ISME

www.nature.com/ismecomms

ARTICLE OPEN

Phage-encoded ribosomal protein S21 expression is linked to late-stage phage replication

Lin-Xing Chen[1,2], Alexander L. Jaffe[3], Adair L. Borges[2], Petar I. Penev[2], Tara Colenbrander Nelson[4], Lesley A. Warren[4] and Jillian F. Banfield[1,2,3,5,6]✉

The ribosomal protein S21 (bS21) gene has been detected in diverse viruses with a large range of genome sizes, yet its in situ expression and potential significance have not been investigated. Here, we report five closely related clades of bacteriophages (phages) represented by 47 genomes (8 curated to completion and up to 331 kbp in length) that encode a bS21 gene. The bS21 gene is on the reverse strand within a conserved region that encodes the large terminase, major capsid protein, prohead protease, portal vertex proteins, and some hypothetical proteins. Based on CRISPR spacer targeting, the predominance of bacterial taxonomic affiliations of phage genes with those from Bacteroidetes, and the high sequence similarity of the phage bS21 genes and those from Bacteroidetes classes of Flavobacteriia, Cytophagia and Saprospiria, these phages are predicted to infect diverse Bacteroidetes species that inhabit a range of depths in freshwater lakes. Thus, bS21 phages have the potential to impact microbial community composition and carbon turnover in lake ecosystems. The transcriptionally active bS21-encoding phages were likely in the late stage of replication when collected, as core structural genes and bS21 were highly expressed. Thus, our analyses suggest that the phage bS21, which is involved in translation initiation, substitutes into the Bacteroidetes ribosomes and selects preferentially for phage transcripts during the late-stage replication when large-scale phage protein production is required for assembly of phage particles.

ISME Communications; https://doi.org/10.1038/s43705-022-00111-w

INTRODUCTION

Only recently, ribosomal proteins have been recognized in the genomes of viruses [1–3], including those that infect bacteria (i.e., bacteriophages, or phages for short) and archaea. Ribosomal protein S21 (bS21) is the most ubiquitous of the several ribosomal protein-encoding genes that have been reported in virus genomes [1, 2] that range up to 642 kbp in length [2]. In one study, viral bS21 was exclusively (over 90%) detected from aquatic samples [1]. Functional assay experiments confirmed that the bS21 from Pelagibacter phage HTVC008M is incorporated into the 70S ribosomes in *Escherichia coli* [1], yet the in situ expression of bS21, and its potential significance to viral growth remain unclear. One hypothesis is that the viral bS21 protein will substitute for their host equivalent and may preferentially initiate translation of phage mRNA over bacterial mRNA [2]. It has also been noted that phage bS21 homologs may contribute to the specialized translation and/or help phages evade bacterial defenses [4].

The bS21 protein is small (8.5 kD), highly basic, and specific for bacterial ribosomes. It comprises two α-helices connected by a coiled region [4]. It locates between the "head" and the "body" of the small ribosomal subunit (SSU) [4], in contact with the RNA helix formed between the mRNA and the 3′ terminus of the SSU ribosomal RNA (rRNA). This SSU region, also known as the anti-Shine-Dalgarno (ASD) sequence, is crucial for translational initiation by binding the mRNA Shine-Dalgarno (SD) sequence [5]. Generally, when bS21 is missing, translation initiation is disturbed and the mRNA has lower association rates with the SSU [6, 7].

In this study, we report 47 phage genomes that we assigned to five closely related clades of phages whose genomes consistently encode a copy of bS21. Notably, the bS21 gene colocates with genes for structural proteins that are responsible for virion assembly including the large terminase (TerL), portal vertex protein (PVP), prohead protease, and major capsid protein (MCP), but is encoded on the opposite strand. We manually curated all genomes and two outgroup phage genomes (thus 49 in total) to ensure accurate protein sequence prediction and to determine overall genome structure and genome sizes (when complete genomes were achieved). Nine genomes were completed, the largest of which is 331 kbp in length. CRISPR-Cas spacer targeting, the taxonomic similarity of phage proteins to bacterial proteins, including bS21, all predicted that these phages infect freshwater Bacteroidetes species. We find that phage bS21 gene expression is significant during late-stage phage replication, likely specifically translating genes encoding core structural proteins that are essential to virion assembly and the lytic cycle.

[1]Department of Earth and Planetary Science, University of California, Berkeley, CA, USA. [2]Innovative Genomics Institute, University of California, Berkeley, CA, USA. [3]Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA. [4]Department of Civil and Mineral Engineering, University of Toronto, Toronto, ON, Canada. [5]Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA, USA. [6]Chan Zuckerberg Biohub, San Francisco, CA, USA. ✉email: jbanfield@berkeley.edu

Received: 10 November 2021 Revised: 31 January 2022 Accepted: 3 February 2022
Published online: 30 March 2022

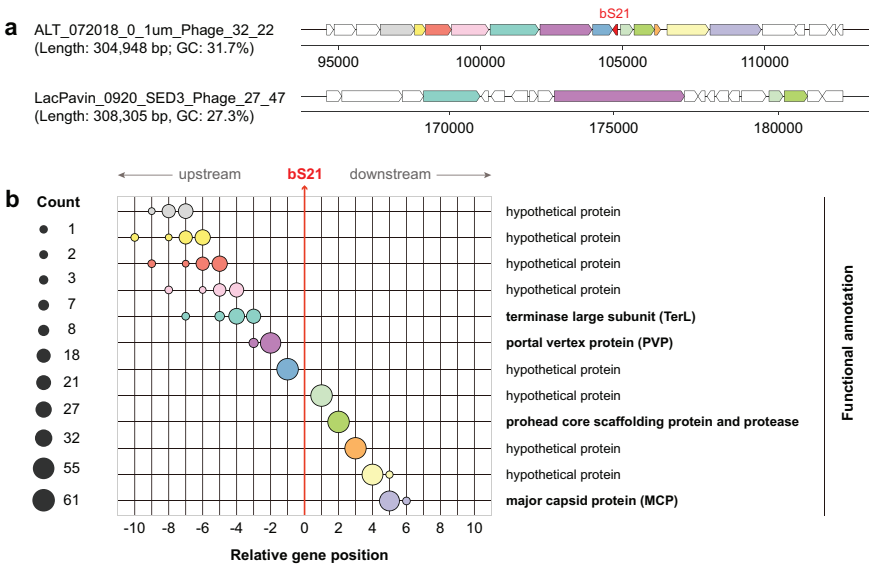

Fig. 1 Genetic context of the genes encoding bS21 in the phage genomes. a Examples of genetic context of phage genomes with and without bS21. The annotation of protein-coding genes is the same as indicated in **b** by different colors. Those in white are genes not shown in subfigure (**b**). **b** Summary of genetic context of all phage genomes encoding bS21. The relative position of genes near the bS21 gene is shown, and the size of circles indicates the number of phages with a gene belonging to a given protein family (annotation shown on right) at that relative position. Only the 12 most frequent families are shown. The details of the genetic context are shown in Supplementary Fig. 1.

RESULTS

Discovery of closely related phage sequences with the conserved genetic context of bS21

Multiple phage-related sequences with a conserved genomic context were detected from several freshwater metagenome-assembled datasets (see Methods). Genes for bS21, TerL, PVP, prohead core scaffolding, and protease protein (hereafter prohead protease for short), and MCP are encoded in the genomic region. BLASTp search of the TerL sequences against the ggKbase sequences (ggkbase.berkeley.edu) obtained a total of 47 unique scaffolds with the conserved genomic region (Supplementary Table 1). Two related phages were included as outgroups for comparative analyses. The corresponding samples were collected from freshwater lakes or reservoirs (one from a wastewater treatment plant), and all but three were from the oxic layer (see Methods for details).

General features of manually curated genomes

All the 49 phage sequences were manually curated to fill scaffolding gaps and fix the assembly errors, and nine of them (including one outgroup phage) were curated to completion (circular and no gaps or local assembly errors) (Supplementary Table 1). A total of 14 related phage genomes from IMG/VR were also included for further analyses. The eight bS21-encoding complete genomes had genome lengths of 293–331 kbp, GC contents of 31.0–33.7% and encoded 350–413 protein-coding genes (coding density, 91.1–94.9%), with 5–25 (average 17) tRNA genes. No alternative coding signal (i.e., stop codon reassignment) was detected in any genome. In comparison, the outgroup complete genome has a size of 308 kbp (450 protein-coding genes, 6 tRNAs, 94.7% coding density) and GC content of 27.3%.

Genomic context of bS21 in phages

Genomic context analyses for bS21 genes showed a highly conserved gene architecture across phage genomes in proximity to the region encoding bS21 (see Fig. 1a for example). Specifically, we found that bS21 was consistently located in between two hypothetical protein families (positions 1 and −1 in Fig. 1b and Supplementary Table 2), with core structural proteins—including the TerL, PVP, prohead protease, and MCP—generally located within five genes in both the upstream and downstream DNA. Other hypothetical proteins were also consistently found in this region, although their positions were more variable upstream (positions −4 through −10, Fig. 1b). Importantly, the bS21 gene was consistently encoded in the reverse strand relative to the conserved hypothetical and structural protein genes (Fig. 1a and Supplementary Fig. 1).

Phylogeny of bS21-encoding phages

Phylogenetic analyses based on TerL suggested the phages belonging to several groups, we thus assigned them to clades a–e (Fig. 2 and Supplementary Table 1). Most of the phages belong to clades c, d, and e, and they have a broader environmental distribution than clades a and b. Interestingly, we found that some phages within a single clade were from distant sampling sites. Closer inspection indicated they also shared large genomic fragments with high similarity (82–98% for nucleotide sequences; Supplementary Fig. 2). Comparative genome-wide analyses of the complete genomes from the same site but sampled at different time points showed sequence variations in some genes (Supplementary Fig. 3).

TerL phylogeny, constructed using sequences from this study and NCBI RefSeq sequences, indicated the most closely related classified phages belong to Caudovirales of either the Myoviridae or Ackermannviridae (Supplementary Fig. 4). A phage baseplate assembly protein was encoded in most curated genomes. This is an important building block for members of Siphoviridae and Myoviridae [8], so we concluded that the bS21-encoding phages are myoviruses.

Predicted bacterial hosts of bS21-encoding phages

To predict host-phage relationships we first used CRISPR-Cas spacers targeting. While none of the 16.5k unique spacers from the relevant metagenomes targeted any of the curated phage genomes from the same sampling sites, a single cross-site target was detected. Specifically, MIW1_072018_0_1um_scaffold_78 was targeted by a spacer (24 nt and no mismatch) from a MIW2 *Flavobacterium* genome (affiliation: Bacteroidetes, Flavobacteria). We then predicted the bacterial hosts based on the bacterial taxonomic affiliations of the phage gene inventories as previously described [2] (Supplementary Table 3). The results indicated that all of the phages infect members of Bacteroidetes, which were detected in 43 out of 45 samples (Fig. 3 and Supplementary Table 4). The two metagenomic samples without Bacteroidetes identified were both collected via filtering through 0.2

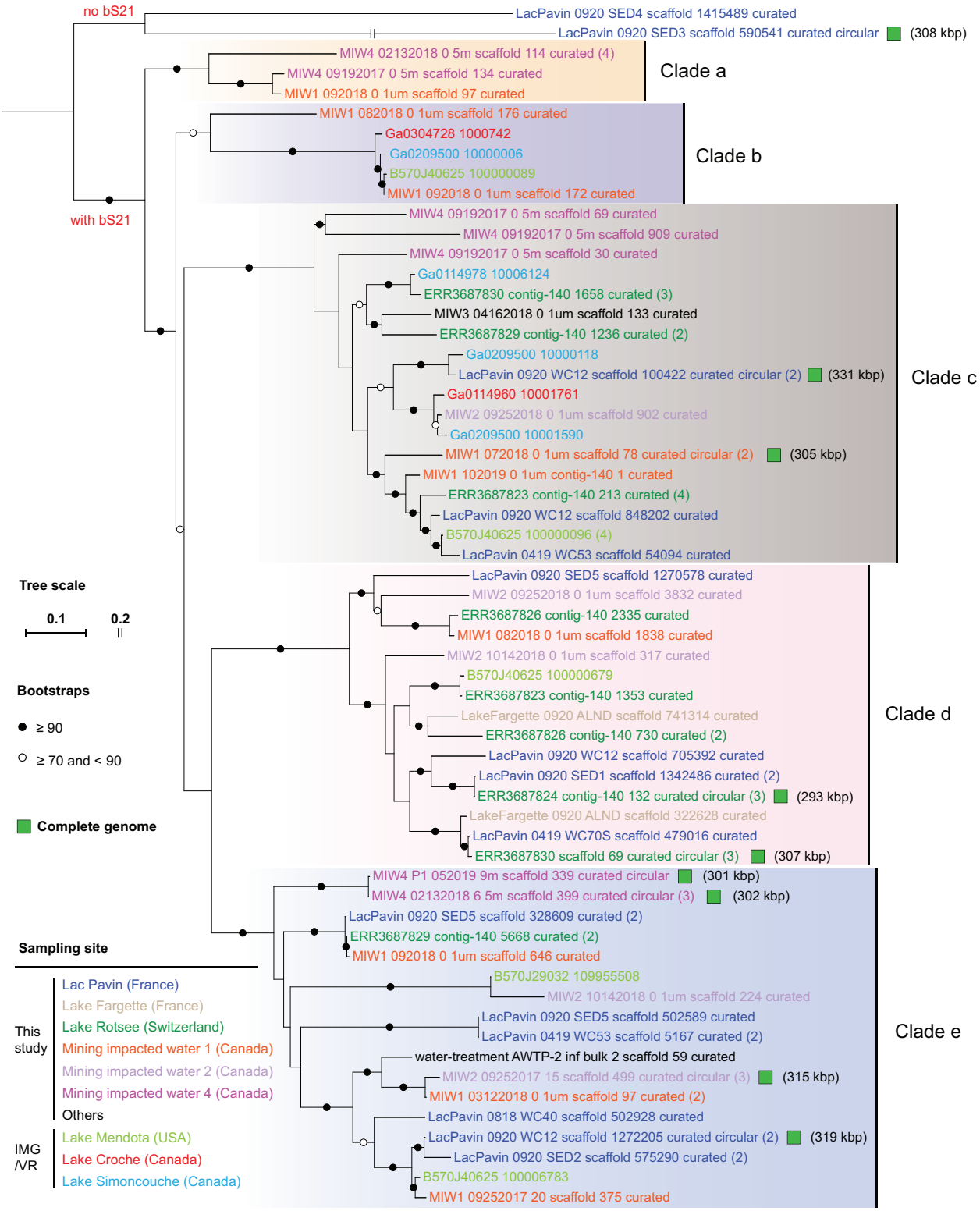

Fig. 2 The phylogeny of bS21 phages based on the large terminal (TerL) protein sequences. Two closely related phages without bS21 encoded were included as outgroups (shown at the top of the tree). The genomes are assigned to five clades (a, b, c, d, and e) based on the topology of the phylogenetic tree. The numbers in the brackets following the scaffold names indicate the total counts of the same scaffold detected from the corresponding sampling sites. The genomes that were manually curated to completion (circular and no gap) are indicated by squares, and the genome sizes are shown in brackets.

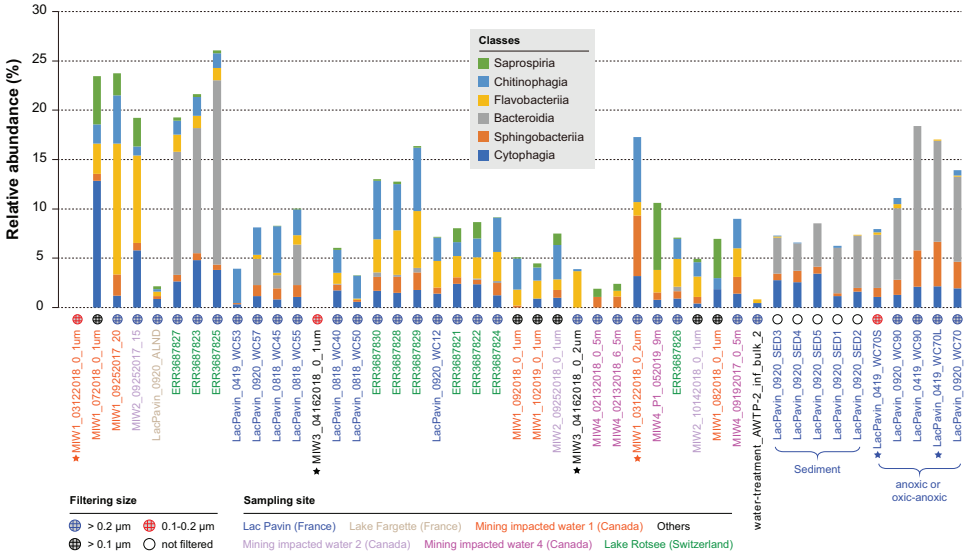

Fig. 3 The relative abundance of the Bacteroidetes classes in all the analyzed samples in this study. The microbial communities were profiled based on ribosomal protein S3 (rpS3) assigned to the Bacteroidetes classes. The sampling sites were indicated by colored names, and the filter sizes used during sampling are shown by circles. The three pairs of filter samples are indicated by colored stars.

µm and onto 0.1 µm pore size filters. Bacteroidetes were detected in both of the corresponding 0.2 µm fraction samples (Fig. 3).

We profiled the co-detection of phage clades and Bacteroidetes classes to test for specific connections (Supplementary Fig. 5). However, this was uninformative because most samples contained more than one class. However, phages from clades a and b are unlikely to infect class Bacteroidia members, as they did not co-occur in any sample.

Comparison of bacterial and phage-encoded bS21

Phylogenetic analyses revealed that bS21 protein sequences from phages (this study) and the bacterial bS21 sequences (from the corresponding samples and NCBI RefSeq) clustered separately (Supplementary Fig. 6). The bacterial bS21 sequences that are most similar to phage bS21 were from Bacteroidetes, mostly from the Flavobacteriia class (Supplementary Table 5). We aligned and compared the Bacteroidetes and phage bS21 sequences and mapped the divergent and non-divergent residues to the model of the ribosome of *Flavobacterium johnsoniae* (Fig. 4a). Multiple divergent positions are located at the beginning of the bS21 sequences and four residues (Arg21, Phe23, Asp25, and Thr28) were significantly divergent (Fig. 4b).

Bacteroidetes usually lack the SD sequences. It was recently reported that the bS21 Tyr54 (numbering in *F. johnsoniae*) is an important residue for blocking the ASD in the 16S rRNA within the ribosome [9]. Our analyses predict that all the analyzed bacterial and phage bS21 in this study have an amino acid with an aromatic ring (often Tyr54 but in a few cases His54, and in one case Phe54) at the position of Tyr54 in *F. johnsoniae* (Fig. 4c, d and Supplementary Fig. 6). This conservation of the aromatic property in phage bS21 should ensure stacking interaction with Adenine 1534 (numbering in F. johnsoniae 16S) from the ASD. In that way, phage bS21 mimics Bacteroidetes bS21 in the region where it binds the ribosome but differs from it in the region where the mRNA would bind.

In contrast, the C-terminal regions of both the bacterial and phage bS21 sets were highly divergent (Fig. 4d). However, the phage C-terminal regions are generally conserved within the clades defined based on TerL phylogeny (Fig. 2 and Supplementary Fig. 7).

Metabolic potentials of bS21-encoding phages

Functional annotation of the predicted protein-coding genes revealed that in addition to bS21, these phages carry other genes related to protein production and stability (Supplementary Table 6). Examples include protein folding chaperones and Clp protease, suggesting the importance of controlling the proteostasis network of the cell. Interestingly, we also identified many genes involved in sugar-related chemistry and polysaccharide biosynthesis. Many of these genes were predicted to perform chemical transformations related to the biosynthesis of lipopolysaccharide, a major component of the Gram-negative bacterial outer membrane. We interpret this as a potential mechanism to remodel the cell surface and prevent superinfection by competitor phages, a strategy common to the phage lysogenic cycle. These phages lack detectable integration machinery (no gene for integrase or resolvase was detected), suggesting the possibility of a non-integrative long-term infection state such as pseudolysogeny [10].

Clustering analyses of 22 phages with a minimum genome size of 100 kbp (including the two outgroup genomes) based on the presence/absence of protein families indicated they shared a total of 16 protein families (Supplementary Fig. 8 and Supplementary Table 7). Phosphate starvation-inducible protein PhoH ("fam582") was the only predicted protein detected in all 22 phages (excluding the shared predicted proteins in the conserved rpS21-encoding region described above). Other common protein families include those related to DNA replication (e.g., DNA primase/helicase, DNA polymerase, HNH endonuclease, thymidylate synthase (EC:2.1.1.45), deoxyuridine 5′-triphosphate nucleotidohydrolase (EC:3.6.1.23)), those associated with virion assembly (e.g., a phage tail sheath protein, phage baseplate assembly protein W), and those for other functions (e.g., chaperone ATPase, alpha-amylase, DegT/DnrJ/EryC1/StrS aminotransferase).

Temporal and spatial distribution and activity of bS21-encoding phages in Lake Rotsee

To reveal the spatial and temporal distribution of the bS21-encoding phages, we focused on the Lake Rotsee data and profiled phage occurrence based on the sequencing coverage in the metagenomic datasets. The Lake Rotsee samples were collected from the oxic (7 samples) and anoxic (3 samples) layers of the water column. The bS21-encoding phages were readily detected in oxic samples, especially in the under-ice samples when the whole water column was oxic (Fig. 5a).

Rotsee Lake RNA reads were mapped to the phage genomes curated from this site to reveal the transcriptional activities of bS21-encoding phages (Fig. 5b). In general, the phages were likely

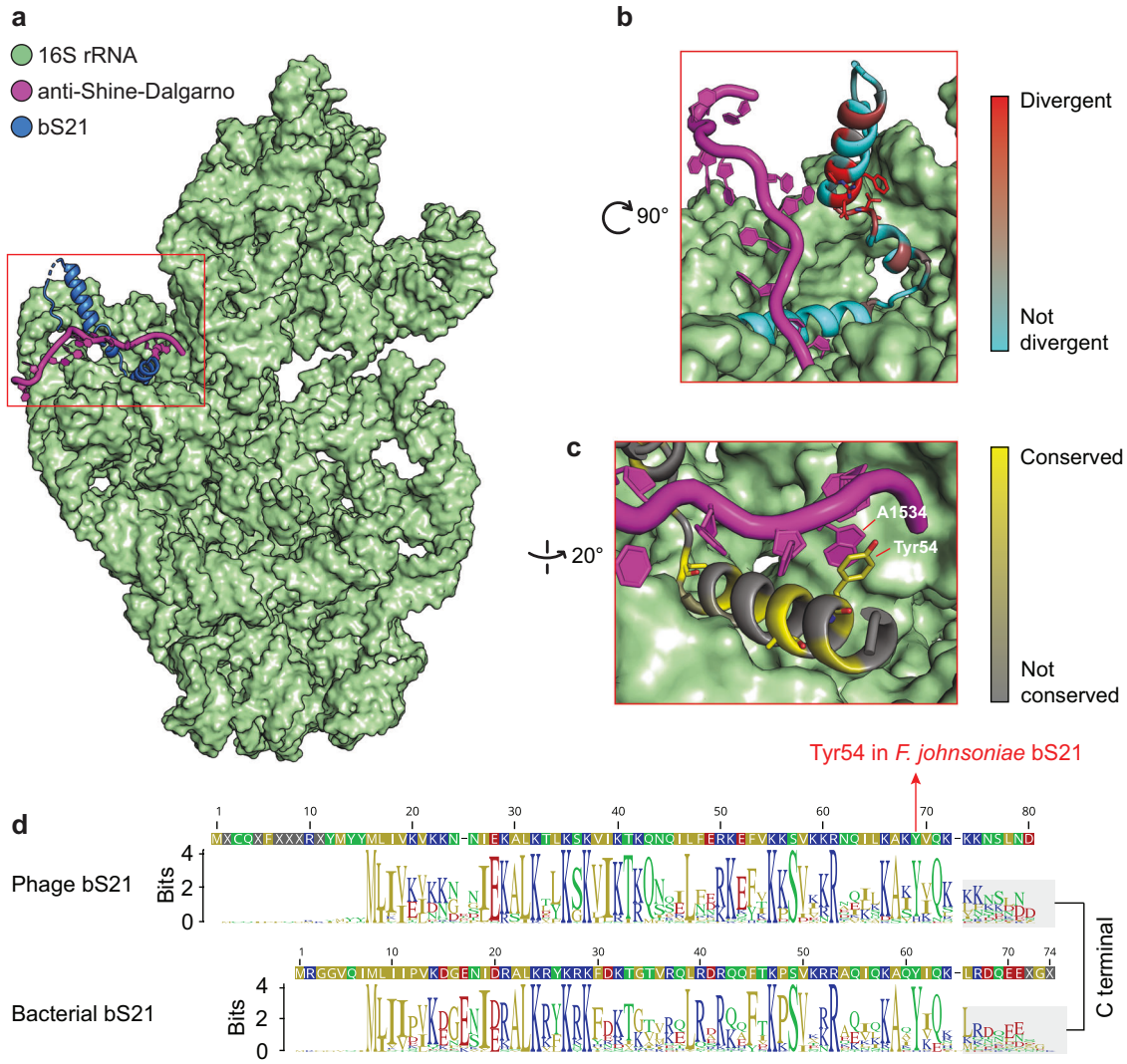

Fig. 4 Conservation and differences between phage and bacterial bS21. a Location of bS21 (blue) within the 16S rRNA (green) and the ASD (magenta) of the *F. johnsoniae* ribosome (PDB ID: 7JIL) [9]. bS21 is in the neck region of the 16S rRNA, interacting closely with the 3′ end of the 16S rRNA, where the ASD is located. The 16S rRNA is shown from the subunit interface direction. **b** Zebra2 divergency results from an alignment of phage and bacterial bS21 sequences mapped on *F. johnsoniae* bS21. Divergent positions between phage and bacterial bS21 are shown with red. **c** Zebra2 conservation results from the same alignment as in (**b**) mapped on *F. johnsoniae* bS21 with conserved residues shown in yellow. The stacking interaction between Tyr54 and Adenine 1534 is indicated. **d** The sequence logo and consensus sequences of phage and bacterial bS21 alignments and the corresponding position of Tyr54 in *F. johnsoniae* bS21 in the alignment are highlighted. The C-terminal parts are highlighted with gray backgrounds.

to be most transcriptionally active in the oxic water columns. A total of 736 genes were transcribed in at least one sample (Supplementary Table 8), those for MCP, an AAA ATPase, tail sheath protein, bS21, FKBP-type peptidyl-prolyl cis-trans isomerase, and a methyltransferase FkbM domain protein are among the top 100 most highly transcribed. The high transcriptional activities of MCP in five phages indicated they were in the late stage of replication at the time of sampling.

The transcriptional behavior of phage bS21 genes

To seek evidence of a transcriptional relationship involving bS21 and other genes we focused on the three phages that were most active based on the transcriptional level of their 19 shared single-copy genes (Fig. 6a). bS21 had very similar (but slightly lower) transcriptional activities as a neighboring gene (hereafter, bS21_CN gene) encoded on the opposite strand. The bS21_CN gene encodes a hypothetical protein (protein family: fam498) and was not detected in the two outgroup phages without bS21 (Supplementary Table 6). Interestingly, a comparison of the phylogenies of bS21 and bS21_CN showed a very similar evolutionary pattern (Supplementary Fig. 9), likely suggesting their potential functional relationship in the bS21-encoding phages.

Inspection of the RNA reads mapping profiles indicated that the conserved region encoding bS21 and core structural proteins was not transcribed as an operon, whereas bS21 and bS21_CN, MCP and its upstream hypothetical protein gene, and prohead protease and its downstream hypothetical protein gene may each be transcribed together (Fig. 6b). Given the observed RNA expression patterns, we conclude that the phage-encoded bS21 genes were actively transcribed during late-stage replication, along with other core structural proteins.

Genomic context of bS21 genes in published phage genomes

To determine whether the phage bS21 genes are generally co-located with those for core structural proteins in diverse phages, we profiled the genomic context of bS21 in 900 published bS21-encoding phages

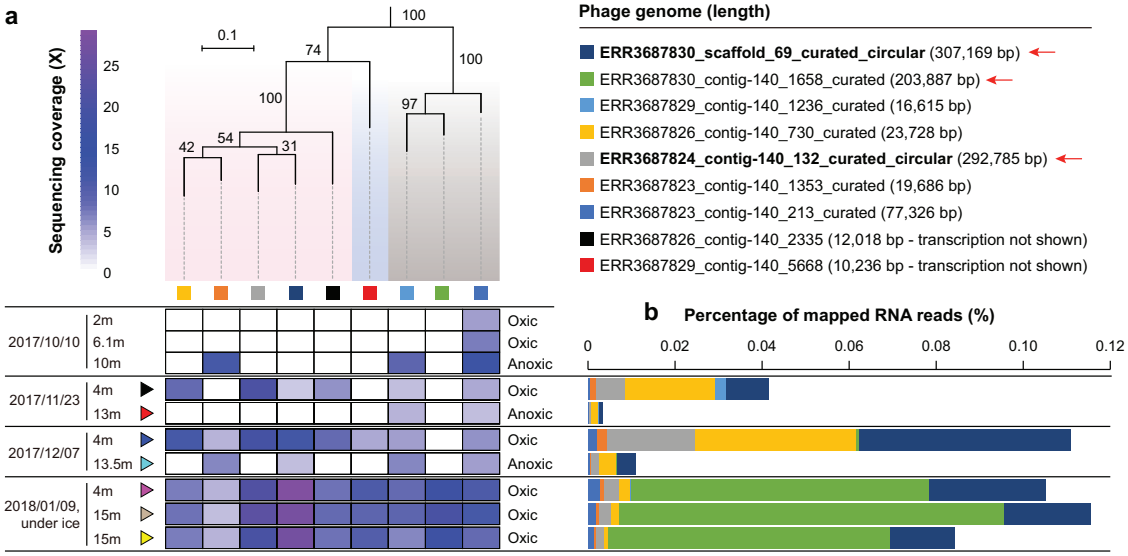

Fig. 5 The spatial and temporal distribution and activity of bS21 phages at Lake Rotsee. a The sequencing coverage of each phage genome in each metagenomic dataset is shown in the heatmaps. The phages are phylogenetically clustered based on their TerL protein sequences (bootstraps shown in numbers), the colored backgrounds are the same as shown in Fig. 2 for different clades. The sampling time points and depths are shown on the left, and the oxygen conditions are indicated by colored circles on the right. Two replicates were sequenced from the 15 m sample collected in 2018. **b** The percentage of mapped RNA reads to the phage genomes in the corresponding samples (rows labeled in (**a**)). The mapped RNA reads had a minimum similarity of 98% to the phage genomes. No RNA data were generated for the three samples collected on October 10, 2017. See the figure legend for each genome in the upper right, the circular genomes have names in bold font.

[2, 11] (Supplementary Table 9). Functional annotations were performed for the upstream and downstream ten genes of the bS21 genes using pVOG (Supplementary Table 10). Of the 20 most abundant pVOGs, 6 were related to core structural assembly (Fig. 7a), i.e., prohead protease ($n = 310$), MCP ($n = 154$), PVP ($n = 120$), TerL ($n = 78$), neck protein ($n = 70$), and a tail sheath protein ($n = 29$). A total of 388 genomes contained at least one of these genes within ten genes of bS21, and eight had all of these six core structural proteins in close proximity. Three pVOGs were related to DNA processing, i.e., an exonuclease ($n = 37$), an endonuclease ($n = 32$), DNA helicase ($n = 30$). Other pVOGs included Hsp20 heat shock protein ($n = 127$), two ATP-dependent CLP proteases ($n = 50$ and 47, respectively), and lysozyme (for lysis; $n = 29$). Interestingly, the prohead protease and the MCP pVOG genes are very close to the bS21 gene (generally 2–4 genes; Fig. 7b), as in the bS21-encoding phage genomes analyzed in this study (2–6 genes away; Fig. 1 and Supplementary Fig. 1).

We respectively predicted the hosts of the bS21-encoding phages with the four most dominant pVOGs within ten genes of bS21 (Fig. 7c and Supplementary Table 11). The bacterial hosts are diverse and include Proteobacteria, Bacteroidetes, and Firmicutes.

DISCUSSION

The Bacteroidetes-infecting bS21-encoding phages are abundant and active in oxic water columns

Bacteroidetes, including Bacteroidia or Flavobacteriia, were ubiquitous in the analyzed samples and are the predicted hosts of most of the newly reported bS21-encoding phages (Fig. 3). Bacteroidia spp. are strictly anaerobic [12] whereas Flavobacteriia spp. are strictly aerobic [13, 14], in line with the general detection of Bacteroidia in anoxic samples and Flavobacteriia spp. in oxic samples (Fig. 3). The majority (51/61) of the phage-encoding bS21 genes in this study are most similar to those from Flavobacteriia (Supplementary Table 5), explaining why most of the bS21 phages, and the most transcriptionally active subset, were detected in the oxic water column of Lake Rotsee. Flavobacteriia spp. likely degrade high molecular weight compounds such as polysaccharides and proteins [13, 15]. Based on the detection of phage genes for these functions (Supplementary Table 6), we conclude that the bS21-encoding phages may primarily impact the abundance of Flavobacteriia and thus alter its impact in the community.

Features of phage bS21 that may enable substitution for bacterial bS21 in ribosomes

Some highly similar phages detected in lakes separated by thousands of miles (for example, Supplementary Fig. 2b), share identical bS21 genes in conserved genomic context, despite sequence divergence throughout the rest of the genome. This points to the high functional importance of bS21 in the phages. The conservation of the C termini of phage bS21 proteins across all of the phage clades that we defined using TerL phylogeny (Supplementary Fig. 8) may indicate that the phage bS21 C termini are important for the phage proteins to substitute for the bacterial bS21 in the ribosomes.

bS21 is composed of two α-helices that interact in different ways with the 16S rRNA. The N-terminal α-helix is situated on top of the ASD where the mRNA would bind, whereas the C-terminal α-helix is tucked between the ASD and the rest of the 16S rRNA (Fig. 4a) and anchors bS21 to the 16S rRNA [4]. Our results are congruent with these observations since sequences of phage bS21 show strong divergences from the bacterial sequences in the N-terminal α-helix (Fig. 4b). These changes should not alter the binding of bS21 to the 16S rRNA but may provide specificity for attracting phage-specific mRNA.

All bS21 phages had either Tyr54 (54 out of 61) as occurs in *Flavobacteriia johnsoniae*, or a residue with an aromatic ring (7 out of 61) near the C-terminus of bS21 (Supplementary Fig. 6). It has been suggested that this helps to block the ASD sequences [9]. Thus, the phage bS21 may function in essentially the same way as that of their Bacteroidetes hosts. We infer that once phage bS21 proteins are available, the bS21 incorporates into the bacterial ribosomes, potentially enabling the phages to have their mRNA transcripts translated preferentially over the host transcripts.

Why might phages use bS21 to hijack the ribosome in the late stage of replication?

Our RNA analyses showed the simultaneous transcription of the genes for bS21 and core structural proteins (e.g., capsid proteins,

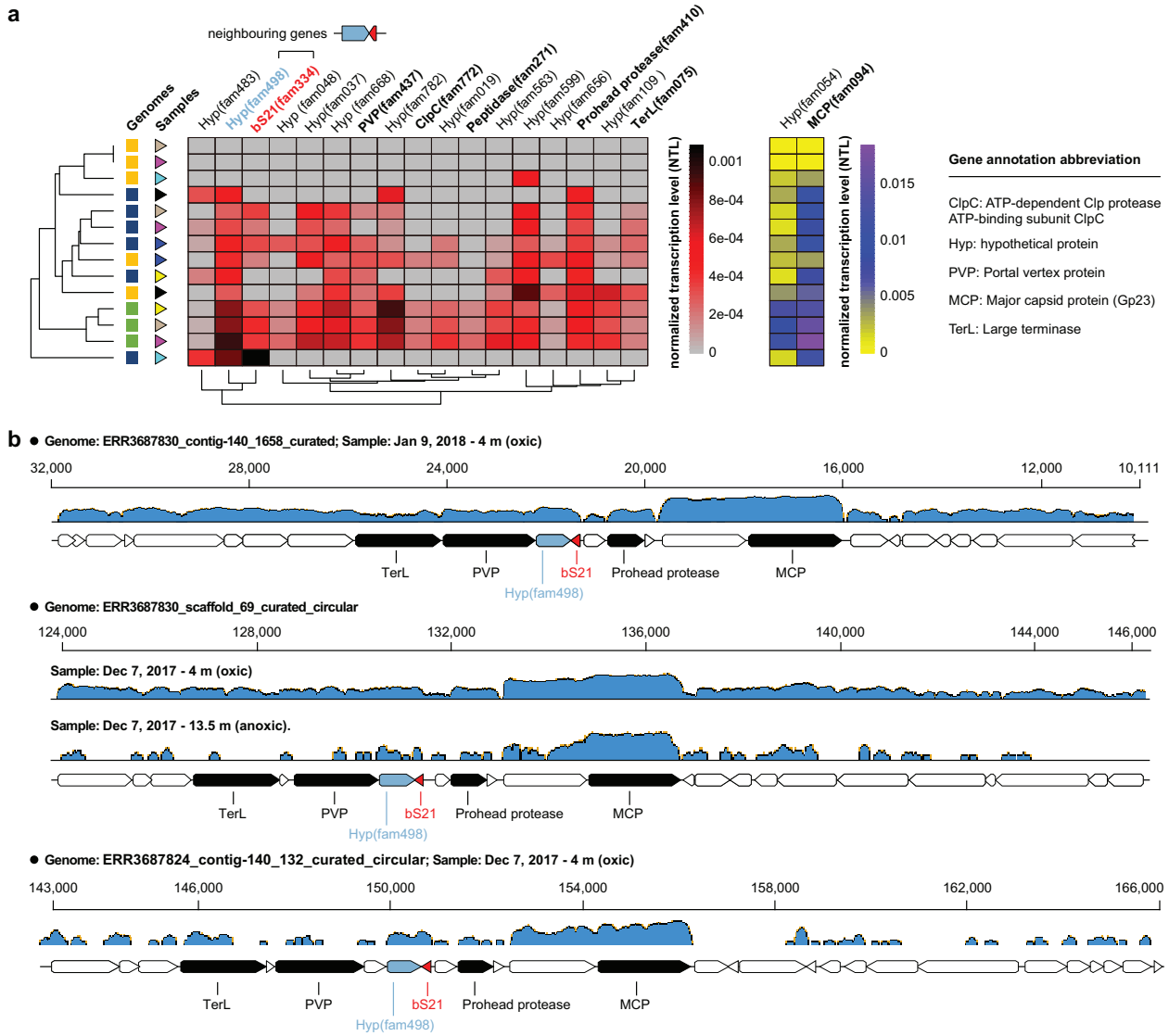

Fig. 6 The transcription levels of bS21 and core structural protein genes. a The normalized transcriptional level (NTL) of shared single-copy protein families of three phages (indicated by arrows in Fig. 5b) with ≥1000 RNA reads mapped. Two families (including MCP) are listed on a different scale due to their much higher transcription levels. Refer to Fig. 5 for shape symbols that designate phage genomes and samples. **b** Examples of RNA mapping profiles indicating the co-transcription of some genes neighboring bS21. Hypothetical protein genes are shown in white.

prohead protease, TerL, scaffolding proteins, and tail proteins; Fig. 6 and Supplementary Table 8), suggesting the potential significance of bS21 in the late-stage replication [16]. The genomic placement and timing of transcription of bS21 genes make sense given the huge number of proteins needed for assembly and packaging. On the other hand, a previous study showed that stalling of phage protein synthesis is an important defense strategy for a Bacteroidetes species, i.e., *Cellulophaga baltica* (class Flavobacteriia [17]). The replacement of bacterial bS21 with phage bS21 may be a mechanism that counters this defense strategy. It is also possible that the phage bS21 protein is encapsulated within the viral particles and delivered into host cells along with the viral genomes to modulate translation from the onset of infection. This hypothesis could be tested by performing proteomics analyses on the concentrate of encapsulated phage particles.

Our genetic context analyses of published bS21-encoding phage genomes showed that many bS21 genes are co-located with genes for core structural proteins, and sometimes with genes for DNA replication and lysis. Thus, we suggest that the acquisition and timing of expression of bS21 may be a more general and consistently evolved phenomenon across diverse phage lineages. This motivates future analyses that could experimentally investigate phage-encoded bS21, especially when the genes co-occur with DNA processing or lysis genes, and test the hypothesis that bS21 may be important for efficient translation of the nearby genes.

CONCLUSION

By carefully manual curating nine huge phage (also sometimes called jumbo phages) [18] genomes to completion, we accurately determined genome sizes, genome organization, and gene inventories (e.g., lack of other genes encoding ribosomal proteins). Partial curation further constrained the genome sizes of other phages and ensured that all key protein sequences are correct. Given RNA expression of bS21 and the flanking structural proteins in several transcriptionally active phages, we suggest that the bS21 genes in phages that infect some freshwater Bacteroidetes species are important in the late-stage replication. Our analyses of publicly

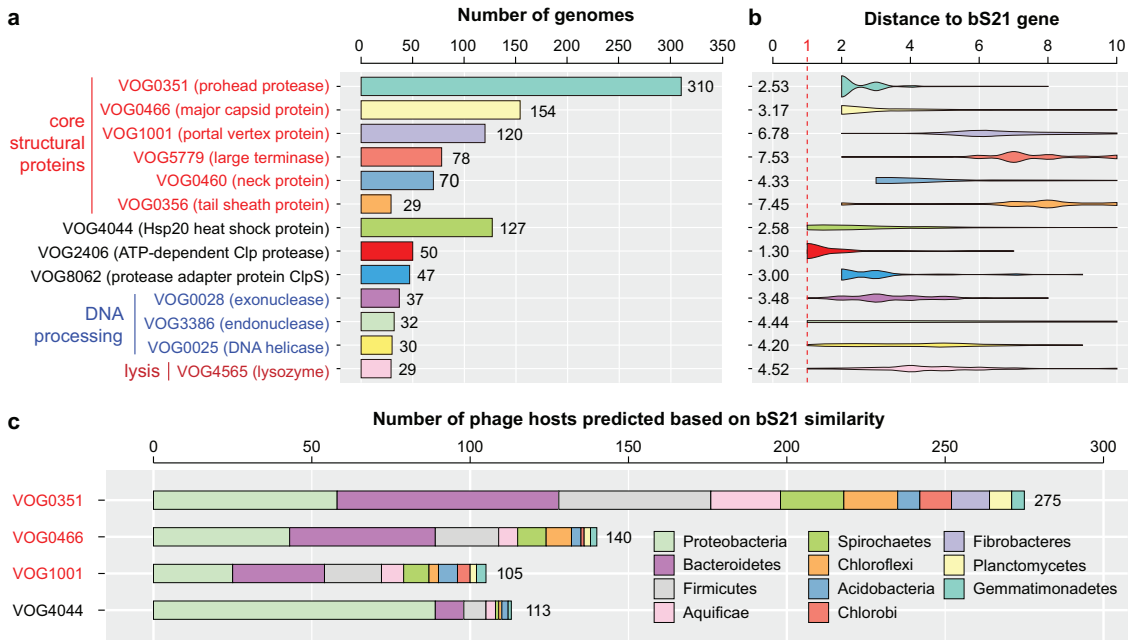

Fig. 7 Neighboring genes within 10 genes of bS21 in published bS21-encoding phage genomes. a The annotation and corresponding functional category (if assigned) of the 20 most commonly detected pVOG genes and their predicted functions are shown on the left, the total number of genomes with the gene are shown on the right. **b** The distribution of the distance of each gene to bS21 in the genomes. The position of genes next to bS21 (thus distance = 1) is highlighted using a red dashed line. The average distance of each gene to bS21 is shown on the left. **c** The predicted hosts of bS21-encoding phages with the top 4 most abundant genes detected within 10 genes of bS21. The total count of hosts is shown on the right.

available genomes suggest that this phenomenon may be more general across phage groups that infect diverse bacterial hosts.

MATERIALS AND METHODS

A global search for related phage sequences encoding bS21

To detect related phage sequences encoding the detected bS21 gene, a BLASTp search using the TerL protein was performed against all metagenomic datasets in ggKbase (ggkbase.berkeley.edu). All the contigs/scaffolds bearing the top BLASTp top hits were manually checked for bS21. It was confirmed that when the TerL similarity dropped to below ~70% similarity, the bS21 that located near the TerL was no longer observed. Two of the scaffolds with the highest TerL similarity to that of bS21 containing contigs/scaffolds but without bS21 were included in this study as an outgroup for phylogenetic and comparative genomic analyses.

Site description of the samples with bS21-encoding phages detected

The samples with bS21-encoding phages detected were collected from several freshwater lakes, including Lac Pavin (45°30'0"N, 2°52'60"E) and Lake Fargette (45°44'24"N; 3°27'39"E) in France, Lake Rotsee (47°04'11"N, 8°18'51"E) in Switzerland, and three freshwater reservoirs of mining-impacted water in Canada. The freshwater samples were conservation with dry ice before genomic DNA was extracted. See related publications for details of sampling procedures [19–21].

Manual curation of phage genomes

To ensure the phage genomes reported in this study were without any assembly errors that could be detected, manual curation was conducted on each of them individually as described previously [22]. Firstly, the corresponding metagenomic quality paired-end reads were mapped to the scaffold using bowtie2 [23] with default parameters, and the sam file was filtered to remove unmapped pairs using shrinksam (https://github.com/bcthomas/shrinksam). The shrunk sam file was imported into Geneious Prime version 2021.1.1 (https://www.geneious.com). The manual check was performed to identify any assembly errors (regions lacking paired read support), which were subsequently fixed using unplaced paired reads as previously described [22]. Those scaffolds with sufficient coverage (generally ≥20×) were selected for curation to completeness. For all of the curated scaffolds, the paired high-quality reads were re-mapped to the genome sequences for a final check. Regions that could not be covered by reads are scaffolding gaps and indicated by 10 Ns.

Retrievement of related phage genomes from IMG/VR

To reveal the distribution of related phages encoding bS21 in public databases, the IMG/VR database (version 2020-10-12_5.1) was downloaded [24]. The IMG/VR proteins were searched against the TerL protein sequences predicted from the curated genomes (see above) using BLASTp. The BLASTp hits with a minimum similarity of 97% and longer than 400 aa in length were filtered for manual inspection to include only those from IMG datasets that have been published. As a result, a total of 14 IMG/VR genomes from three sampling sites [25, 26] which are highly similar to our bS21-encoding phages were included in our analyses (Supplementary Table 1), which were determined to be public by checking "published" on IMG/VR website for usage.

Phylogenetic analyses

To reveal the phylogenetic relatedness of the phages, phylogenetic trees were built using TerL. The protein sequences were aligned using MUSCLE [27] with default parameters and filtered using trimal-trimAl v1.4.rev15 [28] to remove columns comprising ≥90% gaps. The trees were built by IQtree version 1.6.12 [29] with 1000 bootstraps using the "LG+G4" model. For the phylogeny of bS21, the alignment, filtering and tree construction of bacterial and phage bS21 protein sequences were performed the same as did for the TerL sequences.

Comparative analysis between phage and bacterial bS21

To understand the differences between phage and bacteria encoded bS21 we built a sequence alignment of phage bS21 sequences, most closed bacterial bS21 sequences in the corresponding metagenomic samples, and publicly available bacterial bS21 sequences. The alignment was built using MUSCLE v3.8.31 with default parameters [27]. Zebra2 [30] and TwinCons [31] were used to perform a search for divergent positions in the generated alignment. Results were mapped on the bS21 structure from the Bacteroidetes representative *F. johnsoniae* (PDB ID: 7JIL) [9].

Genomic context analysis

Protein sequences for the combined phage set were predicted using Prodigal version 2.6.3 using the "-m" model [32]. To investigate the genomic context of bS21 in phage genomes, we gathered protein sequences within a 10 open

reading frame (ORF) distance (or, to the scaffold end) in both genomic directions of the identified bS21 genes. Each ORF was assigned a genomic position relative to the bS21 (position 0). All "neighboring" proteins were subjected to a two-part, de novo protein clustering pipeline in which proteins are first clustered into "subfamilies" and highly similar/overlapping subfamilies are merged using and HMM-HMM comparison approach (--coverage 0.75) (Méheust et al. 2019). We next searched all neighboring proteins against Pfam (pfam.xfam.org) and pVOG (dmk-brain.ecn.uiowa.edu/pVOGs) [33] HMM collections and retained hits with *e* value <1e–5. Consensus annotations for each family were obtained by computing the HMM with the most above-threshold hits among member sequences of the family (minimum 5% of member sequences). If no HMMs met these thresholds, the protein family was labeled "hypothetical" (hyp). Finally, we plotted the frequency at which each protein family was found as a function of its relative position to the focal bS21 gene (Fig. 1). In addition, we plotted genomic diagrams of individual regions of interest using gggenes (wilkox.org/gggenes).

Microbial community composition analyses

To reveal the community composition of the samples analyzed in this study, the protein-coding genes were predicted using Prodigal [32] (-m = meta) from all metagenomic assembled sequences with a minimum length of 1000 bp. The ribosomal protein S3 (rpS3) was predicted using hmmsearch (version HMMER 3.3) [34] with the hmm database from TIGRFAM [35]. For taxonomic information, the predicted rpS3 protein sequences (minimum length, 100 aa) were searched using BLASTp against the rpS3 proteins from NCBI RefSeq and those of Candidate Phyla Radiation reported previously [36], and taxonomy of the best hits was used. The nucleotide sequences of the predicted rpS3 genes (minimum length, 300 nt) were clustered using cd-hit-est (parameters: -c = 0.97, -aS = 0.5, -aL = 0.5, -G = 1) [37] to generate a non-redundant dataset. The quality metagenomic reads from each sample were individually mapped to the non-redundant dataset and filtered allowing ≤3% mismatch. The coverage of each rpS3 sequence was determined by the total mapped bases divided by its length.

Host prediction of phages

To predict the bacterial hosts of the phages, CRISPR-Cas spacers were searched for targeting. Firstly, all the scaffolds with a minimum length of 1 kbp from all the corresponding samples with the bS21 phages were predicted for CRISPR repeat arrays using PILER-CR [38] with default parameters. The protein-coding genes within 10 kbp of both upstream and downstream of each repeat array were predicted and searched for Cas proteins using the TIGRFAM HMM database [35] with hmmsearch (version HMMER 3.3) [34]. For the scaffolds with both CRISPR repeat arrays and at least one cas protein, spacer sequences were extracted. Spacers were also extracted from the mapped reads and unplaced paired reads that may carry divergent spacers. The extracted spacers were searched against the manually curated phage sequences using blastn-short [39] with parameters as follows: -*e* value = 1e–3, -perc_identity = 70. The search results were parsed to retain those hits with a minimum match of 24 nt and no more than one mismatch [2], or 30 bp with no more than three mismatches. The phages whose scaffolds had matches to the spacers were considered as the likely bacterial hosts.

The temporal and spatial distribution of bS21 phages by read mapping

The quality reads of Rotsee Lake samples were mapped to all the curated bS21 phage genomes from this site using Bowtie2 [23]. A given bS21 phage was considered to be detected in a given sample if 90% of its genome was covered with at least one read (minimum nucleotide sequence similarity of 97%), and the coverage of this phage in the sample was accordingly determined as the total length of mapped reads dividing by the total covered genome length. A custom python script (dCov.py) was prepared to perform the analyses.

Gene annotation and metabolic prediction

The tRNAs encoded on all the curated phage genomes and retrieved IMG/VR sequences were predicted using tRNAscanSE (version 2.0.3) [40]. The predicted protein-coding genes were annotated by searching against the databases of Kyoto Encyclopedia of Genes and Genomes [41], UniRef100 [42], and UniProt [43] using Usearch (version v10.0.240_i86linux64) [44] with an *e* value threshold of 10^{-4}. For the specific metabolic potential of interest, the predicted protein-coding genes were also investigated using online HMM search tools.

Metatranscriptomic analyses

The seven raw metatranscriptomic RNA datasets from Lake Rotsee were downloaded [45] and filtered to remove sequencing contamination, adaptors and low-quality bases/reads as performed for metagenomic reads (see above). To profile the transcription of protein-coding genes, the quality RNA reads were mapped to the curated phage genomes reconstructed from Lake Rotsee using Bowtie2 [23] allowing three mismatches each read (i.e., 98% similarity). The normalized transcriptional level of a given protein-coding gene (*gene_a*) in a given sample was determined by calculating as follows: $\text{total\_base}_{gene\_a}/\text{length}_{gene\_a}/\text{total\_read}_{genome\_a}$, in which $\text{total\_base}_{gene\_a}$ means the total bases mapped to a given gene, length_{gene_a} means the length of the nucleotide sequence of the gene, and $\text{total\_read}_{genome\_a}$ means the total number of reads mapped to the corresponding genome. Only those protein-coding genes with at least 80% of the bases covered were calculated for normalized transcriptional level. To evaluate competitive mapping to regions of bacterial and phage genomes that are shared due to horizontal gene transfer, we also mapped RNA reads to datasets including both bacterial and phages and found no difference.

Genomic context of bS21 in published bS21-encoding phage genomes

To reveal if the bS21 genes in the published viral genomes are also co-located with these for core structural proteins, we checked all the huge phages genomes reported by Al-Shayeb et al. [2], and also the viral genomes reconstructed from the Global Ocean Virome (GOV) [11]. The protein-coding genes were searched against the bS21 HMM from TIGRFAM [35] using hmmsearch (version HMMER 3.3) [34] with the parameters of "--cut_tc". The results were parsed using cath-resolve-hits [46]. We identified bS21 genes in 68 huge phages [2] and 832 GOV viral genomes. The genomic context of the bS21 genes in these 900 genomes was performed as described above (see section "Genomic context analysis"). For the phage bS21 genes with genes for some specific proteins within ten genes, the taxonomy of their most similar bacterial bS21 was evaluated by comparison against the NCBI RefSeq bS21 proteins using BLASTp [39], the hits with the highest bit scores were retained for further analyses. When we checked the IMG/VR genomes for bS21 there were more than 14,000 bS21 hits. However, given the policy of IMG data use, we restricted our analyses to the published genomes.

DATA AVAILABILITY

The genomes of the bS21-encoding and outgroup phages are available at ggkbase https://ggkbase.berkeley.edu/PS21/organisms (please sign in by providing your email address to download) and at figshare (https://figshare.com/articles/dataset/bS21_encoding_phages/16744504).

REFERENCES

1. Mizuno CM, Guyomar C, Roux S, Lavigne R, Rodriguez-Valera F, Sullivan MB, et al. Numerous cultivated and uncultivated viruses encode ribosomal proteins. Nat Commun. 2019;10:752.
2. Al-Shayeb B, Sachdeva R, Chen LX, Ward F, Munk P. Clades of huge phages from across Earth's ecosystems. Nature. 2020;578:425–31.
3. Liu Y, Demina TA, Roux S, Aiewsakun P, Kazlauskas D, Simmonds P, et al. Diversity, taxonomy, and evolution of archaeal viruses of the class Caudoviricetes. PLoS Biol. 2021;19:e3001442.
4. Watson ZL, Ward FR, Méheust R, Ad O, Schepartz A, Banfield JF, et al. Structure of the bacterial ribosome at 2 Å resolution. Elife. 2020;9:e60482.
5. Shine J, Dalgarno L. Determinant of cistron specificity in bacterial ribosomes. Nature. 1975;254:34–38.
6. Held WA, Nomura M, Hershey JW. Ribosomal protein S21 is required for full activity in the initiation of protein synthesis. Mol Gen Genet. 1974;128:11–22.
7. Van Duin J, Wijnands R. The function of ribosomal protein S21 in protein synthesis. Eur J Biochem. 1981;118:615–9.
8. Dion MB, Oechslin F, Moineau S. Phage diversity, genomics and phylogeny. Nat Rev Microbiol. 2020;18:125–38.
9. Jha V, Roy B, Jahagirdar D, McNutt ZA, Shatoff EA, Boleratz BL, et al. Structural basis of sequestration of the anti-Shine-Dalgarno sequence in the Bacteroidetes ribosome. Nucleic Acids Res. 2021;49:547–67.
10. Ripp S, Miller RV. The role of pseudolysogeny in bacteriophage-host interactions in a natural freshwater environment. Microbiology. 1997;143:2065–70.
11. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature. 2016;537:689–93.

12. Krieg NR. Bacteroidia class. nov. Bergey's Manual of Systematics of Archaea and Bacteria. Chichester, UK: John Wiley & Sons, Ltd; 2015. p. 1.
13. Thomas F, Hehemann J-H, Rebuffet E, Czjzek M, Michel G. Environmental and gut bacteroidetes: the food connection. Front Microbiol. 2011;2:93.
14. Bernardet J-F. Flavobacteriia class. nov. Bergey's manual of systematics of archaea and bacteria. Chichester, UK: John Wiley & Sons, Ltd; 2015. p. 1.
15. Liu J, Xue C-X, Sun H, Zheng Y, Meng Z, Zhang X-H. Carbohydrate catabolic capability of a Flavobacteriia bacterium isolated from hadal water. Syst Appl Microbiol. 2019;42:263–74.
16. Aksyuk AA, Rossmann MG. Bacteriophage assembly. Viruses. 2011;3:172–203.
17. Howard-Varona C, Roux S, Dore H, Solonenko NE, Holmfeldt K, Markillie LM, et al. Regulation of infection efficiency in a globally abundant marine Bacteriodetes virus. ISME J. 2017;11:1942.
18. Yuan Y, Gao M. Jumbo bacteriophages: an overview. Front Microbiol. 2017;8:403.
19. Jaffe AL, Fuster M, Schoelmerich MC, Chen L-X, Colombet J, Billard H, et al. Long-term incubation of lake water enables genomic sampling of consortia involving Planctomycetes and Candidate Phyla Radiation bacteria. bioRxiv:2021.09.01.458585 [Preprint]. 2021.
20. Whaley-Martin KJ, Chen L-X, Nelson TC, Gordon J, Kantor R, Twible LE, et al. Acidity and sulfur oxidation intermediate concentrations controlled by O2-driven partitioning of sulfur oxidizing bacteria in a mine tailings impoundment. bioRxiv:2021.09.16.460096 [Preprint]. 2021.
21. Mayr MJ, Zimmermann M, Dey J, Wehrli B, Bürgmann H. Lake mixing regime selects apparent methane oxidation kinetics of the methanotroph assemblage. Biogeosciences. 2020;17:4247–59.
22. Chen L-X, Anantharaman K, Shaiber A, Eren AM, Banfield JF. Accurate and complete genomes from metagenomes. Genome Res. 2020;30:315–33.
23. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012;9:357–9.
24. Paez-Espino D, Roux S, Chen I-MA, Palaniappan K, Ratner A, Chu K, et al. IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. Nucleic Acids Res. 2019;47:D678–D686.
25. Linz AM, He S, Stevens SLR, Anantharaman K, Rohwer RR, Malmstrom RR, et al. Freshwater carbon and nutrient cycles revealed through reconstructed population genomes. PeerJ. 2018;6:e6075.
26. Tran P, Ramachandran A, Khawasik O, Beisner BE, Rautio M, Huot Y, et al. Microbial life under ice: metagenome diversity and in situ activity of Verrucomicrobia in seasonally ice-covered Lakes. Environ Microbiol. 2018;20:2568–84.
27. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
28. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.
29. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37:1530–4.
30. Suplatov D, Sharapova Y, Geraseva E, Švedas V. Zebra2: advanced and easy-to-use web-server for bioinformatic analysis of subfamily-specific and conserved positions in diverse protein superfamilies. Nucleic Acids Res. 2020;48:W65–W71.
31. Penev PI, Alvarez-Carreño C, Smith E, Petrov AS, Williams LD. TwinCons: conservation score for uncovering deep sequence similarity and divergence. PLoS Comput Biol. 2021;17:e1009541.
32. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
33. Grazziotin AL, Koonin EV, Kristensen DM. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res. 2017;45:D491–D498.
34. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7:e1002195.
35. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic Acids Res. 2003;31:371–3.
36. Jaffe AL, Castelle CJ, Matheus Carnevali PB, Gribaldo S, Banfield JF. The rise of diversity in metabolic platforms across the Candidate Phyla Radiation. BMC Biol. 2020;18:69.
37. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010;26:680–2.
38. Edgar RC. PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics. 2007;8:18.
39. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
40. Chan PP, Lowe TM. tRNAscan-SE: searching for tRNA genes in genomic sequences. Methods Mol Biol. 2019;1962:1–14.
41. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45:D353–D361.
42. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007;23:1282–8.
43. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45:D158–D169.
44. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26:2460–1.
45. Mayr MJ, Zimmermann M, Dey J, Wehrli B, Bürgmann H. Lake mixing regime selects apparent methane oxidation kinetics of the methanotroph assemblage. Biogeosciences. 2020;17:4247–59.
46. Lewis TE, Sillitoe I, Lees JG. cath-resolve-hits: a new tool that resolves domain matches suspiciously quickly. Bioinformatics. 2019;35:1766–7.

ACKNOWLEDGEMENTS

The study was supported by the Moore Foundation Grant 71785. We also thank the Chan Zuckerberg Biohub and the Innovative Genomics Institute at the University of California, Berkeley for funding support. We thank Anne-Catherine Lehours, Cindy Castelle, Corinne Bardot, Hermine Billard, Jonathan Colombet, and Fanny Perriere for assistance collecting and preparing the Lac Pavin samples. We acknowledge the researchers who generated the public data used in this study, especially Magdalena J. Mayr et al. who published the Lake Rotsee data.

AUTHOR CONTRIBUTIONS

The study was initiated by JFB. Manual genome curation was performed by L-XC and JFB. Genomic context and protein family analyses were conducted by ALJ, ALB, and L-XC. Comparative genomic analyses, microbial composition, host-phage relationship determination, phylogenetic analyses, and tRNA analyses were performed by L-XC. The distribution of phages in samples by read mapping was conducted by L-XC and ALJ. Metabolic potential and metatranscriptomic analyses were conducted by L-XC and ALB. The bS21-related analyses were performed by L-XC and PIP. TCN and LAW collected and prepared the Canada samples for sequencing. The manuscript was written by L-XC and had input from all authors.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43705-022-00111-w.

Correspondence and requests for materials should be addressed to Jillian F. Banfield.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.