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

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.

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 various wavelength sensitivities for binary programs of circadian rhythm, and forming specialized cells which can sometimes be terminally differentiated and lead to multicellularity [9, 10]. To regulate these complex programs, cyanobacteria have an extensive repertoire of genes governing signal transduction including proteases, kinases, and nucleases. Notably, paralogs of these regulatory proteins are more abundant among the more complex species of cyanobacteria (i.e. those belonging to subsections III-V) [11–15]. However, the mechanisms to diversify and adapt specific functionality in these duplicated genes remain largely unexplored. One mechanism may involve diversity-generating retroelements (DGRs), known to accelerate the evolution of the proteins they target.

Diversity-generating retroelements (DGRs) have been noted within the genomes of several genera of cyanobacteria [16–18]. In experimentally investigated bacterial and viral systems, DGRs drive site-specific hypermutation of a subset of codons in target genes [19, 20], while metagenomic and metatranscriptomic evidence also points to functional DGRs in archaea [21]. These retroelements utilize a uniquely targeted form of retrotransposition. To this end, DGRs insert variants into a flexible coding scaffold, while avoiding non-specific variation in conserved portions of a gene [22]. The essential features of a DGR are most often found within a single genomic locus spanning ~ 5 – 10 kbp (Fig. 1a), though the synteny and organization of DGR components can vary [17]. Diversification is mechanistically carried out by a reverse transcriptase (RT), which acts upon a non-coding RNA transcribed from the template repeat (TR) region in the locus [23]. This region is nearly identical to a variable region (VR) that typically resides in a nearby gene, which encodes a DGR-variable protein (VP). The TR-RNA intermediate is

reverse transcribed into cDNA wherein A → N mutation is highly favored by the error-prone RT. This cDNA then replaces VR, whose sequence commonly corresponds to flexible residues in ligand binding structural domains belonging to the C-type lectin or immunoglobulin-like protein families [19].

The first DGR variable protein was characterized from the bacteriophage, BPP-1. In these phage, DGRs diversify tail fiber tip proteins that recognize and bind to *Bordetella* host receptors [16, 24]. Other cellular DGRs have been characterized in bacterial pathogens, including *Legionella pneumophila* [20] and *Treponema denticola* [25], where DGRs target genes that encode for cellular surface proteins, presumably involved in cell-cell attachment. The conserved function of cell-cell or viral-cell attachment in these target genes lends to a perspective of DGRs for broad use in host recognition for symbiosis or infection. Moreover, several genera of cyanobacteria were identified in recent genomic and metagenomic surveys of DGRs [17, 26]. The essential components of DGRs can be found across most lineages of prokaryotic life [17, 21, 26–29], suggesting broad utility of this form of localized mutation.

Whereas previously characterized DGR target proteins appear to share a functional role in extracellular attachment to ligands displayed on foreign cells, these retroelements could potentially diversify other cellular proteins with entirely distinct functions. The intermediate RNA, which presents a template for DGR mutagenesis, has been shown to be highly expressed in lab isolates of *Trichodesmium erythraeum* IMS101 [18] and in *Nodularia spumigena* CCY9414 under light and oxidative stress [30, 31]. Here, a systematic analysis of DGRs and their variable proteins in cyanobacterial genomes leads

109 to a new perspective on the utility of diversification and optimization of modular protein
110 domains in paralogs that appear linked to signaling and transcriptional control.

111 Results and Discussion

112 A Conserved Subclass of Retroelements in Cyanobacteria

113 Our analysis identified 58 DGRs that include 90 target genes (i.e. encoding VPs) in 52
114 genomes of cyanobacteria spanning 31 different genera. These include filamentous,
115 colonial, and symbiotic organisms (Fig. 1b and Additional file 1: Table S1). Sequence
116 clustering of the 58 DGRs was performed with RT amino acid sequences (at 95% identity)
117 to generate a non-redundant subset of 49 distinct RT genes for phylogenetic analysis,
118 while the full set of 58 were also examined further. All DGRs were identified by presence
119 of diagnostic and essential components: an RT gene; one or more VP genes with VR
120 regions; and a TR region. Our initial RT search was conducted with the UniprotKB coding
121 sequence database, which is in turn linked to complete and draft genomes in
122 EMBL/GenBank/DDBJ databases. The resulting 52 cyanobacterial genomes represent all
123 sequences where complete DGR cassettes were positively identified. Among the 52
124 genomes analyzed, four contain duplicate DGR cassettes, based on clustering, while one
125 contains two unique DGR-RTs. Moreover, several individual DGRs have multiple target
126 genes, and some VP genes have VRs with homology to other genes dispersed throughout
127 the genome (paralogs) (Fig. 1c).

128 To evaluate the diversity of cyanobacteria-encoded DGRs, we first compared these
129 representatives to a recently developed, global metagenomic DGR dataset [26].
130 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 [26]. Given that the cyanobacterial DGRs appear to cluster tightly together, we next sought to analyze phylogenetic relationships within this set.

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

Intragenomic Dispersal of Conserved Domains with Local Hypervariable Regions

DGR variable proteins often contain multiple distinct structural domains [17, 21, 22]. To investigate the specific functions of cyanobacterial DGR-targeted proteins (i.e. containing the VR scaffold), we first separately analyzed the ligand-binding CLec domains in all DGR-VPs. This approach identified a conserved module (i.e. a putative C-terminal domain) with a localized region of hypervariable residues found in each of the 52 cyanobacterial VP representatives (Additional file 1: Table S1). The entire set of VR-containing modules share sequence homology with 50.5% average identity and, moreover, all of these protein sequences were clustered together with >30% pairwise amino acid identity. Structural prediction of the representative C-terminal domain sequence (i.e. obtained from clustering) determined that each module most closely resembles the C-type Lectin domain, which is represented by the CLec-like superfamily (InterPro: IPR016187). In each of these proteins, the DGR variable region (VR) occurs within the C-terminal region of the otherwise conserved CLec-like domain. A search for similar proteins in the Uniprot database identified sequences from an array of other genomes among which 92% belong to cyanobacterial phyla (Additional file 2: Table S2). The similarity between CLec domains found in diverse DGRs may underlie a conserved utility for diversifying this module across different cyanobacterial taxa. The CLec-like superfamily has been linked to a variety of molecular processes in cells and viruses spanning the tree of life, with a common functional role in ligand binding generally predicted for this fold [32–34]. Thus, the modular and dispersed nature of a highly conserved CLec subclass may further point to multifaceted functional significance in cyanobacteria.

We next sought to address whether hypervariable CLec modules might arise from gene duplication and intragenomic dispersal, resulting in recognizable sets of paralogs in cyanobacterial genomes. This search was limited to 21 high-quality genomes of the 52-genome total, such that draft genomes composed of >50 scaffolds were removed from the analysis. This approach uncovered 21 genomes that have multiple genes encoding CLec domain-containing proteins, with varying degrees of VR/TR homology (Fig. 4 and Additional file 3: Table S3). These paralogs occur both within DGR loci and dispersed throughout the genome and most often consist of either a single CLec domain or the C-terminal CLec grafted to an N-terminal putative serine kinase domain. Taken together, the multi-genome set of 219 cyanobacterial orthologs across 21 genomes share average pairwise identity of 50.5% within their CLec domains. The complete set of 219 orthologs comprises 121 genes that appear to be DGR-diversified based on VR/TR homology, including 45 VP genes encoded within a DGR. The additional 76 remote targets were associated with their respective genome's DGR(s) using a threshold of TR identity greater than 50%; these matches were exclusively found near the 3' terminus of CLec-encoding genes. The proximity to 3'-termini suggests that conserved, cis-acting features - such as DNA cruciforms or initiation of mutagenic homing sites required for cDNA integration [35] - may play a role in activating remote targets.

The genome of Nostoc sp. PCC 7120 (formerly Anabaena), contains two DGRs (RT accessions: Alr3497, All5014) and several dispersed VP paralogs (Fig. 4), providing the opportunity to examine the evolutionary history of these genes in an extensively-studied model organism. Within this genome, we identified three highly similar VP homologs (≥ 60% amino acid identity) in dispersed loci, wherein these genes may have proliferated by

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

To assess whether transposable elements were found in proximity to DGRs, we analyzed neighborhoods surrounding each hypervariable protein, including remote VPs with respect to a DGR-RT (i.e. > 5 kbp upstream/downstream). This search uncovered transposase genes belonging to various families in DGR-proximal loci which may be responsible for VP dispersal throughout the genome (Additional file 4: Table S4 and Additional file 5: Table S5). Within the subset of 21 high-quality genomes, *Trichodesmium erythraeum* IMS101 has the greatest number of proximal transposase genes, spanning six different insertion sequence (IS) families. The most widely-distributed transposases were those belonging to the IS200/IS605 family, found nearby 9 VPs from 6 distinct species. Transposases belonging to this family employ a single-stranded DNA intermediate for a “peel-and-paste” mechanism of transposition [36, 37]. The genome of *Anabaena* sp. Strain 90 contains remnants of a putative degraded DGR cassette – containing only the RT with no other detectable features – and notably, the RT gene is flanked by proximal transposase genes. This provides a potential mechanism for

select components of the DGR to be mobilized within the genome. DGR recruitment to one gene from another would allow favorably diversified genes to become conserved while targeting hypervariation elsewhere in the genome. Selective pressures can then influence the recruitment of DGRs to genes wherein hypervariation for ligand-binding residues offers selective advantages. Through this mechanism of transposition, cyanobacterial DGRs may provide a newly-diversified, modular, ligand binding domain to signaling genes.

Function of Multidomain Variable Proteins

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

The most common functional domain of DGR-internal target protein (VPs) in cyanobacteria have similarity to the protein kinase superfamily (Additional file 1: Table S1). Multidomain DGR-external VPs and paralogs of DGR VPs are also most-often predicted to be kinases (Fig. 4). The VP and VP paralog kinase proteins are further predicted to be serine/threonine kinases (STKs) based on the following factors: 1) identification of Hanks and Hunter-type Motifs I through IX [38] (Additional file 6: Figure S1); 2) common NCBI CDS annotations of “serine/threonine protein kinase CDS”; or 3) identification of an STK in previous literature [14]. STKs are mostly associated with

eukaryotic signal transduction pathways. In prokaryotes, two-component regulation controls most phosphorylation pathways with a receptor histidine kinase paired with various response regulators phosphorylated on aspartic acid residues. These kinases often control the expression of certain genes [39]. However, Hanks-type STKs have been found in an array of prokaryotic organisms where their genomic abundance is often correlated with genome size, physiological and ecophysiological complexity, and ability to tolerate complex environments [14, 38, 40]. These STKs are implicated in the regulation of various aspects of bacterial physiology through post-translational modification of proteins, which may themselves be components of phosphorelay and transcriptional regulatory pathways [40–43]. Serine/threonine protein kinases were first associated with the pknA gene of *Nostoc* sp. PCC7120, which is involved in growth and differentiation [14], and in other bacteria their activity regulates processes such as cell growth, segregation, virulence, metabolism, stress adaptation, and cell wall/envelope biogenesis [40]. Ser/Thr kinases in cyanobacteria are usually associated with three different processes: developmental regulation, stress response, and pathogenicity [14]. Slight changes, not in function but in the strength of substrate recognition to a variety of phosphorylation targets, may contribute to the ability to finely tune networks of signal transduction.

Compared to histidine kinases of two-component systems, which exhibit strong substrate discrimination, STKs have relaxed substrate specificities. This has been linked to a lack of co-evolution between the kinase and its cognate target [44, 45]. Accelerated evolution of the substrate-binding domain of these kinases may have resulted in the further expansion of this class of proteins in the Cyanobacteria phylum, contributing to a wide range of

adaptability to external stimuli and challenging environments. We hypothesize that the VR-containing CLec domain could be autoinhibitory, and activation of kinase activity would occur upon binding a small molecule or protein ligand. In this case, DGR-mediated diversity could allow rapid recognition of various ligands for activating phosphorylation cascades. Alternatively, the CLec domain could function in ligand recognition (i.e. determining what protein(s) are phosphorylated). In this case CLec variants could have different substrate specificities. Segregation of phosphorylation targets between paralogous kinases has been shown to play a strong selective pressure in their evolution [46]. In turn, DGR-driven hypervariation of binding components in signaling proteins may offer additional selective advantages in cyanobacteria through preventing cross-talk, which is characteristic of this of kinase class.

We also identified orthocaspase-like peptidase domains in VP N-termini, which are also common among their paralogs (Fig. 4). Caspase proteins are proteases involved in the initiation of programmed cell death in metazoans [47]. The peptidase domains that we identified in many VPs and their paralogs were predicted as orthocaspases, which are the prokaryotic homologs of eukaryotic caspase-type proteases [48]. While these protein types are homologous to metazoan caspases, current evidence supports a broader role in cell homeostasis during normal cellular conditions, programs of cellular differentiation, or ageing as well as potential apoptosis [49, 13, 50]. Previous studies have found orthocaspases to be enriched in morphologically complex filamentous cyanobacteria of subsections III-V (e.g. *Trichodesmium erythraeum* IMS 101, *Anabaena* spp., and *Nostoc* spp.) as well as various strains of the unicellular toxin-producing species, *Microcystis aeruginosa*. Conversely, orthocaspases are entirely absent from unicellular genera

291 *Synechococcus*, *Prochlorococcus*, *Cyanobium*, and *Cyanothece* and are
292 underrepresented in the genomes of cyanobacteria belonging to subsections I-II. This
293 suggests their utility in enabling the complex signal response and regulatory programs
294 that exist in cyanobacteria capable of cellular differentiation, toxin production, and
295 diazotrophy [13].

296 In addition to the serine/threonine kinase and orthocaspases-like peptidase domains, we
297 identified less-common features including repeat motifs, toll/interleukin receptor (TIR)-
298 like, GAF-like, GUN4-like, and CHAT-like domains. Repeat motifs may have a role in
299 protein-protein interactions (e.g., TPR, WD40 repeat, ARM repeat, and VWA-CoxE) [51–
300 54], while the other domains have been linked to intracellular signal transduction [55–
301 59] (Fig. 5).

302 A common feature for nearly all of the N-terminal domains, including the prevalent
303 protein kinase-containing paralogs, is their potential to serve a functional role in signal
304 transduction in response to external stimuli (e.g. light, nutrient deprivation, and general
305 stress response) [9]. A previous study found that genes encoding complex multidomain
306 proteins involved in signal transduction are highly enriched in the filamentous
307 cyanobacterium *Anabaena* sp. PCC 7120 when compared to the genomes of unicellular
308 *Synechocystis* sp. PCC 6803 and *Pseudomonas aeruginosa* [60]. Moreover, regulatory
309 proteins involved in signal transduction could lend to the complex regulation necessary
310 for the physiology of filamentous cyanobacteria. These physiologies include a capacity for
311 cell-differentiation, producing heterocysts during nitrogen deprivation and akinetes
312 under environmental stress, as well as programmed apoptosis [49, 61]. The presence of

DGRs in cyanobacteria follows this trend in the abundance of specialized signal transduction proteins – being seemingly enriched in filamentous nitrogen-fixing taxa and absent from genomes of unicellular taxa, *Synechococcus* spp. and *Prochlorococcus* spp., though they are present in other unicellular species.

DGR-programmed variation of the ligand-binding domain of receptor-binding proteins in Bordetella bacteriophage has been shown to increase the capacity of these proteins to recognize a vast array of molecules. Moreover, diversification of oligomeric structures appeared to confer an amplification of binding affinity, or avidity [19, 62]. Specifically, the existence of 12 DGR-variable target protein trimers in each bacteriophage virion was shown to increase the binding strengths of these proteins to their ligand, pertactin, by relaxing the requirement for optimal binding between the ligand and any single monomer. This multivalent binding was also shown to lead to more distinction in binding events, contributing to enhanced selectivity [19]. These two properties of avidity through multivalency are hypothesized to be characteristic of other DGR systems as a means to provide ligand-recognition flexibility to evolve under constrained conditions, while maintaining selectivity. We hypothesize that, in cyanobacteria, DGR-programmed variation might have a role in providing multimeric avidity in terms of ligand binding for signal response. In the case of autoinhibitory variable proteins attached to a kinase, rather than providing flexibility in host-receptor binding, as in Bordetella bacteriophage, increased avidity may hold a kinetic advantage for substrate binding, whereby flexible activation accelerates signal transduction and regulation. More generally, the available genomic evidence is consistent with a phenomenon of targeted diversification acting to tune cyanobacterial regulatory networks.

Conclusions

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

Declarations

Ethics Approval and Consent to Participate

Not Applicable

Consent for Publication

We hereby give consent for publication in *BMC Genomics*.

Abbreviations

ARM: Armadillo repeat

CHAT: Caspase HetF Associated with Tprs

CLec: C-type lectin

Cox: CO oxidizing

DGR: Diversity-generating retroelement

GAF: cGMP-specific phosphodiesterases, adenylyl cyclases and FhlA

IS: Insertion sequence

RT: Reverse transcriptase

RVP: Remote variable protein

STK: Serine/threonine kinase

S/T: Serine/Threonine

TIR: Toll/interleukin receptor

TPR: Tetratricopeptide repeat

TR: Template region

VP: Variable protein

VR: Variable region

VWA: Von Willebrand factor type A

Availability of Data and Materials

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

Competing Interest Statement

J.F.M. is a cofounder, equity holder and a member of the Board of Directors of Pylum Biosciences, Inc., a biotherapeutics company in South San Francisco, CA, USA.

Funding

This research was funded by a Challenge Grants from the California NanoSystems Institute (CNSI-UCSB) and the US National Science Foundation – NSF OCE-1635562. 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.

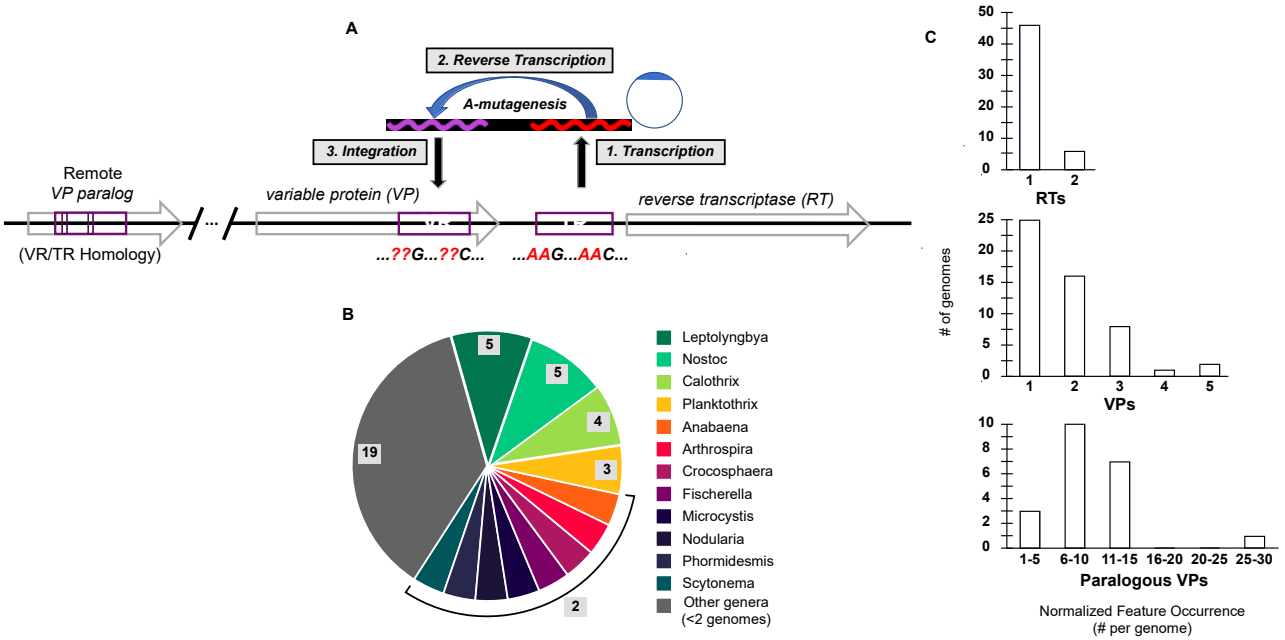
Authors' Contributions

AVE and BP designed the study and carried out comparative genomic analyses of retroelements in cyanobacteria. EA conducted phylogenomic analysis for cyanobacterial genomes. SM, SR, KD, SR participated in clustering and annotation of the retroelements. AVE, DV, and BP wrote the manuscript. All authors read and approved the manuscript.

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.



Figures

Figure 1. Schematic overview of a DGR and their 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.

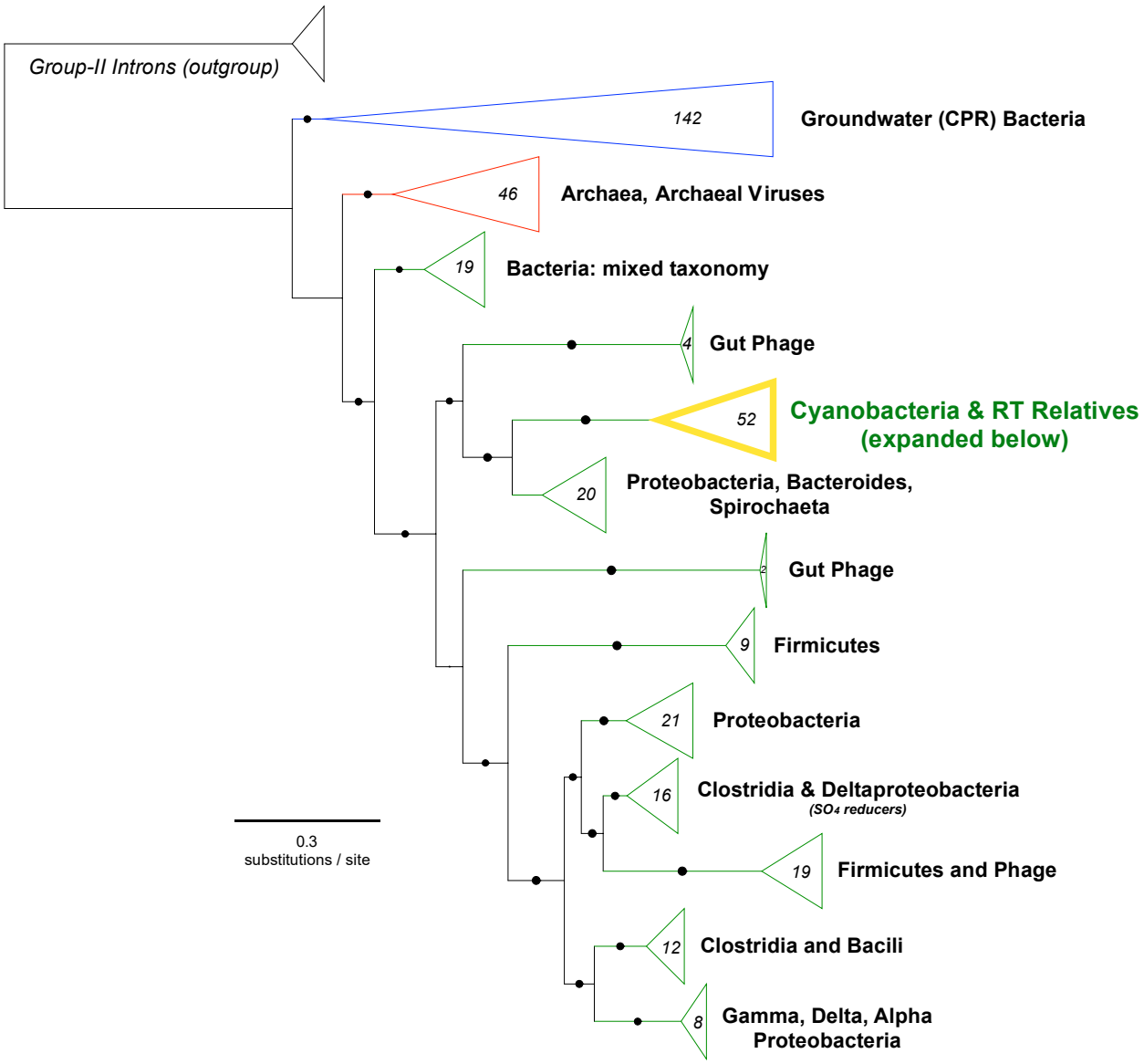


Figure 2. Broad RT phylogeny compared amongst all known DGR-containing lineages.

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

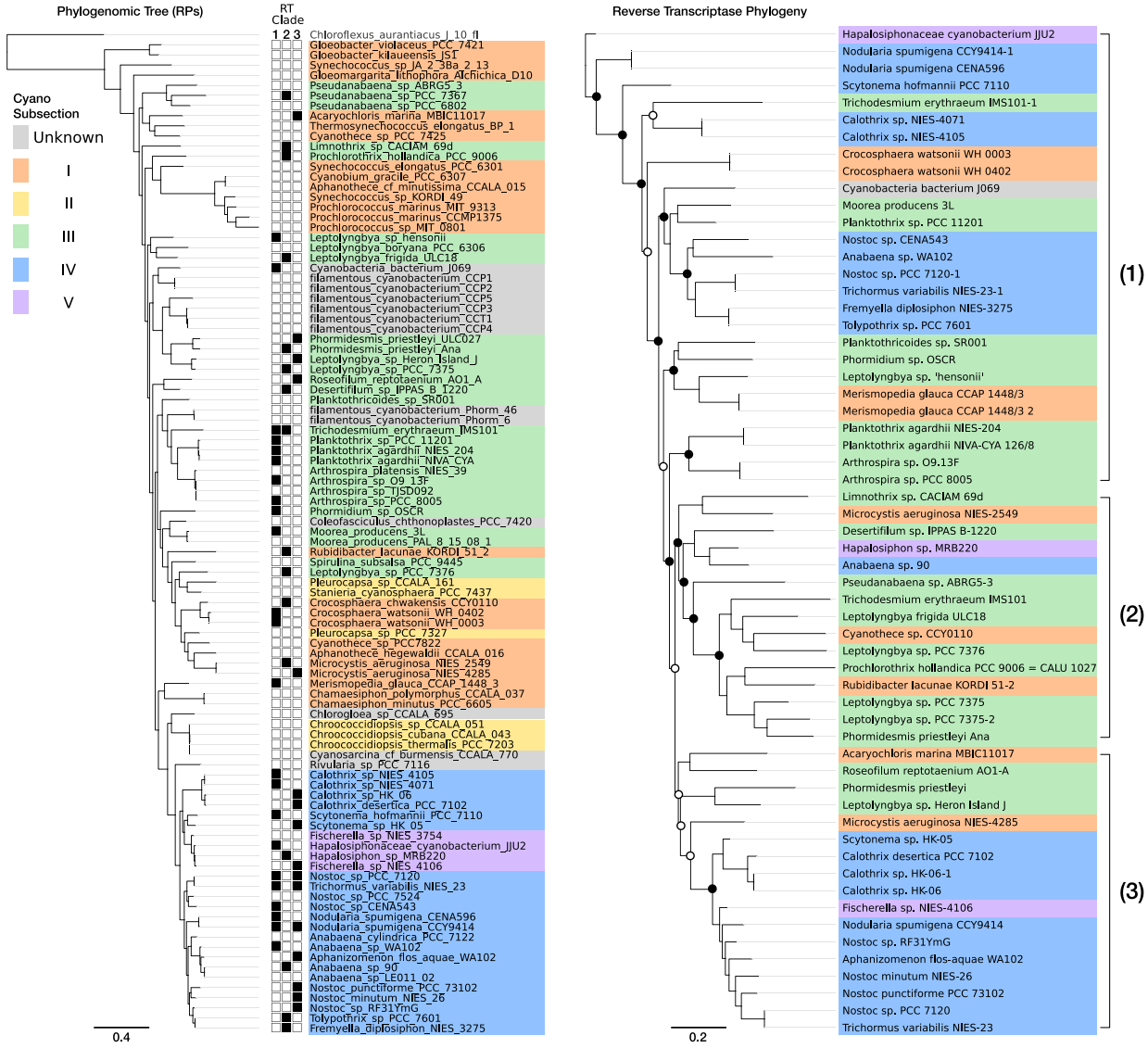


Figure 3. Phylogenetic reconstruction for DGR-containing cyanobacteria and DGR-RT phylogeny.

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. RT clades 1 to 3 were defined based on basal branch support. **b** DGR-RT phylogeny with cyanobacterial physiological subsections highlighted in color. Circles indicate branch support values (hollow >50%; filled >70%).

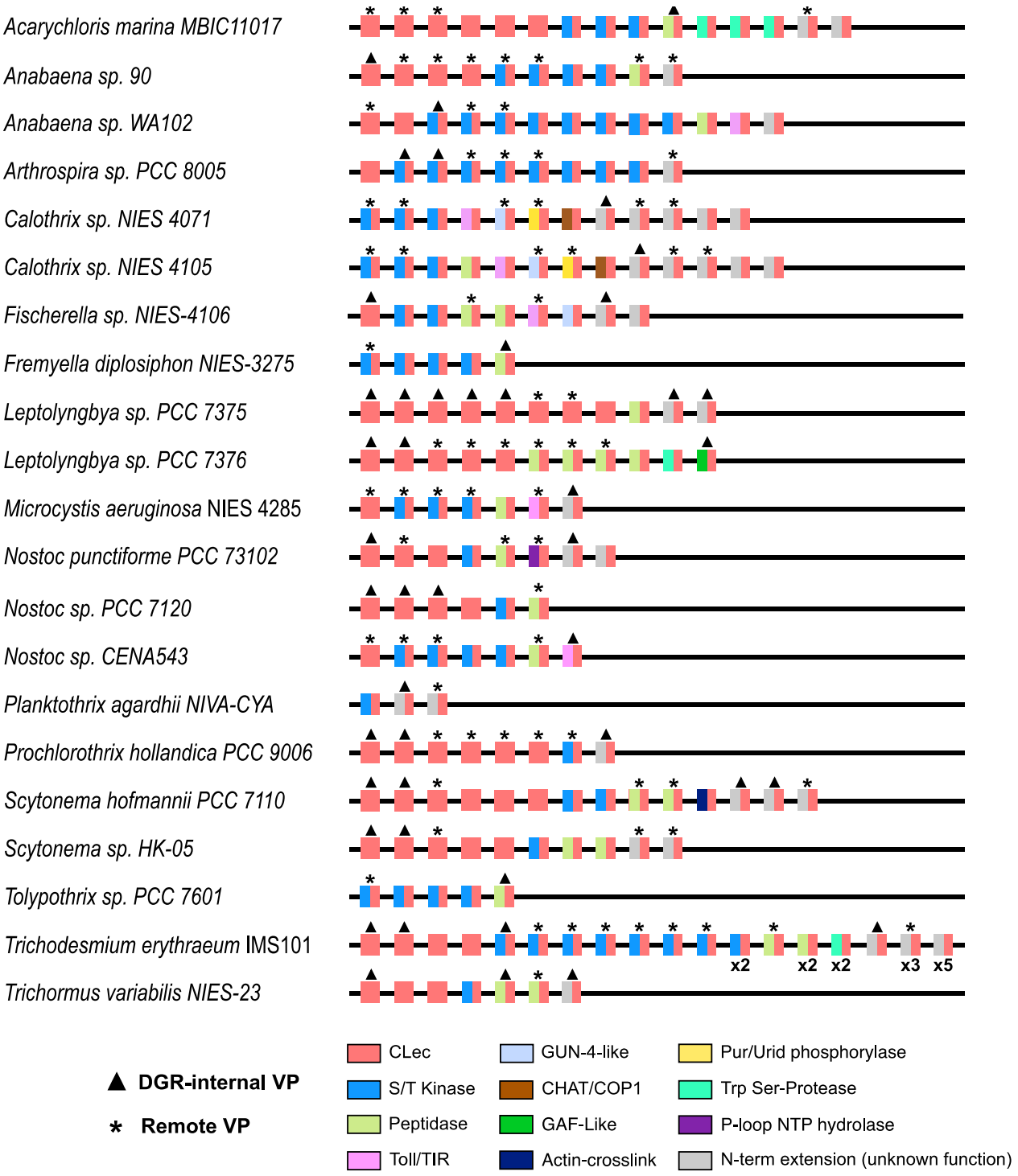


Figure 4. Distribution of CLec variable protein paralogs in cyanobacterial genomes.

DGR-internal 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 DGR-internal variable protein (VP) domain architectures.
Representative domain architectures for DGR-internal variable proteins. 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.

Methods

DGR Identification and Annotation

First, we identified all cyanobacterial genomes containing a DGR-RT-like coding sequence by comparing a consensus sequence for previously-identified cyanobacterial DGR-RT sequences against protein databases using pHMMER. All matches were linked to corresponding genome or nucleotide sequences, which were then downloaded from NCBI. A set of potential DGR candidates was first developed using a workflow with Python and Geneious Prime v 2019.2.3 (Biomatters) as previously described [21]. Briefly, RT genes were manually inspected for core NTP-binding site motifs, before searching for near-repeats in a 10-kbp proximal region (i.e., RT +/- 5-kbp). Repeats in this region were then aligned and inspected for: i) random mismatches in one sequence (VR), which predominantly occur in 1st and 2nd codon positions of an ORF, and ii) >80% of mismatches correspond to adenines in the non-coding near-repeat sequence (TR). Next, retroelements were further analyzed using myDGR [63] which is especially effective at identifying putative trans-acting accessory DGR components, and separately, remote VP and VP homolog genes.

The entire DGR dataset contains several RT and VP sequences that are near-identical, but shared by distinct genomes (Additional file 1: Table S1). To generate a representative subset of these redundant DGRs, we used CD-HIT [64] to cluster RT amino acid sequences using the following settings: 0.9 global alignment; 95% identity threshold. For comparison with the global metagenomic DGR dataset developed by Roux et al. [18], we conducted pairwise alignments with RT sequences using blastp [65] and identified similar representatives at ≥ 50% amino acid identity.

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

Neighborhood Analyses

In order to identify potential transposons, we first examined existing genomic annotations in the neighborhood (i.e. +/- 10 kbp) of each VP and Remote VP for the following features: transposase, integrase, mobile element. Next, we conducted a transposon search using ISFinder [66] using expanded VP loci (60 kbp) that contain one or more annotations associated with mobile elements.

Phylogenetic Analyses

To construct a phylogenetic tree of cyanobacteria, we used a set of 16 ribosomal proteins often used for phylogenomic analysis (RpL2, 3, 4, 5, 6, 14, 15, 16, 18, 22, and 24, and RpS3, 8, 10, 17, 19 [67]. Each ribosomal protein was identified using HMMER [68] and hidden Markov models from the Pfam [69] database (accessed September 2018). Each individual marker gene was aligned using MUSCLE [70], trimmed using TrimAL [71], manually assessed to remove end gaps and ambiguously aligned regions and concatenated. A maximum likelihood tree was constructed using RAxML v. 8.2.9 [72] with the PROTCATLG model.

To reconstruct RT phylogeny, putative DGR-RT coding sequences were identified, as described above, then translated. Sequences were de-replicated and non-redundant candidates were chosen using CD-Hit [64] with a global alignment threshold of 99% identity. All DGR-RT sequences and a set of Group-II intron RT sequences from Bacteria, Archaea, plastids, and mitochondria were aligned with a hidden markov model of the reverse transcriptase protein family (PF00078) using HMMalign [68]. A phylogenetic tree of DGR-RTs was constructed using FastTree2 [73] with the WAG substitution matrix, and the CAT approximation to optimize branch lengths. The cyanobacterial DGR-RT representatives were extracted from the complete alignment, realigned using Clustal Omega [74] and used to construct an unrooted phylogenetic tree.

Protein Function Analysis

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

Supplementary Information

Additional File 1: Table S1. RT/Species Table.
Summary of all DGR-RTs found in cyanobacterial genomes with taxonomic and known physiological information noted. Domain annotations for all DGR-internal VPs and

Remote VPs identified as having VR/TR homology are also summarized. 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 [14].

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) [75] based on the Genome Taxonomy Database (GTDB) [76].

References

1. Sagan L. On the origin of mitosing cells. J Theor Biol. 1967;14.

2. Giovannoni SJ, Turner S, Olsen GJ, Barns S, Lane DJ, Pace NR. Evolutionary relationships among cyanobacteria and green chloroplasts. J Bacteriol. 1988;170:3584–92.

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

4. Mullineaux CW, Mariscal V, Nenninger A, Khanum H, Herrero A, Flores E, et al. Mechanism of intercellular molecular exchange in heterocyst-forming cyanobacteria. EMBO J. 2008;27:1299–308.

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

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

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

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

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

10. Wiltbank LB, Kehoe DM. Diverse light responses of cyanobacteria mediated by phytochrome superfamily photoreceptors. Nature Reviews Microbiology. 2019;17:37–50.

11. Jiang Q, Qin S, Wu Q. Genome-wide comparative analysis of metacaspases in unicellular and filamentous cyanobacteria. BMC Genomics. 2010;11:198.

12. Asplund-Samuelsson J, Sundh J, Dupont CL, Allen AE, McCrow JP, Celepli NA, et al. Diversity and expression of bacterial metacaspases in an aquatic ecosystem. Front Microbiol. 2016;7:1043.

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

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

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

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

17. Wu L, Gingery M, Abebe M, Arambula D, Czornyj E, Handa S, et al. Diversity-generating retroelements: natural variation, classification and evolution inferred from a large-scale genomic survey. Nucleic Acids Res. 2018;46:11–24.

18. Pfreundt U, Kopf M, Belkin N, Berman-Frank I, Hess WR. The primary transcriptome of the marine diazotroph Trichodesmium erythraeum IMS101. Sci Rep. 2015;4:6187.

19. Miller JL, Coq J Le, Hodes A, Barbalat R, Miller JF, Ghosh P. Selective Ligand Recognition by a Diversity-Generating Retroelement Variable Protein. PLoS Biol. 2008;6:e131.

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

21. Paul BG, Burstein D, Castelle CJ, Handa S, Arambula D, Czornyj E, et al. Retroelement-guided protein diversification abounds in vast lineages of Bacteria and Archaea. Nat Microbiol. 2017;2:17045.

22. Guo H, Arambula L, Ghosh P, Miller JF. Diversity-generating Retroelements in Phage and Bacterial Genomes. In: Mobile DNA III. Washington, DC, USA: ASM Press; 2015. p. 1237–52.

23. Naorem SS, Han J, Wang S, Lee WR, Heng X, Miller JF, et al. DGR mutagenic transposition occurs via hypermutagenic reverse transcription primed by nicked template RNA. Proc Natl Acad Sci U S A. 2017;114:E10187–95.

24. Liu M, Deora R, Doulatov SR, Gingery M, Eiserling FA, Preston A, et al. Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. Science. 2002;295:2091–4.

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

26. Roux S, Paul BG, Bagby SC, Allen MA, Attwood G, Cavicchioli R, et al. Ecology and molecular targets of hypermutation in the global microbiome.

27. Schillinger T, Zingler N. The low incidence of diversity-generating retroelements in sequenced genomes. Mob Genet Elem. 2012;2:287–91.

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

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

610 30. Kopf M, Möke F, Bauwe H, Hess WR, Hagemann M. Expression profiling of the bloom-
611 forming cyanobacterium Nodularia CCY9414 under light and oxidative stress conditions. ISME
612 J. 2015;9:2139–52.

613 31. Voß B, Bolhuis H, Fewer DP, Kopf M, Möke F, Haas F, et al. Insights into the Physiology
614 and Ecology of the Brackish-Water-Adapted Cyanobacterium Nodularia spumigena CCY9414
615 Based on a Genome-Transcriptome Analysis. PLoS ONE. 2013;8:e60224.

616 32. Hoving JC, Wilson GJ, Brown GD. Signalling C-type lectin receptors, microbial recognition
617 and immunity. Cellular Microbiology. 2014;16:185–94.

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

620 34. Zelensky AN, Gready JE. The C-type lectin-like domain superfamily. FEBS J.
621 2005;272:6179–217.

622 35. Guo H, Tse L V., Nieh AW, Czornyj E, Williams S, Oukil S, et al. Target Site Recognition
623 by a Diversity-Generating Retroelement. PLoS Genet. 2011;7:e1002414.

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

627 37. He S, Corneloup A, Guynet C, Lavatine L, Caumont-Sarcos A, Siguier P, et al. The
628 IS200/IS605 Family and "Peel and Paste" Single-strand Transposition Mechanism. Microbiol
629 Spectr. 2015;3:609-30.

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

632 39. Stock AM, Robinson VL, Goudreau PN. Two-Component Signal Transduction. Annu Rev
633 Biochem. 2000;69:183–215.

634 40. Janczarek M, Vinardell J-M, Lipa P, Karaś M. Hanks-Type Serine/Threonine Protein
635 Kinases and Phosphatases in Bacteria: Roles in Signaling and Adaptation to Various
636 Environments. Int J Mol Sci. 2018;19:2872.

637 41. Libby EA, Goss LA, Dworkin J. The Eukaryotic-Like Ser/Thr Kinase PrkC Regulates the
638 Essential WalRK Two-Component System in Bacillus subtilis. PLoS Genet. 2015;11:e1005275.

639 42. Mijakovic I, Macek B. Impact of phosphoproteomics on studies of bacterial physiology.
640 FEMS Microbiology Reviews. 2012;36:877–92.

641 43. Dworkin J. Ser/Thr phosphorylation as a regulatory mechanism in bacteria. Current Opinion
642 in Microbiology. 2015;24:47–52.

643 44. Shi L, Pigeonneau N, Ravikumar V, Dobrinic P, Macek B, Franjevic D, et al. Cross-
644 phosphorylation of bacterial serine/threonine and tyrosine protein kinases on key regulatory
645 residues. Front Microbiol. 2014;5:495.

646 45. Shi L, Ji B, Kolar-Znika L, Boskovic A, Jadeau F, Combet C, et al. Evolution of Bacterial
647 Protein-Tyrosine Kinases and Their Relaxed Specificity Toward Substrates. Genome Biol Evol.
648 2014;6:800–17.

649 46. Capra EJ, Perchuk BS, Skerker JM, Laub MT. Adaptive mutations that prevent crosstalk
650 enable the expansion of paralogous signaling protein families. Cell. 2012;150:222–32.

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

653 48. Klemenčič M, Novinec M, Dolinar M. Orthocaspases are proteolytically active prokaryotic
654 caspase homologues: The case of Microcystis aeruginosa. Mol Microbiol. 2015;98:142-50.

655 49. Spungin D, Bidle KD, Berman-Frank I. Metacaspase involvement in programmed cell death
656 of the marine cyanobacterium Trichodesmium. Environ Microbiol. 2018;21:667-81.

657 50. Asplund-Samuelsson J. The art of destruction: revealing the proteolytic capacity of bacterial
658 caspase homologs. Mol Microbiol. 2015;98:1–6.

659 51. Kenneth Allan R, Ratajczak T. Versatile TPR domains accommodate different modes of
660 target protein recognition and function. Cell Stress and Chaperones. 2011;16:353–67.

661 52. van der Voorn L, Ploegh HL. The WD-40 repeat. FEBS Lett. 1992;307:131–4.

662 53. Tewari R, Bailes E, Bunting KA, Coates JC. Armadillo-repeat protein functions: Questions
663 for little creatures. Trends in Cell Biology. 2010;20:470–81.

664 54. Colombatti A, Bonaldo P, Doliana R. Type A Modules: Interacting Domains Found in
665 Several Non-Fibrillar Collagens and in Other Extracellular Matrix Proteins. Matrix.
666 1993;13:297–306.

667 55. Swiderski MR, Birker D, Jones JDG. The TIR domain of TIR-NB-LRR resistance proteins is
668 a signaling domain involved in cell death induction. Mol Plant Microbe Interact. 2009;22:157–
669 65.

670 56. Spear AM, Loman NJ, Atkins HS, Pallen MJ. Microbial TIR domains: not necessarily agents
671 of subversion? Trends Microbiol. 2009;17:393–8.

672 57. Ho YS, Burden LM, Hurley JH. Structure of the GAF domain, a ubiquitous signaling motif
673 and a new class of cyclic GMP receptor. EMBO J. 2000;19:5288–99.

674 58. Sobotka R, Dühring U, Komenda J, Peter E, Gardian Z, Tichy M, et al. Importance of the
675 cyanobacterial Gun4 protein for chlorophyll metabolism and assembly of photosynthetic
676 complexes. J Biol Chem. 2008;283:25794–802.

677 59. Aravind L, Koonin E V. Classification of the caspase-hemoglobinase fold: Detection of new
678 families and implications for the origin of the eukaryotic separins. Proteins Struct Funct Genet.
679 2002;46:355–67.

680 60. Ohmori M, Ikeuchi M, Sato N, Wolk P, Kaneko T, Ogawa T, et al. Characterization of Genes
681 Encoding Multi-domain Proteins in the Genome of the Filamentous Nitrogen-fixing
682 Cyanobacterium Anabaena sp. Strain PCC 7120. 2001;8:271-84.

683 61. Asplund-Samuelsson J, Bergman B, Larsson J. Prokaryotic Caspase Homologs: Phylogenetic
684 Patterns and Functional Characteristics Reveal Considerable Diversity. PLoS ONE.
685 2012;7:e49888.

686 62. Mammen M, Choi S-K, Whitesides GM. Polyvalent Interactions in Biological Systems:
687 Implications for Design and Use of Multivalent Ligands and Inhibitors. Angew Chem Int Ed.
688 1998;37:2754–94.

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

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

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

695 66. Siguier P. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res.
696 2006;34:D32–6.

697 67. Hug LA, Castelle CJ, Wrighton KC, Thomas BC, Sharon I, Frischkorn KR, et al. Community
698 genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and
699 indicate roles in sediment carbon cycling. Microbiome. 2013;1:1–17.

700 68. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity
701 searching. Nucleic Acids Res. 2011;39 Web Server issue:W29–37.

702 69. Bateman A. The Pfam Protein Families Database. Nucleic Acids Res. 2002;30:276–80.

703 70. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput.
704 Nucleic Acids Res. 2004;32:1792–7.

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

707 72. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
708 phylogenies. Bioinformatics. 2014;30:1312–3.

709 73. Price MN, Dehal PS, Arkin AP. FastTree 2 - Approximately maximum-likelihood trees for
710 large alignments. PLoS ONE. 2010;5:e9490.

711 74. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of
712 high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol.
713 2011;7:539.

714 75. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify
715 genomes with the Genome Taxonomy Database. Bioinformatics. 2019;36:1925–7.

716 76. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A
717 standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of
718 life. Nat Biotechnol. 2018;36:996–1004.