Communities of T4-like bacteriophages associated with bacteria in Lake Baikal: diversity and biogeography (#63663)

First submission

Guidance from your Editor

Please submit by **7 Aug 2021** for the benefit of the authors (and your $200 publishing discount) .

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Custom checks
Make sure you include the custom checks shown below, in your review.

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files

Download and review all files from the materials page.

6 Figure file(s)
6 Table file(s)
1 Other file(s)

Custom checks

DNA data checks

- Have you checked the authors data deposition statement?
- Can you access the deposited data?
- Has the data been deposited correctly?
- Is the deposition information noted in the manuscript?

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Communities of T4-like bacteriophages associated with bacteria in Lake Baikal: diversity and biogeography

Sergey Anatoljevich Potapov [Corresp., 1], **Irina Vasilievna Tikhonova** [1], **Andrey Yurjevich Krasnopeev** [1], **Mariya Yurjevna Suslova** [1], **Natalia Albertovna Zhuchenko** [1], **Valentin Valerianovich Drucker** [1], **Olga Ivanovna Belykh** [1]

[1] Limnological Institute Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia

Corresponding Author: Sergey Anatoljevich Potapov
Email address: poet1988@list.ru

We assessed the genetic diversity of T4-like bacteriophages in the fraction greater than 0.2 μm from the pelagic zone, coastal zone and shallow bays of Lake Baikal by the gene fragment of the major capsid protein, gp23. High throughput sequencing allowed us to obtain from 12454 to 41802 sequences of the *g23* gene in the fraction associated with bacteria. The results revealed that the sequences found in this study together with the sequences that we had previously retrieved from the plankton samples (fraction less than 0.2 μm) and biofilms (without separation onto fractions) formed the Baikal cluster. The sequences from shallow bays largely differed from those in the pelagic and coastal samples and formed individual subcluster in the UPGMA tree. According to the RefSeq database, most OTUs had the cultivated closest relatives belonging to cyanophages.

Communities of T4-like bacteriophages associated with bacteria in Lake Baikal: diversity and biogeography

Sergey Anatoljevich Potapov [1], Irina Vasilievna Tikhonova[1], Andrey Yurjevich Krasnopeev[1], Maria Yurjevna Suslova[1], Natalia Albertovna Zhuchenko[1], Valentin Valerianovich Drucker[1], Olga Ivanovna Belykh[1]

[1] Limnological Institute Siberian Branch of the Russian Academy of Sciences, Irkutsk, Ulan-Batorskaya 3, 664033, Russia

Corresponding Author:
Sergey Anatoljevich Potapov[1]
Ulan-Batorskaya 3, Irkutsk, 664033, Russia
Email address: poet1988@list.ru

Abstract

We assessed the genetic diversity of T4-like bacteriophages in the fraction greater than 0.2 μm from the pelagic zone, coastal zone and shallow bays of Lake Baikal by the gene fragment of the major capsid protein, gp23. High throughput sequencing allowed us to obtain from 12454 to 41802 sequences of the *g23* gene in the fraction associated with bacteria. The results revealed that the sequences found in this study together with the sequences that we had previously retrieved from the plankton samples (fraction less than 0.2 μm) and biofilms (without separation onto fractions) formed the Baikal cluster. The sequences from shallow bays largely differed from those in the pelagic and coastal samples and formed individual subcluster in the UPGMA tree. According to the RefSeq database, most OTUs had the cultivated closest relatives belonging to cyanophages.

Introduction

Viruses are obligate intracellular parasites consisting of a single-stranded or double-stranded RNA or DNA molecule enclosed within a protein capsid; enveloped viruses have an additional membrane envelope (supercapsid). Viruses are distinguished by a huge number and high genetic diversity (Suttle, 2007), thereby representing an inexhaustible pool for research. To date, according to ICTV Master Species List (ICTV, 2019), 3973 species belong to DNA viruses and 2617 – to RNA viruses, but most of the sequences obtained to date from viromes are known to be “viral dark matter“ (Krishnamurthy & Wang, 2017), and there are much more real biological species of viruses (Gregory et al., 2019).

The bulk of the sequences obtained by metagenomic sequencing of DNA-containing viruses in aquatic ecosystems, which can be identified from databases, belongs to the order Caudovirales (Cai et al., 2016; Garin-fernandez et al., 2018; Gong et al., 2018; Taboada et al., 2018; Gregory et al., 2019; Wu et al., 2020). Due to the rapid transformation of viral taxonomy, the order Caudovirales expanded to 14 families. Among them, the family *Myoviridae* is the best known and most studied. It includes 8 subfamilies and 153 genera (https://talk.ictvonline.org/taxonomy/).

The family *Myoviridae* contains DNA phages that are genetically and morphologically similar to the well-studied coliphage T4 (Ackermann & Krisch, 1997). At the same time, myoviruses have a relatively wide range of hosts (Sullivan, Waterbury & Chisholm, 2003).

Due to the lack of universal genes in viruses, signature genes are studied using primers targeting a specific group. The *g23* gene fragment encoding major capsid protein is the most reliable marker for the analysis of the biodiversity of T4-like phages of the family *Myoviridae* (Tetart et al., 2001; Adriaenssens & Cowan, 2014). Based on the analysis of gp23 sequences, T4-like phages are divided into several groups: “true” T-evens represented by bacteriophage T4 and closely related phages infecting enterobacteria (e.g. T2, T6), PseudoT-evens and SchizoT-evens (phages of the genera *Aeromonas*, *Vibrio*, etc.) as well as more distant ExoT-evens (cyano- and pelagiphages, etc.) (Desplats & Krisch, 2003). Moreover, T4-like viruses are divided into three subgroups: Far T4 (including *Rhodothermus* phage RM378 (Hjorleifsdottir et al., 2014)), Near T4 (including T-evens, PseudoT-evens and SchizoT-evens) and Cyano T4 (including ExoT-evens) (Comeau & Krisch, 2008).

It is widely acknowledged that viruses play a global role in the Earth’s biosphere, influencing many ecological processes and biogeochemical cycles (Suttle, 2007). In the aquatic environment, viruses are an important factor in the regulation of the number and structure of microbial communities (Kutter & Sulakvelidze, 2005). In freshwater ecosystems such as Lake Biwa (Japan), the percentage of daily bacterial production destroyed by viruses was estimated as high and accounted for 52.7 ± 16.2% in the upper layer and 13.6 ± 5.2% in the deeper layer (Pradeep Ram et al., 2010). In Lake Pavin (France), the average seasonal contribution from bacteriophages to bacterial lysis reached 16.2% (Sime-Ngando et al., 2016).

Currently, the diversity of T4-like phages in the viral fraction (less than 0.4 and 0.2 μm) is mainly studied (López-Bueno et al., 2009; Butina et al., 2010; Jamindar et al., 2012; Parvathi, Zhong & Jacquet, 2012; Bellas & Anesio, 2013; Goldsmith et al., 2015; Wang et al., 2015; Millard, Pearce & Zwirglmaier, 2016; Liu, Cai & Zhang, 2017; Potapov et al., 2018). Organisms larger than 0.2 or 0.4 μm (bacterial fraction) are removed using various methods because it is methodologically more preferable to work with a viral fraction that does not contain bacterial cells. Information about the composition and role of viruses associated with microbial communities is scarce and covered mainly in metagenomic studies (De Cárcer et al., 2016; Zeigler Allen et al., 2017; Aylward et al., 2017; Palermo et al., 2019; Okazaki et al., 2019; Coutinho et al., 2020).

Previously, it was shown that the filtration of samples through filters with a pore size of 0.2 μm reduces the number of phages in the filtrate by an average of two-thirds (transmission electron microscopy counting method) (Paul, Jiang & Rose, 1991). In a later study, the proportion of viruses retained on a 0.2 μm filter did not exceed 15% of the overall number of virus-like particles (epifluorescence microscopy counting method) (Auguet, Montanié & Lebaron, 2006). For lytic phages including myoviruses, three states of life cycle were described: i) free during extracellular search; ii) located in a certain space with no host, which is, for example, associated with an inert particle; iii) actively infects bacteria (Kutter & Sulakvelidze, 2005). T4-like viruses having strong lytic properties likely experience rapid exchange of intracellular and free phages. For instance, phages T4 and λ have latent periods of 20 min and 50 min, respectively (De Paepe & Taddei, 2006). Owing to the rapid change in the phage state from free to associated and vice versa, determination of solely viral fraction (less than 0.2 μm) can distort the results of genetic analysis of diversity. Furthermore, the study of a bacterial fraction can reveal the relationship between bacteria and bacteriophages infecting them, this will give an understanding about the phages that are in the propagation stage.

This study aims to check the difference of the phage sequences i) from samples with a viral fraction, ii) without separation into fractions and iii) from phages with a bacterial fraction. Moreover, based on all obtained *g23* sequences and using cluster analysis, we try to understand the influence of the geographical distance on the diversity of bacteriophages in Lake Baikal and compare them with other ecosystems.

Materials & Methods

Sampling sites.

Samples were collected at five sites of Lake Baikal in August 2019. The sampling sites were as follows: Mukhor Bay near the Kuchelga River (MK), a centre of Mukhor Bay (MC), a coastal zone near the Turka (Turk) settlement (328 m off the coast), and Posolsk Sor Bay (Posol_S); the sampling depth at these stations was 0 m. At the central station of the Listvyanka settlement – the Tankhoy settlement section, sampling was carried out in the layer from 0 to 5 m (LT_05) and from 10 to 15 m (LT_1015) (the sample was taken with a Niskin bathometer). Figure 1 shows the sampling map. Coordinates are shown in Supplementary Table 1.

Figure 1. Map of the sampling area.

Water Chemistry Analysis

For determinations of chlorophyll *a* concentration, 1 L of water was filtered through a 0.4 μm polycarbonate filter (Sartorius, Germany). Algal pigments were extracted with acetone (90%) overnight in the dark at 4 °C (after ultrasonic treatment). The supernatant was centrifuged and chlorophyll *a* was measured with a Cintra-2020 spectrophotometer (GBC, Scientific Equipment,

Australia) at 664, 647 and 630 nm and calculated based on equations provided by Parsons et al. (Parsons, Maita & Lalli, 1984).
Total phosphorus content was determined using a photoelectric colorimeter KFK-2, (ZOMZ, Zagorskii optiko-mekhanicheskii zavod [Zagorskii optical mechanics factory], Russia) after persulfate oxidation. Total nitrogen content was determined by persulfate oxidation in an alkaline medium using a spectrophotometer PE-5300VI (Ekroskhim (previously "Ekohim"), Russia) (Wetzel & Likens, 2000).

DNA extraction and preparation of amplicons

Water samples (1 liter) from each site (MK, MC, Turk, Posol_S, LT_05, LT_1015) were filtered through sterile polycarbonate filters with a pore size of 0.2 µm (Sartorius, Germany) without using prefilters, and they were frozen onboard the research vessel during the expeditions. DNA was extracted by the standard phenol-chloroform method in the laboratory. Primers MZIA1bis and MZIA6 were used (Filée et al., 2005). The PCR mixture consisted of the following components: Master mix 2x Taq M (Alkor Bio, Russia), 0.1 µM primers, nuclease-free water, and DNA template. PCRs were performed with the following PCR cycle parameters: denaturation at 95°C for 15 min, 30 cycles of denaturation at 95°C for 30 s, annealing at 50°C for 30 s, extension at 72°C for 1 min, and a final extension at 72°C for 10 min. DNA was purified from the PCR mixture using a suspension of magnetic particles CleanMag DNA (Evrogen, Russia). Library preparation and sequencing on Illumina MiSeq 2*300 were performed in the "Genomics Core Facility" (ICBFM SB RAS, Novosibirsk, Russia).

Bioinformatic analysis

Sequence quality analysis was carried out using FastQC software tool (Andrews, 2010). Trimming was performed using the Trimmomatic v. 0.36 tool (Bolger, Lohse & Usadel, 2014) using the following parameters: SLIDINGWINDOW:4:20 LEADING:3 TRAILING:3 MINLEN:50. Further processing was performed using the Usearch v. 11.0.667 tool (Edgar, 2010). Paired-end reads were combined using the *-fastq_mergepairs* command. The primers were not removed because, in contrast to bacterial sequences, viral sequences are degenerate and thus carry information about diversity. Then unique sequences (*-derep_fulllength*) were sorted out. The next step was clustering at the 90% identity level, UPARSE-OTU algorithm (*-cluster_otus*), as well as the removal of chimeras, singletons and doubletons. The chosen level was previously substantiated (Potapov et al., 2018). Nucleotide sequences were converted into amino acid ones using the BioEdit v. 7.0.9.0 program (Hall, 1999). The annotation was performed using the BLASTP analysis (default expect threshold) based on the RefSeq and GenBank (NR) databases.
Amino acids were aligned in the Mega 7 software (Kumar, Stecher & Tamura, 2016) using the Muscle algorithm. A phylogenetic tree was constructed through Bayesian analysis using the MrBayes software (v. 3.2.6). Two independent Markov chain Monte Carlo (MCMC) analyses were launched for 15 million generations with 25% burn-in (rejection of initial generations) and

four chains (one cold and three hot ones). All calculations were performed on HPC-cluster "Akademik V.M. Matrosov" ("Irkutsk Supercomputer Center of SB RAS, http://hpc.icc.ru"). Based on the amino acid sequences, the distance matrix was obtained through the unweighted UniFrac metric (multiple sequence alignment – Muscle, model - Blosum62, normalized=TRUE), followed by a hierarchical cluster analysis (hclust) (Murtagh, 1992) method "average" (=UPGMA), using phyloseq (v. 1.21.0) and phangorn (v. 2.2.0) packages, implemented in the R software (v. 3.2.4).

Non-metric multidimensional scaling (NMDS) was based on abundance of *g23* amino acid sequences and physicochemical parameters. We used an unweighted (qualitative) UniFrac method with the phyloseq (v. 1.21.0), phangorn (v. 2.2.0) and vegan (v. 2.4-3) packages implemented in the R software (v. 3.2.4).

Nucleotide diversity values were calculated using DNASP v. 6.12 (Rozas et al., 2017). The sequences were aligned by MUSCLE v. 3.8.1551 with default settings (Edgar, 2004).

The intersection graph was constructed using the UpSetR (v. 1.4.0) package (Lex et al., 2014) implemented in the R software, combination matrix – intersect. The OTU table generated in QIIME v. 1.9.1 (Caporaso et al., 2010) (make_otu_table.py) based on the OTU sequences obtained from Usearch served as the input file.

Results

Environment parameters

Table 1 shows the results of physicochemical analysis. According to the indicators given in the table, the productivity of shallow bays corresponds to the mesotrophic status, and the productivity of the waters of the pelagic stations corresponds to the oligotrophic status according to R. A. Vollenweider (Vollenweider & Kerekes, 1982).

Table 1:
Physical and chemical parameters of water.

An NMDS plot with samples from this study and physicochemical parameters are shown in Figure 2. Chlorophyll *a* and total nitrogen are positively associated with the sample from Posolsk Sor Bay (Posol_S); the water temperature, pH and total phosphorus – with the samples from Mukhor Bay (MC, MK). All parameters have a significance level of less than 0.011. The highest concentration of the total phosphorus, total nitrogen and chlorophyll *a* has been recorded in shallow bays, which is not surprising for Lake Baikal because the maximum rate of formation of organic matter occurs here due to a large number of primary producers (cyanobacteria) (Watanabe & Drucker, 1999; Belykh & Sorokovikova, 2003). Temperature, in turn, is an important factor, influencing the growth rate of bacteria and having a significant positive effect on bacterial production (Straškrábová et al., 2005). An increase in temperature stimulates the development of phytoplankton, enhances its photosynthetic activity, and the water body is

enriched with dissolved organic matter. The pH level is slightly shifted to the alkaline side, which is typical for the waters of Lake Baikal. As shown previously, a key factor in determining the infectivity of a virus is the pH value. For example, a low pH (<4) significantly reduces phage survival (Jurczak-Kurek et al., 2016).

Figure 2. NMDS analysis based on *g23* sequences and physicochemical parameters. pH - potential of hydrogen, P – total phosphorus, Chl – chlorophyll *a*, N – total nitrogen.

Analysis of *g23* sequences

Overall, we obtained 250 representative viral sequences (OTU) of the *g23* gene fragment based on the 97% clustering. The lengths of the sequences at the nucleotide level, taking into account the primers, ranged from 355 to 547 nucleotides (average length 439 nuc.). The main data obtained from processing during the preparation of sequences are shown in Table 2.

Table 2:
Summary information about the sampling site and results of the processing stages.
Sampling date: August 2019

Among the cultivated bacteriophages, the highest identity is at amino acid level (from 64.4 to 77.9%), and the minimum e-value was with cyanophages: *Synechococcus* phage S-SSM7, *Synechococcus* phage S-SM2, *Synechococcus* phage Bellamy, and *Synechococcus* phage S-CAM1 (Table 3). The bulk of the sequences was annotated as belonging to cyanophages. This fact likely indicates the active infection of their hosts with this group of viruses in the study period and reflects the associated interaction of phages with cyanobacteria. In summer, especially in early August, there is a seasonal peak of cyanobacterial bloom in Lake Baikal (Belykh & Sorokovikova, 2003). *Pelagibacter* phage HTVC008M was the second closest relative in terms of the frequency of occurrence: MK – 13.7%, MC – 29.2%, Turk – 20%, Posol_S – 10.8%, LT_05 – 16.7%, and LT_1015 – 14.3%. Moreover, *Serratia* phage BF (YP_009599751), *Agrobacterium* phage Atu_ph07 (YP_009611880), *Caulobacter* phage Cr30 (YP_009098938), *Sinorhizobium* phage phiN3 (YP_009212304), *Acidovorax* phage ACP17 (YP_009609699), and *Cronobacter* phage vB_CsaM_leB (YP_009831235) are among the closest relatives.

Table 3:
The closest relatives with the lowest e-value for each sample (database RefSeq, amino acid level).

The protein sequences had uncultivated relatives from various ecosystems (Table 4), and most relatives from the GenBank database were similar to the Baikal sequences that had been previously obtained from biofilms (from 19.6 to 28.6%) and pelagic zone (from 6.8 to 30%) of

Lake Baikal. The largest number of the closest relatives for samples LT_05, LT_1015 and Turk was from lakes Bourget and Annecy (from 10.8 to 25.7%). An interesting result is the identity of a large number of sequences from small eutrophic bays (MK, MC and Posol_S) with sequences from wetland sediment (from 18.9 to 25.5%) (Li et al., 2018), whereas pelagic representatives from samples Turk, LT_05 and LT_1015 had only from 2.8 to 3.3% of similar sequences.

Table 4:
The number of sequences in this study similar to the sequences from other sources, % (database GenBank, amino acid level).

To assess alpha diversity, we conducted a comparative analysis of sequences based on nucleotides (nucleotide diversity, π) (Table 5). In addition to the sequences from this study, the sequences were taken from various sources available in the NCBI database. Based on microdiversity data, among the sequences obtained in this study, Posol_S is the least diverse *g23* community in Lake Baikal, and MK is the most diverse one. In contrast, the sequences from the polar Lake Limnopolar and dairy water (Ireland) were distinguished by a larger nucleotide diversity. In general, noteworthy is the high nucleotide diversity of the Baikal sequences.

Table 5:
Nucleotide diversity of the *g23* gene fragment.

Figure 3 shows a graph of OTU intersections between samples. Samples LT_05 and LT_1015 have the largest number of intersections (31), which is natural because they were taken from different layers at the same station. The larger number of intersections with the LT_05 (23), LT_1015 (20) and Turk (14) sequences is typical of the BSOTU (Baikal Sample OTU, pelagic water) sequences (Potapov et al., 2018). Samples MK and MC have 22 intersections. It was expectable, as these samples were taken from the same bay (the Maloye More Strait). Posol_S and MK have 12 shared OTUs. Therefore, the sequences from the pelagic zone have a common pool of similar sequences, and those from bays are more similar to each other. The greater identity of LT_05 and LT_1015 with sample Turk (near the Turka settlement) is likely because, on the day of the expedition, the water of the littoral sample Turk collected 328 m from the coast was mixed with pelagic waters during a storm caused by north-westerly wind.

Figure 3. UpSetR plot created by representative nucleotide sequences of OTUs. The dots indicate the total OTU count in a sample; the dots connected by lines – shared OTUs. Intersection size ≥ 8. Mode = "intersect".

Phylogenetic analysis

Phylogenetic analysis of the amino acid *g23* sequences with the cultivated representative T4-like viruses of the family *Myoviridae* and with the sequences from various natural sources revealed that the sequences from this study are mixed in the tree and do not form separate clades consistent with fraction or ecotope (Fig. 4). We conditionally divided them into 13 clusters (marked with Roman numerals). The cluster was considered formed if there were more than three sequences. The bulk of the clusters contained *g23* sequences of phages from natural sources, whose hosts are still unknown because there are no cultivated representatives and hence the confirmation of phylogenetic affiliation.

Cluster I included 51 OTUs from all samples in our study. This cluster had the only cultivated bacteriophage, *Caulobacter* phage Cr30, obtained from the culture of Gram-negative oligotrophic bacteria, *Caulobacter crescentus*, widespread in freshwater lakes.

Cluster II (37 OTUs) contained four cultivated representatives: the thermophilic bacterium phage RM378 assigned to the group Far T4; *Serratia* phage BF obtained from the strain of the *Serratia marcescens* UCC2017 opportunistic bacteria; *Escherichia* phage 121Q, the laboratory host of which is *Escherichia coli* MuLB70.1; *Agrobacterium* phage Atu ph07 isolated from the *Agrobacterium tumefaciens* bacteria opportunistic for humans (Adnan et al., 2013). This cluster probably contained phages of opportunistic bacteria. Cluster III was represented by the group Exo T-evens with isolates of picocyanobacterial and SAR11 phages. Clusters IV-IX, XI, and XII did not contain any cultivated phages and represented a mixture of sequences from various ecosystems. For example, Cluster VI contained a single sequence from the marine ecosystem. Interestingly, the previous Baikal *g23* sequences were closely related to marine T4 cyanophages (Butina et al., 2010), which is probably due to the filling of the databases with the studies on freshwater ecosystems. Cluster X (38 OTUs) consisted of one cultivated representative, *Acidovorax* phage ACP17 (Rahimi-Midani et al., 1970), infecting the *Acidovorax citrulli* bacteria. Cluster XIII comprised phages of pathogenic and opportunistic bacteria belonging to the group Near T4.

Two OTUs (Posol_S_OTU59 and Posol_S_OTU37) were not included in any cluster as well as *Sinorhizobium* phage phiN3 represented by a separate branch.

Figure 4. Bayesian phylogenetic tree based on alignment of 388 gp23 major capsid protein sequences. Dots mark cultivated bacteriophages.

Biogeography of gp23 sequences

For comparative biogeographical analysis we selected sequences available in the GenBank database: Kongsfjorden, proglacial lake (Svalbard, Norway) (Bellas & Anesio, 2013), coral colony (Orpheus Island, Australia) (Buerger et al., 2018), hydrothermal vent (East Scotia Ridge) (Millard, Pearce & Zwirglmaier, 2016), Chesapeake Bay (USA) (Jamindar et al., 2012), Lake Baikal (Russia) (Butina et al., 2010; Potapov et al., 2018, 2020), Lake Annecy and Bourget (France) (Zhong & Jacquet, 2014), wetland (China) (Zheng et al., 2013), Lake Kotokel (Russia) (Butina et al., 2013), Lake Donghu (China) (Huang, Cheng & Xu, 2011), Lake East (China)

(Wang et al., 2015), Lake Limnopolar (Antarctica, Livingston Island) (López-Bueno et al., 2009), dairy wastewater (Ireland) (Knapik & Prentice, 2012), the Selenga River (Russia) (Butina et al., 2015), wetland sediments (China) (Li et al., 2018), sediments of the Pearl River estuary (China) (He et al., 2017), paddy field (China) (Li et al., 2019), and paddy field (Japan) (Cahyani et al., 2009) (Fig. 5).

Figure 5. Map showing the samples included in the analysis for this study.

The *g23* sequences in this study and those represented in the GenBank database formed three groups on the dendrogram: marine, freshwater, soil, and sediments (Fig. 6). LT_05 and LT_1015, sequences from the samples collected at the central station of the Listvyanka settlement – the Tankhoy settlement section, layer from 0 to 5 m and from 10 to 15 m, are located closer to the site with the sequences from sample Turk as mentioned above, which is due to the mixing with pelagic waters. LT_05, LT_1015 and Turk form a common cluster together with the sequences previously obtained from the pelagic zone of Lake Baikal (BSOTU) (Potapov et al., 2018) and the sequences from sponge biofilms (Potapov et al., 2020). Three samples from shallow Baikal bays (MK, MC and Posol_S) group together, confirming clustering by the similar trophic conditions within the Baikal group.

Figure 6. Cluster dendrogram (UPGMA). Samples from this study are marked in bold. SB – southern basin, Lake Baikal, NB – northern basin, Lake Baikal, BSOTU – Baikal samples OTU, LAB – lakes Bourget and Annecy. Untagged samples are either unpublished or have no open access articles.

The sequences from biofilms of stones, which we named GreenStone and CyaStone (Potapov et al., 2020), form a joint cluster with planktonic *g23* sequences from the pelagic zone of the southern basin of Lake Baikal. Samples from Lake Kotokel, Lake East, Lake Donghu, and the wetland of China form a cluster that we designated as eutrophic.
The greatest identity of the Baikal *g23* is observed with the sample from subalpine lakes Bourget (oligo-mesotrophic) and Annecy (oligotrophic) (Zhong & Jacquet, 2014), which we also previously indicated in our study (Potapov et al., 2018).
The sequences from the Arctic proglacial lakes and Lake Limnopolar are closer to the sequences from the Selenga River and dairy water (Ireland).
Despite the difference in the sequencing method and fraction greater than 0.2 μm, the sequences in this study are located in the Baikal cluster with the sequences obtained from the fraction less than 0.2 μm and without separation into fractions.

Discussion

Here, we studied the *g23* communities of bacteriophages associated with bacteria in a fraction greater than 0.2 μm from shallow bays, coastal zone and pelagic zone of Lake Baikal. The sampling sites were chosen not only to carry out a comparative analysis in terms of

geographical distance but also to reveal the differences in the pools of *g23* sequences selected in different zones of Lake Baikal with different trophic statuses.

The shallow bay of Posolsk Sor Bay (maximum depth 3.5 m) is located on the southeast coast of Lake Baikal, about 20 km south of the Selenga River delta. Mukhor Bay is one of the warmest and shallowest bays located in the Maloye More Strait that is situated between the mainland and Olkhon Island. Both bays warm well in July and August and are the most visited tourist sites of Lake Baikal. In August 2019, there was a mass development of diazotrophic cyanobacteria in the waters of the bays: *Gloeotrichia echinulata* bloomed in Posolsk Sor Bay, and the species of the genus *Dolichospermum* – in Mukhor Bay.

Late July and early August is a period of bloom of both pico- and nanoplanktonic cyanobacteria in Lake Baikal. Picocyanobacteria are found in huge numbers in Lake Baikal, reaching an abundance of 1.5 million cells/ml (Belykh & Sorokovikova, 2003). Intensive vegetation of *Dolichospermum* species has been recorded for a long time from June to September in all parts of Lake Baikal, with the maximum concentration of up to 10 million cells/l in bays (Popovskaya, 2000). In this regard, it is logical to assume the presence of cyanophages, natural regulators of the number of cyanobacteria. As mentioned above, many sequences belonged to cyanophages according to RefSeq, but phylogenetic analysis revealed only one cluster with cultivated cyanophages. Eight clusters did not have cultivated representatives; therefore, their taxonomic identification is unknown. They may be similar to cultivated Baikal cyanophages that have not been obtained so far.

In addition to cyanophages, *Pelagibacter* phage HTVC008M (the family *Myoviridae*) infecting *Candidatus Pelagibacter* ubique (Alphaproteobacteria, freshwater SAR11) was the closest relative of the Baikal representatives of T4-phages. Previously, high synteny with this strain of the Baikal sequences was determined (Cabello-Yeves et al., 2018).

Noteworthy is a great number of similar sequences in samples from shallow areas of Lake Baikal (from 18.9 to 25.5%) and wetland sediment (Li et al., 2018), which is much greater than in the samples from the pelagic zone of Lake Baikal and wetland sediments (from 2.9 to 3.3%). This possibly indicates the similar composition and conditions for the existence of bacterial communities from the Baikal bays (shallow water, high productivity and elevated temperature) and the shallow, well-warmed productive wetland sediments.

A large number of *g23* sequences from lakes Bourget and Annecy similar to the Baikal sequences in this study is likely owing to the identity of the lakes' hydrophysical and hydrochemical parameters (altitude above sea level, total P, total N, nitrates and pH), as we previously indicated in the analysis of *g23* sequences from the pelagic zone of Lake Baikal (Potapov et al., 2018).

Although the cluster analysis compared sequences obtained by the Sanger method and high throughput sequencing (HTS), as well as in two different fractions, they were separated by the trophic status of the study sites of Lake Baikal and other water bodies. Therefore, according to our previous studies (Butina et al., 2013), the *g23* communities form regions in the tree

depending on the productivity of the waters. Perhaps, trophic status of water determines the composition of both bacteria and phages.

UPGMA analysis revealed that viral diversity does not follow the latitudinal gradient, and the geographical distance does not influence the composition of bacteriophages according to the "everything is everywhere, but the environment selects" theory. At the same time, it corresponds to one of the distribution models: the limitation of distribution and environmental conditions determine the community composition. This conclusion confirms the isolated formation of viral *g23* communities and, possibly, endemism of bacteriophages of the Earth's oldest lake. Analysis *g23* genes indicates that separation by fractions is not of key importance, and the quantitative loss of bacteriophages during filtration does not significantly change the biodiversity of communities. At the same time, a limited dataset was used in the analysis; thereby further research is required.

Conclusions

In this study, we obtained *g23* sequences of bacteriophages from the pelagic and coastal zones and bays of Lake Baikal, in a fraction greater than 0.2 μm. Our results revealed a high diversity of *g23* communities of bacteriophages of the family *Myoviridae* in Lake Baikal. Cluster analysis confirmed the uniqueness of the Baikal sequences, as evidenced by their grouping with each other rather than with sequences from other ecosystems. Viral communities from different aquatic ecosystems rather cluster by the productivity of the water bodies than by their geographical confinement.

Acknowledgements

The Authors are grateful to the crew of R/V "G. Titov" for their assistance in sampling. We thank Andrey Fedotov and Yuliya Vitushenko for their helpful advice. The authors are grateful to Irkutsk Supercomputer Center SB RAS for providing the access to HPC-cluster "Akademik V. M. Matrosov" (MrBayes calculations).

References

Ackermann H-W, Krisch HM. 1997. A catalogue of T4-type bacteriophages. Archives of Virology 142:2329–2345. DOI: 10.1007/s007050050246.
Adnan M, Khan S, Patel M, Al-Shammari E, Ashankyty IMA. 2013. Agrobacterium: a potent human pathogen. Reviews in Medical Microbiology 24:94–97. DOI: 10.1097/MRM.0b013e3283642449.
Adriaenssens EM, Cowan DA. 2014. Using signature genes as tools to assess environmental viral ecology and diversity. Applied and Environmental Microbiology 80:4470–4480. DOI: 10.1128/AEM.00878-14.
Andrews S. 2010.FastQC: a quality control tool for high throughput sequence data. Available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Auguet JC, Montanié H, Lebaron P. 2006. Structure of virioplankton in the charente estuary (France): Transmission electron microscopy versus pulsed field gel electrophoresis. In: Microbial Ecology. Springer, 197–208. DOI: 10.1007/s00248-005-0043-0.
Aylward FO, Boeuf D, Mende DR, Wood-Charlson EM, Vislova A, Eppley JM, Romano AE, DeLong EF. 2017. Diel cycling and long-term persistence of viruses in the ocean's euphotic zone. Proceedings of the National Academy of Sciences of the United States of America 114:11446–11451. DOI: 10.1073/pnas.1714821114.
Bellas CM, Anesio AM. 2013. High diversity and potential origins of T4-type bacteriophages on the surface of Arctic glaciers. Extremophiles 17:861–870. DOI: 10.1007/s00792-013-0569-x.
Belykh OI, Sorokovikova EG. 2003. Autotrophic picoplankton in Lake Baikal: abundance, dynamics, and distribution. Aquatic Ecosystem Health & Management 6:251–261. DOI: 10.1080/14634980301489.
Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England) 30:2114–2120. DOI: 10.1093/bioinformatics/btu170.
Buerger P, Weynberg KD, Wood-Charlson EM, Sato Y, Willis BL, van Oppen MJH. 2018. Novel T4 bacteriophages associated with black band disease in corals. Environmental Microbiology. DOI: 10.1111/1462-2920.14432.
Butina TV, Belykh OI, Maksimenko SY, Belikov SI. 2010. Phylogenetic diversity of T4-like bacteriophages in Lake Baikal, East Siberia. FEMS Microbiology Letters 309:122–129. DOI: 10.1111/j.1574-6968.2010.02025.x.
Butina TV, Belykh OI, Potapov SA, Sorokovikova EG. 2013. Diversity of the major capsid genes (g23) of T4-like bacteriophages in the eutrophic Lake Kotokel in East Siberia, Russia. Archives of Microbiology 195:513–520. DOI: 10.1007/s00203-013-0884-8.
Butina TV, Usova OS, Potapov SA, Belykh OI, Fedotov AP, Belikov SI. 2015. Molecular genetic study of T4-like bacteriophages in plankton of Selenga River. Seriya «Biologiya. Ekologiya» of the «Izvestiya Irkutskogo gosudarstvennogo universiteta» 12:12–22.
Cabello-Yeves PJ, Zemskay TI, Rosselli R, Coutinho FH, Zakharenko AS, Blinov V V., Rodriguez-Valera F. 2018. Genomes of novel microbial lineages assembled from the sub-ice waters of Lake Baikal. Applied and Environmental Microbiology 84. DOI: 10.1128/AEM.02132-17.

Cahyani VR, Murase J, Ishibashi E, Asakawa S, Kimura M. 2009. T4-type bacteriophage communities estimated from the major capsid genes (g23) in manganese nodules in Japanese paddy fields. Soil Science and Plant Nutrition 55:264–270. DOI: 10.1111/j.1747-0765.2009.00363.x.

Cai L, Zhang R, He Y, Feng X, Jiao N. 2016. Metagenomic analysis of Virioplankton of the subtropical Jiulong river estuary, China. Viruses 8:1–13. DOI: 10.3390/v8020035.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley G a, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone C a, Mcdonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters W a, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data Intensity normalization improves color calling in SOLiD sequencing. Nature Publishing Group 7:335–336. DOI: 10.1038/nmeth0510-335.

De Cárcer DA, Pedrós-Alió C, Pearce DA, Alcamí A. 2016. Composition and interactions among bacterial, microeukaryotic, and T4-like viral assemblages in lakes from both polar zones. Frontiers in Microbiology 7:1–11. DOI: 10.3389/fmicb.2016.00337.

Comeau AM, Krisch HM. 2008. The capsid of the T4 phage superfamily: The evolution, diversity, and structure of some of the most prevalent proteins in the biosphere. Molecular Biology and Evolution 25:1321–1332. DOI: 10.1093/molbev/msn080.

Coutinho FH, Cabello-Yeves PJ, Gonzalez-Serrano R, Rosselli R, López-Pérez M, Zemskaya TI, Zakharenko AS, Ivanov VG, Rodriguez-Valera F. 2020. New viral biogeochemical roles revealed through metagenomic analysis of Lake Baikal. Microbiome 8:163. DOI: 10.1186/s40168-020-00936-4.

Desplats C, Krisch HM. 2003. The diversity and evolution of the T4-type bacteriophages. Research in Microbiology 154:259–267. DOI: http://dx.doi.org/10.1016/S0923-2508(03)00069-X.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792–1797. DOI: 10.1093/nar/gkh340.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461. DOI: 10.1093/bioinformatics/btq461.

Filée J, Tétart F, Suttle CA, Krisch HM. 2005. Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. Proceedings of the National Academy of Sciences of the United States of America 102:12471–6. DOI: 10.1073/pnas.0503404102.

Fujihara S, Murase J, Tun CC, Matsuyama T, Ikenaga M, Asakawa S, Kimura M. 2010. Low diversity of T4-type bacteriophages in applied rice straw, plant residues and rice roots in Japanese rice soils: Estimation from major capsid gene (g23) composition. Soil Science and Plant Nutrition 56:800–812. DOI: 10.1111/j.1747-0765.2010.00513.x.

Garin-fernandez A, Pereira-flores E, Oliver F, Wichels A. 2018. Marine Genomics The North Sea goes viral : Occurrence and distribution of North Sea bacteriophages. Marine Genomics 41:31–41. DOI: 10.1016/j.margen.2018.05.004.

Goldsmith DB, Brum JR, Hopkins M, Carlson CA, Breitbart M. 2015. Water column stratification structures viral community composition in the Sargasso Sea. Aquatic Microbial Ecology 76:85–94. DOI: 10.3354/ame01768.

Gong Z, Liang Y, Wang M, Jiang Y, Yang Q, Xia J, Mcminn A. 2018. Viral diversity and its relationship with environmental factors at the surface and deep sea of Prydz Bay , Antarctica. 9:1–17. DOI: 10.3389/fmicb.2018.02981.

Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, Ardyna M, Arkhipova K, Carmichael M, Cruaud C, Dimier C, Domínguez-Huerta G, Ferland J, Kandels S, Liu Y, Marec C, Pesant S, Picheral M, Pisarev S, Poulain J, Tremblay JÉ, Vik D, Acinas SG, Babin M, Bork P, Boss E, Bowler C, Cochrane G, de Vargas C, Follows M, Gorsky G, Grimsley N, Guidi L, Hingamp P, Iudicone D, Jaillon O, Kandels-Lewis S, Karp-Boss L, Karsenti E, Not F, Ogata H, Poulton N, Raes J, Sardet C, Speich S, Stemmann L, Sullivan MB, Sunagawa S, Wincker P, Culley AI, Dutilh BE, Roux S. 2019. Marine DNA viral macro- and microdiversity from pole to pole. Cell 177:1109–1123.e14. DOI: 10.1016/j.cell.2019.03.040.

Hall T. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41:95–98. DOI: citeulike-article-id:691774.

He M, Cai L, Zhang C, Jiao N, Zhang R. 2017. Phylogenetic diversity of T4-type phages in sediments from the subtropical Pearl River Estuary. Frontiers in Microbiology 8. DOI: 10.3389/fmicb.2017.00897.

Hjorleifsdottir S, Aevarsson A, Hreggvidsson GO, Fridjonsson OH, Kristjansson JK. 2014. Isolation, growth and genome of the Rhodothermus RM378 thermophilic bacteriophage. Extremophiles 18:261–270. DOI: 10.1007/s00792-013-0613-x.

Huang H-Z, Cheng K, Xu M. 2011. Genetic diversity of T4 virioplankton, inferred from g23 gene, in Wuhan Donghu Lake. CHINA ENVIRONMENTAL SCIENCECE 31:443–447.

ICTV. 2019.Master Species List 2019.v1. Available at https://talk.ictvonline.org/files/master-species-lists/m/msl/9601

Irkutsk Supercomputer Center of SB RAS, http://hpc.icc.ru.

Jamindar S, Polson SW, Srinivasiah S, Waidner L, Wommack KE. 2012. Evaluation of two approaches for assessing the genetic similarity of virioplankton populations as defined by genome size. Applied and Environmental Microbiology 78:8773–8783. DOI: 10.1128/AEM.02432-12.

Jia Z, Ishihara R, Nakajima Y, Asakawa S, Kimura M. 2007. Molecular characterization of T4-type bacteriophages in a rice field. Environmental Microbiology 9:1091–1096. DOI: 10.1111/j.1462-2920.2006.01207.x.

Jurczak-Kurek A, Gąsior T, Nejman-Faleńczyk B, Bloch S, Dydecka A, Topka G, Necel A, Jakubowska-Deredas M, Narajczyk M, Richert M, Mieszkowska A, Wróbel B, Węgrzyn G, Węgrzyn A. 2016. Biodiversity of bacteriophages: morphological and biological properties of a large group of phages isolated from urban sewage. Scientific Reports 6:34338. DOI: 10.1038/srep34338.

Kavagutti VS, Andrei A-Ş, Mehrshad M, Salcher MM, Ghai R. 2019. Phage-centric ecological interactions in aquatic ecosystems revealed through ultra-deep metagenomics. Microbiome 7:135. DOI: 10.1186/s40168-019-0752-0.

Knapik K, Prentice MB. 2012.Wide bacteriophage diversity in a single terrestrial aquatic site. Available at https://www.ncbi.nlm.nih.gov/popset?LinkName=protein_popset&from_uid=396575579

Krishnamurthy SR, Wang D. 2017. Origins and challenges of viral dark matter. Virus Research 239:136–142. DOI: 10.1016/j.virusres.2017.02.002.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular biology and evolution 33:msw054. DOI: 10.1093/molbev/msw054.

Kutter E, Sulakvelidze A. 2005. Bacteriophages: Biology and Applications. Boca Raton-London-New York- Washington: CRC Press.

Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. 2014. UpSet: Visualization of intersecting sets. IEEE Transactions on Visualization and Computer Graphics 20:1983–1992. DOI: 10.1109/TVCG.2014.2346248.

Li Y, Liu H, Pan H, Zhu X, Liu C, Zhang Q, Luo Y, Di H, Xu J. 2019. T4-type viruses: Important impacts on shaping bacterial community along a chronosequence of 2000-year old paddy soils. Soil Biology and Biochemistry 128:89–99. DOI: 10.1016/j.soilbio.2018.10.007.

Li X, Sun Y, Liu J, Yao Q, Wang G. 2018.Myoviridae environmental samples. Available at https://www.ncbi.nlm.nih.gov/popset?LinkName=protein_popset&from_uid=1496533459

Liu L, Cai L, Zhang R. 2017. Co-existence of freshwater and marine T4-like myoviruses in a typical subtropical estuary. FEMS Microbiology Ecology 93. DOI: 10.1093/femsec/fix119.

López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcamí A. 2009. High diversity of the viral community from an Antarctic lake. Science 326:858–861. DOI: 10.1126/science.1179287.

Mabizela NB, Litthauer D. 2016.Unculturable T4-like and T7-like phages from South African deep gold mines. Available at https://www.ncbi.nlm.nih.gov/popset?LinkName=protein_popset&from_uid=256033955

Millard AD, Pearce D, Zwirglmaier K. 2016. Biogeography of bacteriophages at four hydrothermal vent sites in the Antarctic based on g23 sequence diversity. FEMS Microbiology Letters 363:1–12. DOI: 10.1093/femsle/fnw043.

Murtagh F. 1992. Comments on “Parallel algorithms for hierarchical clustering and cluster validity.” IEEE Transactions on Pattern Analysis and Machine Intelligence 14:1056–1057. DOI: 10.1109/34.159908.

Okazaki Y, Nishimura Y, Yoshida T, Ogata H, Nakano S. 2019. Genome - resolved viral and cellular metagenomes revealed potential key virus - host interactions in a deep freshwater lake. Environmental Microbiology 21:4740–4754. DOI: 10.1111/1462-2920.14816.

De Paepe M, Taddei F. 2006. Viruses' life history: towards a mechanistic basis of a trade-off between survival and reproduction among phages. PLoS Biology 4:e193. DOI: 10.1371/journal.pbio.0040193.
Palermo CN, Fulthorpe RR, Saati R, Short SM. 2019. Metagenomic analysis of virus diversity and relative abundance in a eutrophic freshwater Harbour. Viruses 11:792. DOI: 10.3390/v11090792.
Parsons TR, Maita Y, Lalli C. 1984. A Manual of Chemical and Biological Methods for Seawater Analysis. New York: Pergamon Press.
Parvathi A, Zhong X, Jacquet S. 2012. Dynamics of various viral groups infecting autotrophic plankton in Lake Geneva. Advances in Oceanography and Limnology 3:171–191. DOI: 10.1080/19475721.2012.738157.
Paul JH, Jiang SC, Rose JB. 1991. Concentration of viruses and dissolved DNA from aquatic environments by vortex flow filtration. Applied and Environmental Microbiology 57:2197–2204. DOI: 10.1128/aem.57.8.2197-2204.1991.
Popovskaya G. 2000. Ecological monitoring of phytoplankton in Lake Baikal. Aquatic Ecosystem Health and Management 3:215–225. DOI: 10.1016/S1463-4988(00)00021-X.
Potapov S, Belykh O, Krasnopeev A, Galachyants A, Podlesnaya G, Khanaev I, Tikhonova I. 2020. Diversity and biogeography of bacteriophages in biofilms of Lake Baikal based on g23 sequences. Journal of Great Lakes Research 46. DOI: 10.1016/j.jglr.2019.06.007.
Potapov S, Belykh O, Krasnopeev A, Gladkikh A, Kabilov M, Tupikin A, Butina T. 2018. Assessing the diversity of the g23 gene of T4-like bacteriophages from Lake Baikal with high-throughput sequencing. FEMS microbiology letters 365. DOI: 10.1093/femsle/fnx264.
Pradeep Ram A, Nishimura Y, Tomaru Y, Nagasaki K, Nagata T. 2010. Seasonal variation in viral-induced mortality of bacterioplankton in the water column of a large mesotrophic lake (Lake Biwa, Japan). Aquatic Microbial Ecology 58:249–259. DOI: 10.3354/ame01381.
Rahimi-Midani A, Lee YS, Kang S-W, Kim M-K, Choi T-J. 1970. First isolation and molecular characterization of bacteriophages infecting acidovorax citrulli, the causal agent of bacterial fruit blotch. The Plant Pathology Journal 34:59–64. DOI: 10.5423/PPJ.NT.08.2017.0190.
Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A. 2017. DnaSP 6: DNA sequence polymorphism analysis of large data sets. Molecular Biology and Evolution 34:3299–3302. DOI: 10.1093/molbev/msx248.
Sandaa R-A, Kristiansen H. 2016.T4-type bacteriophages from different marine environments. Available at https://www.ncbi.nlm.nih.gov/popset?LinkName=protein_popset&from_uid=157273680
Sime-Ngando T, Bettarel Y, Colombet J, Palesse S, Ram ASP, Charpin M, Amblard C. 2016. Lake Pavin: a pioneer site for ecological studies of freshwater viruses. In: Lake Pavin. Cham: Springer International Publishing, 229–244. DOI: 10.1007/978-3-319-39961-4_14.
Straškrábová V, Izmest'yeva LRR, Maksimova EAA, Fietz S, Nedoma J, Borovec J, Kobanova GII, Shchetinina EV V., Pislegina EV V., Stra??kr??bov?? V, Izmest'yeva LRR, Maksimova EAA, Fietz S, Nedoma J, Borovec J, Kobanova GII, Shchetinina EV V., Pislegina EV V. 2005.

629 Primary production and microbial activity in the euphotic zone of Lake Baikal (Southern Basin)
630 during late winter. Global and Planetary Change 46:57–73. DOI:
631 10.1016/j.gloplacha.2004.11.006.
632 Sullivan MB, Waterbury JB, Chisholm SW. 2003. Cyanophages infecting the oceanic
633 cyanobacterium Prochlorococcus. Nature 424:1047–1051. DOI: 10.1038/nature01929.
634 Suttle CA. 2007. Marine viruses — major players in the global ecosystem. 5:801–812. DOI:
635 10.1038/nrmicro1750.
636 Taboada B, Isa P, Gutiérrez-Escolano AL, del Ángel RM, Ludert JE, Vázquez N, Tapia-Palacios
637 MA, Chávez P, Garrido E, Espinosa AC, Eguiarte LE, López S, Souza V, Arias CF. 2018. The
638 Geographic Structure of Viruses in the Cuatro Ciénegas Basin, a Unique Oasis in Northern
639 Mexico, Reveals a Highly Diverse Population on a Small Geographic Scale. Applied and
640 Environmental Microbiology 84. DOI: 10.1128/AEM.00465-18.
641 Tetart F, Desplats C, Kutateladze M, Monod C, Desplats C, Kutateladze M, Monod C,
642 Ackermann H-WW, Krisch HM. 2001. Phylogeny of the major head and tail genes of the wide-
643 ranging T4-type bacteriophages. Journal of bacteriology 183:358–366. DOI:
644 10.1128/JB.183.1.358-366.2001.
645 Vollenweider RA, Kerekes J. 1982. Eutrophication of waters. Monitoring assessment and
646 control. Paris.
647 Wang MN, Ge XY, Wu YQ, Yang X Lou, Tan B, Zhang YJ, Shi ZL. 2015. Genetic diversity and
648 temporal dynamics of phytoplankton viruses in East Lake, China. Virologica Sinica 30:290–300.
649 DOI: 10.1007/s12250-015-3603-6.
650 Watanabe Y, Drucker VV. 1999. Phytoplankton blooms in Lake Baikal, with reference to the
651 lake's present state of eutrophication. In: Kawanabe H, Coulter GG, Roosevelt AC eds. Ancient
652 Lakes: Their Cultural and Biological Diversity. Kenobi Productions: Belgium, 217–225.
653 Wetzel RG, Likens GE. 2000. Limnological Analyses. New York, NY: Springer New York.
654 DOI: 10.1007/978-1-4757-3250-4.
655 Wu S, Zhou L, Zhou Y, Wang H, Xiao J, Yan S, Wang Y. 2020. Diverse and unique viruses
656 discovered in the surface water of the East China Sea. BMC Genomics 21:441. DOI:
657 10.1186/s12864-020-06861-y.
658 Zeigler Allen L, McCrow JP, Ininbergs K, Dupont CL, Badger JH, Hoffman JM, Ekman M,
659 Allen AE, Bergman B, Venter JC. 2017. The Baltic Sea virome: diversity and transcriptional
660 activity of DNA and RNA viruses. mSystems 2:e00125-16. DOI: 10.1128/mSystems.00125-16.
661 Zheng C, Wang G, Liu J, Song C, Gao H, Liu X. 2013. Characterization of the major capsid
662 genes (g23) of T4-type bacteriophages in the wetlands of Northeast China. Microbial Ecology
663 65:616–625. DOI: 10.1007/s00248-012-0158-z.
664 Zhong X, Jacquet S. 2014. Differing assemblage composition and dynamics in T4-like
665 myophages of two neighbouring sub-alpine lakes. Freshwater Biology 59:1577–1595. DOI:
666 10.1111/fwb.12365.
667

Figure 1

Map of the sampling area.

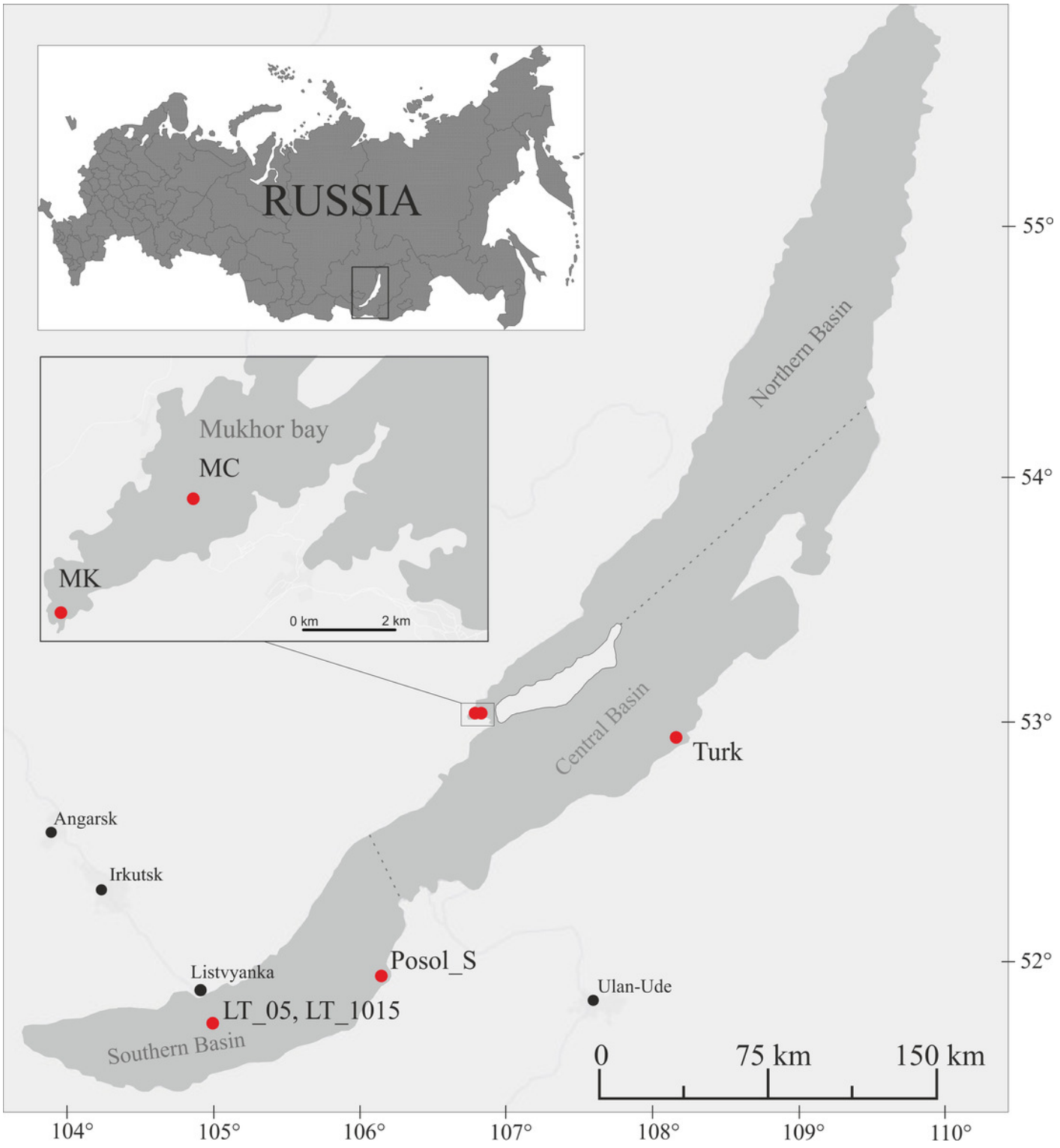
RUSSIA
Mukhor bay
MC
MK
0 km
2 km
Northern Basin
Central Basin
Southern Basin
Turk
Angarsk
Irkutsk
Listvyanka
Posol_S
LT_05, LT_1015
Ulan-Ude
0
75 km
150 km
55°
54°
53°
52°
104°
105°
106°
107°
108°
109°
110°

Figure 2

NMDS analysis based on *g23* sequences and physicochemical parameters.

pH - potential of hydrogen, P – total phosphorus, Chl – chlorophyll *a*, N – total nitrogen.

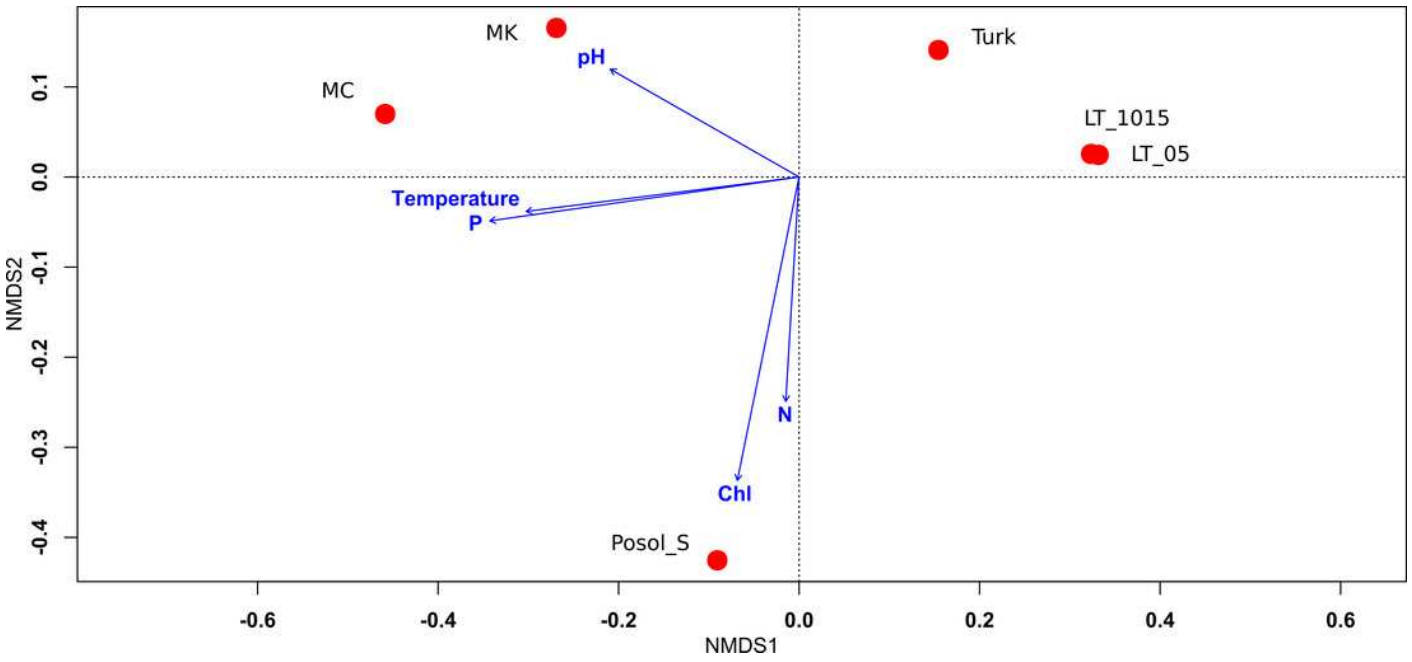

Figure 3

UpSetR plot created by representative nucleotide sequences of OTUs.

The dots indicate the total OTU count in a sample; the dots connected by lines – shared OTUs. Intersection size ≥ 8. Mode = "intersect".

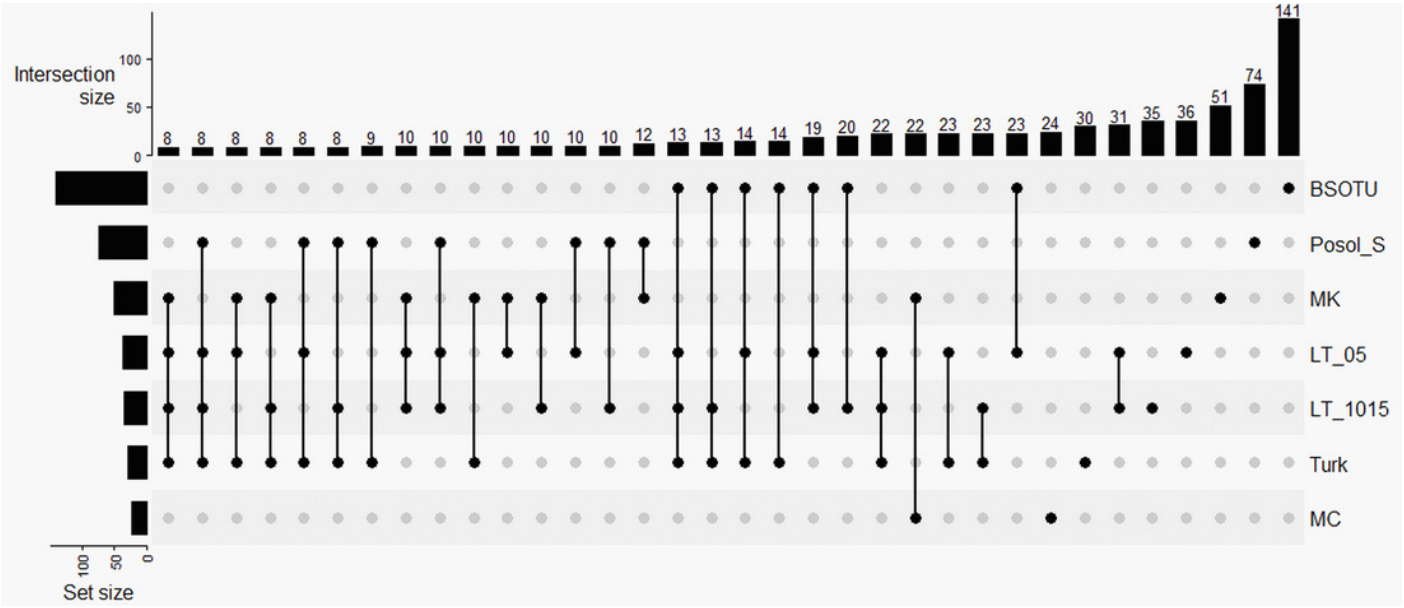

Figure 4

Bayesian phylogenetic tree based on alignment of 388 gp23 major capsid protein sequences.

Dots mark cultivated bacteriophages.

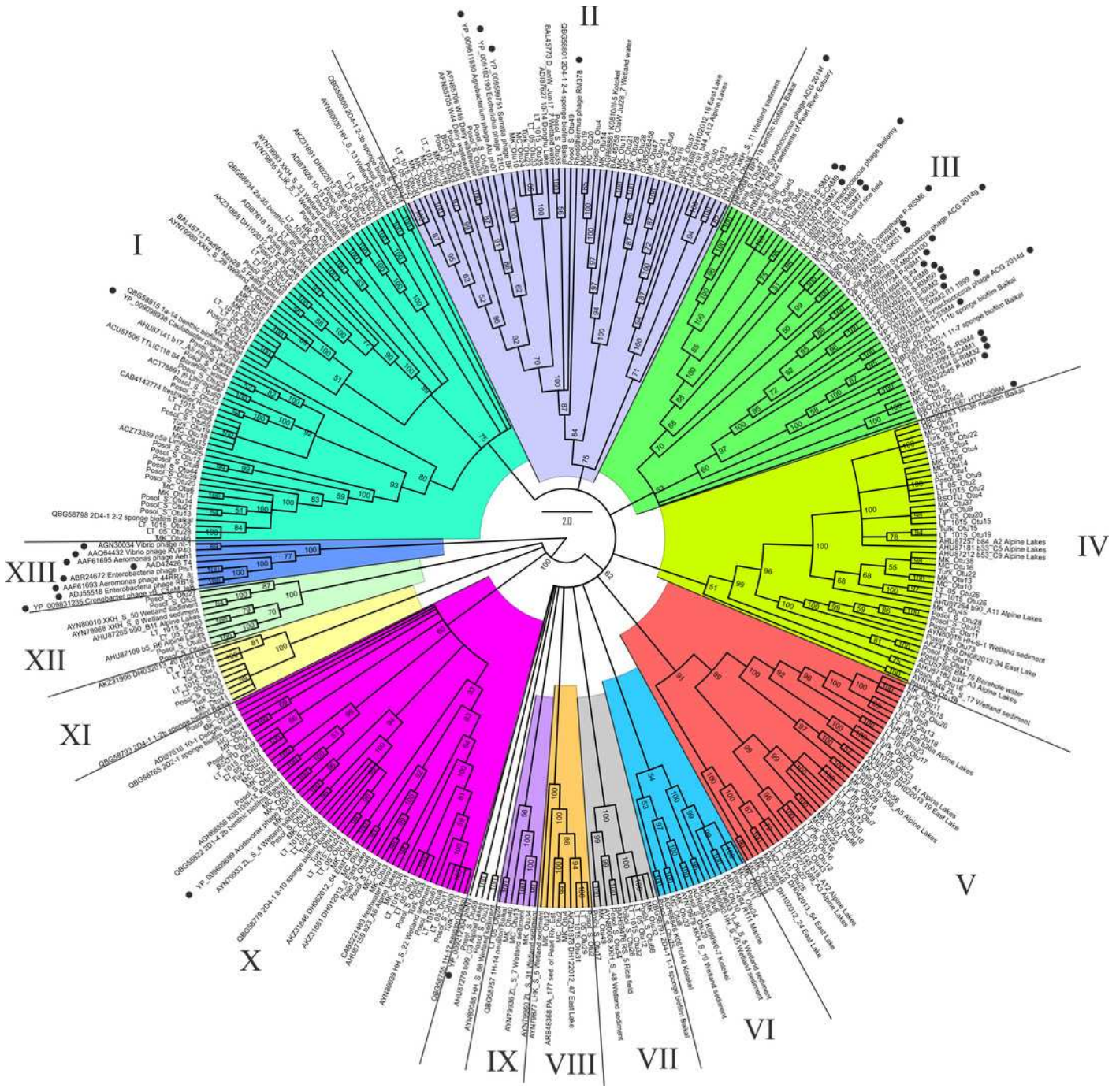

Figure 5

Map showing the samples included in the analysis for this study.

Figure 6

Cluster dendrogram (UPGMA).

Samples from this study are marked in bold. SB – southern basin, Lake Baikal, NB – northern basin, Lake Baikal, BSOTU – Baikal samples OTU, LAB – lakes Bourget and Annecy. Untagged samples are either unpublished or have no open access articles.

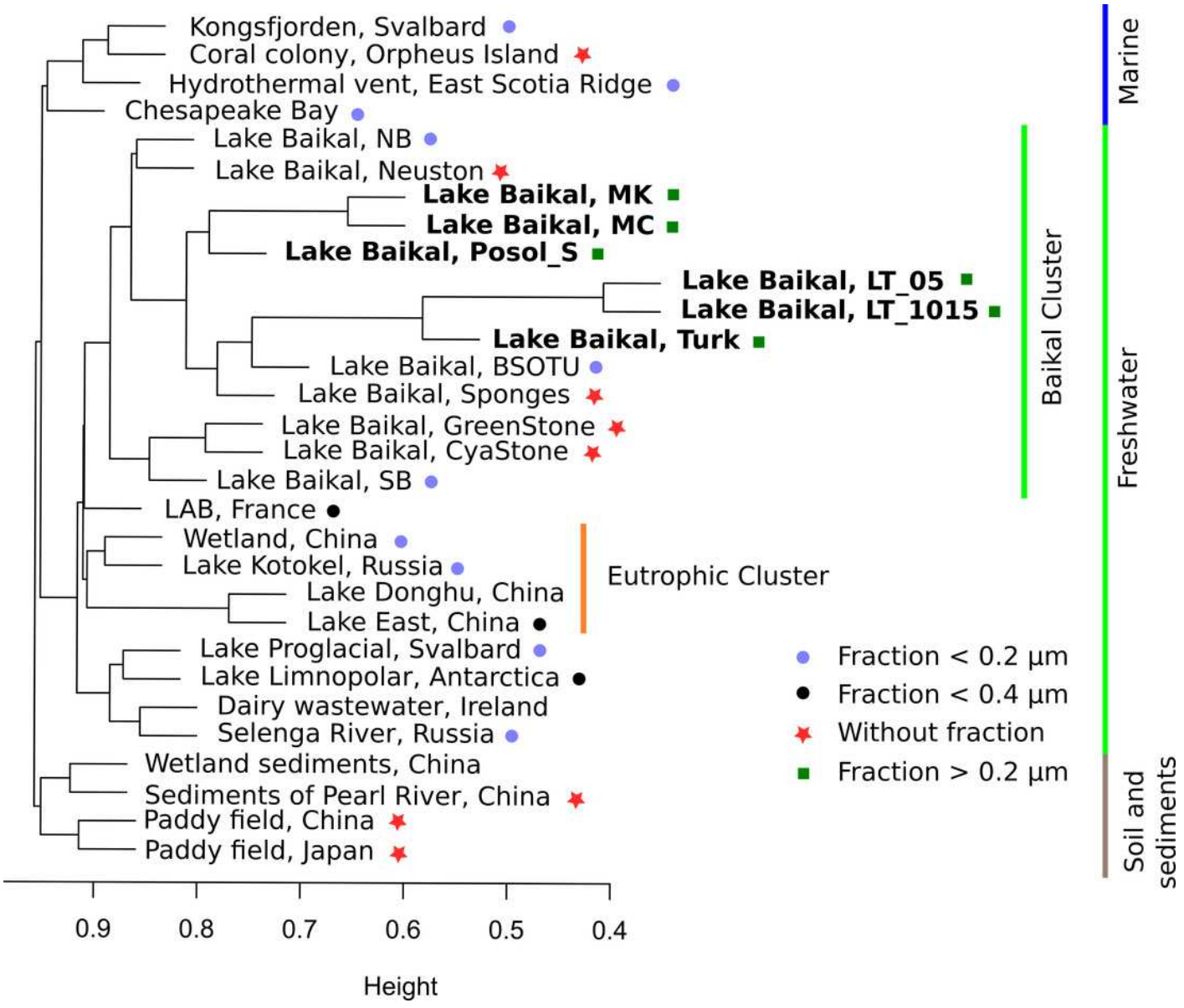

Table 1(on next page)

Physical and chemical parameters of water.

1

	MK	MC	Turk	Posol_S	LT_05	LT_1015
Temperature, °C	19.6	21.5	17.3	19	12	6.2
pH	8.2	8.6	8.3	8.2	8.2	8.2
P_{total}, µg/l	12.8	13.5	10	12	7.8	8.2
N_{total}, mg/l	0.23	0.26	1.3	2	0.11	0.13
Chl *a*, µ/l	4.8	4.8	1.1	25	1.8	2.1

2

Table 2(on next page)

Summary information about the sampling site and results of the processing stages.
Sampling date: August 2019

1

Sample	GC – content, %	Total raw reads	After trimming (reads)	Singletons / doubletons	Chimeras	Number of OTUs after removal of chimeras, singletons and doubletons, 97% identity
MK	47	41802	37745	13579/989	1	51
MC	46	16462	15446	5777/443	1	24
Turk	47	12454	11394	5062/333	1	30
Posol_S	45	19940	18044	8146/470	0	74
LT_05	46	15368	13818	5090/370	0	36
LT_1015	47	13180	11902	4552/317	0	35

2

Table 3(on next page)

The closest relatives with the lowest e-value for each sample (database RefSeq, amino acid level).

1

Sample	Sequences belonging to cyanophages, %	Sequence	The closest relative	Query cover, %	Identity, %	E-value
MK	60.8	Otu13	*Synechococcus* phage S-SSM7	99.3	64.4	5.56E-60
MC	58.3	Otu10	*Synechococcus* phage S-SSM7	99.3	64.4	5.56E-60
Turk	63.3	Otu6	*Synechococcus* phage S-SM2	99.4	72	1.67E-73
Posol_S	73	Otu51	*Synechococcus* phage Bellamy	99.4	72.6	1.93E-75
LT_05	66.7	Otu31	*Synechococcus* phage S-CAM1	99.4	77.9	5.85E-78
LT_1015	65.7	Otu29	*Synechococcus* phage S-CAM1	99.4	77.9	3.06E-78

2

Table 4(on next page)

The number of sequences in this study similar to the sequences from other sources, % (database GenBank, amino acid level).

1

Isolation source	MK	MC	Turk	Posol_S	LT_05	LT_1015	Reference
Biofilms, Lake Baikal (Russia)	19.6	20.8	20	27	27.8	28.6	(Potapov et al., 2020)
Pelagic water, Lake Baikal (Russia)	9.8	12.5	30	6.8	25	22.9	(Potapov et al., 2018)
Lakes Bourget and Annecy (France)	17.6	16.7	23.3	10.8	22.2	25.7	(Zhong & Jacquet, 2014)
Lake Donghu (China)	3.9	4.7	3.3	2.7	8.3	5.7	(Huang, Cheng & Xu, 2011)
Lake East (China)	9.8	8.3	6.7	8.1	5.6	5.7	(Wang et al., 2015)
Wetland sediment (China)	25.5	20.8	3.3	18.9	2.8	2.9	(Li et al., 2018)
Borehole water (South Africa)	2	4.2	3.3	4.1	2.8	2.9	(Mabizela & Litthauer, 2016)
Dairy wastewater (Ireland)	-	-	3.3	2.7	2.8	2.9	(Knapik & Prentice, 2012)
Paddy water (China)	-	-	-	1.4	2.8	2.9	(Zheng et al., 2013)
Rimov reservoir (Czech Republic)	2	-	3.3	4.1	-	-	(Kavagutti et al., 2019)
Marine environment	-	-	3.3	-	-	-	(Sandaa & Kristiansen, 2016)
Wetland water (China)	5.9	4.2	-	1.4	-	-	(Zheng et al., 2013)
Sediments of Pearl River Estuary (China)	2	4.2	-	1.4	-	-	(He et al., 2017)
Lake Kotokel (Russia)	2	4.2	-	5.4	-	-	(Butina et al., 2013)
Lake Limnopolar (Antarctica)	-	-	-	2.7	-	-	(López-Bueno et al., 2009)
Paddy field soil (Japan)	-	-	-	1.4	-	-	(Fujihara et al., 2010)
Surface soil of rice field (Japan)	-	-	-	1.4	-	-	(Jia et al., 2007)

2

Table 5(on next page)

Nucleotide diversity of the *g23* gene fragment.

1

Sample	Nucleotide diversity, π	References
MK	0.36	This study
MC	0.35	-//-
Turk	0.35	-//-
Posol_S	0.31	-//-
LT_05	0.33	-//-
LT_1015	0.34	-//-
Kongsfjorden (Svalbard)	0.33	(Bellas & Anesio, 2013)
Coral colony (Orpheus Island)	0.23	(Buerger et al., 2018)
Hydrothermal vent (East Scotia Ridge)	0.30	(Millard, Pearce & Zwirglmaier, 2016)
Chesapeake Bay	0.29	(Jamindar et al., 2012)
Lake Baikal, pelagic water (South basin)	0.28	(Potapov et al., 2018)
Lake Baikal (South basin)	0.33	(Butina et al., 2010)
Lake Baikal (North basin)	0.30	-//-
Lake Baikal, Neuston	0.27	(Potapov et al., 2020)
Lake Baikal, Sponges	0.36	-//-
Lake Baikal, CyanoStone	0.32	-//-
Lake Baikal, GreenStone	0.31	-//-
Lake Limnopolar (Antarctica)	0.37	(López-Bueno et al., 2009)
Lake Proglacial (Svalbard)	0.34	(Bellas & Anesio, 2013)
Lake Kotokel (Russia)	0.34	(Butina et al., 2013)
Lake East (China)	0.29	(Wang et al., 2015)
Lake Donghu (China)	0.31	(Huang, Cheng & Xu, 2011)
Lake Annecy and Bourget (France)	0.28	(Zhong & Jacquet, 2014)
Dairy water (Ireland)	0.37	(Knapik & Prentice, 2012)
Wetland, water (China)	0.32	(Zheng et al., 2013)
Sediments of Pearl River Estuary (China)	0.28	(He et al., 2017)
Wetland sediments (China)	0.29	(Li et al., 2018)
Paddy field (Japan)	0.33	(Cahyani et al., 2009)
Paddy field (China)	0.35	(Li et al., 2019)

2