# VIRAL DIVERSITY AND DYNAMICS, AND CRISPR-CAS MEDIATED IMMUNITY IN A ROBUST ALKALIPHILIC CYANOBACTERIAL CONSORTIUM

**Varada Khot[1*], Marc Strous[1], Xiaoli Dong[1,2], Alyse K. Kiesser[3]**

[1]Department of Geoscience, University of Calgary, Calgary, Alberta, Canada

[2] Public Health Laboratory, Alberta Precision Laboratories, Foothills Medical Centre, Calgary, Alberta, Canada

[3]School of Engineering, University of British Columbia Okanagan, Kelowna, BC, Canada

***Correspondence:**

Varada Khot

vmkhot@ucalgary.ca

## ABSTRACT

In many industries, from food to biofuels, contamination of production systems with predators is a costly problem and requires the maintenance of sterile operating conditions. In this study, we look at the robustness of one such alkaliphilic consortium, comprised largely of a

cyanobacterium *Candidatus* Phormidium alkaliphilum, to viral predation. This consortium has existed without a community crash over several years in laboratory and pilot scale environments. We look at CRISPR-Cas systems and viral dynamics in this consortium at four conditions using metagenomic analyses. Results show that while there are active viral members in this community, viral predation of the cyanobacteria is low and does not affect the community dynamics. The multiple CRISPR arrays within the Phormidium were found to be static following initial lab establishment of consortium. Multiple cryptic CRISPR-Cas systems were detected with uncertain viral protection capacity. Our results suggest that dynamics of potential viruses and CRISPR-Cas mediated immunity likely play an important role in the initial establishment of consortia and may continue to support the functional robustness of engineered microbial communities throughout biotechnology applications.

## IMPORTANCE

Biotechnology applications utilizing the function of microbial communities have become increasingly important solutions as we strive for sustainable applications. Although viral infections are known to have significant impact on microbial turnover and nutrient cycling, viral dynamics have remained largely overlooked in these engineered communities. Predatory perturbations to the functional stability of these microbial biotechnology applications must be investigated in order to design more robust applications. In this study, we closely examine virus-microbe dynamics in a model microbial community used in a biotechnology application. Our findings suggest that viral dynamics change significantly with environmental conditions and that microbial immunity may play an important role in maintaining functional stability. We present

this study as a comprehensive template for other researchers interested in exploring predatory dynamics in engineered microbial communities.

# INTRODUCTION

Recent large scale viral metagenomic studies have revealed many discoveries of virus-bacteria interactions across diverse environments. For example, these studies showed the dynamics and impact of viruses on global biogeochemical cycles (1–5) and discovered new bacterial defence mechanisms (6–8), unravelling the critical roles viruses play in microbial community dynamics. Community dynamics refers to the constant microbe-microbe interplay in terms of sharing metabolites, predation and living arrangements such as biofilms. Although viruses have been recognized as important members in all microbial communities, most studies of engineered microbial systems have so far mainly focused on bacteria-bacteria (9,10) or environment-bacteria interactions (11). Engineered microbial environments differ from natural environments in that they are often closed or partially closed systems with more stable environmental conditions such as temperature, salinity, pH and nutrient concentrations, and generally lower microbial diversity. As engineered systems are designed for specific functional outcomes, loss of specific community members due to viral predation, can destabilize otherwise successful processes (12,13). Examples of this have been observed in the food fermentation industry where viral contamination has been an on-going issue (14). A recent experimental study on viruses of microalgae also showed a significant reduction in algal growth in lab cultures infected with viruses, showing potential for economic losses in this industry (15).

In the growing applications of microalgae biotechnology (16), the overall robustness of algae-based systems is of central interest. In particular, viral-associated lysis may lead to major losses of cultivated cyanobacteria (15,17). This phenomenon is commonly observed in cyanobacterial and phytoplanktonic blooms in oceans (1,18,19) and freshwater lakes (20). Yet, little is known about the extent of the impact that viruses and their dynamics have within cyanobacterial microbial communities used in biotechnology or methods to mitigate their impacts. Here we explore the dynamics of viral diversity and CRISPR-Cas mediate immunity in an alkaliphilic cyanobacterial consortium for use in biotechnology applications (21–23). A microbial consortium or community can be defined as a group of microorganisms that fulfil unique ecological roles living together. In addition to the cyanobacterium *Candidatus* Phormidium alkaliphilum, this consortium includes heterotrophs affiliated with *Bacteroidota*, Proteobacteria (Gamma and Alpha), *Verrucomicrobiota*, *Patescibacteria*, *Planctomycetota*, and Archaea (24). The consortium has displayed high productivity and robustness to environmental change, both in the lab and in larger outdoor pilot scale trials (21,25). A microbial consortium is considered robust if external perturbations do not disturb the overall function and growth of the community. The microbial community was originally sourced from alkaline Soda Lakes on the Cariboo Plateau in British Columbia, Canada (26) and has been maintained without a community collapse in a non-sterile lab environment since 2015.

Current literature on viruses from alkaline environments such as soda lakes suggests that viruses are abundant entities there and distinct to viruses from other aquatic environments. For example, tailed viruses featuring large capsids showed abundances that shifted with depth and season (27,28). Novel viruses such as those infecting alkaliphilic *Paracoccus* and giant viruses have also been recently isolated from East African Rift Valley soda lakes (29) and soda lakes in Brazil

82 respectively (30). Additionally, high pH conditions can impact the stability of the capsid protein
83 structure (31,32), and may pose a unique evolutionary pressure for viruses in high pH, high salt
84 environments.

85 Studying viral dynamics within highly diverse microbial communities poses many challenges
86 because of the added complexity of strain-level diversity and the multitude of microbial
87 interactions that are present. Challenges associated with viral sequence identification, historically
88 low rates of host predictions to the genus level (33) and the magnitude of the network connecting
89 multiple viruses associated with multiple hosts (34) make spatiotemporal dynamics difficult to
90 follow. As such, less diverse microbial consortia that exist as established communities or
91 consortia make a more tractable system to study viral community dynamics as microbial
92 populations and dynamics can be tracked over time and under a variety of experimental
93 conditions. In the context of biotechnology, such a model system could be useful to effectively
94 predict functional outcomes and community robustness.

95 To characterize alkaliphilic viral diversity in this engineered system, we used metagenomes of
96 the consortium previously obtained at high (pH 10-11) and low (pH 8-9) pH and with three
97 different nitrogen sources (ammonia, nitrate, and urea) (23,24). A consensus approach was used
98 to identify DNA sequences associated with viruses (35), classify their taxonomy (36) and to infer
99 their hosts (37) from these metagenomes. To explore the bacterial defenses against viruses, we
100 focused our analyses on CRISPR-Cas (Clustered Regularly Interspaced Short Palindromic
101 Repeat) systems. CRISPR-Cas systems are a microbial adaptive “immune system” present in
102 36% of bacteria and 75% of archaea (38). They carry a history of interactions with viruses and
103 mobile genetic elements in the form of ‘spacers’ – snippets of 25-35 base pairs inserted into the
104 CRISPR array, which can be useful for host predictions. CRISPR arrays are typically located

adjacent to CRISPR-Associated (Cas) proteins, which are required for the systems to be functional and therefore also important to identify.

# METHODS

## Dataset

Eleven metagenomic datasets were used for this study as described in Supplementary Table 1. These included four metagenomes from the soda lakes (GLM, DLM, PLM, LCM) (26), six metagenomes of the laboratory photobioreactor cultures (Year 0, Year 1, Year 2) (23,24) and one complete genome of *Candidatus* Phormidium alkaliphilum (23). In all analyses, the Year 1 metagenomic assembly was used as a reference point to compare other datasets to. Metagenomic reads from the consortium grown under four experimental conditions in a photobioreactor described by Ataeian et. al. (23,24) (Supplementary Table 1), were used to create the Year 1 assembly. These conditions, varied by pH and nitrogen source, consisted of low-pH ammonium, low-pH nitrate, high-pH nitrate, and high-pH urea. The total number of paired-end reads from each metagenome was between 40-45 million and the average read length was 145-150 base pairs. Reads from the four conditions were co-assembled using MEGAHIT (34) and are referred to "Year 1 metagenome". The resulting contigs were binned into Metagenome-Assembled-Genomes (MAGs) using MetaBAT v2.12.1 (35). These MAGs are the same as referenced in Ataeian et. al. (24).

## CRISPR Array and *Cas* gene Identification

### CRISPR Array Identification

CRISPR arrays were identified from Year 1 metagenomic reads and assembled contigs using Crass (39) and MinCED (40) respectively, both with default settings. Reads identified by Crass as belonging to CRISPR arrays were mapped to contigs using BBMap (41). An in-house script was then used to retrieve the taxonomic associations for the MAGs that contained the CRISPR reads. Crass identified CRISPR arrays were compared to those identified by MinCED by the associated MAG and number of spacers in each array using an in-house script. These arrays were then dereplicated to keep only 1 unique set of CRISPR arrays per MAG.

### CRISPR Array Comparison

CRISPR arrays from the Year 1 co-assembly were then compared to other consortium metagenomic assemblies (Year 0, Year 1 and Year 2) and assembled metagenomes of microbial mats from soda lakes on the Cariboo Plateau (GEM, PLM, DLM, LCM) (Supplementary Table 1). using Blastn (42) with the options " -task blastn-short -outfmt 6". CRISPR arrays from bin00 (*Candidatus* Phormidium alkaliphilum MAG) from the Year 1 assembly was also compared to the *Candidatus* Phormidium alkaliphilum whole genome sequence (Year 3) using Blastn (same parameters as above). Blast results were then filtered for 100% identity and 100% query cover to find exact matches of spacers over time within the consortium.

### Cas Gene Identification

Cas genes were annotated in the Year 1 co-assembly using the annotation pipeline MetaErg (43), which uses Hidden Markov Models (HMMs) to identify Cas genes (44). Identified clusters of Cas genes were then classified as specific "Cas systems" according to Makarova et al. (45).

## Viral Sequence Identification

Potential viral DNA sequences within the contigs from the Year 1 assembly were identified using Virsorter2 (35) with the options "–keep-original-seq –include-groups dsDNAphage,ssDNA –min-length 5000 –min-score 0.5 all". The resulting 213 viral contigs (VC) were then quality-checked, and host genes trimmed from the ends of identified contigs using CheckV (46). The VCs were filtered using criteria in Viral Sequence Identification Standard Operating Procedure (SOP (47)) and only contigs in the "keep" category with a sequencing depth greater than or equal to 5 in at least 1 sample were kept for downstream analyses. The final number of viral contigs was 83. Dramv (48) was used to annotate genes on these 83 contigs. Reads from all other metagenomic datasets were mapped to the viral contigs identified in the Year 1 assembly using BBMap (41) to trace the origin and temporal dynamics of potential viruses.

## Viral Phylogenomics and taxonomy

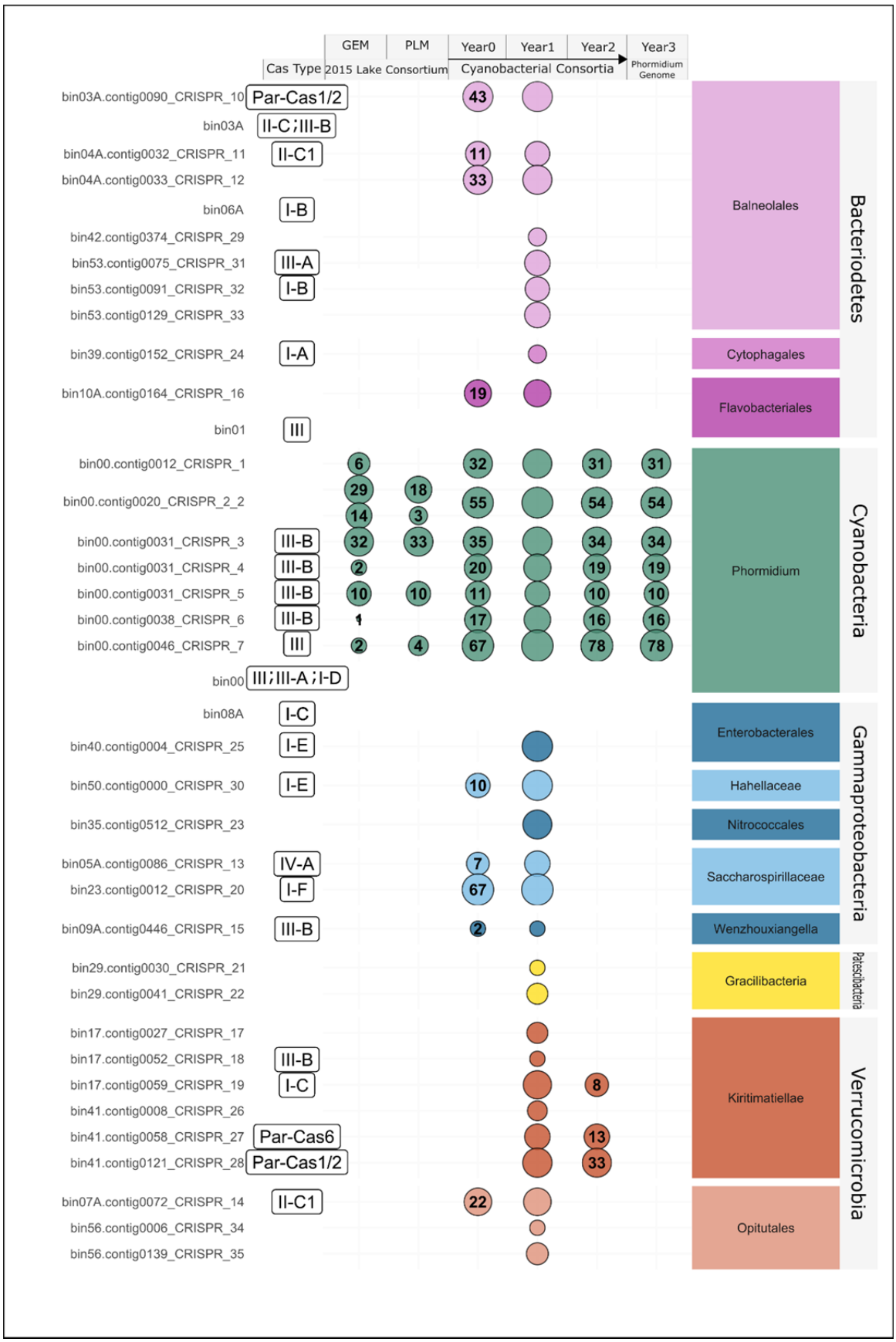
The viral contig dataset was then compared to reference viral genomes from NCBI Viral Refseq Release 208 (2021) (49) using Blastp (42). The bitscores of these matches were used to calculate a Dice coefficient (50,51) to measure the distances between the viral contigs and reference viral genomes. Based on Dice coefficients, sequences were clustered using neighbour-joining to generate a phylogram showing relationships between viral contigs from the consortium and reference viruses. In parallel, contigs were also taxonomically classified using VConTACT2 (36), which uses a reference database of archaeal and bacterial viruses from Viral Refseq v86. The top classification for each viral contig from VConTACT2 was retained.

## Virus-Host Predictions

Four methods were used to determine the most probable host for a given viral contig. 1) Previously identified CRISPR spacers from binned metagenomic contigs were dereplicated and aligned to the viral contigs using BLASTn with the options "-task blastn-short -dust no -outfmt 6" (42). Results were filtered for <= 2 mismatches, alignment length > =20bp and an evalue <= $10^{-5}$. 2) Viral contigs were aligned to the entire metagenomic assembly using BLASTn and alignments > 100bp were considered for host prediction. 3) Tetranucleotide frequency profiles between viral contigs and MAGs were calculated using an in-house script and the cosines of the angle between the tetranucleotide vectors were used to determine proximity of vectors. Top two matches per VC, scoring >= 0.95 were retained for host prediction. 4) read depth profiles of viral contigs and host genomes were generated using bbmap to map reads from each sample to the CC2018 assembly and providing the subsequent .bam files to CheckM coverage (default settings) (52). These sequencing depth profiles were then clustered using hierarchical clustering with a "Bray-Curtis" distance matrix and an "average" clustering algorithm in Python, resulting in a dendrogram of the viral contigs and MAGs. An in-house script was used to parse the dendrogram to find the closest match between each VC and MAG. Viral contigs clustering closely with MAGs were identified as potential virus-host pairs. All four methods for host predictions were used in conjunction to identify a "most probable host" for each viral contig.

# RESULTS

## 1. HOST ADAPTIVE IMMUNITY IN THE CONSORTIUM



189 Figure 1: CRISPR arrays, Cas-system types and spacers from the Year 1 reference metagenome
190 MAGs. Each line shows a single CRISPR array and associated Cas proteins. For each MAG and
191 metagenome, the number within the bubble indicates the number of CRISPR spacers matched to
192 the Year 1 reference metagenome.

193 CRISPR arrays detected by Crass (39) and MinCED (40) were found in 18 out of 29
194 metagenome-assembled genomes (MAGs) from the Year 1 consortium metagenome, sampled
195 after one year of laboratory enrichment (Figure 1). Spacers from the Year 1 CRISPR arrays were
196 compared to previous metagenomes using BLASTn (42) from the enrichment time points (Year
197 0, Year 2, Year 3) and alkaline soda lakes (GEM, PLM, DLM and LCM, Supplementary Table
198 1). The number of spacers within the genome of *Cand*. P. alkaliphilum appears highly consistent
199 across multiple years within the consortia, suggesting these spacers have remained static
200 throughout multiple years of enrichment. When compared to the lakes, spacers from *Cand* P.
201 alkaliphilum CRISPR arrays were only found in the GEM and PLM samples, although the
202 original consortium was seeded from all four soda lakes. The increase in spacers from the
203 original lakes to the enrichment indicates that the cyanobacteria acquired new spacers in all
204 CRISPR arrays (except CRISPR5) during the initial adaptation to the laboratory environment
205 and these have been carried forwards over multiple years.

206 To further explore the potential origin of these spacers, spacers from *Cand* P. alkaliphilum
207 CRISPR arrays were also searched against the Year 1 assembly using BLASTn ("-task blastn-
208 short"). Conclusive matches from this BLAST search (100% identity and (e-value < 1E-3) were
209 only to unbinned contigs, for which taxonomy could not be determined and did not include any
210 viral proteins or contigs. The same spacers were also searched against *Cand* P. alkaliphilum
211 whole genome (Year 3) to look for self-targeting spacers using BLASTn (short task). No

212 conclusive matches for self-targeting spacers were found as they were either short in length
213 (length < 19bp) or with high e-values (e-value > 1E-3).

214 For the heterotrophic consortium members, CRISPR arrays were more dynamic during three
215 years of laboratory cultivation and CRISPR arrays were not conserved between time points of
216 laboratory cultivation. This could be most easily explained by rapid turnover of organisms in the
217 community, rather than a turnover of CRISPR spacers, as MAGs from Year 1 were not found in
218 Year 0 or Year 2 metagenomes. High turnover of heterotrophic consortium members was also
219 reported previously for this consortium (24).

220 As CRISPR-Cas systems often function along side associated *Cas* genes, Cas genes were
221 identified in the MAGs from Year 1 using Hidden Markov Models (44). Cas genes were found
222 adjacent to 20 out of 33 CRISPR arrays (Figure 1). Thirteen out of 31 MAGs had no CRISPR
223 arrays or *Cas* genes.

224 Type I and Type III Cas systems were the most common in the consortium. These are also the
225 most commonly found Cas systems in bacteria, with genes for both systems often found in the
226 same genome (53,54). This was also the case in our dataset where *Cand*. P. alkaliphilum and a
227 species of Kiritimatiellae (Figure 1, bin17), each contained both Type I and Type III systems.
228 Overall, the presence of CRISPR-Cas systems in most of the consortium members would be
229 consistent with the robust nature and long-term stability of our system overall. Yet, the observed
230 high turnover in heterotrophic community members (24) suggests multiple factors that may be at
231 play in the modulation of the heterotrophic populations.

## 2. VIRAL DIVERSITY AND DYNAMICS IN THE CONSORTIUM

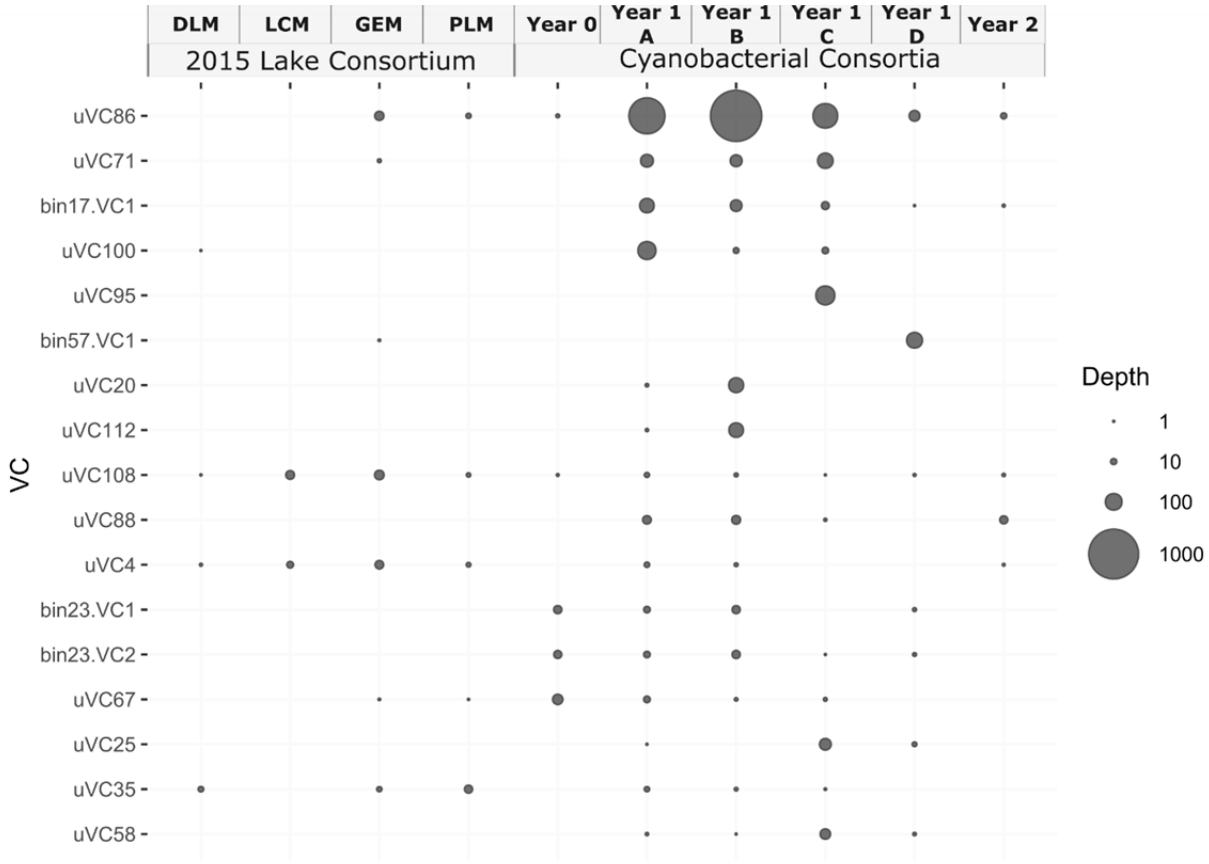



Figure 2. Viral Contig abundance. Numbers of reads mapped to the seventeen most abundant viral contigs (VCs) from all studied metagenomes. For year 1, viral contigs were found in the Year 1 metagenome from the following samples: A) Low-pH Ammonia, B) Low-pH Nitrate, C) High-pH Nitrate, D) High-pH Urea. Abundances of all 83 VCs are presented in Supplementary Figure 2.

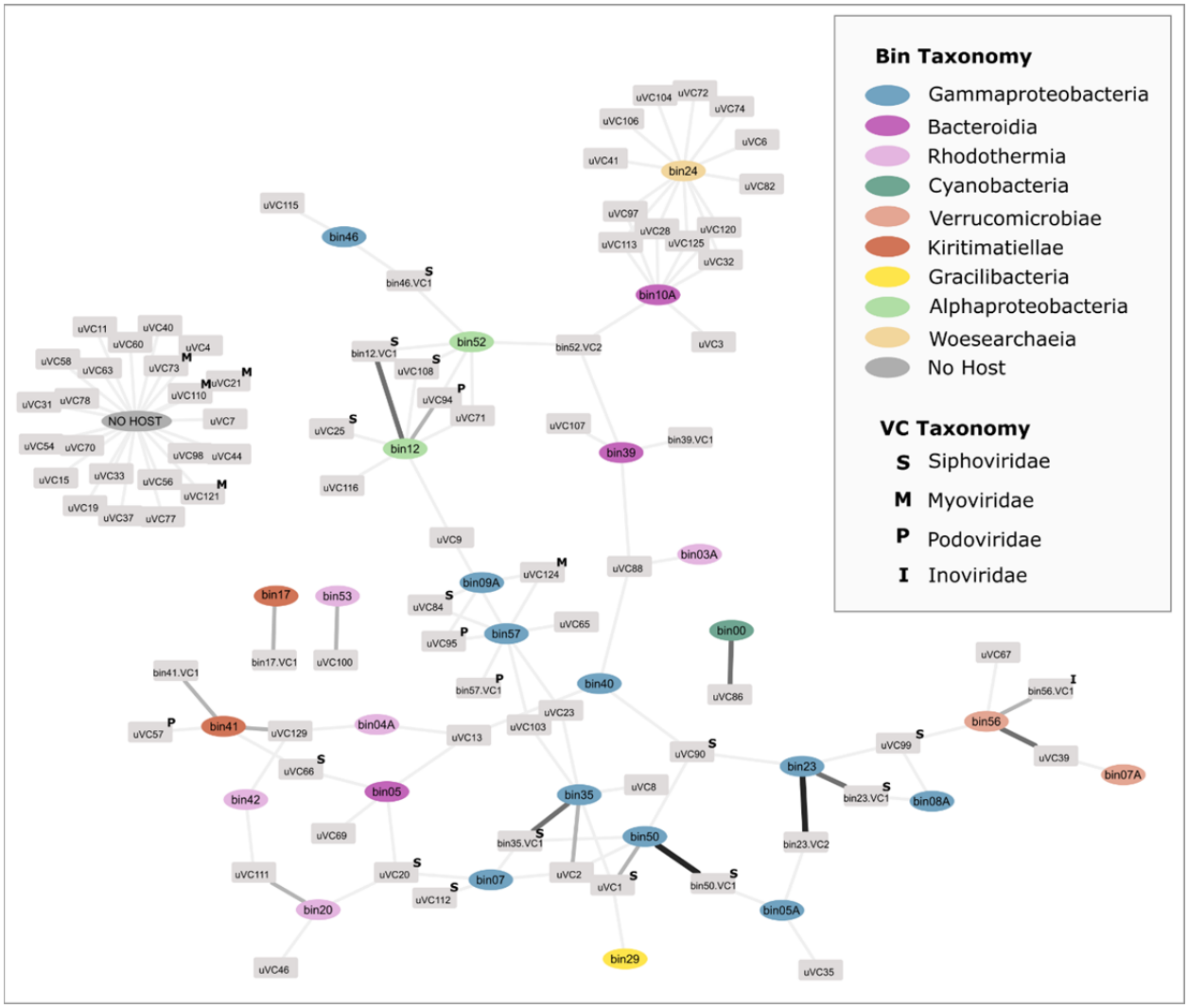
A survey of the Year 1 metagenome for potential viruses with VirSorter2 (35) characterized the viral abundance and diversity in the consortium. The identified contigs were filtered based on their length (>5 kb) and sequencing depth (>5x in any sample), resulting in 83 identified viral contigs (VCs). Twenty-six of the 83 VCs could be taxonomically classified using VConTACT2

(36), with 25 out of 83 viral contigs affiliated with the order Caudovirales (Supplementary Table 3). Sequencing depth of the viral contigs was used to differentiate viral populations in each of the samples (Low-pH Ammonia, Low-pH Nitrate, High-pH Nitrate, High-pH Urea) contributing to the Year 1 co-assembly. A mean sequencing depth of the VCs was within the same order of magnitude of the heterotrophic MAGs (Supplemental Table 3) at 15.32, 14.90, 7.61 and 2.86 in the Low-pH Ammonia, Low-pH Nitrate, High-pH Nitrate, High-pH Urea samples respectively. The notable exception was uVC86 with a sequencing depth an over order of magnitude higher than the mean (Figure 2).

To explore the potential relationship of the viral particles with the consortium community, hosts were computationally assigned to the VCs by four methods: 1) matching CRISPR spacers, 2) high sequence similarity 3) similar tetranucleotide frequencies and 4) similar read depth profiles across samples. Sixty VCs had host predictions using at least one method (Figure 3, Supplemental Table 4) (55). MAGs identified as potential hosts included a range of heterotrophic bacteria from the community, affiliated with *Alphaproteobacteria*, *Bacteriodia*, *Gammaproteobacteria*, *Phycisphaerae*, *Rhodothermia* and *Verrucomicrobiae* (Figure 3). Although many host MAGs possessed CRISPR arrays, no CRISPR spacers matched to their associated VCs at high confidence (100% identity and e-value <= $10^5$). This indicated that the information stored in bacterial CRISPR arrays was likely left over from older, no-longer-relevant viral infections or had a non-viral origin (56). Overall, the host prediction methods used here in parallel linked VCs to multiple hosts (Figure 3), rather than linking each VC with a specific organism in the community. This interpretation of host prediction, while broad, reflects the potential of current computational methods within this dataset. Therefore, in cases where

multiple prediction methods agreed on the host, it resulted in a more robust overall virus-host linkage.

For sixteen VCs, a robust virus-host linkage was determined with two or more methods agreeing on the host prediction. This included 7 VCs that were binned with the host contigs. Binning criteria of MetaBAT2 (35), used for binning, included tetranucleotide frequencies and sequencing depth profiles to group contigs into a bin. Thus, it was unsurprising that binning worked as a host prediction method, regardless of whether the virus was integrated into the host genome or not. In four cases (bin12.VC1, bin23.VC1, bin23.VC2 and bin41.VC1), host genes were found flanking the VC, indicating that it was, indeed, a prophage (Supplemental Table 4, Tab 2). Conversely, for VCs bin35.VC1, bin50.VC1 and bin75.VC1, it was not possible to determine whether these were prophages or free viral fragments as no host genes were found on the same contig. Sequencing depth profiles of bacterial MAGs and viral contigs were also compared to investigate whether viruses and their hosts follow a similar pattern of abundance in the different experimental conditions (Supplemental Figure 2). This method provided host predictions for 27 viral contigs. Generally, the sequencing depth between bacterial contigs and viral contigs within the same MAG co-varied across conditions, indicating that the VC might have originated from a prophage (Supplementary Figure 2), though the viral contigs had lower overall sequencing depth than their identified potential hosts. This could be interpreted that the VC originated from a virus (temperate or lytic) that only infected a portion of the host population. Future efforts to further explore the temporal dynamics of these virus host interactions could include a time series study with closely spaced time intervals or perturbations to conditions and could yield significant insights to how theses interactions integrate into the overall robustness of the system.



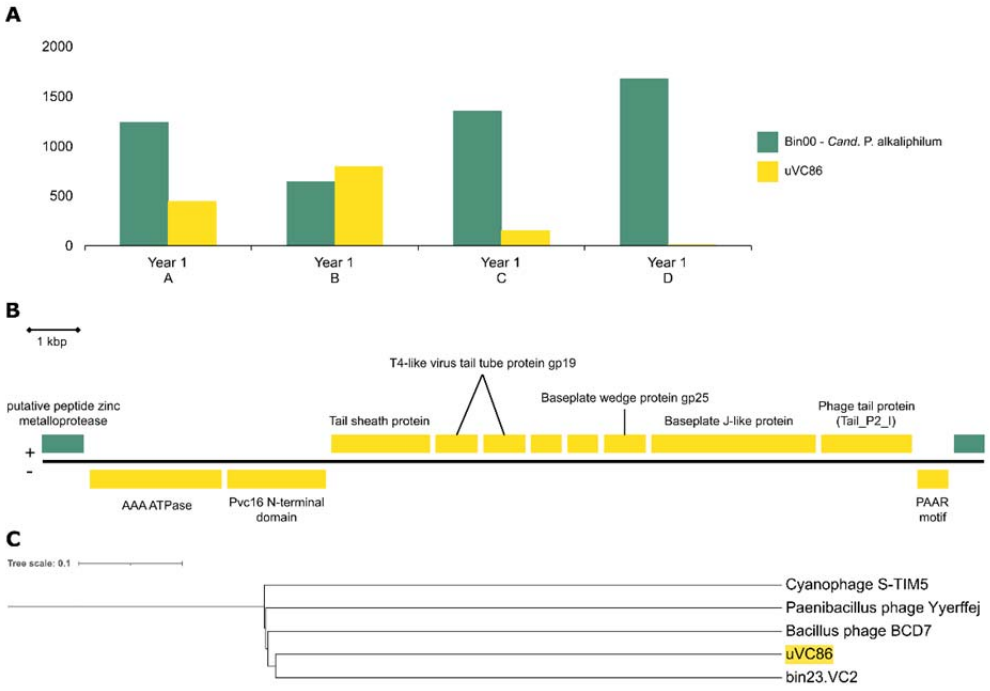

288

289 Figure 3. Network of viral contigs (VCs) linked to computationally predicted hosts. The width
290 and colour of line represent the confidence of the host association. VCs with no associated hosts
291 are clustered together. MAGs are coloured based on taxonomy at the class level.

292 Several studies have documented the existence of CRISPR-Cas systems and integrated phages in
293 a single genome (54,55,57). One such study found that bacteria are likely to have self-targeting
294 spacers if they also have an integrated phage, although this was not found to be the case in our
295 study.

One viral contig of particular interest was bin57.VC1, which was binned with contigs from its host population a Gammaproteobacteria classified as a species of Wenzhouxiangella. This VC only appeared in the high pH-urea condition with 11x higher sequencing depth than the identified host of bin57. The viral contig Bin57.VC1 was relatively large (~41kbp) and did not contain bacterial genes on either end. This indicated that bin57.VC1 likely originated from an active virus in the lytic cycle in the high-pH urea conditions. Although binned, bin57.VC1 did not have hits to any other Wenzhouxiangella genomes or viruses, when searched using Blastn. DRAMv annotations identified genes for both putative capsid and head-tail connecter proteins (Supplemental table 2). Based on previously studied populations of Wenzhouxiangella from high-pH environments, this organism is considered to have a predatory lifestyle, with genes for proteolytic activities and lantibiotic biosynthesis pathways (21,55). Previous stable isotope probing and proteomics of the consortium suggest the bin57 Wenzhouxiangella uptakes macromoleucles from *Cand*. P. alkaliphilum. An activate viral infection of this predatory organism may result in an increase in dissemination of fixed carbon from the Cyanobacteria to the heterotrophic community members (Supplementary figure 2). Indeed, viral activity is known to play a significant role in carbon dynamics on a global scale (18). Additional studies of the dynamics of Bin57.VC1 within the consortia could explore the role of the virus and carbon cycling and its potential to disrupt the overall stability of the consortia in biotechnological applications.

315 ## 3. ROBUSTNESS OF *CAND*. P ALKALIPHILUM TO VIRAL PREDATION




316

317 Figure 4. A) Sequencing depths of *Cand*. P. alkaliphilum and uVC86 in four experimental
318 conditions in Year 1 of growth for Y1 A (Low pH-Ammonia), Y1 B (lLow pH-Nitrate), Y1 C
319 (High pH-Nitrate), Y1 D (High pH-Urea. B) Placement of uVC86 with its closest relatives from
320 RefSeq v208 (49) and the VCs from this study on a neighbor-joining tree calculated from Dice
321 distances. C) Genes on contig uVC86 annotated by Dramv (48) in yellow and cyanobacterial
322 genes in green.

323 As the cyanobacterial consortium depended on the primary productivity of *Cand.* P. alkaliphilum
324 (22,23), its robustness against viral predators was further investigated. We found significant
325 evidence for adaptive immunity in the genome in the form of seven CRISPR arrays and five
326 clusters of Cas genes (Figure 1). As mentioned above, it appeared that spacers in some of these
327 arrays were acquired during initial adaption to the lab. After the initial adaptation, the lack of

change in CRISPR arrays of the *Cand.* P. alkaliphilum (Figure 1) could be attributed to the lack of new viral predators in the laboratory environment, or the presence of other, more effective defence mechanisms that prevented activation of the CRISPR-Cas systems and subsequent incorporation of spacers. Alternatively, the cyanobacterium's CRISPR systems may serve a different, unknown purpose unrelated to viral predation and that does not require updating the arrays (58). For example, given partial spacer self-targets, CRISPR-Cas systems may be involved in gene regulation, DNA repair or programmed cell death (58,59).

Clusters of Cas genes found in the *Cand.* P. alkaliphilum genome belonged to type I and III systems and included a complete type III-B Cas system with a reverse transcriptase adjacent to a Cas1 gene and three CRISPR arrays. Other Cas systems in this genome included a type I-D, with no associated CRISPR array and missing adaption modules (Cas1 and Cas2 genes) as well as a type III-B with a CRISPR array but a missing expression module (Cas6). The *Cand.* P. alkaliphilum genome also contained a 'hybrid' type III system with Cas genes from A, B and D subtypes, including all the modules needed for a functional system (CRISPR array plus adaption, expression, and interference modules (45)). Other Cas genes, Cas3 and Csa3, were also present, with no association with the CRISPR-Cas loci. Multiple and incomplete Cas systems with missing Cas1 and Cas2 genes have been discovered in previous surveys of Cyanobacteria (60).

Type I and type III systems, both found in the Phormidium genome, have different modes of defense via spacer acquisition and target DNA cleavage (61–63) and may be providing synergistic immunity through redundant systems (54) or by sharing Cas modules and CRISPR arrays. Type III Cas systems have broad target specificity as they do not use protospacer adjacent motifs (PAMs) to recognize cleavage sites like type I and II systems, leading to the cleavage of even partial target matches (63). A combination of type I (specific) and type III (broad) Cas

351 systems may ensure that viruses that escape the type I system are captured by the type III system,
352 particularly if CRISPR arrays are shared between the systems. This in part may explain the lack
353 of the spacers with viral sequence targets found in consortium metagenomes as these viruses may
354 have become extinct during the transition from the lakes. In the case of the Phormidium, the type
355 I-D system may be using CRISPR arrays or adaption modules *in trans* from the type III systems
356 as no CRISPR arrays were closely associated with the I-D loci (45). Type I and type III systems
357 are often observed together in genomes and are speculated to share Cas genes (particularly Cas1,
358 Cas2, Cas6) (53,56,60,64). The diverse Cas systems also ensure adaptive immunity to double
359 stranded DNA (all systems), single stranded RNA (type III-B), and single stranded DNA (type I-
360 D effector Csc2). Orphan CRISPR arrays, present in the Phormidium genome, are arrays which
361 are not adjacent to Cas genes and are often found in genomes with multiple CRISPR-Cas
362 systems. The functional significance of these arrays is not well understood; however, some Cas
363 systems are able to use orphan CRISPR arrays in *trans*. Orphaned CRISPR arrays could be a
364 result of the loss of Cas genes, the formation of a new CRISPR array or an array inserted by
365 transposable elements (65). Several recent studies show that multiple remote CRISPR arrays are
366 still functional and could play a role in long-term infection memory (66,67). CRISPR dormancy
367 could be beneficial for evolving the genome through DNA uptake and competency (59). Further
368 studies which specifically test the functionality of the CRISPR arrays to determine whether the
369 unchanging CRISPR arrays in *Cand*. P. alkaliphilum are dormant, have been orphaned or
370 perform other cellular functions. These could include, removal of the CRISPR arrays or Cas
371 genes, insertion of Cas genes into another organism or applying viral predatory pressure.

372 In addition to adaptive immunity, we also looked specifically for viral contigs which came from
373 viruses that infected *Cand*. P. alkaliphilum MAG (bin00). Based on the “Kill-the-Winner” theory

374 (68), we expected the Phormidium to have many viral predators as it was the most abundant
375 organism in the community at 80-90% abundance. However only one viral contig (uVC86)
376 matched with the *Cand.* P. alkaliphilum during host prediction (Figure 2). Other viral predators
377 may have been RNA viruses, which were likely missed during DNA extraction. The virus-host
378 association between the Phormidium and uVC86 was supported by sequencing depth profiles
379 (Figure 4A), tetranucleotide frequencies and the presence of identical sequence fragments
380 (Figure 4C).

381 However, despite the large number of Phormidium CRISPR spacers present, no spacers matched
382 uVC86. This relatively small viral contig, of only 13kbp, had the highest abundance of all the
383 viral contigs. The metagenomes from Good Enough and Probe lakes contained a low number of
384 reads that mapped to uVC86 (26 and 7respectively) (Figure 2). Reads mapping from the lakes
385 covered ~70% of the uVC86 contig from the enrichment, suggesting the lakes as the origin for
386 this viral sequence. This indicated that any potential viruses representing uVC86 found in the
387 consortium at least partially originated in the lakes. The Year 1 metagenomes contained
388 hundreds of reads that mapped to uVC86 and interestingly, abundances of uVC86 and *Cand.* P.
389 Alkaliphilum in the four experiments were negatively correlated (Figure 4A). While the
390 cyanobacterium's abundance was highest in the two high pH experiments, uVC86's abundance
391 was highest in the two low pH experiments and very low or below detection limit at high pH.
392 Anticorrelation of abundances of the viral contig and its putative host might, for example, be
393 explained by 1) Successful viral infections in the low-pH conditions leading to increased uVC86
394 and decreased *Cand.* P. Alkaliphilum abundance; 2) Destabilization of viral (capsid) proteins in
395 the two high pH experiments, leading to lower uVC86 abundance and success (32).

To identify potential relatives of uVC86, we compared the encoded proteins to those encoded on the other viral contigs as well as all Viral Refseq proteins, via Blastp. The resulting Dice-distance based phylogenetic reconstruction showed the closest similar protein sequences found for uVC86 was bin23.VC2, a viral contig predicted to have a different host within the same consortium, a Saccharospirillum species (Figure 4B). The Dice-distance phylogenetic reconstruction was based on homology of individual predicted protein coding sequences, allowing for rearrangement of protein coding sequences rather than overall nucleotide identity which requires gene order to be conserved (69). Despite being neighbours in this tree, viral contigs uVC86 and bin23.VC2 did not appear to be closely related, having an average amino acid identity of only 12% between shared proteins, when compared using Blastp (42). Taxonomic assignment for uVC86 with vConTACT2 was also not possible. Thus, it appeared that uVC86 is a unique cyanophage with no closely related known viruses.

Functional gene annotations by DRAMv (48) provided some clues about this potential virus's lifestyle and taxonomy (Figure 4C, Supplementary Table 3, Tab 2). Several genes were shared with T4-like phages (tail tube and baseplate wedge proteins), and some had distant similarity to P2-like phage proteins (baseplate J-like and phage tail). Both these viruses are affiliated with Myoviridae. uVC86 encoded no capsid, integrase, or replication genes, so it is likely an incomplete viral genome, unable to replicate by itself. Viral contigs of comparable sequencing depth that could potentially complement these missing genes in uVC86 were not found. Furthermore, both ends of uVC86, encoded putative host genes (bin00.contig 4394, gene 1 and gene 12, Supplementary Table 3, Tab 2) flanking the cluster of viral genes (Figure 3C). One of these genes was a M50 metallopeptidase MER0890051; MEROPS peptidase database (70), while the other was of unknown function. Both genes had high similarity to genes of a

Geitlerinema species, a close relative of *Cand*. P. alkaliphilum. As there was no evidence of Geitlerinema in the consortium, we assumed that these genes originated from *Cand*. P. alkaliphilum, even though, they were not present in the (circular) whole genome sequence of *Cand*. P. alkaliphilum (23). They most likely originated from a *Cand*. P. alkaliphilum subpopulation. Thus, it appears that uVC86 was a remnant of a prophage present in the genome of a subpopulation of Cand. P. alkaliphilum, a subpopulation that was enriched at low pH. The differences in uVC86 abundances between experiments were likely caused by differences in abundances of the host subpopulation. Coexistence of multiple Cand. P. alkaliphilum subpopulations with different ecological fitness, dependent on pH and nitrogen source, was previously reported and is a key aspect of the robustness of the consortium. The existence of a remnant of a potential phage genome within these subpopulations is consistent with a dynamic nature of phage-host dynamics within natural environments.

# CONCLUSION

From a biotechnology perspective, success of an autotrophic-based consortium is reliant on the survival of the primary producer, in this case *Cand*. P. alkaliphilum. Previous research has demonstrated the consortium's robustness in the context of environmental change (21,24), as well as the different ecological roles filled by the heterotrophic members of the consortia and the characteristics of *Cand*. P. alkaliphilum. Here, we explored further the potential robustness of the consortium to viral predation. Many consortium members, including the *Cand*. P alkaliphilum, displayed adaptive immunity in the form of one or more CRISPR-Cas systems. Based on the diversity and abundance of its Cas genes, we expect the Phormidium to have comprehensive adaptive immunity that targets both RNA and DNA viruses. However, the observed lack of

dynamicity of the CRISPR arrays points to greater complexity of purpose for these systems. The survey of potential viral predators in the consortium highlighted the presence of several diverse putative viruses that exist in this high pH system. The recovery of only one partial viral genome of a potential cyanophage, despite the high abundance of *Cand*. P alkaliphilum, suggests an insufficient predatory pressure to maintain cyanobacterial innate immunity within the established consortia. The addition of new CRISPR spacers within the first year of the consortia being established strongly suggests a critical role of adaptive immunity of *Cand*. P. alkaliphilum populations in the *initial establishment* of the robust consortium.

## FUNDING

This study was supported by the Natural Sciences and Engineering Research Council (NSERC), Canada Foundation for Innovation (CFI), Canada First Research Excellence Fund (CFREF), Alberta Innovates, the Government of Alberta, the University of Calgary. Additional funding for VK was provided by Alberta Graduate Excellence Scholarship and the Faculty of Graduate Studies Doctoral Scholarship.

## ACKNOWLEDGMENTS

We would like to thank Maryam Ataeian and Jackie Zorz for providing access to and help with sequencing data.

# REFERENCES

1. Coutinho FH, Gregoracci GB, Walter JM, Thompson CC, Thompson FL. Metagenomics Sheds Light on the Ecology of Marine Microbes and Their Viruses. Trends Microbiol. 2018;26(11):955–65.

2. Suttle CA. Viruses in the sea. Nature. 2005 Sep 15;437(7057):356–61.

3. Jarett JK, Džunková M, Schulz F, Roux S, Paez-Espino D, Eloe-Fadrosh E, et al. Insights into the dynamics between viruses and their hosts in a hot spring microbial mat. ISME J. 2020;14(10):2527–41.

4. Trubl G, Jang HB, Roux S, Emerson JB, Solonenko N, Vik DR, et al. Soil Viruses Are Underexplored Players in Ecosystem Carbon Processing. mSystems . 2018 Oct 2 [cited 2021 Apr 12];3(5).

5. Roux S, Hawley AK, Torres Beltran M, Scofield M, Schwientek P, Stepanauskas R, et al. Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. Dubilier N, editor. eLife. 2014 Aug 29;3:e03125.

6. Goldfarb T, Sberro H, Weinstock E, Cohen O, Doron S, Charpak-Amikam Y, et al. BREX is a novel phage resistance system widespread in microbial genomes. EMBO J. 2015 Jan 14;34(2):169–83.

7. Makarova KS, Wolf YI, Snir S, Koonin EV. Defense Islands in Bacterial and Archaeal Genomes and Prediction of Novel Defense Systems. J Bacteriol. 2011 Nov 1;193(21):6039–56.

8. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, et al. Systematic discovery of antiphage defense systems in the microbial pangenome. Science . 2018 Mar 2 [cited 2021 Apr 12];359(6379).

9. Xia J, Ye L, Ren H, Zhang XX. Microbial community structure and function in aerobic granular sludge.

10. Ferro L, Hu YOO, Gentili FG, Andersson AF, Funk C. DNA metabarcoding reveals microbial community dynamics in a microalgae-based municipal wastewater treatment open photobioreactor. Algal Res. 2020 Oct 1;51:102043.

11. Wells GF, Park HD, Eggleston B, Francis CA, Criddle CS. Fine-scale bacterial community dynamics and the taxa–time relationship within a full-scale activated sludge bioreactor. Water Res. 2011 Nov 1;45(17):5476–88.

12. Louca S, Doebeli M. Taxonomic variability and functional stability in microbial communities infected by phages. Environ Microbiol. 2017;19(10):3863–78.

13. Marx V. Engineers embrace microbiome messiness. Nat Methods. 2019 Jul;16(7):581–4.

14. Samson JE, Moineau S. Bacteriophages in Food Fermentations: New Frontiers in a Continuous Arms Race. Httpdxdoiorg101146annurev-Food-030212-182541. 2013 Mar 6;4(1):347–68.

15. Kraft K, Alum A, Abbaszadegan M. Environmental algal phage isolates and their impact on production potential for food and biofuel applications. J Appl Microbiol. 2020 Jan 3;128(1):182–90.

16. Shurin JB, Abbott RL, Deal MS, Kwan GT, Litchman E, McBride RC, et al. Industrial-strength ecology: trade-offs and opportunities in algal biofuel production. Ecol Lett. 2013 Nov;16(11):1393–404.

17. Molina D, de Carvalho JC, Júnior AIM, Faulds C, Bertrand E, Soccol CR. Biological contamination and its chemical control in microalgal mass cultures. Appl Microbiol Biotechnol. 2019 Dec 1;103(23–24):9345–58.

18. Wilhelm SW, Suttle C a. Viruses and Nutrient Cycles in the Sea. BioScience. 1999;49(10):781–8.

19. Proctor LM, Fuhrman JA. Viral mortality of marine bacteria and cyanobacteria. Nature. 1990;343(6253):60–2.

20. Gons HJ, Ebert J, Hoogveld HL, Van Den Hove L, Pel R, Takkenberg W, et al. Observations on cyanobacterial population collapse in eutrophic lake water. Antonie Van Leeuwenhoek Int J Gen Mol Microbiol. 2002;81(1–4):319–26.

21. Sharp CE, Urschel S, Dong X, Brady AL, Slater GF, Strous M. Robust , high - productivity phototrophic carbon capture at high pH and alkalinity using natural microbial communities. Biotechnol Biofuels. 2017;1–13.

22. Ataeian M, Liu Y, Canon-Rubio KA, Nightingale M, Strous M, Vadlamani A. Direct capture and conversion of CO2 from air by growing a cyanobacterial consortium at pH up to 11.2. Biotechnol Bioeng. 2019 Jul 1;116(7):1604–11.

517 23. Ataeian M, Vadlamani A, Haines M, Mosier D, Dong X, Kleiner M, et al. Whole Genome
518 Sequence and Experimental Proteome of Candidatus “Phormidium Alkaliphilum”, an
519 Abundant and Widespread Cyanobacterium With Applications in Biotechnology. iScience.
520 2021;

521 24. Ataeian M, Liu Y, Kouris A, Hawley AK, Strous M. Ecological Interactions of
522 Cyanobacteria and Heterotrophs Enhances the Robustness of Cyanobacterial Consortium for
523 Carbon Sequestration. Front Microbiol . 2022;13.

524 25. Haines M, Vadlamani A, Daniel Loty Richardson W, Strous M. Pilot-scale outdoor trial of a
525 cyanobacterial consortium at pH 11 in a photobioreactor at high latitude. Bioresour Technol.
526 2022 Jun 1;354.

527 26. Zorz JK, Sharp C, Kleiner M, Gordon PMK, Pon RT, Dong X, et al. A shared core
528 microbiome in soda lakes separated by large distances. Nat Commun. 2019;10(1):1–10.

529 27. Brum JR, Steward GF. Morphological Characterization of Viruses in the Stratified Water
530 Column of Alkaline, Hypersaline Mono Lake. Microb Ecol. 2010;60(3):636–43.

531 28. Jiang S, Steward G, Jellison R, Chu W, Choi S. Abundance, distribution, and diversity of
532 viruses in alkaline, hypersaline Mono Lake, California. Microb Ecol. 2004;47(1):9–17.

533 29. Van Zyl LJ, Nemavhulani S, Cass J, Cowan DA, Trindade M. Three novel bacteriophages
534 isolated from the East African Rift Valley soda lakes. Virol J. 2016 Dec 3;13(1):1–14.

30. Abrahão J, Silva L, Silva LS, Khalil JYB, Rodrigues R, Arantes T, et al. Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. Nat Commun 2018 91. 2018 Feb 27;9(1):1–12.

31. Michel JP, Ivanovska IL, Gibbons MM, Klug WS, Knobler CM, Wuite GJL, et al. Nanoindentation studies of full and empty viral capsids and the effects of capsid protein mutations on elasticity and strength. Proc Natl Acad Sci U S A. 2006;103(16):6184–9.

32. Cuellar JL, Meinhoevel F, Hoehne M, Donath E. Size and mechanical stability of norovirus capsids depend on pH: A nanoindentation study. J Gen Virol. 2010;

33. Paez-Espino D, Chen IA, Palaniappan K, Ratner A, Chu K, Szeto E, et al. IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. Nucleic Acids Res. 2017;45(D1):D457–65.

34. Albery GF, Becker DJ, Brierley L, Brook CE, Christofferson RC, Cohen LE, et al. The science of the host–virus network. Nat Microbiol. 2021 Dec;6(12):1483–92.

35. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-huerta G, Delmont TO, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. 2021;1–13.

36. Bolduc B, Jang HB, Doulcier G, You ZQ, Roux S, Sullivan MB. vConTACT: An iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. PeerJ. 2017;2017(5):1–26.

554 37. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict
555 bacteriophage-host relationships. FEMS Microbiol Rev. 2016;40(2):258–72.

556 38. Pourcel C, Touchon M, Villeriot N, Vernadet JP, Couvin D, Toffano-Nioche C, et al.
557 CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from
558 complete genome sequences, and tools to download and query lists of repeats and spacers.
559 Nucleic Acids Res. 2020 Jan 8;48(D1):D535–44.

560 39. Skennerton CT, Imelfort M, Tyson GW. Crass: Identification and reconstruction of CRISPR
561 from unassembled metagenomic data. Nucleic Acids Res. 2013;41(10).

562 40. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, et al. CRISPR
563 Recognition Tool (CRT): A tool for automatic detection of clustered regularly interspaced
564 palindromic repeats. BMC Bioinformatics. 2007;8:1–8.

565 41. Bushnell B. BBMap: a fast, accurate, splice-aware aligner. Vol. 13, Joint Genome Institute,
566 department of energy. 2014.

567 42. Camacho C, Coulouris G, Avagyan V, Papadopoulos MNJ, Bealer K, Madden TL. BLAST+:
568 architecture and applications. BMC Bioinformatics. 2009;10(421).

569 43. Dong X, Strous M. An Integrated Pipeline for Annotation and Visualization of Metagenomic
570 Contigs. Front Genet. 2019;

571 44. Burstein D, Harrington LB, Strutt SC, Probst AJ, Anantharaman K, Thomas BC, et al. New
572 CRISPR–Cas systems from uncultivated microbes. Nat 2017 5427640. 2016 Dec
573 22;542(7640):237–41.

574 45. Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, Brouns SJJ, et al.
575 Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants.
576 Nat Rev Microbiol. 2020;18(2):67–83.

577 46. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV assesses
578 the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol 2020
579 395. 2020 Dec 21;39(5):578–85.

580 47. Guo J, Vik D, Pratama AA, Roux S, Sullivan M. Viral sequence identification SOP with
581 VirSorter2. protocols.io. 2021;

582 48. Shaffer M, Borton MA, McGivern BB, Zayed AA, Leanti La Rosa S, Solden LM, et al.
583 DRAM for distilling microbial metabolism to automate the curation of microbiome function.
584 Nucleic Acids Res. 2020;48(16):8883–900.

585 49. JR B, D AA, Y B, O B. NCBI viral genomes resource. Nucleic Acids Res. 2015;43(database
586 issue):571–7.

587 50. Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. Expanding the Marine Virosphere
588 Using Metagenomics. PLoS Genet. 2013;9(12).

589 51. Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, Brussaard CPD, et
590 al. Marine viruses discovered via metagenomics shed light on viral strategies throughout the
591 oceans. Nat Commun. 2017;8(May):1–12.

592 52. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the
593 quality of microbial genomes recovered from isolates, single cells, and metagenomes.
594 Genome Res. 2015 Jul 1;25(7):1043.

595 53. Bernheim A, Bikard D, Touchon M, Rocha EPC. Atypical organizations and epistatic
596 interactions of CRISPRs and cas clusters in genomes and their mobile genetic elements.
597 Nucleic Acids Res. 2020;48(2):748–60.

598 54. Silas S, Lucas-Elio P, Jackson SA, Aroca-Crevillén A, Hansen LL, Fineran PC, et al. Type
599 III CRISPR-Cas systems can provide redundancy to counteract viral escape from type I
600 systems. eLife. 2017;6.

601 55. Nobrega FL, Walinga H, Dutilh BE, Brouns SJJ. Prophages are associated with extensive
602 CRISPR-Cas auto-immunity. Nucleic Acids Res. 2020;48(21):12074–84.

603 56. Martínez Arbas S, Narayanasamy S, Herold M, Lebrun LA, Hoopmann MR, Li S, et al.
604 Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics
605 revealed using time-series integrated meta-omics. Nat Microbiol . 2020;

606 57. Rollie C, Chevallereau A, Watson BNJ, Chyou T yuan, Fradet O, McLeod I, et al. Targeting
607 of temperate phages drives loss of type I CRISPR–Cas systems. Nat 2020 5787793. 2020 Jan
608 22;578(7793):149–53.

609 58. Faure G, Makarova KS, Koonin EV. CRISPR–Cas: Complex Functional Networks and
610 Multiple Roles beyond Adaptive Immunity. J Mol Biol. 2019;431(1).

611 59. Westra ER, Buckling A, Fineran PC. CRISPR-Cas systems: Beyond adaptive immunity. Nat
612 Rev Microbiol. 2014;12(5):317–26.

613 60. Cai F, Axen SD, Kerfeld CA. Evidence for the widespread distribution of CRISPR-Cas
614 system in the Phylum Cyanobacteria. RNA Biol. 2013;10(5):687–93.

615 61. Fineran PC, Gerritzen MJH, Suárez-Diez M, Künne T, Boekhors J, Hiju SAFT van, et al.
616 Degenerate target sites mediate rapid primed CRISPR adaptation. Proc Natl Acad Sci U S A.
617 2014;111:1629–38.

618 62. Nicholson TJ, Jackson SA, Croft BI, Staals RHJ, Fineran PC, Brown CM. Bioinformatic
619 evidence of widespread priming in type I and II CRISPR-Cas systems. RNA Biol.
620 2019;16(4).

621 63. Pyenson NC, Gayvert K, Varble A, Elemento O, Marraffini LA. Broad Targeting Specificity
622 during Bacterial Type III CRISPR-Cas Immunity Constrains Viral Escape. Cell Host
623 Microbe. 2017;22(3):343-353.e3.

624 64. Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, et al.
625 Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol. 2011;9(6).

626 65. Shmakov SA, Utkina I, Wolf YI, Makarova KS, Severinov KV, Koonin EV. CRISPR Arrays
627 Away from cas Genes.

628 66. Reimann V, Ziemann M, Li H, Zhu T, Behler J, Lu X, et al. Specificities and functional
629 coordination between the two Cas6 maturation endonucleases in Anabaena sp. PCC 7120
630 assign orphan CRISPR arrays to three groups. RNA Biol. 2020 Oct 2;17(10):1442–53.

631 67. Weissman JL, Fagan WF, Johnson PLF. Selective Maintenance of Multiple CRISPR Arrays
632 Across Prokaryotes. CRISPR J. 2018;1(6):405–13.

633 68. Thingstad TF. Elements of a theory for the mechanisms controlling abundance, diversity,
634 and biogeochemical role of lytic bacterial viruses in aquatic systems. Am Soc Limnography
635 Oceanogr. 2000;45(6):1320–8.

636 69. Coutinho FH, Edwards RA, Rodríguez-Valera F. Charting the diversity of uncultured viruses
637 of Archaea and Bacteria. BMC Biol. 2019;17(1):1–16.

638 70. Rawlings ND, Waller M, Barrett AJ, Bateman A. MEROPS: the database of proteolytic
639 enzymes, their substrates and inhibitors. Nucleic Acids Res . 2014 Jan 1