Kavagutti et al. Microbiome (2019) 7:135
https://doi.org/10.1186/s40168-019-0752-0

Microbiome

RESEARCH Open Access

Phage-centric ecological interactions in aquatic ecosystems revealed through ultra-deep metagenomics

Vinicius S. Kavagutti[1], Adrian-Ştefan Andrei[1], Maliheh Mehrshad[1], Michaela M. Salcher[1,2] and Rohit Ghai[1*]

Abstract

The persistent inertia in the ability to culture environmentally abundant microbes from aquatic ecosystems represents an obstacle in disentangling the complex web of ecological interactions spun by a diverse assortment of participants (pro- and eukaryotes and their viruses). In aquatic microbial communities, the numerically most abundant actors, the viruses, remain the most elusive, and especially in freshwaters their identities and ecology remain unknown. Here, using ultra-deep metagenomic sequencing from pelagic freshwater habitats, we recovered complete genomes of > 2000 phages, including small "miniphages" and large "megaphages" infecting iconic freshwater prokaryotic lineages. For instance, abundant freshwater *Actinobacteria* support infection by a very broad size range of phages (13–200 Kb). We describe many phages encoding genes that likely afford protection to their host from reactive oxygen species (ROS) in the aquatic environment and in the oxidative burst in protist phagolysosomes (phage-mediated ROS defense). Spatiotemporal abundance analyses of phage genomes revealed evanescence as the primary dynamic in upper water layers, where they displayed short-lived existences. In contrast, persistence was characteristic for the deeper layers where many identical phage genomes were recovered repeatedly. Phage and host abundances corresponded closely, with distinct populations displaying preferential distributions in different seasons and depths, closely mimicking overall stratification and mixis.

Introduction

Freshwater planktonic communities are complex and dynamic, exhibiting distinct, recurrent patterns driven by both biotic and abiotic environmental factors [1]. However, in practice, the accurate resolution of recurrence of individual pelagic components is challenging, and from small to large (viruses, prokaryotes, eukaryotes) our discriminative ability to quantify each participant varies greatly. Freshwaters typically contain 10^2–10^4 unicellular eukaryotes and 10^5–10^7 prokaryotes per milliliter, but viruses are clearly the most abundant entities, with up to 10^6–10^8 viruses per milliliter [2]. Moreover, viruses are extraordinarily diverse, and complete genomic contexts are essential to understand the nature and dynamics of this diversity. The viral collective influences microbial community ecology by increasing carbon and phosphorous transfer to microbes [3–6], modulating individual lifestyles and evolutionary histories of microbial lineages [7, 8] and maintaining the diversity of the community at large [9]. Viruses in aquatic habitats are responsible for the mortality of nearly 20–40% prokaryotes every day [10], yet freshwater viruses remain largely understudied and untouched by advances in microbial culturing techniques and environmental genomics. Only a handful of isolate phage genomes are available from freshwater habitats [11–14], and only a few metagenomic studies are available [15–19]. However, the host-virus community responses to the establishment of the characteristic vertical zones in the water column of seasonally stratified water bodies, (a relatively warmer, light-exposed epilimnion and a deeper and colder hypolimnion) remain uncharacterized. Even more importantly, a representative collection of complete viral genomes from freshwater has so far remained out of reach.

* Correspondence: ghai.rohit@gmail.com
[1]Department of Aquatic Microbial Ecology, Institute of Hydrobiology, Biology Centre of the Academy of Sciences of the Czech Republic, Na Sádkách 7, 370 05 České Budějovice, Czech Republic
Full list of author information is available at the end of the article

Here, we exploit the potential of ultra-deep metagenomic time series sequencing to simultaneously recover phage (*Caudovirales*) and host genomic data from two common freshwater habitats (a drinking water reservoir and a humic pond). In doing so, we reconstructed 2034 complete genomes of phages infecting freshwater prokaryotes. These phage genomes are predicted to infect freshwater *Actinobacteria*, *Betaproteobacterales*, *Alphaproteobacteria*, *Bacteroidetes*, *Chloroflexi*, and other phyla for which no phages have been described before. Using the abundant freshwater *Actinobacteria* and their phages as models, we show that not only do phage genome abundances in deep water bodies mirror the abundance of their hosts, but they also reflect the classical patterns in thermal cycles of the water column, i.e., stratification and mixis. High abundances for both phages and hosts in the epilimnion are transitory, and persistence at lower abundances in the hypolimnion, the far larger niche, is the rule.

Results and discussion

Metagenomic sequencing, assembly, and complete phage genome recovery

We chose for our study two sites that serve as models for two distinct freshwater habitat types: meso-eutrophic Římov reservoir, a typical man-made, canyon-shaped reservoir, common to north temperate regions [20], and Jiřická pond, a shallow, humic mountain pond habitat found across the world [21]. The Římov reservoir is dimictic [22] and begins to mix at the onset of spring (March–April). It is stratified in summer when a distinct, warm epilimnion develops (May–October) above a colder hypolimnion. At the onset of winter, the colder waters sink, and the reservoir mixes again when water temperature throughout the water column drops to ca. 4 °C (Additional file 1: Figure S1). It is ice-covered during winter for at least 2 months (see the "Methods" for more site details).

We generated metagenomic time series from both sites, producing 18 metagenomes from Římov (both epi- and hypolimnion) and 5 from Jiřická (12.97 billion reads, ca. 1.9 Tb, Additional file 2: Table S1). While in most samples, we sequenced were ca. 54 Gb in size (ranging from 190 to 482 million reads, average 368 million reads), in two Římov samples (epi and hypolimnion), we sequenced ca. 380 Gb each (2.5 billion reads each). An overview of the microbial community using 16S rRNA abundances for both sites is shown in Additional file 1: Figure S2.

We also collected an additional 149 publicly available freshwater metagenomes (Additional file 2: Table S1, total of 4.04 billion reads, 1.09 Tb data) to search for complete phage genomes. All datasets were assembled independently (no co-assembly). In total, we analyzed ca. 3 Tb of metagenomic sequences from freshwater (ca. 17 billion reads).

The number of complete phage genomes recovered from any sample increased with sequencing depth, but with diminishing returns (Fig. 1a), with genome recovery maintaining linearity up to 100 Gb (ca. 1 phage genome for every additional gigabase) before tapering off at a maximum of 160 genomes from 400 Gb sequence data. While a total of 1677 genomes were assembled from the sequence data generated from the two study sites, (Římov and Jiřická), 357 genomes were recovered from all other available freshwater metagenomes. This suggests that the potential of ultra-deep sequencing to recover far more phage genomes has not yet been fully realized. We also recovered a number of metagenome-assembled genomes (MAGs) from the Římov metagenome time series dataset (see below). We denominate this entire collection of genomes as the uncultured freshwater organisms (UFO) dataset, where the UFOv subset refers to viruses and the UFOp subset to prokaryotic genomes.

Phage genome analyses

A total of 598 complete phage genomes were recovered from the Římov epilimnion (10 samples), 800 from the hypolimnion (8 samples), and 279 from Jiřická (5 samples). Upon dereplication (genomes with > 95% identity and > 95% coverage treated as one, see the "Methods" section), these numbers reduced by nearly threefold for the hypolimnion suggesting repeated capture of nearly identical genomes from multiple samplings. We found only a single instance of a phage that was nearly identical in two habitats (Římov and Jiřická).

The comparison of recovered freshwater phage genomes to representative sets of phages from Viral RefSeq (1996 genomes) and the marine habitat (1335 genomes) [23–25] is shown in Fig. 1c. Intriguingly, the genome size distributions of marine and freshwater phage genome sizes appear similar, except for a pronounced peak at small genome size (ca. 15 Kb) in the freshwater datasets. We recovered 155 "miniphage" genomes that were < 15 Kb in length (minimum length 13.5 Kb). This somewhat bimodal distribution is remarkably reminiscent of cell size distributions of prokaryotes themselves in freshwater [26] and while not conclusive in itself, this suggests that phage size distribution mirrors host cell size. That such peaks are not visible in the size distributions from isolate phages (Additional file 1: Figure S3) also points towards a more ecological explanation for the bimodal distribution in freshwater datasets. On the other hand, we also recovered 27 "megaphage" genomes (> 200 Kb in length, maximum length 446 Kb) that are similar in genome size to some recently described phages from the human gut microbiome [27].

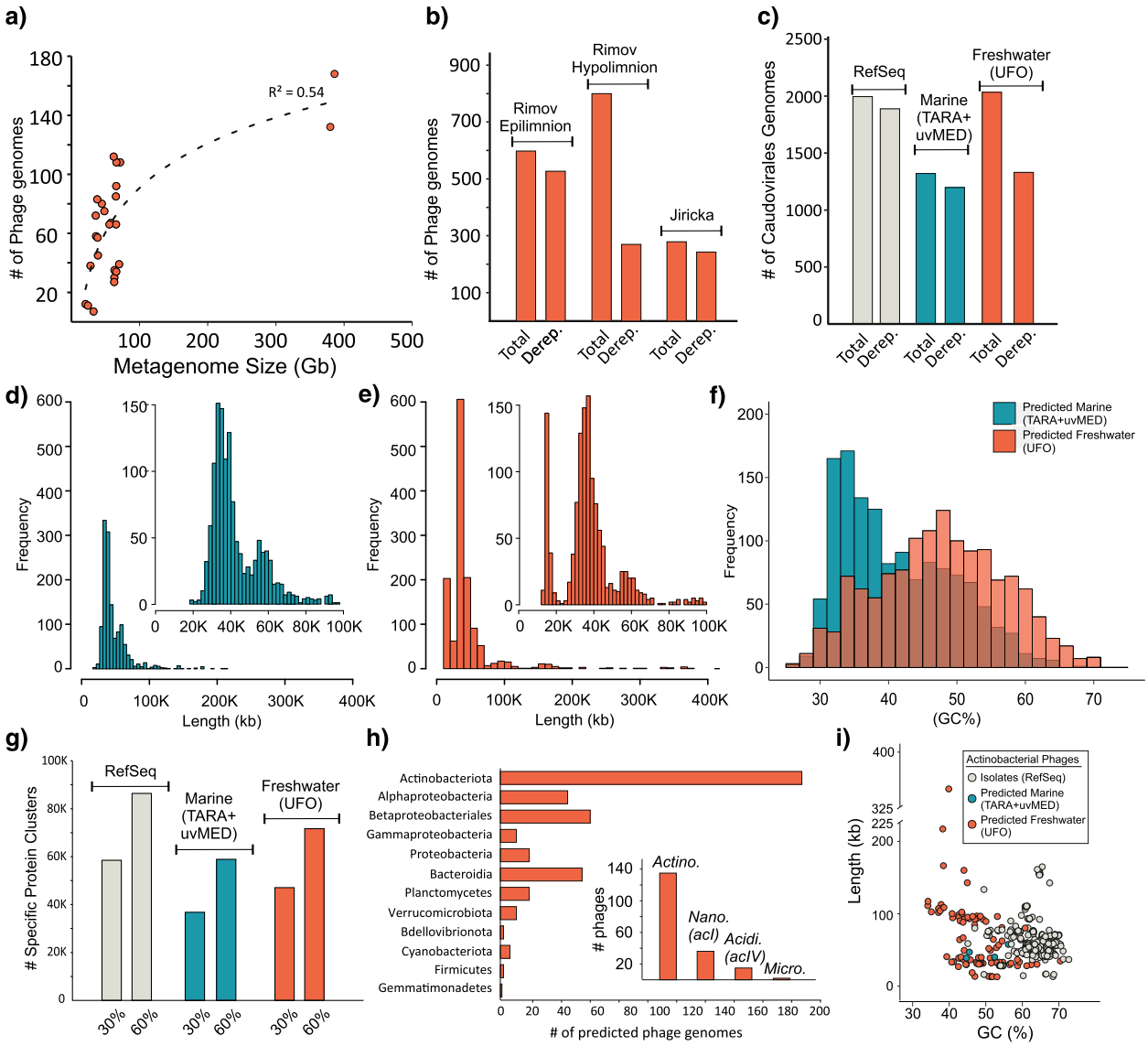

Fig. 1 Genome statistics. **a** Complete phage genome recovery as a function of metagenome size. **b** Total number of phage genomes vs dereplicated genomes from Římov reservoir (epi- and hypolimnion) and Jiřicka Pond. **c** Comparison of all phage genomes and dereplicated phage genomes from Viral Refseq, Marine habitat (TARA+uvMED) and the freshwater habitat (UFO). **d** Length distribution of marine phage genomes (inset, enlarged view of phages up to length 100 K). **e** Length distribution of recovered freshwater phage genomes (inset, enlarged view of phages up to length 100 K). **f** Comparative GC% distributions of marine and freshwater phage genomes. **g** Number of unique protein clusters in RefSeq, Marine, and Freshwater phage genomes at 30% and 60% identity. **h** Number of phage genomes with predicted host of the freshwater dataset (inset, the same for lower taxonomic ranks within *Actinobacteria*; Actino, Actinobacteriota; Nano, Nanopelagicales; Acidi, Acidimicrobiales; Micro, Microbacteriales). **i** GC% vs length for predicted (freshwater, marine) and known actinobacterial phages (RefSeq)

Common to both freshwater and marine habitats, two frequently recovered genome sizes appear to be ca. 40 Kb and 60 Kb, with 40 Kb being the most frequent. Unsurprisingly perhaps, genomic GC% of recovered phage genomes mirrors the GC% of the habitats (Fig. 1f and Additional file 1: Figure S4). As a measure of how many novel proteins are available in our freshwater phage dataset, we clustered all proteins in these datasets at two percentage identity levels (30% and 60%) [28]. The RefSeq dataset has the maximum number of unique protein clusters (not found in the others), followed closely by the UFO dataset (Fig. 1g). Additionally, the number of distinct Pfam domains detected in each dataset were 1761, 927, and 932 for RefSeq, marine, and freshwater datasets, respectively. These statistics suggest the UFO complete phage genome dataset adds significant novelty in phage sequence space. We also applied vContact2 (that uses Viral RefSeq phage genomes as references) to

assess the novelty of the recovered phages [29]. Of the 1330 freshwater and 1202 marine phage genomes (both dereplicated), 775 freshwater phages could not be assigned to any known RefSeq or marine phage cluster, thus remaining either unclassified or clustering only with freshwater phages. Nearly half of marine phages (n = 553) also did not cluster with any freshwater or RefSeq phage suggesting the existence of extremely divergent phage populations in these habitats. This is also seen in an all-vs-all comparison of all phage genomes, where the freshwater phages form large clusters that are only weakly related to other known groups (Additional file 1: Figure S5).

Host predictions and lifestyle strategies

Using multiple methods (host genes, similarity of tRNA integration sites, and presence of CRISPR spacers), we were able to predict hosts for 404 phage genomes (ca. 20%). The maximum number were predicted to be actinophages, largely owing to the presence of the characteristic *whiB* gene (sometimes even in multiple copies, similar to their hosts) that are taxonomically restricted to members of the actinobacterial phylum [19]. These actinophage genomes show an extremely broad size distribution, with multiple "miniphages" (n = 15, 6 dereplicated clusters) and "megaphages" (n = 3, 2 dereplicated clusters) (Fig. 1h). It appears that the most abundant microbes in the freshwater water column are infected by the full-size range of tailed phages ranging from as small as 14 Kb to as large as 347 Kb.

While for most actinophages we could not specifically pinpoint the host, it was possible for a few, and we predict at least 36 to infect the most abundant lineage in freshwaters ('Ca. Nanopelagicales', acI lineage [30]). In addition, others that possibly infect freshwater *Acidimicrobia* (acIV lineage [31]) and *Microbacteraceae* (Luna cluster [32]) were also found (Additional file 3: Table S2). In comparison to known actinobacterial phages (from cultured isolates), those from the metagenomes display a wide range of GC content and lengths (Fig. 1f), with a somewhat lower genomic GC% corresponding to the relatively lower GC% of freshwater *Actinobacteria* (esp. 'Ca. Nanopelagicales', 42% GC, Neuenschwander et al. 2018).

We also present phages predicted to infect abundant freshwater groups, e.g., *Betaproteobacterales*, *Alpha-* and *Gammaproteobacteria*, *Bacteroidetes*, *Planctomycetes*, *Cyanobacteria*, and *Gemmatimonadetes* (Fig. 1h, Additional file 1: Figure S5, Additional file 3: Table S2). We recovered only six recognizable cyanophages from our freshwater datasets. We attribute this to the high abundances of filamentous *Cyanobacteria* in freshwaters that are excluded by our filtration and phage recovery methodology. Amongst the other predicted hosts are several well-known freshwater genera (e.g., *Limnohabitans*, *Polynucleobacter*, *Fonsibacter*, *Flavobacterium*, *Novosphingobium*, *Sphingomonas*, Additional file 3: Table S2) with no described phages so far, except for 'Ca. Methylopumilus' [13].

Remarkably, 24 freshwater phages were found to encode the toxin ADP-ribosyltransferase. These are eukaryotic toxins related to VIP2 like toxins (a special class of AB toxins), that inhibit actin polymerization [33] and have been previously suggested to function as a "trojan horse", effecting targeted killing of eukaryotic predators that phagocytose phage-infected microbes [19, 34, 35] (Fig. 2). These are frequently encoded in phage genomes and have been found in both free-living phages or inserted prophages, e.g., Shiga toxin [36]. We also found evidence of the same toxin in six marine phage genomes and 116 phages from RefSeq infecting isolates of *Escherichia*, *Mycobacterium*, *Aeromonas*, etc. (Additional file 4: Table S3). Some of the freshwater phages encoding this toxin are predicted to infect *Actinobacteria*. Recently, phages have been shown to also encode ribosomal proteins that likely assist phage protein translation during infection [37]. We found 20 freshwater phages encoding at least one ribosomal protein (either S21 or L12) (Additional file 5: Table S4).

More than 10% of phages (i.e., 254 phage genomes) harbored genes involved in oxidative stress mitigation (Additional file 6: Table S5). Aquatic bacteria routinely experience the damaging effects of reactive oxygen species (ROS) (superoxide, hydrogen peroxide, and hydroxyl radicals) produced by their own metabolic machinery, released by other community members or generated by UV-induced photochemical reactions [38]. The widespread occurrence of multiple ROS defense mechanisms in bacteria designates oxidative stress as one of the main threats to their fitness and a major culprit in mortality [38–40]. Thus, based on the plethora of ROS defense mechanisms found in these freshwater phages, e.g., ferritin prevents ROS formation, superoxide dismutases and glutathione peroxidases inhibit ROS, thioredoxins and glutaredoxins repair oxidized amino acids particularly cysteine, and methionine and PAPS reductases boost reduced sulfur group assimilation [40, 41] (Additional file 6: Table S5), we consider that phages could provide their hosts the means to combat the harmful effects of oxidative stress. Such a strategy could be beneficial for phages, as it ensures the survival (during the lytic cycle) and proliferation of the hosts (during the lysogenic cycle) and protects their own proteins and DNA against oxidative damage. Given that nearly 10% of all recovered phages encode some ROS defense genes and the high rate of infections in the natural environments (estimated to be up to ca. 25% [42]), it also appears that this strategy is

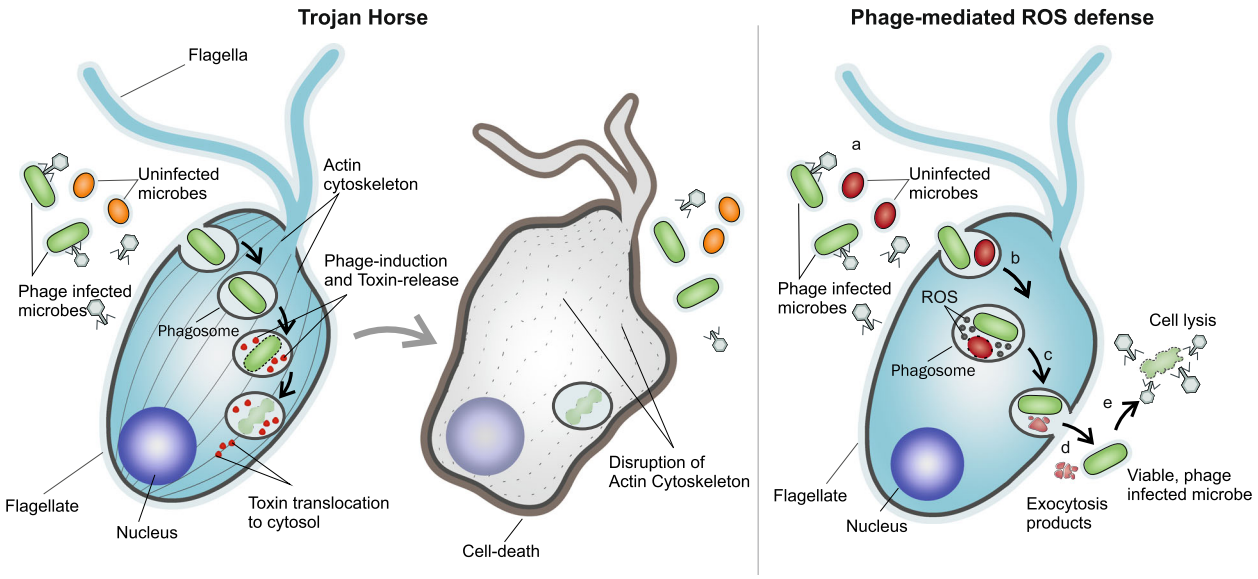

Fig. 2 Viral life strategies in freshwater environments. Left: The Trojan horse strategy. A phage encoding the eukaryotic toxin ADP-ribosyltransferase (VIP2 family) infects a microbe that is then ingested by a eukaryotic predator (a flagellate is shown). In the phagolysosome, the toxin is expressed and released after cell lysis into the phagolysosome from where it translocates to the cytosol. In the cytosol, the toxin inhibits actino polymerization leading to cell death. Right: Phage-mediated ROS defense. **a**, **b** A phage encoding ROS defense genes, e.g., thioredoxin, glutaredoxin is ingested by a flagellate. **c**, **d** The damage produced to the microbe by the ROS present in the phagolysosome will be reduced by the expression of phage-encoded genes, which will favor survival during the oxidative burst. **e** When the phage-infected microbe is released outside, cell lysis by the phage can proceed

commonly employed. However, unicellular eukaryotes (primarily flagellates) may consume up to 50% of bacteria in freshwater habitats on a daily basis [43]. This significant number, coupled with high phage infection frequencies suggests that multiple phage-infected bacteria must be ingested by these flagellates. Not all microbes are fully digested in the food vacuoles and many are expelled outside again [44, 45]. It is quite likely that ROS defense mechanisms improve the odds for surviving the phagolysosome where reactive oxygen species are discharged to destroy bacteria (oxidative burst). We postulate that a bacterium infected by a phage containing a ROS defense mechanism will have a selective advantage during phagocytic flagellate grazing. Thus, the phage-encoded proteins could help the host survive the high ROS environment that characterizes the phagolysosomes [46], i.e., a phage-mediated ROS defense (Fig. 2).

Phage abundance time series

Owing to the samples from multiple time points in the epi- and hypolimnion of Římov reservoir, we were able to recover several, nearly identical phage genomes repeatedly. The most extreme case was that of a predicted actinophage that was recovered 12 times, in distinct seasonal phases (spring bloom, summer, and winter). Remarkably, even though it was retrieved at different times of the year, only seven "variant" locations are seen (Additional file 1: Figure S6). Six of these are present in hypothetical genes and one in an intergenic region. Exhaustive sequence searches using jackhmmer [47] and HHpred server [48] revealed little clues to their functions. However, the retrieval of multiple, nearly identical phage genomes from multiple time points and strata suggests that some lineages are persistent, likely also owing to the constant presence of the host, in this case, *Actinobacteria* that are always abundant in the Římov reservoir [20] (Additional file 1: Figure S2). Similar to this phage, we also recovered multiple other examples that remain unchanged during our sampling efforts (Additional file 3: Table S2).

Remarkably, the abundance patterns of phages recovered from the Římov reservoir ($n = 1398$) (Fig. 3) suggest seasonality in their appearance. Distinct sets of phages peak in different layers during summer stratification (epi- or hypolimnion) or mixis (both spring and early winter). In the hypolimnion, some phages are persistent throughout the year while others appear only during stratification. Moreover, the abundances (coverage per gigabase) for each phage across the entire timeline of the Římov reservoir (18 samples) show that most phages in the epilimnion transiently achieve high abundances followed by near disappearance (a boom and bust scenario), while several in the hypolimnion appear to be more persistent and are recovered at multiple time points (Fig. 3). The shorter timeline of the Jiřická

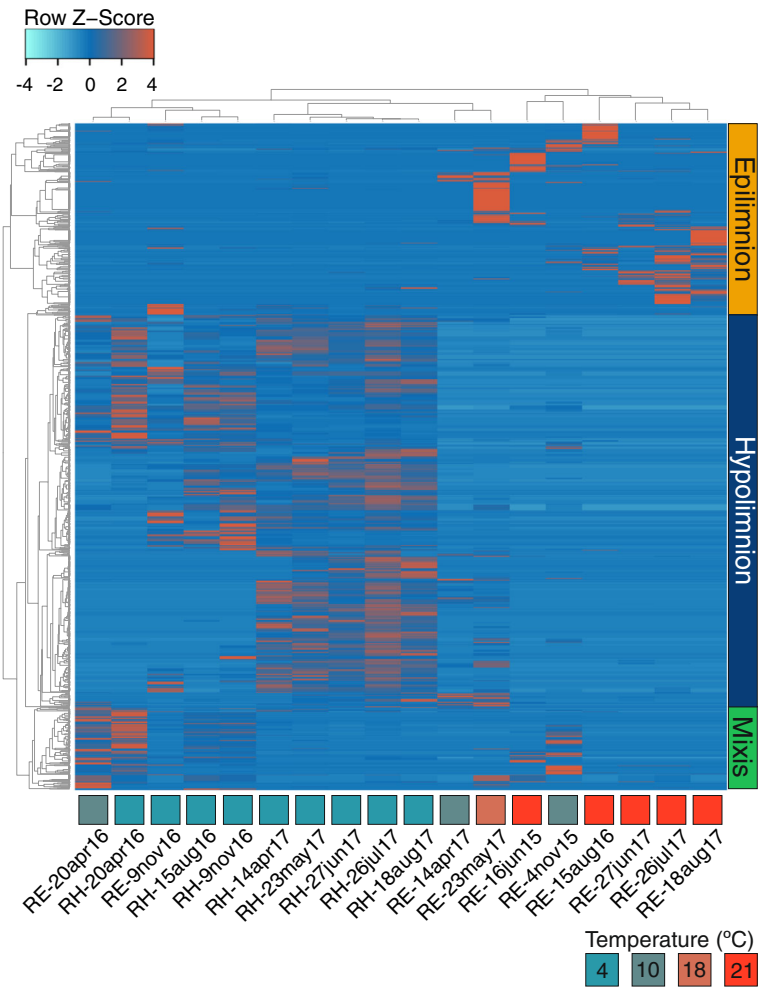

Fig. 3 Relative abundance of 1398 Římov phages in 18 metagenomes. A heatmap of abundances is shown (coverage/Gb of metagenome normalized by *Z*-score). Phages are clustered by sample and abundance (average linkage, Spearman rank correlation). Vertical bars on the right side indicate the classification of the clusters. Columns are annotated with the temperature, depth of the sample, and the sampling date (RE = 0.5 m and RH = 30 m). Temperature color key is shown at the top right

samples also shows a near continuous replacement of the abundant phages comparable to the Římov epilimnion (Additional file 1: Figure S7) in the relatively warm time period sampled here. However, as this time series is far shorter (lasting only a few summer months, Additional file 1: Figure S2), it remains to be seen if in such dynamic systems anything resembling persistence as observed in the colder hypolimnion of the Římov reservoir.

Spatiotemporal dynamics of actinophages and their hosts

As actinobacterial phages were the largest identifiable group (owing to the presence of the *whiB* gene), and that *Actinobacteria* are known to be dominant members of the community throughout the year (Additional file 1: Figure S2), we chose to focus subsequent analyses on both recovered actinophages and actinobacterial MAGs. Abundance profiles of all recovered actinophages encoding *whiB* (confident predictions, $n = 125$) are shown in Fig. 4. Phages that are nearly identical, i.e., persistent phages (> 95% identity and > 95% coverage) are shown as part of a cluster. The profiles within a cluster appear homogenous but different clusters show distinct preferences for either the epilimnion or the hypolimnion, suggesting that their hosts (*Actinobacteria*) would also show similarly distinct patterns.

We recovered 444 actinobacterial metagenome-assembled genomes (MAGs) from the Římov Reservoir datasets, whereof 305 MAGs fulfilled the criteria for further analyses (see the "Methods"). A phylogenomic analysis of the recovered actinobacterial MAGs in context of known isolate genomes is shown in Fig. 5. Most MAGs are placed within the three known groups of *Actinobacteria* frequently found in freshwater habitats, 'Ca. Nanopelagicales' ($n = 280$), *Acidimicrobiia* ($n = 114$), and *Microbacteriaceae* ($n = 37$)

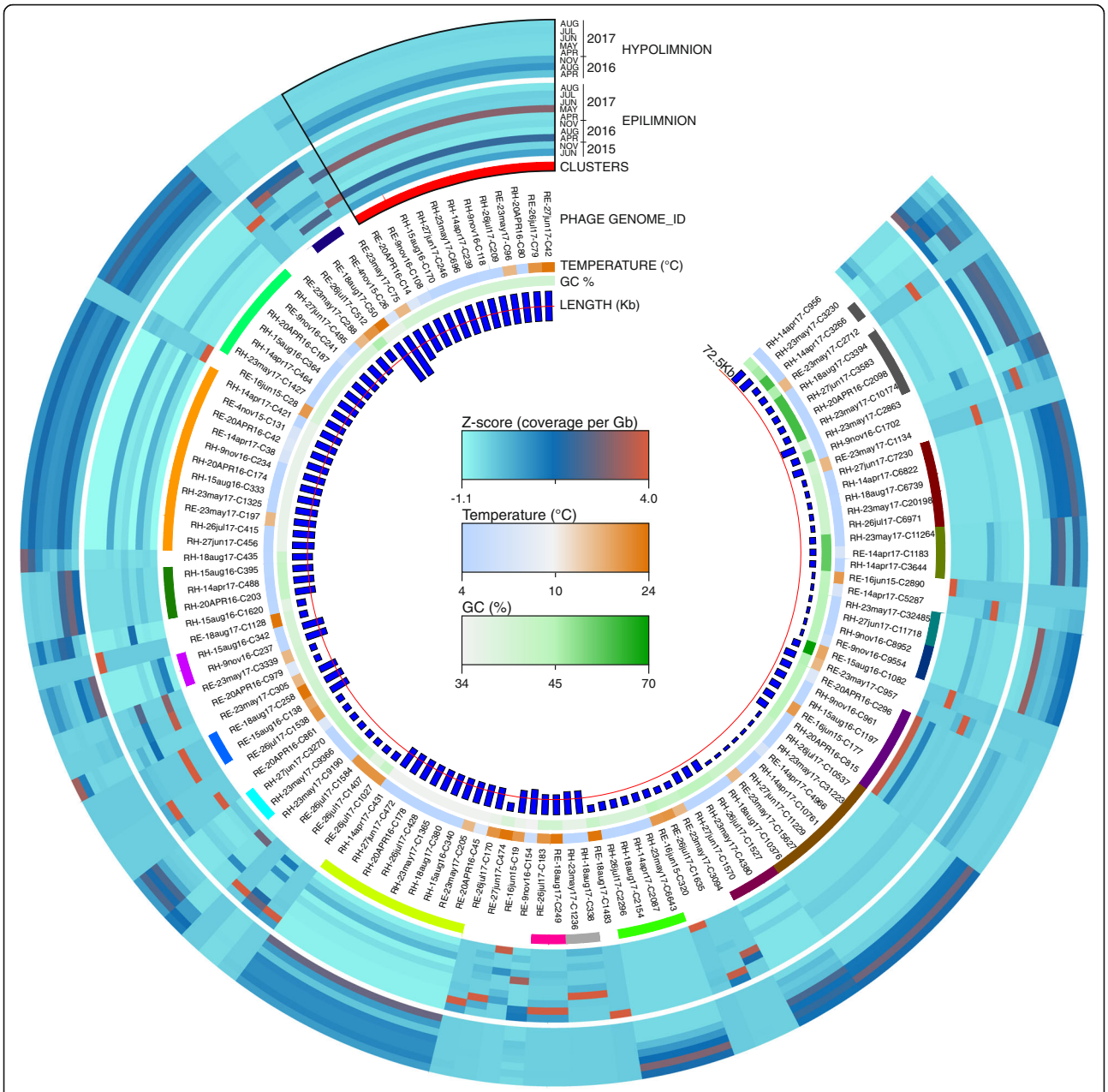

Fig. 4 Abundance profiles of the WhiB encoding actinophages recovered from Římov reservoir. Actinophages ($n = 125$) are arranged based on average amino acid identity (dendrogram not shown). From inside out the rings represent: phage genome size in kilobytes (red line indicates the average genome length, 72.5 Kb), phage genome GC%, water temperature of samples whereof phages were assembled, span of clusters of phage genomes grouped by similarity (nucleotide identity > 95%), abundance profiles of each actinophage in the epi- and hypolimnion time series datasets of Římov reservoir (coverage per gigabase of metagenome normalized by *Z*-score). Color keys are shown at the center. Phages of Cluster1 are outlined in black (top left of the circle)

(Additional file 7: Table S6). In particular, the order 'Ca. Nanopelagicales' are the most cosmopolitan microbes from freshwater [30, 49, 50] and have only recently been brought into culture [30, 51]. Isolates from *Microbacteriaceae* are also available [32, 52] but no cultured representatives exist yet for *Acidimicrobiia*, and these are described only from metagenome-assembled genomes [53]. Within the order 'Ca. Nanopelagicales', two genera are defined, 'Ca. Planktophila' (formerly acI-A) and 'Ca. Nanopelagicus' (formerly acI-B) [30, 54]. However, several other lineages have been described from 16S rRNA-based surveys, e.g., acI-C [54], acSTL [55], and acTH1 [56]. While recently a single genome from acI-C isolate has become available [51], no isolates or genomes have been described for either acSTL or acTH1. Based on

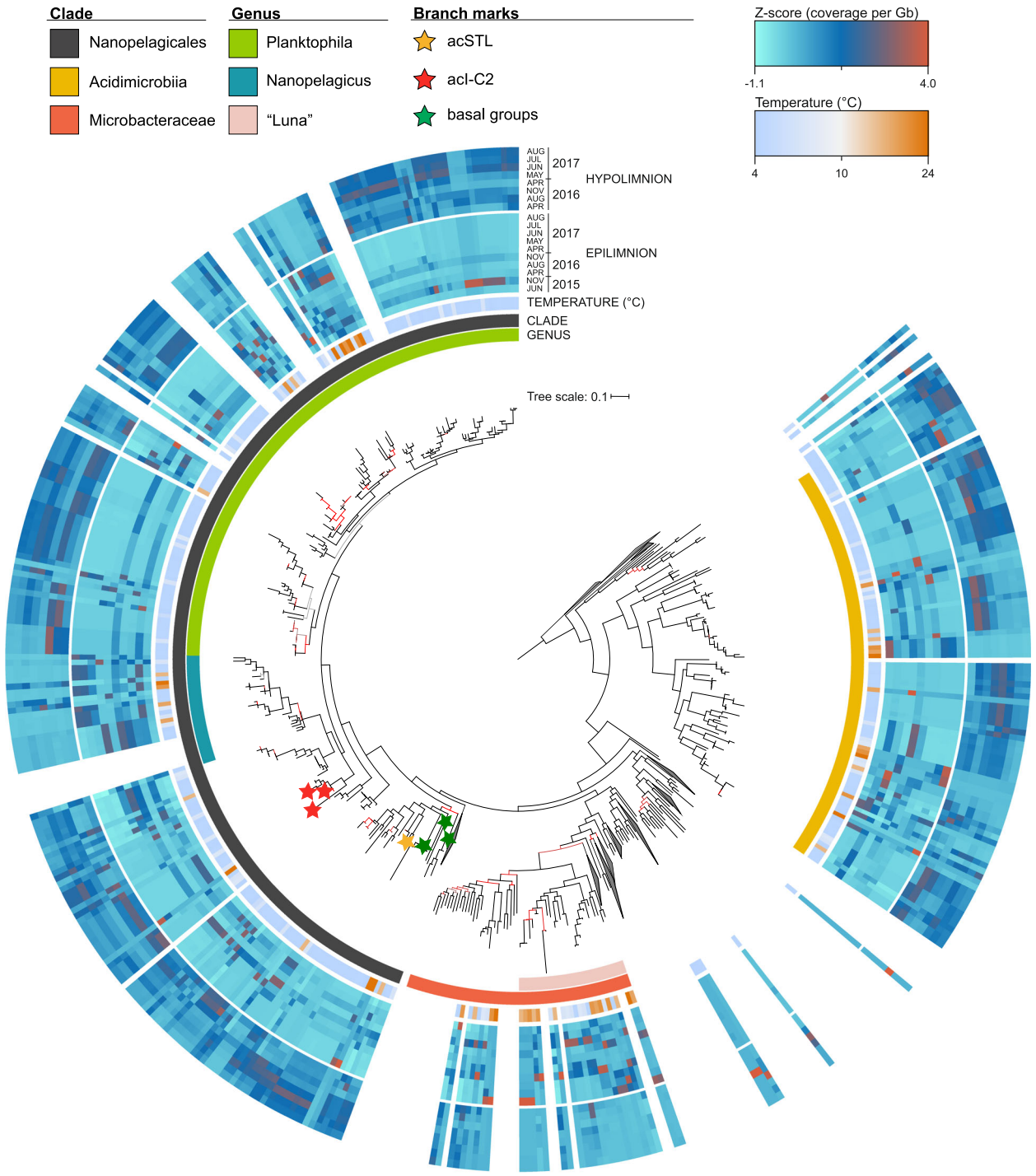

Fig. 5 Abundance profiles of actinobacterial genomes and MAGs combined with phylogenomic analysis. The tree was obtained using complete reference genomes ($n = 205$) and recovered MAGs ($n = 350$) of Římov reservoir time series metagenomes (see "Methods" section). From inside out the rings represent the following: the larger ring covering some tree branches highlights detailed taxonomic levels of *Microbacteraceae* and Nanopelagicales, the second ring shows class or order of *Actinobacteria*, followed by water temperature of samples whereof MAGs were assembled and abundance profiles for each MAG in the epi- and hypolimnion time series datasets of Římov reservoir (coverage per gigabase of metagenome normalized by *Z*-score). The abundance profiles are shown only for MAGs recovered from the Římov reservoir. The red and yellow stars indicate acSTL and acTH1 MAGs with 16S rRNA sequences, respectively. Green stars indicate deep-branching basal groups within the order Nanopelagicales. Branch colors reflect bootstrap support (UFboot): black, ≥ 95; red, 45–95; and gray, < 45. Color keys are shown at the top

the presence of 16S rRNA sequences, we identified MAGs that belong to acI-C2, acSTL, and acTH1 lineages. The acI-C2-related MAGs branch outside the acI-A and acI-B in accordance with known phylogeny. Both acSTL and acTH1 appear as a deep-branching sister group to other 'Ca. Nanopelagicales' but still belong to the same order (as classified by GTDB [57]). In addition, we also recovered 'Ca. Nanopelagicales' MAGs that are basal to all known groups (Fig. 5).

The abundance profiles of these *Actinobacteria* MAGs revealed that the vast majority are more abundant in the hypolimnion than the epilimnion, especially 'Ca. Nanopelagicales' and *Acidimicrobiia*, while the reverse is true for *Microbacteriaceae*, that are nearly always more abundant in the epilimnion (Fig. 5). MAGs that are in close phylogenetic proximity (for instance within the genus 'Ca. Nanopelagicus') do not necessarily show similar abundance patterns implying niche divergence even within closely related organisms (Neuenschwander et al. 2018) and several show only peaks in epilimnion or hypolimnion alone. Remarkably, the temporal abundances of both phages and their hosts mirror three distinct states of the reservoir, warm and stratified epilimnion, cold hypolimnion, and mixed water column. This is seen even more clearly in the predicted actinophages (Fig. 4) and reflected in the abundances of their hosts (Fig. 5).

The sporadic peaks of phages observed in the epilimnion reflect the transient niche that it truly is, in comparison to the apparently more stable environment of the hypolimnion (with little temperature variation). The ground state for Římov is a low-temperature regime, lasting for nearly two thirds of the year. Only at the end of the spring overturn until onset of winter a shallow, peripheral zone with higher temperature and light intensity (i.e., epilimnion) is established within which periodic blooms of photosynthetic organisms are observable [58]. Regardless of the sporadic peaks, we captured a complete genome of an actinophage in summer (at high abundance), and the same genome was recovered repeatedly at lower abundances, from both epi- and hypolimnion samples and at widely different temperatures (4 °C to 24 °C) at different times of the year (cluster 1 in Fig. 4). This suggests its host experiences conditions favorable for its increased abundance in the warmer epilimnion which leads to the higher abundance of its phage. The persistent recovery of such a phage also suggests that its host remains available throughout the year. In line with these observations, many actinobacterial MAGs show short-lived maxima in the epilimnion followed by lower abundances in the hypolimnion as observed earlier via fluorescence in situ hybridization with species to genus-specific probes [30]. It has also been shown recently that in the absence of the optimal host, phages may switch to sub-optimal hosts [59], further driving diversification and likely helping extend the longevity of phage lineages.

The major observable dynamic in the hypolimnion is the remarkably similar abundance patterns of the hosts and their phages, i.e., both show persistence from the onset of stratification till next spring. It appears that the hypolimnion maintains a large pool of highly related host genomes most of which are well-adapted to the long-lasting low-temperature regime of the reservoir. A fraction of these might find favorable niches in warmer temperatures, blooming, and then retreating within the hypolimnion at winter onset. However, even with the observable "persistent" abundances of phages, the hypolimnion is not without its perturbations (not in temperature but in other environmental variables, e.g., hypoxia, irregular nutrient input). Less obvious but sporadic peaks are observed in the hypolimnion as well (Fig. 3), which would suggest clonal expansions of the host, as has been reported for some *Planctomycetes* [60], suggesting similar dynamics are also played out in deeper waters.

Conclusions

Freshwater habitats are relatively accessible to monitoring and the use of time series metagenomes allows sensitive, genome-based surveys to capture both recurrent and anomalous changes in community composition. In this work, we used a deep-sequencing time series approach to viral ecology aimed towards the recovery of a representative freshwater phage genome collection that significantly expands the known viral sequence space and revealed viruses infecting many different freshwater phyla for which none were known before. This collection of complete phage genomes should serve as a critical reference in boosting environmental genomics of viruses in freshwater habitats at large. While this study was focused only upon phages, the UFO dataset is expected to shed light on many other viral groups as well, e.g., ssDNA viruses, phycodnaviruses, and virophages. With the availability of relatively inexpensive and increasingly higher throughput in sequencing technologies and the advent of even longer reads, freshwater viral ecology can transition from a gene-based to a genome-centric view of the viral world around us.

Methods

Sampling site and collection

Site I

Římov reservoir (Czech Republic, 48.846361 N 14.487639 E) is 2.06 km^2, with volume 34.5×10^6 m^3, length 13.5 km, circum-neutral pH, maximum depth 40 m, average depth

16.5 m, retention time ~ 100 days, dimictic, mesoeutrophic, moderately humic. Built in 1979, on the Malše River, it is part of the Czech Long Term Ecological Research network [22, 61, 62]. Eighteen water samples were collected from epilimnion (0.5 m, $n = 10$) and hypolimnion (30 m, $n = 8$) from June 2015 to August 2017. Two of these have been published previously [60, 63] and the rest were generated in this study. Vertical profiles of the physicochemical characteristics of the water column (temperature, pH, oxygen; GRYF XBQ4, Havlíčkův Broc, CZ) and chlorophyll *a* (FluoroProbe TS-16-12, bbe Moldaenke, Kiel, Germany) were also taken.

Site II

Jiřická pond (Czech Republic, 48.616034 N 14.676594 E) is 0.0356 km^2, with volume 6.59×10^3 m^3, pH 5.6–6.2, maximum depth 3.7 m, retention time ~ 5–7 days, dystrophic, located in the Novohradské mountains of Southern Bohemia [21]. Five samples were collected from the epilimnion (0.5 m) from May 2016 to August 2017.

Filtration and DNA extraction

All water samples (ca. 10 L each) from Římov reservoir and Jiřická pond were sequentially filtered through 20 μm, 5 μm, and 0.22 μm polycarbonate membrane filters (Sterlitech, USA). The 0.22 μm filters (containing the 5–0.22 μm microbial size fraction) were cut in small pieces (≅ 3–5 mm) using sterile scissors and processed for DNA extraction using the ZR Soil Microbe DNA MiniPrep kit (Zymo Research, Irvine, CA, USA), following the manufacturer's instructions.

Preprocessing of metagenomic datasets

Shotgun sequencing was performed using Illumina HiSeq4000 (for samples of 2015 and 2016—2 × 151 bp) (BGI HongKong, China) and Novaseq 6000 (for samples of 2017 and Jiřická 2016—2 × 151 bp) (Novogene, HongKong, China). Raw Illumina metagenomic reads were preprocessed in order to remove low-quality bases/reads and adaptor sequences using the bbmap package [64]. Briefly, the PE reads were interleaved by *reformat.sh* and quality trimmed by *bbduk.sh* (using a Phred quality score of 18). Subsequently, *bbduk.sh* was used for adapter trimming and identification/removal of possible PhiX and p-Fosil2 contamination. Additional checks (i.e., de novo adapter identification with *bbmerge.sh*) were performed in order to ensure that the datasets meet the quality threshold necessary for assembly. The preprocessed reads were assembled independently with MEGAHIT (v1.1.5) [65] using the k-mer sizes: 49, 69, 89, 109, 129, 149, and default settings.

Publicly available freshwater metagenomes (total 149 datasets) were downloaded and assembled as described above. Basic metadata (sampling date, location, depth, Bioproject identifiers, SRA accessions), and sequence statistics of all metagenomes generated or used in this study are provided in Additional file 2: Table S1.

16S rRNA abundance-based taxonomic classification

Twenty million reads were randomly sampled from each metagenome and compared to the SILVA database (version 132) [66] to identify candidate 16S reads using an *e*-value cutoff of 1e− 3 using MMSeqs2 [67]. The candidate reads were further screened with ssu-align [68] to find bona fide 16S rRNA sequences. The 16S rRNA sequences were compared to the SILVA database using blastn and taxonomy of the best hits was used to obtain the final taxonomic classification.

Gene prediction, phage detection, and annotation

Prodigal was used for gene prediction in metagenomic mode [69]. To retrieve complete phage genomes, we selected assembled contigs > 10 Kb that provided evidence of a circular genome as described before [19] ($n = 3576$ circles). These sequences were scanned with the VirSorter tool (default settings, Virome and RefSeq decontamination mode, scores 1 and 2) (https://de.iplantcollaborative.org/de/) [70] and MARVEL [71], and those contigs that were detected by either method to be of phage origin were retained. Finally, a set of 2034 genomes were judged to be complete. Previously described marine phage genomes were recovered from the Tara Oceans metavirome assemblies [24, 25] and the uvMED dataset [72] and rerun through VirSorter (using the same criteria as for freshwater phage genomes). Additional manual curation was performed using NCBI Batch CDD server to minimize errors in phage identification [73].

Recovery of actinobacterial genomes

The curated metagenomic datasets of Římov reservoir and Jiřická Pond were mapped using bbwrap.sh [74] (kfilter = 31, subfilter = 15, maxindel = 80) against the assembled contigs (longer than 3 Kb) in a lake-dependent fashion. The resulting BAM files (324 for Římov Reservoir, 25 for Jiřická Pond, respectively) were used to generate contig abundance files with *jgi_summarize_bam_contig_depths* [75] (--percentIdentity 97). The contigs and their abundance files were used for binning with MetaBAT2 [75] (default settings). Bin completeness, contamination, and strain heterogeneity were estimated using CheckM [76] (with default parameters). Bins with estimated completeness above 40% and contamination below 5% were denominated as metagenome-assembled genomes (MAGs). MAGs were taxonomically classified with GTDB-Tk [57] with default settings. MAGs belonging to

Actinobacteria (444), together with reference genomes (205) recovered from public repositories (Additional file 7: Table S6) were annotated using the TIGRFAMs database [77]. Thirty-five conserved marker proteins (Additional file 8: Table S7) were extracted from the annotated *Actinobacteria* genomes. MAGs that had more than 19 markers present and reference genomes were used for phylogenetic reconstruction. Briefly, homologous proteins were independently aligned with PRANK [78] (default settings), trimmed with BMGE [79] (-t AA -g 0.5 -b 3 -m BLOSUM30) and concatenated. A maximum-likelihood phylogeny was constructed using IQ-TREE [80] with the VT+F+R10 substitution model (chosen as the best-fitting model by ModelFinder [81]) and 1000 ultrafast bootstrap replicates [82]. These samples also allowed us to assemble and bin ca. 2400 microbial genomes in order to maximize chances for host prediction for the recovered phages.

Sequence annotation

All selected viral contigs were compared to the NCBI non-redundant protein database. Protein domains in coding sequences were annotated with Interproscan [83, 84]. Searches were performed locally using HMMER3 package [85] with *e*-value = 1e− 3 for Clusters of Orthologous Groups (COGS) [86] and trusted score cutoffs for TIGR Families—TIGRfams [77]. Pfam domains were identified in all datasets using the script pfam_scan.pl (ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools/) with the PFAM database release 31 [87].

Host prediction

Multiple methods were used to assign a host to a phage genome. Host-specific genes were used to predict the host, e.g., photosystem genes to link to cyanophages [88] and *whiB* for *Actinobacteria* [19]. Integration sites in the host genome (termed *attB*), usually a tRNA locus, were checked using BLASTN [89] for assigning a specific association between host and phage as described before [72]. CRISPR spacers in microbial genomes were detected using minced (https://github.com/ctSkennerton/minced) and for host prediction spacers were compared to phage genomes using BLASTN with stringent cutoffs (alignment length ≥ 30 bp, ≥ 97% nucleotide identity, ≥ 97% query coverage, ≤ 1e− 5). Additionally, a direct comparison of phage genomes to host genomes was made using BLASTN to identify shared nucleotide sequences (alignment length ≥ 30 bp, ≥ 97% nucleotide identity, ≥ 97% query coverage, ≤ 1e− 5) [90].

Selecting representative phage genomes

Caudovirales (tailed phages) phage genomes were divided into three sets, from NCBI Viral RefSeq (n = 1996), marine *Caudovirales* (Tara Oceans and uvMED, n = 1335), and freshwater *Caudovirales* (n = 2034). Within each set all-vs-all blastn comparisons were made retaining significant matches at *e*-value < 1e− 3 and > 95% nucleotide identity. For each comparison, two phages were considered as belonging to a cluster if the genome coverage of both phages in a pairwise comparison was ≥ 95%. Clusters of phages that meet these criteria were then merged together if they shared a phage genome in common (single linkage). The longest phage in each cluster was selected as a representative phage genome for this cluster. Clustering at these relatively high nucleotide identity levels and genome coverages is expected to retain phage genomes that are very closely related at the genomic level. At these cutoffs, the number of representatives in each dataset was RefSeq 1887, marine 1202, and freshwater 1330, making a total of 4419 representative phage genomes.

Phage proteomic tree

An all-vs-all tblastx comparison was performed for all 4419 phages (-M BLOSUM45 –e 1e-3) and the scores of all significant hits were added together to provide a comparison score for all pairwise comparisons. The comparison scores between two phage genomes were normalized by the self-comparison of both phages to provide a similarity metric (Dice coefficient). For example, 2 × comparison score of A and B/(self-comparison score of A + self-comparison score B). The Dice coefficient was subtracted from 1 to provide a distance measure [25, 72]. The resultant distance matrix of the all-vs-all comparison was used to generate 10,000 alternative, but equally valid tree topologies using clearcut [91] (clearcut –in distance_matrix.txt –out 10000.trees –n 10000), and a consensus tree was computed using IQ-TREE [80] (iqtree -t 10000.trees –con consensus.tre) that contains confidence values for each comparison. The final tree was visualized in iTOL [92] (http://itol.embl.de).

vContact2 classification

vContact2 was run on the Cyberverse infrastructure using diamond, freshwater phages only (1330 genomes, dereplicated), marine phages only (1202 genomes, dereplicated), and finally both freshwater and marine phages together (2532 genomes).

Fragment recruitment

All phage genomes were compared to all metagenomic datasets using RazerS 3 [93] (using cutoffs of > 95% identity and alignment lengths ≥ 50 bp) to compute coverage per gigabase. For microbial genomes, all rRNA sequences (5S, 16S, and 23S) were identified using rrna_hmm [94] and were masked prior to comparisons with

metagenomic sequences. At most, 20 million reads were used from metagenomic datasets for computing abundances of microbial genomes while a full set of reads were used for phage genomes. Raw data (coverage per gigabase) is given in Additional file 3: Table S2 (phages) and Additional file 7: Table S6 (*Actinobacteria*). A phage genome or microbial genome was considered to be present only when it presented > 80% genome coverage. Heatmaps were created using http://heatmapper.ca [95].

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s40168-019-0752-0.

Additional file 1: Figure S1. a) Water temperature along a depth profile from June 2015 to December 2017 in Římov reservoir. b) Oxygen concentrations along a depth profile from March 2015 to November 2017 in Římov reservoir. Red circles indicate the time and depth of the samples. Oxygen measurements were not available for all time points. **Figure S2.** 16S rRNA-based abundances of prokaryotic groups. **Figure S3.** Length distribution of Viral RefSeq phage genomes (inset: enlarged view of phages up to length 100 K). **Figure S4.** Comparative GC% distributions of freshwater metagenomic data for Římov epilimnion ($n = 10$ datasets), Římov hypolimnion ($n = 8$ datasets) and Jiřická ($n = 5$ datasets). **Figure S5.** Phage proteomic tree showing relationships of freshwater ($n = 1330$), marine ($n = 1202$) and RefSeq ($n = 1887$) phages used in this study. **Figure S6.** Genomic variations in a persistent actinophage (Cluster1). **Figure S7.** Relative abundance of 279 Jiřická phages in 5 metagenomes (coverage per Gb of metagenome normalized by Z-score).

Additional file 2: Table S1. Metagenomic datasets used in this study.

Additional file 3: Table S2. Freshwater phage genome information.

Additional file 4: Table S3. Phage Genomes encoding ADP-ribosyltransferase toxin (pfam domain PF03496).

Additional file 5: Table S4. Freshwater phage genomes encoding ribosomal proteins.

Additional file 6: Table S5. Oxidative Stress Related Pfam Domains.

Additional file 7: Table S6. Genome statistics and abundances of 444 actinobacterial bins.

Additional file 8: Table S7. TIGRFAMs markers used for phylogenomic analyses.

Abbreviations

MAGs: Metagenome-assembled genomes; ROS: Reactive oxygen species; UFO: Uncultured freshwater organisms

Acknowledgments

The authors thank Petr Znachor and Pavel Rychtecký for help with the sampling of Římov Reservoir and Petr Porcal for sampling Jiřická Pond.

Authors' contributions

RG designed the study. VSK, ASA, MM, MMS, and RG analyzed and interpreted the data. RG, VSK, and ASA wrote the manuscript. All authors commented on and approved the manuscript.

Funding

VSK was supported by the research grant 17-04828S (Grant Agency of the Czech Republic) and The Grant Agency of the Faculty of Science, University of South Bohemia, grant GAJU 158/2016/P. A-Ş. A was supported by the research grants 17-04828S (Grant Agency of the Czech Republic) and MSM200961801 (Academy of Sciences of the Czech Republic). MM was supported by the postdoctoral program PPPLZ (application number L200961651) provided by the Academy of Sciences of the Czech Republic. MMS was supported by research grant 19-23469S (Grant Agency of the Czech Republic). RG was supported by the research grants 17-04828S (Grant Agency of the Czech Republic) and CZ.02.1.01/0.0/0.0/16_025/0007417 (ERDF/ESF). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

Sequence data for all metagenomes generated in this work are archived at DDBJ/EMBL/GenBank and can be accessed under the Bioproject PRJNA429141 for the Římov reservoir and Bioproject PRJNA429145 for Jiřická pond. All phage genomes and the actinobacterial metagenome-assembled genomes are available in NCBI Bioproject PRJNA449258. The actinobacterial and phage genomic data could be accessed also on figshare: https://doi.org/10.6084/m9.figshare.9752813.v1 .

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Aquatic Microbial Ecology, Institute of Hydrobiology, Biology Centre of the Academy of Sciences of the Czech Republic, Na Sádkách 7, 370 05 České Budějovice, Czech Republic. [2]Limnological Station, Institute of Plant and Microbial Biology, University of Zurich, Seestrasse 187, 8802 Kilchberg, Switzerland.

Received: 21 June 2019 Accepted: 24 September 2019
Published online: 20 October 2019

References

1. Sommer U, Adrian R, De Senerpont DL, Elser JJ, Gaedke U, Ibelings B, et al. Beyond the Plankton Ecology Group (PEG) model: mechanisms driving plankton succession. Annu Rev Ecol Evol Syst Annual Reviews. 2012;43:429–48 Available from: http://www.annualreviews.org/doi/10.1146/annurev-ecolsys-110411-160251.
2. Wetzel RG. Freshwater ecosystems. Encycl Biodivers. Elsevier; 2001. p. 560–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780123847195000605
3. Suttle CA. The significance of viruses to mortality in aquatic microbial communities. Microb Ecol. 1994;28:237–43 Available from: http://link.springer.com/10.1007/BF00166813.
4. Fuhrman JA, Noble RT. Viruses and protists cause similar bacterial mortality in coastal seawater. Limnol Oceanogr. 1995;40:1236–42 Available from: http://doi.wiley.com/10.4319/lo.1995.40.7.1236.
5. Gobler CJ, Hutchins DA, Fisher NS, Cosper EM, Saňudo-Wilhelmy SA. Release and bioavailability of C, N, P Se, and Fe following viral lysis of a marine chrysophyte. Limnol Oceanogr. 1997;42:1492–504 Available from: https://www.infona.pl//resource/bwmeta1.element.elsevier-26a30eee-fd79-3046-bd11-335bab11446f.
6. Middelboe M, Jørgensen NOG. Viral lysis of bacteria: an important source of dissolved amino acids and cell wall compounds. J Mar Biol Assoc. 2006;86:605–12 Available from: www.sartorius.com.
7. Waldor MK, Mekalanos JJ. Lysogenic conversion by a filamentous phage encoding cholera toxin. Science. 1996;272:1910–4 Available from: http://www.ncbi.nlm.nih.gov/pubmed/8658163.
8. Zeidner G, Bielawski JP, Shmoish M, Scanlan DJ, Sabehi G, Béjà O. Potential photosynthesis gene recombination between Prochlorococcus and Synechococcus via viral intermediates. Environ Microbiol. 2005;7:1505–13.
9. Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pašić L, Thingstad TF, Rohwer F, et al. Explaining microbial population genomics through phage predation. Nat Rev Microbiol. 2009;7:828–36 Available from: http://www.nature.com/articles/nrmicro2235.
10. Suttle CA. Marine viruses — major players in the global ecosystem. Nat Rev Microbiol. 2007;5:801–12 Available from: http://www.ncbi.nlm.nih.gov/pubmed/17853907.
11. Yoshida T, Nagasaki K, Takashima Y, Shirai Y, Tomaru Y, Takao Y, et al. Ma-LMM01 infecting toxic *Microcystis aeruginosa* illuminates diverse

Cyanophage genome strategies. J Bacteriol. 2008;190:1762–72 Available from: http://jb.asm.org/cgi/doi/10.1128/JB.01534-07.
12. Chénard C, Wirth JF, Suttle CA. Viruses infecting a freshwater filamentous cyanobacterium (Nostoc sp.) encode a functional CRISPR array and a proteobacterial DNA polymerase B. MBio. 2016;7:e00667–16 Available from: http://mbio.asm.org/lookup/doi/10.1128/mBio.00667-16.
13. Moon K, Kang I, Kim S, Kim S-J, Cho J-C. Genome characteristics and environmental distribution of the first phage that infects the LD28 clade, a freshwater methylotrophic bacterial group. Environ Microbiol. 2017;19:4714–27 Available from: http://doi.wiley.com/10.1111/1462-2920.13936.
14. Moon K, Kang I, Kim S, Kim S-J, Cho J-C. Genomic and ecological study of two distinctive freshwater bacteriophages infecting a Comamonadaceae bacterium. Sci Rep. 2018;8:7989 Available from: http://www.nature.com/articles/s41598-018-26363-y.
15. Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, et al. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. PLoS One. 2012;7:e33641 Available from: https://dx.plos.org/10.1371/journal.pone.0033641.
16. Roux S, Chan L-K, Egan R, Malmstrom RR, McMahon KD, Sullivan MB. Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. Nat Commun. 2017;8:858 Available from: http://www.nature.com/articles/s41467-017-01086-2.
17. Skvortsov T, De Leeuwe C, Quinn JP, McGrath JW, Allen CCR, McElarney Y, et al. Metagenomic characterisation of the viral community of lough neagh, the largest freshwater lake in Ireland. PLoS One. 2016;11:e0150361.
18. Arkhipova K, Skvortsov T, Quinn JP, McGrath JW, Allen CCR, Dutilh BE, et al. Temporal dynamics of uncultured viruses: a new dimension in viral diversity. ISME J. 2018;12:199–211. https://doi.org/10.1038/ismej.2017.157.
19. Ghai R, Mehrshad M, Mizuno CM, Rodriguez-Valera F. Metagenomic recovery of phage genomes of uncultured freshwater actinobacteria. ISME J. 2017;11:304–8 Available from: http://www.nature.com/articles/ismej2016110.
20. Šimek K, Horňák K, Jezbera J, Nedoma J, Znachor P, Hejzlar J, et al. Spatio-temporal patterns of bacterioplankton production and community composition related to phytoplankton composition and protistan bacterivory in a dam reservoir. Aquat Microb Ecol. 2008;51:249–62.
21. Gabaldón C, Devetter M, Hejzlar J, Šimek K, Znachor P, Nedoma J, et al. Repeated flood disturbance enhances rotifer dominance and diversity in a zooplankton community of a small dammed mountain pond. J Limnol. 2016;76:292–304 Available from: http://www.jlimnol.it/index.php/jlimnol/article/view/jlimnol.2016.1544.
22. Znachor P, Hejzlar J, Vrba J, Nedoma J, Seda J, Simek K, et al. Brief history of long-term ecological research into aquatic ecosystems and their catchments in the Czech Republic: Part I : Manmade reservoirs. Institute of Hydrobiology, BC CAS, České Budějovice; 2016. Available from: http://www.gap2017.alga.cz/UserFiles/files/Manmade%20reservoirs.pdf.
23. Ghai R, Mizuno CM, Picazo A, Camacho A, Rodriguez-Valera F. Metagenomics uncovers a new group of low GC and ultra-small marine Actinobacteria. Sci Rep. 2013;3:2471 Available from: http://www.nature.com/articles/srep02471.
24. Brum JR, Sullivan MB, Ignacio-espinoza JC, Roux S, Doulcier G, Acinas SG, et al. Patterns and ecological drivers of ocean viral communities. Science. 2015;348:1261498 1–11. Available from: http://science.sciencemag.org/.
25. Nishimura Y, Watai H, Honda T, Mihara T, Omae K, Roux S, et al. Environmental viral genomes shed new light on virus-host interactions in the Ocean. mSphere. 2017:2 Available from: http://msphere.asm.org/lookup/doi/10.1128/mSphere.00359-16.
26. Pernthaler J, Sattler B, Šimek K, Schwarzenbacher A, Psenner R. Top-down effects on the size-biomass distribution of a freshwater bacterioplankton community. Aquat Microb Ecol. 1996;10:255-263. Available from: https://www.int-res.com/abstracts/ame/v10/n3/p255-263/.
27. Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, Tung J, et al. Megaphages infect Prevotella and variants are widespread in gut microbiomes. Nat Microbiol. 2019;4:693–700 Available from: http://www.nature.com/articles/s41564-018-0338-9.
28. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9 Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl158.
29. Bolduc B, Jang HB, Doulcier G, You ZQ, Roux S, Sullivan MB. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect archaea and bacteria. PeerJ. 2017;5:e3243.
30. Neuenschwander SM, Ghai R, Pernthaler J, Salcher MM. Microdiversification in genome-streamlined ubiquitous freshwater Actinobacteria. ISME J. 2018;12:185–98 Available from: http://www.nature.com/doifinder/10.1038/ismej.2017.156.
31. Warnecke F, Amann R, Pernthaler J. Actinobacterial 16S rRNA genes from freshwater habitats cluster in four distinct lineages. Environ Microbiol. 2004;6:242–53 Available from: http://doi.wiley.com/10.1111/j.1462-2920.2004.00561.x.
32. Hahn MW, Lunsdorf H, Wu Q, Schauer M, Hofle MG, Boenigk J, et al. Isolation of novel *Ultramicrobacteria* classified as *Actinobacteria* from five freshwater habitats in Europe and Asia. Appl Environ Microbiol. 2003;69:1442–51 Available from: http://aem.asm.org/cgi/doi/10.1128/AEM.69.3.1442-1451.2003.
33. Barth H, Aktories K, Popoff MR, Stiles BG. Binary bacterial toxins: biochemistry, biology, and applications of common *Clostridium* and *Bacillus* proteins. Microbiol Mol Biol Rev. 2004;68:373–402 Available from: http://mmbr.asm.org/cgi/doi/10.1128/MMBR.68.3.373-402.2004.
34. Lainhart W, Stolfa G, Koudelka GB. Shiga toxin as a bacterial defense against a eukaryotic predator, Tetrahymena thermophila. J Bacteriol. 2009;191:5116–22 Available from: http://jb.asm.org/cgi/doi/10.1128/JB.00508-09.
35. Arnold JW, Koudelka GB. The Trojan Horse of the microbiological arms race: phage-encoded toxins as a defence against eukaryotic predators. Environ Microbiol. 2014;16:454–66 Available from: http://doi.wiley.com/10.1111/1462-2920.12232.
36. Casas V, Miyake J, Balsley H, Roark J, Telles S, Leeds S, et al. Widespread occurrence of phage-encoded exotoxin genes in terrestrial and aquatic environments in Southern California. FEMS Microbiol Lett. 2006;261:141–9 Available from: https://academic.oup.com/femsle/article-lookup/doi/10.1111/j.1574-6968.2006.00345.x.
37. 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.
38. Imlay JA. Where in the world do bacteria experience oxidative stress? Environ Microbiol. 2019;21:521–30 Available from: http://doi.wiley.com/10.1111/1462-2920.14445.
39. Imlay JA. The molecular mechanisms and physiological consequences of oxidative stress: lessons from a model bacterium. Nat Rev Microbiol. 2013;11:443–54 Available from: http://www.nature.com/articles/nrmicro3032.
40. Ezraty B, Gennaris A, Barras F, Collet J-F. Oxidative stress, protein damage and repair in bacteria. Nat Rev Microbiol. 2017;15:385–96 Available from: http://www.nature.com/doifinder/10.1038/nrmicro.2017.26.
41. Mizuno CM, Rodriguez-Valera F, Garcia-Heredia I, Martin-Cuadrado AB, Ghai R. Reconstruction of novel cyanobacterial siphovirus genomes from mediterranean metagenomic fosmids. Appl Environ Microbiol. 2013;79:688–95.
42. Weinbauer MG, Höfle MG. Significance of viral lysis and flagellate grazing as factors controlling bacterioplankton production in a eutrophic lake. Appl Environ Microbiol. 1998;64:431-38.
43. Šimek K, Nedoma J, Znachor P, Kasalický V, Jezbera J, Hornňák K, et al. A finely tuned symphony of factors modulates the microbial food web of a freshwater reservoir in spring. Limnol Oceanogr. 2014;59:1477–92 Available from: http://doi.wiley.com/10.4319/lo.2014.59.5.1477.
44. Siegmund L, Burmester A, Fischer MS, Wöstemeyer J. A model for endosymbiosis: interaction between Tetrahymena pyriformis and Escherichia coli. Eur J Protistol. 2013;49:552–63 Available from: https://linkinghub.elsevier.com/retrieve/pii/S0932473913000369.
45. Gourabathini P, Brandl MT, Redding KS, Gunderson JH, Berk SG. Interactions between food-borne pathogens and protozoa isolated from lettuce and spinach. Appl Environ Microbiol. 2008;74:2518–25 Available from: http://aem.asm.org/cgi/doi/10.1128/AEM.02709-07.
46. Rehfuss MYM, Parker CT, Brandl MT. Salmonella transcriptional signature in Tetrahymena phagosomes and role of acid tolerance in passage through the protist. ISME J. 2011;5:262–73 Available from: http://www.nature.com/articles/ismej2010128.
47. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. Nucleic Acids Res. 2018;46:W200–4.
48. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, et al. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its Core. J Mol Biol. 2018;430:2237–43 Available from: https://linkinghub.elsevier.com/retrieve/pii/S0022283617305879.
49. Glockner FO, Zaichikov E, Belkova N, Denissova L, Pernthaler J, Pernthaler A, et al. Comparative 16S rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of Actinobacteria. Appl Environ Microbiol. 2000;66:5053–65.

50. Ghai R, McMahon KD, Rodriguez-Valera F. Breaking a paradigm: cosmopolitan and abundant freshwater actinobacteria are low GC. Environ Microbiol Rep. 2012;4:29–35 Available from: http://doi.wiley.com/10.1111/j.1758-2229.2011.00274.x.
51. Kang I, Kim S, Islam MR, Cho J-C. The first complete genome sequences of the acI lineage, the most abundant freshwater Actinobacteria, obtained by whole-genome-amplification of dilution-to-extinction cultures. Sci Rep. 2017;7:46830 Available from: http://www.nature.com/articles/srep46830.
52. Hahn MW, Schmidt J, Taipale SJ, Doolittle WF, Koll U. Rhodoluna lacicola gen. nov., sp. nov., a planktonic freshwater bacterium with stream-lined genome. Int J Syst Evol Microbiol. 2014;64:3254–63 Available from: http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.065292-0.
53. Ghai R, Mizuno CM, Picazo A, Camacho A, Rodriguez-Valera F. Key roles for freshwater Actinobacteria revealed by deep metagenomic sequencing. Mol Ecol. 2014;23:6073–90 Available from: http://doi.wiley.com/10.1111/mec.12985.
54. Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. A guide to the natural history of freshwater lake bacteria. Microbiol Mol Biol Rev. 2011;75:14–49 Available from: http://mmbr.asm.org/cgi/doi/10.1128/MMBR.00028-10.
55. Allgaier M, Grossart H-P. Diversity and seasonal dynamics of Actinobacteria populations in four lakes in northeastern Germany. Appl Environ Microbiol. 2006;72:3489–97 Available from: http://aem.asm.org/cgi/doi/10.1128/AEM.72.5.3489-3497.2006.
56. Wu QL, Zwart G, Wu J, Kamst-van Agterveld MP, Liu S, Hahn MW. Submersed macrophytes play a key role in structuring bacterioplankton community composition in the large, shallow, subtropical Taihu Lake. China Environ Microbiol. 2007;9:2765–74 Available from: http://doi.wiley.com/10.1111/j.1462-2920.2007.01388.x.
57. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36:996–1004 Available from: http://www.ncbi.nlm.nih.gov/pubmed/30148503.
58. Znachor P, Visocká V, Nedoma J, Rychtecký P. Spatial heterogeneity of diatom silicification and growth in a eutrophic reservoir. Freshw Biol. 2013;58:1889–902 Available from: http://doi.wiley.com/10.1111/fwb.12178.
59. Enav H, Kirzner S, Lindell D, Mandel-Gutfreund Y, Béjà O. Adaptation to sub-optimal hosts is a driver of viral diversification in the ocean. Nat Commun. 2018;9:4698 Available from: http://www.nature.com/articles/s41467-018-07164-3.
60. Andrei AŞ, Salcher MM, Mehrshad M, Rychtecký P, Znachor P, Ghai R. Niche-directed evolution modulates genome architecture in freshwater Planctomycetes. ISME J. 2019;13:1056.
61. Simek K, Pernthaler J, Weinbauer MG, Hornak K, Dolan JR, Nedoma J, et al. Changes in bacterial community composition and dynamics and viral mortality rates associated with enhanced flagellate grazing in a mesoeutrophic reservoir. Appl Environ Microbiol. 2001;67:2723–33 Available from: http://aem.asm.org/cgi/doi/10.1128/AEM.67.6.2723-2733.2001.
62. Šimek K, Weinbauer MG, Hornák K, Jezbera J, Nedoma J, Dolan JR. Grazer and virus-induced mortality of bacterioplankton accelerates development of Flectobacillus populations in a freshwater community. Environ Microbiol. 2007;9:789–800 Available from: http://doi.wiley.com/10.1111/j.1462-2920.2006.01201.x.
63. Mehrshad M, Salcher MM, Okazaki Y, Nakano S, Šimek K, Andrei A-S, et al. Hidden in plain sight—highly abundant and diverse planktonic freshwater *Chloroflexi*. Microbiome. 2018;6:176.
64. Bushnell B. BBMap (version 35.14) [Software]. Available at https://sourceforge.net/projects/bbmap/. 2015.
65. Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, et al. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods. 2016;102:3–11 Available from: https://linkinghub.elsevier.com/retrieve/pii/S1046202315301183.
66. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41:590–6.
67. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017; Available from: http://www.nature.com/doifinder/10.1038/nbt.3988.
68. Nawrocki EP, Eddy SR. ssu-align: a tool for structural alignment of SSU rRNA sequences; 2010.
69. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
70. Bolduc B, Youens-Clark K, Roux S, Hurwitz BL, Sullivan MB. IVirus: facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. ISME J. 2017;11:7–14. https://doi.org/10.1038/ismej.2016.89.
71. Amgarten D, Braga LPP, da Silva AM, Setubal JC. MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. Front Genet. 2018;9 Available from: https://www.frontiersin.org/article/10.3389/fgene.2018.00304/full.
72. Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. Expanding the marine virosphere using metagenomics. PLoS Genet. 2013;9:e1003987 Available from: http://dx.plos.org/10.1371/journal.pgen.1003987.
73. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 2017;45:D200-D203.
74. Bushnell B. BBMap short-read aligner, and other bioinformatics tools. Bioinformatics. 2016. Available from: https://sourceforge.net/projects/bbmap.
75. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015;3:e1165.
76. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55 Available from: http://www.ncbi.nlm.nih.gov/pubmed/25977477.
77. Haft DH. TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res. 2001;29:41–3 Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/29.1.41.
78. Löytynoja A. Phylogeny-aware alignment with PRANK. Methods Mol Biol. 2014;1079:155-170.
79. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol. 2010;10:210 Available from: http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-10-210.
80. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268-74.
81. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587-89.
82. Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol. 2018;35:518-522.
83. Bateman A. The Pfam protein families database. Nucleic Acids Res. 2004;32:138D–141 Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=308855%7B&%7Dtool=pmcentrez%7B&%7Drendertype=abstract.
84. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;9:1236-40.
85. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7:e1002195 Available from: https://dx.plos.org/10.1371/journal.pcbi.1002195.
86. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003;4:41 Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-4-41.
87. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42:D222-D230.
88. Mann NH, Cook A, Millard A, Bailey S, Clokie M. Bacterial photosynthesis genes in a virus. Nature. 2003;424:741.
89. Altschul S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402 Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/25.17.3389.
90. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict bacteriophage–host relationships. FEMS Microbiol Rev. 2016;40:258–72 Available from: https://academic.oup.com/femsre/article-lookup/doi/10.1093/femsre/fuv048.
91. Evans J, Sheneman L, Foster J. Relaxed neighbor joining: a fast distance-based phylogenetic tree construction method. J Mol Evol. 2006;62:785-92.
92. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019; Available from: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz239/5424068.
93. Weese D, Holtgrewe M, Reinert K. RazerS 3: faster, fully sensitive read mapping. Bioinformatics. 2012;28:2592–9 Available from: https://

academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts505.
94. Huang Y, Gilna P, Li W. Identification of ribosomal RNA genes in metagenomic fragments. Bioinformatics. 2009;25:1338–40 Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp161.
95. Babicki S, Arndt D, Marcu A, Liang Y, Grant JR, Maciejewski A, et al. Heatmapper: web-enabled heat mapping for all. Nucleic Acids Res. 2016;44:W147–53.

Publisher's Note