Tracking Double-Stranded DNA Bacteriophages and Their Hosts in a Deep Freshwater Lake by Integrating Metagenomics and The Hi-C Technique

Ye Tao

Nanjing Institute of Geography and Limnology Chinese Academy of Sciences

Peixin Gao

Nanjing Institute of Geography and Limnology Chinese Academy of Sciences

Biao Li

Nanjing Institute of Geography and Limnology Chinese Academy of Sciences

Peng Xing (✉ pxing@niglas.ac.cn)

Nanjing Institute of Geography and Limnology Chinese Academy of Sciences <https://orcid.org/0000-0002-4917-5181>

Qinglong L. Wu

Nanjing Institute of Geography and Limnology Chinese Academy of Sciences

Research

Keywords: Freshwater metagenome, Bacteriophage, Hi-C, Bacteriophage-host association

Posted Date: December 17th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-129104/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: In aquatic ecosystems, bacteriophages play key roles in species diversity, host population dynamics and functional gene transfer. Due to a lack of high-throughput experimental tools to obtain robust associations between bacteriophages and hosts, the current metagenome-based phage community research generally predicts interactions based on various computational pipelines. In this study, by introducing an *in vivo* proximity-ligation chromosomal confirmation capture (Hi-C) technique combined with existing sequencing data from a freshwater metagenome study, we investigated double-stranded DNA (dsDNA) bacteriophages distribution, bacteriophage-host associations and abundance profiles at various depths in a deep alpine lake.

Results: A total of 985 nonredundant viral genomes (containing 1,464 viral contigs measuring at least 10 kb) were identified using rigorous bioinformatic pipelines, and 67.3% of bacteriophages had never been reported previously. The bacteriophages had a significant depth-dependent distribution, and 56.9% of phage populations were only present in the epilimnion. Overall, 567 auxiliary metabolic genes (AMGs) were identified in 239 phages (24.26% of the total population). Hi-C reconstructed 58 bacteriophage-host association events encompassing eight bacterial phyla, among which 14 AMG-coding bacteriophages were clearly linked to their hosts. Three bacteriophages contained the *mec* gene, which was exactly complementary to their hosts in the sulfur-related pathway, and three other bacteriophages contained AMGs that likely participated in host fructose and mannose metabolism.

Conclusions: The present study identifies bacteriophages in Lake Fuxian and describes the bacteriophages and their host dynamics along the depth profile. Our results indicate that metagenomics combined with the Hi-C technique has promise for future applications in high-throughput bacteriophage identification and bacteriophage-host association detection, which could further extend our knowledge of the functional roles played by bacteriophages in aquatic ecosystems.

Background

Bacteriophages (a group of viruses that infect bacteria and archaea) representing the largest constituents of the microbiome community are ubiquitous in various environments and are likely to play highly important roles in preying hosts, mediating horizontal gene transfer and altering host metabolism and the cycling of geochemical elements [1-3]. However, traditional culture-based methods are not suitable for recovering all information regarding these bacteriophages because of the strict constraints of viral cultivation. In addition, unlike prokaryotes and eukaryotes, bacteriophages do not have conserved shared genes across multiple families taxonomically, which means that amplicon-based high-throughput survey technology is unsuited for bacteriophage communities. Thus, viral abundance and ecological importance are largely underestimated [4]. However, the rapid development of viral metagenomics and bioinformatics has provided an opportunity to overcome these obstacles [5, 6]. In recent years, large amounts of viral genomes, most of which are novel bacteriophages, have been identified from a variety of environments, such as marine environments [7, 8], freshwater environments [9], hot spring mats [10], rhizosphere soils

[11], and mangroves [12]. The findings of these studies indicated that bacteriophages interacting with their hosts (bacteria and archaea), acted as key partners in multiple ecosystems.

Bacteriophages can affect the diversity and function of aquatic microbial populations through the incorporation and expression of auxiliary metabolic genes (AMGs). Currently, phage-encoded AMGs are reported to include multiple types of genes involved in not only main nutrient cycle pathways, such as carbon metabolism [12-14], nitrogen [15], phosphorus [16] and sulfur cycling [17, 18], but also some specific pathways, such as acetate fermentation [19], methane oxidation [14], nucleotide metabolism [20] and oxidative stress responses [21]. Taken together, these findings imply that viral AMGs influence most, if not all, pathways of microbial metabolism. For phages, carrying and expressing their own AMGs related to energy metabolism provides advantages by generating ATP and/or reducing power for the synthesis of viral proteins, which places the largest energy expense on relatively small bacteriophages, including cyanophages [22, 23]. For hosts, expanding their ecological communities based on receiving AMGs through horizontal gene transfer (HGT) can be mediated by AMG-encoding bacteriophages [24, 25]. It follows that the importance of bacteriophages to nutrient cycling is not limited to killing their hosts but also includes altering the central pathways of host metabolism. Nevertheless, a comprehensive map of the AMG content of aquatic viral communities and how viral infection transforms host metabolism, partly through phage-encoded metabolic genes, has not been presented to date [26].

Currently, it remains a major challenge to associate bacteriophages and their hosts in viromics studies, although sequencing-based metagenomic technology has dramatically expanded our knowledge regarding the environmental virome, especially viral genome contents and micro- and macrodiversity [7, 27]. Common bioinformatic methods for predicting bacteriophage-host associations are categorized into [28] (i) clustered regularly interspaced short palindromic repeat (CRISPR) spacer mining, wherein the spacer sequences should be nearly identical between bacteriophages and their hosts at the nuclear acid level [29, 30], (ii) sequence composition similarity (such as tRNA, codon usage and restriction-modification systems) [27], (iii) homology to other known bacteriophages [31, 32], (iv) abundance correlations between bacteriophages and their hosts [26], (v) host-specific genes located in the viral genome, such as *whiB* for Actinobacteria [25] and photosystem D1/D2 proteins for Cyanobacteria [33]. However, the rate of correct bacteriophage-host signals at the host species level was determined to be notably low (< 5%) by the above computational approaches [28]. Clearly, bacteriophage-host associations predicted only *in silico* were limited in further analysis; moreover, traditional experimental approaches, such as plaque assays, liquid assays and viral tagging methods, did not meet the needs of high-throughput detection of bacteriophage-host associations [28]. Recently, a new experimental method called AdsorpSeq (adsorption sequencing) was developed to measure bacteriophages that are preferentially adsorbed to specific host cell envelopes and to detect unknown bacteriophage-host associations [34]. However, this novel technology was limited in the types of host cell envelopes and bacteriophage isolations that could be analyzed. Single-cell genome sequencing technology could represent direct experimental evidence of bacteriophage-host links and was used in previous aquatic virome studies [10, 35, 36]. However, with this method, bacteriophages in active replication might not be detected effectively, and single amplified host genomes could not be reconstructed on a large scale.

High-throughput chromosomal confirmation capture (Hi-C), which was developed ten years ago, is a fixation-based proximity-ligation method for estimating the probability of close physical proximity between DNA fragments in the same cell [37]. In the early stage, this technique was employed to study DNA folding status (3D map for DNA) and chromosome regulatory landscapes [38, 39]. During the past decade, Hi-C has been widely employed in contig orientation and scaffolding to construct chromosome-level genomes for many eukaryotes [40-42]. More broadly, this technique was able to reconstruct strain- and species-level microbial genomes (metagenome-assembled genomes, MAGs) from meta-communities [43]. The notable advantage of Hi-C-based MAG construction was the determination of the source of plasmid contigs that were always overlooked in traditional pipelines [44, 45]. Bacteriophages need to enter host cells to complete their life histories, and it is possible to theoretically track all bacteriophage-host interactions simultaneously by metagenomic Hi-C [46, 47]. Hi-C facilitated the association of extrachromosomal DNA (phages and plasmids) with microbial host genomes in the human gut, mouse gut and synthetic bacterial communities but not in environmental field samples [48-51].

Freshwater ecosystems are small components of the world's surface ecosystems, but these areas are very important to humans for drinking water supply, and they matter for global carbon cycling [52]. There have been several freshwater virome studies related to spatiotemporal viral community dynamics and the identification of phages with antibiotic resistance genes and virophages [9, 53-56]. However, few freshwater virome studies have explored the viral AMG content and distribution across environmental gradients, such as the water depth, or identified bacteriophage-host interactions at the species level by relatively reliable methods. Lake Fuxian (24° 35'N, 102° 50'E, 1788.5 m a.s.l.) is the third deepest lake in the China Yunnan-Guizhou Plateau (its maximum depth is 158.9 m) and is a warm, oligotrophic monomictic lake [57]. Previous metagenomic deep sequencing studies took this lake as a model system for studying the microbiome community in the holomictic period. For Lake Fuxian, which had a microbiome stratification pattern in the epilimnion and hypolimnion, we obtained 440 MAGs across 10 phyla and found that the epilimnion had a relatively high phage abundance [58]. In that study, the phage community was classified at the read level (< 0.5% of total reads) based on the NCBI RefSeq database, which only contained cultivated species, while most bacteriophages detected by metagenomic methods were novel and uncultivated [58]. Considering the wealth of good-quality and deeply sequenced metagenomic data for Lake Fuxian, this research provided an opportunity to thoroughly characterize the virome in this deep freshwater lake. Therefore, the primary aim of this study was to fill the research gap regarding the interactions between bacteriophages and their hosts by introducing Hi-C technology to environmental metagenome-based virome research and to provide a model for studying bacteriophages-host associations in natural ecosystems.

Methods

Hi-C sample collection and sequencing

Approximately 15 liters of Lake Fuxian raw water was collected from a depth of 20 m in December 2019 (in the holomictic period) when the sampling time was the same as that in a previous ultradeep

metagenome study [58]. Subsequently, 15 liters of water was filtered through 0.1- μ m filters (Isopore™ Membrane Filters, Merck Millipore) and used for the Hi-C experiment.

Microbe cells were lysed from one piece of membrane by glass bead disruption in detergent buffer. Next, we collected approximately 300 μ l of solid material from the filtered membrane. The Hi-C library was constructed using a ProxiMeta™ kit according to the ProxiMeta Hi-C protocols provided by Phase Genomics (Seattle, WA, USA). Importantly, Hi-C DNA was fragmented using the Sau3AI (5'- $\wedge$ GATC-3') restriction enzyme. Finally, more than 200 million read pairs of Hi-C sequencing data were generated by NovaSeq 6000 sequencing systems (Illumina, USA).

Reanalysis of published Lake Fuxian metagenome datasets

First, raw reads for each sample were trimmed using the JAVA program Trimmomatic (version 0.33) to remove sequencing adapters and low-quality sequences (default parameters) [69]. To decrease the potential contamination from sampling and sequencing library construction procedures, we mapped the remaining sequences to the human genome (NCBI GRCh38) by BWA-MEM (v.0.7.17) with default parameters. This quality control step used in this study was different from a previous study, lacking the removal of known nonenvironmental bacterial genomes where prophage fragments might have existed to maximize the virome detection probability [58, 70]. Second, MEGAHIT (v.1.1.1; parameter options: `-min-contig-len 500 -k-min 21 -k-max 141`) was introduced to perform metagenome assembly for each sample and generate a total of 7.11 Gb contigs (metagenome size was from 1.11 Gb to 1.87 Gb) [71]. Moreover, a coassembly strategy (five samples along the depth profile) was also applied to improve the completeness and quality of Lake Fuxian metagenomes, while more than 6.53 Gb contigs were put into identify viral sequences (Additional File 2: Table S2).

MAGs were created by multiple binning software programs with both single-sample assemblies and coassembly results. Sequencing depth for each contig (minimum contig length $\geq$ 1500 bp) was calculated using the functional script 'jgi_summarize_bam_contig_depths' from the bbtools suite (<https://sourceforge.net/projects/bbmap/>) based on the sorted bam files generated by BWA-MEM (v.0.7.17) and SAMtools (v1.546) together [72, 73]. Then, MetaBAT2 (v.2.12.1), CONCOCT (v0.4.0) and MaxBin (v2.2.4) were applied to bin the assemblies with contig depth results under default parameters [74-76]. Bins generated by the above three methods were considered as the input for DAS Tool (v1.1.0) to obtain high-quality recovered MAGs [77]. CheckM (v.1.0.7) with lineage_wf workflow was used to estimate the quality of MAGs (completeness and contamination) [78]. To reduce the replication of MAGs, we used dRep to group MAGs under ANI > 95% and picked the final host MAGs with the highest QualityScore value (defined as completeness – 5X contamination) as Lake Fuxian Metagenome-assembled Genomes (FXLMGs) [79, 80]. For FXLMG taxonomy identification, the GTDB Toolkit (Genome Taxonomy Database, downloaded July 2020, version r95) was introduced to obtain the taxonomy information for each FXLMG [81].

Bacteriophage genomes identification

Following a previous virome study in hot spring, we used VirSorter and sequence BLASTN comparison to the IMG/VR database to detect viral genomes in the unbinned fraction of the assembled contigs [10, 60, 82]. First, the contigs (≥ 10 kb) classified as VirSorter categories 1-2 (viral genomes) and 4-5 (integrated viral contigs) were first picked as candidate viral contigs. To improve the viral quality from VirSorter, contigs whose Pfam hit scores were higher than those of viral hits were discarded before further analysis. The second total unbinned contigs were BLASTN against the IMG/VR database (downloaded July 2020, version 2 IMG_VR_2018-07-01_4) with an e-value less than $1e-5$. Only best subject contigs remained, and corresponding hits were merged into consensus coordinate ranges by custom PERL scripts. Only contigs with at least 90% identity and 75% alignment coverage to IMG/VR viral genomes were assigned as IMG/VR viral clusters. Contigs with IMG/VR- and/or VirSorter-positive hits were considered final viral sequences. Then, the final viral sequences were aligned against each other by nucmer in the mummer package (v.3) [83]. Viral sequences sharing more than 95% nucleotide identity across more than 80% of the whole genome were dereplicated into groups, and the longest sequence within each group was chosen as the representative nonredundant Lake Fuxian viral genome (FXLVG). The remaining unbinned and non-FXLVG contigs were annotated by CAT (version 2020-06-18) with default parameters [84].

Bacteriophage taxonomic assignments and functional annotation

vConTACT2 was used to classify the taxonomy for FXLVGs [85]. For each contig of FXLVGs, ORFs were predicted by Prodigal with ‘-p meta’ parameter [86]. The corresponding protein sequences were treated as the input for vConTACT2, and DIAMOND was used to compare the similarity between all gene pairs based on amino acid homology [87]. The phage references used in vConTACT2 were from the NCBI RefSeq database (released on November 01, 2019, totaling 12,155 viral genomes). ClusterONE within the vConTACT2 pipeline was performed in the viral population clustering step (-d 0.3 -s 2 -max-overlap 0.9 -haircut 0.55). Networks representing the viral clusters were graphed using Gephi (v.0.9.2, <https://gephi.org/>). Auxiliary metabolic genes were annotated by VIBRANT (v1.2.1) with default parameters [64].

Abundance calculation

Sequencing depth has great impacts on the metagenomic gene abundance profile, especially for ultralow- and/or high-abundance objects [88]; therefore, 290 M reads (the lowest sequencing read number among samples) were randomly subsampled without replacement from five metagenome-sequencing samples. The relative abundance of FXLMGs and FXLVGs was calculated as the genome coverage ratio using BBDMap (a pipeline from the BBTools package, <https://jgi.doe.gov/data-and-tools/bbtools/>) and the SAMtools ‘depth’ module (v1.546) with default parameters. Representative FXLMGs and FXLVGs described above were selected for abundance calculation, and only genomes with reads across more than 75% of the corresponding genome length were treated as positive hits; otherwise, the relative abundance of objects was considered to be zero. The habitat specificity of FXLMGs and FXLVGs (using epilimnion as an example, S_{epi}) was measured as the quotient of relative abundance in the epilimnion (20 m) versus the sum of abundance in the epilimnion and hypolimnion (20, 60, 80, 120 and 140 m)

[$S_{\text{epi}} = \text{epilimnion abundance} / (\text{epilimnion} + \text{hypolimnion abundance}) * 100\%$]. Epilimnion-specific FXLMGs and FXLVGs were defined as those whose S_{epi} value was larger than 90%. So were hypolimnion-specific FXLMGs and FXLVGs. BHR (bacteriophage/host ratio) was defined as the bacteriophage genome abundance divided by the host abundance.

CRISPR-based bacteriophage-host linking

CRISPR spacers in FXLMGs and unbinned contigs were identified by three software programs: CRT, CRISPRFinder and CRISPRDetect [89-91]. Spacers from one FXLMG and contigs were grouped by cd-hit-est (v.4.8.1; parameter options: -g 1 -c 1 -G 0 -M 0 -d 0 -aS 1) [92]. Dereplicated spacer sequences were compared against FXLVGs by BLASTN, and only matched pairs with $\geq 97\%$ alignment coverage and $\geq 97\%$ nucleotide identity were considered positive bacteriophage-host matches.

Hi-C-based bacteriophage-host linking

Hi-C data using the Sau3AI (5'-[^]GATC-3') restriction enzyme were first trimmed to low-quality reads by Trimmomatic (default setting), and filtered high-quality reads were mapped to FXLMGs and FXLVGs simultaneously by BWA-MEM (v.0.7.17) and SAMtools (v1.546) with parameters recommended by Phase Genomics, Inc. (US Seattle, hic_qc.py script, <https://github.com/phasegenomics>). Hi-C contacts were extracted from mapping results by ALLHiC software (setting: "-RE GATC") [41]. The number of restriction enzyme cutting sites ($RECS$) for each genome and the observed link number (OLN) between each pair of genomes were recovered by the ALLHiC-extract module. To obtain highly reliable bacteriophage-host links from Hi-C contacts, we used the abundance mentioned above, as the normalized read coverage M and H is the assembly set of FXLMGs and FXLVGs respectively. We defined the expected link number $ELN(ij) = RECS(i) * M(i) * RECS(j) * H(j)$. We modeled the statistic (D) of spurious contact between all pairs of FXLMGs and FXLVGs as $D = \text{rank of } OLN(ij) - \text{rank of } ELN(ij)$. The Kolmogorov-Smirnov test was used to test the normal distribution of statistic D . The threshold for Hi-C significant contacts between FXLMGs and FXLVGs was $P < 0.05$ (single-tailed test).

Results And Discussion

Overview of viral genome in Lake Fuxian

From the five shotgun metagenomes and coassembly results, 1,464 viral sequences (> 10 kb) were recovered from Lake Fuxian. Of these sequences, 942 were identified by the *de novo* method (VirSorter), 211 showed high similarity with reported bacteriophage sequences in the IMG/VR database, and 311 were recovered by both methods (Additional File 1: Table S1). Both approaches for bacteriophage identification employed in this study (VirSorter and comparison to IMG/VR) were based on an *a priori* algorithm with known bacteriophage sequences and core bacteriophage proteins, reducing the number of FXLVGs detected but increasing the accuracy of the results [59, 60]. After clustering at least 95% sequence similarity and 85% aligned coverage, 985 nonredundant Lake Fuxian viral genomes (FXLVGs) were obtained for further analysis, with genome lengths ranging from 10 kb to 451.7 kb and guanine-

cytosine (GC) contents ranging from 19.37% to 74.05% (Additional File 1: Table S1). Four bacteriophages measured more than 200 kb in length and were classified as prophages by VirSorter; therefore, we did not consider them as giant viruses that were reported in previous study [61]. FXLVG richness at 20 m in depth was the highest (three times larger than the average richness at other depths), while the richness from 60 m to the bottom decreased rapidly (Fig. 1a; Additional File 2: Table S2). The FXLVG richness status also supported the profile that not only microorganisms, including bacteria and archaea, but also phages had stratified features during the holomictic period [58].

The genomic heterogeneity of the freshwater virome dataset was evaluated in our study. FXLVGs were compared with viral data obtained from other freshwater habitats at the nucleic acid level with the same sequence similarity cutoff used in FXLVG construction. There was no FXLVG matching with either the UFO (Uncultured Freshwater Organisms) or LBVC (Lake Biwa Viral Contigs) viral datasets, indicating that the freshwater virome had high habitat heterogeneity. A total of 655 FXLVGs were novel sequences that were first reported in this study (Additional File 3: Fig. S1) [27, 55]. It might not be productive to compare viral genomes from similar freshwater habitats even with the same bacteriophage identification pipeline, let alone virome dataset from different habitats, such as TARA Ocean, the large ocean virome GOV2.0 dataset generated by at least three kinds of viral sequences detection software [7]. Remarkably, among 322 IMG/VR-like FXLVGs, 245 (76.09%) were from freshwater, and 59 (18.32%) were from other aquatic ecosystems, such as marine and wastewater (Additional File 4: Table S3). A total of 94.41% of IMG/VR-like FXLVGs were traced from aquatic environments, suggesting that our viral detection pipeline was reliable.

A gene-sharing network was constructed by vConTACT2 to classify FXLVGs (985 accessions) with publicly available taxonomy-known viral sequences in RefSeq (12,155 accessions, March 2020). The classification using vConTACT2 was able to cluster viral sequences with known genomes. A total of 57.44% (n=7547, 7547/13,140) of viral accessions were clustered, and the remaining accessions did not cluster with any viral genomes. Of the 985 FXLVGs, 778 could be assigned to 158 viral clusters (VCs), of which 21 VCs contained taxonomy-known viral genome members (Additional File 4: Table S3; Additional File 5: Fig. S2). Finally, 66 (6.63%) FXLVGs clustered within the VCs belonging to the families *Podoviridae* (n=20), *Siphoviridae* (n=20), *Myoviridae* (n=19), *Mimiviridae* (n=5) and *Lavidaviridae* (n=2). Except for five *Mimiviridae* FXLVGs, other known-family FXLVGs belong to Caudovirales classified as double-stranded DNA viruses (dsDNA viruses). *Mimiviridae* and *Lavidaviridae* were detected only in the epilimnion, while the ratio of *Podoviridae* in the hypolimnion was two times that in the epilimnion. These results indicated that the majority of the FXLVGs obtained from Lake Fuxian were completely unclassified, and the bacteriophage taxonomic structure showed stratification between the epilimnion and hypolimnion during the holomictic period (Additional File 6: Fig. S3). Interestingly, the bacteriophages whose taxonomic percentages were in the top three were *Podoviridae*, *Siphoviridae* and *Myoviridae* in Lake Fuxian, which coincided with those identified from the ocean virome GOV2.0 and wastewater viromes [7, 18]. In Lake Fuxian, putative hosts for *Myoviridae*, *Podoviridae* and *Siphoviridae* were members of *Caulobacter*, *Myxococcus* and *Enterobacteria*. In wastewater plants, the most abundant hosts were *Acinetobacter*,

Arcobacter, and *Moraxella* [18]. This finding demonstrated that host structure could be more diverse than viral taxonomy structure among different habitats.

Depth profile for bacteriophage abundance and gene function

The relative abundance of FXLVGs was determined based on the metagenomic average depth, calculated as read coverage per base under the same read input for each sample. Among the 985 FXLVGs, 695 (70.56%) were epilimnion-specific, and 161 (16.35%) were hypolimnion-specific, which was consistent with the viral richness results (Fig. 1b). With respect to the GC content of the clean metagenomic reads, an increasing GC content of FXLVGs was observed between the epi- and hypolimnion ($40.75\pm 9.09\%$ and $51.98\pm 12.31\%$), whose pattern coincided with those of MAGs (Fig. 1c; Additional File 1: Table S1) [58]. In the hypolimnion, 17 FXLVGs were depth-specific (7 at 120 m, 6 at 140 m, 2 at 60 m and 2 at 80 m). These numbers were considerably lower than those observed at 20 m (epilimnion-specific FXLVGs). We observed that epilimnion-specific MAGs (higher-temperature habitats) had considerably smaller genome sizes than hypolimnion-specific MAGs (Additional File 7: Fig. S4). However, no depth profile was detected for FXLVG genome length based on the current genome completeness. Bacteria and archaea growing in a high-optimum-temperature environment were found to have smaller genomes in a previous report [62]. These results demonstrated that the genome evolutionary process might differ between bacteriophages and their prokaryotic hosts.

To further explore how bacteriophages might affect the biogeochemistry and the potential relationship between viral abundance specificity and function at different lake depths, AMGs in FXLVGs, which are host metabolic genes carried by viral genomes and function during the infection process, were categorized [63]. Overall, 567 AMGs were detected in 239 (24.26%) FXLVGs (2.37 AMGs per bacteriophage) (Fig. 2; Additional File 8: Table S4). Similar to previous freshwater and soil virome studies, carbohydrate metabolism-related AMGs were the most popular in FXLVGs [9, 12]. The proportion of AMG-encoding epilimnion-specific FXLVGs (198/695, 28.49%) was significantly higher than that of hypolimnion-specific FXLVGs (18/161, 11.18%) (Fisher's exact test, $p < 0.01$). The proportion differences of aromatic compounds, glycans, nucleotides and sulfur relay metabolism-related AMGs between epi- and hypolimnion-specific FXLVGs exceeded more than 2-fold (Fig. 2). Among the four metabolism categories, epilimnion-specific FXLVGs exhibited more AMGs in aromatic compound and glycan metabolism, while hypolimnion-specific FXLVGs exhibited more AMGs in nucleotide and sulfur relay metabolism.

Notably, there were two AMGs involved in sulfur relay metabolism, *moeB* (molybdopterin-synthase adenylyltransferase, EC:2.7.7.80) and *mec* ([CysO sulfur-carrier protein]-S-L-cysteine hydrolase, EC:3.13.1.6). Both genes were detected in epi- and hypolimnion-specific FXLVGs, indicating that bacteriophages might universally be involved in the sulfur relay cycle in the host genome. Moreover, the proportional differences in amino acid, carbohydrate, cofactor/vitamin, energy, lipid, terpenoid/polyketide metabolism and secondary metabolite-related AMGs between epi- and hypolimnion-specific FXLVGs were minor (< 2 -fold difference) (Fig. 2). However, regarding the detailed gene functions, there were large

differences. For example, the glycine hydroxymethyltransferase (*glyA*, EC:2.1.2.1) gene involved in methane metabolism only existed in epilimnion-specific FXLVGs. In contrast, *pdhB* (pyruvate dehydrogenase E1 component beta subunit) and *aceF* (pyruvate dehydrogenase E2 component), two members of the citrate cycle (TCA cycle), were only detected in the hypolimnion-specific FXLVG867 (Additional File 8: Table S4). These results demonstrated that bacteriophages might play important roles in basic carbohydrate metabolism in deep freshwater ecosystems.

All these results demonstrated that the epi- and hypolimnion-specific lineages were clearly differentiated by viral abundance and GC content, and depth-specific AMGs might engage in favorable functional niches.

CRISPR-based bacteriophage-host link events

Since the introduction of new bioinformatic tools in the data analysis pipeline, the MAG dataset in this study has improved considerably compared with a previous report [58]. Coassembly and binning of metagenomes initially generated 2,431 metagenome-assembled genomes (FXLMGs) for Lake Fuxian, including 405 high-quality (> 90% completeness and < 5% contamination), 578 medium-quality ($\geq$ 50% completeness and < 10% contamination) and 1,448 low-quality (others) FXLMGs, doubling the number of FXLMGs in the former dataset (Additional File 9: Table S5). After clustering at 95% ANI, 431 high- or medium-quality representative FXLMGs, including diverse bacterial and archaeal phyla, remained (Additional File 10: Table S6). Among 431 representative FXLMGs, 423 (98.14%) were bacterial FXLMGs spanning 22 phyla, where Proteobacteria (n = 134), Actinobacteriota (n = 71), Planctomycetota (n = 51), Verrucomicrobiota (n = 34), Bacteroidota (n = 30) and Patescibacteria (n = 22) were highly represented (> 5%). Within the domain Archaea, 8 FXLMGs, Crenarchaeota (n = 2) and Nanoarchaeota (n = 6), were recovered (Fig. 3).

As a relatively reliable computational method for detecting bacteriophage-host association signals, the CRISPR spacer-based approach has been applied in many virome studies [10, 14, 26-28]. A total of 7,075 CRISPR spacer sequences were identified from 393 representative MAGs covering a total of 24 phyla (Additional File 10: Table S6). The spacer sequencer distribution at the phylum level was similar to the FXLMG numbers across each phylum. The top five phyla (Proteobacteria, Actinobacteriota, Planctomycetota, Verrucomicrobiota and Bacteroidota) also had the highest numbers of spacer sequencers (Fig. 3). The spacer sequence number varied notably at the single-FXLMG level; for example, one Myxococcota FXLMG had 526 spacer sequences, while 49 FXLMGs only had 1 for each (Fig. 3; Additional File 10: Table S6; Additional File 11: Fig. S5). Furthermore, only 10 CRISPR-based virus-host events (10 FXLVGs and 8 FXLMGs) were identified from 985 FXLVGs detected in Lake Fuxian (Additional File 12: Table S7). Among these sequences, one Chloroflexota FXLMG (FXLMG137) linked with three FXLMGs, but the other FXLMGs linked one-by-one with a unique FXLVG. We also searched CRISPR spacer sequences in unbinned or no-FXLVG fractions. A total of 2,895 spacer sequences and 19 link events were recovered, but only 10 hosts could be classified as bacteria, while the others were unclassified (Additional File 13: Table S8).

Similarly, in cone pool virome research, only three bacteriophage-host link events were detected by CRISPR [10]. Twenty-six (1.28%, 26/2034 total viral genomes) bacteriophages had host information identified by the CRISPR-based method in two freshwater habitats [27]. These results suggested that most MAGs lack a CRISPR defense system to combat current phages, and existing CRISPR elements across the host genomes might be evolutionary traces of host bacteriophage-fighting machinery in the history of host genomes [28]. On the other hand, the bioinformatic principle of FXLMG reconstruction could not be constructed efficiently for very low-abundance microbial genomes [6, 22]. Meanwhile, bacteriophages could also control host abundance to an extent. Nevertheless, the low efficiency of CRISPR-based link event detection in Lake Fuxian suggests the need to use other strategies for linking bacteriophages to their hosts.

Hi-C-based bacteriophage-host link events

Because of the highest richness and abundance of FXLVGs in the epilimnion, we chose the epilimnion sample as a model to investigate the bacteriophage-host relationship by Hi-C technology. A total of 3,118 candidate bacteriophage-host link events were initially identified (Additional File 14: Table S9). Considering the bacteriophage and host abundance and the number of restriction sites in each sequence, 58 link events (confidence > 95%) were eventually detected across 31 representative FXLMGs spanning 8 phyla (Fig. 3). Among 58 link events, 8 FXLVGs had more than one host, and 2 of them might infect FXLMGs across different phyla. In addition, 11 FXLMGs could be infected by multiple FXLVGs (Additional File 15: Fig. S6). There were 24 (24/58, 41.38%) events (among 8 bacteriophage-host clusters), where the corresponding FXLVGs contained AMGs (Fig. 4a). Interestingly, link events detected by CRISPR and Hi-C were completely different in our study. The low number of CRISPR-based events decreased the possibility of overlap event detection, and the Hi-C-based method was focused on the infection events within each host cell *in situ*. Therefore, it was reasonable that there was no link event detected by either strategy.

Strong correlation signals for functional complementarity between FXLVGs and FXLMGs were observed in the Hi-C-based results. The infection signals of 11 Cyanobacteriota-FXLVG link events were identified, including 2 epilimnion-specific FXLMGs and 8 epilimnion-specific FXLVGs (Additional File 12: Table S7). Four of the eight FXLVGs (FXLVG66, FXLVG309, FXLVG472 and FXLVG500) contained D1 or D2 proteins, which were members of the photosystem II P680 reaction center (*psbA* and *psbD* in the VIBRANT software classification system). Among these four FXLVGs, proteins PC and Fd, which participate in photosynthetic electron transport, were also identified (Additional File 16: Fig. S7). Remarkably, one of 8 cyanophages, FXLVG608, had the *mec* gene ([CysO sulfur-carrier protein]-S-L-cysteine hydrolase, EC:3.13.1.6), and its infected host was FXLMG141, which lacked the *mec* gene in the sulfur relay system (Fig. 4b). These results indicated that cyanophages might provide other important genes to Cyanobacteriota in addition to the most widely known D1 and D2. In previous reports, D1/D2 proteins in viral AMGs were generally used to predict potential Cyanobacteriota hosts [26, 27, 64, 65], but the limitation was that only phylum information of the host could be predicted, and the corresponding bacteriophages were uniformly classified as cyanophages.

Another bacteriophage-host functional complementary event was found in the mannose-fucose metabolism pathway (KEGG ko00051) but with different strategies of cooperation. Mannose and fucose are sugars, with the former being an essential endoplasmic reticulum component in prokaryotes, while the latter is distributed in complex carbohydrates of all species. In prokaryotes, GDP fucose is synthesized from GDP-mannose in a three-step reaction catalyzed by two enzymes, GMDS (GDP-mannose 4,6-dehydratase, EC:4.2.1.47) and TSTA3 (GDP-L-fucose synthase, EC:1.1.1.271) (Fig. 4c) [66]. In the FXLVG611-FXLMG349-FXLMG350 cluster (VHC4), two key enzymes (GMDS and TSTA3) were identified in the bacteriophage, and both hosts contained a PPM gene (phosphomannomutase, EC:5.4.2.8), which was used to transform mannose-6-phosphate into mannose-1-phosphate. In this case, it appears that the host could utilize the fucose synthesized within the bacteriophage. However, in the FXLVG794-FXLMG254 cluster (VHC5), only TSTA3 was identified in FXLVG794, and GMDS was observed in the host FXLMG254 (Planctomycetota). These results demonstrated that bacteriophages might play an important role in fucose synthesis, which is necessary in bacterial cell construction, although cooperation can be very diverse. Nevertheless, these findings demonstrate the considerable power of Hi-C in bacteriophage-host link event detection in the future.

Bacteriophage-host interactive abundance analysis

Based on the identified bacteriophage-host pairs and genome coverage information obtained by the read mapping approach along the depth profile, this research provided an opportunity to evaluate the abundance relationship between bacteriophages and hosts at species level. For 58 Hi-C-based bacteriophage-host pairs in five samples (290 events), 170 (58.72%) events were discarded in further analysis because of undetectable coverage of the bacteriophage and/or host genome (Additional File 17: Table S10). All 170 events were from hypolimnion samples, which might be observed because only the epilimnion sample was selected to perform the Hi-C experiment, and no hypolimnion-specific FXLVG was identified in the Hi-C-based bacteriophage-host relationship results. Among the remaining 120 events, more than half of the hosts (62/120, 51.67%) had higher genome coverage than their bacteriophages (bacteriophage/host abundance ratios, BHR < 1.5). A total of 22.5% (27/120) of events had nearly equal genome coverages ($-1.5 \leq \text{BHR} \leq 1.5$), and 25.83% (31/120) had higher bacteriophage coverage (Additional File 18: Fig. S8). Notably, these results were significantly different from those obtained in a previous virome study in which most bacteriophages had greater abundance than their hosts [10, 67]. The uncertainty provides another opportunity to use the Hi-C-based method to detect the bacteriophage-host relationship within prokaryotic cells instead of counting bacteriophages at lysis time.

Similar to FXLVG abundance, BHR showed a specific pattern along the depth profile and host diversity. BHRs were significantly ($p < 0.01$) higher in the epilimnion than in the hypolimnion (Fig. 5a). The BHRs in the epilimnion and hypolimnion were both negatively correlated with host abundance (Fig. 5b), and BHRs in the epilimnion decreased considerably more quickly with lower host abundance. Remarkably, BHRs varied among different hosts at the phylum level (Additional File 19: Fig. S9). For Cyanobacteriota, total FXLVGs had a higher abundance than FXLMGs (> 6.97-fold on average). For Planctomycetota, FXLMGs in hypolimnion had higher abundance where the corresponding BHRs were in considerably lower

abundance, while in epilimnion, the opposite distribution was observed. These results were in keeping with the “Piggyback-the-Winner” theory that lysogeny was the predominant viral lifestyle in a high-host-abundance environment where bacteriophages (integrated into the host genome or coexisting within the host cell) increased the ability of the host to resist other bacteriophage infections and gain competitive advantages compared with other microbes [68].

Conclusion

Metagenomic techniques and decreasing sequencing costs provided an opportunity to survey the depth profile of bacteriophage distribution and ecological functions in an alpine freshwater lake. Our study contributes nearly one thousand nonredundant viral genomes to the current collection of freshwater phage communities. The results of bacteriophage abundance profiles and AMG distribution along the depth provide further evidence of the incomplete mixture in Lake Fuxian during the holomictic period. To the best of our knowledge, this study was the first to employ the Hi-C technique to characterize the bacteriophage-host association for real field ecological samples, and we identified 58 bacteriophage-host association events. The results of AMG association analysis within bacteriophage-host pairs not only suggest that bacteriophages and their hosts might collaborate closely and have complementary functions at the gene level but also provide reliable evidence to indicating the robustness of the Hi-C-based bacteriophage-host detection method. We believe that it will be useful to design virome studies in extensive ecosystems combining metagenomics and the Hi-C technique in future research.

Abbreviations

dsDNA: Double-Stranded DNA; AMGs: Auxiliary Metabolic Genes; HGT: Horizontal Gene Transferring; CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats; AdsorpSeq: Adsorption Sequencing; Hi-C: High-throughput Chromosomal Confirmation Capture; MAGs: Metagenome-Assembled Genomes; FXLVGs: Lake Fuxian Viral Genomes; GC: guanine-cytosine; UFO: Uncultured Freshwater Organisms; LBVC: Lake Biwa Viral Contigs; VCs: Viral Clusters; moeB: molybdopterin-synthase adenyltransferase; mec: [CysO sulfur-carrier protein]-S-L-cysteine hydrolase; glyA: glycine hydroxymethyltransferase; pdhB: pyruvate dehydrogenase E1 component beta subunit; aceF: pyruvate dehydrogenase E2 component; FXLMGs: Lake Fuxian Metagenome-assembled Genomes; KEGG: Kyoto Encyclopedia of Genes and Genomes; MDS: GDP-mannose 4,6-dehydratase; TSTA3: GDP-L-fucose synthase; PPM: phosphomannomutase; psbA: photosystem II P680 reaction center D1 protein; psbD: photosystem II P680 reaction center D2 protein; BHR: bacteriophage/host ratio.

Declarations

Acknowledgements

We would like to thank Wenlei Luo and Qinghua Liu for their help in the field sampling organization. We appreciated the support from Fuxianhu Laboratory for Alpine Deep Lake Research, Nanjing Institute of

Geography and Limnology, Chinese Academy of Sciences.

Funding

This work was funded by National Natural Science Foundation of China (31722008), Science & Technology Basic Resources Investigation Program of China (2017FY100300), the Second Tibetan Plateau Scientific Expedition and Research (STEP) program (Grant No. 2019QZKK0503), Chinese Academy of Sciences (QYZDJ-SSW-DQC030), and the Youth Innovation Promotion Association of CAS (2014273).

Availability of data and materials

Raw sequence reads for Hi-C sample are available under China National GeneBank (CNGB) with accession CNP0001442. All viral sequences and MAGs are available at <https://figshare.com/s/4c3ed1590ae56b0d23db>.

Authors' contributions

PX and QLW designed the experiments. YT, PG and BL performed the experiments. YT analysed the data. YT wrote the main manuscript. All authors read and approved the final manuscript.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

1State Key Laboratory of Lake Science and Environment, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing 210008, China.

2University of Chinese Academy of Sciences, Beijing 100049, China.

3Sino-Danish Centre for Education and Research, University of Chinese Academy of Sciences, Beijing 100049, China.

References

1. Suttle CA. Marine viruses—major players in the global ecosystem. *Nat Rev Microbiol.* 2007;5(10):801-812.
2. Jacquet S, Miki T, Noble R, Peduzzi P, Wilhelm S. Viruses in aquatic ecosystems: Important advancements of the last 20 years and prospects for the future in the field of microbial

- oceanography and limnology. *Adv Oceanogr Limnol*. 2010;1:97–141.
3. Salmond GP, Fineran PC. A century of the phage: past, present and future. *Nat Rev Microbiol*. 2015;13(12):777-786.
 4. Bergh Ø, Børsheim KY, Bratbak G, Haldal M. High abundance of viruses found in aquatic environments. *Nature*. 1989;340(6233):467-468.
 5. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. *PeerJ*. 2015;3:e985.
 6. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*. 2017;5(1):69.
 7. Gregory AC, Zayed AA, Conceicao-Neto N, Temperton B, Bolduc B, Alberti A, Ardyna M, Arkhipova K, Carmichael M, Cruaud C et al. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell*. 2019;177(5):1109-1123.
 8. Zaragoza-Solas A, Rodriguez-Valera F, López-Pérez M. Metagenome Mining Reveals Hidden Genomic Diversity of Pelagimyophages in Aquatic Environments. *mSystems*. 2020;5(1):e00905-00919.
 9. Moon K, Jeon JH, Kang I, Park KS, Lee K, Cha CJ, Lee SH, Cho JC. Freshwater viral metagenome reveals novel and functional phage-borne antibiotic resistance genes. *Microbiome*. 2020;8(1):75.
 10. Jarett JK, Dzunkova M, Schulz F, Roux S, Paez-Espino D, Eloë-Fadrosch E, Jungbluth SP, Ivanova N, Spear JR, Carr SA et al. Insights into the dynamics between viruses and their hosts in a hot spring microbial mat. *The ISME Journal*. 2020;14(10):2527-2541.
 11. Bi L, Yu DT, Du S, Zhang LM, Zhang LY, Wu CF, Xiong C, Han LL, He JZ. Diversity and potential biogeochemical impacts of viruses in bulk and rhizosphere soils. *Environ Microbiol*. 2020 [cited 2020 April 05]. Available from: <https://doi.org/10.1111/1462-2920.15010>.
 12. Jin M, Guo X, Zhang R, Qu W, Gao B, Zeng R. Diversities and potential biogeochemical impacts of mangrove soil viruses. *Microbiome*. 2019;7(1):58.
 13. Hurwitz BL, Hallam SJ, Sullivan MB. Metabolic reprogramming by viruses in the sunlit and dark ocean. *Genome Biol*. 2013;14(11):R123.
 14. Chen L-X, Méheust R, Crits-Christoph A, McMahon KD, Nelson TC, Slater GF, Warren LA, Banfield JF. Large freshwater phages with the potential to augment aerobic methane oxidation. *Nature Microbiology*. 2020;5(12):1504-1515.
 15. Ahlgren NA, Fuchsman CA, Rocap G, Fuhrman JA. Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. *The ISME Journal*. 2019;13(3):618-631.
 16. Zeng Q, Chisholm Sallie W. Marine Viruses Exploit Their Host's Two-Component Regulatory System in Response to Resource Limitation. *Current Biology*. 2012;22(2):124-128.
 17. Anantharaman K, Duhaime MB, Breier JA, Wendt KA, Toner BM, Dick GJ. Sulfur oxidation genes in diverse deep-sea viruses. *Science*. 2014;344(6185):757-760.

18. Gulino K, Rahman J, Badri M, Morton J, Bonneau R, Ghedin E. Initial Mapping of the New York City Wastewater Virome. *mSystems*. 2020;5(3):e00876-00819.
19. Mara P, Vik D, Pachiadaki MG, Suter EA, Poulos B, Taylor GT, Sullivan MB, Edgcomb VP. Viral elements and their potential influence on microbial processes along the permanently stratified Cariaco Basin redoxcline. *The ISME Journal*. 2020;14(12):3079-3092.
20. Dwivedi B, Xue B, Lundin D, Edwards RA, Breitbart M. A bioinformatic analysis of ribonucleotide reductase genes in phage genomes and metagenomes. *BMC Evol Biol*. 2013;13:33.
21. Enav H, Mandel-Gutfreund Y, Béjà O. Comparative metagenomic analyses reveal viral-induced shifts of host metabolism towards nucleotide biosynthesis. *Microbiome*. 2014;2(1):9-9.
22. Mahmoudabadi G, Phillips R. A comprehensive and quantitative exploration of thousands of viral genomes. *eLife*. 2018;7:e31955.
23. Puxty RJ, Evans DJ, Millard AD, Scanlan DJ. Energy limitation of cyanophage development: implications for marine carbon cycling. *The ISME Journal*. 2018;12(5):1273-1286.
24. Mya Breitbart, Luke R. Thompson, Curtis A. Suttle, Matthew B. Sullivan. Exploring the Vast Diversity of Marine Viruses. *Oceanography*. 2007;20(2):135-139.
25. Mann NH, Cook A, Millard A, Bailey S, Clokie M. Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature*. 2003;424(6950):741.
26. Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, Brussaard CPD, Dutilh BE, Thompson FL. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nature Communications*. 2017;8(1):15955.
27. Kavagutti VS, Andrei A-Ş, Mehrshad M, Salcher MM, Ghai R. Phage-centric ecological interactions in aquatic ecosystems revealed through ultra-deep metagenomics. *Microbiome*. 2019;7(1):135.
28. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol Rev*. 2016;40(2):258-272.
29. Koonin EV, Wolf YI. Evolution of the CRISPR-Cas adaptive immunity systems in prokaryotes: models and observations on virus–host coevolution. *Molecular BioSystems*. 2015;11(1):20-27.
30. Puschnik AS, Majzoub K, Ooi YS, Carette JE. A CRISPR toolbox to study virus-host interactions. *Nat Rev Microbiol*. 2017;15(6):351-364.
31. Villarroel J, Kleinheinz K, Jurtz VI, Zschach H, Lund O, Nielsen M, Larsen MV. HostPhinder: A Phage Host Prediction Tool. *Viruses*. 2016;8(5):116.
32. Galiez C, Siebert M, Enault F, Vincent J, Söding J. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics*. 2017;33(19):3113-3114.
33. Ghai R, Mehrshad M, Mizuno CM, Rodriguez-Valera F. Metagenomic recovery of phage genomes of uncultured freshwater actinobacteria. *ISME Journal*. 2017;11(1):304-308.
34. de Jonge PA, von Meijenfeldt FAB, Costa AR, Nobrega FL, Brouns SJJ, Dutilh BE. Adsorption Sequencing as a Rapid Method to Link Environmental Bacteriophages to Hosts. *iScience*. 2020;23(9):101439.

35. Labonte JM, Swan BK, Poulos B, Luo H, Koren S, Hallam SJ, Sullivan MB, Woyke T, Wommack KE, Stepanauskas R. Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton. *ISME Journal*. 2015;9(11):2386-2399.
36. Munson-McGee JH, Field EK, Bateson M, Rooney C, Stepanauskas R, Young MJ. Nanoarchaeota, Their Sulfolobales Host, and Nanoarchaeota Virus Distribution across Yellowstone National Park Hot Springs. *Appl Environ Microbiol*. 2015;81(22):7860-7868.
37. van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES. Hi-C: a method to study the three-dimensional architecture of genomes. *Journal of visualized experiments*. 2010;(39):1869.
38. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009;326(5950):289-293.
39. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*. 2012;485(7398):381-385.
40. Bickhart DM, Watson M, Koren S, Panke-Buisse K, Cersosimo LM, Press MO, Van Tassell CP, Van Kessel JAS, Haley BJ, Kim SW et al. Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation. *Genome Biology*. 2019;20(1):153.
41. Zhang J, Zhang X, Tang H, Zhang Q, Hua X, Ma X, Zhu F, Jones T, Zhu X, Bowers J et al. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nature Genetics*. 2018;50(11):1565-1573.
42. Belser C, Istace B, Denis E, Dubarry M, Baurens F-C, Falentin C, Genete M, Berrabah W, Chèvre A-M, Delourme R et al. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature Plants*. 2018;4(11):879-887.
43. Stewart RD, Auffret MD, Warr A, Wiser AH, Press MO, Langford KW, Liachko I, Snelling TJ, Dewhurst RJ, Walker AW et al. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nature Commun*. 2018;9(1):870.
44. Stalder T, Press MO, Sullivan S, Liachko I, Top EM. Linking the resistome and plasmidome to the microbiome. *The ISME Journal*. 2019;13(10):2437-2446.
45. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*. 2017;49(4):643-650.
46. Beitel CW, Froenicke L, Lang JM, Korf IF, Micheltmore RW, Eisen JA, Darling AE. Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ*. 2014;2:e415.
47. Burton JN, Liachko I, Dunham MJ, Shendure J. Species-Level Deconvolution of Metagenome Assemblies with Hi-C–Based Contact Probability Maps. *G3: Genes|Genomes|Genetics*.

2014;4(7):1339.

48. Marbouty M, Cournac A, Flot J-F, Marie-Nelly H, Mozziconacci J, Koszul R. Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *eLife*. 2014;3:e03318-e03318.
49. Marbouty M, Baudry L, Cournac A, Koszul R. Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay. *Science advances*. 2017;3(2):e1602105-e1602105.
50. Press MO, Wiser AH, Kronenberg ZN, Langford KW, Shakya M, Lo C-C, Mueller KA, Sullivan ST, Chain PSG, Liachko I. Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. *bioRxiv*. 2017;198713.
51. Stalder T, Press MO, Sullivan S, Liachko I, Top EM. Linking the resistome and plasmidome to the microbiome. *The ISME Journal*. 2019;13(10):2437-2446.
52. Raymond PA, Hartmann J, Lauerwald R, Sobek S, McDonald CP, Hoover M, Butman D, Striegl RG, Mayorga E, Humborg C et al. Global carbon dioxide emissions from inland waters. *Nature*. 2013;503:355-359.
53. Aguirre de Cárcer D, López-Bueno A, Pearce DA, Alcamí A. Biodiversity and distribution of polar freshwater DNA viruses. *Science Advances*. 2015;1(5):e1400127.
54. Gong C, Zhang W, Zhou X, Wang H, Sun G, Xiao J, Pan Y, Yan S, Wang Y. Novel Virophages Discovered in a Freshwater Lake in China. *Frontiers in Microbiology*. 2016;7:5-5.
55. Okazaki Y, Nishimura Y, Yoshida T, Ogata H, Nakano S-i. Genome-resolved viral and cellular metagenomes revealed potential key virus-host interactions in a deep freshwater lake. *Environmental Microbiology*. 2019;21:4740-4754.
56. Malki K, Rosario K, Sawaya NA, Szekely AJ, Tisza MJ, Breitbart M. Prokaryotic and Viral Community Composition of Freshwater Springs in Florida, USA. *mBio*. 2020;11(2):e00436-00420.
57. Wang, S and H. Dou. China lakes chorography. Science Press, [in Chinese]. 1998.
58. Xing P, Tao Y, Luo J, Wang L, Li B, Li H, Wu QL. Stratification of microbiomes during the holomictic period of Lake Fuxian, an alpine monomictic lake. *Limnology and Oceanography*. 2020;65:S134-S148.
59. Vik DR, Roux S, Brum JR, Bolduc B, Emerson JB, Padilla CC, Stewart FJ, Sullivan MB. Putative archaeal viruses from the mesopelagic ocean. *PeerJ*. 2017;5:e3428.
60. Chen IMA, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, Huntemann M, Varghese N, White JR, Seshadri R et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Research*. 2019;47(D1):D666-D677.
61. Al-Shayeb B, Sachdeva R, Chen LX, Ward F, Munk P, Devoto A, Castelle CJ, Olm MR, Bouma-Gregson K, Amano Y et al. Clades of huge phages from across Earth's ecosystems. *Nature*. 2020;578(7795):425-431.

62. Sorensen JW, Dunivin TK, Tobin TC, Shade A. Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. *Nat Microbiol*. 2019;4(1):55-61.
63. Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, Chisholm SW. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proc Natl Acad Sci USA*. 2011;108(39):E757-764.
64. Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*. 2020;8(1):90.
65. Hevroni G, Enav H, Rohwer F, Beja O. Diversity of viral photosystem-I *psaA* genes. *ISME Journal*. 2015;9(8):1892-1898.
66. Somers WS, Stahl ML, Sullivan FX. GDP-fucose synthetase from *Escherichia coli*: structure of a unique member of the short-chain dehydrogenase/reductase family that catalyzes two distinct reactions at the same active site. *Structure*. 1998;6(12):1601-1612.
67. Li Z, Pan D, Wei G, Pi W, Wang J-H, Peng Y, Zhang L, Wang Y, Hubert CRJ, Dong X. Deep sea sediments associated with cold seeps are a subsurface reservoir of viral diversity. *bioRxiv*. 2020;284018.
68. Silveira CB, Rohwer FL. Piggyback-the-Winner in host-associated microbial communities. *NPJ Biofilms Microbiomes*. 2016;2:16010.
69. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120.
70. Fouts DE. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res*. 2006;34(20):5839-5851.
71. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. 2015;31(10):1674-1676.
72. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.
73. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.
74. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*. 2019;7:e7359.
75. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. Binning metagenomic contigs by coverage and composition. *Nature Methods*. 2014;11:1144.
76. Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*. 2016;32(4):605-607.

77. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol.* 2018;3(7):836-843.
78. 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 research.* 2015;25(7):1043-1055.
79. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME Journal.* 2017;11(12):2864-2868.
80. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, Lawley TD, Finn RD. A new genomic blueprint of the human gut microbiota. *Nature.* 2019;568(7753):499-504.
81. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology.* 2018;36(10):996-1004.
82. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of Molecular Biology.* 1990;215(3):403-410.
83. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. *Genome Biology.* 2004;5(2):R12.
84. von Meijenfeldt FAB, Arkhipova K, Cambuy DD, Coutinho FH, Dutilh BE. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biology.* 2019;20(1):217.
85. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR, Kropinski AM, Krupovic M, Lavigne R et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol.* 2019;37(6):632-639.
86. 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;11:119-119.
87. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nature Methods.* 2015;12(1):59-60.
88. Pereira MB, Wallroth M, Jonsson V, Kristiansson E. Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC genomics.* 2018;19(1):274-274.
89. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P. CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics.* 2007;8(1):209.
90. Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research.* 2007;35(2):W52-W57.
91. Biswas A, Staals RHJ, Morales SE, Fineran PC, Brown CM. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics.* 2016;17(1):356.

92. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28(23):3150-3152.

Figures

Figure 1

The overview of bacteriophage genomes. (a) Richness and taxonomy of viral genomes along vertical profile. 20 m depth sample was in the epilimnion and other samples were in the hypolimnion. (b) A heatmap of relative abundance (normalized by Z-score) of 985 FXLVGs in five metagenome samples. Columns are annotated by the abundance feature. FXLVGs with Sepi > 90% were epilimnion-specific (light green) and hypolimnion-specific FXLVGs had Sepi < 10% (dark green), others were defined as the mixture type. (c) GC content vs genome length for epi- and hypolimnion-specific FXLVGs.

Figure 1

The overview of bacteriophage genomes. (a) Richness and taxonomy of viral genomes along vertical profile. 20 m depth sample was in the epilimnion and other samples were in the hypolimnion. (b) A heatmap of relative abundance (normalized by Z-score) of 985 FXLVGs in five metagenome samples. Columns are annotated by the abundance feature. FXLVGs with Sepi > 90% were epilimnion-specific (light green) and hypolimnion-specific FXLVGs had Sepi < 10% (dark green), others were defined as the mixture type. (c) GC content vs genome length for epi- and hypolimnion-specific FXLVGs.

Figure 2

AMGs numbers and relative ratio between epi- (light green) and hypolimnion-specific (dark green) FXLVGs. 11 functional subcategories of AMGs were classified by VIBRANT. “*”: > 2x fold difference for the corresponding ratio of subcategory.

Figure 2

The overview of bacteriophage genomes. (a) Richness and taxonomy of viral genomes along vertical profile. 20 m depth sample was in the epilimnion and other samples were in the hypolimnion. (b) A heatmap of relative abundance (normalized by Z-score) of 985 FXLVGs in five metagenome samples. Columns are annotated by the abundance feature. FXLVGs with Sepi > 90% were epilimnion-specific (light green) and hypolimnion-specific FXLVGs had Sepi < 10% (dark green), others were defined as the mixture type. (c) GC content vs genome length for epi- and hypolimnion-specific FXLVGs.

Figure 3

AMGs numbers and relative ratio between epi- (light green) and hypolimnion-specific (dark green) FXLVGs. 11 functional subcategories of AMGs were classified by VIBRANT. “*”: > 2x fold difference for the corresponding ratio of subcategory.

Figure 4

AMGs numbers and relative ratio between epi- (light green) and hypolimnion-specific (dark green) FXLVGs. 11 functional subcategories of AMGs were classified by VIBRANT. “*”: > 2× fold difference for the corresponding ratio of subcategory.

Figure 5

Phylogeny, genome size information, number of CRISPR spacers, relative abundance of 431 FXLMGs and their associated bacteriophage. Like FXLVGs, FXLMGs were also categorized into epi- and hypolimnion-specific types. FXLMGs with Sepi > 90% were epilimnion-specific (light green) and hypolimnion-specific FXLMGs had Sepi < 10% (dark green). Pentagram means FXLMGs having more than one associated bacteriophage.

Figure 6

Phylogeny, genome size information, number of CRISPR spacers, relative abundance of 431 FXLMGs and their associated bacteriophage. Like FXLVGs, FXLMGs were also categorized into epi- and hypolimnion-specific types. FXLMGs with Sepi > 90% were epilimnion-specific (light green) and hypolimnion-specific FXLMGs had Sepi < 10% (dark green). Pentagram means FXLMGs having more than one associated bacteriophage.

Figure 7

Phylogeny, genome size information, number of CRISPR spacers, relative abundance of 431 FXLMGs and their associated bacteriophage. Like FXLVGs, FXLMGs were also categorized into epi- and hypolimnion-specific types. FXLMGs with Sepi > 90% were epilimnion-specific (light green) and hypolimnion-specific FXLMGs had Sepi < 10% (dark green). Pentagram means FXLMGs having more than one associated bacteriophage.

Figure 7

Profile of bacteriophage-host association events identified by Hi-C approach. (a) The overview of total bacteriophages with AMGs and their potential host genomes. Triangle: FXLVGs; circular: FXLMGs. (b-c) The examples of functional complementarity between FXLVGs and FXLMGs in sulfur relay system and mannose-fucose metabolism. Genes and substrates in light blue background were found in host genomes, while those in light brown color were identified in bacteriophage genomes.

Figure 8

Abundance profile of bacteriophage and their hosts. (a) The bacteriophage-host ratio comparison between epilimnion (light green) and hypolimnion (dark green). (b) The associations between the microbial host (FXLMGs) abundance and the bacteriophage-host ratio (BHR).

Figure 9

Profile of bacteriophage-host association events identified by Hi-C approach. (a) The overview of total bacteriophages with AMGs and their potential host genomes. Triangle: FXLVGs; circular: FXLMGs. (b-c) The examples of functional complementary between FXLVGs and FXLMGs in sulfur relay system and mannose-fucose metabolism. Genes and substrates in light blue background were found in host genomes, while those in light brown color were identified in bacteriophage genomes.

Figure 10

Abundance profile of bacteriophage and their hosts. (a) The bacteriophage-host ratio comparison between epilimnion (light green) and hypolimnion (dark green). (b) The associations between the microbial host (FXLMGs) abundance and the bacteriophage-host ratio (BHR).

Figure 11

Abundance profile of bacteriophage and their hosts. (a) The bacteriophage-host ratio comparison between epilimnion (light green) and hypolimnion (dark green). (b) The associations between the microbial host (FXLMGs) abundance and the bacteriophage-host ratio (BHR).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [FigureS1.pdf](#)
- [FigureS1.pdf](#)
- [FigureS2.png](#)
- [FigureS2.png](#)
- [FigureS3.png](#)

- [FigureS3.png](#)
- [FigureS4.pdf](#)
- [FigureS5.pdf](#)
- [FigureS5.pdf](#)
- [FigureS4.pdf](#)
- [FigureS6.pdf](#)
- [FigureS7.pdf](#)
- [FigureS6.pdf](#)
- [FigureS7.pdf](#)
- [FigureS8.pdf](#)
- [FigureS9.pdf](#)
- [FigureS8.pdf](#)
- [FigureS9.pdf](#)
- [TableS1.xlsx](#)
- [TableS1.xlsx](#)
- [TableS10.xlsx](#)
- [TableS10.xlsx](#)
- [TableS2.xlsx](#)
- [TableS2.xlsx](#)
- [TableS3.xlsx](#)
- [TableS3.xlsx](#)
- [TableS4.xlsx](#)
- [TableS4.xlsx](#)
- [TableS5.xlsx](#)
- [TableS5.xlsx](#)
- [TableS6.xlsx](#)
- [TableS6.xlsx](#)
- [TableS7.xlsx](#)
- [TableS7.xlsx](#)
- [TableS8.xlsx](#)
- [TableS8.xlsx](#)
- [TableS9.xlsx](#)
- [TableS9.xlsx](#)