***Legionella pneumophila* CRISPR-Cas suggests recurrent encounters with *Gokushovirinae*.**

Shayna R. Deecker¹, Malene L. Urbanus¹, Beth Nicholson¹ and Alexander W. Ensminger^{1,2}

¹ Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada.

² Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada.

Running title: Targets of *Legionella* CRISPR-Cas

ABSTRACT

Legionella pneumophila is a ubiquitous freshwater pathogen and the causative agent of Legionnaires' disease. This pathogen and its ability to cause disease is closely tied to its environmental encounters. From phagocytic protists, *L. pneumophila* has “learned” how to avoid predation and exploit conserved eukaryotic processes to establish an intracellular replicative niche. Legionnaires' disease is a product of these evolutionary pressures as *L. pneumophila* uses the same molecular mechanisms to replicate in grazing protists and in macrophages of the human lung. *L. pneumophila* growth within protists also provides a refuge from desiccation, disinfection, and other remediation strategies. One outstanding question has been whether this protection extends to phages. *L. pneumophila* isolates are remarkably devoid of prophages and to date no *Legionella* phages have been identified. Nevertheless, many *L. pneumophila* isolates maintain active CRISPR-Cas defenses. So far, the only known target of these systems has been an episomal element that we previously named *Legionella* Mobile Element-1 (LME-1). In this study, we have identified over 150 CRISPR-Cas systems across 600 isolates, to establish the

clearest picture yet of *L. pneumophila*'s adaptive defenses. By leveraging the sequence of 1,500 unique spacers, we can make two main conclusions: current data argue against CRISPR-Cas targeted integrative elements beyond LME-1 and the heretofore “missing” *L. pneumophila* phages are most likely lytic gokushoviruses.

IMPORTANCE

The causative agent of Legionnaires' disease, an often-fatal pneumonia, is an intracellular bacterium, *Legionella pneumophila*, that normally grows inside amoebae and other freshwater protists. Unfortunately for us, this has two major consequences: the bacterium can take what it has learned in amoebae and use similar strategies to grow inside our lungs; and these amoebae can protect *Legionella* from various forms of chemical and physical disinfection regimes. *Legionella* are ubiquitous in the environment and frequently found in man-made water systems. Understanding the challenges to *Legionella* survival before it reaches the human lung is critical to preventing disease.

We have leveraged our earlier discovery that *L. pneumophila* CRISPR-Cas systems are active and adaptive – meaning that they respond to contemporary threats encountered in the environment. In this way, CRISPR arrays can be considered genomic diaries of past encounters, with spacer sequences used to identify elements that may impinge on the pathogen's survival. One outstanding question in the field is whether *L. pneumophila* is susceptible to phage, given the presumptive protection provided by intracellular replication within its eukaryotic hosts. In this work, we use CRISPR spacer sequences to suggest that the heretofore “missing” *L. pneumophila* phage are most likely lytic gokushoviruses. Such information is critical to the long-

term goal of developing of new strategies for preventing colonization of our water systems by *Legionella* and subsequent human exposure to the pathogen.

INTRODUCTION

Legionella pneumophila is a Gram-negative, intracellular bacterium that is ubiquitous in freshwater environments (1-3), where it replicates within a wide range of protist hosts (4). If a contaminated water source becomes aerosolized and inhaled, *L. pneumophila* can infect human lung macrophages and cause a severe pneumonia known as Legionnaires' disease (5, 6). Replication in the accidental human host uses similar molecular strategies to those used to infect protists (7). As humans are an evolutionary dead end to the pathogen (8), understanding how *L. pneumophila* is able to persist and replicate in environmental reservoirs is critical to limiting its ability to cause human disease.

Protozoan hosts not only serve as a replicative niche to *L. pneumophila* (1, 4), but also provide protection from desiccation (9), temperature changes (10, 11) and disinfectants (10-13). One outstanding question is whether these hosts also protect *L. pneumophila* from predation by foreign genetic elements, such as viral bacteriophages (phages). Notably, phages for the obligate intracellular pathogens *Chlamydia psittaci* and *Chlamydia pecorum* have been described (14-17), raising the possibility that phages may access *L. pneumophila* even within the protection of the host.

The first report of *Legionella* infection by lytic phages isolated from freshwater samples is equal parts promise and obfuscation (18). Preliminary analysis suggested these phages were members of the *Myoviridae* family (18), but inefficient phage enrichment prevented the preservation of stocks for validation and further study. Around the same time, another group

reported visualization of temperate *Legionella* phage, but this occurred in a well-characterized, fully sequenced type strain with no potential to generate phage particles (19). Now a decade later, no subsequent studies have confirmed either laboratory's findings or identified any other type of *L. pneumophila* phage. Lysogeny also seems uncommon in the *Legionella* genus as the only known "prophage-like" elements described to date are *Legionella* mobile element-1 (LME-1), which has been proposed to descend from an ancestral phage (20, 21), and a putative prophage in one *Legionella micdadei* isolate (22).

Many *L. pneumophila* isolates maintain active and adaptive CRISPR-Cas systems (21, 23, 24), providing another avenue by which to explore the species' relationship with phage. Because *L. pneumophila* CRISPR-Cas protects against foreign threats, the CRISPR array in each *L. pneumophila* genome serves as a sequence-based diary of past environmental encounters. When a bacterium encounters a foreign element (e.g. a plasmid or a phage), a short DNA sequence (spacer) can be obtained and incorporated into a CRISPR array (25-28). The transcribed CRISPR RNA forms a complex with associated *cas* genes, and uses complementary base-pairing and endonuclease activity to cleave the foreign element and protect the bacterium (29-32). Given the sequence-based functionality of CRISPR-Cas systems, sequence similarity can be applied to identify targets of these systems for individual bacterial species (21, 33) and in a global context (34-37).

We previously used spacer sequence similarity to identify the first known target of *L. pneumophila* CRISPR-Cas systems, a mobile genetic element known as LME-1 (21). In the same study, we showed that LME-1 restricts host range and that *L. pneumophila* CRISPR-Cas successfully defends against transfer of the element between strains (21). However, we found that only ~3% of *L. pneumophila* spacers match to LME-1 (21), leaving the vast majority of

spacers without defined targets. In this study we have expanded our analyses to include a collection of over 600 *L. pneumophila* isolates and 1,500 unique spacers. Leveraging this expanded dataset, we identify only two recurrent targets of *L. pneumophila* CRISPR-Cas: LME-1 and the *Gokushovirinae* family of phages.

MATERIALS AND METHODS

Legionella genomes used in this study

Legionella pneumophila genomes (draft and completed) were downloaded from the European Nucleotide Archive (at: <https://www.ebi.ac.uk/ena/browser/home>) (38) and from NCBI (at: <https://www.ncbi.nlm.nih.gov/>) (39). A complete list of accession numbers can be found in Table S1.

Some target sequence data were produced by the US Department of Energy Joint Genome Institute (at: <http://www.jgi.doe.gov/>) in collaboration with the user community (Table S7).

Prophage identification

The downloaded *L. pneumophila* genomes were annotated with Prokka (v.1.14.6) (40) and the resulting GenBank formatted files were analyzed using PhiSpy (41) with the default settings. The original nucleotide FASTA files were analyzed using VirSorter (42) with the default settings.

CsrT amino acid sequences from mobile elements (43) identified by PhiSpy and VirSorter were extracted and aligned with MUSCLE (44) and a tree was generated using FastTree (v 2.1.11) with the default settings (45). The tree was rooted to group I, Trb (43), and

visualized using the Interactive Tree of Life (iTOL) web server (46). Reference sequences were annotated as per Abbott and colleagues (43).

The 29 nucleotide LME-1 attachment site (GTCTGATTATCAAAATAATCAGACTTAAT) (21) was queried by BLASTn against *L. pneumophila* genomes using the following parameters: -gapopen 10 -gapextend 2 -reward 1 -penalty -1 -evaluate 1 -word_size 7 (47).

Core genome identification across LME-1 variants

The FASTA sequences for unique LME-1 variants were subjected to an OrthoMCL analysis to look at the overlap in gene content. Briefly, the sequences were annotated using Prodigal (48), then were analyzed using OrthoMCL (v.2.0.9) with the default settings (49). The resulting clusters were queried by BLASTp using the NCBI viral database (taxid:10239) (47) and using HHpred (at: <https://toolkit.tuebingen.mpg.de/tools/hhpred>) (50) against “UniProt-SwissProt-viral70_23_Aug_2020” using default settings to determine if any had sequence similarity to known phage genes.

Core genomes for each LME-1 variant, containing only genes present within all the variants, were aligned using MUSCLE (44) and a tree was generated using FastTree (v 2.1.11) using the default settings and the GTR model (45). The tree was visualized with ggtree (v. 2.2.4) (51), and the orthologous group presence/absence map was plotted with ggplot2 (v. 3.3.3) (52).

CRISPR-Cas identification and spacer cataloguing

CRISPRCasFinder (53) was used to detect putative CRISPR-Cas systems in *L. pneumophila* genomes (Table S1) and CRISPRDetect (54) was used to determine the

transcriptional direction of each CRISPR array. Spacers were extracted from each array and compiled to form spacer libraries. Redundant spacers were removed from the libraries to create a unique spacer library for downstream analyses. Rarefaction curves for the spacer library of each CRISPR-Cas sub-type were generated using Analytic Rarefaction (v. 2.2.1) using the default parameters (available from <https://strata.uga.edu/software/index.html>). The curves were plotted using ggplot2 (v. 3.3.2) (52).

CRISPR-Cas target search

A pipeline to search for targets of *L. pneumophila* CRISPR-Cas was written as a Unix shell script. *L. pneumophila* genomes were first searched for putative type I-C, I-F and II-B CRISPR arrays using the repeat structure as a guide. If a CRISPR array was detected, it was subsequently masked to minimize false positive hits to the input genomes. A BLAST database was then constructed containing the *Legionella* genomes with masked CRISPR arrays to look for integrated genetic elements (see Table 1). The unique spacer library (see above) was queried against the BLAST database (BLASTn parameters: -gapopen 10 -gapextend 2 -reward 1 -penalty -1 -evaluate 0.01 -word_size 7) (47). The BLAST search was done twice, once for each strand. The BLAST output was then processed to score each hit. The scoring metric subtracted the length of the alignment between a spacer and a hit from the length of the spacer, then added the number of mismatches between the two sequences. For example, an alignment length of 30, a spacer length of 32 and 0 mismatches would generate a score of 2. This was done to take both alignment length and number of mismatches into account when ordering the hits for downstream analyses (for example, so a hit with 0 mismatches and a short alignment length would not be ranked higher than a hit with 2 mismatches but an alignment length across the entire spacer).

After scoring, the BLAST output was converted into a BED format. During this step, the putative target sequences were extended to account for alignment length, followed by the addition of a 3 nucleotide extension on either end of the sequence for downstream protospacer adjacent motif (PAM) filtering. SAMtools (55) and BEDTools (56) were used to extract the putative target sequences and their flanking regions from the input genomes. Finally, the pipeline searched for canonical PAMs for the type I-C, I-F and II-B systems in the flanking region of the putative target sequence. Although our pipeline searched for type I-C, I-F and II-B PAMs, we considered this secondary information and were intentionally PAM agnostic when assigning spacers to target groups to avoid missing potential targets that have since escaped through PAM mutations or those targeted by non-canonical PAMs. Analysis of the PAM sequences present in the target sequences showed that many target sequences did contain the canonical PAM (Figure S3).

The pipeline was also used to search for exogenous threats to the bacterium in viral, plasmid and metagenomic sequences and run as described above (data sets listed in Table 1). The output from the pipeline can be found in Table S6 for complete/draft genomes, and Table S7 for metagenomic data. CRISPR arrays containing spacers with ascribed targets were visualized with CRISPRStudio (57). PAM sequence logos for the hits were plotted using ggseqlogo (v 0.1) (58).

Major capsid protein phylogeny of targeted *Microviridae* genomes

A phylogeny of *Microviridae* major capsid proteins was generated using the amino acid sequences from 3014 genomes (Table S8). The amino acid sequences were aligned using MUSCLE (44) and the tree was generated with FastTree (v 2.1.11) using the default settings (45). The resulting tree was rooted to the *Bullavirinae* clade and visualized using the iTOL web

server (46). The sub-family designations for each phage based on the literature were plotted as a colour-coded ring circling the tree (Table S8).

Sequence conservation and spacer mapping for repeat *L. pneumophila* CRISPR-Cas targets

Sequence conservation analysis and spacer mapping were performed on the *Microviridae* genomes that contained a top-hit target for a spacer that met the following stringency criteria: 5 or fewer mismatches with its protospacer, a canonical PAM for its CRISPR subtype, and no mismatches in the seed sequence region (positions 1-5, 7, 8). The VP1 (major capsid protein), VP2 (DNA pilot protein) and VP4 (replication initiation protein) genes were extracted from these genomes using Geneious Prime 2021 (available from: <https://www.geneious.com/prime/>), and the sequences for each gene were aligned using the MUSCLE plug-in (44). The alignment, consensus sequence, sequence identity profiles and spacer location for each gene were plotted using GViz (v. 1.32.0) (59), binning the sequence conservation with a sliding window of 5.

Code availability

The script used to search for targets of *L. pneumophila* CRISPR-Cas systems is available on GitHub at <https://github.com/EnsmingerLab/LegionellaCRISPRTargetSearch>.

RESULTS

Legionella integrated elements identified by prophage prediction

The only target of *L. pneumophila* CRISPR-Cas identified to date is an integrated mobile element, LME-1 (21). It has been previously reported that LME-1 contains a number of predicted

proteins of phage origin (20, 21), so we decided to revisit whether variants of LME-1 or other integrated *L. pneumophila* phage-like elements could be identified using established prophage prediction tools. We subjected a set of over 600 publicly available *L. pneumophila* genomes to analysis by two complementary prophage prediction programs, PhiSpy (41) and VirSorter (42) (Table S1), which use different methods to identify potential phage-like sequences. PhiSpy predictions are based on a variety of metrics such as AT/GC skew, phage sequence signatures, strand directionality and the presence of potential attachment sites (41). VirSorter identifies viral sequences by similarity to phage proteins (Hmsearch and BlastP) and metrics associated with virus-like genome structures, such as the presence of hallmark viral genes, enrichment in uncharacterized genes and depletion in strand switching between consecutive genes (42).

Of 35 predictions that overlapped between VirSorter and PhiSpy, 7 were LME-1 variants - flanked by the previously described LME-1 attachment site and containing nearly identical gene content (Table S2). Each of the remaining 28 predicted regions contained *csrT*, a regulatory gene which is co-inherited with type IV secretion systems from *Legionella* integrative conjugative elements (ICEs) (43) (Table S2). By CsrT phylogeny, these ICEs all fall into the previously described group IV Lgi elements (Figure S1) (43).

To determine if any additional LME-1 variants might have been missed by our analyses, we searched each *L. pneumophila* genome for a signature of LME-1 integration: the presence of two LME-1 attachment (*att*) sites. This analysis identified 8 *att*-flanked insertions of >29 kb (Table S3). These integrants corresponded to all the previously identified LME-1 variants above, along with one additional LME-1 variant missed by PhiSpy. We also identified 24 instances of a short, 469 bp insertion at the LME-1 *att* site that shares no sequence similarity to LME-1, phages, or any other known sequence (Table S3). Unlike LME-1 insertions, the second *att* site

for these instances contain several substitutions and insertions, suggesting against a contemporary integration event.

We next examined this expanded set of LME-1 variants for differences in gene content and sequence similarity to phage genes. Based on core genome phylogeny, the LME-1s segregate into 5 distinct variants with a set of 32 core genes and 17 accessory genes (Figure 1, Table S4). We examined each LME-1 predicted gene product for similarity to phage proteins

Figure 1 | Core orthologous genes in LME-1 enriched for sequence similarity to known phage genes. Orthologous groups for LME-1 genes were determined using OrthoMCL. The core LME-1 genomes were aligned with MUSCLE; a phylogeny was constructed using FastTree and subsequently visualized in ggtree. Orthologous group presence/absence in each LME-1 genome was visualized with ggplot2. Red dots adjacent to each orthologous group indicate sequence similarity to known phage genes via BLASTp and/or HHPred predictions (Table S4).

using HHpred and BlastP. Strikingly, 27/32 of the core LME-1 genes share similarity to phage proteins (and 6/17 accessory genes). The preponderance of phage-like genes in the LME-1 core (including key phage structural proteins) clearly points to a phage origin (20, 21). It also raises the intriguing possibility that these sequences may generate phage-like particles under one or more heretofore untested experimental conditions.

Cataloging the scope of *L. pneumophila*

CRISPR-Cas defenses

Based on our analysis of 18 *L. pneumophila* isolates, we previously identified LME-1 as the only recurrent target of *L. pneumophila* CRISPR-Cas (21).

Leveraging the availability of a drastically expanded genome set, we revisited our analysis of *L. pneumophila* CRISPR-Cas to determine if one or more additional targets might emerge. To begin, we systematically searched for CRISPR-Cas systems in over 600 *L. pneumophila* isolates (Table S1, Figure 2A) using CRISPRCasFinder (53) and CRISPRDetect (54). Combined with previously identified systems (21, 24, 60, 61), we identified a total of 13 isolates with type I-C CRISPR systems, 47 with type I-F systems and 108 with type II-B systems. These systems contain a total of 1589 unique spacers (Table S5), a four-fold increase relative to our earlier search for targets of *L. pneumophila* CRISPR-Cas (21). Nevertheless, rarefaction curves clearly show the continued value of further sequencing, especially for genomes containing type I-C systems (Figure 2B). This is consistent with our previous observations that at least two of these systems (I-C and I-F) are adaptive (21), suggesting that additional sampling of spacer sequences will continue to reveal new encounters between *L. pneumophila* and its invaders.

LME-1 is the only integrative element targeted by *Legionella* CRISPR-Cas systems

With our expanded spacer library, we next sought to identify the matching protospacer sequences targeted by these systems. Our previous analysis, limited in both spacer sequences (queries) and *L. pneumophila* genomes (to search), nevertheless identified two variants of a LME-1 as frequent targets of all 3 types of *L. pneumophila* CRISPR-Cas (21). Backed by our prophage analysis above, we sought to determine if LME-1 remained uniquely targeted relative to other integrated genomic elements. We queried our expanded spacer set against a local BLAST database of over 600 *L. pneumophila* genomic sequences to search for any integrated elements that overlap with our prophage analysis or may have been missed (Table 1, Figure 3A) (38, 39).

Even within the expanded spacer catalog, LME-1 remains the only integrated target of *L. pneumophila* CRISPR-Cas. Thirty-two spacers target the element, representing approximately 2% of the unique spacer library (Table S6). LME-1 targeting spacers are present in 12 distinct CRISPR-Cas arrays, covering all three CRISPR-Cas sub-types (Table 2, Figure 3B). Within individual arrays, we observe multiple instances of adjacent LME-1 targeting spacers, consistent with repeat exposure and primed spacer acquisition (28) (Figure 3B). Of the 32 spacers that target LME-1, the vast majority (28) target all variants.

While two additional spacers mapped back to other *L. pneumophila* genomic targets (Table 2, Figure 3B, Table S6), neither fall within regions predicted by PhiSpy or VirSorter. Close inspection of the genomic regions surrounding these singleton hits does not suggest the presence of a mobile element or prophage based on the presence of phage-like proteins or %GC content. Notably absent from the collection of spacer hits are sequences matching the 5 endogenous plasmids (Paris (61), Lorraine (61), Lens (60, 61), Mississauga-2006 (21) and

OLDA (62)) previously described in *L. pneumophila*. Likewise, we observe no matches between spacers and 22 additional endogenous *L. pneumophila* plasmids deposited into NCBI (Table S1) or even a larger, general database of plasmid sequences (63) (Table 1). Somewhat surprisingly, LME-1 remains the only integrative element (prophage-like or otherwise) targeted by *L. pneumophila* CRISPR-Cas.

***L. pneumophila* CRISPR-Cas targets phages**

The observation that LME-1 is the only prophage-like element targeted by CRISPR-Cas argues against widespread lysogeny in *L. pneumophila* and is consistent with prior observations that the *Legionella* genus is remarkably prophage-poor (22). The question remains, however, as to whether one or more lytic phages might pose an exogenous threat to *L. pneumophila* survival. To look for evidence of past *L. pneumophila*-phage encounters, we used the unique spacer library to query an extensive database of viral and phage sequences with over 28,000 viral genomes from NCBI (Table 1) (39). Within this collection, only one class of targets presented a signature similar to that of LME-1 (both in the number of unique spacers and the diversity of systems involved). Phages from the *Microviridae* family were, like LME-1, frequent targets of *L. pneumophila* CRISPR-Cas: targeted by 28 spacers across 11 distinct arrays (Table 2, Figure 3B, Table S6). Many of these spacers are found near the 5' end of the array, suggesting a recent acquisition event as this is where new spacers are generally acquired (28) (Figure 3B). As with LME-1, several instances of adjacent *Microviridae*-targeting spacers can be found within individual arrays – again suggestive of multiple exposures and consistent with primed spacer acquisition (28) (Figure 3B).

Since many of the targeted *Microviridae* genomes were assembled from metagenomic reads and subsequently deposited into NCBI, we reasoned that there could be additional targets of *L. pneumophila* CRISPR-Cas present in metagenomic data. As such, we next searched for spacer hits across several metagenomic datasets (Table 1) (64-69). These analyses identified 7 additional spacers that targeted contigs of likely *Microviridae* origin, along with one additional spacer that targeted a contig of likely *Caudovirales* origin (Table 2, Figure 3B, Table S7).

In total, 35 *L. pneumophila* spacers target sequences of *Microviridae* origin: 25 spacers target the conserved major capsid protein VP1, 4 target the pilot protein VP2, and 5 target the DNA replication protein VP4. One spacer targets a non-coding region (Table 3). As with the spacers against LME-1, *Microviridae*-targeting spacers come from a genetically and geographically diverse set of *L. pneumophila* isolates. As with LME-1, this suggests that one or more *Microviridae* phage may represent a commonly encountered threat to *L. pneumophila* in the environment. The apparent frequency of these exposures, along with the lack of repeat spacers against any additional class of phage, significantly narrows the hunt for *L. pneumophila* phages.

L. pneumophila* phages most likely belong to *Gokushovirinae

The *Microviridae* family consists of the recognized subfamilies *Bullavirinae*, which includes well-characterized members such as phi-X174, alpha3 and G4 (70), and *Gokushovirinae*, first characterized as phages of *Chlamydia* (14-17), *Spiroplasma* (71-74) and *Bdellovibrio* (75). Several additional subfamilies have been proposed in recent years including: *Pichovirinae* (76), *Alpavirinae* (77), *Sukshmavirinae* (78), Dragonfly-associated Microvirus “Group D” (78, 79), *Parabacteroidetes* group based on *Parabacteroidetes* prophages (80, 81) and Pequeñovirus (80, 82).

To determine if we could further narrow the targets of *L. pneumophila* CRISPR-Cas to one or more of these subfamilies, we first generated a major capsid (VP1) phylogeny with all 3000 available *Microviridae* genomes (Figure 4, Table S8). We also included 14 spacer-targeted

metagenomic contigs (identified above) with overlaps of 99-135 bp on the ends, as this overlap suggests these to be complete, circularizable *Microviridae* genomes (Table S7). We then indicated the top target for each spacer onto its corresponding phage genome on the tree. Strikingly, nearly all the spacer targets belong to the *Gokushovirinae* sub-family (Figure 4). First identified as phages for intracellular pathogens such as *Chlamydia* and *Spiroplasma*, *Gokushovirinae* is a very abundant *Microviridae* sub-family found in diverse

environments including marine, freshwater, soil, fecal and animal tissue environments (78-90), and in hosts ranging from marine bacteria (65) to enterobacteria (91). Within the *Microviridae*, the only additional hit is to genomes clustering with the proposed *Pichovirinae* sub-family; a single spacer notable mainly because it targets a non-coding region which one would not expect to be well conserved (76) (Fig S2).

Extension of spacer sequence

similarity into non-conserved

regions of *Gokushovirinae*

We next asked whether

one or more of the

Gokushovirinae-targeting spacers

mapped back to regions of low

conservation between phages. In

particular, spacer sequences that

mapped back to predicted host-

determinant regions of the capsid

(74, 76) might be highly

informative, suggestive of a close

match. In contrast, highly

conserved sequences are present in

many related phage genomes,

reducing the discriminatory power

of an individual spacer to identify

the specific phage whose

encounter led to its acquisition.

Not surprisingly, the vast majority of *Gokushovirinae*-targeting spacers map back to regions of

high conservation, likely reflective of the overrepresentation of conserved sequences within any

database (Figure 5A). Nevertheless, two spacers flank the predicted host-determinant variable

region and extend into it (Figure 5B) and one spacer maps to internal variable sequence (Figure 5C). This is potentially quite informative, given the contribution of the capsid variable region to *Gokushovirinae* host range. In short, these spacers likely provide the first glimpse into specific phage sequences required to infect *L. pneumophila*.

DISCUSSION

We previously demonstrated that a 30+ kb episomal element, LME-1, is a frequent target of an evolutionarily diverse set of *L. pneumophila* CRISPR-Cas arrays. We followed this observation by demonstrating that these systems actively defend against the transfer of this element between strains (21). We report here that LME-1 represents the only integrated element targeted by *L. pneumophila*'s CRISPR-Cas defenses. Unexpectedly, we also show that *L. pneumophila* CRISPR-Cas defenses are also directed against sequences with strong similarity to the *Gokushovirinae* family of phages. While the specific phages that target *L. pneumophila* remain uncharacterized, *Gokushovirinae* meet the criteria expected for a heretofore undiscovered *L. pneumophila* phage: they are largely lytic, consistent with the dearth of prophage sequences across *Legionellae*; and they are known to infect other intracellular bacteria, such as *Chlamydia* (14-17).

Identification of *L. pneumophila* phages would be of practical importance for at least two reasons. First, once established within water systems, *L. pneumophila* can be notoriously difficult to eliminate. Several studies have observed widespread, decades-long colonization of hospital water systems and stable bacterial genotypes, despite remediation efforts (92-96). A targeted biological remediation strategy (such as phage) would likely pose no threat to human health and could support the frequent treatment of high-risk water systems (to either eliminate or prevent

colonization by *Legionellae*) (97). Secondly, the suitability of phage therapy for intracellular pathogens remains an open question. Compared to *Chlamydia*, *L. pneumophila* is much more genetically tractable, can be cultured outside of host cells in the laboratory, and grows inside a staggering diversity of eukaryotic hosts (4). As such *L. pneumophila* would represent a powerful model system by which to understand the feasibility and constraints of using phages to target intracellular pathogens.

By leveraging the active, adaptive nature of *L. pneumophila* CRISPR-Cas, we have developed an updated, sequence-based composite sketch of *L. pneumophila* phage. Unlike LME-1, which we were able to isolate from defenseless “lysogens,” the *Gokushovirinae* sequences identified in this study are exclusively metagenomic contigs from mixed environmental samples. Our findings suggest that isolation of phages from *Legionella*-containing water sources should be performed under conditions likely to capture gokushoviruses (64, 65, 98), using a diverse collection of CRISPR-mutant strains. The size of gokushoviruses (4-5 kb) also raises the intriguing possibility that synthetic biology could be used to develop one or more *Legionella*-infecting virions (99), starting with the sequences we have identified to date.

During infection, *L. pneumophila* and other intracellular pathogens trade diminished access to nutrients for the protection and isolation provided by the vacuolar environment. Once inside the vacuole, *L. pneumophila* recruits nutrients, avoids lysosomal fusion, and is protected from environmental stresses and host antimicrobial activities (9, 100). Do gokushoviruses access *L. pneumophila* (and *Chlamydia*) inside these “cozy niches?” Is the pathogen’s exposure to phages limited to a transient extracellular period (during cell-to-cell spread) or indicative of a more complicated environmental lifestyle? One thing is clear, 100 years after their discovery, phages continue to hold important secrets about the bacteria upon which they prey.

ACKNOWLEDGEMENTS

The authors thank Eric Bastien and Melissa Duhaime for early access to their metagenomic data to search for additional *L. pneumophila* CRISPR-Cas targets. In particular, we want to acknowledge the entire community of researchers who have shared data through JGI, especially Karthik Anantharaman, Vincent Deneff, Katherine McMahon, David Walsh, and Erica Young for access to their unpublished data. We thank Félix Croteau and Alexander Hynes for their suggestions regarding *Legionella* prophage analysis, Jordan Lin for his help with OrthoMCL analyses and David Faguy for discussions regarding coding, along with members of the Ensminger laboratory for their suggestions and careful reading of the manuscript. SRD is supported by a fellowship from the Department of Biochemistry, University of Toronto and an Ontario Graduate Scholarship. This work was supported by a Project Grant from the Canadian Institutes of Health Research (PHT-148819).

TABLES

Table 1 | Target databases. The databases used for spacer target searches for integrated elements and exogenous elements.

Database Name	Target Type	Date Downloaded	Number of Genomes, Reads or Contigs	Genome Completion	Reference
European Nucleotide Archive - <i>Legionella pneumophila</i>	Integrated elements	2019-01-09	519	Draft	38
NCBI - <i>Legionella pneumophila</i>	Integrated elements	2019-04-08	83	Complete	39
Sequence Read Archive - <i>Legionella pneumophila</i>	Integrated elements	2019-04-08	9	Raw reads	21
DataONE Dash Plasmids	Exogenous elements	2019-04-05	10,892	Complete	63
NCBI - <i>Microviridae</i>	Exogenous elements	2019-12-02	2,864	Complete & draft	39
NCBI - Viruses	Exogenous elements	2019-04-08	28,652	Complete	39
IMG/VR	Exogenous elements	2020-01-29	734,969	Metagenomic	66, 67
<i>Microviridae</i> associated contigs from IMG/M	Exogenous elements	2020-02-04	14,724	Metagenomic	68
<i>Microviridae</i> associated contigs from marine sources	Exogenous elements	2020-07-30	12,775	Metagenomic	64, 65
Oxford Nanopore read from an aquifer	Exogenous elements	2020-06-25	2,380,279	Metagenomic	69

Table 2 | CRISPR spacer distribution. A breakdown of the number of unique spacers for each CRISPR-Cas subtype, the number of spacers with identified targets and by target class.

CRISPR-Cas Subtype	Type I-C	Type I-F	Type II-B
Number of Unique Spacers	270	774	545
Number of Spacers with Target	13	52	6
Percent of Library with Target	4.81	6.72	1.10
Number of Spacers Targeting <i>L. pneumophila</i> genome: LME-1	6	21	5
Number of Spacers Targeting <i>Microviridae</i>	6	29	0
Number of Spacers Targeting <i>L. pneumophila</i> genome: other	0	1	1
Number of Spacers Targeting <i>Myoviridae</i>	0	1	0
Number of Spacers Targeting <i>Caudovirales</i>	1	0	0

Table 3 | List of *Microviridae* target regions. The name, sequence and predicted target region for the unique spacers targeting *Microviridae*.

Spacer ID	Spacer Sequence (5' to 3')	Predicted Target Region
Alcoy-spacer32-IF	TGCTTTGAGTCATATTGGAGAGCAAGCTGTTC	VP1 (Major capsid protein)
Calgary2017-spacer6-IF	TTTGAACCGTATGTTTTCTCGTCAGACACGTT	VP1 (Major capsid protein)
FJAB01-spacer12-IF	TAGTATTGTAACGGTCAGTTTAGGTCTAATT	VP1 (Major capsid protein)
FJAB01-spacer53-IF	AAGATGAGACAACGCAGGCCAATAAAAAATCAA	VP1 (Major capsid protein)
FJDB01-spacer29-IF	TTATCTGCCGCGCTTCGCTTGCTGCTACTTT	Putative non-coding region
FJDB01-spacer38-IF	TAAAGTAGCTGCCAGCATAGGATTAATACCCG	VP2 (DNA pilot protein)
FJDB01-spacer39-IF	TGATGATTCTACGGATTATTTAGTTCCACGA	VP1 (Major capsid protein)
FJDB01-spacer69-IF	AATTAATGTTAATCCTGTACCTCAGACTAGTT	VP1 (Major capsid protein)
FJDB01-spacer70-IF	AATTATCCCAACAAGACGATACGGAACAGCA	VP1 (Major capsid protein)
FJEJ01-spacer1-IF	TAGATAATTGTTTTGCACCTGTTGCAAGTCTA	VP1 (major capsid protein)
JFII01-spacer14-IC	CTGATGTTTATGCTGATTTGTCAAAATGCTACTGCTG	VP1 (Major capsid protein)
JFII01-spacer19-IC	ACACCATACTCTCCACAATGATAATAACGAATAC	VP4 (Replication initiation protein)
JFII01-spacer30-IC	ATAATATAACGAGCAACATAAGCAGCAGAATTAAA	VP4 (Replication initiation protein)
JFII01-spacer31-IC	AAATTAAAAGTATCACCAGGTAAAGCTTCATCAA	VP1 (Major capsid protein)
JFII01-spacer32-IC	TCCTGAATAAACGATGAATTAAGAACAGGCAAAG	VP1 (Major capsid protein)
JFII01-spacer5-IC	TGAATTTGAAACGCCTCTCGCAACTGATTAATAG	VP1 (Major capsid protein)
JFIM01-spacer10-IF	TACCATTAAATAACAACAAAATCACCATTATAG	VP4 (Replication initiation protein)
JFIM01-spacer13-IF	TGATAATCCTGCTGATTATGTCCTTTTGAGGC	VP1 (Major capsid protein)
JFIM01-spacer15-IF	ATCATACGTTGATATGATGTATTAGACATACG	VP2 (DNA pilot protein)
JFIM01-spacer16-IF	ACAAGCAAATTTGGCTGCTTTTGGTACTGCTA	VP1 (Major capsid protein)
JFIM01-spacer18-IF	CAAAACAAGTCGATAAGGAACAGCAAAAAGAA	VP1 (Major capsid protein)
JFIM01-spacer19-IF	TCAAGATAAACAGAATCACCTTCTGGTAAGA	VP1 (major capsid protein)
LBAK01-spacer49-IF	AGCAGCCTTCATATCAGCCACCATACGCTGAT	VP2 (DNA pilot protein)
LBAK01-spacer52-IF	TCTTGGCACTACTGCTCCAGTTATTCGTGATG	VP1 (Major capsid protein)
LBAK01-spacer53-IF	AACGAGTTTGACGACTAAACATCCTATTCAAA	VP1 (Major capsid protein)
LBAN01-spacer4-IC	ATTCTCCTTCTTTTGCTTTTTCTACTAATAGCGA	VP1 (major capsid protein)
Lens-Plasmid-spacer28-IF	ACTGTGGTGAGTATGGTGAGAAGTACGGAAGA	VP4 (Replication initiation protein)
Lens-Plasmid-spacer5-IF	AAAATAATCCGACAAACTACCATTTCGCATAAC	VP1 (Major capsid protein)
Lens-Plasmid-spacer6-IF	TTGGCCTCAGAAAGGCCCTGCTGTTGAGTTAC	VP1 (Major capsid protein)
Miss.2006-spacer1-IF	AAATCAAATTGTACGCACGATGAAACAAACTA	VP1 (Major capsid protein)
Miss.2006-spacer2-IF	ACGACGAAGAAGAACATAATCAGCAGGATTGT	VP1 (Major capsid protein)
Miss.2006-spacer34-IF	TCAAAAAGGTCTGCTGTTGAGTTGCCTTTGG	VP1 (Major capsid protein)
Miss.2006-spacer43-IF	TTGTGGTCAGTGATTGGTTGTAAGTTACGCA	VP4 (Replication initiation protein)
Miss.2006-spacer44-IF	TCCTATTGATCGTATTAAAGCTCTTGGTGATG	VP1 (major capsid protein)
Miss.2006-spacer69-IF	TTCTCAAGTTGGTGCTCCTATGCAAACTGGTG	VP2 (DNA pilot protein)

REFERENCES

1. **Rowbotham TJ.** 1980. Preliminary report on the pathogenicity of *Legionella pneumophila* for freshwater and soil amoebae. *Journal of Clinical Pathology* **33**:1179–1183.
2. **Fliermans C.** 1996. Ecology of *Legionella*: From Data to Knowledge with a Little Wisdom. *Microb Ecol* **32**:203–228.
3. **van Heijnsbergen E, Schalk JAC, Euser SM, Brandsema PS, Boer den JW, de Roda Husman AM.** 2015. Confirmed and Potential Sources of *Legionella* Reviewed. *Environ Sci Technol* **49**:4797–4815.
4. **Boamah DK, Zhou G, Ensminger AW, O'Connor TJ.** 2017. From Many Hosts, One Accidental Pathogen: The Diverse Protozoan Hosts of *Legionella*. *Front Cell Infect Microbiol* **7**:21660–17.
5. **McDade JE, Shepard CC, Fraser DW, Tsai TR, Redus MA, Dowdle WR.** 1977. Legionnaires' disease: isolation of a bacterium and demonstration of its role in other respiratory disease. *N Engl J Med* **297**:1197–1203.
6. **Muder RR, Yu VL, Woo AH.** 1986. Mode of transmission of *Legionella pneumophila*. A critical review. *Arch Intern Med* **146**:1607–1612.
7. **Gao LY, Harb OS, Kwaik YA.** 1997. Utilization of similar mechanisms by *Legionella pneumophila* to parasitize two evolutionarily distant host cells, mammalian macrophages and protozoa. *Infect Immun* **65**:4738–4746.
8. **Ensminger AW.** 2016. *Legionella pneumophila*, armed to the hilt: justifying the largest arsenal of effectors in the bacterial world. *Current Opinion in Microbiology* **29**:74–80.
9. **Steinert M, Hentschel U, Hacker J.** 2002. *Legionella pneumophila*: an aquatic microbe goes astray. *FEMS Microbiology Reviews* **26**:149–162.
10. **Storey MV, Winiecka-krusnell J, Ashbolt NJ, Stenström T-A.** 2009. The Efficacy of Heat and Chlorine Treatment against Thermotolerant *Acanthamoebae* and *Legionellae*. *Scandinavian Journal of Infectious Diseases* **36**:656–662.
11. **Cervero-Aragó S, Rodríguez-Martínez S, Puertas-Bennasar A, Araujo RM.** 2015. Effect of Common Drinking Water Disinfectants, Chlorine and Heat, on Free *Legionella* and Amoebae-Associated *Legionella*. *PLoS ONE* **10**:e0134726–18.
12. **Kilvington S, Price J.** 1990. Survival of *Legionella pneumophila* within cysts of *Acanthamoeba polyphaga* following chlorine exposure. *J Appl Bacteriol* **68**:519–525.
13. **Barker J, Brown MR, Collier PJ, Farrell I, Gilbert P.** 1992. Relationship between *Legionella pneumophila* and *Acanthamoeba polyphaga*: physiological status and

- susceptibility to chemical inactivation. *Applied and Environmental Microbiology* **58**:2420–2425.
14. **Liu BL, Everson JS, Fane B, Giannikopoulou P, Vretou E, Lambden PR, Clarke IN.** 2000. Molecular characterization of a bacteriophage (Chp2) from *Chlamydia psittaci*. *Journal of Virology* **74**:3464–3469.
15. **Hsia R, Ohayon H, Gounon P, Dautry-Varsat A, Bavoil PM.** 2000. Phage infection of the obligate intracellular bacterium, *Chlamydia psittaci* strain guinea pig inclusion conjunctivitis. *Microbes Infect* **2**:761–772.
16. **Hsia RC, Ting LM, Bavoil PM.** 2000. Microvirus of *Chlamydia psittaci* strain guinea pig inclusion conjunctivitis: isolation and molecular characterization. *Microbiology* **146** (Pt 7):1651–1660.
17. **Garner SA, Everson JS, Lambden PR, Fane BA, Clarke IN.** 2004. Isolation, molecular characterisation and genome sequence of a bacteriophage (Chp3) from *Chlamydomonas pecorum*. *Virus Genes* **28**:207–214.
18. **Lammertyn E, Vande Voorde J, Meyen E, Maes L, Mast J, Anné J.** 2008. Evidence for the presence of *Legionella* bacteriophages in environmental water samples. *Microb Ecol* **56**:191–197.
19. **Grigor'ev AA, Bondarev VP, Borisevich IV, Darmov IV, Mironin AV, Zolotarev AG, Pogorel'skiĭ IP, Ianov DS.** 2008. [Temperate *Legionella* bacteriophage: discovery and characteristics]. *Zh Mikrobiol Epidemiol Immunobiol* 86–88.
20. **Luneberg E, Mayer B, Daryab N, Kooistra O, Zahringer U, Rohde M, Swanson J, Frosch M.** 2001. Chromosomal insertion and excision of a 30 kb unstable genetic element is responsible for phase variation of lipopolysaccharide and other virulence determinants in *Legionella pneumophila*. *Molecular Microbiology* **39**:1259–1271.
21. **Rao C, Guyard C, Pelaz C, Wasserscheid J, Bondy-Denomy J, Dewar K, Ensminger AW.** 2016. Active and adaptive *Legionella* CRISPR-Cas reveals a recurrent challenge to the pathogen. *Cellular Microbiology* **18**:1319–1338.
22. **Gomez-Valero L, Rusniok C, Rolando M, Neou M, Dervins-Ravault D, Demirtas J, Rouy Z, Moore RJ, Chen H, Petty NK, Jarraud S, Etienne J, Steinert M, Heuner K, Gribaldo S, Digue CM, Glöckner G, Hartland EL, Buchrieser C.** 2014. Comparative analyses of *Legionella* species identifies genetic features of strains causing Legionnaires disease. *Genome Biology* **15**:505.
23. **Rao C, Chin D, Ensminger AW.** 2017. Priming in a permissive type I-C CRISPR–Cas system reveals distinct dynamics of spacer acquisition and loss. *RNA* **23**:1525–1538.
24. **Deecker SR, Ensminger AW.** 2020. Type I-F CRISPR-Cas Distribution and Array Dynamics in *Legionella pneumophila*. *G3* **10**:1039–1050.

25. **Mojica FJM, Díez-Villaseñor C, Soria E, Juez G.** 2000. Biological significance of a family of regularly spaced repeats in the genomes of archaea, bacteria and mitochondria. *Molecular Microbiology* **36**:244–246.
26. **Jansen R, van Embden JD, Gaastra W, Schouls LM.** 2002. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology* **43**:1565–1575.
27. **Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P.** 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**:1709–1712.
28. **McGinn J, Marraffini LA.** 2019. Molecular mechanisms of CRISPR-Cas spacer acquisition. *Nature Reviews Microbiology* **17**:7–12.
29. **Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuys RJH, Snijders APL, Dickman MJ, Makarova KS, Koonin EV, van der Oost J.** 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**:960–964.
30. **Jore MM, Lundgren M, van Duijn E, Bultema JB, Westra ER, Waghmare SP, Wiedenheft B, Pul Ü, Wurm R, Wagner R, Beijer MR, Barendregt A, Zhou K, Snijders APL, Dickman MJ, Doudna JA, Boekema EJ, Heck AJR, van der Oost J, Brouns SJJ.** 2011. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat Struct Mol Biol* **18**:529–536.
31. **Wiedenheft B, van Duijn E, Bultema JB, Waghmare SP, Zhou K, Barendregt A, Westphal W, Heck AJR, Boekema EJ, Dickman MJ, Doudna JA.** 2011. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc Natl Acad Sci USA* **108**:10092–10097.
32. **Westra ER, van Erp PBG, Künne T, Wong SP, Staals RHJ, Seegers CLC, Bollen S, Jore MM, Semenova E, Severinov KV, de Vos WM, Dame RT, de Vries R, Brouns SJJ, van der Oost J.** 2012. CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. *Molecular Cell* **46**:595–605.
33. **Soto-Perez P, Bisanz JE, Berry JD, Lam KN, Bondy-Denomy J, Turnbaugh PJ.** 2019. CRISPR-Cas System of a Prevalent Human Gut Bacterium Reveals Hyper-targeting against Phages in a Human Virome Catalog. *Cell Host and Microbe* **26**:325–335.e5.
34. **Stern A, Mick E, Tirosh I, Sagy O, Sorek R.** 2012. CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Research* **22**:1985–1994.
35. **Zhang Q, Rho M, Tang H, Doak TG, Ye Y.** 2013. CRISPR-Cas systems target a diverse collection of invasive mobile genetic elements in human microbiomes. *Genome Biology* **14**:R40–15.

36. **Shmakov SA, Sitnik V, Makarova KS, Wolf YI, Severinov KV, Koonin EV.** 2017. The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes. *mBio* **8**:e01397–17–18.
37. **Shmakov SA, Wolf YI, Savitskaya EE, Severinov KV, Koonin EV.** 2020. Mapping CRISPR spaceromes reveals vast host-specific viromes of prokaryotes. *Commun Biol* **3**:321–9.
38. **Amid C, Alako BTF, Balavenkataraman Kadhivelu V, Burdett T, Burgin J, Fan J, Harrison PW, Holt S, Hussein A, Ivanov E, Jayathilaka S, Kay S, Keane T, Leinonen R, Liu X, Martinez-Villacorta J, Milano A, Pakseresht A, Rahman N, Rajan J, Reddy K, Richards E, Smirnov D, Sokolov A, Vijayaraja S, Cochrane G.** 2020. The European Nucleotide Archive in 2019. *Nucleic Acids Research* **48**:D70–D76.
39. **Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I.** 2019. GenBank. *Nucleic Acids Research* **47**:D94–D99.
40. **Seemann T.** 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**:2068–2069.
41. **Akhter S, Aziz RK, Edwards RA.** 2012. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Research*, 2nd ed. **40**:e126–e126.
42. **Roux S, Enault F, Hurwitz BL, Sullivan MB.** 2015. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**:e985–20.
43. **Abbott ZD, Flynn KJ, Byrne BG, Mukherjee S, Kearns DB, Swanson MS.** 2016. *csrT* Represents a New Class of *csrA*-Like Regulatory Genes Associated with Integrative Conjugative Elements of *Legionella pneumophila*. *Journal of Bacteriology* **198**:553–564.
44. **Edgar RC.** 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**:113.
45. **Price MN, Dehal PS, Arkin AP.** 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**:e9490.
46. **Letunic I, Bork P.** 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research* **47**:W256–W259.
47. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.** 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421–9.
48. **Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ.** 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**:119–11.

49. **Li L, Stoeckert CJ, Roos DS.** 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13**:2178–2189.
50. **Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V.** 2018. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology* **430**:2237–2243.
51. **Yu G, Smith DK, Zhu H, Guan Y, Lam TTY.** 2016. ggtree: an rpackage for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* **8**:28–36.
52. **Wickham H.** 2016. ggplot2: Elegant graphics for data analysis. Springer-Verlag New York. ISBN: 978-3-319-24277-4.
53. **Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Néron B, C Rocha EP, Vergnaud G, Gautheret D, Pourcel C.** 2018. CRISPRCasFinder, an update of CRISPRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research* **99**:7536–6.
54. **Biswas A, Staals RHJ, Morales SE, Fineran PC, Brown CM.** 2016. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* **17**:356–14.
55. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup.** 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079.
56. **Quinlan AR, Hall IM.** 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841–842.
57. **Dion M, Labrie S, Shah S, Moineau S.** 2018. CRISPRStudio: A User-Friendly Software for Rapid CRISPR Array Visualization. *Viruses* **10**:602–11.
58. **Wagih O.** 2017. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**:3645–3647.
59. **Hahne F, Ivanek R.** 2016. Visualizing Genomic Data Using Gviz and Bioconductor. *Methods Mol Biol* **1418**:335–351.
60. **D'Auria G, Jimenez-Hernandez N, Peris-Bondia F, Moya A, Latorre A.** 2010. Legionella pneumophila pangenome reveals strain-specific virulence factors. *BMC Genomics* **11**:181.
61. **Gomez-Valero L, Rusniok C, Jarraud S, Vacherie B, Rouy Z, Barbe V, Medigue C, Etienne J, Buchrieser C.** 2011. Extensive recombination events and horizontal gene transfer shaped the Legionella pneumophila genomes. *BMC Genomics* **12**:536.

62. **Mercante JW, Morrison SS, Raphael BH, Winchell JM.** 2016. Complete Genome Sequences of the Historical Legionella pneumophila Strains OLDA and Pontiac. *Genome Announc* **4**:e00866–16.
63. **Brooks L, Kaze M, Siström M.** 2019. A Curated, Comprehensive Database of Plasmid Sequences. *Microbiol Resour Announc* **8**:711–2.
64. **Labonté JM, Suttle CA.** 2013. Metagenomic and whole-genome analysis reveals new lineages of gokushoviruses and biogeographic separation in the sea. *Front Microbiol* **4**:404.
65. **Labonté JM, Hallam SJ, Suttle CA.** 2015. Previously unknown evolutionary groups dominate the ssDNA gokushoviruses in oxic and anoxic waters of a coastal marine environment. *Front Microbiol* **6**:1–10.
66. **Paez-Espino D, Eloë-Fadrosch EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC.** 2016. Uncovering Earth's virome. *Nature* **536**:425–430.
67. **Paez-Espino D, Pavlopoulos GA, Ivanova NN, Kyrpides NC.** 2017. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. *Nature Protocols* **12**:1673–1682.
68. **Chen I-MA, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, Huntemann M, Varghese N, White JR, Seshadri R, Smirnova T, Kirton E, Jungbluth SP, Woyke T, Eloë-Fadrosch EA, Ivanova NN, Kyrpides NC.** 2018. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Research* **47**:D666–D677.
69. **Overholt WA, Hölzer M, Geesink P, Diezel C, Marz M, Küsel K.** 2020. Inclusion of Oxford Nanopore long reads improves all microbial and viral metagenome-assembled genomes from a complex aquifer system. *Environ Microbiol* **578**:425–14.
70. **Jaschke PR, Dotson GA, Hung KS, Liu D, Endy D.** 2019. Definitive demonstration by synthesis of genome annotation completeness. *Proc Natl Acad Sci USA* **116**:24206–24213.
71. **Pascarel-Devilder MC, Renaudin J, Bove JM.** 1986. The spiroplasma virus 4 replicative form cloned in Escherichia coli transfects spiroplasmas. *Virology* **151**:390–393.
72. **Renaudin J, Pascarel MC, Bove JM.** 1987. Spiroplasma virus 4: nucleotide sequence of the viral DNA, regulatory signals, and proposed genome organization. *Journal of Bacteriology* **169**:4950–4961.
73. **Stamburski C, Renaudin J, Bove JM.** 1990. Characterization of a promoter and a transcription terminator of Spiroplasma melliferum virus SpV4. *Journal of Bacteriology* **172**:5586–5592.

74. **Chipman PR, Agbandje-McKenna M, Renaudin J, Baker TS, McKenna R.** 1998. Structural analysis of the Spiroplasma virus, SpV4: implications for evolutionary variation to obtain host diversity among the Microviridae. *Structure* **6**:135–145.
75. **Brentlinger KL, Hafenstein S, Novak CR, Fane BA, Borgon R, McKenna R, Agbandje-McKenna M.** 2002. Microviridae, a Family Divided: Isolation, Characterization, and Genome Sequence of MH2K, a Bacteriophage of the Obligate Intracellular Parasitic Bacterium *Bdellovibrio bacteriovorus*. *Journal of Bacteriology* **184**:1089–1094.
76. **Roux S, Krupovic M, Poulet A, Debroas D, Enault F.** 2012. Evolution and Diversity of the Microviridae Viral Family through a Collection of 81 New Complete Genomes Assembled from Virome Reads. *PLoS ONE* **7**:e40418–12.
77. **Krupovic M, Forterre P.** 2011. Microviridae Goes Temperate: Microvirus-Related Proviruses Reside in the Genomes of Bacteroidetes. *PLoS ONE* **6**:e19893–9.
78. **Quaiser A, Dufresne A, Ballaud F, Roux S, Zivanovic Y, Colombet J, Sime-
Ngando T, Francez A-J.** 2015. Diversity and comparative genomics of Microviridae in Sphagnum- dominated peatlands. *Front Microbiol* **6**:1–10.
79. **Rosario K, Dayaram A, Marinov M, Ware J, Kraberger S, Stainton D, Breitbart M, Varsani A.** 2012. Diverse circular ssDNA viruses discovered in dragonflies (Odonata: Epiprocta). *Journal of General Virology* **93**:2668–2681.
80. **Creasy A, Rosario K, Leigh B, Dishaw L, Breitbart M.** 2018. Unprecedented Diversity of ssDNA Phages from the Family Microviridae Detected within the Gut of a Protochordate Model Organism (*Ciona robusta*). *Viruses* **10**:404–18.
81. **Orton JP, Morales M, Fontenele RS, Schmidlin K, Kraberger S, Leavitt DJ, Webster TH, Wilson MA, Kusumi K, Dolby GA, Varsani A.** 2020. Virus Discovery in Desert Tortoise Fecal Samples: Novel Circular Single-Stranded DNA Viruses. *Viruses* **12**:143–19.
82. **Bryson SJ, Thurber AR, Correa AMS, Orphan VJ, Vega Thurber R.** 2015. A novel sister clade to the enterobacteria microviruses (family Microviridae) identified in methane seep sediments. *Environ Microbiol* **17**:3708–3721.
83. **Tucker KP, Parsons R, Symonds EM, Breitbart M.** 2011. Diversity and distribution of single-stranded DNA phages in the North Atlantic Ocean. *ISME J* **5**:822–830.
84. **Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-
Ngando T, Debroas D.** 2012. Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. *PLoS ONE* **7**:e33641–12.
85. **Reavy B, Swanson MM, Cock PJA, Dawson L, Freitag TE, Singh BK, Torrance L, Mushegian AR, Taliansky M.** 2015. Distinct circular single-stranded DNA viruses exist in different soil types. *Applied and Environmental Microbiology* **81**:3934–3945.

- 717 86. **Zhong X, Guidoni B, Jacas L, Jacquet S.** 2015. Structure and diversity of ssDNA
718 Microviridae viruses in two peri-alpine lakes (Annecy and Bourget, France). *Research in*
719 *Microbiology* **166**:644–654.
- 720 87. **Han L-L, Yu D-T, Zhang L-M, Shen J-P, He J-Z.** 2017. Genetic and functional
721 diversity of ubiquitous DNA viruses in selected Chinese agricultural soils. *Scientific*
722 *Reports* **7**:45142–10.
- 723 88. **Tikhe CV, Husseneder C.** 2017. Metavirome Sequencing of the Termite Gut Reveals
724 the Presence of an Unexplored Bacteriophage Community. *Front Microbiol* **8**:2548.
- 725 89. **Kraberger S, Waits K, Ivan J, Newkirk E, VandeWoude S, Varsani A.** 2018.
726 Identification of circular single-stranded DNA viruses in faecal samples of Canada lynx
727 (*Lynx canadensis*), moose (*Alces alces*) and snowshoe hare (*Lepus americanus*)
728 inhabiting the Colorado San Juan Mountains. *Infect Genet Evol* **64**:1–8.
- 729 90. **Wang H, Ling Y, Shan T, Yang S, Xu H, Deng X, Delwart E, Zhang W.** 2019. Gut
730 virome of mammals and birds reveals high genetic diversity of the family Microviridae.
731 *Virus Evol* **5**:vez013.
- 732 91. **Kirchberger PC, Ochman H.** 2020. Resurrection of a global, metagenomically defined
733 gokushovirus. *eLife* **9**:690–18.
- 734 92. **Lawrence C, Reyrolle M, Dubrou S, Forey F, Decludt B, Goulvestre C, Matsiota-**
735 **Bernard P, Etienne J, Nauciel C.** 1999. Single clonal origin of a high proportion of
736 *Legionella pneumophila* serogroup 1 isolates from patients and the environment in the
737 area of Paris, France, over a 10-year period. *Journal of Clinical Microbiology* **37**:2652–
738 2655.
- 739 93. **Darelid J, Bernander S, Jacobson K, Löfgren S.** 2004. The presence of a specific
740 genotype of *Legionella pneumophila* serogroup 1 in a hospital and municipal water
741 distribution system over a 12-year period. *Scandinavian Journal of Infectious Diseases*
742 **36**:417–423.
- 743 94. **Perola O, Kauppinen J, Kusnetsov J, Kärkkäinen U-M, Lück PC, Katila M-L.**
744 2005. Persistent *Legionella pneumophila* colonization of a hospital water supply:
745 efficacy of control methods and a molecular epidemiological analysis. *APMIS* **113**:45–
746 53.
- 747 95. **Garcia-Nuñez M, Sopena N, Ragull S, Pedro-Botet ML, Morera J, Sabria M.** 2008.
748 Persistence of *Legionellain* hospital water supplies and nosocomial Legionnaires'
749 disease. *FEMS Immunol Med Microbiol* **52**:202–206.
- 750 96. **Oberdorfer K, Müssigbrodt G, Wendt C.** 2008. Genetic diversity of *Legionella*
751 *pneumophila* in hospital water systems. *International Journal of Hygiene and*
752 *Environmental Health* **211**:172–178.

- 753 97. **Mathieu J, Yu P, Zuo P, Da Silva MLB, Alvarez PJJ.** 2019. Going Viral: Emerging
754 Opportunities for Phage-Based Bacterial Control in Water Treatment and Reuse. *Acc*
755 *Chem Res* **52**:849–857.
- 756 98. **Suttle CA, Chan AM, Cottrell MT.** 1991. Use of ultrafiltration to isolate viruses from
757 seawater which are pathogens of marine phytoplankton. *Applied and Environmental*
758 *Microbiology* **57**:721–726.
- 759 99. **Smith HO, Hutchison CA, Pfannkoch C, Venter JC.** 2003. Generating a synthetic
760 genome by whole genome assembly: phi X174 bacteriophage from synthetic
761 oligonucleotides. *Proc Natl Acad Sci USA* **100**:15440–15445.
- 762 100. **Isberg RR, O'Connor TJ, Heidtman M.** 2009. The *Legionella pneumophila*
763 replication vacuole: making a cosy niche inside host cells. *Nature Reviews Microbiology*
764 **7**:13–24.

Supplemental Figure 1. A phylogeny of CsrT sequences reveals mobile elements uncovered by PhiSpy and VirSorter belong to the group IV Lgi elements. CsrT amino acid sequences from mobile elements identified by PhiSpy and VirSorter were extracted and aligned using MUSCLE. A phylogeny was constructed using FastTree and rooted to group I, Trb. Reference sequences were annotated as per Abbott and colleagues (Abbott et al., J Bacteriol. 198, 553–564, 2016) and coloured as indicated.

Supplemental Figure 2. Genome organization of metagenomic contigs that cluster with *Pichovirinae* in the capsid phylogeny. Circularizable metagenomic contigs that cluster with *Pichovirinae* were annotated using Prokka and plotted in GViz with the reference pichovirus Pavin_723. The metagenomic contigs have the unique genome organization VP1-VP4-VP2 characteristic for *Pichovirinae* (S. Roux et al., Plos One. 7, e40418,2012) suggesting that they are indeed part of the *Pichovirinae* sub-family. VP1 (major capsid protein) is shown in purple, VP4 (replication initiation protein) is shown in pink, VP2 (DNA pilot protein) is shown in orange and hypothetical proteins are shown in blue.

Supplemental Figure 3. PAM analysis from putative targets reveals that many targets possess the canonical PAM for their CRISPR sub-type. A sequence logo of PAMs from putative targets shows that many targets possess the canonical PAM for their CRISPR sub-type. This suggests spacers were not over-designated to target groups despite being PAM agnostic during data analysis.