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.

Conclusion

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

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

46 Background

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

61 Certain members of the cyanobacterial phylum possess an extensive capacity to adapt to
62 various environmental pressures through tightly-controlled regulation of complex
63 cellular programs for signal response. This is exemplified by abilities for metabolic

64 switching (i.e. CO_2/N_2 fixation), maintaining photoreceptors of varying intensity for
65 binary programs of circadian rhythm, and forming specialized cells which can sometimes
66 be terminally differentiated and lead to multicellularity [9]. To regulate these complex
67 programs, cyanobacteria have an extensive repertoire of genes governing signal
68 transduction including proteases, kinases, and nucleases. Notably, paralogs of these
69 regulatory proteins are more abundant among the more complex species of cyanobacteria
70 (i.e. those belonging to subsections III-V) [10,11]. However, the mechanisms to diversify
71 and adapt specific functionality in these duplicated genes remain largely unexplored. One
72 mechanism may be the application of diversity-generating retroelements, known to
73 accelerate the evolution of the proteins they target.

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

87 corresponds to flexible residues in ligand binding structural domains belonging to the C-
88 type lectin or Immunoglobulin (Ig)-like protein families.

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

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

Results and discussion

A Conserved Subclass of Retroelements in Cyanobacteria

Our analysis identified 58 DGRs that include 90 target genes (i.e. encoding VPs) in 52 genomes of cyanobacteria spanning 31 different genera. These include filamentous, colonial, and symbiotic organisms (Fig. 1b and Additional file 1: Table S1). Sequence clustering of the 58 DGRs (RT sequences at 95% identity) suggests they represented 49 distinct DGRs, although the full set of 58 were examined further. All DGRs were identified by presence of diagnostic and essential components: an RT gene; one or more VP genes with VR regions; and a TR region. Among the 52 genomes analyzed, four contain duplicate DGR cassettes, based on clustering, while one contains two unique DGR-RTs. Moreover, several individual DGRs have multiple target genes, and some VP genes have VRs with homology to other genes dispersed throughout the genome (paralogs) (Fig. 1c).

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

130 Phylogenetic analysis of cyanobacterial DGR-RTs revealed a monophyletic clade, unique
131 from all other bacterial DGR-RTs (Fig. 2). The cyanobacterial DGR-RT clade comprises
132 sequences that span nearly all major cyanobacterial genera within morphological
133 subclasses I, III, IV, and V (Fig. 3a). None of the DGR-containing genomes correspond to
134 genera within subclass II. Strikingly, cyanobacterial reverse transcriptases within the
135 monophyletic cyanobacterial DGR clade share an average global sequence identity of 67%
136 (minimum 55%; amino acid sequence). Whereas members of this DGR-RT subgroup do
137 not appear to be shared with other bacteria or archaea, their phylogenetic relationships
138 suggest a complex evolutionary history punctuated by horizontal exchange within the
139 cyanobacterial phylum (Fig. 3b). Although none of the cyanobacterial DGRs could be
140 definitively assigned to prophage elements, they were identified on plasmids of *Anabaena*
141 sp. 90 (CP003287) and *Fischerella* sp. NIES-4106 (AP018301), which may indicate a
142 vehicle for retroelement transfer between closely related populations. Among members
143 of this RT clade, each corresponding DGR-VP contains a ligand-binding C-type lectin-like
144 domain (CLec) with additional functional domains described in detail below.

145 Intragenomic Dispersal of Hypervariable Binding Domains

146 DGR variable proteins often contain multiple distinct structural domains [13,21,25]. To
147 investigate the specific functions of cyanobacterial DGR-targeted proteins (i.e. containing
148 the VR scaffold), we first separately analyzed the ligand-binding CLec domains in all
149 DGR-VPs. This approach identified a conserved module (i.e. a putative C-terminal
150 domain) found in each of the 52 cyanobacterial VP representatives (Additional file 1:
151 Table S1). The entire set of VR-containing modules share sequence homology with 50.5%
152 average identity and, moreover, all of these protein sequences were clustered together

153 with >30% pairwise amino acid identity. Structural prediction of the representative C-
154 terminal domain sequence (i.e. obtained from clustering) determined that each module
155 most closely resembles the C-type Lectin fold, which is represented by the CLec-like
156 superfamily (InterPro: IPR016187). A search for similar proteins in the Uniprot database
157 identified sequences from an array of other genomes among which 92% belong to
158 cyanobacterial phyla (Additional file 2: Table S2). The similarity between CLec domains
159 found in diverse DGRs may underlie a conserved utility for diversifying this module
160 across different cyanobacterial taxa. The CLec-like superfamily has been linked to a
161 variety of molecular processes in cells and viruses spanning the tree of life, with a common
162 functional role in ligand binding generally predicted for this fold [26–28]. Thus, the
163 modular and dispersed nature of a highly conserved CLec subclass may further point to
164 multifaceted functional significance in cyanobacteria.

165 We next sought to address whether hypervariable CLec modules might arise from gene
166 duplication and intragenomic dispersal, resulting in recognizable sets of paralogs in
167 cyanobacterial genomes. This search was limited to 21 high-quality genomes of the 52-
168 genome total, such that draft genomes composed of >50 scaffolds were removed from the
169 analysis. This approach uncovered 21 genomes that have multiple genes encoding CLec
170 domain-containing proteins, with varying degrees of VR/TR homology (Fig. 4 and
171 Additional file 3: Table S3). These paralogs occur both within DGR loci and dispersed
172 throughout the genome and most often consist of either a single CLec domain or the C-
173 terminal CLec grafted to an N-terminal putative serine kinase domain. Taken together,
174 the multi-genome set of 219 cyanobacterial orthologs across 21 genomes-share average
175 pairwise identity of 50.5% within their CLec domains. The complete set of 219 orthologs

176 comprises 121 genes that appear to be DGR-diversified based on VR/TR homology,
177 including 45 VP genes encoded within a DGR. The additional 76 remote targets were
178 associated with their respective genome's DGR(s) using a threshold of TR identity greater
179 than 50%; these matches were exclusively found near the 3' terminus of CLec-encoding
180 genes. The proximity to 3'-termini suggests that conserved, cis-acting features - such as
181 DNA cruciforms or initiation of mutagenic homing sites required for cDNA integration -
182 may play a role in activating remote targets.

183 The genome of *Nostoc* sp. PCC 7120 (formerly *Anabaena*), contains two DGRs and several
184 dispersed VP paralogs (Fig. 4), providing the opportunity to examine the evolutionary
185 history of these genes in an extensively-studied model organism. Within this genome, we
186 identified three highly similar VP homologs (≥ 60% amino acid identity) in dispersed loci,
187 wherein these genes may have proliferated by duplication and transposition from a
188 common ancestral gene. Notably, one of these paralogs (*all3226*) contains remnant TR-
189 VR homology, despite an absence of proximal RT genes or pseudogenes. Taken together,
190 this suggests a capacity for intragenomic dispersal of DGR-targeted variable proteins, and
191 perhaps removal of diversification components once an optimal variant is selected. In
192 addition to its tractability, the common constellation of DGR VPs that occurs in PCC 7120,
193 as observed in other cyanobacteria, make this species an ideal representative for further
194 analysis of the physiological, ecological and evolutionary ramifications of DGR VP
195 functionality and modularity in cyanobacteria.

196 To assess whether transposable elements were found in proximity to DGRs, we analyzed
197 neighborhoods surrounding each hypervariable protein, including remote VPs with

198 respect to a DGR-RT (i.e. > 5 kbp upstream/downstream). This search uncovered
199 transposase genes belonging to various families in DGR-proximal loci which may be
200 responsible for VP dispersal throughout the genome (Additional file 4: Table S4 and
201 Additional file 5: Table S5). Within the subset of 21 high-quality genomes,
202 *Trichodesmium erythraeum* IMS101 has the greatest number of proximal transposase
203 genes, spanning six different insertion sequence (IS) families. The most widely-
204 distributed transposases were those belonging to the IS200/IS605 family, found nearby
205 9 VPs from 6 distinct species. Transposases belonging to this family employ a single-
206 stranded DNA intermediate for a "peel-and-paste" mechanism of transposition [29,30].
207 The genome of *Anabaena* sp. Strain 90 contains remnants of a putative degraded DGR
208 cassette – containing only the RT with no other detectable features – and notably, the RT
209 gene is flanked by proximal transposase genes. This provides a potential mechanism for
210 select components of the DGR to be mobilized within the genome. DGR recruitment to
211 one gene from another would allow favorably diversified genes to become conserved while
212 targeting hypervariation elsewhere in the genome. Selective pressures can then influence
213 the recruitment of DGRs to genes wherein hypervariation for ligand-binding residues
214 offers selective advantages. Through this mechanism of transposition, cyanobacterial
215 DGRs may provide a newly-diversified, modular, ligand binding domain to signaling
216 genes.

217 Function of Multidomain Variable Proteins

218 In part, functional diversity of DGR variable proteins is found in their multidomain
219 complexity. We examined cyanobacterial VPs and their paralogs, which consist of N-
220 terminal domains that are grafted to the C-terminal CLec domain (Fig. 4, Fig. 5). Toward

221 assessing cellular localization, transmembrane and/or signal peptide regions were
222 predicted for 4 DGR-associated VPs and 9 remote VPs, spanning 11 of the 21 high-quality
223 genome set (Additional file 3: Table S3). Most cyanobacterial DGR VPs are predicted to
224 be cytosolic, however evidence exists for TM localization and secretion as well.

225 The most common N-terminal domains of cyanobacterial VPs and their paralogs have
226 similarity to the protein kinase superfamily. These proteins are further predicted to be
227 serine/threonine kinases (STKs) based on the following factors: 1) identification of Hanks
228 and Hunter-type Motifs I through IX [31] (Additional file 6: Figure S1); 2) common NCBI
229 CDS annotations of "serine/threonine protein kinase CDS"; or 3) identification of an STK
230 in previous literature [32]. STKs are mostly associated with eukaryotic signal
231 transduction pathways. In prokaryotes, two-component regulation controls most
232 phosphorylation pathways with a receptor histidine kinase paired with various response
233 regulators phosphorylated on aspartic acid residues. These kinases often control the
234 expression of certain genes [33]. However, Hanks-type STKs have been found in an array
235 of prokaryotic organisms where their genomic abundance is often correlated with genome
236 size, physiological and ecophysiological complexity, and ability to tolerate complex
237 environments [31,32,34]. These STKs are implicated in the regulation of various aspects
238 of bacterial physiology through post-translational modification of proteins, which may
239 themselves be components of phosphorelay and transcriptional regulatory pathways [34–
240 37]. Serine/threonine protein kinases were first associated with the pknA gene of *Nostoc*
241 sp. PCC7120, which is involved in growth and differentiation [32], and in other bacteria
242 their activity regulates processes such as cell growth, segregation, virulence, metabolism,
243 stress adaptation, and cell wall/envelope biogenesis [34]. Ser/Thr kinases in

244 cyanobacteria are usually associated with three different processes: developmental
245 regulation, stress response, and pathogenicity [32]. Slight changes, not in function but in
246 the strength of substrate recognition to a variety of phosphorylation targets, may
247 contribute to the ability to finely tune networks of signal transduction.

248 Compared to histidine kinases of two-component systems, which exhibit strong substrate
249 discrimination, STKs have relaxed substrate specificities. This has been linked to a lack of
250 co-evolution between the kinase and its cognate target [38,39]. Accelerated evolution of
251 the substrate-binding domain of these kinases may have resulted in the further expansion
252 of this class of proteins in the Cyanobacteria phylum, contributing to a wide range of
253 adaptability to external stimuli and challenging environments. We hypothesize that the
254 VR-containing CLec domain could be autoinhibitory, and activation of kinase activity
255 would occur upon binding a small molecule or protein ligand. In this case, DGR-mediated
256 diversity could allow rapid recognition of various ligands for activating phosphorylation
257 cascades. Alternatively, the CLec domain could function in ligand recognition (i.e.
258 determining what protein(s) are phosphorylated). In this case CLec variants could have
259 different substrate specificities. Segregation of phosphorylation targets between
260 paralogous kinases has been shown to play a strong selective pressure in their evolution
261 [40]. In turn, DGR-driven hypervariation of binding components in signaling proteins
262 may offer additional selective advantages in cyanobacteria through preventing cross-talk,
263 which is characteristic of this of kinase class.

264 We also identified orthocaspase-like peptidase domains in VP N-termini, which are also
265 common among their paralogs (Fig. 4). Caspase proteins are proteases involved in the

266 initiation of programmed cell death in metazoans [41]. The peptidase domains that we
267 identified in many VPs and their paralogs were predicted as orthocaspases, which are the
268 prokaryotic homologs of eukaryotic caspase-type proteases [42]. While these protein
269 types are homologous to metazoan caspases, current evidence supports a broader role in
270 cell homeostasis during normal cellular conditions, programs of cellular differentiation,
271 or ageing as well as potential apoptosis [43–45]. Previous studies have found
272 orthocaspases to be enriched in morphologically complex filamentous cyanobacteria of
273 subsections III-V (e.g. *Trichodesmium erythraeum* IMS 101, *Anabaena* spp., and *Nostoc*
274 spp.) as well as various strains of the unicellular toxin-producing species, *Microcystis*
275 *aeruginosa*. Conversely, orthocaspases are entirely absent from unicellular genera
276 *Synechococcus*, *Prochlorococcus*, *Cyanobium*, and *Cyanothece* and are
277 underrepresented in the genomes of cyanobacteria belonging to subsections I-II. This
278 suggests their utility in enabling the complex signal response and regulatory programs
279 that exist in cyanobacteria capable of cellular differentiation, toxin production, and
280 diazotrophy [46].

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

286 A common feature for nearly all of the N-terminal domains, including the prevalent
287 protein kinase-containing paralogs, is their potential to serve a functional role in signal

288 transduction in response to external stimuli (e.g. light, nutrient deprivation, and general
289 stress response) [9]. A previous study found that genes encoding complex multidomain
290 proteins involved in signal transduction are highly enriched in the filamentous
291 cyanobacterium *Anabaena* sp. PCC 7120 when compared to the genomes of unicellular
292 *Synechocystis* sp. PCC 6803 and *Pseudomonas aeruginosa* [56]. Moreover, regulatory
293 proteins involved in signal transduction could lend to the complex regulation necessary
294 for the physiology of filamentous cyanobacteria. These physiologies include a capacity for
295 cell-differentiation, producing heterocysts during nitrogen deprivation and akinetes
296 under environmental stress, as well as programmed apoptosis [43,57]. The presence of
297 DGRs in cyanobacteria follows this trend in the abundance of specialized signal
298 transduction proteins – being seemingly enriched in filamentous nitrogen-fixing taxa and
299 absent from genomes of unicellular taxa, *Synechococcus* spp. and *Prochlorococcus* spp.,
300 though they are present in other unicellular species.

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

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

320 Conclusions

321 The DGR-enabled diversification of proteins involved in host attachment should lead to
322 selective advantages, as this offers an offensive selective advantage to variation by the
323 host cell. By genomic inference, other DGR-containing prokaryotes seem to have adopted
324 DGR function to mechanisms of virulence, and other cell-cell or phage/cell binding
325 interactions. By contrast, our findings suggest a selective use of DGRs for purposes of
326 isolated hyper-diversification of a small pocket in the C-terminal binding domains of
327 multidomain proteins broadly involved in signal transduction within cyanobacteria. This
328 class of DGR-target proteins is thus-far unique to the cyanobacterial phylum.
329 Diversification of the binding site of these proteins, paired with natural selection over
330 iterations of diversity generation and the ability to segregate resulting beneficial mutants
331 via transposition, may contribute to the complexity and adaptability of cellular regulation
332 amongst cyanobacterial taxa. In developing a better grasp on the functional significance

of DGR hypervariation, it is clear that the phenomenon adds new layers of complexity in the expansion of bacterial protein networks.

Figures

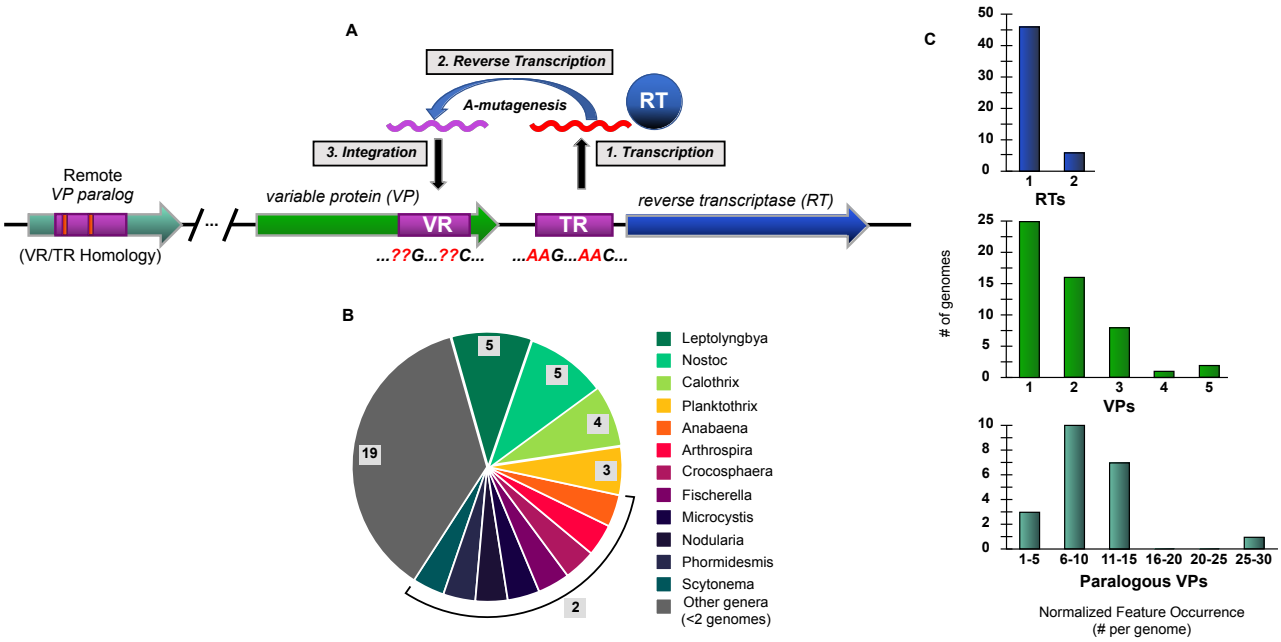


Figure 1. Schematic overview of a DGR and prevalence in cyanobacterial genomes.
a Three primary steps in the process of mutagenic homing are shown: **1)** conserved template region (TR) in the DGR cassette is transcribed into intermediate, non-coding RNA, which is the substrate for DGR reverse transcriptase (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.

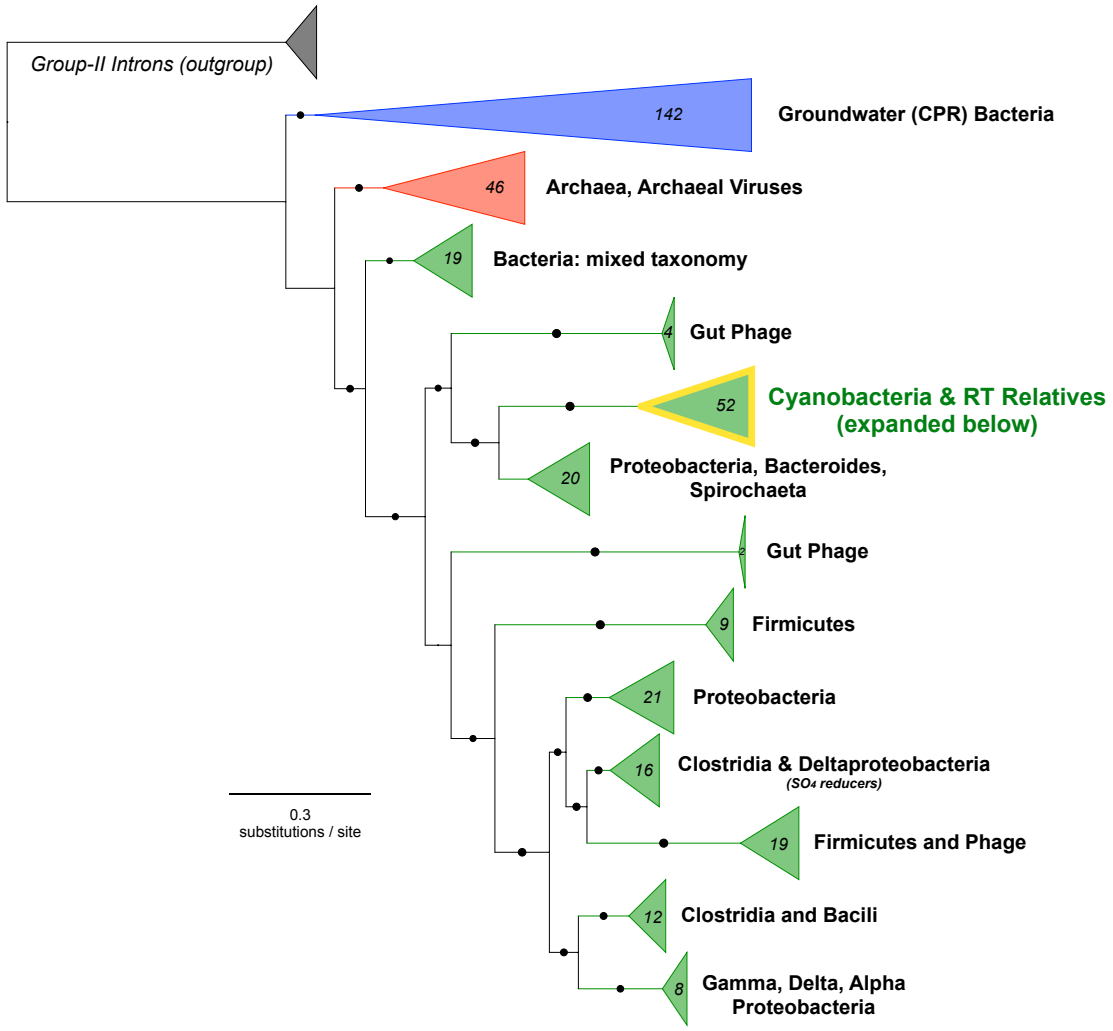


Figure 2. Broad RT phylogeny compared amongst all known DGR-containing lineages.
RT phylogeny compared amongst all known DGR-containing lineages with Group-II Introns as the outgroup (highlighted cyanobacterial clade expanded in Figure 3).

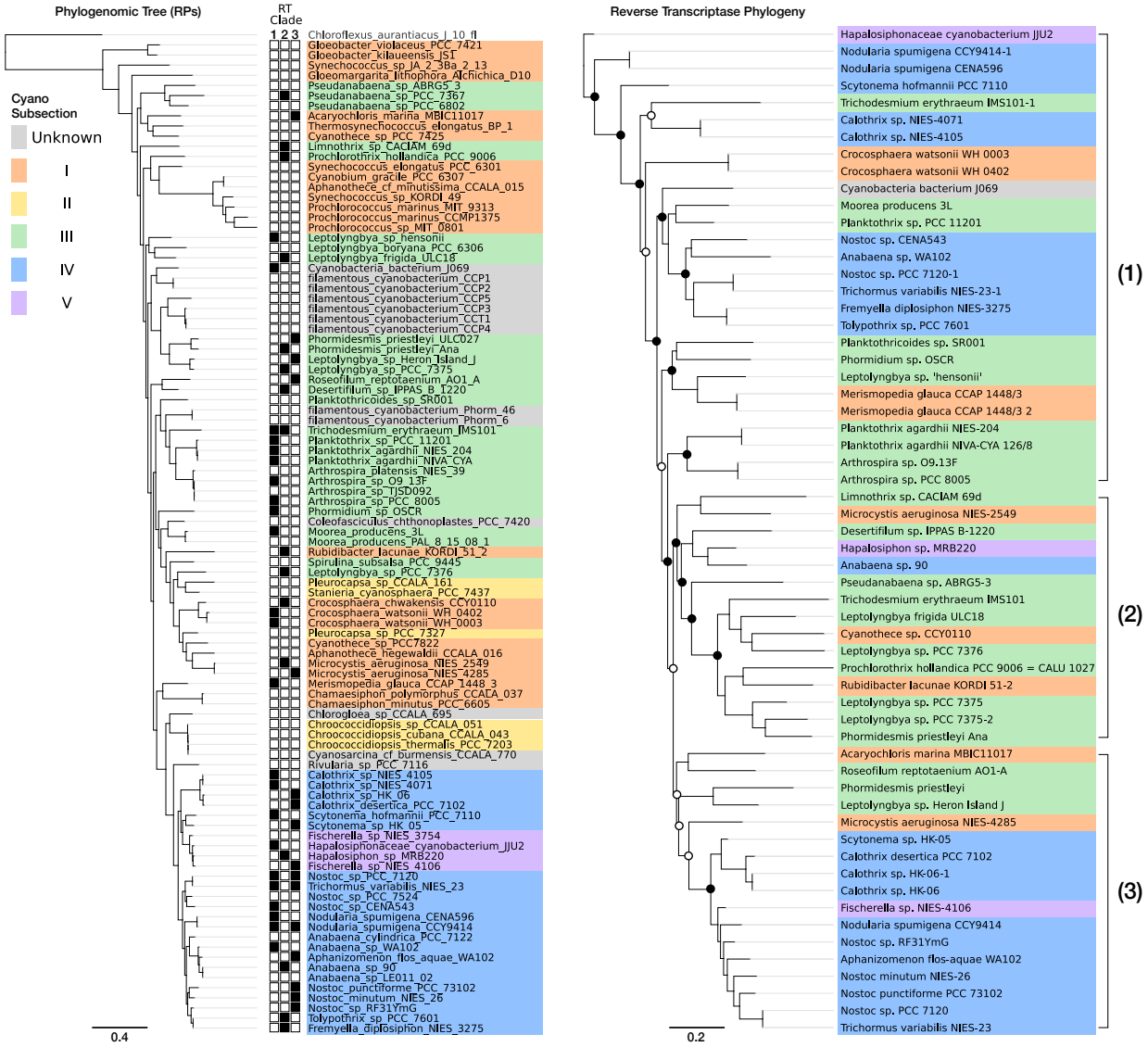


Figure 3. Phylogenetic reconstruction for DGR-containing cyanobacteria and DGR-RT phylogeny.
a Phylogeny for concatenated ribosomal protein alignments, including all DGR-containing species. Filled boxes (left) indicate DGR-RT containing species and the corresponding RT clade. **b** DGR-RT phylogeny with cyanobacterial physiological subsections highlighted in color. Circles indicate branch support values (hollow >50%; filled >70%).

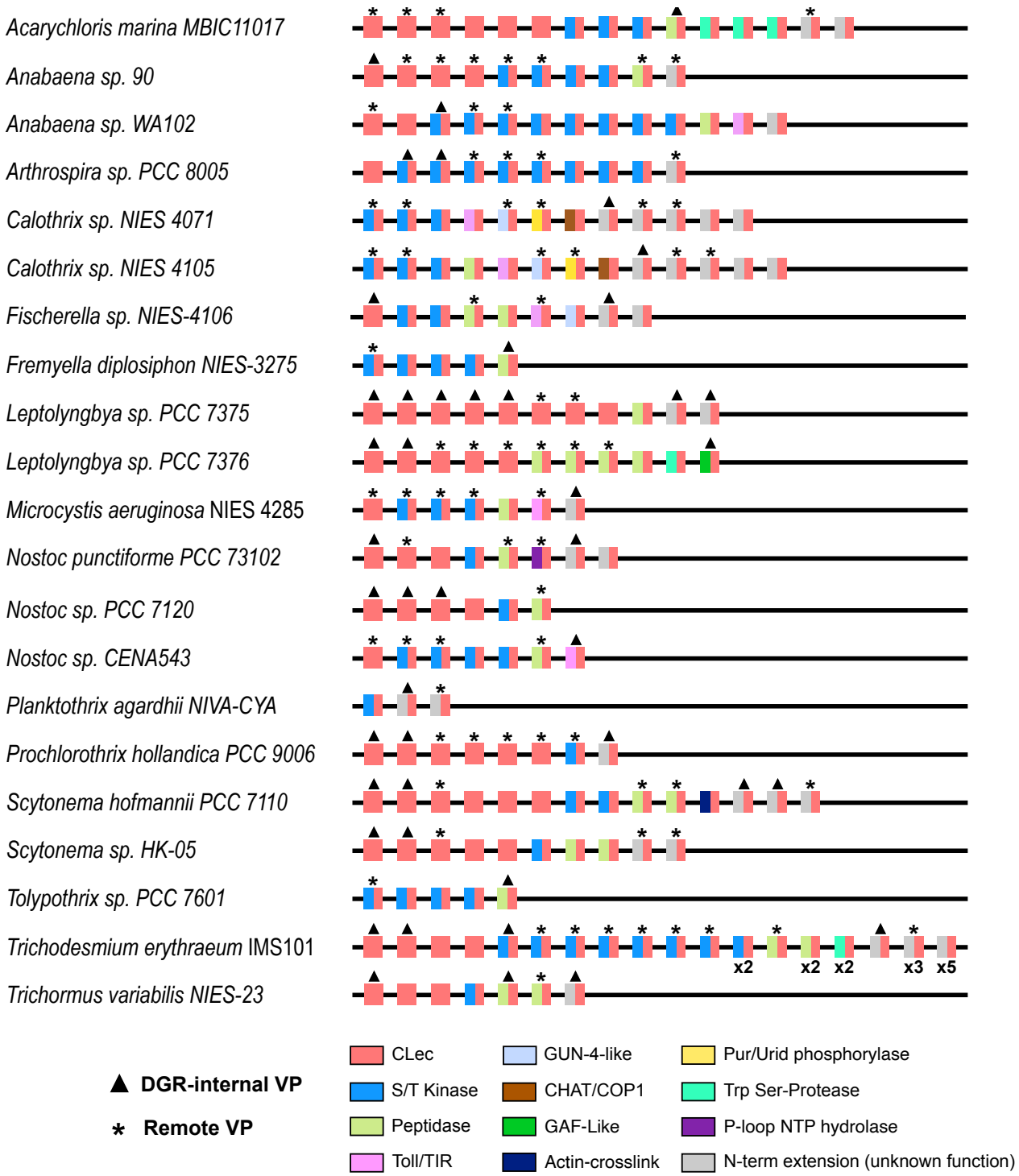
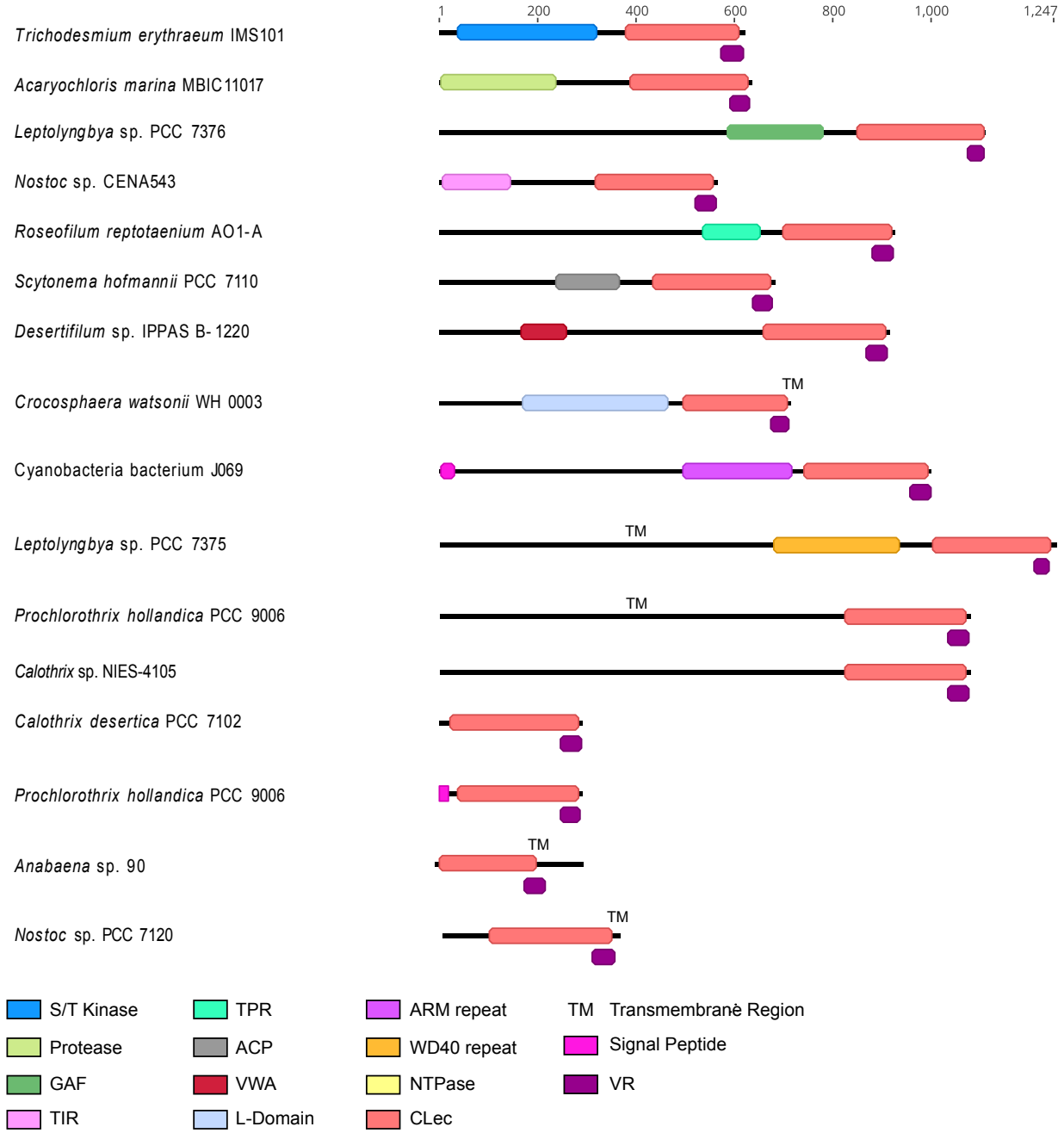


Figure 4. Distribution of CLec variable protein paralogs in cyanobacterial genomes.

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

369 same domain architecture in *Trichodesmium erythraeum* IMS101 are indicated below
370 the representative protein (e.g. x2). Note: only a representative subset of DGR-
371 containing genomes is shown (15 of 52 genomes).



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

377 # Methods

378 ## DGR Identification and Annotation

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

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

400 Genes, homologous to VPs within the DGR cassette, were inspected by aligning amino
401 acid sequences for the CLec domain of each putative remote VP to the DGR-VP within the
402 same organism using Clustal Omega. Genes with CLec domains having a putative VR with
403 ≥50% nucleotide identity to a DGR-VP were designated as remote VPs, while those with
404 <50% were designated as VP homologs. DGR and remote VP neighborhood regions were
405 defined as regions 10kb upstream and downstream from the DGR cassette, remote VP, or
406 VP homolog.

407 Neighborhood Analyses

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

413 Phylogenetic Analyses

414 To construct a phylogenetic tree of cyanobacteria, we used a set of 16 ribosomal proteins
415 often used for phylogenomic analysis (RpL2, 3, 4, 5, 6, 14, 15, 16, 18, 22, and 24, and RpS3,
416 8, 10, 17, 19 [64]. Each ribosomal protein was identified using HMMER [65] and hidden
417 Markov models from the Pfam [66] database (accessed September 2018). Each individual
418 marker gene was aligned using MUSCLE [67], trimmed using TrimAL [68], manually
419 assessed to remove end gaps and ambiguously aligned regions and concatenated. A
420 maximum likelihood tree was constructed using RAxML v. 8.2.9 [69] with the
421 PROTCATLG model.

422 To reconstruct RT phylogeny, putative DGR-RT coding sequences were identified, as
423 described above, then translated. Sequences were de-replicated and non-redundant
424 candidates were chosen using CD-Hit [61] with a global alignment threshold of 99%
425 identity. All DGR-RT sequences and a set of Group-II intron RT sequences from Bacteria,
426 Archaea, plastids, and mitochondria were aligned with a hidden markov model of the
427 reverse transcriptase protein family (PF00078) using HMMalign [65]. A phylogenetic
428 tree of DGR-RTs was constructed using FastTree2 [70] with the WAG substitution matrix,
429 and the CAT approximation to optimize branch lengths. The cyanobacterial DGR-RT
430 representatives were extracted from the complete alignment, realigned using Clustal
431 Omega [71] and used to construct an unrooted phylogenetic tree.

432 Protein Function Analysis

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

442 Availability of data and materials

443 The datasets supporting the conclusions of this article are included within the article and
444 its additional files.

445 ## Acknowledgements

446 This research was funded by a Challenge Grant from the California NanoSystems Institute
447 439 (CNSI-UCSB). AVE is supported by a National Science Foundation Graduate
448 Research 440 Fellowship Program under Grant No. 1650114, and by the NSF California
449 LSAMP Bridge to the 441 Doctorate Fellowship under Grant No. HRD-1701365. The work
450 conducted by the U.S. 442 Department of Energy Joint Genome Institute is supported by
451 the Office of Science of the U.S. 443 Department of Energy under contract no. DE-AC02-
452 05CH11231. Computational analyses for 444 this work were supported by an NSF-XSEDE
453 resource allocation DEB170007.

454 ## Supplementary information

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

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

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

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

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

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

480 **Additional File 7: Table S6. Updated Cyanobacterial Taxonomy**.
481 Taxonomic assignment for each cyanobacterial genome was generated using relative
482 evolutionary divergence and average nucleotide identity with the Genome Taxonomy
483 Database Toolkit (GTDB-Tk) [72] based on the Genome Taxonomy Database (GTDB)
484 [73].

485 # References

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

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

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

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

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

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

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

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

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

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

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

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

519 13. Wu L, Gingery M, Abebe M, Arambula D, Czornyj E, Handa S, et al. Diversity-

520 generating retroelements: natural variation, classification and evolution inferred from a
521 large-scale genomic survey. Nucleic Acids Res. Narnia; 2018;46:11–24.

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

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

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

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

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

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

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

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

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

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

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

555 25. Guo H, Arambula D, Ghosh P, Miller JF. Diversity-generating Retroelements in

556 Phage and Bacterial Genomes. Microbiol Spectr. NIH Public Access; 2014;2.

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

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

562 28. Zelensky AN, Gready JE. The C-type lectin-like domain superfamily. FEBS J.
563 2005;272:6179–217.

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

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

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

572 32. Zhang X, Zhao F, Guan X, Yang Y, Liang C, Qin S. Genome-wide survey of putative
573 Serine/Threonine protein kinases in cyanobacteria. BMC Genomics. BioMed Central;
574 2007;8:395.

575 33. Stock AM, Robinson VL, Goudreau PN. Two-Component Signal Transduction. Annu
576 Rev Biochem. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA
577 94303-0139, USA; 2000;69:183–215.

578 34. Janczarek M, Vinardell J-M, Lipa P, Karaś M. Hanks-Type Serine/Threonine
579 Protein Kinases and Phosphatases in Bacteria: Roles in Signaling and Adaptation to
580 Various Environments. Int J Mol Sci [Internet]. MDPI AG; 2018 [cited 2020 Mar
581 22];19:2872. Available from: http://www.mdpi.com/1422-0067/19/10/2872

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

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

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

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

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

596 40. Capra EJ, Perchuk BS, Skerker JM, Laub MT. Adaptive mutations that prevent
597 crosstalk enable the expansion of paralogous signaling protein families. Cell. Cell Press;
598 2012;150:222–32.

599 41. Aravind L, Dixit VM, Koonin E V, Aravind L, Dixit VM, Koonin E V., et al. The
600 domains of death: evolution of the apoptosis machinery. Trends Biochem Sci. Elsevier;
601 1999;24:47–53.

602 42. Klemenčič M, Novinec M, Dolinar M. Orthocaspases are proteolytically active
603 prokaryotic caspase homologues: The case of Microcystis aeruginosa. Mol Microbiol.
604 2015;98.

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

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

609 45. Asplund-Samuelsson J. The art of destruction: revealing the proteolytic capacity of
610 bacterial caspase homologs. Mol Microbiol. John Wiley & Sons, Ltd (10.1111);
611 2015;98:1–6.

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

614 47. Kenneth Allan R, Ratajczak T. Versatile TPR domains accommodate different modes
615 of target protein recognition and function. Cell Stress Chaperones. Springer; 2011. p.
616 353–67.

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

620 49. Tewari R, Bailes E, Bunting KA, Coates JC. Armadillo-repeat protein functions:
621 Questions for little creatures. Trends Cell Biol. Elsevier Current Trends; 2010. p. 470–
622 81.

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

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

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

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

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

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

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

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

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

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

650 59. Miller JL, Le Coq J, Hodes A, Barbalat R, Miller JF, Ghosh P. Selective ligand
651 recognition by a diversity-generating retroelement variable protein. PLoS Biol.
652 2008;6:1195–207.

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

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

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

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

661 64. Hug LA, Castelle CJ, Wrighton KC, Thomas BC, Sharon I, Frischkorn KR, et al.
662 Community genomic analyses constrain the distribution of metabolic traits across the
663 Chloroflexi phylum and indicate roles in sediment carbon cycling. Microbiome. BioMed

664 Central Ltd.; 2013;1:1–17.

665 65. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity
666 searching. Nucleic Acids Res [Internet]. 2011/05/18. Oxford University Press;
667 2011;39:W29–37. Available from: https://pubmed.ncbi.nlm.nih.gov/21593126

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

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

672 68. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated
673 alignment trimming in large-scale phylogenetic analyses. Bioinformatics.
674 2009;25:1972–3.

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

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

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

684 72. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify
685 genomes with the Genome Taxonomy Database. Bioinformatics [Internet].
686 2019;36:1925–7. Available from: https://doi.org/10.1093/bioinformatics/btz848

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