183 VPs that occurs in PCC 7120, as observed in other cyanobacteria, make this species an ideal
184 representative for further analysis of the physiological, ecological and evolutionary ramifications
185 of DGR VP functionality and modularity in cyanobacteria.

186 To assess whether transposable elements were found in proximity to DGRs, we analyzed
187 neighborhoods surrounding each hypervariable protein, including remote VPs with respect to a
188 DGR-RT (i.e. > 5 kbp upstream/downstream). This search uncovered transposase genes belonging
189 to various families in DGR-proximal loci which may be responsible for VP dispersal throughout
190 the genome (Additional file 4: Table S4 and Additional file 5: Table S5). Within the subset of 21
191 high-quality genomes, *Trichodesmium erythraeum* IMS101 has the greatest number of proximal
192 transposase genes, spanning six different insertion sequence (IS) families. The most widely-
193 distributed transposases were those belonging to the IS200/IS605 family, found nearby 9 VPs from
194 6 distinct species. Transposases belonging to this family employ a single-stranded DNA
195 intermediate for a "peel-and-paste" mechanism of transposition [29,30]. The genome of *Anabaena*
196 sp. Strain 90 contains remnants of a putative degraded DGR cassette – containing only the RT with
197 no other detectable features – and notably, the RT gene is flanked by proximal transposase genes.
198 This provides a potential mechanism for select components of the DGR to be mobilized within the
199 genome. DGR recruitment to one gene from another would allow favorably diversified genes to
200 become conserved while targeting hypervariation elsewhere in the genome. Selective pressures

201 can then influence the recruitment of DGRs to genes wherein hypervariation for ligand-binding
202 residues offers selective advantages. Through this mechanism of transposition, cyanobacterial
203 DGRs may provide a newly-diversified, modular, ligand binding domain to signaling genes.

204 **Function of Multidomain Variable Proteins**

205 In part, functional diversity of DGR variable proteins is found in their multidomain
206 complexity. We examined cyanobacterial VPs and their paralogs, which consist of N-terminal
207 domains that are grafted to the C-terminal CLec domain (Fig. 4, Fig. 5). Toward assessing cellular
208 localization, transmembrane and/or signal peptide regions were predicted for 4 DGR-associated
209 VPs and 9 remote VPs, spanning 11 of the 21 high-quality genome set (Additional file 3: Table
210 S3). Most cyanobacterial DGR VPs are predicted to be cytosolic, however evidence exists for TM
211 localization and secretion as well.

212 The most common N-terminal domains of cyanobacterial VPs and their paralogs have
213 similarity to the protein kinase superfamily. These proteins are further predicted to be
214 serine/threonine kinases (STKs) based on the following factors: 1) identification of Hanks and
215 Hunter-type Motifs I through IX [31] (Additional file 6: Figure S1); 2) common NCBI CDS
216 annotations of "serine/threonine protein kinase CDS"; or 3) identification of an STK in previous
217 literature [32]. STKs are mostly associated with eukaryotic signal transduction pathways. In
218 prokaryotes, two-component regulation controls most phosphorylation pathways with a receptor
219 histidine kinase paired with various response regulators phosphorylated on aspartic acid residues.
220 These kinases often control the expression of certain genes [33]. However, Hanks-type STKs have
221 been found in an array of prokaryotic organisms where their genomic abundance is often correlated
222 with genome size, physiological and ecophysiological complexity, and ability to tolerate complex
223 environments [31,32,34]. These STKs are implicated in the regulation of various aspects of

224 bacterial physiology through post-translational modification of proteins, which may themselves be
225 components of phosphorelay and transcriptional regulatory pathways [34–37]. Serine/threonine
226 protein kinases were first associated with the pknA gene of *Nostoc* sp. PCC7120, which is involved
227 in growth and differentiation [32], and in other bacteria their activity regulates processes such as
228 cell growth, segregation, virulence, metabolism, stress adaptation, and cell wall/envelope
229 biogenesis [34]. Ser/Thr kinases in cyanobacteria are usually associated with three different
230 processes: developmental regulation, stress response, and pathogenicity [32]. Slight changes, not
231 in function but in the strength of substrate recognition to a variety of phosphorylation targets, may
232 contribute to the ability to finely tune networks of signal transduction.

233 Compared to histidine kinases of two-component systems, which exhibit strong substrate
234 discrimination, STKs have relaxed substrate specificities. This has been linked to a lack of co-
235 evolution between the kinase and its cognate target [38,39]. Accelerated evolution of the substrate-
236 binding domain of these kinases may have resulted in the further expansion of this class of proteins
237 in the Cyanobacteria phylum, contributing to a wide range of adaptability to external stimuli and
238 challenging environments. We hypothesize that the VR-containing CLec domain could be
239 autoinhibitory, and activation of kinase activity would occur upon binding a small molecule or
240 protein ligand. In this case, DGR-mediated diversity could allow rapid recognition of various
241 ligands for activating phosphorylation cascades. Alternatively, the CLec domain could function in
242 ligand recognition (i.e. determining what protein(s) are phosphorylated). In this case CLec
243 variants could have different substrate specificities. Segregation of phosphorylation targets
244 between paralogous kinases has been shown to play a strong selective pressure in their evolution
245 [40]. In turn, DGR-driven hypervariation of binding components in signaling proteins may offer

246 additional selective advantages in cyanobacteria through preventing cross-talk, which is
247 characteristic of this of kinase class.

248 We also identified orthocaspase-like peptidase domains in VP N-termini, which are also
249 common among their paralogs (Fig. 4). Caspase proteins are proteases involved in the initiation of
250 programmed cell death in metazoans [41]. The peptidase domains that we identified in many VPs
251 and their paralogs were predicted as orthocaspases, which are the prokaryotic homologs of
252 eukaryotic caspase-type proteases [42]. While these protein types are homologous to metazoan
253 caspases, current evidence supports a broader role in cell homeostasis during normal cellular
254 conditions, programs of cellular differentiation, or ageing as well as potential apoptosis [43–45].
255 Previous studies have found orthocaspases to be enriched in morphologically complex filamentous
256 cyanobacteria of subsections III-V (e.g. *Trichodesmium erythraeum* IMS 101, *Anabaena* spp., and
257 *Nostoc* spp.) as well as various strains of the unicellular toxin-producing species, *Microcystis*
258 *aeruginosa*. Conversely, orthocaspases are entirely absent from unicellular genera *Synechococcus*,
259 *Prochlorococcus*, *Cyanobium*, and *Cyanothece* and are underrepresented in the genomes of
260 cyanobacteria belonging to subsections I-II. This suggests their utility in enabling the complex
261 signal response and regulatory programs that exist in cyanobacteria capable of cellular
262 differentiation, toxin production, and diazotrophy [46].

263 In addition to the serine/threonine kinase and orthocaspases-like peptidase domains, we
264 identified less-common features including repeat motifs, Toll-like/TIR receptors, GAF, GUN4-
265 like, and CHAT domains. Repeat motifs may have a role in protein-protein interactions (e.g., TPR,
266 WD40 repeat, ARM repeat, and VWA-CoxE) [47–50], while the other domains have been linked
267 to intracellular signal transduction [51–55] (Fig. 5).

268 A common feature for nearly all of the N-terminal domains, including the prevalent protein
269 kinase-containing paralogs, is their potential to serve a functional role in signal transduction in
270 response to external stimuli (e.g. light, nutrient deprivation, and general stress response) [9]. A
271 previous study found that genes encoding complex multidomain proteins involved in signal
272 transduction are highly enriched in the filamentous cyanobacterium *Anabaena* sp. PCC 7120 when
273 compared to the genomes of unicellular *Synechocystis* sp. PCC 6803 and *Pseudomonas aeruginosa*
274 [56]. Moreover, regulatory proteins involved in signal transduction could lend to the complex
275 regulation necessary for the physiology of filamentous cyanobacteria. These physiologies include
276 a capacity for cell-differentiation, producing heterocysts during nitrogen deprivation and akinetes
277 under environmental stress, as well as programmed apoptosis [43,57]. The presence of DGRs in
278 cyanobacteria follows this trend in the abundance of specialized signal transduction proteins –
279 being seemingly enriched in filamentous nitrogen-fixing taxa and absent from genomes of
280 unicellular taxa, *Synechococcus* spp. and *Prochlorococcus* spp., though they are present in other
281 unicellular species.

282 DGR-programmed variation of the ligand-binding domain of receptor-binding proteins in
283 Bordetella bacteriophage has been shown to increase the capacity of these proteins to recognize a
284 vast array of molecules. Moreover, diversification of oligomeric structures appeared to confer an
285 amplification of binding affinity, or avidity [58,59]. Specifically, the existence of 12 DGR-variable
286 target protein trimers in each bacteriophage virion was shown to increase the binding strengths of
287 these proteins to their ligand, pertactin, by relaxing the requirement for optimal binding between
288 the ligand and any single monomer. This multivalent binding was also shown to lead to more
289 distinction in binding events, contributing to enhanced selectivity [59]. These two properties of
290 avidity through multivalency are hypothesized to be characteristic of other DGR systems as a

291 means to provide ligand-recognition flexibility to evolve under constrained conditions, while
292 maintaining selectivity. We hypothesize that, in cyanobacteria, DGR-programmed variation might
293 have a role in providing multimeric avidity in terms of ligand binding for signal response. In the
294 case of autoinhibitory variable proteins attached to a kinase, rather than providing flexibility in
295 host-receptor binding, as in Bordetella bacteriophage, increased avidity may hold a kinetic
296 advantage for substrate binding, whereby flexible activation accelerates signal transduction and
297 regulation. More generally, the available genomic evidence is consistent with a phenomenon of
298 targeted diversification acting to tune cyanobacterial regulatory networks.

299 Conclusions

300 The DGR-enabled diversification of proteins involved in host attachment should lead to
301 selective advantages, as this offers an offensive selective advantage to variation by the host cell.
302 By genomic inference, other DGR-containing prokaryotes seem to have adopted DGR function to
303 mechanisms of virulence, and other cell-cell or phage/cell binding interactions. By contrast, our
304 findings suggest a selective use of DGRs for purposes of isolated hyper-diversification of a small
305 pocket in the C-terminal binding domains of multidomain proteins broadly involved in signal
306 transduction within cyanobacteria. This class of DGR-target proteins is thus-far unique to the
307 cyanobacterial phylum. Diversification of the binding site of these proteins, paired with natural
308 selection over iterations of diversity generation and the ability to segregate resulting beneficial
309 mutants via transposition, may contribute to the complexity and adaptability of cellular regulation
310 amongst cyanobacterial taxa. In developing a better grasp on the functional significance of DGR
311 hypervariation, it is clear that the phenomenon adds new layers of complexity in the expansion of
312 bacterial protein networks.

313 # Methods

314 ## DGR Identification and Annotation

315 First, we identified all cyanobacterial genomes containing a DGR-RT-like coding sequence
316 by comparing a consensus sequence for previously-identified cyanobacterial DGR-RT sequences
317 against protein databases using pHMMER. All matches were linked to corresponding genome or
318 nucleotide sequences, which were then downloaded from NCBI. A set of potential DGR candidates
319 was first developed using a workflow with Python and Geneious Prime v 2019.2.3 (Biomatters)
320 as previously described [21]. Briefly, RT genes were manually inspected for core NTP-binding
321 site motifs, before searching for near-repeats in a 10-kbp proximal region (i.e., RT +/- 5-kbp).
322 Repeats in this region were then aligned and inspected for: i) random mismatches in one sequence
323 (VR), which predominantly occur in 1st and 2nd codon positions of an ORF, and ii) >80% of
324 mismatches correspond to adenines in the non-coding near-repeat sequence (TR). Next,
325 retroelements were further analyzed using myDGR [60] which is especially effective at identifying
326 putative trans-acting accessory DGR components, and separately, remote VP and VP homolog
327 genes.

328 The entire DGR dataset contains several RT and VP sequences that are near-identical, but
329 shared by distinct genomes (Additional file 1: Table S1). To generate a representative subset of
330 these redundant DGRs, we used CD-HIT [61] to cluster RT amino acid sequences using the
331 following settings: 0.9 global alignment; 95% identity threshold. For comparison with the global
332 metagenomic DGR dataset developed by Roux et al. [18], we conducted pairwise alignments with
333 RT sequences using blastp [62] and identified similar representatives at ≥ 50% amino acid
334 identity.

335 Genes, homologous to VPs within the DGR cassette, were inspected by aligning amino
336 acid sequences for the CLec domain of each putative remote VP to the DGR-VP within the same
337 organism using Clustal Omega. Genes with CLec domains having a putative VR with ≥50%
338 nucleotide identity to a DGR-VP were designated as remote VPs, while those with <50% were
339 designated as VP homologs. DGR and remote VP neighborhood regions were defined as regions
340 10kb upstream and downstream from the DGR cassette, remote VP, or VP homolog.

341 **Neighborhood Analyses**

342 In order to identify potential transposons, we first examined existing genomic annotations
343 in the neighborhood (i.e. +/- 10 kbp) of each VP and Remote VP for the following features:
344 transposase, integrase, mobile element. Next, we conducted a transposon search using ISFinder
345 [63] using expanded VP loci (60 kbp) that contain one or more annotations associated with mobile
346 elements.

347 **Phylogenetic Analyses**

348 To construct a phylogenetic tree of cyanobacteria, we used a set of 16 ribosomal proteins
349 often used for phylogenomic analysis (RpL2, 3, 4, 5, 6, 14, 15, 16, 18, 22, and 24, and RpS3, 8,
350 10, 17, 19 [64]. Each ribosomal protein was identified using HMMER [65] and hidden Markov
351 models from the Pfam [66] database (accessed September 2018). Each individual marker gene
352 was aligned using MUSCLE [67], trimmed using TrimAL [68], manually assessed to remove end
353 gaps and ambiguously aligned regions and concatenated. A maximum likelihood tree was
354 constructed using RAxML v. 8.2.9 [69] with the model

355 To reconstruct RT phylogeny, putative DGR-RT coding sequences were identified, as
356 described above, then translated. Sequences were de-replicated and non-redundant candidates
357 were chosen using CD-Hit [61] with a global alignment threshold of 99% identity. All DGR-RT

358 sequences and a set of Group-II intron RT sequences from Bacteria, Archaea, plastids, and
359 mitochondria were aligned with a hidden markov model of the reverse transcriptase protein family
360 (PF00078) using HMMalign [65]. A phylogenetic tree of DGR-RTs was constructed using
361 FastTree2 [70] with the WAG substitution matrix, and the CAT approximation to optimize branch
362 lengths. The cyanobacterial DGR-RT representatives were extracted from the complete alignment,
363 realigned using Clustal Omega [71] and used to construct an unrooted phylogenetic tree.

364 **Protein Function Analysis**

365 VP domain architecture was annotated using InterProScan, pHMMER, and HMMScan
366 tools. CD-HIT analysis was performed on CLec domains for all VPs using the following settings:
367 0.3 global alignment; 30% identity threshold. Amino acid sequences for the CLec domain of all
368 VPs were aligned using Clustal Omega. The C-terminal sequence of all DGR-VP CLecs was
369 extracted based on the InterProScan feature positions, then further aligned using Clustal Omega
370 and a consensus sequence was picked at 75% sequence similarity (Additional File 6: Figure S1).
371 This consensus sequence was used to further identify homologous domains. Using hmmscan,
372 1,579 hits were returned using an E-value cutoff of 10^{-40} to generate Table S2 (Additional file 2:
373 Table S2).

Figures

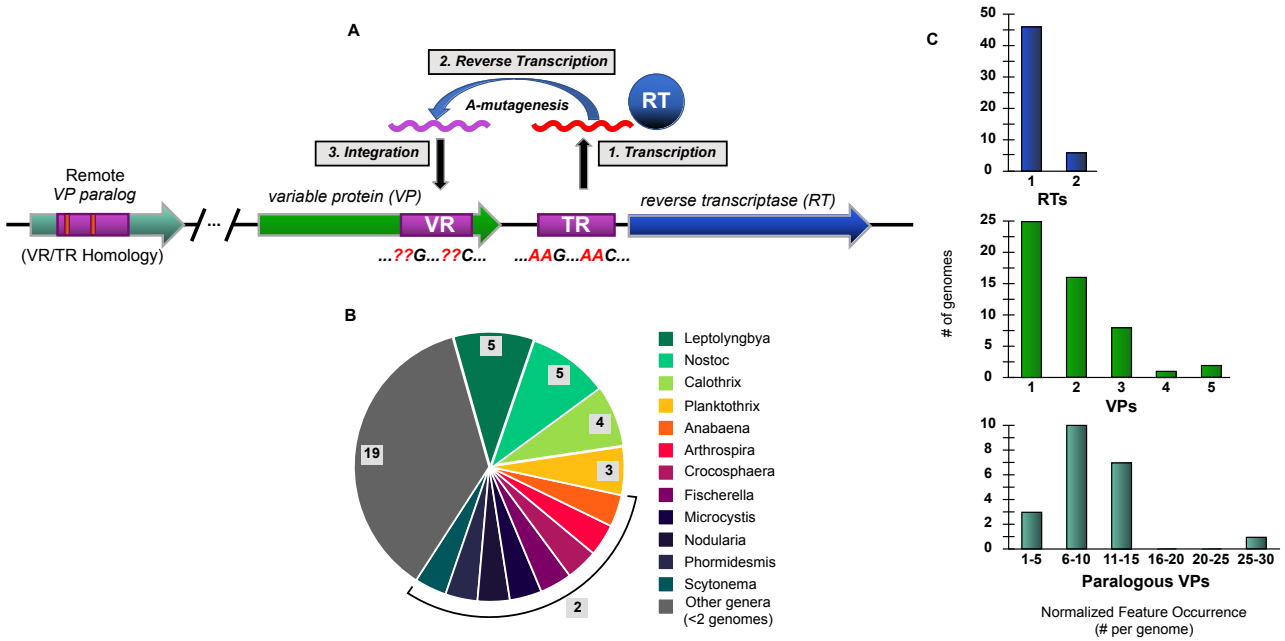

Figure 1. Schematic overview of a DGR and prevalence in cyanobacterial genomes. **a** Three primary steps in the process of mutagenic homing are shown: **1)** conserved template region (TR) in the DGR cassette is transcribed into intermediate, non-coding RNA, which is the substrate for DGR reverse transcriptase (DGR-RT). **2)** Template-primed reverse transcription of TR-RNA is highly error-prone at adenines, which thus incorporates random nucleotides at specific positions in the resulting cDNA. **3)** The new cDNA molecule is integrated into the variable region (VR) at a fixed locus, resulting in the replacement of a portion of the target gene (~100 – 200bp). Genomic surveys suggest that VRs occur almost exclusively near the 3' terminus of a target gene. Additional "remote" VP genes (i.e. paralogs) may be found in non-DGR loci throughout the genome, which have detectable TR vs VR homology. **b** Summary of 52 cyanobacteria genomes known to have DGR components (in Fig. 1a) spanning 31 genera. Genera with ≥ 2 DGR-containing genomes annotated. **c** DGR feature occurrence normalized to genome number.

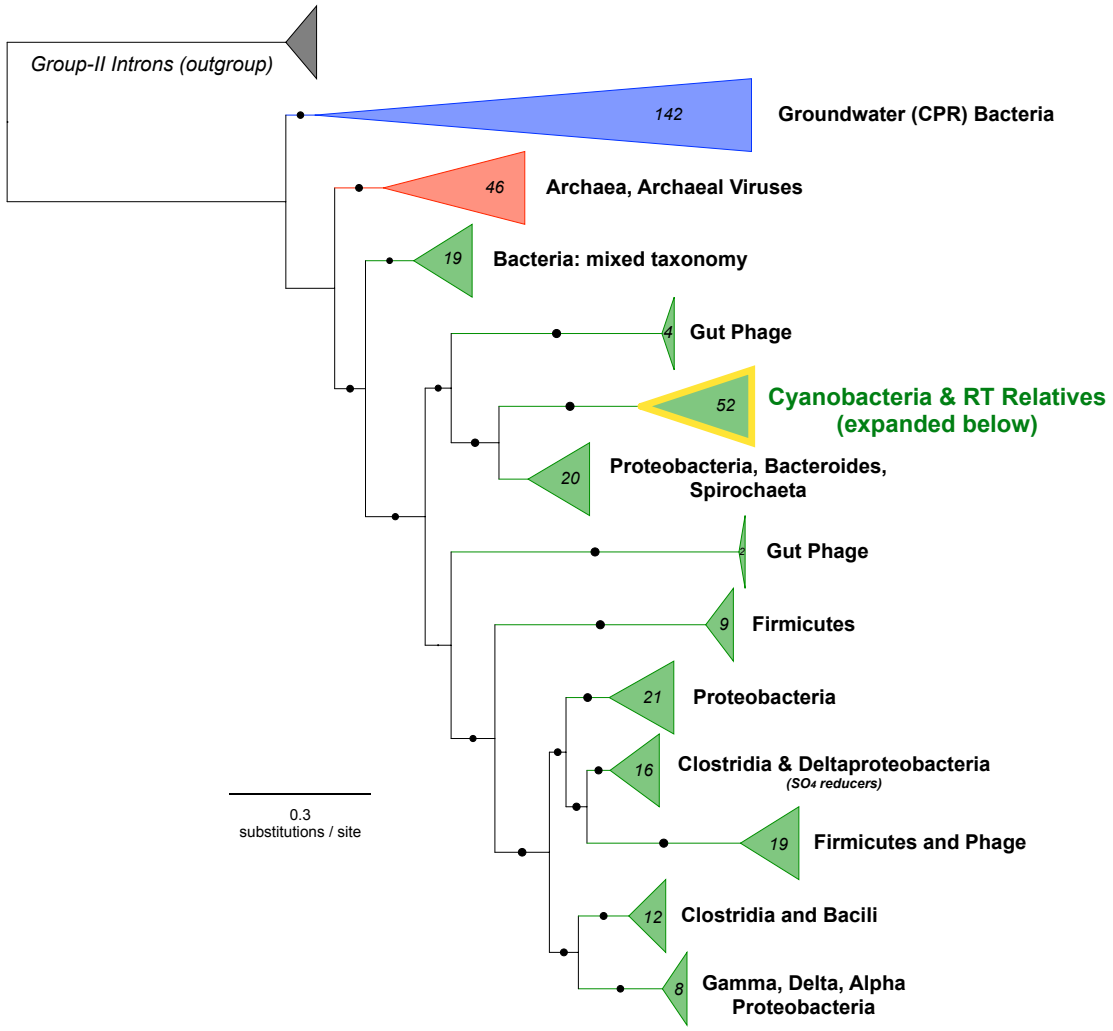

Figure 2. Broad RT phylogeny compared amongst all known DGR-containing lineages. RT phylogeny compared amongst all known DGR-containing lineages with Group-II Introns as the outgroup (highlighted cyanobacterial clade expanded in Figure 3).

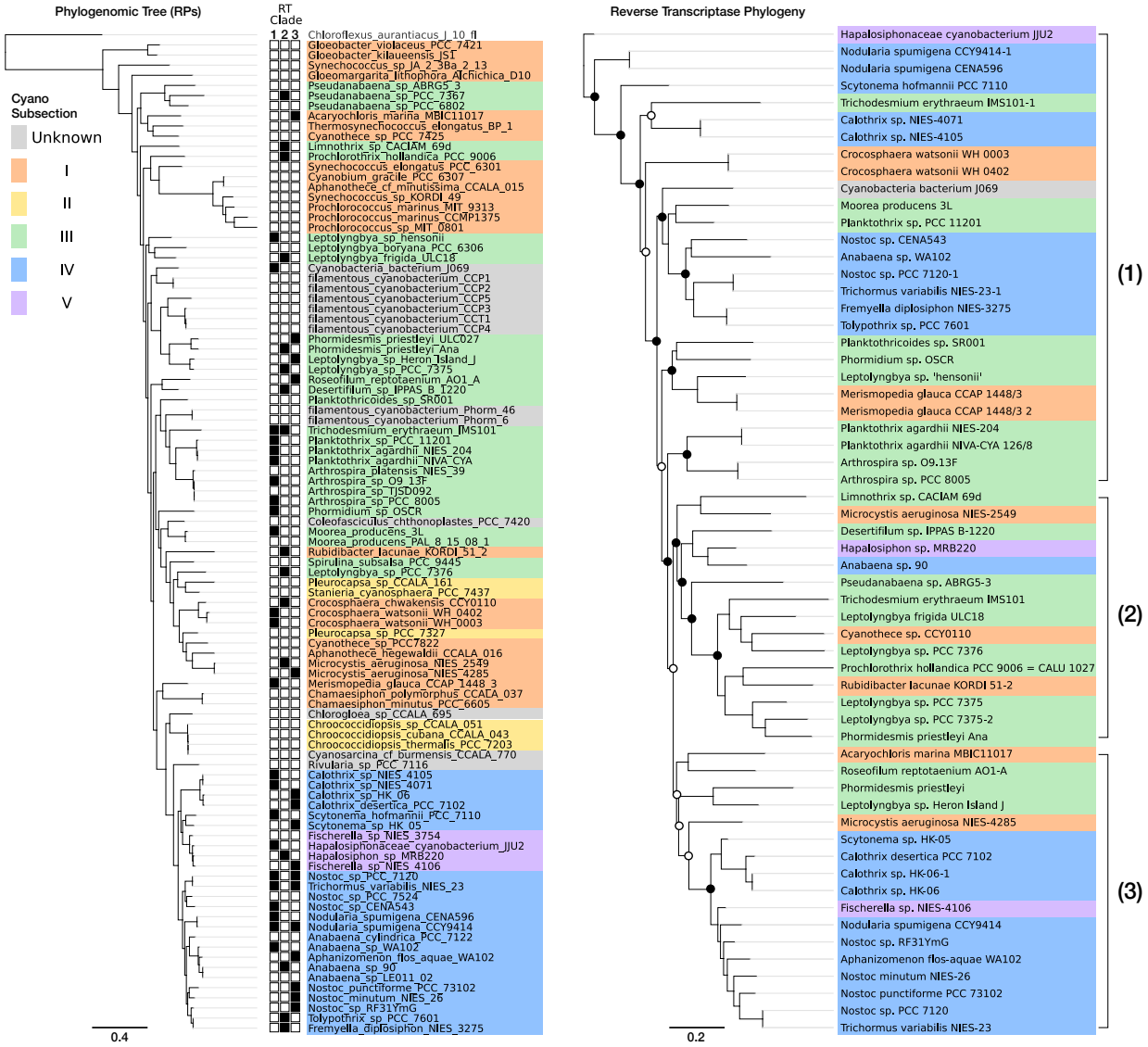

Figure 3. Phylogenetic reconstruction for DGR-containing cyanobacteria and DGR-RT phylogeny.

a Phylogeny for concatenated ribosomal protein alignments, including all DGR-containing species. Filled boxes (left) indicate DGR-RT containing species and the corresponding RT clade. **b** DGR-RT phylogeny with cyanobacterial physiological subsections highlighted in color. Circles indicate branch support values (hollow >50%; filled >70%).

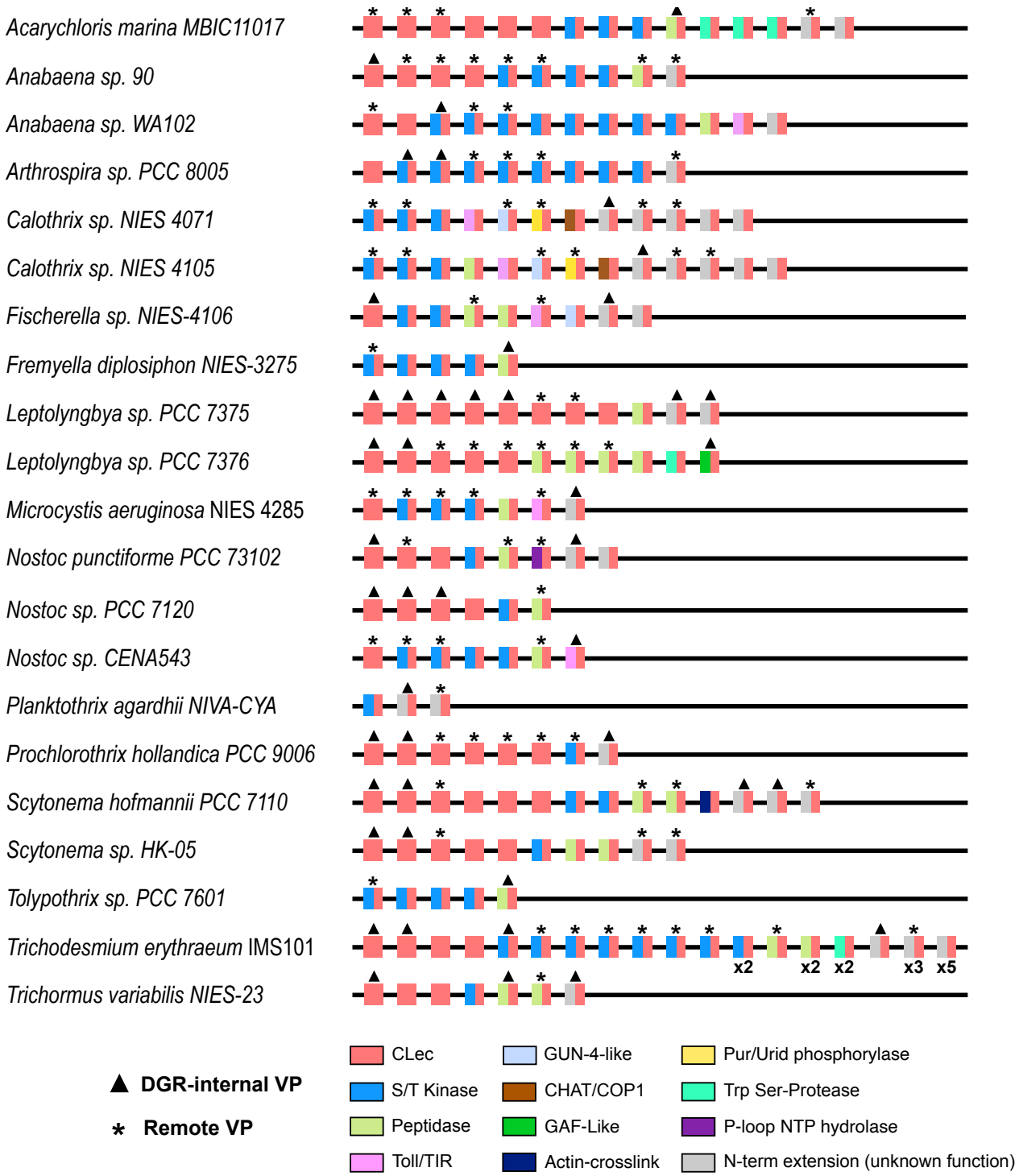

Figure 4. Distribution of CLec variable protein paralogs in cyanobacterial genomes. DGR variable proteins and remote variable proteins are indicated by triangles and asterisks over corresponding representatives. The CLec paralogs of VPs and RVPs are also shown for each genome. The CLec domain found in all paralogs is indicated by either a red square (CLec only) or two grafted domains shown as adjacent rectangles (C-terminal CLec in red; N-terminal domains in various colors). For clarity, paralogs of the same domain architecture in *Trichodesmium erythraeum* IMS101 are indicated below the representative protein (e.g. x2). Note: only a representative subset of DGR-containing genomes is shown (15 of 52 genomes).

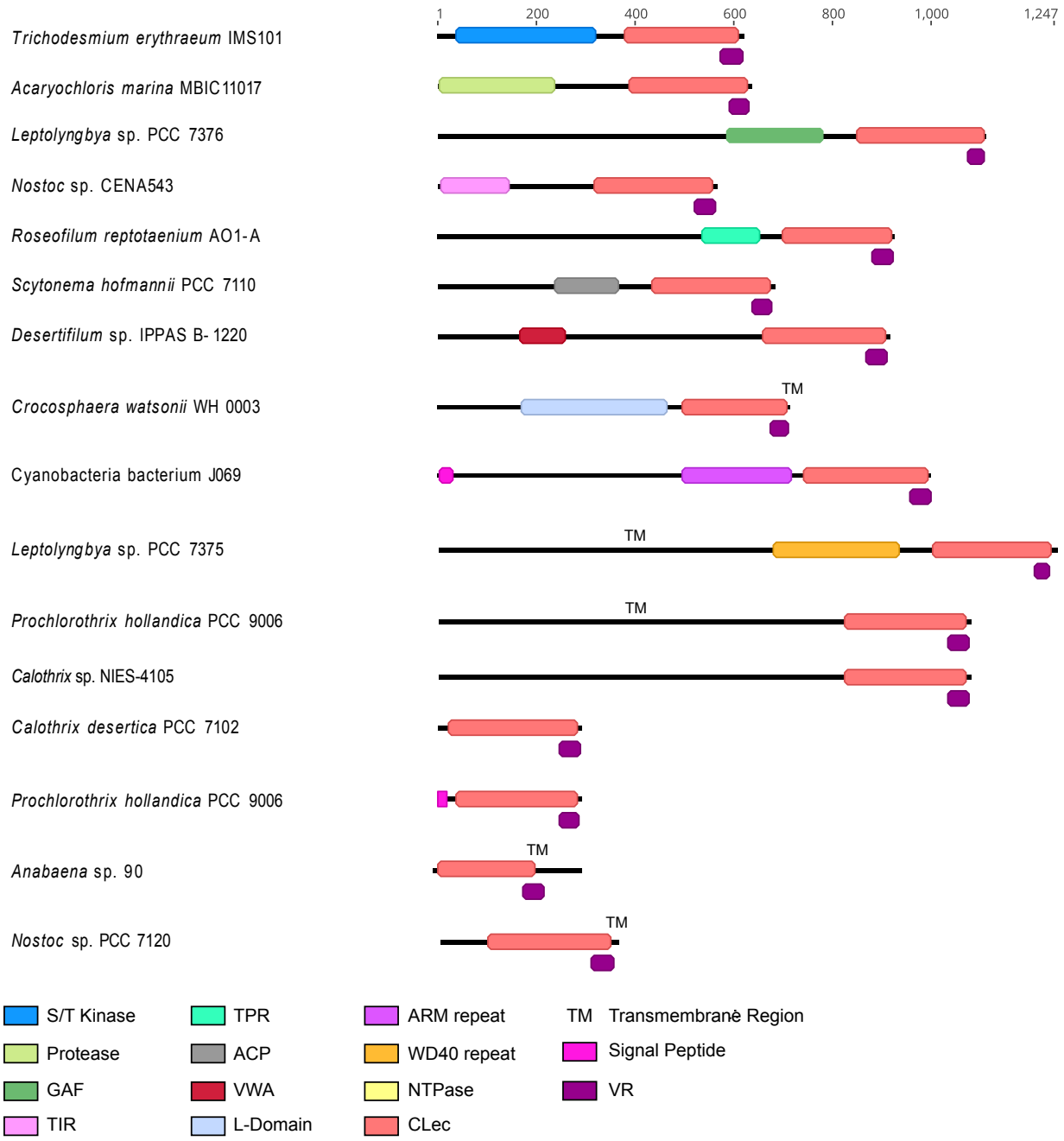

Figure 5. Representative variable protein (VP) domain architectures. Protein domains are colored according to pHMMR domain assignment. Variable regions are shown in deep purple. Additional features, including predicted signal peptides and transmembrane helices, are also indicated. An example species is shown to left of each VP architecture.

Supplement

Additional File 1: Table S1. RT/Species Table.
Summary of all DGR-RTs found in cyanobacterial genomes with taxonomic and known physiological information noted. DGR-internal VPs and Remote VPs identified as having VR/TR homology are also noted. See table S6 for information regarding taxonomic affiliations.

Additional File 2: Table S2. Taxonomy of C-Type Lectin-Like HMM Hits.
The consensus sequence of cyanobacterial C-type lectin-like predicted domains was generated via alignment of DGR-associated C-termini (up to 200 amino acids) and confirmed with pHMM scan. The taxonomy is summarized for those genomes that contain orthologs.

Additional File 3: Table S3. Clec Variable Protein Paralogs.
A subset of 21 high quality genomes were chosen to assess the presence of DGR-VP paralogs. This is a summary of VP paralogs identified and their domains.

Additional File 4: Table S4. DGR-Proximal Transposable Elements.
All transposable elements found in proximity to DGRs are listed, including transposase family and mechanisms of integration.

Additional File 5: Table S5. Genes within DGR Neighborhoods.
Genes found proximal to cyanobacterial DGR cassettes are annotated with predicted function and counts for each annotated function displayed.

Additional File 6: Figure S1. Hanks-type Kinase Motif Characterization in VPs.
Alignment of known "Hanks and Hunter-type" (S/T) kinase domains to the kinase domains of all DGR-VPs, Remote VPs, and VP Paralogs from this dataset. Motifs I-XI highlighted in blue. The top eight sequences denoted "STKII" are known Type II S/T kinases from Zhang et al. 2007 [32].

Additional File 7: Table S6. Updated Cyanobacterial Taxonomy.
Taxonomic assignment for each cyanobacterial genome was generated using relative evolutionary divergence and average nucleotide identity with the Genome Taxonomy Database Toolkit (GTDB-Tk) [72] based on the Genome Taxonomy Database (GTDB) [73].

Acknowledgements:
This research was funded by a Challenge Grant from the California NanoSystems Institute (CNSI-UCSB). AVE is supported by a National Science Foundation Graduate Research Fellowship Program under Grant No. 1650114, and by the NSF California LSAMP Bridge to the Doctorate Fellowship under Grant No. HRD-1701365. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. Computational analyses for this work were supported by an NSF-XSEDE resource allocation DEB170007.

445 **References:**

446 1. Sagan L. On the origin of mitosing cells. J Theor Biol. 1967;14:255–74.

447 2. Giovannoni SJ, Turner S, Olsen GJ, Barns S, Lane DJ, Pace NR. Evolutionary
448 relationships among cyanobacteria and green chloroplasts. J Bacteriol. American
449 Society for Microbiology (ASM); 1988;170:3584–92.

450 3. Flores E, Herrero A. Compartmentalized function through cell differentiation in
451 filamentous cyanobacteria. Nat Rev Microbiol. 2010;8:39–50.

452 4. Mullineaux CW, Mariscal V, Nenninger A, Khanum H, Herrero A, Flores E, et al.
453 Mechanism of intercellular molecular exchange in heterocyst-forming cyanobacteria.
454 EMBO J [Internet]. John Wiley & Sons, Ltd; 2008 [cited 2020 May 25];27:1299–308.
455 Available from: http://emboj.embopress.org/cgi/doi/10.1038/emboj.2008.66

456 5. Flores E, Herrero A, Wolk CP, Maldener I. Is the periplasm continuous in filamentous
457 multicellular cyanobacteria? Trends Microbiol. Elsevier; 2006;14:439–43.

458 6. Giddings TH, Staehelin LA. Observation of microplasmodesmata in both heterocyst-
459 forming and non-heterocyst forming filamentous cyanobacteria by freeze-fracture
460 electron microscopy. Arch Microbiol. Springer-Verlag; 1981;129:295–8.

461 7. Castenholz RW, Wilmotte A, Herdman M, Rippka R, Waterbury JB, Iteman I, et al.
462 Phylum BX. Cyanobacteria. Bergey's Manual® Syst Bacteriol. New York, NY: Springer
463 New York; 2001. p. 473–599.

464 8. Stanier RY, Deruelles J, Rippka R, Herdman M, Waterbury JB. Generic Assignments,
465 Strain Histories and Properties of Pure Cultures of Cyanobacteria. Microbiology.
466 Microbiology Society; 1979;111:1–61.

467 9. Sarma TA. Handbook of Cyanobacteria. CRC Press, Taylor and Francis; 2012.

468 10. López-Igual R, Wilson A, Leverenz RL, Melnicki MR, de Carbon CB, Sutter M, et al.
469 Different functions of the paralogs to the N-terminal domain of the orange carotenoid
470 protein in the cyanobacterium Anabaena sp. PCC 7120. Plant Physiol. American
471 Society of Plant Biologists; 2016;171:1852–66.

472 11. Asselman J, Pfrender ME, Lopez JA, De Coninck DIM, Janssen CR, Shaw JR, et al.
473 Conserved transcriptional responses to cyanobacterial stressors are mediated by
474 alternate regulation of paralogous genes in Daphnia. Mol Ecol. Blackwell Publishing Ltd;
475 2015;24:1844–55.

476 12. Doulatov S, Hodes A, Dai L, Mandhana N, Liu M, Deora R, et al. Tropism switching
477 in Bordetella bacteriophage defines a family of diversity-generating retroelements.
478 Nature. Nature Publishing Group; 2004;431:476–81.

479 13. Wu L, Gingery M, Abebe M, Arambula D, Czornyj E, Handa S, et al. Diversity-
480 generating retroelements: natural variation, classification and evolution inferred from a

481 large-scale genomic survey. Nucleic Acids Res. Narnia; 2018;46:11–24.

482 14. Pfreundt U, Kopf M, Belkin N, Berman-Frank I, Hess WR. The primary transcriptome
483 of the marine diazotroph Trichodesmium erythraeum IMS101. Sci Rep. Nature
484 Publishing Group; 2015;4:6187.

485 15. Liu M, Deora R, Doulatov SR, Gingery M, Eiserling FA, Preston A, et al. Reverse
486 transcriptase-mediated tropism switching in Bordetella bacteriophage. Science (80-).
487 American Association for the Advancement of Science; 2002;295:2091–4.

488 16. Arambula D, Wong W, Medhekar BA, Guo H, Gingery M, Czornyj E, et al. Surface
489 display of a massively variable lipoprotein by a Legionella diversity-generating
490 retroelement. Proc Natl Acad Sci. 2013;110:8212 LP – 8217.

491 17. Le Coq J, Ghosh P. Conservation of the C-type lectin fold for massive sequence
492 variation in a Treponema diversity-generating retroelement. Proc Natl Acad Sci U S A.
493 National Academy of Sciences; 2011;108:14649–53.

494 18. Roux S, Paul BG, Bagby SC, Allen MA, Attwood G, Cavicchioli R, et al. Ecology and
495 molecular targets of hypermutation in the global microbiome. [cited 2020 May 25];
496 Available from: https://doi.org/10.1101/2020.04.01.020958

497 19. Schillinger T, Zingler N. The low incidence of diversity-generating retroelements in
498 sequenced genomes. Mob Genet Elements. Informa UK Limited; 2012;2:287–91.

499 20. Yan F, Yu X, Duan Z, Lu J, Jia B, Qiao Y, et al. Discovery and characterization of
500 the evolution, variation and functions of diversity-generating retroelements using
501 thousands of genomes and metagenomes. BMC Genomics. 2019;20:595.

502 21. Paul BG, Burstein D, Castelle CJ, Handa S, Arambula D, Czornyj E, et al.
503 Retroelement-guided protein diversification abounds in vast lineages of Bacteria and
504 Archaea. Nat Microbiol. Nature Publishing Group; 2017;2.

505 22. Paul BG, Bagby SC, Czornyj E, Arambula D, Handa S, Sczyrba A, et al. Targeted
506 diversity generation by intraterrestrial archaea and archaeal viruses. Nat Commun.
507 Nature Publishing Group; 2015;6:1–8.

508 23. Kopf M, Möke F, Bauwe H, Hess WR, Hagemann M. Expression profiling of the
509 bloom-forming cyanobacterium Nodularia CCY9414 under light and oxidative stress
510 conditions. ISME J. Nature Publishing Group; 2015;9:2139–52.

511 24. Voß B, Bolhuis H, Fewer DP, Kopf M, Möke F, Haas F, et al. Insights into the
512 Physiology and Ecology of the Brackish-Water-Adapted Cyanobacterium Nodularia
513 spumigena CCY9414 Based on a Genome-Transcriptome Analysis. Janssen PJ, editor.
514 PLoS One. Public Library of Science; 2013;8:e60224.

515 25. Guo H, Arambula D, Ghosh P, Miller JF. Diversity-generating Retroelements in
516 Phage and Bacterial Genomes. Microbiol Spectr. NIH Public Access; 2014;2.

517 26. Hoving JC, Wilson GJ, Brown GD. Signalling C-type lectin receptors, microbial
518 recognition and immunity. Cell. Microbiol. 2014. p. 185–94.

519 27. del Fresno C, Iborra S, Saz-Leal P, Martínez-López M, Sancho D. Flexible Signaling
520 of Myeloid C-Type Lectin Receptors in Immunity and Inflammation. Front Immunol.
521 Frontiers; 2018;9:804.

522 28. Zelensky AN, Gready JE. The C-type lectin-like domain superfamily. FEBS J.
523 2005;272:6179–217.

524 29. Barabas O, Ronning DR, Guynet C, Hickman AB, Ton-Hoang B, Chandler M, et al.
525 Mechanism of IS200/IS605 Family DNA Transposases: Activation and Transposon-
526 Directed Target Site Selection. Cell. Elsevier; 2008;132:208–20.

527 30. He S, Corneloup A, Guynet C, Lavatine L, Caumont-Sarcos A, Siguier P, et al. The
528 IS200/IS605 Family and “Peel and Paste” Single-strand Transposition Mechanism.
529 Microbiol Spectr. American Society for Microbiology; 2015;3.

530 31. Hanks SK, Hunter T. Protein kinases 6. The eukaryotic protein kinase superfamily:
531 kinase (catalytic) domain structure and classification. FASEB J. FASEB; 1995;9:576–
532 96.

533 32. Zhang X, Zhao F, Guan X, Yang Y, Liang C, Qin S. Genome-wide survey of putative
534 Serine/Threonine protein kinases in cyanobacteria. BMC Genomics. BioMed Central;
535 2007;8:395.

536 33. Stock AM, Robinson VL, Goudreau PN. Two-Component Signal Transduction. Annu
537 Rev Biochem. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA
538 94303-0139, USA; 2000;69:183–215.

539 34. Janczarek M, Vinardell J-M, Lipa P, Karaś M. Hanks-Type Serine/Threonine Protein
540 Kinases and Phosphatases in Bacteria: Roles in Signaling and Adaptation to Various
541 Environments. Int J Mol Sci [Internet]. MDPI AG; 2018 [cited 2020 Mar 22];19:2872.
542 Available from: http://www.mdpi.com/1422-0067/19/10/2872

543 35. Libby EA, Goss LA, Dworkin J. The Eukaryotic-Like Ser/Thr Kinase PrkC Regulates
544 the Essential WalRK Two-Component System in Bacillus subtilis. PLoS Genet. Public
545 Library of Science; 2015;11.

546 36. Mijakovic I, Macek B. Impact of phosphoproteomics on studies of bacterial
547 physiology. FEMS Microbiol. Rev. FEMS Microbiol Rev; 2012. p. 877–92.

548 37. Dworkin J. Ser/Thr phosphorylation as a regulatory mechanism in bacteria. Curr.
549 Opin. Microbiol. Elsevier Ltd; 2015. p. 47–52.

550 38. Shi L, Pigeonneau N, Ravikumar V, Dobrinic P, Macek B, Franjevic D, et al. Cross-
551 phosphorylation of bacterial serine/threonine and tyrosine protein kinases on key
552 regulatory residues. Front Microbiol. Frontiers Research Foundation; 2014;5.

553 39. Shi L, Ji B, Kolar-Znika L, Boskovic A, Jadeau F, Combet C, et al. Evolution of
554 Bacterial Protein-Tyrosine Kinases and Their Relaxed Specificity Toward Substrates.
555 Genome Biol Evol [Internet]. 2014;6:800–17. Available from:
556 https://doi.org/10.1093/gbe/evu056

557 40. Capra EJ, Perchuk BS, Skerker JM, Laub MT. Adaptive mutations that prevent
558 crosstalk enable the expansion of paralogous signaling protein families. Cell. Cell Press;
559 2012;150:222–32.

560 41. Aravind L, Dixit VM, Koonin E V, Aravind L, Dixit VM, Koonin E V., et al. The
561 domains of death: evolution of the apoptosis machinery. Trends Biochem Sci. Elsevier;
562 1999;24:47–53.

563 42. Klemenčič M, Novinec M, Dolinar M. Orthocaspases are proteolytically active
564 prokaryotic caspase homologues: The case of Microcystis aeruginosa. Mol Microbiol.
565 2015;98.

566 43. Spungin D, Bidle KD, Berman-Frank I. Metacaspase involvement in programmed
567 cell death of the marine cyanobacterium Trichodesmium. Environ Microbiol. 2018;

568 44. Klemenčič M, Funk C. Structural and functional diversity of caspase homologues in
569 non-metazoan organisms. Protoplasma. Springer-Verlag Wien; 2018. p. 387–97.

570 45. Asplund-Samuelsson J. The art of destruction: revealing the proteolytic capacity of
571 bacterial caspase homologs. Mol Microbiol. John Wiley & Sons, Ltd (10.1111);
572 2015;98:1–6.

573 46. Klemenčič M, Funk C. Structural and functional diversity of caspase homologues in
574 non-metazoan organisms. Protoplasma. Springer Vienna; 2018;255:387–97.

575 47. Kenneth Allan R, Ratajczak T. Versatile TPR domains accommodate different
576 modes of target protein recognition and function. Cell Stress Chaperones. Springer;
577 2011. p. 353–67.

578 48. van der Voorn L, Ploegh HL. The WD-40 repeat. FEBS Lett [Internet]. John Wiley &
579 Sons, Ltd; 1992 [cited 2020 Mar 23];307:131–4. Available from:
580 http://doi.wiley.com/10.1016/0014-5793%2892%2980751-2

581 49. Tewari R, Bailes E, Bunting KA, Coates JC. Armadillo-repeat protein functions:
582 Questions for little creatures. Trends Cell Biol. Elsevier Current Trends; 2010. p. 470–
583 81.

584 50. Colombatti A, Bonaldo P, Doliana R. Type A Modules: Interacting Domains Found in
585 Several Non-Fibrillar Collagens and in Other Extracellular Matrix Proteins. Matrix.
586 Elsevier; 1993;13:297–306.

587 51. Swiderski MR, Birker D, Jones JDG. The TIR domain of TIR-NB-LRR resistance
588 proteins is a signaling domain involved in cell death induction. Mol Plant-Microbe

589 Interact. The American Phytopathological Society ; 2009;22:157–65.

590 52. Spear AM, Loman NJ, Atkins HS, Pallen MJ. Microbial TIR domains: not necessarily
591 agents of subversion? Trends Microbiol. Elsevier Current Trends; 2009;17:393–8.

592 53. Ho YS, Burden LM, Hurley JH. Structure of the GAF domain, a ubiquitous signaling
593 motif and a new class of cyclic GMP receptor. EMBO J [Internet]. Oxford University
594 Press; 2000;19:5288–99. Available from: https://pubmed.ncbi.nlm.nih.gov/11032796

595 54. Sobotka R, Dühring U, Komenda J, Peter E, Gardian Z, Tichy M, et al. Importance
596 of the cyanobacterial Gun4 protein for chlorophyll metabolism and assembly of
597 photosynthetic complexes. J Biol Chem. American Society for Biochemistry and
598 Molecular Biology; 2008;283:25794–802.

599 55. Aravind L, Koonin E V. Classification of the caspase-hemoglobinase fold: Detection
600 of new families and implications for the origin of the eukaryotic separins. Proteins Struct
601 Funct Genet. 2002;46:355–67.

602 56. Ohmori M, Ikeuchi M, Sato N, Wolk P, Kaneko T, Ogawa T, et al. Characterization
603 of Genes Encoding Multi-domain Proteins in the Genome of the Filamentous Nitrogen-
604 fixing Cyanobacterium Anabaena sp. Strain PCC 7120. DNA Res. 2001.

605 57. Asplund-Samuelsson J, Bergman B, Larsson J. Prokaryotic Caspase Homologs:
606 Phylogenetic Patterns and Functional Characteristics Reveal Considerable Diversity.
607 Driks A, editor. PLoS One. Public Library of Science; 2012;7:e49888.

608 58. Mammen M, Choi S-K, Whitesides GM. Polyvalent Interactions in Biological
609 Systems: Implications for Design and Use of Multivalent Ligands and Inhibitors. Angew
610 Chemie Int Ed. 1998;37:2754–94.

611 59. Miller JL, Le Coq J, Hodes A, Barbalat R, Miller JF, Ghosh P. Selective ligand
612 recognition by a diversity-generating retroelement variable protein. PLoS Biol.
613 2008;6:1195–207.

614 60. Sharifi F, Ye Y. MyDGR: a server for identification and characterization of diversity-
615 generating retroelements. Nucleic Acids Res. 2019;47:W289–94.

616 61. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of
617 protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.

618 62. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search
619 tool. J Mol Biol. J Mol Biol; 1990;215:403–10.

620 63. Siguier P. ISfinder: the reference centre for bacterial insertion sequences. Nucleic
621 Acids Res. Oxford University Press (OUP); 2006;34:D32–6.

622 64. Hug LA, Castelle CJ, Wrighton KC, Thomas BC, Sharon I, Frischkorn KR, et al.
623 Community genomic analyses constrain the distribution of metabolic traits across the

624 Chloroflexi phylum and indicate roles in sediment carbon cycling. Microbiome. BioMed
625 Central Ltd.; 2013;1:1–17.

626 65. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity
627 searching. Nucleic Acids Res [Internet]. 2011/05/18. Oxford University Press;
628 2011;39:W29–37. Available from: https://pubmed.ncbi.nlm.nih.gov/21593126

629 66. Bateman A. The Pfam Protein Families Database. Nucleic Acids Res. Oxford
630 University Press (OUP); 2002;30:276–80.

631 67. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high
632 throughput. Nucleic Acids Res. 2004;32:1792–7.

633 68. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated
634 alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–
635 3.

636 69. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis
637 of large phylogenies. Bioinformatics [Internet]. 2014;30:1312–3. Available from:
638 https://doi.org/10.1093/bioinformatics/btu033

639 70. Price MN, Dehal PS, Arkin AP. FastTree 2 - Approximately maximum-likelihood
640 trees for large alignments. PLoS One. Public Library of Science; 2010;5:e9490.

641 71. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable
642 generation of high-quality protein multiple sequence alignments using Clustal Omega.
643 Mol Syst Biol [Internet]. John Wiley & Sons, Ltd; 2011 [cited 2020 Apr 23];7:539.
644 Available from: https://onlinelibrary.wiley.com/doi/abs/10.1038/msb.2011.75

645 72. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify
646 genomes with the Genome Taxonomy Database. Bioinformatics [Internet].
647 2019;36:1925–7. Available from: https://doi.org/10.1093/bioinformatics/btz848

648 73. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al.
649 A standardized bacterial taxonomy based on genome phylogeny substantially revises
650 the tree of life. Nat Biotechnol [Internet]. 2018;36:996–1004. Available from:
651 https://doi.org/10.1038/nbt.4229