GRIFFITH UNIVERSITY

Cylindrospermopsis raciborskii Virus and host: genomic characterization and ecological relevance

Author

Martin, Robbie M, Moniruzzaman, Mohammad, Mucci, Nicholas C, Willis, Anusuya, Woodhouse, Jason N, Xian, Yuejiao, Xiao, Chuan, Brussaard, Corina PD, Wilhelm, Steven W

Published

2019

Journal Title

Environmental Microbiology

Version

Accepted Manuscript (AM)

DOI

10.1111/1462-2920.14425

Rights statement

© 2018 Society for Applied Microbiology and John Wiley & Sons Ltd. This is the peer reviewed version of the following article: Cylindrospermopsis raciborskii Virus and host: genomic characterization and ecological relevance, Environmental Microbiology, Volume 21, Issue 6, Pages 1942-1956, which has been published in final form at https://doi.org/10.1111/1462-2920.14425. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving (http://olabout.wiley.com/WileyCDA/Section/id-828039.html)

Downloaded from

http://hdl.handle.net/10072/391619

Griffith Research Online

https://research-repository.griffith.edu.au

Accepted Article

Cylindrospermopsis raciborskii Virus and host: genomic characterization and ecological relevance

Robbie M. Martin[1], Mohammad Moniruzzaman[1,6], Nicholas C. Mucci[1], Anusuya Willis[2], Jason N. Woodhouse[3], Yuejiao Xian[4], Chuan Xiao[4], Corina P. D. Brussaard[5], Steven W. Wilhelm[1*]

[1]Department of Microbiology, University of Tennessee, Knoxville, Tennessee, USA
[2]Australian National Algae Culture Collection, CSIRO, Hobart, Australia
[3]Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany
[4]Department of Chemistry, University of Texas at El Paso, El Paso, Texas, USA
[5]Royal Netherlands Institute for Sea Research, Department of Marine Microbiology and Biogeochemistry, and Utrecht University, Texel, The Netherlands

*author for correspondence: Dr. Steven W. Wilhelm, Department of Microbiology, The University of Tennessee, Knoxville TN, 37996 USA,
wilhelm@utk.edu
865-974-0665

Current addresses: [6] Monterey Bay Aquarium Research Institute, Moss Landing, California

Running title: *Cylindrospermopsis raciborskii* Virus genome

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1462-2920.14425

This article is protected by copyright. All rights reserved

Accepted Article

Abstract

Cylindrospermopsis (Raphidiopsis) raciborskii is an invasive, filamentous, nitrogen-fixing cyanobacterium that forms frequent blooms in freshwater habitats. While viruses play key roles in regulating the abundance, production, and diversity of their hosts in aquatic ecosystems, the role(s) of viruses in the ecology of *C. raciborskii* is almost unexplored. Progress in this field has been hindered by the absence of a characterized virus-host system in *C. raciborskii*. To bridge this gap, we sequenced the genome of CrV-01T, a previously isolated cyanosiphovirus, and its host *C. raciborskii* strain Cr2010. Analyses suggest that CrV-01T represents a distinct clade of siphoviruses infecting, and perhaps lysogenizing, filamentous cyanobacteria. Its genome contains unique features that include an intact CRISPR array and a 12-kb inverted duplication. Evidence suggests CrV-01T recently gained the ability to infect Cr2010 and recently lost the ability to form lysogens. The cyanobacterial host contains a CRISPR-Cas system with CRISPR spacers matching protospacers within the inverted duplication of the CrV-01T genome. Examination of metagenomes demonstrates that viruses with high genetic identity to CrV-01T, but lacking the inverted duplication, are present in *C. raciborskii* blooms in Australia. The unique genomic features of the CrV/Cr2010 system offers opportunities to investigate in more detail virus-host interactions in an ecologically important bloom-forming cyanobacterium.

This article is protected by copyright. All rights reserved

Accepted Article

Originality – significance statement

The paper describes the first genomic sequence of a virus (CrV-01T) infecting a strain of *Cylindrospermopsis* spp.; the paper also describes the genome of the host and establishes a newly characterized virus/host pair in a bloom-forming cyanobacterium.. We demonstrate the presence of a CRISPR-array within the virus, providing a new model system for the study of potential CRISPR benefits within viruses. We use the genomic information from CrV-01T to demonstrate that other filamentous cyanobacteria in culture collections contain previously unseen prophage remnants, including, in at least one case, nearly intact virus genomes. The genome information of CrV-O1T is used to show that highly similar viruses are found within a *Cylindrospermopsis* bloom occurring in Australia, demonstrating an apparent wide distribution of CrV-like viruses and their presumed ecologic importance.

This article is protected by copyright. All rights reserved

Accepted Article

Introduction

Cylindrospermopsis raciborskii (recently proposed to be renamed *Raphidiopsis raciborskii* (Aguilera et al., 2018)) is a filamentous, nitrogen-fixing cyanobacterium that forms frequent blooms in freshwater habitats (Antunes et al., 2015). Some strains produce potent cytotoxins and neurotoxins (cylindrospermopsins, saxitoxins), while others are non-toxic. Identification of *C. raciborskii* as a producer of cytotoxins (Hawkins et al., 1985) and its increasing importance as a nuisance bloom-former has driven a considerable research effort aimed at understanding the biological underpinnings of its success. Its high phosphorus uptake capacity (Istvanovics et al., 2000), opportunistic nitrogen-fixing ability (Burford et al., 2006; Willis et al., 2016), and tolerance of a wide range of temperatures (Briand et al., 2004) are key physiological adaptations supporting its success as an invasive and competitive species (reviewed in Burford *et al.*, 2016). *C. raciborskii* was initially described as a tropical/subtropical species (Padisák, 1997), but over the past two decades, the number of studies reporting the presence of *C. raciborskii* in temperate locations has increased. It is now a persistent member of freshwater phytoplankton communities in temperate regions of Europe, North and South America, Asia, Australia, and New Zealand (reviewed in Antunes, *et al.*, 2015; Sinha, *et al.*, 2012; Steenhauer, *et al.*, 2016). The consensus is that increasing global temperatures are contributing to a range expansion of *C. raciborskii* into temperate ecosystems (Sinha et al., 2012).

While the physiological advantages of *C. raciborskii* over other phytoplankton are now clearer, an unanswered ecological question is how viruses shape *C. raciborskii* populations. Viruses play key roles in regulating the abundance, production, and diversity of their hosts in both marine and freshwater ecosystems and heavily influence biogeochemical cycles (Wilhelm and Suttle, 1999; Weitz and Wilhelm, 2012). It is becoming increasingly accepted that to

This article is protected by copyright. All rights reserved

Accepted Article

understand the ecology of any aquatic microorganism, one must also study the viruses that infect them (Mojica and Brussaard, 2014; Sullivan et al., 2017). This may be of particular importance for microorganisms that produce toxins that can be released during virus-mediated cell lysis (*e.g.* Steffen et al., 2017).

To date only three reports describing viruses infecting *C. raciborskii* are available. In two reports from Australia, a siphovirus was isolated from Lake Samsonvale (Pollard and Young, 2010), and four culture collection strains and two strains isolated from natural blooms were suggested to harbor inducible prophages (Steenhauer et al., 2014). In the third report, a cyanophage (CrV-01T, referred to as CrV from here on) infecting *C. raciborskii* was isolated from a temperate lake (Reeuwijkse Plassen) in the Netherlands (Steenhauer et al., 2016). Thus far only one study has examined the influence of environmental factors (*e.g.*, temperature, light) on *C. raciborskii* host-cyanophage interactions (Steenhauer et al., 2016). With so little known, the role of viruses in shaping *C. raciborskii* populations remains an open and exciting question.

Genome sequences provide access to methods that allow deeper exploration of interactions between a virus and its host(s). Genome sequences can also be used to leverage metagenomic datasets to provide valuable ecological insight on abundance, distribution, or activity (from metatranscriptomes) of viruses. For example, using metatranscriptomic data, Stough et al. (2017) were able to demonstrate oscillating cycles of lytic and lysogenic gene expression in Ma-LMM01-like phages during a *Microcystis* bloom in Lake Tai (China). Yet, despite availability of a few isolates we have no sequences of cyanophages infecting *C. raciborskii*. To close this knowledge gap, we sequenced and analyzed the genome of the cyanosiphovirus CrV (Steenhauer et al., 2016), and its host *C. raciborskii* strain Cr2010. Along with annotating the genome, we have used this sequence information to demonstrate the presence

This article is protected by copyright. All rights reserved

Accepted Article

of CrV-like viral genes within other filamentous cyanobacteria as well as the presence of CrV-like viruses within large blooms of *C. raciborskii.*

Results/Discussion

Genome Structure and Content of CrV

CrV is a dsDNA bacteriophage with features that place it in the family Siphoviridae (Steenhauer et al., 2016). The complete genome is 104,363 bp, making it considerably larger than the average size (~70.6 kb) of cyanosiphoviruses in the NCBI RefSeq database, but only half the size of the largest cyanosiphovirus genome, *Synechococcus* phage S-SKS1 (208 kb. NCBI Genbank NC_020851, unpublished). The assembled CrV genome size is consistent with an estimate of size using pulse-field gel electrophoresis (PFGE) (Steenhauer et al., 2016). Its GC content is 39%.

The CrV genome contains 123 predicted open reading frames (ORFs) (Figure 1; Table 1). Forty-nine of these had recognizable homologues (defined as having a BLASTx E-value < 0.001) in the NCBI nr protein database: all but one homologue produced an alignment E-value < 10^{-6}. Twenty-nine genes (~23%) were annotated with a predicted function; twenty ORFs (~16%) were annotated only as hypothetical proteins (Table 1). As is often seen in viruses, many of the predicted ORFs (74, ~60%) had no hits in nr below the specified E-value cutoff. The percentage of CrV ORFs with no recognizable homologues is higher than the mean (~47%) but within the range (33-71%) of seven other characterized cyansiphoviruses, six of which were isolated from marine or estuarine habitats (Sullivan et al., 2009; Huang et al., 2012; Ponsero et al., 2013) and one from the Baltic Sea (Coloma et al., 2017).

This article is protected by copyright. All rights reserved

Accepted Article

Eighteen genes could be identified as of viral origin. Of these, 10 code for structural proteins. Six genes similar to lambdoid tail proteins cluster together, illustrating a genomic synteny conserved in siphoviruses (Figure 1). Two genes (*gp014, gp015*) located on the negative strand, had similarity to virion structural proteins of the cyanophages S-TIM5 (Sabehi et al., 2012) and S-CBS4 (Huang et al., 2012), respectively, and are likely capsid proteins. One other structural protein, the lambdoid minor tail protein M (*gp036*), was encoded on the negative strand. CrV has a long non-contractile tail of ~600 nm (Figure S2); the gene encoding the tape measure protein (*gp064*) comprise ~13% of the CrV genome and codes for a predicted 4,416 amino acids. Larger tape measure proteins are known from only two other viruses: *Thermus thermophilus* phages P74-26 and P23-45 contain proteins at 5,006 and 5,002 amino acids, respectively (Minakhin et al., 2008).

Three viral genes (*gp017, gp046, gp095*) encode proteins involved in transcriptional regulation. All appear to be lambda-like repressors with each containing a Cro/C1-type DNA binding domain. Two viral genes function in DNA replication/packaging: the terminase large subunit (*gp003*) and an adenine-specific methyltransferase (*gp096*). The three remaining viral genes encode a putative phage-associated protein and two copies of an unusual putative class 3 lipase that is similar to genes found in cyanophages P-SSM2 (Sullivan et al., 2010) and P-SSM5 (NCBI Genbank HQ632825.1, unpublished).

Evidence of Genetic Exchange Between CrV and Cyanobacteria

The CrV genome is highly mosaic, with large segments that appear to be of cyanobacterial origin (Figure S3), consistent with frequent genetic exchange between viruses and hosts (Hatfull and Hendrix, 2011). Twenty-six of the predicted ORFs in the CrV genome (~20%) are similar to

This article is protected by copyright. All rights reserved

Accepted Article

genes in cyanobacteria, while five are similar to genes in heterotrophic bacteria. Genes involved in DNA metabolism are well represented in the subset of those cyanobacteria-like genes with known function. The alpha subunit of the ribonucleoside triphosphate reductase (RNR) protein is more similar to homologues found in strains of Nostocalean cyanobacteria than to the RNR homologues found in other viruses (Figure S4). From its phylogenetic placement, the CrV RNR gene appears ancestral to those of the Nostocalean species, consistent with an early genetic exchange and a long interaction between CrV and Nostocalean ancestors. Within the Nostocalean group, the RNR gene of CrV is most highly divergent from those of *Cylindrospermopsis* strains, including that of the host, making the host a doubtful source of this gene. RNR genes have been reported in cyanosiphoviruses, e.g., *Synechococcus* phages S-SKS1 and S-CBS2 (Huang et al., 2012), but less frequently than in marine cyanomyoviruses and cyanopodoviruses (Rohwer et al., 2000; Sakowski et al., 2014; Perez Sepulveda et al., 2016).

The gene encoding deoxycytidine triphosphate (dCTP) deaminase is more similar to orthologues in the host than those in other viruses or other Nostocalean species (Figure S5). In this case, the topology of the trees suggests that a divergence in the ancestral dCTP deaminase gene occurred between *C. raciborskii* and other Nostocalean species prior to acquisition by CrV, Another interpretation is that CrV transferred the gene to an ancestor of the *Cylindrospermopsis* group. Other CrV genes involved in DNA metabolism that are similar to cyanobacteria homologues include RNA polymerase sigma F, DNA translocase, and DNA helicase, although the phylogenetic relationship among these genes is less clear (Figures S6-8).

Terminase Large Subunit Phylogeny

This article is protected by copyright. All rights reserved

Accepted Article

Amino acid sequence of the terminase large subunit protein (TerL) can be used to predict DNA packaging strategy (Casjens et al., 2005) and is commonly used to examine phylogeny of phages (Sullivan et al., 2009; Huang et al., 2012; Chenard et al., 2016). Placement of the CrV terminase sequence into a phylogenetic tree of sequences of phages whose DNA packaging strategies are known indicates that the CrV uses a T4-like headful packaging strategy and that the genome is circularly permuted (Figure S9) (Merrill et al., 2016).

To place CrV into an evolutionary context, its TerL amino acid sequence was included in a phylogenetic tree with those of other cyanoviruses (Figure S10). Comprising the tree are all sequenced freshwater cyanophages included in the NCBI RefSeq Viral Genome database (as of 12/3/2017), representative marine cyanophages from each of the Caudovirales families, and several phages infecting cyanobacteria but not yet included in RefSeq. Marine cyanomyoviruses group tightly together into an almost monophyletic group while myoviruses infecting freshwater *Microcystis* form a distinctive cluster. Marine and freshwater podoviruses each form well supported but distinct clusters. Cyanosiphoviruses fail to group together and are found distributed throughout this phylogeny in a manner similar to that shown by others (Huang et al., 2015). CrV forms a highly supported and distinct group together with a siphovirus isolated from the Baltic Sea that infects a strain of *Nodularia* (Coloma et al., 2017), suggesting that CrV and NpeS-2AV2 are members of a possibly widely distributed and under-sampled clade of cyanosiphoviruses.

As part of our efforts, homologues of CrV *terL* were found in the genome sequences of 15 other cyanobacteria. Ten of these species are Nostocalean, two are filamentous but non-heterocystous, and three are unicellular. These lingering terminase genes likely represent remnant prophages (Sullivan et al., 2009; Huang et al., 2012) or perhaps failed infections.

This article is protected by copyright. All rights reserved

Accepted Article

Notably, no CrV-like terminase gene were found in any of the draft *C. raciborskii* genomes, including that of the host isolate used in this study. The sequences of these CrV-like terminases plus two terminases identified in a metagenome (see below) were included in the phylogenetic tree and form a strongly supported monophyletic clade (Figure 2). Moreover, sequences from CrV, the metagenomes, and the *Calothrix* prophage form a distinct branch within this monophyletic group, appearing to represent the different evolutionary trajectory of *terL* genes harbored in hosts *vs.* that of actively infecting viruses, or of recently infecting viruses in the case of *Calothrix*. To better understand evolutionary relationships, we checked the topology of a tree of 16S rRNA genes from the cyanobacterial "hosts" of the CrV-like terminases to that of the primary branch containing the CrV-like terminases. The topology of the major branches of the two trees are largely congruent (Figure 3), indicating that the viruses represented by the remnant *terL* genes and the cyanobacterial "hosts" likely shared a long evolutionary history.

Evidence is Consistent With a Recent Abandonment of the Lysogenic Lifestyle by CrV

The CrV-like *terL* gene in the genome of *Calothrix* PCC7103 led us to investigate the region of the *Calothrix* genome harboring the gene. This region was annotated as primarily hypothetical proteins (Shih et al., 2013). Re-annotation revealed a ~105-kb segment containing 20 phage-specific genes including terminases, prophage-like repressors and anti-repressors, and structural proteins, all of which appear to be a nearly intact prophage (Figure S11). No integrase gene could be recognized. We obtained a culture of PCC7103 from the University of Texas culture collection and verified the presence of the prophage element *via* PCR. Attempts to induce the prophage with protocols using mitomycin C (Steenhauer et al., 2014) failed.

This article is protected by copyright. All rights reserved

Accepted Article

CrV-like terminases in 15 cyanobacteria plus the remnant prophage in *Calothrix* PCC7103 support the idea that CrV may represent a clade of siphoviruses that includes members that can lysogenize filamentous cyanobacteria. The CrV genome contains genes indicative of temperate phages (Sullivan et al., 2009), including Cro/C1-like repressors, although it is missing others, like integrase. This could be due either to our inability to identify recognizable homologues, or more likely that CrV lost the ability to form a lysogenic relationship. Absence of CrV-like genes in sequenced strains of *C. raciborskii* and absence of key lysogeny-associated genes within CrV are consistent with the idea that CrV gained the ability to infect *C. raciborskii* by means of lytic-only infection in recent evolutionary times. Along these lines, Steenhauer et al. (2016) tested the host range of CrV against 10 species/strains of filamentous cyanobacteria and found CrV to have a restricted range as only *C. raciborskii* Cr2010 supported lytic infections. It is important to note that the method of isolation of CrV (Steenhauer et al., 2016) and continued culturing technique have favored or selected for the lytic lifestyle. That closely related CrV-like viruses in nature have lysogenic ability is a possibility.

The loss (or gain) of ability by an infecting phage to establish lysogeny alters the role played by the phage in influencing population structure of bloom-forming cyanobacteria. The significance of this gain or loss is illustrated by a mechanism that may help explain how cyanobacterial blooms maintain high cell concentrations amid a putative community of infecting phages. When infective viruses are present, formation of dense blooms increases the likelihood of virus-host contact and subsequent infection. In the presence of lytic phages, blooms of susceptible cyanobacteria should rapidly yield to infection *via* the "Kill-the-Winner" mechanism (Thingstad and Lignell, 1997), yet observations seemingly demonstrate otherwise. In contrast, supported by observations that virus to microbe ratios frequently decrease at high cell

This article is protected by copyright. All rights reserved

Accepted Article

concentrations, Knowles et al. (2016) proposed the alternative "Piggyback-the-Winner" model in which the lysogenic cycle of temperate phages is favored at high cell abundance. With evidence of cycles in lytic *vs.* lysogenic viral gene expression during a *Microcystis* bloom and by application of the "Piggyback-the-Winner" model, Stough et al. (2017) postulated a link between lysogeny and *Microcystis* success, positing that shifts to lysogenic infection might offer resistance against superinfection or infection by related lytic phages. In this scenario, blooms could temporarily escape density dependent lysis, allowing prolonged periods of high cell concentration.

Exploration into the relative contributions of lytic and lysogenic infection to viral and phytoplankton dynamics is in the early stages, and cyanobacterial blooms seem to be a promising system in which to test various models. Advances in this field will require access to relevant sequence information on related viruses and their hosts. If CrV represents a clade of siphoviruses that includes members that lysogenize filamentous cyanobacteria, then availability of the CrV genome potentially provides the sequences to investigate this concept employing molecular and metagenomic methods. While hypothetical at this juncture, availability of the CrV genome provides the sequence necessary to investigate this concept in a group of phages that infect filamentous bloom-forming cyanobacteria.

CrV Genome Contains a 12-kb Inverted Duplication

The genome of CrV contains a 12,754-bp-long inverted duplication. The repeat segments occupy positions 71,319-84,072 (arbitrarily designated as segment B) and 91,853-1:243 (designated as segment B') (Figure 1, Figure S3). A 7,780 bp non-duplicated segment (designated as C) separates the inverted repeats (Figure 1). Four independent lines of evidence

This article is protected by copyright. All rights reserved

Accepted Article

were used to confirm the presence of the inverted duplications and rule out their being an artifact of assembly. Initial assembly resulted in four contigs arbitrarily labeled A-D. In all cases, mapping reads to the original assembled contigs consistently produced a depth of coverage precisely 2-fold greater in contig B than in other contigs. This was the first indication of a possible collapsed repeat. Primers were designed to amplify outward from the end of all contigs to deduce contig connectivity. Sanger sequencing was used to sequence PCR amplicons, which were aligned to the contigs. In each case, the amplicons produced exact alignment overlaps establishing contig connectivity. If the inverted duplications and their predicted arrangement were correct, then a single primer targeting the proximal ends of segments B and B' should amplify across segment C and produce an amplicon ~8,000 bp long (Figure 1), which would represent the length of segment C plus twice the distance of the primer annealing site from the ends of segments B/B'. A primer was designed to test this conjecture. PCR using this single primer and an aliquot of the same CrV DNA used in sequencing produced the predicted amplicon size of 8,356 bp. If segment B is duplicated in the genome, then unique genes located in segment B should exist in a copy number twice that of unique genes located outside of this segment. qPCR primers were designed to target unique gene in segment B (*gp083* or *gp108*), segment C (*gp097*), and in a region outside of either of these (*gp064*). Using qPCR across a range of CrV DNA template concentrations, primers targeting segment B consistently produced a threshold cycle ~1 less than either of the other primers. This difference in threshold cycles was significant for all tested conditions (Table S2). Finally, the size of the CrV genome was estimated at ~110 kb via PFGE by Steenhauer et al. (2016) in their initial characterization work. Segment B existing as inverted repeats results in a genome size consistent with the Steenhauer

This article is protected by copyright. All rights reserved

Accepted Article

estimate. The length of contigs A-D without a duplicated B would result in a genome size of ~91.6 kb, notably less than the Steenhauer PFGE estimate.

The genetic content of the duplicated segment is intriguing. Segment B' contains 24 predicted ORFs. Twenty-three of these genes (all except *gp125*) are repeated in segment B. The sequence of *gp125* is repeated in segment B but was not included in a predicted ORF by GeneMark due to the difference in its genomic location and sequence context. Three genes with predicted function are present in the B/B' segment: dCTP deaminase, lipase, and an essential recombination protein. The deaminase and recombination protein resolved as cyanobacterial-like (see above). The origin of the lipase was assigned as viral but is uncertain. Homologues occur in other viral genomes, but may actually be of bacterial origin. The B/B' segment also contains 19 predicted ORFs with no hits in the nr database, providing no information on origin. It is possible these are virus-like but with no annotated homologues yet present in the nr database.

CrV Genome Encodes a CRISPR Array

A CRISPR array was identified in the genome of CrV using the online tool CRISPRFinder (Grissa et al., 2007a). Presence of the CRISPR sequence was confirmed with PCR and Sanger sequencing. The CRISPR was located just downstream of a putative phage associated protein (*gp097*) in the region located between the inverted duplications (Figure 1). The five direct repeats (DR) were 37-bp long and highly conserved, with the distal three repeats suffering degeneration and increasing departure from the consensus sequence. The DR sequences had no significant matches to RNA families in the Rfam database (Kalvari et al., 2017) and had no hits in either the CRISPRdb (Grissa et al., 2007b) or NCBI non-redundant (nr) nucleotide database.

This article is protected by copyright. All rights reserved

Accepted Article

The DR sequence was not similar to CRISPR arrays in the host (see below), making the host an unlikely source of the array. The first two spacers were 39 bp long and had identical sequences. The third and fourth spacers were 31 and 36 bp and unrelated to the others. None of the spacer sequences had significant alignments to sequences in the nr nucleotide database or to the host genome. No Cas-like proteins were identified in CrV.

A CRISPR array has been reported in the genome cyanophage N-1, a myovirus which infects *Nostoc* sp. PCC7210 (Chenard et al., 2016). To our knowledge, this is the only other instance of a CRISPR being found in a cyanophage genome. The CRISPR array in N-1 was structurally similar to that of CrV. The N-1 array contains four spacers with 37 bp DRs. Like CrV, the N-1 spacers had no significant matches in the nr nucleotide database, a common observation across the entirety of CRISPR spacers analyzed (Shmakov et al., 2017). No *cas* genes were found in the N-1 genome. In contrast, the two viral CRISPRs were dissimilar in DR sequence. The DRs in N-1 were similar to the DR5 family, which is common in cyanobacteria, and were most similar to those in filamentous cyanobacteria, indicating the likely source of the N-1 array and providing additional evidence for the horizontal transfer of CRISPRs *via* viruses (Makarova et al., 2015; Chenard et al., 2016).

CRISPR arrays have been reported in viruses and prophages in a couple of additional systems. In a remarkable example, a myovirus infecting *Vibrio cholera* encoded its own CRISPS-Cas system that targeted a phage inducible genomic island in the host (Seed et al., 2013). Inhibition of this island by the virus CRISPR-Cas system allowed successful phage replication. Mutating the spacers targeting the inducible island prevented plaque formation by the virus. In this case, the myovirus encoded a complete CRISPR-Cas locus containing two CRISPR arrays and six *cas* genes. In a second system, strains of *Clostridium difficile* are

This article is protected by copyright. All rights reserved

Accepted Article

commonly lysogenized and a study found that at least 10 prophages (and all ϕC2-like prophages) identified across several *C. difficile* strains contained CRISPR arrays (Hargreaves et al., 2014). In a compelling contrast, no ϕC2-like phages isolated from lytic propagation contained arrays. The prophage arrays contained spacers with exact matches to protospacers of other *C. difficile*-infecting phages.

Benefit to the *Vibrio* phage harboring CRISPRs was directly demonstrated, but this system is unique in that the phage also contained a complement of *cas* genes (Seed et al., 2013). However, processing by a host Cas system of CRISPR RNAs from prophage arrays has been demonstrated (Soutourina et al., 2013), setting up a straightforward and easy to visualize mechanism in which CRISPRs could afford a fitness advantage to the possessing prophage by influencing and perhaps preventing infections by other phage (Hargreaves et al., 2014). In the case of lytic viruses, use of the host Cas system to process viral CRISPR RNAs could potentially protect against co-infection (Chenard et al., 2016). Indeed, Chenard and colleagues demonstrated that the N-1 CRISPR was transcribed during active infection and that the host contains a CRISPR-Cas system.

If possessing CRISPRs offers a distinct fitness advantage, one might expect the presence of CRISPRs within genomes of phages to be more prevalent. It is possible, though, that a CRISPR affords little if any fitness advantage to the possessor. In this scenario, CRISPRs might simply be ephemeral occupants of phage genomes, which could explain their apparently low prevalence. Demonstrating a fitness advantage in lytic phages would provide exciting clarification on the role of CRISPRs in phage genomes.

Host CRISPR-Cas System Targets Regions of CrV Genome

This article is protected by copyright. All rights reserved

Accepted Article

To establish a characterized virus/host experimental system, we sequenced the genome of *C. raciborskii* strain Cr2010, the only known host of CrV. The assembled genome is 3.55 Mb comprising 89 contigs. General features of the draft genome are shown in Table S3.

Cr2010 contains Type I-D and Type III-B CRISPR-Cas loci and also harbors six CRISPR arrays (Figure 4). General features of the host CRISPR arrays are provided in Figure S12. CRISPR-2 was found adjacent to Type III-B *cas* genes (Figure 5). Spacer #1 in CRISPR-2 (spacer 2.1) is 35 bp long and is an exact match to a protospacer in the CrV genome. CRISPR-5 was orphaned on its respective contig with no flanking *cas* genes (Figure 5). As such, we were unable to classify this CRISPR type. Spacer #1 in CRISPR-5 (spacer 5.1) is 38 bp long and aligns to a putative protospacer with two mismatches that occur at spacer positions 17 and 34. Type I systems typically require protospacer adjacent motifs (PAMs) and exact matches of spacer to protospacer within a seed region for successful interference, while Type III systems do not (Semenova et al., 2011; Rath et al., 2015). In identifying protospacer matches, the more conservative approach is to assume CRISPR-5 is of Type I. Even with this assumption, the mismatches in spacer 5.1 are found in the distal region of the spacer outside of the potential seed sequence, where multiple mismatches can be tolerated without interfering with the immunity function of CRISPR-Cas (Semenova et al., 2011). However, with a single match sequence, we cannot identify or confirm the presence of PAM consensus sequences. We therefore identify the CrV protospacer match to host spacer 5.1 as putative.

Both protospacer locations fall within the large inverted repeat of the CrV genome. Thus there are four putative protospacer locations in the CrV genome that are targeted by the Cr2010 CRISPR-Cas system. Protospacer 2.1 is found in *gp085/gp106*, while protospacer 5.1 is located

This article is protected by copyright. All rights reserved

Accepted Article

in *gp077/gp114*. Neither of these predicted genes had significant BLASTx hits in the nr protein database. No other host CRISPR spacers had significant hits in the nr nucleotide database.

In well characterized virus/host experimental systems, a functional CRISPR-Cas system targeting a single protospacer of a virus can reduce host sensitivity to infection by ~10^5-10^6-fold (Semenova et al., 2011; Maniv et al., 2016), while a system targeting two or more spacers can reduce host sensitivity by an additional order of magnitude (Barrangou et al., 2007). In our case, CrV efficiently and reproducibly lyses cultures of Cr2010. We were therefore surprised to observe that Cr2010 has a CRISPR-Cas system with CRISPR spacers putatively targeting four sites in the CrV genome. However, the consistency of infection and Cr2010's apparent lack of immunity to CrV hints at something unusual in this virus/host system. The simplest explanation is that the Cr2010 CRISPR-Cas immune system does not function properly. Yet, the adaptation module proteins (*cas1* and *cas2*) are present and appear functional based on evidence of spacer incorporation. Incorporation of new spacers into a CRISPR array generally occurs in a directional manner, with the newest spacers being added to the array proximal to the leader, thus providing a chronologic record of exposure to foreign DNA (Barrangou et al., 2007; Rath et al., 2015). The most recent spacer incorporated into CRISPR-2 (and putatively CRISPR-5) is from a CrV protospacer, demonstrating that this part of the system was operable in the recent history of Cr2010.

The adaptation module can function independently of the expression and interference modules, so recent spacer incorporation is not evidence of a functional CRISPR-Cas system (Yosef et al., 2012). CRISPR-2 is found within a Type III-B locus. Cmr5 is a component of the Type III-B multi-subunit effector complex and no homolog to the *cmr5* gene was found in Cr2010, raising speculation about the competence of this system. It is entirely possible that the

This article is protected by copyright. All rights reserved

Accepted Article

function of Cmr5 is provided by an as yet poorly characterized *cas* gene or by a domain fused to other Cas proteins (Makarova et al., 2015). There were no reports of the *cmr5* gene among the Type III-B loci in the nine Australian *C. raciborskii* genomes recently sequenced (Willis et al., 2018), suggesting that a cryptic protein may provide this function.

For Type I-D systems, the Csc3/Cas10d protein is the large subunit of the effector complex. No homologous gene encoding this protein could be identified in Cr2010. An important difference exists here relative to *cmr5* and that is that *csc3/cas10d* genes are found among Type I-D loci in all nine Australian *C. raciborskii* strains. This increases the likelihood that absence of *csc3/cas10d* represents a gene loss that renders the Type I-D system non-functional. CRISPR-Cas systems are highly diverse and modular. Enzymatic or structural functions essential to a system can be provided by proteins, perhaps cryptic, encoded elsewhere in a genome (Makarova et al., 2015). Additionally, there is always the possibility that a given gene is located in a region of the genome that fails to assemble and is thus masked from analysis. While it is difficult to draw firm conclusions from analsyes of this nature, the collective evidence outlined above raises doubt about the functionality of at least the Type I-D system and any immunity it may offer.

The Presence of CrV-like Viruses in Nature

To gauge the prevalence of CrV-like viruses in nature, we examined publicly available metagenomes and metatranscriptomes from freshwater lakes. We found no evidence of CrV-like viruses in metagenomes targeting the viral fraction in freshwater lakes in France (Roux et al., 2012) or Taiwan (Tseng et al., 2013). Likewise, metatranscriptomes targeting the cell-fraction of *Microcystis*-dominated blooms in Lake Tai (China) (Stough et al., 2017) and in Lake Erie

This article is protected by copyright. All rights reserved

Accepted Article

(Davenport, 2016; Steffen et al., 2017) provided no evidence of CrV-like viruses. We assumed the probability of finding CrV-like viruses would be greatest in environments with a significant presence of *Cylindrospermopsis* or closely related species, but were able to identify only one dataset meeting this profile, that of a metagenome collected from Lake Samsonvale (Australia) during a *C. raciborskii* bloom (Woodhouse and Willis *et al.*, in prep). From this dataset, we identified six viral contigs ranging in size from ~4,500 to 12,500 bp and with high similarity to CrV. The annotated contigs contained a total of 53 predicted genes (Table S4) and aligned with almost perfect synteny and congruent annotations across nearly the entire first half of the CrV genome (Figure 5). In contig 2, two small genes had no homologues in CrV while two homologous genes were inverted relative to CrV (Figure 5). The 51 homologous genes share an average amino acid identity of ~77% (range of 46-96%) with cognate genes in CrV (Table S4). The metagenome contigs failed to align to three small segments across the first ~56 kbp of the CrV genome. No contigs aligning to the latter half of the CrV genome were identified. This curious observation points perhaps to the uniqueness of half of the CrV genome and begs the question as to the possible additional genomic content of CrV-like viruses contributing to the metagenome. Characterizing viruses from *Cylindrospermopsis* blooms in Lake Samsonvale would seem to offer exciting opportunities in comparative viral genomics.

Collectively, the Lake Samsonvale metagenome demonstrates convincingly that CrV-like viruses occur in *Cylindrospermopsis* blooms in Australia. Closely related viruses occurring in blooms on opposite sides of the globe, while remarkable, is consistent with reports from both marine and freshwater systems (Short and Suttle, 2005; Sabehi et al., 2012; Holmfeldt et al., 2013). A recent study demonstrated that viruses are dispersed across distant ecosystems *via* atmospheric transport and deposition of dust and aerosolized particles of marine origin,

This article is protected by copyright. All rights reserved

providing a compelling explanation for this commonly observed phenomenon (Reche et al., 2018).

Accepted Article

Conclusion

CrV is a siphovirus that infects *C. raciborskii* strain Cr2010. CrV is unique in that its genome contains both a 12-kb inverted duplication and an intact CRISPR array. The CrV host contains a CRISPR-Cas system with spacers targeting protospacers in the duplicated region of the CrV genome. The host CRISPR-Cas system appears ineffective against CrV infection for, as yet, undetermined reasons, but possibly due to lack of key *cas* genes. CrV appears to represent a distinct clade of siphoviruses that infects and in some cases can lysogenize filamentous cyanobacteria. The CrV genome sequence may help identify a clade of phages that may prove useful in metagenomic studies investigating the role of lysogeny *vs.* lytic infections during cyanobacterial blooms. Furthermore, viruses with high genetic identity to CrV, albeit lacking the 12-kb inverted duplicated region, are present in metagenomes of *C. raciborskii* blooms in Australia, suggesting a wide distribution of CrV-like viruses..

Experimental Procedures

Host Culturing and DNA Extraction

Cylindrospermopsis raciborskii host strain Cr2010, which was isolated from Reeuwijkse Lakes (Netherlands) in 2010 (Steenhauer et al., 2016), was grown in 50 mL glass tubes in standard MLA medium (Bolch and Blackburn, 1996) at 26 °C under a 12 h light:dark cycle with an illumination of ~140 µmol of quanta m^{-2} s^{-1} provided by cool-white fluorescent bulbs (GE T12

This article is protected by copyright. All rights reserved

Accepted Article

Ecolux). Chlorophyll *a* fluorescence was measured using a TD-700 fluorometer (Turner Designs) as a proxy for biomass and cell concentration (Steenhauer et al., 2016).

To extract host DNA, cells were collected on 5 µm pore-size polycarbonate filters (GE Water and Process Technologies) and washed in 2 volumes of sterile MLA medium to reduce the quantity of contaminating heterotrophic bacteria following the method of Sinha et al. (2014). DNA was extracted following the protocol described in Martin and Wilhelm (2018). Briefly, cells were disrupted with lysozyme and treated with proteinase K. DNA was extracted with phenol/chloroform, precipitated with sodium acetate and 100% ethanol, and washed with 70% alcohol. DNA was quantified with a NanoDrop ND-1000 spectrophotometer and stored at -20 °C until sequencing.

CrV Infection, Virus Purification, and DNA Extraction

CrV was isolated from Reeuwijkse Lakes (Netherlands) in 2012. The methods for isolation and initial characterization of the virus are detailed in Steenhauer et al. (2016). Cultures of *C. raciborskii* were infected with CrV by adding ~5 x 10^9 virus particles to 25 mL of exponentially growing cultures. Lysis was determined visually by cultures losing green color and becoming clear, which occurred in 2-3 days. Viruses were enumerated using epifluorescence microscopy (Leica DM5500 B) after SYBR Green staining (Suttle and Fuhrman, 2010). Viruses from 500 mL of pooled lysate were concentrated to ~50 mL by tangential flow filtration using an Ultracel 30 kDa filter (EMD Millipore, Billerica, MA), then further concentrated by centrifugal filtration using Amicon Ultra-15 30-kDa filters (Merck Millipore, Cork, Ireland). Viruses were dislodged from Amicon filters following the method of Brum (2016). Approximately ~1.5 mL of SM

This article is protected by copyright. All rights reserved

Accepted Article

phage buffer (Sambrook and Russell, 2001) were vortexed across the filter, removed, and the process repeated twice with fresh buffer.

Viruses were separated from contaminating bacteria and cellular debris *via* centrifugation in CsCl buoyant density gradients using established methods (Sambrook and Russell, 2001; Lawrence and Steward, 2010). CsCl was added to the Amicon filter concentrate (~4.5 mL) at a density of 1.45 g mL^{-1}, which was loaded into OptiSeal 4.9 mL ultracentifuge tubes (Beckman Coulter, Brea, CA). Tubes were centrifuged at 240,200 x g (RCFmax) overnight in a Beckman Coulter VTi80 rotor. Viral bands (Fig. S1) were extracted by puncturing the bottom of the tube and controlling the drip *via* an air hose attached to the top of the tube (Lawrence and Steward, 2010). CsCl was dialyzed away from particles against two changes of phage dialysis buffer (Sambrook and Russell, 2001) using Slide-A-Lyzer MINI dialysis devices (Thermo Scientific).

Dialyzed particles (in ~400 µL residual dialyzed buffer) were treated with 36 units of Turbo DNase and incubated at 37 °C for 1 h to remove contaminating bacterial DNA. Virus capsids were digested by incubating with 10 units of proteinase K for 1 h at 37 °C. DNA was extracted with phenol/chloroform and precipitated with ethanol using standard methods (Sambrook and Russell, 2001). DNA was quantified with a NanoDrop ND-1000 spectrophotometer and stored at -20 °C until sequencing.

Transmission Electron Microscopy of CrV Particles

A sample solution of 3.5 µL was loaded onto a glow-discharged carbon-coated copper grid and incubated for 3 min. Excess solution was blotted with filter paper. The grid was stained three times with 10 µL of freshly prepared 2% uranyl acetate and dried overnight at room temperature. Samples were imaged at room temperature using a Hitachi H-7650 transmission electron

This article is protected by copyright. All rights reserved

Accepted Article

microscope operated at an acceleration voltage of 80 kV with nominal magnification of 30,000x. Images were recorded on a CCD camera with 1.5 s exposure and printed with magnification of 124,000x. Sample preparation was performed at the University of Texas El Paso. Electron microscopy was performed at the Microscopic Imaging Core Suite at New Mexico State University (Las Cruces, New Mexico).

Sequencing and Genome Assembly

DNA from host and CrV were sequenced at HudsonAlpha Institute for Biotechnology (Huntsville, AL). DNA libraries were prepared by HudsonAlpha using the NEBNext® DNA Library Prep kit for sample preparation and the KAPA HiFi HotStart master mix for amplification following manufacturer's recommendations. DNA from the host and CrV were multiplexed and sequenced in a single lane as 250 bp paired-end reads on the Illumina™ MiSeq platform. All reads from host and virus were trimmed for quality in CLC Genomics Workbench (v. 10.1.1) using a quality score of 0.02 and allowing zero ambiguous base calls. The CrV genome was assembled using the CLC Genomics Workbench *de novo* assembler (v. 10.1.1). Sanger sequencing of PCR amplicons for CrV genome closure was performed at the University of Tennessee Genomics Core using an Applied Biosystems 3730 capillary sequencer. For assembly of the host genome, quality trimmed host reads were assembled with the metaSPAdes assembler (Nurk et al., 2017). Nucleotide sequences of CrV-01T and host Cr2010 have been deposited in GenBank under the accession numbers MH636380 and PVMC01000000, respectively.

Genome Annotation

This article is protected by copyright. All rights reserved

Accepted Article

Genes in CrV were predicted with the GeneMark heuristic model (Besemer and Borodovsky, 1999). Blast2go (Conesa et al., 2005) was used to annotate predicted genes in CrV and to re-annotate segments of sequenced cyanobacterial genomes found to harbor remnant prophages closely related to CrV. Genes in the host genome were predicted and annotated using NCBI's prokaryotic genome annotation pipeline (Tatusova et al., 2016).

Analysis of Metagenomic Data

We examined metagenomic/metatranscriptomic datasets for the presence of CrV-like viruses. We selected datasets targeting freshwater viromes and cyanobacterial blooms. Datasets examined include metagenomes targeting viral fractions in Lakes Pavin and Bourget (France) (Roux et al., 2012) (Sequence Read Archive (SRA) accession number PRJEB2280) and in Feitsui Reservoir (Taiwan) (Tseng et al., 2013) (SRA accession SRP009395). Metatranscriptomes examined include those targeting the cell fractions of: Lake Erie 2014 (Davenport, 2016) (SRA accession numbers PRJNA405733, 405772, 405774, 405775, 429068, 429071, 429072); Lake Erie 2014 (Steffen et al., 2017) (SRA accession PRJNA354726); Lake Tai (China) 2013 (Krausfeldt et al., in prep.) (MG-RAST project number mgp82644); and Lake Tai 2014 (Stough et al., 2017) (MG-RAST project number mgp14658). Details of library preparation for the above datasets are provided in the respective publications. Libraries for Lake Tai 2013 were prepared following the method of Stough et al. (2017).

We also examined metagenomes targeting a *Cylindrospermopsis* spp. bloom in Lake Samsonvale (Queensland, Australia) (Woodhouse and Willis, *et al.*, in prep). Lake Samsonvale metagenomes originated from 5 m integrated water samples collected near North Pine Dam Wall from December 2014 to May 2015. 1 L samples were filtered through 110 mm Whatman GFF.

This article is protected by copyright. All rights reserved

Accepted Article

Filters were stored frozen (-80 °C) until extraction. DNA was extracted using a standard lysozyme/proteinase K/phenol/chloroform method (Sambrook and Russell, 2001). Samples were treated with lysozyme for 20 min at 37° C; proteinase K and SDS was added and incubated at 50° C for 2 h; samples were extracted with phenol:chloroform:isoamyl alcohol (25:24:1); DNA was washed with chloroform:isoamyl alcohol (24:1) and precipitated with ethanol. DNA libraries were prepared using the Nextera Library Preparation Kit (Illumina) following the manufacturer's protocol. Libraries were sequenced using the Illumina MiSeq platform at the Walter and Eliza Hall Institute Genome Sequencing Facility (Parkville, Victoria, Australia). Lake Samsonvale libraries have been deposited in SRA (Bioproject accession number PRJNA482263).

We used a conservative multi-step screening approach in examining metagenomic datasets for the presence of CrV-like viruses. Preliminary screening was conducted by mapping library reads to the CrV genome in CLC Genomics Workbench using a moderate stringency of length fraction = 0.7 and similarity = 0.85. Mapping results were manually inspected. Metagenome mappings showing no coverage or coverage of conserved genes only were considered as "no evidence" and not examined further. Metagenomes showing promise were assembled and contigs screened by BLAST. Reads from Lake Samsonvale were assembled using MEGAHIT v1.0 with kmers = 33, 55, 77, and 99 (Li et al., 2016). Reads from other metagenomes were assembled using the CLC Genomics Workbench *de novo* assembler using default settings. Contigs producing tBLASTx E-values $<10^{-20}$ were manually inspected for length and alignment across CrV ORFs. Manually curated contigs were annotated with Blast2go as described above. Amino acid identity between reciprocal BLAST hit pairs from CrV and

This article is protected by copyright. All rights reserved

Accepted Article

metagenome contigs was calculated with the Enveomics online toolbox (Rodriguez-R and Konstantinidis, 2016).

qPCR Assessment

Relative CrV contig copy number was determined by comparison of threshold cycles (Ct) generated from amplifying fragments of genes located in three distinct regions (contigs A, B, or C) of the CrV genome. See results for detailed explanation of contigs. qPCR assays were performed on a BioRad DNA Engine Opticon 2 system using 25 µL reaction mixtures formulated with Absolute qPCR SYBR Green Mix (Thermo Scientific). Thermal cycling conditions were 95° C for 15 min followed by 40 cycles at 95° C for 15 s, 55° C for 30 s, and 72° C for 30 s. Melting curves were examined across the range of 50-95 ° C at 1° C intervals. Primer sequences are listed in Table S1. Reactions were performed in triplicate across four template concentrations. Significance was determined by standard T-test using GraphPad Prism software (v. 7.03). Welch's correction was applied when sample variances were unequal.

This article is protected by copyright. All rights reserved

Accepted Article

Acknowledgements.

We thank Lisa Steenhauer and Gary LeCleir for assistance. Supported by funds from the National Science Foundation (IOS 141528, DEB 1240870) and the *Kenneth & Blaire Mossman Endowment* to the University of Tennessee (SWW). AW was funded by ARC-Linkage grant LP130100311. CPDB was funded by NIOZ.

The authors have no conflicts of interest.

This article is protected by copyright. All rights reserved

Accepted Article

References

Aguilera, A., Gómez, E.B., Kaštovský, J., Echenique, R.O., and Salerno, G.L. (2018) The polyphasic analysis of two native *Raphidiopsis* isolates supports the unification of the genera *Raphidiopsis* and *Cylindrospermopsis* (Nostocales, Cyanobacteria). *Phycologia* **57**: 130-146.

Antunes, J.T., Leao, P.N., and Vasconcelos, V.M. (2015) *Cylindrospermopsis raciborskii*: review of the distribution, phylogeography, and ecophysiology of a global invasive species. *Front Microbiol* **6**: 473.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S. et al. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**: 1709-1712.

Besemer, J., and Borodovsky, M. (1999) Heuristic approach to deriving models for gene finding. *Nucleic Acids Res* **27**: 3911-3920.

Bolch, C.J.S., and Blackburn, S.I. (1996) Isolation and purification of Australian isolates of the toxic cyanobacterium *Microcystis aeruginosa* Kütz. *J Appl Phycol* **8**: 5-13.

Briand, J.F., Leboulanger, C., Humbert, J.F., Bernard, C., and Dufour, P. (2004) *Cylindrospermopsis raciborskii* (cyanobacteria) invasion at mid‐latitudes: election, wide physiological tolerance, or global warming? *J Phycol* **40**: 231-238.

Brum, J. (2016). Concentrating viruses with an amicon or nanosep centrifugal ultrafiltration device. *protocols.io.* doi: dx.doi.org/10.17504/protocols.io.c54y8v. Accessed: April 24, 2018.

Burford, M.A., McNeale, K.L., and McKenzie-Smith, F.J. (2006) The role of nitrogen in promoting the toxic cyanophyte *Cylindrospermopsis raciborskii* in a subtropical water reservoir. *Freshw Biol* **51**: 2143-2153.

Burford, M.A., Beardall, J., Willis, A., Orr, P.T., Magalhaes, V.F., Rangel, L.M. et al. (2016) Understanding the winning strategies used by the bloom-forming cyanobacterium *Cylindrospermopsis raciborskii*. *Harmful Algae* **54**: 44-53.

Casjens, S.R., Gilcrease, E.B., Winn-Stapley, D.A., Schicklmaier, P., Schmieger, H., Pedulla, M.L. et al. (2005) The generalized transducing *Salmonella* bacteriophage ES18: complete genome sequence and DNA packaging strategy. *Journal of Bacteriolgy* **187**: 1091-1104.

Chenard, C., Wirth, J.F., and Suttle, C.A. (2016) Viruses infecting a freshwater filamentous cyanobacterium (*Nostoc* sp.) encode a functional crispr array and a proteobacterial DNA polymerase B. *MBio* **7**.

Coloma, S.E., Dienstbier, A., Bamford, D.H., Sivonen, K., Roine, E., and Hiltunen, T. (2017) Newly isolated *Nodularia* phage influences cyanobacterial community dynamics. *Environ Microbiol* **19**: 273-286.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**: 3674-3676.

Davenport, E.J. (2016) Diel regulation of metabolic functions of a western Lake Erie *Microcystis* bloom informed by metatranscriptomic analysis. Master of Science Thesis. Bowling Green State University. Bowling Green, Ohio.

This article is protected by copyright. All rights reserved

Accepted Article

Grissa, I., Vergnaud, G., and Pourcel, C. (2007a) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35**: W52-W57.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007b) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**: 172.

Hargreaves, K.R., Flores, C.O., Lawley, T.D., and Clokie, M.R. (2014) Abundant and diverse clustered regularly interspaced short palindromic repeat spacers in *Clostridium difficile* strains and prophages target multiple phage types within this pathogen. *MBio* **5**: e01045-01013.

Hatfull, G.F., and Hendrix, R.W. (2011) Bacteriophages and their genomes. *Curr Opin Virol* **1**: 298-303.

Hawkins, P., Runnegar, M., Jackson, A., and Falconer, I. (1985) Severe hepatotoxicity caused by the tropical cyanobacterium (blue-green alga) *Cylindrospermopsis raciborskii* (Woloszynska) Seenaya and Subba Raju isolated from a domestic water supply reservoir. *Appl Environ Microbiol* **50**: 1292-1295.

Holmfeldt, K., Solonenko, N., Shah, M., Corrier, K., Riemann, L., VerBerkmoes, N.C., and Sullivan, M.B. (2013) Twelve previously unknown phage genera are ubiquitous in global oceans. *Proc Natl Acad Sci U S A* **110**: 12798-12803.

Huang, S., Wang, K., Jiao, N., and Chen, F. (2012) Genome sequences of siphoviruses infecting marine *Synechococcus* unveil a diverse cyanophage group and extensive phage–host genetic exchanges. *Environ Microbiol* **14**: 540-558.

Huang, S., Zhang, S., Jiao, N., and Chen, F. (2015) Comparative genomic and phylogenomic analyses reveal a conserved core genome shared by estuarine and oceanic cyanopodoviruses. *PLoS One* **10**: e0142962.

Istvanovics, V., Shafik, H., Presing, M., and Juhos, S. (2000) Growth and phosphate uptake kinetics of the cyanobacterium *Cylindrospermopsis raciborskii* (Cyanophyceae) in throughflow cultures. *Freshw Biol* **43**: 257-275.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R. et al. (2017) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res* **46**: D335-D342.

Knowles, B., Silveira, C.B., Bailey, B.A., Barott, K., Cantu, V.A., Cobian-Guemes, A.G. et al. (2016) Lytic to temperate switching of viral communities. *Nature* **531**: 466-470.

Lawrence, J.E., and Steward, G.F. (2010) Purification of viruses by centrifugation. In *Manual of Aquatic Viral Ecology*. Wilhelm, S.W., Weinbauer, M.G., and Suttle, C.A. (eds): ASLO, pp. 166-181.

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K. et al. (2016) MEGAHIT v1. 0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**: 3-11.

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J. et al. (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nature Reviews Microbiology* **13**: 722-736.

Maniv, I., Jiang, W., Bikard, D., and Marraffini, L.A. (2016) Impact of different target sequences on Type III CRISPR-Cas immunity. *J Bacteriol* **198**: 941-950.

This article is protected by copyright. All rights reserved

Accepted Article

Martin, R.M., and Wilhelm, S.W. (2018). Phenol/chloroform extraction of DNA from cyanobacteria. *protocols.io.* doi: dx.doi.org/10.17504/protocols.io.ptndnme. Accessed: April 24, 2018.

Merrill, B.D., Ward, A.T., Grose, J.H., and Hope, S. (2016) Software-based analysis of bacteriophage genomes, physical ends, and packaging strategies. *BMC Genomics* **17**: 679.

Minakhin, L., Goel, M., Berdygulova, Z., Ramanculov, E., Florens, L., Glazko, G. et al. (2008) Genome comparison and proteomic characterization of *Thermus thermophilus* bacteriophages P23-45 and P74-26: siphoviruses with triplex-forming sequences and the longest known tails. *J Mol Biol* **378**: 468-480.

Mojica, K.D., and Brussaard, C.P. (2014) Factors affecting virus dynamics and microbial host–virus interactions in marine environments. *FEMS Microbiol Ecol* **89**: 495-515.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**: 824-834.

Padisák, J. (1997) *Cylindrospermopsis raciborskii* (Woloszynska) Seenayya et Subba Raju, an expanding, highly adaptive cyanobacterium: worldwide distribution and review of its ecology. *Archiv Für Hydrobiologie Supplementband Monographische Beitrage* **107**: 563-593.

Perez Sepulveda, B., Redgwell, T., Rihtman, B., Pitt, F., Scanlan, D.J., and Millard, A. (2016) Marine phage genomics: the tip of the iceberg. *FEMS Microbiol Lett* **363**.

Pollard, P.C., and Young, L.M. (2010) Lake viruses lyse cyanobacteria, *Cylindrospermopsis raciborskii*, enhances filamentous-host dispersal in Australia. *Acta Oecologica* **36**: 114-119.

Ponsero, A.J., Chen, F., Lennon, J.T., and Wilhelm, S.W. (2013) Complete genome sequence of cyanobacterial siphovirus KBS2A. *Genome Announcements* **1**: e00472-00413.

Rath, D., Amlinger, L., Rath, A., and Lundgren, M. (2015) The CRISPR-Cas immune system: biology, mechanisms and applications. *Biochimie* **117**: 119-128.

Reche, I., D'Orta, G., Mladenov, N., Winget, D.M., and Suttle, C.A. (2018) Deposition rates of viruses and bacteria above the atmospheric boundary layer. *ISME J* **12**: 1154-1162.

Rodriguez-R, L.M., and Konstantinidis, K.T. (2016) The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Preprints*.

Rohwer, F., Segall, A., Steward, G., Seguritan, V., Breitbart, M., Wolven, F., and Azam, F. (2000) The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. *Limnol Oceanogr* **45**: 408-418.

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S. et al. (2012) Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* **7**: e33641.

Sabehi, G., Shaulov, L., Silver, D.H., Yanai, I., Harel, A., and Lindell, D. (2012) A novel lineage of myoviruses infecting cyanobacteria is widespread in the oceans. *Proc Natl Acad Sci U S A* **109**: 2037-2042.

Sakowski, E.G., Munsell, E.V., Hyatt, M., Kress, W., Williamson, S.J., Nasko, D.J. et al. (2014) Ribonucleotide reductases reveal novel viral diversity and predict biological and ecological features of unknown marine viruses. *Proc Natl Acad Sci U S A* **111**: 15786-15791.

This article is protected by copyright. All rights reserved

Accepted Article

Sambrook, J., and Russell, D.W. (2001) *Molecular Cloning, A Laboratory Manual*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

Seed, K.D., Lazinski, D.W., Calderwood, S.B., and Camilli, A. (2013) A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* **494**: 489-491.

Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B. et al. (2011) Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc Natl Acad Sci U S A* **108**: 10098-10103.

Shih, P.M., Wu, D., Latifi, A., Axen, S.D., Fewer, D.P., Talla, E. et al. (2013) Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc Natl Acad Sci U S A* **110**: 1053-1058.

Shmakov, S.A., Sitnik, V., Makarova, K.S., Wolf, Y.I., Severinov, K.V., and Koonin, E.V. (2017) The CRISPR spacer space is dominated by sequences from species-specific mobilomes. *mBio* **8**: e01397-01317.

Short, C.M., and Suttle, C.A. (2005) Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Appl Environ Microbiol* **71**: 480-486.

Sinha, R., Pearson, L.A., Davis, T.W., Burford, M.A., Orr, P.T., and Neilan, B.A. (2012) Increased incidence of *Cylindrospermopsis raciborskii* in temperate zones-is climate change responsible? *Water Res* **46**: 1408-1419.

Sinha, R., Pearson, L.A., Davis, T.W., Muenchhoff, J., Pratama, R., Jex, A. et al. (2014) Comparative genomics of *Cylindrospermopsis raciborskii* strains with differential toxicities. *BMC Genomics* **15**: 83.

Soutourina, O.A., Monot, M., Boudry, P., Saujet, L., Pichon, C., Sismeiro, O. et al. (2013) Genome-wide identification of regulatory RNAs in the human pathogen *Clostridium difficile*. *PLoS Genetics* **9**: e1003493.

Steenhauer, L.M., Pollard, P.C., Brussaard, C.P., and Säwström, C. (2014) Lysogenic infection in sub-tropical freshwater cyanobacteria cultures and natural blooms. *Mar Freshw Res* **65**: 624-632.

Steenhauer, L.M., Wierenga, J., Carreira, C., Limpens, R., Koster, A.J., Pollard, P.C., and Brussaard, C.P.D. (2016) Isolation of cyanophage CrV infecting *Cylindrospermopsis raciborskii* and the influence of temperature and irradiance on CrV proliferation. *Aquat Microb Ecol* **78**: 11-23.

Steffen, M.M., Davis, T.W., McKay, R.M.L., Bullerjahn, G.S., Krausfeldt, L.E., Stough, J.M.A. et al. (2017) Ecophysiological examination of the Lake Erie *Microcystis* bloom in 2014: linkages between biology and the water supply shutdown of Toledo, OH. *Environmental Science and Technology* **51**: 6745-6755.

Stough, J.M.A., Tang, X., Krausfeldt, L.E., Steffen, M.M., Gao, G., Boyer, G.L., and Wilhelm, S.W. (2017) Molecular prediction of lytic vs lysogenic states for *Microcystis* phage: metatranscriptomic evidence of lysogeny during large bloom events. *PLoS One* **12**: e0184146.

Sullivan, M.B., Weitz, J.S., and Wilhelm, S.W. (2017) Viral ecology comes of age. *Environ Microbiol Rep* **9**: 33-35.

This article is protected by copyright. All rights reserved

Accepted Article

Sullivan, M.B., Krastins, B., Hughes, J.L., Kelly, L., Chase, M., Sarracino, D., and Chisholm, S.W. (2009) The genome and structural proteome of an ocean siphovirus: a new window into the cyanobacterial 'mobilome'. *Environ Microbiol* **11**: 2935-2951.

Sullivan, M.B., Huang, K.H., Ignacio-Espinoza, J.C., Berlin, A.M., Kelly, L., Weigele, P.R. et al. (2010) Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environ Microbiol* **12**: 3035-3056.

Suttle, C.A., and Fuhrman, J.A. (2010) Enumeration of virus particles in aquatic or sediment samples by epifluorescence microscopy. In *Manual of Aquatic Viral Ecology*. Wilhelm, S.W., Weinbauer, M.G., and Suttle, C.A. (eds): ASLO, pp. 145-153.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E.P., Zaslavsky, L. et al. (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res* **44**: 6614-6624.

Thingstad, T., and Lignell, R. (1997) Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquat Microb Ecol* **13**: 19-27.

Tseng, C.-H., Chiang, P.-W., Shiah, F.-K., Chen, Y.-L., Liou, J.-R., Hsu, T.-C. et al. (2013) Microbial and viral metagenomes of a subtropical freshwater reservoir subject to climatic disturbances. *ISME J* **7**: 2374.

Weitz, J.S., and Wilhelm, S.W. (2012) Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 Biol Rep* **4**: 17.

Wilhelm, S.W., and Suttle, C.A. (1999) Viruses and nutrient cycles in the sea: viruses play critical roles in the structure and function of aquatic food webs. *Bioscience* **49**: 781-788.

Willis, A., Chuang, A.W., and Burford, M.A. (2016) Nitrogen fixation by the diazotroph *Cylindrospermopsis raciborskii* (Cyanophyceae). *J Phycol* **52**: 854-862.

Willis, A., Woodhouse, J.N., Ongley, S.E., Jex, A.R., Burford, M.A., and Neilan, B.A. (2018) Genome variation in nine co-occurring toxic *Cylindrospermopsis raciborskii* strains. *Harmful Algae* **73**: 157-166.

Yosef, I., Goren, M.G., and Qimron, U. (2012) Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res* **40**: 5569-5576.

Figures for Martin *et al.* 2018.

Figure 1. Genome map of *Cylindrospermopsis raciborskii* Virus (CrV). Numbers above the protospacer symbols indicate the host CRISPR spacer to which the CrV sequence is a cognate. For example, protospacer sequence 2.1 is a cognate match to spacer #1 in host CRISPR-2. Abbreviations: RNR, ribonucleoside triphosphate reductase; dCTP, deoxycytidine triphosphate deaminase.

Figure 2. Maximum likelihood tree of the terminase large subunit (TerL) amino acid sequences of cyanophages and cyanobacteria harboring CrV-like terminases. Sequences from cyanobacteria are shown in green, myoviruses in blue, podoviruses in black, siphoviruses in red, and from Lake Samsonvale (Australia) metagenomes in purple. Support values of 100 bootstrap iterations are shown at the nodes.

This article is protected by copyright. All rights reserved

Figure 3. Maximum likelihood tree of the terminase large subunit (TerL) amino acid sequence (left) vs. a tree of 16S rRNA sequences (right) of cyanobacteria harboring CrV-like terminase large subunit genes (*terL*).

Figure 4. Annotated contigs of the *Cylindrospermopsis raciborskii* Cr2010 genome containing either CRISPR arrays or CRISPR associated (*cas*) genes.

Figure 5. Comparison of synteny and annotation of ORFs 1-61 of the CrV genome (above horizontal line) *vs.* annotated contigs of metagenomes (below horizontal line) from Lake Samsonvale (Australia).

Table 1. Summary of those predicted genes with functional annotations in *Cylindrospermopsis raciborskii* Virus (CrV).

Accepted Article

This article is protected by copyright. All rights reserved

Accepted Article

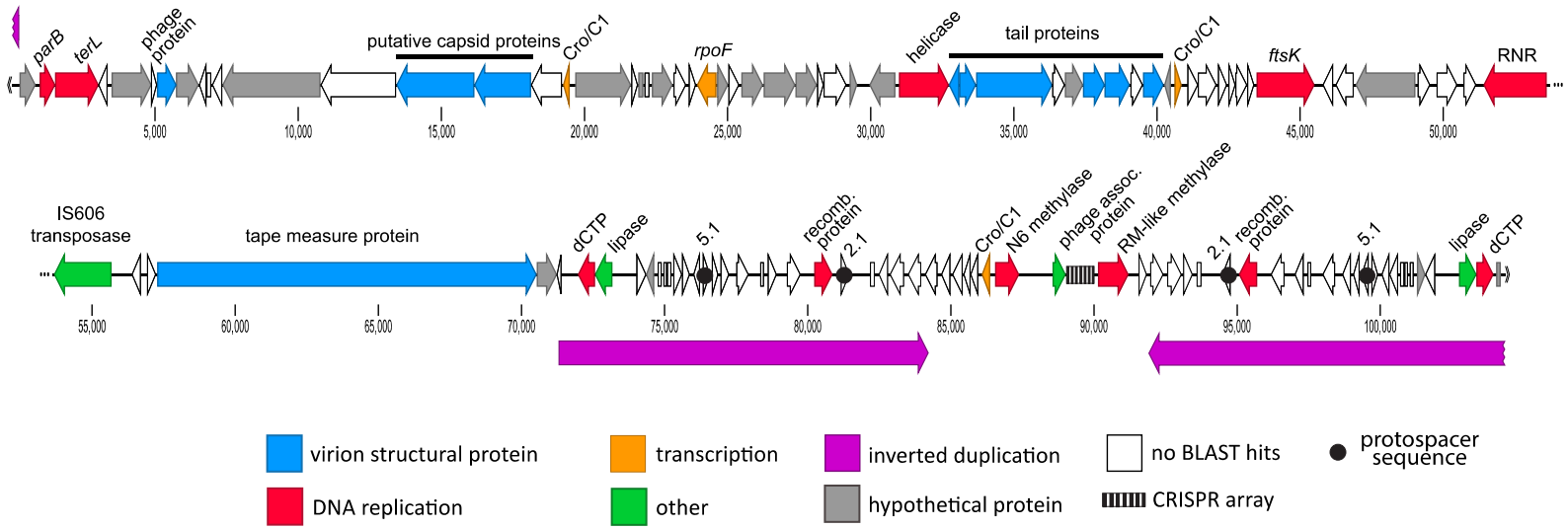

Figure 1.

This article is protected by copyright. All rights reserved

Accepted Article

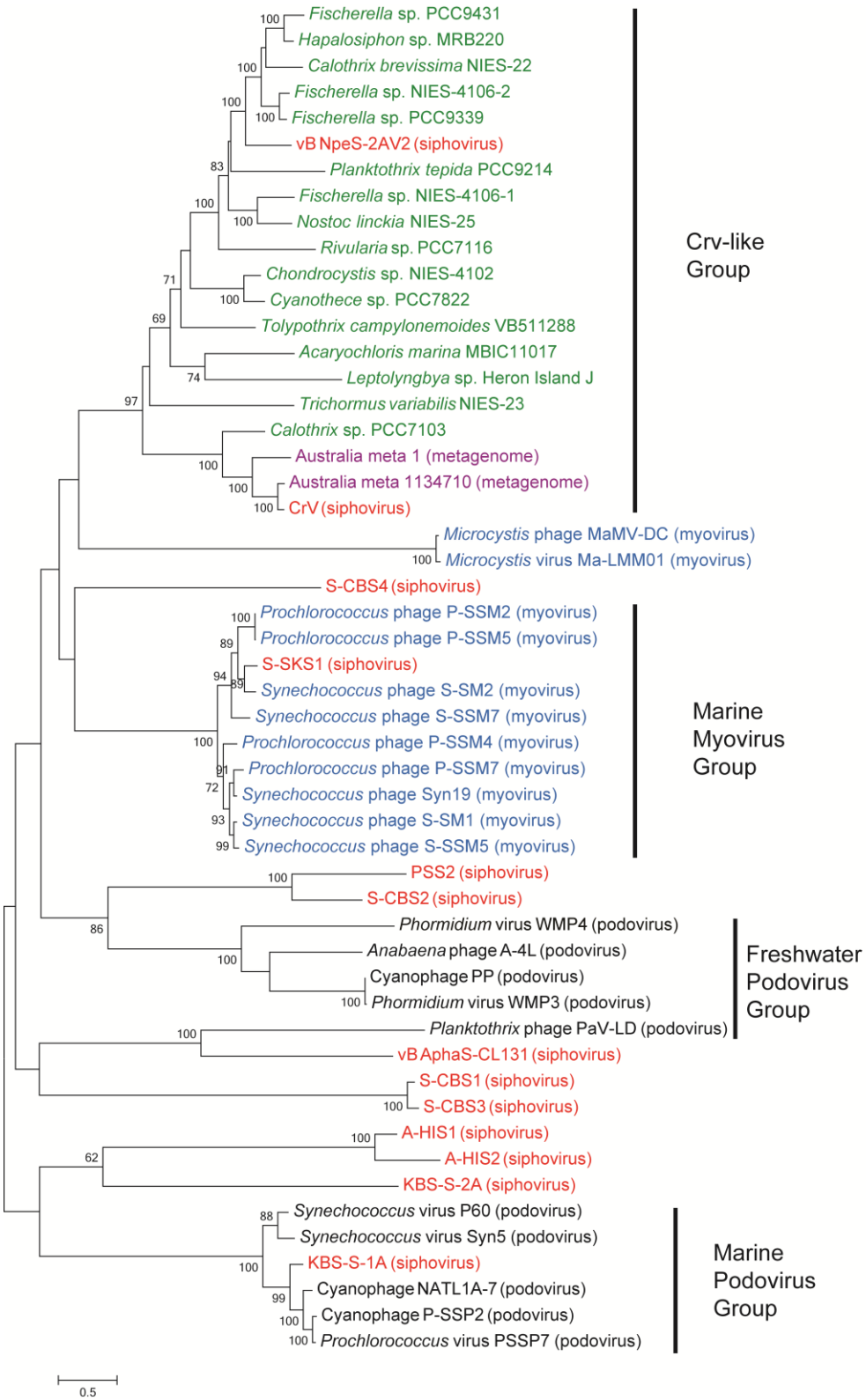

Figure 2.

This article is protected by copyright. All rights reserved

Accepted Article

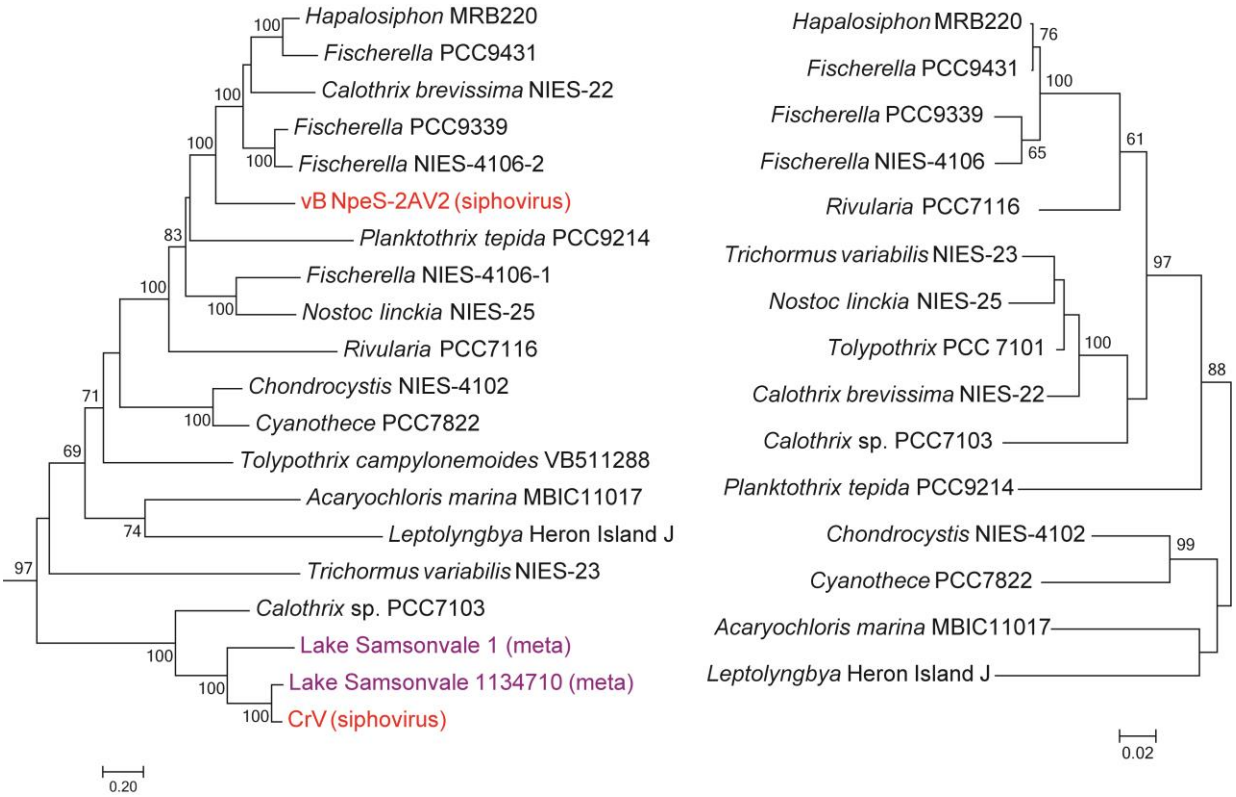

Figure 3.

This article is protected by copyright. All rights reserved

Accepted Article

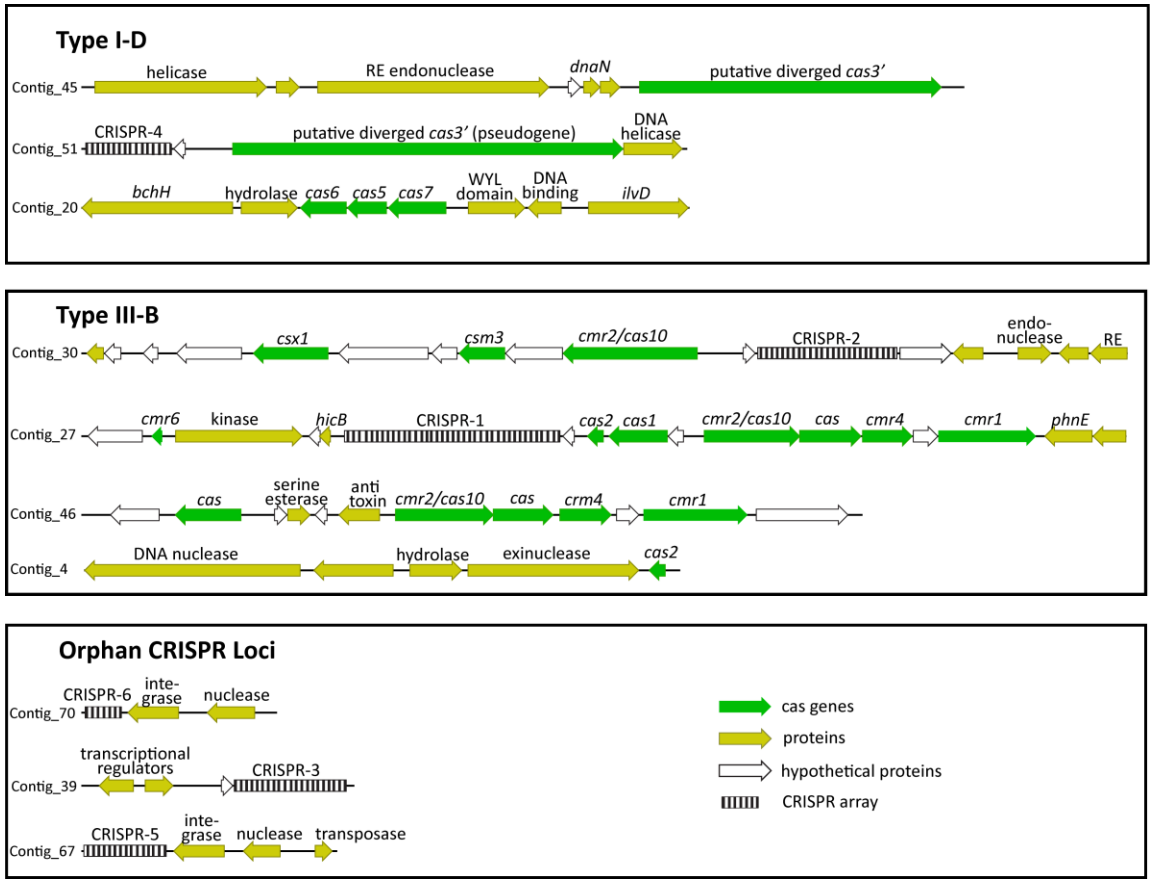

Figure 4.

This article is protected by copyright. All rights reserved

Accepted Article

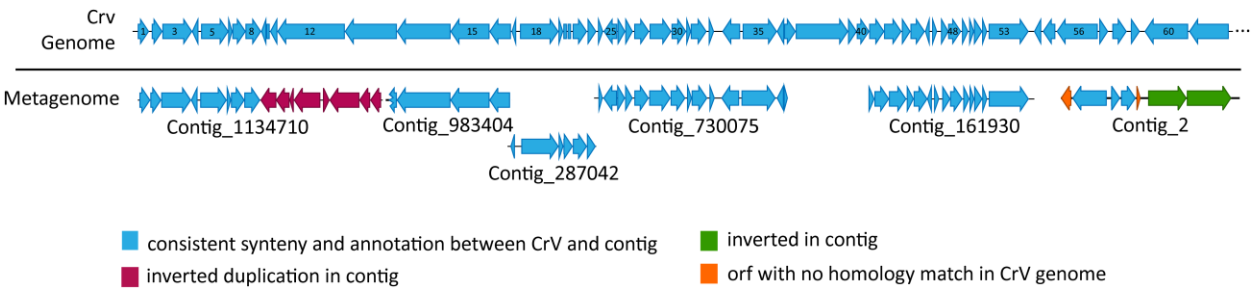

Figure 5.

This article is protected by copyright. All rights reserved

Accepted Article

Table 1. Summary of those predicted genes with functional annotations in *Cylindrospermopsis raciborskii* Virus (CrV)[1].

ORF#	Length (nt)	Strand	Coord.	Coord.	Repeat	Annotation	Functional Class	Likely Origin
CrV_gp002	540	+	974	1,513		ParB; plasmid partition protein	DNA replication	bacterial
CrV_gp003	1,512	+	1,514	3,025		terminase large subunit	DNA replication	viral
CrV_gp007	675	+	5,071	5,745		unnamed phage protein	structural	viral
CrV_gp014	2,718	-	13,430	16,147		virion structural protein	structural	viral
CrV_gp015	1,983	-	16,147	18,129		virion structural protein	structural	viral
CrV_gp017	228	-	19,255	19,482		Cro/C1-type domain; Lambda-like repressor	transcription	viral
CrV_gp025	678	-	23,933	24,610		RNA polymerase sigma-70 factor	transcription	cyanobacterial
CrV_gp035	1,734	+	30,991	32,724		DNA/RNA helicase superfamily II	DNA replication	cyanobacterial
CrV_gp036	366	-	32,727	33,092		Lambda-like phage minor tail protein M	structural	viral
CrV_gp037	588	+	33,095	33,682		Lambda-like tail assembly I	structural	viral
CrV_gp038	2,655	+	33,689	36,343		Lambda-like host specificity tail fiber protein J	structural	viral
CrV_gp041	762	+	37,416	38,177		phage tail protein	structural	viral
CrV_gp042	876	+	38,174	39,049		phage tail assembly protein K	structural	viral
CrV_gp044	723	+	39,514	40,236		Lambda-like phage minor tail protein L	structural	viral
CrV_gp046	249	+	40,612	40,860		Cro/C1-type domain; Lambda-like repressor	transcription	viral
CrV_gp053	2,019	+	43,486	45,504		FtsK; DNA translocase	DNA replication	cyanobacterial
CrV_gp060	2,199	-	51,409	53,607		Ribonucleoside-triphosphate reductase, B12-dependent	DNA replication	cyanobacterial
CrV_gp061	2,007	-	53,668	55,674		transposase, IS605 OrfB family	other	cyanobacterial
CrV_gp064	13,251	+	57,277	70,527		phage tape measure protein	structural	viral
CrV_gp067	594	-	71,972	72,565	repeat	dCTP deaminase	DNA replication	cyanobacterial
CrV_gp068	603	-	72,568	73,170	repeat	putative class 3 lipase	other	viral
CrV_gp084	633	+	80,229	80,861	repeat	essential recombination function protein	DNA replication	cyanobacterial
CrV_gp095	291	-	86,080	86,370		DNA binding protein; Lambda repressor-like domain	transcription	viral
CrV_gp096	834	+	86,549	87,382		N6 Adenine-specific DNA methyltransferase, N12 class	DNA replication	viral
CrV_gp097	459	+	88,556	89,014		putative phage-associated protein	other	viral
CrV_gp100	1,059	+	90,144	91,202		DNA methylase; RM-like system	DNA replication	bacterial
CrV_gp107	633	-	95,064	95,696	repeat	essential recombination function protein	DNA replication	cyanobacterial
CrV_gp123	603	+	102,755	103,357	repeat	putative class 3 lipase	other	viral
CrV_gp124	594	+	103,360	103,953	repeat	dCTP deaminase	DNA replication	cyanobacterial

[1]Predicted genes with no blastx hits to the nr database or those annotated only as hypothetical proteins are excluded from table for space and clarity.

This article is protected by copyright. All rights reserved

View publication stats