Exploring Viral Communities Associated With Terrestrial Cyanobacteria Metagenomes

Cassandra L. Ettinger[a,#], Sudipto R. Paul[b], Neptali Flores[c], Ryan D. Ward[d], Nicole Pietrasiak[e], Jason E. Stajich[a,f,#]

[a]Department of Microbiology and Plant Pathology, University of California, Riverside, Riverside, California, USA.

[b]Department of Biochemistry, University of California, Riverside, Riverside, California, USA.

[c]Department of Biological Sciences, San Jose State University, San Jose, California, USA.

[d]Laboratory of Genetics, University of Wisconsin, Madison, Madison, Wisconsin, USA.

[e]School of Life Sciences, University of Nevada, Las Vegas, Las Vegas, Nevada, USA.

[f]Institute for Integrative Genome Biology, University of California, Riverside, Riverside, CA, USA.

Running Title: Terrestrial Cyanophage Survey

[#]Address correspondence to: Cassandra L. Ettinger, cassande@ucr.edu; Jason E. Stajich, jason.stajich@ucr.edu

Keywords: cyanophage, virus, phage, viral diversity, viral host, terrestrial, Cyanobacteria, desert ecosystem

Abstract

Viruses are exceedingly common, but little is known about how they behave in extreme environments, and whether viruses facilitate adaptation of their hosts to harsh conditions. To understanding of these understudied viral-host interactions, we analyzed metagenomes from 50 unialgal but nonaxenic Cyanobacteria cultures with 47 representing various terrestrial habitats, including desert soil and rock surfaces, tropical soil, and vernal pools. These cultures represent low diversity microbial consortia dominated by the terrestrial Cyanobacteria and its associated cyanosphere microbiome containing heterotrophic microbes. We identified viral sequences in metagenomes, grouped these into viral operational taxonomic units (vOTUs) and then placed vOTUs into viral clusters (VCs). We also calculated vOTU relative abundance and predicted possible bacterial hosts In total we predicted 814 viral sequences representing 726 vOTUs. We assigned putative taxonomy to 72 of the 814 putative viral sequences; these were distributed into 15 VCs — mostly assigned to the Caudovirales order of viruses. We found that metagenomes were dominated by unclassified and unclustered viral sequences. Furthermore, we predicted bacterial hosts for 211 vOTUs, with the majority of viruses predicted to infect a Proteobacteria host, and only 3 vOTUs predicted to infect a Cyanobacteria host. Overall, these results are consistent with the notion that viruses in extreme environments are underrepresented in reference datasets. This work increases knowledge of viral diversity and sets a foundation for future work on viruses associated with terrestrial Cyanobacteria and their heterotroph associates, such as connecting specific viruses to critical cycling processes and investigating their metabolic functions.

Importance

Relatively little is known about Cyanobacteria that live on land, or their associated viral communities. To begin to address this knowledge gap, we identified viruses from metagenomes of multi-species cultures which contain mainly Cyanobacteria, but also associated heterotrophic

bacteria. The majority of surveyed viruses did not share similarity with reference viral genomes, potentially indicating some viral novelty. Further, we found that for the viruses where we could predict a putative bacterial host, most were not associated with the Cyanobacteria members in these cultures. These results highlight how little is known about land-based Cyanobacteria or their associated viruses. Given the role of some viruses in influencing ecosystem services and enhancing host fitness, surveying viruses is foundational to understanding how Cyanobacteria may have adapted to and survive in some of the most extreme environments on land.

Introduction

Viruses infect all domains of life and are thought to be the most abundant entities on Earth (1). Viruses rely on their hosts for replication, and thus are intricately entwined with their host communities. In the lytic life cycle viruses infect their hosts, hijack the cellular machinery to replicate, and induce host cell lysis for dispersal of the newly created viral particles (2). As a host defense mechanism, phage infections can be evidenced in host's CRISPR cassettes which resulted from past infection events building up hosts' phage immunity (3, 4). Frequently viruses can also integrate and persist in their host's genome as prophages, with many bacteria harboring at least one prophage (5). Prophages can have auxiliary metabolic genes (AMGs) or pathogenicity genes which they use to promote host growth (6–9). With such controls on host metabolism and fitness, viruses can have important and often underappreciated effects on bacterial population dynamics and turnover in microbial communities (10–12).

Viral discovery has long been hindered by the lack of universal gene markers, however, modern machine-learning based and k-mer approaches have enabled researchers to robustly search metagenomes for viral sequences (13–17). With these new methodologies, researchers can utilize available public sequencing data to discover and explore complex viral communities across previously uncharted ecosystems. However, virus discovery and potential host interactions and feedbacks to microbial communities in extreme environments, such as cave wall biofilms, biofilms of ephemeral waterfalls, temporal vernal pools, and dryland soil, are still notably understudied.

Cyanobacteria are keystone taxa of such polyextreme terrestrial environments due to their important roles in primary production and microbial organic matter build up, nitrogen fixation, bioweathering, and biofilm, microbial mat or biological soil crust formation (18–21). As essential ecosystem engineers, Cyanobacteria interact with a broad variety of other microbes. For

138 examples, filamentous biological soil crust cyanobacteria house their own set of associated
139 microbes established around the polysaccharide rich sheath material forming the cyanosphere, a
140 microbial community similar to the copiotrophic rhizosphere microbiomes around plant roots (22).
141 Cyanobacteria also influence free-living co-occurring microbes as shown in another study in which
142 cyanobacteria in hypersaline microbial mats generated fermentation byproducts that promoted
143 the metabolism of Chloroflexi and Sulfate-reducing bacteria (23). However, our knowledge on
144 Cyanobacteria phages, viruses that infect Cyanobacteria, is still largely based on prior
145 investigations of samples collected from aquatic ecosystems (24–28). Cyanophages in aquatic
146 ecosystems have shown to carry AMGs related to photosynthesis, phosphorus utilization, and
147 carbon metabolism which are thought to aid phage fitness (29–32). Yet, if Cyanophages have
148 similar dynamic roles in polyextreme terrestrial habitats, or how phages might impact
149 Cyanobacteria associated microbiomes is still virtually unknown.
150
151 Previously, Ward et al. generated 50 metagenomes from nonaxenic Cyanobacterial cultures,
152 which represented unialgal polycultures containing primarily Cyanobacteria and heterotrophic
153 microbial associates living in the cyanosphere. The majority of the metagenomes (47) were
154 obtained from cyanobacteria isolated from terrestrial habitats (24). In this study, we investigated
155 the viral diversity associated with these Cyanobacterial co-culture metagenomes. We sought to
156 (i) create a catalog of viral genomes present in these metagenomes, (ii) assess the relative
157 abundance of viral genomes across co-cultures, and (iii) explore possible bacterial hosts for these
158 viruses.
159
160 **Results and Discussion**
161 *Successful viral identification from Cyanobacterial co-cultures*
162 In total, 814 viral sequences were identified across the 50 cyanobacterial co-culture
163 metagenomes (Table S1), with each metagenome having on average 16.3 viral sequences

164 (range: 1-54). Of the 814 viral sequences, 188 were integrated prophages (Fig. 1A). CheckV
165 assigned viral quality to 712 of the 814 viral genomes with 465 predicted to be low-quality, 208 to
166 be medium-quality, 32 to be high-quality, and 7 to be complete (Fig. 1B). The mean length of
167 predicted viral sequences was 27,594 bp (range: 5,013–199,914 bp), with an average of 35.4
168 predicted genes per viral sequence. We clustered these 814 viral sequences at 95% identity into
169 726 viral operational taxonomic units (vOTUs). These numbers are similar to previous
170 metagenomic viral discovery efforts in cyanolichen metagenomes (33). While this work provides
171 a catalog of the dominant DNA viruses that infect terrestrial Cyanobacteria and associated
172 heterotrophs, it is likely that we are undersampling rare viruses which are better profiled using
173 virome data (34).
174
175 *Clustering of viral sequences with reference data reveals possibly novel genera*
176 To be able to assign putative taxonomy to vOTUs, we further used gene-sharing networks to
177 cluster them with publicly available reference data to form VCs which approximate genus-level
178 groupings. In total 305 vOTUs, representing 341 viral sequences, were placed in 111 VCs. Of
179 these 111 VCs, 26 included representatives from the reference dataset, while 79 represented
180 putatively novel genera. The majority of these 79 unique VCs originated from Cyanobacterial co-
181 cultures isolated from desert environments, which is not unexpected as the majority of co-cultures
182 were from these environments (56% of co-cultures) (Fig. 2A-B). This limited overlap with existing
183 reference data is similar to recent work on cyanolichen metagenomes, where only 28 of 133 VCs
184 included reference genomes (33). As viral exploration of cyanobacteria-dominated and other
185 extreme environments continues, we expect to see more sequences clustering with reference
186 genomes over time.
187
188 The 26 VCs that clustered with the reference data represented 75 of the 814 predicted viral
189 sequences in this dataset. Of these 26 VCs, the majority of these (15 VCs) were clustered with

190 members assigned to the Caudovirales order of tailed double-stranded DNA viruses, which have
191 also been observed to be the dominant viruses in cyanolichens, biocrusts, hypoliths, and aquatic
192 cyanobacteria-dominated metagenomes (33, 35–42). Within this, VCs were further classified as
193 belonging to the families Autographiviridae, Inoviridae, Microviridae, Myoviridae, Podoviridae, and
194 Siphoviridae (Fig. 2C-D). Only 15 of the 39 sequences representing high-quality or complete viral
195 genomes, which were made up of 14 vOTUS, came together to form 13 VCs. Of these, only five
196 VCs included representative viruses from the reference data, while 8 VCs were potentially novel
197 genera. Of the VCs that clustered with reference genomes, three were clustered with members
198 of the order Caudovirales, one with the order Petitvirales, and one with the order Tubulavirales.
199
200 *Most abundant vOTUs are unclassified or unclustered*
201 While the number of viral sequences recovered from each cyanobacterial co-culture varied (Fig.
202 3A), we were also curious about the distribution of viral sequences across co-cultures. Thus, we
203 mapped metagenomic reads from each sample to the entire catalog of vOTUs to calculate their
204 relative abundance across all co-cultures. Of 726 vOTUs, only 617 vOTUs had quality scores
205 from CheckV and were detected above thresholds set for inclusion. Looking at the relative
206 abundance of these remaining 617 vOTUs, we found that the majority of co-cultures were
207 dominated by unclustered vOTUs or by vOTUs that formed VCs that did not cluster with
208 reference genomes (Fig. 3B). Only a small portion, on average 8.2% (range: 0 - 100%), of
209 vOTU relative abundance in samples was from vOTUs in VCs that included reference genomes.
210 The majority of vOTUs, 496 (80.4%), were singletons and only found in one sample. While a
211 total of 121 vOTUs and 86 VCs (13 of which clustered with reference data) were detected in two
212 or more samples. The most prevalent VC (714_0) did not cluster with reference genomes and
213 represented 25 vOTUS across 26 samples spanning different habitats and collection locations.
214 Using ViPTree, we found that vOTUs in this VC form a unique monophyletic clade within the
215 family, Siphoviridae (Fig. S1).

216

217 *Proteobacteria, not Cyanobacteria, dominate host predictions*

218 For the 617 vOTUs, we were able to predict possible bacterial hosts for 211 vOTUs to at least
219 the phylum-level (Fig. 3C). Of these predictions, we were able to classify the hosts for 151 to
220 genus-level. Thus, the majority of viral sequences (65.8%) had no host prediction. For those
221 viruses with a host prediction, largely these were projected to infect Proteobacterial hosts
222 (95.3%). Within the Proteobacteria, almost half (47.3%) of the predicted hosts were members of
223 the order Hyphomicrobiales. Further, 18 of the 25 vOTUs represented in the most shared VC
224 (714_0) were predicted to infect members of the order Hyphomicrobiales. Given that some
225 members of the Hyphomicrobiales fix nitrogen and often form symbiotic relationships (e.g.
226 *Rhizobium (43)*), we wonder if bacteria in this order may be acting in synergy with the
227 Cyanobacteria in these nonaxenic cultures.

228

229 Only 3 vOTUs (vOTU602, vOTU638, vOTU669), representing 3 unclassified viral sequences,
230 had Cyanobacterial host predictions. Of these two were predicted to be medium-quality
231 (vOTU602, vOTU669) and one complete (vOTU638). Two of the three vOTUs formed separate
232 VCs unique to this data (vOTU602 = VC 1342.0; vOTU638 = VC 1447.0). All three vOTUs were
233 predicted to have hosts in the order Nostocales. However for two vOTUs we were able to only
234 predict hosts to the order-level (vOTU602, vOTU638), while the third vOTU was predicted to
235 infect the genus *Calothrix* (vOTU669). These predictions match the taxonomy of the
236 Cyanobacterial member of the samples that they were obtained from. Two of these vOTUs
237 (vOTU602, vOTU669) originated from samples from desert environments, while the third
238 originated from a sample from tropical soil (vOTU638). Each of these three vOTUs was only
239 detected in the sample it originated from. Using ViPTree, we found that these vOTUs were
240 placed on long branches with limited similarity to reference genomes and that while some of the
241 reference genomes in the tree did have aquatic Cyanobacterial host predictions, none were

close relatives to the three vOTUs here (Fig. S2). While finding only three cyanophage vOTUs was surprising given the predominance of Cyanobacteria in these cultures, previous work from biocrusts and hypoliths have similarly reported a low abundance of cyanophage (40, 42). An important caveat to remember is the reliance of host predictions on available genomes; as Cyanobacterial genome availability increases, we expect that ability to link viruses to their hosts in these ecosystems will increase as well.

Conclusion

To expand knowledge of Cyanobacteria viruses largely from terrestrial environments, we employed viral discovery methods on existing metagenomes from nonaxenic Cyanobacterial cultures (24). We were able to identify 814 viral sequences including sequences that formed unique VC and which may represent novel viral genera. While this work provides a catalog of viruses that infect terrestrial Cyanobacteria and associated heterotrophs, due to methodological limitations, we are likely only sampling DNA phage and are undersampling rare viruses in these samples. Future work should seek to profile these cultures using more comprehensive methods such as through viromics sequencing. While we observed no wide scale habitat or Cyanobacterial host patterns driving viral diversity, it is possible that this is due to incomplete sampling of the viral community or the mixed nature of the nonaxenic cultures. Overall, future studies should include a more extensive investigation of viral host interactions in extreme environments and profile the auxiliary metabolic genes of viral genomes which may play roles in various cycling processes in terrestrial Cyanobacteria-dominated environments.

Methods

Data collection

266 Publicly available metagenomic samples were downloaded from NCBI GenBank and imported
267 to the KBase platform using their respective SRA numbers (Table S1) (44). Sample data
268 associated with each metagenome was obtained from Table 1 of Ward et al (24).
269
270 *Identification of viral contigs*
271 Initial viral analysis was performed following the KBase "Viral Analysis End-to-End" workflow
272 (https://kbase.us/n/75811/85/). Sequence quality for each sample was assessed in KBase using
273 FastQC v0.11.5 (45). Afterwards, low-quality regions and residual Illumina adapters were trimmed
274 using Trimmomatic v0.36 (46). This step was followed by a reassessment of the quality of the
275 sequence data again using FastQC. Next, the reads were assembled into contigs with
276 MetaSPAdes v3.13.0 with k-mer sizes 21, 33, 55, 77, 99 and 127 (47). Three samples failed to
277 assemble at this stage using MetaSPAdes due to memory constraints in KBase; for these three
278 samples, MEGAHIT v1.2.9 was utilized instead (48). Subsequently, all contigs shorter than 5000
279 base pairs were filtered out using the KBase application "Filter Assembled Contigs By Length -
280 v1.1.2".
281
282 Next, VirSorter v1.0.5 was used to predict the viral genomes in each sample and categorize them
283 into six categories based on their predicted length and viral characteristics (49). Viral contigs from
284 VirSorter categories 1-2 (high-confidence lytic viruses) and 4-5 (high-confidence prophages) were
285 merged into one file using the KBase application "Modify Bins in BinnedContigs - v1.0.2" and then
286 extracted using the "Extract Bins as Assemblies from BinnedContigs - v1.0.2" application. The
287 assembly file for each sample was then exported from KBase. All assemblies were combined into
288 a single file and then run through VIBRANT v1.2.1 in "virome" mode (14). This second viral
289 identification step was performed to ensure proper trimming of prophage. CheckV v0.8.1 was then
290 run on the resulting file of predicted viral genomes using the "end_to_end" workflow to calculate
291 viral genome quality (50).

292

293 *Clustering of viral contigs across all samples & relative abundance*

294 Viral genomes were then clustered into 95% identity vOTUs using CD-HIT using the following
295 options: -c 0.95 -aS 0.85 -M 0 -d 0 (51). In order to assign a putative taxonomy to vOTUs, we
296 first used Prodigal v. 2.6.3 to predict open reading frames in vOTU representative genomes
297 using the -p meta option (52). We then used VContact2 to cluster vOTUs with the INPHARED
298 August 2022 reference database into viral clusters (VCs) (53, 54) and assigned predicted
299 taxonomy to viruses based on cluster membership as in Santos-Medellin et al (34). For vOTUs
300 in VCs of interest, ViPTree was further used to place vOTUs into proteomic trees with viral
301 reference genomes from RefSeq release 212 (55).

302

303 To quantify the relative abundance of vOTUs across samples, we mapped reads from each
304 metagenome to vOTUs using BBMap with a minid=0.90 (56). We used SAMtools to convert
305 sam files to bam files and BEDTools genomecov to get coverage information for each vOTU
306 across each sample from the bam files (57, 58). We also used bamM to parse the bam files to
307 calculate the trimmed pileup coverage (tpmean), which we used to assess viral relative
308 abundance (59). We then filtered out vOTUs which displayed < 75% coverage over the length of
309 the viral sequence in R (60). Thresholds for analysis of relative abundance of vOTUs were
310 based on previous community guidelines for identity (i.e. ≥ 95% identity) and detection (i.e. ≥
311 75% of the viral genome length covered ≥ 1x by reads at ≥ 90% average nucleotide identity)
312 (61, 62). We further removed vOTUs that CheckV could not make quality predictions for (i.e.
313 where quality was “not-determined”) prior to downstream analyses. Viral relative abundance
314 was visualized in R using many packages including tidyverse and phyloseq (63, 64). Analysis

315 scripts for the post-KBase portions of this work are available on GitHub
316 (https://github.com/stajichlab/Terrestrial_Cyanophage) and archived in Zenodo (65).
317
318 *Putative identification of hosts*
319 In order to predict possible microbial hosts for the viral genomes identified, we downloaded all
320 complete bacterial and archaeal genomes, as of April 7, 2022, from NCBI RefSeq using "ncbi-
321 genome-download" (66). In total this amounted to 25,984 bacterial and 416 archaeal genomes.
322 We then used "makeblastdb" from NCBI BLAST 2.12.0+ to convert these complete genomes
323 into a blast database and "blastn" to compare vOTUs to this reference (67). We filtered the
324 results of this analysis in R based on guidelines used previously (68, 69). Matches had to share
325 ≥ 2000 bp region with ≥ 70% sequence identity to the viral genome and possible host matches
326 needed to have a bit score of ≥ 50 and minimum e value of 0.001. Only the top five hits
327 matching these thresholds were considered when predicting a host for each viral genome, with
328 host predictions made at each taxonomic level only if all hits were in agreement. Discrepancies
329 in taxonomy between hits resulted in a "mixed" assignment for that taxonomic level. Based on
330 this method, the lowest level of host prediction was genus level, with some predictions only able
331 to be made at higher taxonomic levels (e.g. class or order).
332
333 **Acknowledgements**
334 S.P. was supported by UC Riverside's Research in Science Experience (RISE;
335 https://rise.ucr.edu/) program which received funding from UC Riverside College of Natural and
336 Agricultural Sciences and the National Science Foundation (NSF) PERSIST, Award No.
337 1565140. N.F. was supported by the National Summer Undergraduate Research Project
338 (NSURP; https://nsurp.org/) which was supported by NSF RAPID: The National Summer
339 Undergraduate Research Project, Award No. 2128247. J.E.S. is a CIFAR Fellow in the program

Fungal Kingdom: Threats and Opportunities and partially supported by USDA Agriculture Experimental Station at the University of California, Riverside and NIFA Hatch projects CA-R-PPA-5062-H.

We are thankful to DOE Systems Biology Knowledgebase (KBase) which is supported as part of the BER Genomic Science Program, and is funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05-00OR22725, and DE-AC02-98CH10886. Data analyses performed at the High-Performance Computing Cluster at the University of California Riverside in the Institute of Integrative Genome Biology were supported by NSF grant DBI-1429826 and NIH grant S10-OD016290. Metagenomic sequencing of the cyanobacteria cultures was supported by JGI-CSP-503837 awarded to N.P. The work (proposal: 10.46936/10.25585/60001171) conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication. We also thank Jeffrey R. Johansen for being a lifelong mentor, collaborator, and friend to N.P. Jeff initiated a culture collection of terrestrial cyanobacteria at JCU from which subcultures of at least half of the investigated strains were obtained and investigated in Ward et al. (2021).

References

1. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. 2016. Uncovering Earth's virome. Nature 536:425–430.

2. Guttman B, Raya R, Kutter E. 2004. Basic Phage Biology. Bacteriophages

366 https://doi.org/10.1201/9780203491751.ch3.

367 3. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA,
368 Horvath P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes.
369 Science 315:1709–1712.

370 4. Hampton HG, Watson BNJ, Fineran PC. 2020. The arms race between bacteria and their
371 phage foes. Nature 577:327–336.

372 5. Casjens S. 2003. Prophages and bacterial genomics: what have we learned so far? Mol
373 Microbiol 49:277–300.

374 6. Fortier L-C, Sekulovic O. 2013. Importance of prophages to evolution and virulence of
375 bacterial pathogens. Virulence 4:354–365.

376 7. Hurwitz BL, Hallam SJ, Sullivan MB. 2013. Metabolic reprogramming by viruses in the
377 sunlit and dark ocean. Genome Biol 14:R123.

378 8. Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, Chisholm SW. 2011.
379 Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon
380 metabolism. Proc Natl Acad Sci U S A 108:E757–64.

381 9. Bragg JG, Chisholm SW. 2008. Modeling the fitness consequences of a cyanophage-
382 encoded photosynthesis gene. PLoS One 3:e3550.

383 10. Hurwitz BL, Westveld AH, Brum JR, Sullivan MB. 2014. Modeling ecological drivers in
384 marine viral communities using comparative metagenomics and network analyses. Proc
385 Natl Acad Sci U S A 111:10714–10719.

386 11. Crummett LT, Puxty RJ, Weihe C, Marston MF, Martiny JBH. 2016. The genomic content
387 and context of auxiliary metabolic genes in marine cyanomyoviruses. Virology 499:219–

229.

12. Hurwitz BL, U'Ren JM. 2016. Viral metabolic reprogramming in marine ecosystems. Curr Opin Microbiol 31:161–168.

13. Sullivan MB. 2015. Viromes, Not Gene Markers, for Studying Double-Stranded DNA Virus Communities. Journal of Virology https://doi.org/10.1128/jvi.03289-14.

14. Kieft K, Zhou Z, Anantharaman K. 2020. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome 8:90.

15. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, Pratama AA, Gazitúa MC, Vik D, Sullivan MB, Roux S. 2021. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome 9:37.

16. Amgarten D, Braga LPP, da Silva AM, Setubal JC. 2018. MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. Frontiers in Genetics https://doi.org/10.3389/fgene.2018.00304.

17. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. 2017. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome 5:69.

18. Pombubpa N, Pietrasiak N, De Ley P, Stajich JE. 2020. Insights into dryland biocrust microbiome: geography, soil depth and crust type affect biocrust microbial communities and networks in Mojave Desert, USA. FEMS Microbiol Ecol 96.

19. Paerl HW, Pinckney JL, Steppe TF. 2000. Cyanobacterial-bacterial mat consortia: examining the functional unit of microbial survival and growth in extreme environments. Environ Microbiol 2:11–26.

410 20. Perera I, Subashchandrabose SR, Venkateswarlu K, Naidu R, Megharaj M. 2018.
411 Consortia of cyanobacteria/microalgae and bacteria in desert soils: an underexplored
412 microbiota. Appl Microbiol Biotechnol 102:7351–7363.

413 21. Velichko N, Smirnova S, Averina S, Pinevich A. 2021. A survey of Antarctic cyanobacteria.
414 Hydrobiologia 848:2627–2652.

415 22. Couradeau E, Giraldo-Silva A, De Martini F, Garcia-Pichel F. 2019. Spatial segregation of
416 the biological soil crust microbiome around its foundational cyanobacterium, Microcoleus
417 vaginatus, and the formation of a nitrogen-fixing cyanosphere. Microbiome 7:55.

418 23. Lee JZ, Burow LC, Woebken D, Everroad RC, Kubo MD, Spormann AM, Weber PK, Pett-
419 Ridge J, Bebout BM, Hoehler TM. 2014. Fermentation couples Chloroflexi and sulfate-
420 reducing bacteria to Cyanobacteria in hypersaline microbial mats. Front Microbiol 5:61.

421 24. Ward RD, Stajich JE, Johansen JR, Huntemann M, Clum A, Foster B, Foster B, Roux S,
422 Palaniappan K, Varghese N, Mukherjee S, Reddy TBK, Daum C, Copeland A, Chen I-MA,
423 Ivanova NN, Kyrpides NC, Shapiro N, Eloe-Fadrosh EA, Pietrasiak N. 2021. Metagenome
424 Sequencing to Explore Phylogenomics of Terrestrial Cyanobacteria. Microbiol Resour
425 Announc 10:e0025821.

426 25. Pietrasiak N, Reeve S, Osorio-Santos K, Lipson DA, Johansen JR. 2021. *Trichotorquatus*
427 gen. nov. - a new genus of soil cyanobacteria discovered from American drylands. J Phycol
428 57:886–902.

429 26. Pietrasiak N, Osorio-Santos K, Shalygin S, Martin MP, Johansen JR. 2019. When Is A
430 Lineage A Species? A Case Study In *Myxacorys* gen. nov. (Synechococcales:
431 Cyanobacteria) With The Description of Two New Species From The Americas. J Phycol
432 55:976–996.

433 27. Baldarelli LM, Pietrasiak N, Osorio-Santos K, Johansen JR. 2022. *Mojavia aguilerae* and *M.*
434 *dolomitestris* - two new Nostocaceae (Cyanobacteria) species from the Americas. J Phycol
435 58:502–516.

436 28. Patzelt DJ, Hodač L, Friedl T, Pietrasiak N, Johansen JR. 2014. Biodiversity of soil
437 cyanobacteria in the hyper-arid Atacama Desert, Chile. J Phycol 50:698–710.

438 29. Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW. 2005. Three
439 *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations.
440 PLoS Biol 3:e144.

441 30. Bailey S, Clokie MRJ, Millard A, Mann NH. 2004. Cyanophage infection and photoinhibition
442 in marine cyanobacteria. Res Microbiol 155:720–725.

443 31. Sullivan MB, Huang KH, Ignacio-Espinoza JC, Berlin AM, Kelly L, Weigele PR,
444 DeFrancesco AS, Kern SE, Thompson LR, Young S, Yandava C, Fu R, Krastins B, Chase
445 M, Sarracino D, Osburne MS, Henn MR, Chisholm SW. 2010. Genomic analysis of oceanic
446 cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and
447 environments. Environ Microbiol 12:3035–3056.

448 32. Millard AD, Zwirglmaier K, Downey MJ, Mann NH, Scanlan DJ. 2009. Comparative
449 genomics of marine cyanomyoviruses reveals the widespread occurrence of
450 *Synechococcus* host genes localized to a hyperplastic region: implications for mechanisms
451 of cyanophage evolution. Environ Microbiol 11:2370–2387.

452 33. Ponsero AJ, Hurwitz BL, Magain N, Miadlikowska J, Lutzoni F, U'Ren JM. 2021.
453 Cyanolichen microbiome contains novel viruses that encode genes to promote microbial
454 metabolism. ISME Communications 1:1–4.

455 34. Santos-Medellin C, Zinke LA, Ter Horst AM, Gelardi DL, Parikh SJ, Emerson JB. 2021.

456 Viromes outperform total metagenomes in revealing the spatiotemporal patterns of
457 agricultural soil viral communities. ISME J 15:1956–1970.

458 35. Sabehi G, Shaulov L, Silver DH, Yanai I, Harel A, Lindell D. 2012. A novel lineage of
459 myoviruses infecting cyanobacteria is widespread in the oceans. Proc Natl Acad Sci U S A
460 109:2037–2042.

461 36. Guajardo-Leiva S, Pedrós-Alió C, Salgado O, Pinto F, Díez B. 2018. Active Crossfire
462 Between Cyanobacteria and Cyanophages in Phototrophic Mat Communities Within Hot
463 Springs. Front Microbiol 9:2039.

464 37. Skvortsov T, de Leeuwe C, Quinn JP, McGrath JW, Allen CCR, McElarney Y, Watson C,
465 Arkhipova K, Lavigne R, Kulakov LA. 2016. Metagenomic Characterisation of the Viral
466 Community of Lough Neagh, the Largest Freshwater Lake in Ireland. PLoS One
467 11:e0150361.

468 38. Zeigler Allen L, McCrow JP, Ininbergs K, Dupont CL, Badger JH, Hoffman JM, Ekman M,
469 Allen AE, Bergman B, Venter JC. 2017. The Baltic Sea Virome: Diversity and
470 Transcriptional Activity of DNA and RNA Viruses. mSystems 2.

471 39. Adriaenssens EM, Van Zyl L, De Maayer P, Rubagotti E, Rybicki E, Tuffin M, Cowan DA.
472 2015. Metagenomic analysis of the viral community in Namib Desert hypoliths.
473 Environmental Microbiology https://doi.org/10.1111/1462-2920.12528.

474 40. Zablocki O, Adriaenssens EM, Cowan D. 2016. Diversity and Ecology of Viruses in
475 Hyperarid Desert Soils. Appl Environ Microbiol 82:770–777.

476 41. Zablocki O, van Zyl L, Adriaenssens EM, Rubagotti E, Tuffin M, Cary SC, Cowan D. 2014.
477 High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like
478 elements in the metaviromes of antarctic soils. Appl Environ Microbiol 80:6888–6897.

479 42. Van Goethem MW, Swenson TL, Trubl G, Roux S, Northen TR. 2019. Characteristics of
480 Wetting-Induced Bacteriophage Blooms in Biological Soil Crust. MBio 10.

481 43. Poole P, Ramachandran V, Terpolilli J. 2018. *Rhizobia*: from saprophytes to
482 endosymbionts. Nat Rev Microbiol 16:291–303.

483 44. Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, Dehal P, Ware D,
484 Perez F, Canon S, Sneddon MW, Henderson ML, Riehl WJ, Murphy-Olson D, Chan SY,
485 Kamimura RT, Kumari S, Drake MM, Brettin TS, Glass EM, Chivian D, Gunter D, Weston
486 DJ, Allen BH, Baumohl J, Best AA, Bowen B, Brenner SE, Bun CC, Chandonia J-M, Chia J-
487 M, Colasanti R, Conrad N, Davis JJ, Davison BH, DeJongh M, Devoid S, Dietrich E,
488 Dubchak I, Edirisinghe JN, Fang G, Faria JP, Frybarger PM, Gerlach W, Gerstein M,
489 Greiner A, Gurtowski J, Haun HL, He F, Jain R, Joachimiak MP, Keegan KP, Kondo S,
490 Kumar V, Land ML, Meyer F, Mills M, Novichkov PS, Oh T, Olsen GJ, Olson R, Parrello B,
491 Pasternak S, Pearson E, Poon SS, Price GA, Ramakrishnan S, Ranjan P, Ronald PC,
492 Schatz MC, Seaver SMD, Shukla M, Sutormin RA, Syed MH, Thomason J, Tintle NL, Wang
493 D, Xia F, Yoo H, Yoo S, Yu D. 2018. KBase: The United States Department of Energy
494 Systems Biology Knowledgebase. Nat Biotechnol 36:566–569.

495 45. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data.
496 https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Retrieved 20 September 2022.

497 46. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina
498 sequence data. Bioinformatics 30:2114–2120.

499 47. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile
500 metagenomic assembler. Genome Res 27:824–834.

501 48. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast single-node

502 solution for large and complex metagenomics assembly via succinct de Bruijn graph.
503 Bioinformatics 31:1674–1676.

504 49. Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from
505 microbial genomic data. PeerJ 3:e985.

506 50. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. 2021. CheckV
507 assesses the quality and completeness of metagenome-assembled viral genomes. Nat
508 Biotechnol 39:578–585.

509 51. Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of
510 protein or nucleotide sequences. Bioinformatics 22:1658–1659.

511 52. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal:
512 prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics
513 11:119.

514 53. Cook R, Brown N, Redgwell T, Rihtman B, Barnes M, Clokie M, Stekel DJ, Hobman J,
515 Jones MA, Millard A. 2021. INfrastructure for a PHAge REference Database: Identification
516 of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. PHAGE
517 https://doi.org/10.1089/phage.2021.0007.

518 54. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR,
519 Kropinski AM, Krupovic M, Lavigne R, Turner D, Sullivan MB. 2019. Taxonomic assignment
520 of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat
521 Biotechnol 37:632–639.

522 55. Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, Goto S. 2017. ViPTree: the viral
523 proteomic tree server. Bioinformatics 33:2379–2380.

524 56. Bushnell B. 2022. BBMap. SourceForge. https://sourceforge.net/projects/bbmap/. Retrieved
525 20 September 2022.

526 57. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
527 features. Bioinformatics 26:841–842.

528 58. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin
529 R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map
530 format and SAMtools. Bioinformatics 25:2078–2079.

531 59. Ecogenomics/BamM: Metagenomics-focused BAM file manipulation. GitHub.
532 https://github.com/Ecogenomics/BamM. Retrieved 20 September 2022.

533 60. R Core Team. 2021. R: A Language and Environment for Statistical Computing.
534 https://www.r-project.org/. Retrieved 2022.

535 61. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, Kuhn JH,
536 Lavigne R, Brister JR, Varsani A, Amid C, Aziz RK, Bordenstein SR, Bork P, Breitbart M,
537 Cochrane GR, Daly RA, Desnues C, Duhaime MB, Emerson JB, Enault F, Fuhrman JA,
538 Hingamp P, Hugenholtz P, Hurwitz BL, Ivanova NN, Labonté JM, Lee K-B, Malmstrom RR,
539 Martinez-Garcia M, Mizrachi IK, Ogata H, Páez-Espino D, Petit M-A, Putonti C, Rattei T,
540 Reyes A, Rodriguez-Valera F, Rosario K, Schriml L, Schulz F, Steward GF, Sullivan MB,
541 Sunagawa S, Suttle CA, Temperton B, Tringe SG, Thurber RV, Webster NS, Whiteson KL,
542 Wilhelm SW, Wommack KE, Woyke T, Wrighton KC, Yilmaz P, Yoshida T, Young MJ, Yutin
543 N, Allen LZ, Kyrpides NC, Eloe-Fadrosh EA. 2018. Minimum Information about an
544 Uncultivated Virus Genome (MIUViG). Nat Biotechnol 37:29–37.

545 62. Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. 2017. Benchmarking viromics: an
546 evaluation of metagenome-enabled estimates of viral community composition and diversity.

PeerJ 5:e3817.

63. McMurdie PJ, Holmes S. 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS One 8:e61217.

64. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. 2019. Welcome to the Tidyverse. Journal of Open Source Software 4:1686.

65. Ettinger C, Pietrasiak N, Stajich J. stajichlab/Terrestrial_Cyanophage: v1. https://doi.org/10.5281/zenodo.7272374.

66. kblin/ncbi-genome-download: Scripts to download genomes from the NCBI FTP servers. GitHub. https://github.com/kblin/ncbi-genome-download. Retrieved 20 September 2022.

67. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:1–9.

68. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. 2016. Computational approaches to predict bacteriophage-host relationships. FEMS Microbiol Rev 40:258–272.

69. Kothari A, Roux S, Zhang H, Prieto A, Soneja D, Chandonia J-M, Spencer S, Wu X, Altenburg S, Fields MW, Deutschbauer AM, Arkin AP, Alm EJ, Chakraborty R, Mukhopadhyay A. 2021. Ecogenomics of Groundwater Phages Suggests Niche Differentiation Linked to Specific Environmental Tolerance. mSystems e0053721.

570 **Figures and Tables:**

571

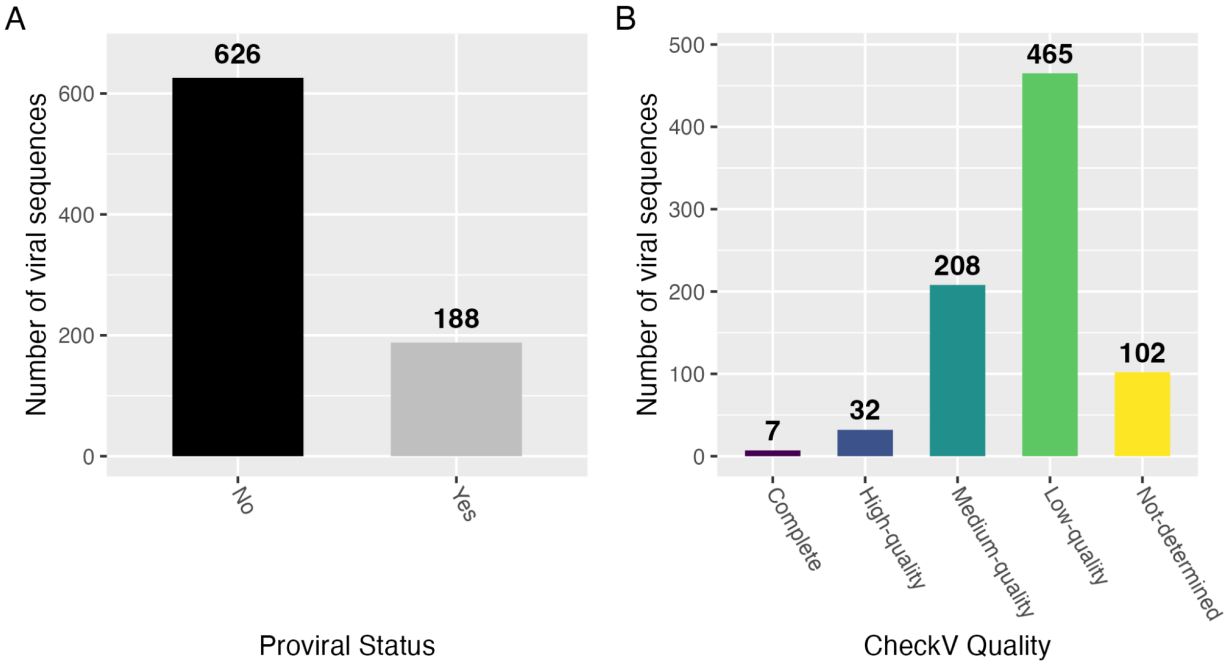
572 **Figure 1:** Viral sequences identified from Cyanobacterial co-cultures. (A) The proviral status of

573 predicted viral genomes is shown as a bar graph, with the number above the bar representing

574 the total number of sequences. (B) CheckV quality predictions for viral sequences are shown as

575 a bar graph with bars colored by quality and the number above the bar representing the total

576 number of sequences in each bar.



577

578

579

580

581

582

583

584

585

Figure 2: Gene-sharing networks of vOTUs that formed viral clusters (VCs). (A) Network of VCs, including VCs unique to this dataset. Each node is a viral sequence and is colored by the habitat of the Cyanobacterial co-culture. Edges connect sequences with overlap in predicted protein content. (B) Bar charts displaying the number of viral sequences in VCs from this study colored by habitat and shown facetted by whether the VC clustered with reference data. Bars are colored by habitat, as in A, and the number next to the bar represents the total number of sequences it represents. (C) Subnetwork of VCs with members that include reference genomes obtained from INPHARED. Each node is a viral sequence and is colored by the predicted viral taxonomic family for the VC, with the exception of sequences identified in this study which are colored in black. Edges connect sequences with overlap in predicted protein content. (D) Bar chart displaying the number of viral sequences from this study assigned to viral families based on VC membership with reference genomes from INPHARED. Bars are colored by viral families, as in C, and the number above the bar represents the total number of sequences it represents. For A-D, only viral sequences generated in this study where CheckV was able to predict quality are shown.

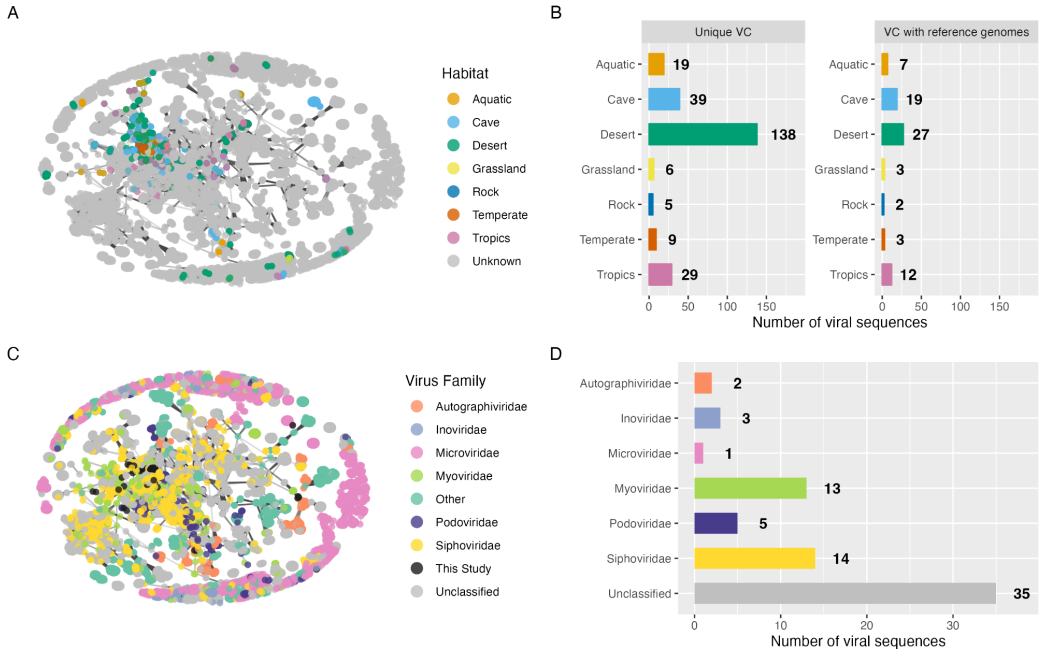
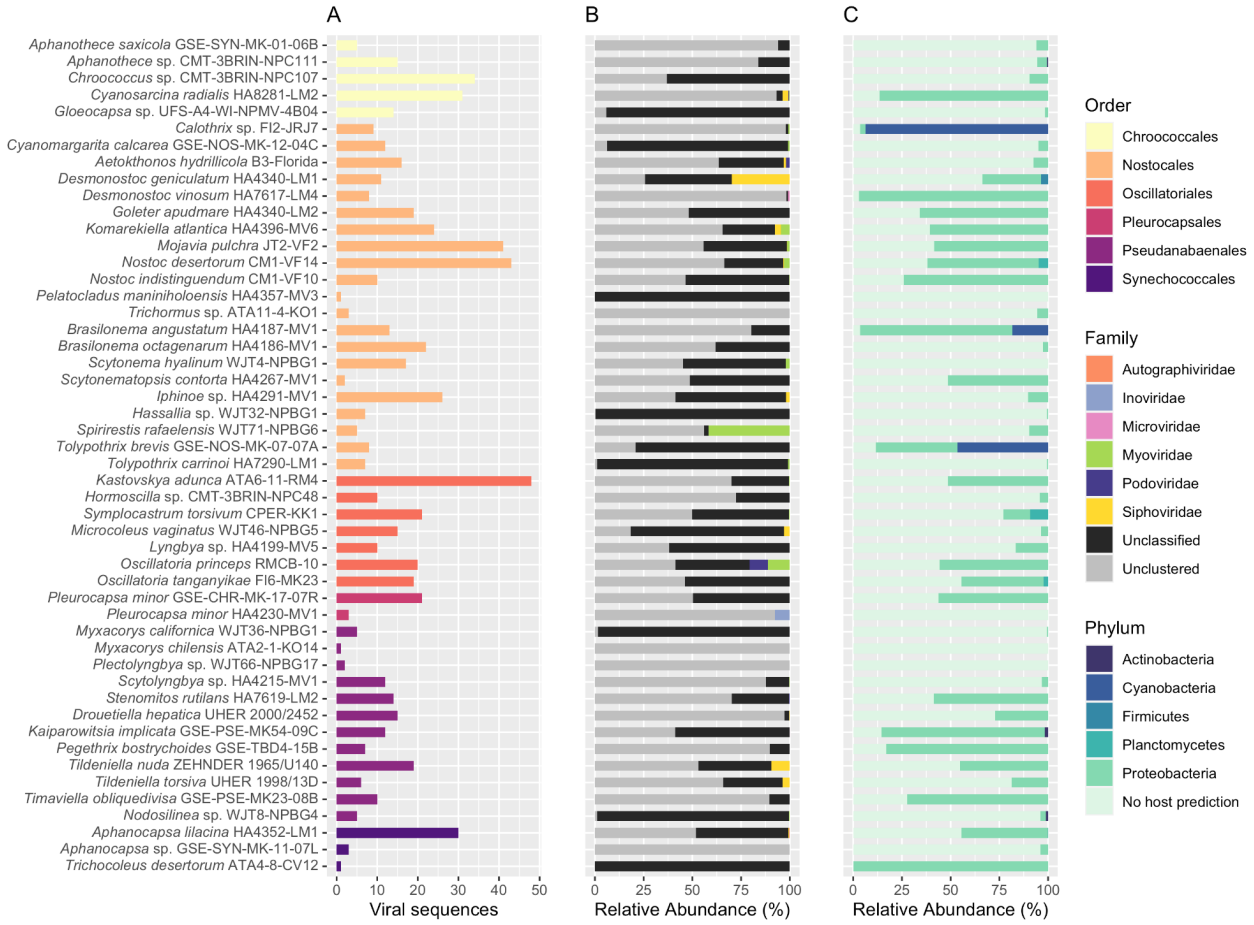



Figure 3: Viral communities across Cyanobacterial co-cultures. (A) The number of viral sequences predicted from each co-culture metagenome. Bars are arranged and colored based on the taxonomic order the Cyanobacteria in the co-culture is taxonomically classified as. (B) Stacked bar charts displaying the relative abundance of vOTUs in each co-culture. Bars are colored based on predicted viral taxonomic families. Viral sequences that clustered into VCs, but that lack taxonomy (i.e., unique VCs that lack reference data, or VCs where reference data taxonomy is unclear) are labeled “Unassigned”, while viral sequences that did not cluster into VCs are labeled “Unclustered”. (C) Stacked bar charts of the relative abundance of vOTUs in each co-culture. Bars are colored based on the predicted host taxonomy, at the phylum level, of each vOTU . For A-C, only viral sequences where CheckV was able to predict quality are shown.