

Isolation and characterization of a novel Lambda-like phage infecting the bloom-forming cyanobacteria *Cylindrospermopsis raciborskii*

Emmanuelle Laloum,^{1†} Esther Cattan-Tsaushu,^{1†}
Daniel A. Schwartz^{1b,2†}, Hanaa Shaalan,¹
Hagay Enav,³ Dikla Kolan¹ and Sarit Avrani^{1b,1*}

¹Department of Evolutionary and Environmental Biology and The Institute of Evolution, University of Haifa, Haifa, Israel.

²Department of Biology, Indiana University, Bloomington, IN.

³Department of Microbiome Science, Max Planck Institute for Developmental Biology, Tübingen, Germany.

Summary

Cylindrospermopsis raciborskii is a central bloom-forming cyanobacteria. However, despite its ecological significance, little is known of its interactions with the phages that infect it. Currently, only a single sequenced genome of a *Cylindrospermopsis*-infecting phage is publicly available. Here we describe the isolation and characterization of Cr-LKS3, a second phage infecting *Cylindrospermopsis*. Cr-LKS3 is a siphovirus with a higher genome similarity to prophages within heterotrophic bacteria genomes than to any other cyanophage/cyano-prophage, suggesting that it represents a novel cyanophage group. The function, order and orientation of the 72 genes in the Cr-LKS3 genome are highly similar to those of Escherichia virus Lambda (hereafter Lambda), despite the very low sequence similarity between these phages, showing high evolutionary convergence despite the substantial difference in host characteristics. Similarly to Lambda, the genome of Cr-LKS3 contains various genes that are known to be central to lysogeny, suggesting it can enter a lysogenic cycle. Cr-LKS3 has a unique ability to infect a host with a dramatically different GC content, without carrying any tRNA genes to compensate for this difference. This ability, together with its

potential lysogenic lifestyle shed light on the complex interactions between *C. raciborskii* and its phages.

Introduction

Non-marine cyanobacteria are highly diverse in their shapes (unicellular, colonial, filamentous, or branching), their ability to differentiate into several specialized cell types (heterocyst, akinete, hormogonia), and in the environments they inhabit (freshwater, brackish, desert crust, lithosphere, halophilic, hyperthermophile, etc.) (Sarma, 2013). Moreover, cyanobacterial genomes vary dramatically in size and contain multiple anti-phage defence systems (Makarova *et al.*, 2011). Thus, it is not surprising that the phages infecting these cyanobacteria (cyanophages) are highly diverse, as well.

It is currently possible to obtain much information regarding phage diversity from metagenomics data. However, an in-depth ecological understanding of phage diversity and its role in shaping the host populations still requires identification of the phage hosts (Coclet and Roux, 2021). Thus, to gain insights into phage–host interactions, we remain restricted to isolated phages whose genome and host are known, or to phages for which hosts were identified by molecular [e.g.: (Deng *et al.*, 2014; Ignacio-Espinoza *et al.*, 2020)] or bioinformatic methods [reviewed in (Coclet and Roux, 2021)]. However, despite significant advances in this field (Coclet and Roux, 2021), our ability to identify non-marine cyanophages in metagenomes is still limited. The small number of sequenced genomes (<30; Table S1), together with the high diversity of these sequenced genomes, restrict the ability to identify novel viral sequences as cyanophages. Moreover, the isolation and genomic characterization of new non-marine cyanophages are essential for PCR or probe-based quantification methods such as Colonies (Baran *et al.*, 2018), and for physiological and co-evolution studies. Identification and characterization of cyanophages infecting bloom-forming cyanobacteria strains are particularly important, as it is a necessary step in enhancing our understanding of the role played by phages in bloom dynamics.

Received 26 August, 2021; accepted 14 January, 2022. *For correspondence. E-mail savrani@univ.haifa.ac.il; Tel. +97248240617.

[†]These authors contributed equally to this work.

Cylindrospermopsis raciborskii is a central bloom-forming cyanobacteria. It is considered an invasive species and, in many cases, forms toxic blooms (Sukenik *et al.*, 2012). Two phages that infect *C. raciborskii* were previously isolated. The first, AR-1, was isolated from polluted pond waters in India (Singh and Singh, 1967). It was host specific but was never fully characterized. The second phage infecting *C. raciborskii* (CrV) was isolated from a lake in the Netherlands (Steenhauer *et al.*, 2016). This phage is host-specific as well, and has a long non-contractile tail and double-stranded DNA, suggesting it belongs to the Siphoviridae family. The genome of CrV was recently sequenced (Martin *et al.*, 2019), and it consists of 104 363 bp and 123 open reading frames (ORFs). CrV was suggested to be obligatory lytic, though it may have relatives that lysogenize natural populations of cyanobacteria (Martin *et al.*, 2019). Here, we describe the isolation of a novel phage named Cr-LKS3, infecting *C. raciborskii*, and the study of its genome to advance our understanding of population dynamics of *C. raciborskii* and its phages.

Results and discussion

Cr-LKS3 is a siphovirus with a linear genome

Cylindrospermopsis phage Cr-LKS3 (hereafter Cr-LKS3) was isolated from Lake Kinneret during a *Cylindrospermopsis* bloom, using *C. raciborskii* st. KLL07, which was previously isolated from the same lake (Alster *et al.*, 2010), as a host. Imaging Cr-LKS3 by scanning electron microscopy (SEM) shows that it has an icosahedral capsid and a non-contractile tail (Fig. S1), suggesting that it is a member of the Siphoviridae family.

The genome of Cr-LKS3 is composed of linear double-stranded DNA 46 251 bp in length, which is 32%–50% the size of genomes of other known siphoviruses infecting bloom-forming freshwater cyanobacteria [92 627–144 444 bp (Morimoto *et al.*, 2020)], but longer than genomes of other siphoviruses [e.g. S-LBS1, which is 34 641 bp long (Zhong *et al.*, 2018)]. It has cohesive ends with 5' extensions that form a *cos* site when annealed (see details in the Supplementary Information text), which is similar to Lambda (Vahanian *et al.*, 2017). The *cos* site is a 9 bp long site with the sequence GGGCTCGCG, which is shorter than the 12 bp *cos* site of Lambda, but longer than the shortest known cohesive ends (7 bp) (Vahanian *et al.*, 2017). The sequence of the first 4 bp and of the nucleotides upstream (2) and downstream (2) to the *cos* site is identical to that of Lambda and other Lambda-like phages (Vahanian *et al.*, 2017), while the other 5 bps are

different. This result is consistent with the high conservation known for these loci (Vahanian *et al.*, 2017).

The GC content of Cr-LKS3 differs from its host

The GC content of Cr-LKS3's genome is 66.3%, which is much higher than the GC content of the host used for isolation (40.2%). The level of compatibility of a phage to its host's GC content (which correlates with levels of similarity in codon usage) was suggested to affect the efficiency of phage infection (Limor-waisberg *et al.*, 2011). It was previously shown in marine cyanobacteria that the GC content of host-specific phages is similar to that of their hosts. In contrast, the GC content of broad host range phages is similar to the lowest value among their hosts, and they carry multiple tRNA genes that could compensate for the difference with other hosts (Limor-waisberg *et al.*, 2011; Enav *et al.*, 2012). Therefore, it was surprising that the GC content of Cr-LKS3 was so much higher than that of the strain used for isolation, while the phage carried no tRNA genes. This may suggest a novel mechanism used by Cr-LKS3 to compensate for large differences in nucleotide and codon composition.

We, therefore, wished to examine whether additional phages exist that can infect hosts with dramatically different GC content than their own, but with no encoded tRNA genes. Previous studies were based on podo- and myo-cyanophages (Limor-waisberg *et al.*, 2011; Enav *et al.*, 2012), probably due to the low number of available sequences for siphoviruses when these studies were published. However, many genomes of siphoviruses were subsequently published, enabling examination of the relevance of the GC correlations described for podo- and myo-cyanophages, in siphoviruses and their hosts. The range of GC content of siphoviruses (35.2%–69.4%) was broader than that of podo- (37.9%–55%) or myo- (34.3%–51.7%) cyanophages (Table S1). In siphoviruses, tRNA genes (1–3) were identified in only 10 of the 27 sequenced genomes (Table S2). Moreover, while 7/10 of the siphoviruses with a sequenced host genome varied from their host in GC content by more than 5%, only one of these seven had tRNA genes (2) in its genome (Fig. 1A; Table S2), suggesting that siphoviruses do not usually use tRNA genes to compensate for the difference in GC content from their host's genome. A striking example for this phenomenon may be seen in two phages (Mic1 and vB_MaeS-yong1), both isolated from different strains of *Microcystis aeruginosa* (GC content of 42.1–43.2), which have GC content of 35.1% and 66.7%, while carrying no tRNA genes (Fig. 1B; Table S2). These results

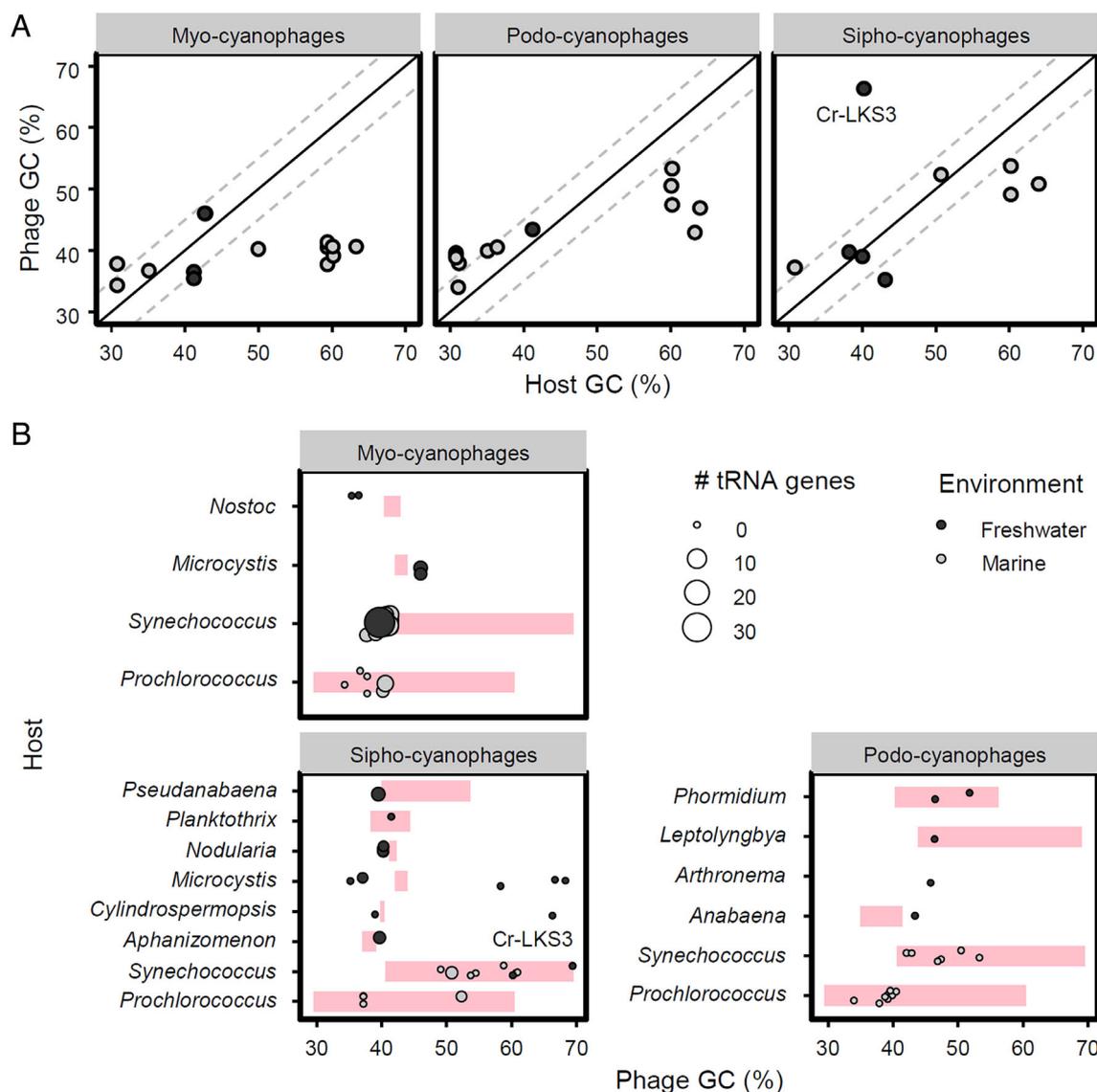


Fig. 1. GC content of siphocyanophages and their hosts.

A. GC content (%) of phage-host couples in all siphocyanophages with sequenced genome for both phage and host, and in a subset of the sequenced podo- and myo-cyanophages. Phages isolated from non-marine (black) and marine (grey) environments. Black line shows equal GC content of the host and phage. Grey dashed lines delineate the region in which the difference in GC content of the phage and host differ by <5%. B. GC content (%) of all sequenced siphocyanophages and a subset of the sequenced podo- and myo-cyanophages ordered based on their host. The size of the circle reflects the number of tRNA genes in the phage genome. Pink areas show the range of GC content (%) of all genomes belonging to the relevant host genus (no data for *Arthonema*). Raw data can be found in Table S2.

suggest that Cr-LKS3, together with other siphocyanophages, resolve the difference in GC content relative to their host, using a mechanism different from the tRNA strategy previously described.

Genes of Cr-LKS3

The genome of Cr-LKS3 contains 72 ORFs, 42 of which have homologues (blastp; E -value ≤ 0.0001) in the non-redundant protein database (Table S3). Twenty of these ORFs have predicted functions (Fig. 2; Table 1). Using

HHpred, predicted functions were assigned to 19 additional ORFs, which reduced the number of hypothetical genes of unknown function to 33 (45.8%).

Cr-LKS3 belongs to a novel cyanophage genomic lineage

The genome of Cr-LKS3 showed no sequence similarity (at either nucleic or amino acid levels) to the genome of CrV, which is the first sequenced phage that infects *C. raciborskii*. We wished to identify other cyanophages

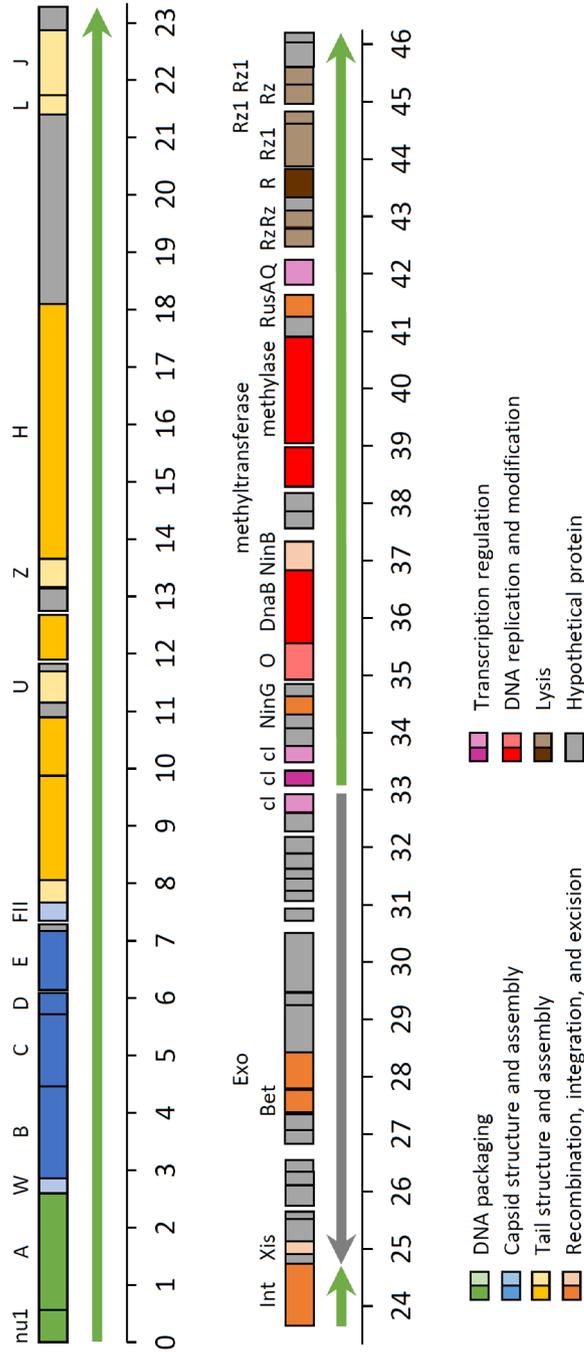


Fig. 2. Genome organization of Cr-LKS3. Each gene in Cr-LKS3's genome is coloured according to its predicted function. Dark colours denote annotations according to blastp against the n/r protein database (E -value<0.0001), and light colours denote annotations based on HHpred (probability >85%; all but one gene had probability of >90%). Green arrows – genes encoded on the plus strand; grey arrow – genes encoded on the minus strand.

Table 1. ORFs with predicted function in the genome of Cr-LKS3.

^a ORF No.	^b Start	^b End	^c No. a.a (strand)	Protein Category	ORF Name	Predicted Function	^d Prediction category
1	95	610	171 (+)	DNA packaging	nu1	Terminase small subunit	1
2	603	2708	701 (+)	DNA packaging	A	Phage terminase large subunit A	1
3	2705	2923	72 (+)	Capsid structure and assembly	W	Possible head completion protein	2
4	2923	4476	517 (+)	Capsid structure and assembly	B	Portal protein	1
5	4473	5747	424 (+)	Capsid structure and assembly	C	Capsid assembly protease C	1
6	5754	6140	128 (+)	Capsid structure and assembly	D	Head decoration protein D	1
7	6153	7175	340 (+)	Capsid structure and assembly	E	Major capsid protein	1
9	7409	7741	110 (+)	Capsid structure and assembly	FII	Possible head-tail joining protein	2
10	7738	8139	133 (+)	Tail structure and assembly		Possible tail fibre protein	4
11	8136	9935	599 (+)	Tail structure and assembly		Possible tail fibre protein	1
12	9932	10,924	330 (+)	Tail structure and assembly		Tail assembly protein	1
14	11 265	11 762	165 (+)	Tail structure and assembly	U	Possible tail tube terminator protein	2
16	11 974	12 801	275 (+)	Tail structure and assembly		Possible major tail protein	1
18	13 221	13 775	184 (+)	Tail structure and assembly	Z	Possible tail completion protein Z	2
19	13 775	18 157	1460 (+)	Tail structure and assembly	H	Phage tail tape measure protein	1
21	21 393	21 719	108 (+)	Tail structure and assembly	L	Possible tail tip protein L	5
22	21 716	22 969	417 (+)	Tail structure and assembly	J	Possible tip attachment protein J	2
24	23 692	24 708	338 (+)	Recombination, integration, and excision	int	Integrase	1
26	25 135	24 938	65 (-)	Recombination, integration, and excision	xis	Possible excisionase	3
34	27 744	27 304	146 (-)	Recombination, integration, and excision	Bet?	Single-stranded DNA-binding protein	1
35	28 397	27 741	218 (-)	Recombination, integration, and excision	exo	Exonuclease	1
46	32 969	32 613	118 (-)	Transcription regulation	cl	Possible repressor protein cl	2
47	33 058	33 330	90 (+)	Transcription regulation	cl	Repressor protein cl	1
48	33 457	33 870	137 (+)	Transcription regulation	cl	Possible repressor protein cl	3
51	34 288	34 644	118 (+)	Recombination, integration, and excision	ninG?	Possible phage RecA-dependent nuclease	1
53	34 890	35 564	224 (+)	DNA replication and modification	O	Possible replication protein O	3
54	35 561	36 862	433 (+)	DNA replication and modification	dnaB	Replicative DNA helicase	1
55	36 862	37 374	170 (+)	Recombination, integration, and excision	ninB	Possible protein ninB	2
58	38 260	38 997	245 (+)	DNA replication and modification		Methyltransferase	1
59	39 044	40 894	616 (+)	DNA replication and modification		DNA methylase	1
61	41 232	41 651	139 (+)	Recombination, integration, and excision	rusA	RusA family crossover junction endodeoxyribonuclease	1
62	41 793	42 212	139 (+)	Transcription regulation	Q	Possible antitermination protein Q	2
63	42 479	42 802	107 (+)	Lysis	Rz	Possible spanin, inner membrane subunit	5
64	42 750	43 076	108 (+)	Lysis	Rz	Possible spanin, inner membrane subunit	5
66	43 336	43 851	171 (+)	Lysis	R	Endolysin	1
67	43 861	44 598	245 (+)	Lysis	Rz1	Possible spanin, outer lipoprotein subunit	4
68	44 595	44 819	74 (+)	Lysis	Rz1	Possible spanin, outer lipoprotein subunit	4
69	44 946	45 281	111 (+)	Lysis	Rz	Possible spanin, inner membrane subunit	4
70	45 278	45 628	116 (+)	Lysis	Rz1	Possible spanin, outer lipoprotein subunit	4

^aORFs with no predicted function were excluded.

^bStart/End – Locus of the first/last nucleotide of the gene respectively.

^cNumber of amino acids in a gene. In parentheses, the directionality of the gene: plus/minus strand.

^dCategory: 1 – BLASTp against the *n/r* protein database (*E*-value<0.0001); 2 – HHpred 99%–100%; 3 – HHpred 98%–99%; 4 – HHpred 90%–98%; 5 – HHpred 80%–90%; HHpred against UniProt-SwissProt_viral70_23_Aug_2020. See details in Table S3.

showing nucleotide-level genomic similarity to Cr-LKS3. To this end, we compared the average nucleotide identity of Cr-LKS3 to a collection of 65 cyanophage genomes (Table S2), using fastANI (Jain *et al.*, 2018), with loose parameters (kmer size = 10, fragment length >200). Our search did not yield nucleotide-level genomic similarity to any cyanophage.

Only eight genes of Cr-LKS3 had homologues in non-marine cyanophages (Table S3), and none of these

cyanophages had more than two homologous genes in their genomes. Moreover, the vast majority of the homologues in non-marine cyanophages showed much lower similarity to Cr-LKS3 than those in heterotrophic bacteria (Fig. S2), and these homologues did not cluster with the genes of Cr-LKS3 (Fig. 3D and E). These results suggest that Cr-LKS3 belongs to a novel group infecting cyanobacteria that evolved separately from other known cyanophages.

To identify additional phage genomes belonging to this group, we searched for homologues of Cr-LKS3 genes in existing metagenomic datasets of microbial populations from six aquatic environments in Israel. We identified (see methods for details) a full circular, 47 137 bp long,

phage genome that highly resembled the genome of Cr-LKS3, in Dalton reservoir. This reservoir contains freshwater mixed with treated wastewater and is located approximately 25 km north-western to the location of Cr-LKS3 isolation. We named this novel phage Cy-LDV1

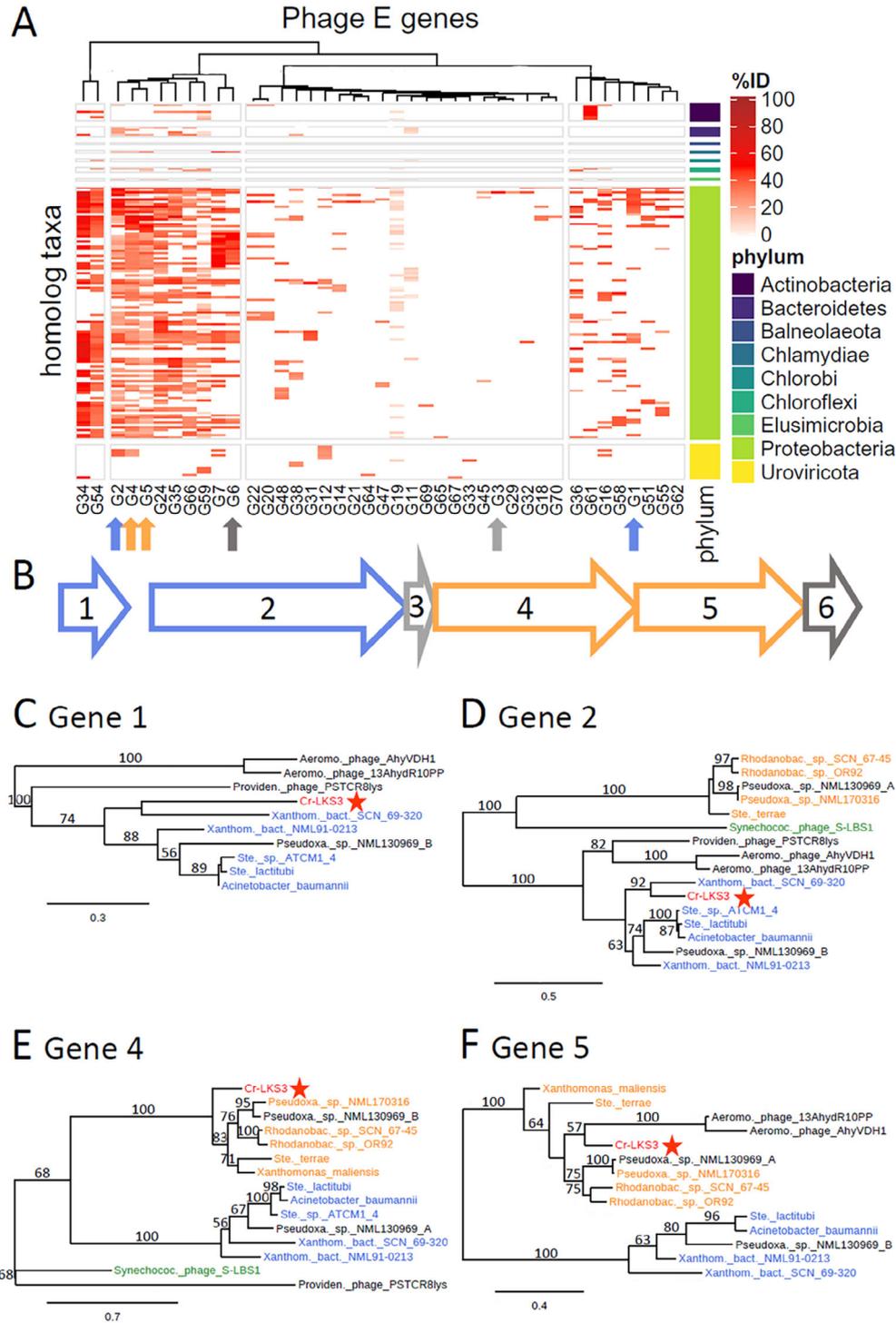


Fig. 3. Legend on next page.

(Fig. S3B). The genome of Cy-LDV1 contains 66 genes, 44 of which are homologues of genes in the Cr-LKS3 genome and they show high synteny with the Cr-LKS3 genome. Nine of the 22 'non-homologous' genes in Cy-LDV1 genome have a predicted function. Interestingly, all these functions are found in the Cr-LKS3 genome and five of them are also located at the parallel locus to that function in Cr-LKS3 (Tables S3 and S4). These results show that the distribution of the Cr-LKS3 lineage is not restricted to Lake Kinneret. Additionally, the differences between Cr-LKS3 and Cy-LDV1 suggest that they are members of a diverse lineage.

Similar functional structure of Cr-LKS3 and Lambda despite low protein-level sequence similarity

The morphology of Cr-LKS3, which resembles the siphoviridae family and the cohesive ends, which resemble the ends of Lambda genome, suggested that Cr-LKS3 may be a lambdoid phage, and thus, we compared their genomes. Although the size of Cr-LKS3's genome (46 251 bp) is similar to that of Lambda (48 502), there is a very low amino acid sequence similarity between their genomes (Fig. 4). Low similarity [$<30\%$ aa identity per gene; (Rost, 1999)] between the genomes was found in Cr-LKS3 ORFs 2, 4, 5, 7 and 35, which correspond to genes A, B, C, E and exo (see Table S5) in Lambda. These genes are involved in DNA packaging (A), procapsid structure and assembly (B, C and E), and homologous recombination (Casjens and Hendrix, 2015). None of the other genes in the genome of Cr-LKS3 (86.3% of the genome) showed any sequence similarity to Lambda.

Although very low sequence similarity was found between the genomes of Cr-LKS3 and Lambda, the order and orientation of the Cr-LKS3 gene clusters highly resemble those of Lambda (Fig. 4). Both genomes start at the cos-site, followed by the small and large subunits of the terminase. The next gene cluster encodes for procapsid related genes, followed by a large cluster that is tail-related. All the above are encoded on the plus strand of the genome. These

structural genes are followed by genes responsible for the ability to lysogenize. These genes, such as those responsible for site-specific and homologous recombination are all found in Lambda on the minus strand, and in Cr-LKS3, all except the integrase (int), are on the minus strand. The final part of the genome, which is responsible, in both phages, for DNA replication and host lysis, starts in both phage genomes with a cro or cro-like gene, followed by central DNA replication genes. In Lambda and other lambdoid phages, DNA replication is initiated by a protein complex formed by the O protein, the host DnaB helicase and the P protein, which is responsible for the recruitment of the host DnaB (Casjens and Hendrix, 2015). In analogy to other lambdoid phages, Cr-LKS3 has its own helicase; thus, this complex in Cr-LKS3 has the same function while built from different phage proteins. In both genomes this cluster is on the plus strand. Both genomes end with genes responsible for host cell lysis. Thus, the genomes of Lambda and Cr-LKS3 are highly similar (where the function of the genes is known) in terms of gene function and orientation, despite the great phylogenetic distance and different hosts.

The high similarity in the functional composition of the genomes of Cr-LKS3 and Lambda strongly suggests that this structure has an unknown advantage. The phages maintained this structure through a substantial evolutionary period, despite marked divergence in their genome sequence (both DNA and amino acid) and in the characteristics of their hosts.

Cr-LKS3 is closest to prophages within heterotrophic bacteria genomes

Forty-two of the ORFs in the Cr-LKS3 genome have homologues in the n/r protein database (Table S3). Surprisingly, the best hits of only four genes were in phage genomes (two in cyanophages and two in phages of heterotrophic bacteria), while the best hits of 38 (90%) of these genes were in genomes of heterotrophic bacteria (Table S3). Similarly, 95% of the hits were found in heterotrophic bacteria when we examined the 10 best

Fig. 3. Cr-LKS3 has a mosaic genome.

A. Heatmap summarizing the amino acid % identity of Cr-LKS3 genes to their homologues. Hierarchical clustering of the genes (G#) was done based on Euclidean distance of percent identity values. Depicted values for the 10 best hits per gene (a single best representative per genus), and the values of hits for all homologues of other genes in the heatmap within those genera. Taxonomic classification at the phylum level is shown on the right side.

B. First six genes in Cr-LKS3 genome coloured according to the gene clustering clade (C–F). Genes with no homologues in the genomes shown in the phylogenetic trees are indicated in grey.

C–F. Maximum likelihood phylogenetic trees of genes 1, 2, 4 and 5. Bootstrap values (out of 100) >50 are shown on the branches. Blue and orange indicate organisms that cluster together in all trees; red marks Cr-LKS3; green marks cyanophage. Accession numbers of the genomes of the organisms shown in the phylogenetic trees (C–F) are listed in Table S6. Aeromo., *Aeromonas*; bact., bacterium; Providen., *Providencia*; Pseudoxa., *Pseudoxanthomonas*; Rhodanobac., *Rhodanobacter*; Ste., *Stenotrophomonas*; Synechococ., *Synechococcus*; Xanthom., *Xanthomonadaceae*.

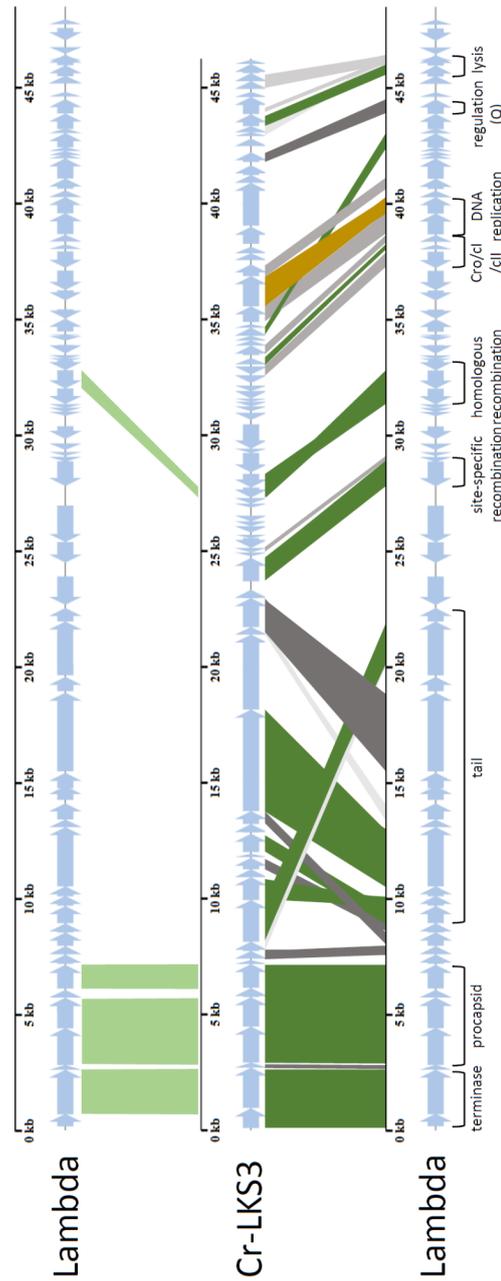


Fig. 4. Cr-LKS3 genomic similarity to Lambda. Amino acid sequence similarity (top panel) versus gene function similarity (bottom panel) between Cr-LKS3 (central genome) and Lambda (top and bottom genome; NCBI accession number: NC_001416.1). Genes in the genome of Cr-LKS3 and Lambda that have a significant amino acid identity (E -value <0.0001) to each other are connected by light green shading in the upper panel. Genes of Cr-LKS3 that are connected to the lower copy of Lambda genome have significant amino acid similarity (E -value <0.0001 ; dark green shading) or predicted structure similarity (HP-pred probability = 99%–100%/98%–99%/90%–98%/80%–90%; grey shading from dark to light respectively) to genes, from various organisms, with the same function as in Lambda. Orange shading connects the *dnaB* helicase gene in Cr-LKS3, which is not found in Lambda, and the Lambda gene *P*, which is responsible for the recruitment of the host DnaB. Main essential gene clusters in Lambda are highlighted at the bottom.

hits (Fig. 3A). At least one of the two best hits in most (30) of these 38 ORFs was located within prophages that are integrated into these genomes (Table S3). Within the other eight genes, six had a distinctive viral function (Portal protein, Tail assembly protein, Repressor protein *cl*, Replication protein *O*, Spanin) also suggesting their origin from a prophage. These results suggest that the closest known relatives of Cr-LKS3 infect heterotrophic bacteria and can enter a lysogenic cycle.

Lysogeny potential

The fact that the closest known relatives of Cr-LKS3 can enter a lysogenic cycle may suggest that Cr-LKS3 has or used to have a similar capacity. The genome of Cr-LKS3 contains various genes that are known to be central to lysogeny, due to their role in recombination, integration, and excision; these encode integrase, excisionase, exonuclease, etc. (see Fig. 2; Table 1). Moreover, these genes are in a similar position as the genes with similar function in the genome of Lambda (Fig. 4), indicating that their location may be important to their function. Additionally, these genes may be regulated in a similar manner as in Lambda, due to the presence of the potential regulatory genes that belong to the *cro/cl* repressor family (gene 46/47) at the same position and orientation as in Lambda (Fig. 4).

One of the genes that is essential for lysogeny is the integrase. The directionality of the integrase gene is highly conserved in prophages recovered from genomes of various Gram-positive and Gram-negative bacteria (Canchaya *et al.*, 2003), as well as in the genome of Cr-LKS3's relative, Cy-LDV1 (Fig. S3). However, in the genome of Cr-LKS3, the gene coding for the integrase protein, while located at the same locus as in Lambda, is encoded on the plus strand instead of the minus strand (Figs 2 and 4). As a result, it is potentially regulated together with genes that take part in the lytic cycle instead of the lysogenic cycle. This suggested that Cr-LKS3 cannot enter a lysogenic cycle. However, the fact that (at least) one phage that carries an 'inverted' integrase gene (Phage Gifsy-1) is lysogenic (Canchaya *et al.*, 2003) suggests that Cr-LKS3 do have the ability to become lysogenic as well. Thus, it is yet to be understood what is the lysogeny potential of Cr-LKS3.

Genome mosaicism in Cr-LKS3

While the function of most of the genes in Cr-LKS3 genome was similar to those of Lambda and they had similar loci and orientation, their nucleotide and amino acid sequence differed dramatically. Moreover, the Cr-LKS3 genome showed low similarity to any known genome, which was manifested in various ways: First, 30 of Cr-LKS3's 72 genes had no homologue in the *n/r* protein database, and these

30 genes were scattered along the Cr-LKS3 genome (Table S3). Second, best hits to viral genomes (blast against non-redundant protein database, taxid: 10239 – viruses) were to siphoviruses (41%), myoviruses (32%), podoviruses (25%), or to unclassified viruses [2%; Table S3]. The hits to the different viral families were scattered all over Cr-LKS3's genome. Moreover, when we looked at the taxonomy of the organisms in which we found the top 10 hits (at the genus level), we found a dispersed pattern with only two pairs of neighbouring genes that clustered together (genes 4 and 5, and genes 6 and 7; Fig. 3A). These latter genes are all related to the capsid structure and assembly, which tend to be more conserved genes in Lambdoid phages. Genes 4 and 5 clustered with the nearby gene 2. A phylogenetic characterization of the genes in this region (Fig. 3) revealed that this region probably experienced multiple recombination events, as the phylogeny of genes 1 and 2 is different from that of genes 4 and 5, and moreover, genes 3 and 6 had no homologues in any of the genomes with high similarity to genes 1 and 2 and genes 4 and 5 (Fig. 3).

These results suggest that while it is possible that Cr-LKS3 and Lambda had a common ancestor, their genomes experienced multiple recombination events with various Caudovirales phages and other unknown organisms during their evolution. This mosaicism is surprising given the high similarity in function location and orientation of the genes of Cr-LKS3 and those of Lambda, which may be a signal for strong and convergent selection. Such an evolutionary pattern has been observed in other lambdoid phages (Casjens and Hendrix, 2015); however, the signal in Cr-LKS3 is uniquely strong, due to the exceptional conservation of gene order and orientation despite the substantial divergence in sequence and host.

Conclusions

Cr-LKS3 is the second identified phage with a sequenced genome that infects the bloom-forming cyanobacteria *Cylindrospermopsis raciborskii*. It belongs to a novel clade of cyanophages, related to Lambda. Unlike marine podo- and myo-cyanophages, each of which appears to have one or two common ancestors that infected cyanobacteria (for T7-like, T4-like, and S-TIM5-like phages), siphocyanophages seem to be acquired multiple times. Cr-LKS3 is an example for such an introduction. Its high GC content, which is considerably different from that of its host, may suggest that it is at an early stage of a long adaptation, and that it might evolve a lower GC content in the future similar to other cyanophages. This also suggests that Cr-LKS3 has an, as yet uncharacterized, mechanism that enables the infection of hosts with highly different GC content, and that a similar mechanism may be present in additional siphocyanophages.

We are still far from understanding the full diversity of non-marine cyanophages, but the characterization of the genome of Cr-LKS3 provides an important step in this direction.

Methods

Cyanobacteria strains used in this study

Cr-LKS3 was isolated using *C. raciborskii* st. KLL07 (Alster *et al.*, 2010), which was kindly given to our lab by Ora Hadas from the Kinneret Limnological Laboratory at the Israel Oceanographic and Limnological Research (IOLR) Institute. This strain is nontoxic and was isolated from Lake Kinneret (Alster *et al.*, 2010). GeneBank accession number of the genome of this strain is CP091284.

Culture conditions

All cultures were grown in BG-11 medium (Stanier *et al.*, 1979) at 24°C, at a light intensity of 5–7.5 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ and under a regime of 14:10 light–dark diel cycle. Relative growth of the cyanobacteria cultures was estimated by chlorophyll autofluorescence, using a BioTek Synergy H1 microplate reader at excitation and emission of 440 and 680 nm respectively. Plaque assays were performed by pour plating of diluted phage-containing lysates with the appropriate cyanobacteria strain and with Invitrogen UltraPure low melting point agarose at a final concentration of 0.28% mixed with BG-11 medium, in 9 cm Petri dishes.

Phage isolation

Cr-LKS3 was isolated from water sampled at a depth of 5 m in Station A (32°82.146'N 35°35.191'E) in the middle of Lake Kinneret (25.9.2017) by the monitoring crew of Lake Kinneret (IOLR). Isolation was performed by four sequential plaque assays using *C. raciborskii* as the host. After the final plaque assay, the phage was used to inoculate the host in liquid medium. After the host population was cleared, the lysate was filtered through a 0.22 μm filter. Host inoculation in liquid medium by the phage was repeated to enlarge the volume of the phage (in a 1:10 phage to host volume ratio). The phage was preserved in 50% SM buffer (100 mM NaCl, 10 mM MgSO_4 50 mM Tris–HCl [pH 7.5], and 0.01% gelatin) for further study. Samples that were preserved at 4°C were revived by inoculating 180 μl of host culture with 20 μl preserved phage culture in a 96 well plate. Samples that were preserved at –80°C were revived by fast thawing (3 min) in a 40°C water bath, followed by plating as above.

Scanning electron microscopy

Morphology of Cr-LKS3 was determined by SEM. Phage lysate was filtered through a 0.22 μm filter and the filtrate was used for imaging and then applied onto a silicon chip. Before phage application, each chip (5 × 5 mm diced silicon wafer, 270c BN16008E) was coated with 0.2 mg ml^{-1} poly-Lysine. After exactly 25 min at room temperature (RT), the chip was washed three times with double distilled water, and air-dried for 5 min at RT. Next, 20 μl of concentrated and filtered lysate was applied on the silicon chip and incubated at RT for 2 h. Prior to fixation, the sample was washed again three times with sample medium (BG-11). For fixation, samples were first incubated in the original sample medium (BG-11) with 2% glutaraldehyde (GA) and 2% paraformaldehyde (PFA) for 10 min at RT, and then transferred to the second buffer containing 0.1 M cacodylate buffer pH = 7.4 with 2% GA and 2% PFA and incubated for an additional hour at RT. For post-fixation, the sample was washed three times with 0.1 M cacodylate buffer pH = 7.4 and incubated in freshly prepared 1% OsO_4 in 0.1 M cacodylate buffer for 10 min at RT, covered with aluminium foil. Then, the chip was dehydrated through a graded ethanol series: 30%, 50%, 70%, 90%, 95% and 100%; 2 × 2 min for each step, with an additional 3 min for the last step. For the final dehydration step before coating with chromium, critical point drying was used. Two-nanometre thick chromium coating was applied using the coating devise Quorum Q150R ES. Chips were examined using a Zeiss Sigma FE-SEM equipped with Gemini column and Oxford Instruments with an INCA Penta FET-X3 detector.

Analysis of cyanophage genomes in NCBI

A list of all full phage genomes in NCBI was created using the pipeline described in Michniewski *et al.* (2019) (Feb 2021). Cyanophage genomes were sorted from this list according to their hosts using Excel, and all duplications were removed manually. Only one entry per phage species was maintained (total of 151 genomes; Table S1). The genomes were then sorted to marine and non-marine by their host genus. The one cyanobacterial genus that has both marine and non-marine strains in this list is *Synechococcus*. Three phages infecting non-marine *Synechococcus* (S-LBS1, S-EIV1 and S-CRM01) (Dreher *et al.*, 2011; Chénard *et al.*, 2015; Zhong *et al.*, 2018) were sorted out of the marine cyanophages based on the reported source of isolation.

Specific host strain was assigned to each phage using the virus-host database (Mihara *et al.*, 2016) or NCBI data. In phages for which this information was not provided, we looked for information in the GeneBank file of the phage genome or in the paper describing the isolation of the

phage. For two phages (Table S2), we were unable to find the specific host strain. GC content of the phage/host genomes was obtained from NCBI data. The genomes were scanned for tRNA genes by tRNAscan-SE 2.0 server, using the bacterial sequence source (Chan and Lowe, 2019).

Whole-genome sequencing

DNA was extracted using the phenol/chloroform method following the protocol in Pickard (2009). Libraries were prepared using Quantabio's sparQ DNA Frag & Library Prep Kit with Illumina's TrueSeq DNA CD indexes. The phage's entire genome was sequenced by an Illumina MiSeq machine (2×150 bp). A quality assessment of the sequences was performed using FastQC (standard parameters) (Andrews, 2010). Quality trimming was performed using Trim Galore 0.4.2 (parameters: phred 33, quality 20), in order to discard the adapter sequences, and the read quality was reassessed. Assembly was performed using SPAdes 3.9.0 genome assembler (phred offset 33, threads 24, only-assembler, K-127, 77, 55,105) (Nurk *et al.*, 2013). Gene prediction was done as in Zhong *et al.* (2018), using GeneMarkS [threshold value = 0.5; (Besemer *et al.*, 2001)] and PROKKA [(Seemann, 2014) with default parameters] to predict ORFs. The amino acid predictions of Cr-LKS3 by these tools were highly similar (92.1%). When the prediction results were different, the longer sequence was kept, as in Zhong *et al.* (2018). The predicted genes were annotated using blastp (Wheeler *et al.*, 2002) against NCBI non-redundant protein database (Metrix Blossom 62, standard, expected threshold = 10). Predicted sequences with e -values $< 10^{-4}$ were considered to be homologues following Zhong *et al.* (2018). For genes that had no significant amino acid similarity to any sequence in the non-redundant protein database, predicted function was assigned using HHpred with default parameters (Zimmermann *et al.*, 2018). The database used for these predictions was UniProt-SwissProt viral70, 23 August 2020. HHpred Probability was indicated (see Table S3). Results with probability lower than 80% are not shown. Additionally, the average nucleotide identity of Cr-LKS3 was compared to a collection of 65 cyanophage genomes (Table S2), using fastANI (Jain *et al.*, 2018), with loose parameters (kmer size = 10, fragment length > 200). Nucleotide sequence of Cr-LKS3 has been deposited in GeneBank under the accession number OM373202.

Identification of homologues in prophages

The genome, which was the source of each of the top two hits of each gene, was identified using NCBI website. Each genome or partial genome was scanned by PHASTER (Arndt *et al.*, 2016) with default parameters to identify prophage regions. We then verified that the

homologous gene was indeed within one of the identified prophages. Genomes, in which no prophage overlapped the homologous gene locus, were scanned in a similar manner by PhageBoost (Sirén *et al.*, 2021). The remaining genes were analyzed manually. The homologous gene was then identified in the protein FASTA file of the genome, and the genes upstream and downstream to the homologous gene were scanned. Genes that were surrounded by at least four structural phage genes (capsid/head/tail related) along with multiple hypothetical genes were classified as located in a probable prophage. These prophages may be non-functional.

Identifying relatives of Cr-LKS3 in environmental datasets

We used metagenomic recruitment to search for relatives of Cr-LKS3 in natural populations sampled from Lake Kinneret (from which Cr-LKS3 was isolated) and additional five freshwater ponds in Israel (PRJNA497963 in the NCBI BioProject database; see Fig. S4). For recruitment we used Cr-LKS3 protein sequences as queries against BLAST databases made of each metagenomic sample (tBLASTn, E -value $< 1e-4$), following Sabehi *et al.* (2012). We found extensive coverage of the Cr-LKS3 genome with high amino acid sequence similarity in two metagenomes (5 and 0.22 μ m fractions) of a single sample (Lake Dalton reservoir, Israel; 33.013770N 35.459780E; November 2015; Fig. S4). We then assembled the metagenomes from the Dalton reservoir, using metaSPADES [v3.14.1; (Nurk *et al.*, 2017)], and then mapped Cr-LKS3 protein sequences to the scaffolds using tBLASTn, as above. We identified a few scaffolds that resembled Cr-LKS3 in gene order and amino acid sequence (Fig. S3A). Analysis code is available at github.com/danschw/phage_e. Using the sequences of scaffolds 124, 371 and 10 281 (Fig. S3) we were able to assemble a full scaffold covering all the genome of Cr-LKS3. This scaffold seemed to be circular, and thus for the ease of further analysis it was cut in the homologous locus of the cos-site in Cr-LKS3, and the regions that were at the ends of the genome were now connected to form a truncated gene homologous to gene 19 of Cr-LKS3 (using 'nnnn' to fill the gap). We then produced a BAM file using Bowtie 2 (Langmead and Salzberg, 2012). The new assembled genome was used as a reference and the SRR8088693 metagenome as the raw data. The BAM file was then visualized using IGV (Robinson *et al.*, 2011). The gap within the homologue of Cr-LKS3 gene 19 was then filled iteratively, using sequence information from reads that were partially covering the gap and their mate reads. The fact that the genome is circular suggests that this DNA belongs to a phage during a lytic infection.

ORFs of Cy-LDV1 were predicted by GeneMarkS (Besemer *et al.*, 2001). Annotation of these ORFs was

performed as described above for Cr-LKS3. Sequence similarity (percent amino acid identity) of phage Cr-LKS3 to phage Cy-LDV1 was calculated by VipTree (Nishimura *et al.*, 2017). Nucleotide sequence of Cy-LDV1 has been deposited in GeneBank.

Heatmap

For each gene in the Cr-LKS3 genome, blastp was used to identify the first 10 000 hits in the n/r protein database. We then selected the 10 genera with the best hits (lowest *E*-value) for each gene (*E*-value ≤ 0.0001), meaning that if there were multiple results from different strains or species of the same genus within these best hits, we chose the one with the lowest *E*-value, and discarded the rest. These 10 genera per gene formed the seed genera of the heatmap, and we then looked for homologues in all of the seed genera, within the best 10 000 hits of each gene. *E*-value and percent identity were assigned for each homologue. The blast identity results were arranged as a matrix of Cr-LKS3 proteins X taxa containing homologues. The matrix was visualized as a heatmap in R 4.03 (R Core Team, 2020) using the ComplexHeatMap package (v2.6.2) (Gu *et al.*, 2016) with complete-linkage hierarchical clustering of phage proteins by Euclidean distance with K-means set to 4.

Phylogeny

A phylogenetic analysis was performed for four Cr-LKS3 genes (gene 1, 2, 4 and 5). Homologous genes used in the phylogeny analysis were selected as follows: (i) for each of the four analyzed Cr-LKS3 genes, hits (blastp against n/r protein database) with lowest *E*-values were selected; (ii) the genome that contained each homologue (accession numbers can be found in Table S6) was scanned for any other homologue of the other three analyzed genes (*E*-value < 0.0001 ; all of these hits were in bacterial genomes). Therefore, the location of each homologous gene was scanned manually to verify that they all are found within the same prophage within the bacterial genome; (iii) where homologous genes were not found for all four genes, homologous regions were screened by tblastn against the nucleotide collection (nr/nt) database with the same threshold; (iv) homologues originating from three phage genomes were identified similarly when blast was performed against all genomes belonging to the taxid 'viruses'; and (v) the only two homologues within non-marine cyanophages were added, as well.

The above sequences were aligned using PROMALS (Pei and Grishin, 2007) with default parameters, followed by manual curation. Trees were constructed from these alignments using the phylogeny.fr web server (Dereeper

et al., 2008). Within the phylogeny.fr web server, PhyML (Guindon *et al.*, 2010) was used to build maximum likelihood trees (100 bootstraps).

Acknowledgements

We thank O. Hadas for the *C. raciborskii* culture provided to us; S. Ninio for providing us the water sample from which Cr-LKS3 was isolated; T. Dagan and R. Hershberg for comments on the manuscript. This work was supported by a GIF YOUNG grant (no. I-2505-413.13/2018) and a startup BSF grant (no. 2017056) to S.A., and by the National Science Foundation DEB (no. 1934554) to D.A.S. Additionally, this research was supported in part by Lilly Endowment, through its support for the Indiana University Pervasive Technology Institute.

References

- Alster, A., Kaplan-Levy, R.N., Sukenik, A., and Zohary, T. (2010) Morphology and phylogeny of a non-toxic invasive *Cylindrospermopsis raciborskii* from a Mediterranean Lake. *Hydrobiologia* **639**: 115–128.
- Andrews, S. (2010) FastQC: a quality control tool for high throughput sequence data. URL <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D.S. (2016) PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* **44**: W16–W21.
- Baran, N., Goldin, S., Maidanik, I., and Lindell, D. (2018) Quantification of diverse virus populations in the environment using the polony method. *Nat Microbiol* **3**: 62–72.
- Besemer, J., Lomsadze, A., and Borodovsk, M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* **29**: 2607–2618.
- Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brüssow, H. (2003) Prophage genomics. *Microbiol Mol Biol Rev* **67**: 473–473.
- Casjens, S.R., and Hendrix, R.W. (2015) Bacteriophage lambda: early pioneer and still relevant. *Virology* **479–480**: 310–330.
- Chan, P.P., and Lowe, T.M. (2019) tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol Biol* **1962**: 1–14.
- Chénard, C., Chan, A.M., Vincent, W.F., and Suttle, C.A. (2015) Polar freshwater cyanophage S-EIV1 represents a new widespread evolutionary lineage of phages. *ISME J* **9**: 1–13.
- Coclet, C., and Roux, S. (2021) Global overview and major challenges of host prediction methods for uncultivated phages. *Curr Opin Virol* **49**: 117–126.
- Deng, L., Ignacio-Espinoza, J.C., Gregory, A.C., Poulos, B. T., Weitz, J.S., Hugenholtz, P., and Sullivan, M.B. (2014) Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. *Nature* **513**: 242–245.

- Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., *et al.* (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* **36**: 465–469.
- Dreher, T.W., Brown, N., Bozarth, C.S., Schwartz, A.D., Riscoe, E., Thrash, C., *et al.* (2011) A freshwater cyanophage whose genome indicates close relationships to photosynthetic marine cyanomyophages. *Environ Microbiol* **13**: 1858–1874.
- Enav, H., Béjà, O., and Mandel-Gutfreund, Y. (2012) Cyanophage tRNAs may have a role in cross-infectivity of oceanic *Prochlorococcus* and *Synechococcus* hosts. *ISME J* **6**: 619–628.
- Gu, Z., Eils, R., and Schlesner, M. (2016) Complex heatmaps reveal patterns and correlations in multi-dimensional genomic data. *Bioinformatics* **32**: 2847–2849.
- Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307–321.
- Ignacio-Espinoza, J.C., Laperriere, S.M., Yeh, Y.-C., Weissman, J., Hou, S., Long, M., and Fuhman, J.A. (2020) Ribosome-linked mRNA-rRNA chimeras reveal active novel virus host associations. *bioRxiv* 2020.10.30.332502. <https://doi.org/10.1101/2020.10.30.332502>
- Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**: 5114.
- Langmead, B., and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.
- Limor-waisberg, K., Carmi, A., Scherz, A., and Pilpel, Y. (2011) Specialization versus adaptation: two strategies employed by cyanophages to enhance their translation efficiencies. *Nucleic Acids Res* **39**: 6016–6028.
- Makarova, K.S., Wolf, Y.I., Snir, S., and Koonin, E.V. (2011) Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J Bacteriol* **193**: 6039–6056.
- Martin, R.M., Moniruzzaman, M., Mucci, N.C., Willis, A., Woodhouse, J.N., Xian, Y., *et al.* (2019) *Cylindrospermopsis raciborskii* virus and host: genomic characterization and ecological relevance. *Environ Microbiol* **21**: 1942–1956.
- Michniewski, S., Redgwell, T., Grigonyte, A., Rihtman, B., Aguilo-Ferretjans, M., Christie-Oleza, J., *et al.* (2019) Riding the wave of genomics to investigate aquatic coliphage diversity and activity. *Environ Microbiol* **21**: 2112–2128.
- Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., *et al.* (2016) Linking virus genomes with host taxonomy. *Viruses* **8**: 10–15.
- Morimoto, D., Šulcius, S., and Yoshida, T. (2020) Viruses of freshwater bloom-forming cyanobacteria: genomic features, infection strategies and coexistence with the host. *Environ Microbiol Rep* **12**: 486–502.
- Nishimura, Y., Yoshida, T., Kuronishi, M., Uehara, H., Ogata, H., and Goto, S. (2017) ViPTree: the viral proteomic tree server. *Bioinformatics* **33**: 2379–2380.
- Nurk, S., Bankevich, A., Antipov, D., Gurevich, A.A., Korobeynikov, A., Lapidus, A., *et al.* (2013) Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J Comput Biol* **20**: 714–737.
- Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017) MetaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**: 824–834.
- Pei, J., and Grishin, N.V. (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics* **23**: 802–808.
- Pickard, D.J.J. (2009) Preparation of bacteriophage lysates and pure DNA. In *Bacteriophages: Methods and Protocols, Volume 2 Molecular and Applied Aspects*, Clokie, M.R.J., and Kropinski, A.M. (eds). Totowa, NJ: Humana Press, pp. 3–9.
- R Core Team. (2020) R: A language and environment for statistical computing.
- Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011) Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26.
- Rost, B. (1999) Twilight zone of protein sequence alignments. *Protein Eng* **12**: 85–94.
- 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.
- Sarma, T.A. (2013) *Handbook of Cyanobacteria*. Boca Raton, FL: CRC Press.
- Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**: 2068–2069.
- Singh, R.N., and Singh, P.K. (1967) Isolation of cyanophages from India. *Nature* **216**: 1020–1021.
- Sirén, K., Millard, A., Petersen, B., Gilbert, M.T.P., Clokie, M. R.J., and Sichert-Pontén, T. (2021) Rapid discovery of novel prophages using biological feature engineering and machine learning. *NAR Genomics Bioinformatics* **3**: 11–13.
- Stanier, R.Y., Deruelles, J., Rippka, R., Herdman, M., and Waterbury, J.B. (1979) Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *Microbiology* **111**: 1–61.
- Steenhauer, L.M., Wierenga, J., Carreira, C., Limpens, R.W. A.L., 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.
- Sukenik, A., Hadas, O., Kaplan, A., and Quesada, A. (2012) Invasion of Nostocales (cyanobacteria) to subtropical and temperate freshwater lakes – physiological, regional, and global driving forces. *Front Microbiol* **3**: 86.
- Vahanian, N., Oh, C.S., Sippy, J., and Feiss, M. (2017) Natural history of a viral cohesive end site: cosN of the λ -like phages. *Virology* **509**: 140–145.
- Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., *et al.* (2002) Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res* **29**: 13–16.
- Zhong, K.X., Suttle, C.A., Baudoux, A.-C., Derelle, E., Colombet, J., Cho, A., *et al.* (2018) A new freshwater cyanosiphovirus harboring integrase. *Front Microbiol* **9**: 1–15.

Zimmermann, L., Stephens, A., Nam, S.Z., Rau, D., Kübler, J., Lozajic, M., *et al.* (2018) A completely Reimplemented MPI bioinformatics toolkit with a new HHpred server at its Core. *J Mol Biol* **430**: 2237–2243.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Table S1. All Sequenced cyanophages in the NCBI database.

Table S2. Cyanophages and cyanobacteria used for host-phage GC content analysis.

Table S3. Gene characterization in Cr-LKS3.

Table S4. Characterization of Cy-LDV1 genes with no homologues in Cr-LKS3.

Appendix S1: Supporting Information.