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] , **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, Russia

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

Lake Baikal phage communities are important for lake ecosystem functioning. Here, we describe the diversity of T4-bacteriophage associated with the bacterial fraction of filtered water samples collected from the pelagic zone, coastal zone and shallow bays. Although the study of the diversity of phages for the *g23* gene has been carried out at Lake Baikal for more than ten years, shallow bays that comprise a significant part of the lake's area have been neglected, and the bacterial fraction for this gene has not been studied either. During this study, we examined six water samples, in which 24 to 74 viral OTUs were obtained. Phage communities were probed using amplicon sequencing methods targeting the gene of major capsid protein (*g23*) and compared phylogenetically across sample locations and with sequences previously retrieved from non-bacterial fractions (<0.2 um) and biofilms (non-fractionated). The sequences from shallow bays largely differed from those in the pelagic and coastal samples and formed individual subcluster in the UPGMA tree, reflecting differences in viral communities depending on the productivity of various sites of Lake Baikal. According to the RefSeq database, from 58.3 to 73% of sequences of each sample had cultivated closest relatives belonging to cyanophages. In this study, for phylogenetic analysis, we chose the closest relatives not only from the RefSeq and GenBank NR databases but also from two marine and one freshwater viromes: eutrophic Osaka Bay (Japan), oligotrophic area of the Pacific Ocean (Station ALOHA) and mesotrophic and ancient Lake Biwa (Japan), which allowed us to more fully compare the diversity of marine and freshwater phages. The identity with marine sequences at the amino acid level ranged from 35 to 80%, and with the sequences from the viral fraction and bacterial one from Lake Biwa – from 35.3 to 98% and from 33.9 to 89.1%, respectively.

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

Lake Baikal phage communities are important for lake ecosystem functioning. Here, describe the diversity of T4-bacteriophage associated with the bacterial fraction of filtered water samples collected from the pelagic zone, coastal zone and shallow bays. Although the study of the diversity of phages for the *g23* gene has been carried out at Lake Baikal for more than ten years, shallow bays that comprise a significant part of the lake's area have been neglected, and the bacterial fraction for this gene has not been studied either. During this study, we examined six water samples, in which 24 to 74 viral OTUs were obtained. Phage communities were probed using amplicon sequencing methods targeting the gene of major capsid protein (*g23*) and compared phylogenetically across sample locations and with sequences previously retrieved from non-bacterial fractions (<0.2 um) and biofilms (non-fractionated). The sequences from shallow bays largely differed from those in the pelagic and coastal samples and formed individual subcluster in the UPGMA tree, reflecting differences in viral communities depending on the productivity of various sites of Lake Baikal. According to the RefSeq database, from 58.3 to 73% of sequences of each sample had cultivated closest relatives belonging to cyanophages. In this study, for phylogenetic analysis, we chose the closest relatives not only from the RefSeq and GenBank NR databases but also from two marine and one freshwater viromes: eutrophic Osaka Bay (Japan), oligotrophic area of the Pacific Ocean (Station ALOHA) and mesotrophic and ancient Lake Biwa (Japan), which allowed us to more fully compare the diversity of marine and freshwater phages. The identity with marine sequences at the amino acid level ranged from

39 35 to 80%, and with the sequences from the viral fraction and bacterial one from Lake Biwa –
40 from 35.3 to 98% and from 33.9 to 89.1%, respectively.

42

Introduction

45   Viruses are obligate intracellular parasites consisting of a single-stranded or double-
46 stranded RNA or DNA molecule enclosed within a protein capsid; enveloped viruses have an
47 additional membrane envelope (supercapsid). Viruses are distinguished by their high abundance
48 and high genetic diversity (Suttle, 2007), thereby representing an inexhaustible pool for research.
49 To date, according to ICTV Master Species List (ICTV, 2019), 3973 species belong to DNA
50 viruses and 2617 species – to RNA viruses, but most of the sequences obtained to date from
51 viromes are known to be "viral dark matter" (Krishnamurthy & Wang, 2017), and there are much
52 more real biological species of viruses (Gregory et al., 2019). The bulk of the sequences obtained
53 by metagenomic sequencing of DNA-containing viruses in aquatic ecosystems, which can be
54 identified from databases, belong to the order Caudovirales (Cai et al., 2016; Garin-Fernandez et
55 al., 2018; Gong et al., 2018; Taboada et al., 2018; Gregory et al., 2019; Wu et al., 2020). Due to
56 the rapid transformation of viral taxonomy, the order Caudovirales has expanded to 14 families.
57 Among them, the family *Myoviridae* is the best known and most studied. It includes 8
58 subfamilies and 153 genera (https://talk.ictvonline.org/taxonomy/).
59 The family *Myoviridae* contains DNA phages that are genetically and morphologically similar to
60 the well-studied coliphage T4 (Ackermann & Krisch, 1997). At the same time, myoviruses have
61 a relatively wide range of hosts; they can infect their hosts belonging to various species or
62 sometimes even various genera of bacteria (Sullivan, Waterbury & Chisholm, 2003).
63 Due to the lack of universal genes in viruses, signature genes for specific groups are targeted
64 using group-specific primers. The *g23* gene fragment encoding major capsid protein is the most
65 reliable marker for the analysis of the biodiversity of T4-like phages of the family *Myoviridae*
66 (Tetart et al., 2001; Adriaenssens & Cowan, 2014). Based on the analysis of *g23* sequences, T4-
67 like phages are divided into several groups: "true" T-evens represented by bacteriophage T4 and
68 closely related phages infecting enterobacteria (e.g. T2, T6), Pseudo T-evens and Schizo T-evens
69 (phages of the genera *Aeromonas*, *Vibrio*, etc.) as well as more distant Exo T-evens (cyano- and
70 pelagiphages, etc.) (Desplats & Krisch, 2003).
71   In the aquatic environment, viruses are an important factor in the regulation of the
72 number and structure of microbial communities (Kutter & Sulakvelidze, 2005). As a result of the
73 viral lysis of bacteria, a significant amount of organic matter does not enter higher trophic levels
74 but is reused by the bacterial community (Suttle, 2007). Viruses are also capable of horizontal
75 gene transfer (Zhang & Gui, 2018). Moreover, viruses can have auxiliary metabolic genes that
76 modulate the metabolism of host cells during infection, making phage replication more efficient
77 (Warwick-Dugdale et al., 2019).

78   In freshwater ecosystems such as Lake Biwa (Japan), the percentage of daily bacterial
79   production destroyed by viruses was estimated as high and accounted for 52.7 ± 16.2% in the
80   upper layer and 13.6 ± 5.2% in the deeper layer (Pradeep Ram et al., 2010). In Lake Pavin
81   (France), the average seasonal contribution from bacteriophages to bacterial lysis reached 16.2%
82   (Sime-Ngando et al., 2016).
83   This article presents the study of the diversity of T4-like viruses for the gene of the main
84   capsid protein (*g23*) in various ecotopes of Lake Baikal, the oldest lake in the world. Previously,
85   a high diversity of planktonic T4-like viruses was detected in the pelagic zone of Lake Baikal,
86   and their difference from bacteriophages of other ecosystems was demonstrated (Butina et al.,
87   2010). Later, this signature gene provided an estimation of the composition of phages in
88   eutrophic Lake Kotokel located close to Lake Baikal, 3 km from the east coast of the lake, and
89   having a direct water connection with it (Butina et al., 2013). Unifrac analysis has revealed that
90   the viral community in Lake Kotokel is closely related to the phages from Lake Donghu in China
91   characterized by a similar trophic state. Subsequently, we have shown that the sequences of the
92   *g23* major capsid protein gene from the eutrophic lakes Kotokel (Russia) and Donghu (China)
93   are rather more similar to the sequences from the shallow Maloye More Strait of Lake Baikal
94   than to the sequences from the pelagic zone of the lake (Potapov et al., 2013). In 2018, using
95   high throughput sequencing for the first time, we studied the diversity of phages in the deep
96   southern basin of Lake Baikal. The array of reads yielded 141 OTUs that differed from other
97   ecosystems and, at the same time, were similar to the sequences from the alpine lakes Bourget
98   and Annecy (Potapov et al., 2018). Comparative study of the *g23* gene sequences of phages from
99   biofilms formed on stones and sponges in the littoral zone of the lake (viriobenthos), as well as
100   from its surface microlayer (virioneuston), has revealed that benthic and previously obtained
101   planktonic bacteriophages from the pelagic zone of Lake Baikal form a shared Baikal cluster
102   (Potapov et al., 2020).
103   Previously, it was shown that the filtration of samples through filters with a pore size of
104   0.2 μm reduces the number of phages in the filtrate by an average of two-thirds (transmission
105   electron microscopy counting method) (Paul, Jiang & Rose, 1991). In a later study, the
106   proportion of viruses retained on a 0.2 μm filter did not exceed 15% of the overall number of
107   virus-like particles (epifluorescence microscopy counting method) (Auguet, Montanié &
108   Lebaron, 2006).
109   For lytic phages, including myoviruses, three states of life cycle were described: i) free
110   during extracellular search; ii) located in a certain space with no host, which is, for example,
111   associated with an inert particle; iii) actively infects bacteria (Kutter & Sulakvelidze, 2005). T4-
112   like viruses having strong lytic properties likely experience rapid exchange of intracellular and
113   free phages. For instance, phages T4 and λ have latent periods of 20 min and 50 min,
114   respectively (De Paepe & Taddei, 2006). However, in natural ecosystems, the latent period can
115   be up to 8 hours (Mruwat et al., 2021) or up to 24 hours in myopelagiphages (Zhao et al., 2013).

116 The study of a bacterial fraction can reveal the relationship between bacteria and bacteriophages
117 infecting them, this will provide an understanding about the phages that are in the propagation
118 stage.
119 Currently, the diversity of T4-like phages in the viral fraction (less than 0.4 and 0.2 μm) has been
120 mainly studied (Parvathi, Zhong & Jacquet, 2012; Goldsmith et al., 2015; Wang et al., 2015;
121 Millard, Pearce & Zwirglmaier, 2016; Liu, Cai & Zhang, 2017). Organisms larger than 0.2 or 0.4
122 μm (bacterial fraction) are removed using various methods because it is methodologically more
123 preferable to work with a viral fraction that does not contain bacterial cells. Information about
124 the composition and role of viruses associated with microbial communities is scarce and
125 discussed mainly in metagenomic studies (De Cárcer et al., 2016; Zeigler Allen et al., 2017;
126 Aylward et al., 2017; Palermo et al., 2019; Okazaki et al., 2019; Coutinho et al., 2020).
127 This study aims to reveal the diversity of phages for the *g23* major capsid protein gene in
128 shallow bays, at the coastal area and the pelagic part of Lake Baikal and to analyse the obtained
129 data in terms of regional and global biogeography by comparing them with the sequences
130 obtained previously from Lake Baikal and other ecosystems.
131

Materials & Methods

133

Sampling sites

135 Lake Baikal is the oldest (25 million years), the deepest lake. It is of tectonic origin and is
136 located in the Baikal rift depression. The length of the lake is 636 km; the width in the widest
137 part of the lake opposite the estuary of the Barguzin River reaches 80 km, and in the narrowest
138 area opposite the delta of the Selenga River – about 25 km. The lake is situated in the
139 northeastern part of Central Asia at an altitude of 455.6 m above sea level. The area of the lake
140 with islands is 31500 km^2, the length of the coastline without islands is approximately 2000 km.
141 The greatest depth of the lake (1637 m) is recorded in its central basin; the maximum depth in
142 the southern basin is 1432 m, and that in the northern basin is 889 m (Galaziy, 1993).
143 Lake Baikal is also the world's largest freshwater body that stores 20% of the land surface
144 waters. The volume of the lake is 23000 km^3. Constant temperature (~ 4°C) below the
145 thermocline throughout the year is one of the main hydrophysical characteristics of the lake.
146 Deep-water renewal does not exceed 19 years (Hohmann et al., 1997). Additionally,
147 hydrochemical characteristics include high levels of dissolved oxygen (more than 80%) and low
148 concentrations of mineral nitrogen and phosphorus as well as of organic carbon (Khodzher et al.,
149 2018). Geological, geographical and hydrological characteristics of the lake reflect its
150 uniqueness and high endemism of aquatic organisms (Kozhova & Izmest'eva, 1998).
151 For this study, we took samples from two shallow bays of Lake Baikal (Mukhor Bay and
152 Posolsk Sor), one coastal zone (near the Turka settlement) and one pelagic zone (in the middle of
153 the Listvyanka settlement-the Tankhoy settlement section). Samples were collected in August
154 2019. The shallow bay of Posolsk Sor (maximum depth 3.5 m) is located on the southeast coast
155 of Lake Baikal, about 20 km south of the Selenga River delta. Mukhor Bay is one of the warmest

156   and shallowest bays (average depth 5 m) located in the Maloye More Strait that is situated
157   between the mainland and Olkhon Island. Both bays warm up well in July and August and are
158   the most visited tourist sites of Lake Baikal.
159   Coordinates and names of sampling sites are as follows: Mukhor Bay near the Kuchelga
160   River – MK (53.020244 N, 106.753739 E), a centre of Mukhor Bay – MC (53.044278 N,
161   106.798343 E), Posolsk Sor Bay – Posol_S (51.96556 N, 106.17181 E), a coastal zone near the
162   Turka settlement – Turk (52.9384 N, 108.21304 E), and the central station of the Listvyanka
163   settlement – the Tankhoy settlement section – LT_05, LT_1015 (51.721883 N, 104.993283 E).
164   Figure 1 shows the sampling map.
165   In the bays and the coastal zone, samples of a 1 L volume were taken from the water
166   surface using sterile bottles. At the pelagic station, samples of a 500 mL volume were taken
167   using a Niskin bathometer from depths of 0, 5, 10, and 15 m. Thereafter, the 500 mL samples
168   from the layers of 0, 5, 10, and 15 m were combined, and two integral samples were obtained: 0
169   to 5 m and 10 to 15 m.
170
171
172
173   **Figure 1.**
174
175   **DNA extraction and preparation of amplicons**
176   Water samples (1 liter) from each site (MK, MC, Turk, Posol_S, LT_05, and LT_1015) were
177   filtered through sterile polycarbonate filters with a pore size of 0.2 μm (Sartorius, Germany)
178   without using prefilters, and they were frozen on board the research vessel during the
179   expeditions. DNA was extracted by the standard phenol-chloroform method in the laboratory.
180   Briefly, filters from each sample were cut and placed into separate test tubes with 300 μl 1 x TE
181   buffer. Then, lysozyme (Thermo Fisher Scientific, USA) was added (final concentration 1
182   mg/ml) and left for 1.5 h at 37°C in a thermostat. After that, proteinase K (Thermo Fisher
183   Scientific, USA) (final concentration 1 mg/ml) and sodium dodecyl sulfate (SDS) were added
184   followed by incubation at 37°C for 1.5 h. The next stage was freezing at -70°C for 12 h; after
185   thawing, 500 μl of phenol and chloroform, isoamyl alcohol (24:1) were added to the mixture,
186   kept for 10 min and centrifuged for 5 min at 13400 rpm. The supernatant was transferred to a
187   new tube, and the extraction was repeated with chloroform and isoamyl alcohol. Thereafter, 1/10
188   of the $NaCH_3COO$ (3M) volume and 2 volumes of ethanol (96%) were added to the extracted
189   supernatant and left overnight at -20°C. Then, it was centrifuged for 20 min at 13400 rpm and
190   washed twice with 70% ethanol each time for 1 min. The residue was dried and dissolved in 50
191   μl of deionized water. Primers MZIA1bis and MZIA6 were used (Filée et al., 2005). The PCR
192   mixture consisted of the following components: Master mix 2x Taq M (Alkor Bio, Russia), 0.1
193   μM primers, nuclease-free water, and DNA template. PCRs were performed with the following
194   PCR cycle parameters: denaturation at 95°C for 15 min, 30 cycles of denaturation at 95°C for 30
195   s, annealing at 50°C for 30 s, extension at 72°C for 1 min, and a final extension at 72°C for 10

196   min. DNA was purified from the PCR mixture using a suspension of magnetic particles
197   CleanMag DNA (Evrogen, Russia). Library preparation and sequencing on Illumina MiSeq
198   2*300 were performed in the "Genomics Core Facility" (ICBFM SB RAS, Novosibirsk, Russia).
199

Bioinformatic analysis

201   Sequence quality analysis was carried out using the FastQC software tool (Andrews,
202   2010). Trimming was performed using the Trimmomatic v. 0.36 tool (Bolger, Lohse & Usadel,
203   2014) with the following parameters: SLIDINGWINDOW:4:20 LEADING:3 TRAILING:3
204   MINLEN:50. Further processing was performed using the Usearch v. 11.0.667 tool (Edgar,
205   2010). Paired-end reads were combined using the *-fastq_mergepairs* command. Then unique
206   sequences (*-derep_fulllength*) were sorted out. The next step was clustering at the 97% identity
207   level, UPARSE-OTU algorithm (*-cluster_otus*) as well as the removal of chimeras, singletons
208   and doubletons. Thus, all six sets of sequences were processed separately. The chosen level was
209   used after the analysis of rarefaction curves. Additionally, we were guided by previous
210   publications (Millard, Pearce & Zwirglmaier, 2016; He et al., 2017) where the clustering at the
211   97% level was also chosen. For further analyses, the primers were removed. The annotation was
212   performed using the online blastp analysis (https://blast.ncbi.nlm.nih.gov/Blast.cgi) with default
213   expected threshold based on the RefSeq (Release 207) and GenBank non-curated (NR) (Release
214   243) databases.
215   Nucleotide diversity values were calculated using DNASP v. 6.12 (Rozas et al., 2017).
216   The sequences were aligned by the ClustalW, Mega 7 software (Kumar, Stecher & Tamura,
217   2016) with default settings. Each set of the *g23* sequences was tested separately. From this study
218   the sequences were analysed individually for each sample based on the resulting OTUs.
219   The network was built using the QIIME v. 1.9.1 scripts (Caporaso et al., 2010). At the
220   first step, all sets (566 sequences) obtained from lakes Baikal and Kotokel (nucleotide level)
221   were clustered at the 97% level (*pick_otus.py*). The second step was the OTU table
222   (*make_otu_table.py*). Next, two files were created for the network (*make_otu_network.py*:
223   real_edge_table and real_node_table), which are required for the analysis in Cytoscape v. 3.7.0
224   (Shannon, 2003).
225   For phylogenetic analysis and Unweighted Pair Group Method with Arithmetic Mean
226   (UPGMA) (Sneath & Sokal, 1973), nucleotide sequences from this study were translated to
227   amino acid ones using the BioEdit v. 7.0.9.0 program (Hall, 1999).
228   Based on the blastp analysis, the nearest neighbours were selected for phylogenetic
229   analysis. These data were supplemented with the *g23* sequences from the sampling site of the
230   Hawaii Ocean Time-series (HOT) program (Luo et al., 2020), Osaka Bay (Japan) (Nishimura et
231   al., 2017) and Lake Biwa (Japan) (Okazaki et al., 2019).
232   Viromes from Lake Biwa were obtained from epilimnion and hypolimnion and from two
233   fractions (bacterial and viral). Datasets available in the SRA archive were combined by fractions
234   into two metagenomes: 3 (epilimnion) from the bacterial fraction and 12 from the viral fraction.
235   Thereafter, quality filtration using Trimmomatic v.0.36 (Bolger, Lohse & Usadel, 2014) with the

236 following parameters: SLIDINGWINDOW:4:20 LEADING:3 TRAILING:3, was carried out
237 separately for the viral fraction and the bacterial fraction. Reads were assembled using the
238 metaSPAdes program (Bankevich et al., 2012). The assembled contigs from Lake Biwa and the
239 contigs available for downloading in NCBI, which were obtained from the sampling site of the
240 Hawaii Ocean Time-series (HOT) program, station ALOHA (PRJNA352737) and Osaka Bay
241 virome (OBV) (PRJDB4437), were translated into CDS using GeneMarkS v. 3.36 (Besemer,
242 2001). Nucleotide sequences were annotated in DIAMOND (Buchfink, Xie & Huson, 2014) with
243 the following settings: database – RefSeq, blastx, min-score – 50, more-sensitive. The sequences
244 having less than 60% amino acid identity with the sequences from the RefSeq database were
245 removed. Short sequences of less than 250 nucleotides were also removed. Clustering of the
246 sequences at the 100% level using CD-HIT (Li & Godzik, 2006) allowed us to exclude identical
247 sequences. Then, the sequences were translated into amino acids, and, on their basis, local
248 databases were created for blastp analysis (e-value 10^{-3}) with the studied sequences. For each set,
249 the search was carried out independently. Unique sequences that were assigned as the closest
250 relatives were included in the phylogenetic analysis. Each sequence was manually checked for
251 the length and *g23* gene. Amino acids were aligned in the Mega 7 software using the ClustalW
252 algorithm. A phylogenetic tree was constructed through Bayesian analysis using the MrBayes
253 software (v. 3.2.6) (Huelsenbeck & Ronquist, 2001). Two independent Markov chain Monte
254 Carlo (MCMC) analyses were launched for 25 million generations with 25% burn-in (rejection
255 of initial generations) and four chains (one cold and three hot ones), amino acid model (Blosum).
256 Analysis was completed with an average standard deviation of split frequencies: 0.02; parameter
257 potential scale reduction factor (PSRF+) was 1.000. All calculations were performed on HPC-
258 cluster "Akademik V.M. Matrosov" ("Irkutsk Supercomputer Center of SB RAS,
259 http://hpc.icc.ru").
260   To assess the β-diversity of the sequences, UPGMA was performed. The datasets were
261 taken from those available in the GenBank NR database. Based on the amino acid sequences, the
262 distance matrix was obtained through the unweighted UniFrac metric (multiple sequence
263 alignment – ClustalW, model - Blosum62, normalized=TRUE) followed by a hierarchical cluster
264 analysis (hclust) (Murtagh, 1992) method "average" (=UPGMA) using phyloseq (v. 1.21.0),
265 vegan (v. 2.5-7) and phangorn (v. 2.2.0) packages implemented in the R software (v. 3.2.4).
266   The raw sequencing data were deposited in the Sequence Read Archive (SRA) of the
267 National Center for Biotechnology Information (NCBI) under the project accession number
268 PRJNA770749; all 250 *g23* sequences (OTUs) were deposited in GenBank NR under accession
269 numbers MW690662 - MW690911.
270
271 **Water Chemistry Analysis**
272 Chlorophyll *a* concentration was determined as described previously (Bondarenko et al., 2019).
273 Total phosphorus content was determined using a KFK-2 photoelectric colorimeter (ZOMZ,
274 Zagorskii optiko-mekhanicheskii zavod [Zagorskii optical mechanics factory], Russia) after
275 persulfate oxidation. Total nitrogen concentration was determined by persulfate oxidation in an

276 alkaline medium according to R. G. Wetzel & G. E . Likens (2000) using a PE-5300VI
277 spectrophotometer (Ekroskhim (the former "Ekohim"), Russia).
278
279 ## Results
280 **Environmental parameters**
281   According to the indicators given in the table, the productivity of shallow bays
282 corresponds to the mesotrophic status, and the productivity of the waters of the pelagic stations
283 corresponds to the oligotrophic status according to R. A. Vollenweider (Vollenweider &
284 Kerekes, 1982). Table 1 shows the results of the physicochemical analysis. Chlorophyll *a* and
285 total nitrogen are positively associated with the sample from Posolsk Sor Bay (Posol_S); the
286 water temperature, pH and total phosphorus – with the samples from Mukhor Bay (MC and
287 MK).
288
289 **Table 1**
290
291
292 **Analysis of *g23* sequences**
293   Overall, we obtained 250 representative viral sequences (OTU) of the *g23* gene fragment
294 based on the 97% clustering. The lengths of the sequences at the nucleotide level, taking into
295 account the primers, ranged from 355 to 547 nucleotides (average length 439 nucleotides). The
296 main data obtained from processing during the preparation of sequences are shown in
297 supplementary Table 1.
298
299 **Nucleotide diversity of the *g23* gene fragment**
300   To assess alpha diversity, we conducted a comparative analysis of sequences based on
301 nucleotides (nucleotide diversity, π) (Table 2). In addition to the sequences from this study, the
302 sequences were taken from various sources available in the NCBI database. Based on
303 microdiversity data, among the sequences obtained in this study, Posol_S is the least diverse *g23*
304 community in Lake Baikal, and MK is the most diverse one. In contrast, the sequences from the
305 polar Lake Limnopolar, dairy water (Ireland) and Lake Baikal (LS, 2D4) were distinguished by a
306 larger nucleotide diversity.
307
308 **Table 2**
309
310 **Blast analysis of the *g23* gene (amino acid level)**
311
312   Among the cultivated bacteriophages, the highest identity is at amino acid level (from
313 64.4 to 77.9%), and the minimum e-value was with cyanophages: *Synechococcus* phage S-
314 SSM7, *Synechococcus* phage S-SM2, *Synechococcus* phage Bellamy, and *Synechococcus* phage
315 S-CAM1. The bulk of the sequences were annotated as belonging to cyanophages (58.3 to 73%).

316 This likely indicates the active infection of their hosts with this group of viruses in the study
317 period and reflects the associated interaction of phages with cyanobacteria. In summer,
318 especially in early August, there is a seasonal peak of cyanobacterial bloom in Lake Baikal
319 (Belykh & Sorokovikova, 2003).
320   *Pelagibacter* phage HTVC008M was the second closest relative (34 to 70.1% amino acid
321 identity) in terms of the frequency of occurrence: MK – 13.7%, MC – 29.2%, Turk – 20%,
322 Posol_S – 10.8%, LT_05 – 16.7%, and LT_1015 – 14.3%.
323   Moreover, *Serratia* phage BF (YP_009599751) with identity of 65.8 to 68%,
324 *Agrobacterium* phage Atu_ph07 (YP_009611880) with identity of 63.7%, *Caulobacter* phage
325 Cr30 (YP_009098938) with identity of 58.8 to 61.5%, *Sinorhizobium* phage phiN3
326 (YP_009212304) with identity of 34.8 to 62.34%, and *Acidovorax* phage ACP17
327 (YP_009609699) with identity of 56.9% were among the closest relatives. There were no
328 sequences with 100% identity with the studied sequences among the cultivated closest relatives.
329   The protein sequences had uncultivated relatives from various ecosystems (Table 3), and
330 most relatives from the GenBank database were similar to the Baikal sequences that had been
331 previously obtained from biofilms (from 19.6 to 28.6%; identity from 67.9 to 100%) and the
332 pelagic zone (from 6.8 to 30%; identity from 75.2% to 100%) of Lake Baikal. The largest
333 number of the closest relatives for samples LT_05, LT_1015 and Turk was from lakes Bourget
334 and Annecy (from 10.8 to 25.7%; identity from 83.9 to 100%). An interesting result is the
335 identity of a large number of sequences from shallow eutrophic bays (MK, MC and Posol_S)
336 with sequences from wetland sediments (from 18.9 to 25.5%; identity from 59.4 to 97.9%) (Li et
337 al., 2018), whereas pelagic representatives from samples Turk, LT_05 and LT_1015 had only
338 from 2.8 to 3.3% of similar sequences.
339   Among uncultivated closest relatives, two sequences from other ecosystems were
340 identified with 100% identity: from wetland water (BAL45773), Posol_S -Otu35, and Lake
341 Annecy (AHU87275), LT_05-Otu22 and LT_1015-Otu25.
342   In marine viromes, 34 (HOT) and 11 (OBV) *g23* gene sequences that were the closest
343 relatives for our OTUs were detected. The identity with the sequences from HOT varied from
344 36.2% to 79.8%; the highest identity of the amino acid sequence from the HOT201114979 contig
345 corresponded to the sequence from the shallow Posolsk Sor Bay (Posol_S-Otu58). The Baikal
346 sequences, LT_05-Otu31 and LT_1015-Otu29, were closely related to sequences OBV (from
347 35.1% to 77.3%), with the highest identity of the amino acid sequence from OBV-Vall-
348 contig_N000719. The coverage for all sequences was 100%. Based on blastp analysis,
349 HOT201114979 had *Escherichia* phage 121Q as the closest relative with an identity of 65.7%
350 and coverage of 95%; for OBV-Vall-contig_N000719, the closest relative was *Synechococcus*
351 phage S-CAM1 with the identity of 96.2% and 100% coverage.
352   The identity of the *g23* sequences from the viral fraction of Lake Biwa with the
353 sequences from this study ranged from 35.3 to 98%, and the coverage – from 88 to 100%.
354 Overall, eight unique sequences were identified, which were the closest relatives. The highest
355 identity was with the MK-Otu49 sequence and a fragment from the Node_13864 contig. In the

356   bacterial fraction from Lake Biwa, 18 sequences were detected with the identity ranging from
357   33.9 to 89.1% and the coverage from 75 to 100%. The highest identity was with the Node 42706
358   fragment and the Posol_S-Otu49 sequence. The most identical *g23* fragments in the RefSeq
359   database (blastp analysis) belonged to *Pelagibacter* phage HTVC008M (YP_007517957), the
360   Node_13864 contig (identity 61.3% and coverage 99%) and *Escherichia* phage 121Q –
361   Node_42706 (identity 60.1% and coverage 100%).

362
363
364
365
366
367   **Table 3**

368
369
370   **Phylogenetic analysis**
371   Phylogenetic analysis of the amino acid *g23* sequences with the cultivated representative
372   T4-like viruses of the family *Myoviridae* and with the sequences from various natural sources
373   revealed that the sequences from this study were mixed in the tree, and 46.4% of them had <90%
374   amino acid sequence identity with the *g23* sequences from the Refseq and Genbank NR
375   databases (Fig. 2). Furthermore, 19 clusters (except for groups T-, Pseudo T- and Schizo T-
376   events) were obtained. Only the sequences from Posol_S formed five monophyletic groups
377   consisting of two or three OTUs. The bulk of the clusters contained the *g23* sequences of phages
378   from natural sources, whose hosts are still unknown because there are no cultivated
379   representatives and, hence, the confirmation of phylogenetic affiliation.
380   Most sequences from this study were similar to the sequences from freshwater
381   ecosystems. None of the sequences was included in groups T-, Pseudo T- and Schizo T-events.
382   Twelve sequences from this study were included in group Exo T-evens containing cultivated
383   cyanophages and the HTVC008M pelagiphages. Also, this group included most of the *g23*
384   sequences from viromes (94.4%).

385
386
387   **Figure 2.**

388
389
390   **Biogeography of *g23* sequences**
391   **Regional level**
392   To analyse relationships of the *g23* sequences at the local level in the Baikal region, we built the
393   network that included various biotopes in Lake Baikal and plankton of Lake Kotokel (Fig. 3).
394   Based on the analysis, clusters from Lake Kotokel and biofilms of stone substrates, 2D1 and
395   1D3Tp, of Lake Baikal are remote from other Baikal clusters. These clusters do not have shared

396   OTUs (clustering of nucleotide sequences at 97% level) with other samples. As mentioned
397   above, the sequences from the eutrophic Lake Kotokel were more similar to the sequences from
398   Lake Donghu (China) than to the sequences from Lake Baikal (Butina et al., 2013). Sequences
399   from the biofilm samples, 1D3Tp and 2D1, taken in the southern basin of Lake Baikal from
400   depths of 9 and 36 m, which consisted mainly of cyanobacteria, formed separate clusters in the
401   tree (Potapov et al., 2020).
402   Samples MK and MC from Mukhor Bay (Maloye More Strait, the west coast) taken at nearby
403   stations with similar hydrochemical and hydrophysical parameters are close to each other and
404   have more shared OTUs than with sample Posol_S taken from Posolsk Sor Bay on the east coast
405   of the lake.
406   Samples from the coastal site (Turk) and the pelagic station (LT_05 and LT_1015) cluster
407   together and have more shared OTUs with the sequences from the pelagic water (BSOTU).
408   Samples LT_05 and LT_1015 have more similar OTUs (ten shared OTUs) with previously
409   obtained pelagic samples BSOTU, S0508 (southern basin) and N0508 (northern basin) than with
410   samples from the shallow bays, MC and MK (three shared OTUs), the coastal site, Turk (six
411   shared OTUs), and biofilms, 2D1, 1D3Tp, 2D2, 2D4, 2a, LS, and Neuston (seven shared OTUs).
412
413
414   **Figure 3.**
415
416
417   **Global level**
418   For comparative biogeographical analysis, we selected sequences available in the
419   GenBank NR database (Table 4). All of them were obtained using primers MZIA1bis and
420   MZIA6. The geography of the samples in the databases covers a wide range of climate zones and
421   habitats, ensuring a rather large-scale analysis of phages for the *g23* gene fragment. Sampling
422   sites are shown in Figure 4.
423
424   **Table 4**
425
426
427   **Figure 4.**
428
429
430   The *g23* sequences in this study and those represented in the GenBank NR database
431   formed groups on the dendrogram: marine, Baikal cluster I, Baikal cluster II, soil, and sediments
432   (Fig. 5).
433   LT_05 and LT_1015, sequences from the samples collected at the central station of the
434   Listvyanka settlement – the Tankhoy settlement section, layer from 0 + 5 m and from 10 + 15 m,
435   are located closer to the site with the sequences from sample Turk as mentioned above, which is

436 due to the mixing with pelagic waters. LT_05, LT_1015 and Turk form a shared cluster together
437 with the sequences previously obtained from the pelagic zone of Lake Baikal (BSOTU) (Potapov
438 et al., 2018). Three samples from shallow Baikal bays (MK, MC and Posol_S) group together,
439 confirming clustering by the similar trophic conditions within the Baikal group.

440
441
442
443 **Figure 5.**

444
445 The sequences from biofilms inhabiting stones (LS, 2a, 2D1, and 1D3Tp), neuston and
446 sponges (2D2 and 2D4) (Potapov et al., 2020) form a shared cluster with planktonic *g23*
447 sequences from the pelagic zone of the southern and northern basins of Lake Baikal. Samples
448 from Lake East, Lake Donghu and Kongsfjorden also form a shared cluster.
449 The greatest identity of the Baikal *g23* is observed with the sample from the subalpine lakes
450 Bourget (oligo-mesotrophic) and Annecy (oligotrophic) (Zhong & Jacquet, 2014), which we also
451 previously indicated in our study (Potapov et al., 2018).
452 The sequences from the Arctic proglacial lakes and Lake Limnopolar are closer to the sequences
453 from the dairy water (Ireland) and wetland water of China.

454
455 **Discussion**

456 Here, we studied bacteriophages associated with bacteria in a fraction greater than 0.2 μm
457 from shallow bays, coastal zone and pelagic zone of Lake Baikal based on the analysis of the *g23*
458 gene. The sampling sites were chosen not only to carry out a comparative analysis in terms of
459 geographical distance but also to reveal the differences in the pools of the *g23* sequences selected
460 in different zones of Lake Baikal with different trophic states.
461 The highest concentration of the total phosphorus, total nitrogen and chlorophyll *a* has been
462 documented in shallow bays, which is expectable for Lake Baikal because the maximum rate of
463 formation of organic matter occurs here due to a large number of primary producers
464 (cyanobacteria) (Watanabe & Drucker, 1999; Belykh & Sorokovikova, 2003). Temperature, in
465 turn, is an important factor, influencing the growth rate of bacteria and having a significant
466 positive effect on bacterial production (Straškrábová et al., 2005). An increase in temperature
467 stimulates the development of phytoplankton, as well as enhances its photosynthetic activity, and
468 the water body is enriched with dissolved organic matter. Previously, it has been shown that
469 temperature controls the number of viruses, their diversity and the activity of hosts (Danovaro et
470 al., 2011).
471 The pH level is slightly shifted to the alkaline side, which is typical for the waters of Lake
472 Baikal. According to long-term observations, the pH value of water in Lake Baikal varies from
473 7.7 to 8.3, and during the intense vegetation of picoplanktonic cyanobacteria, it rises to 8.7
474 (Votintsev, 1961; Khodzher et al., 2018). As shown previously, a key factor in determining the

475   infectivity of a virus is the pH value. For example, a low pH (<4) significantly reduces phage
476   survival (Jurczak-Kurek et al., 2016).
477   In the network analysis, the sequences from Posolsk Sor occupy an intermediate position
478   between planktonic (LT_05 and LT_1015) and coastal sites (Turk) as well as from shallow
479   Mukhor Bay (MK and MC). The distribution pattern can be explained both by the dilution of
480   water in Posolsk Sor with waters of rivers flowing into it and by its closed nature: two sandy
481   spits separate the bay from the lake, and, in contrast to the well-flowing Mukhor Bay, it is
482   connected with the lake only by a narrow 700 m long channel. With this connection, the
483   sequences from Posolsk Sor are separated from the sequences from the Muhor Bay, despite the
484   close trophic state of these bays. At the same time, in the UPGMA analysis, they are rather
485   included in the shared cluster with the samples from Mukhor Bay than with the sequences from
486   the pelagic and coastal sites.
487   Late July and early August is a period of bloom of both pico- and nanoplanktonic
488   cyanobacteria in Lake Baikal. Picocyanobacteria are found in huge numbers in Lake Baikal,
489   reaching an abundance of 1.5 million cells/mL (Belykh & Sorokovikova, 2003). In August 2019,
490   there was a mass development of diazotrophic cyanobacteria in the waters of the bays:
491   *Gloeotrichia echinulata* bloomed in Posolsk Sor Bay, and the species of the genus
492   *Dolichospermum* – in Mukhor Bay. Intensive vegetation of *Dolichospermum* species has been
493   documented for a long time from June to September in all parts of Lake Baikal, with the
494   maximum concentration of up to 10 million cells/L in bays (Popovskaya, 2000). In this regard, it
495   is logical to assume the presence of cyanophages, natural regulators of the number of
496   cyanobacteria. As mentioned above, many sequences belonged to cyanophages according to
497   RefSeq, but phylogenetic analysis revealed only 12 sequences included in the Exo T-evens group
498   (cyano- and pelagiphages). In addition to cyanophages, *Pelagibacter* phage HTVC008M (the
499   family *Myoviridae*) infecting *Candidatus Pelagibacter* ubique (Alphaproteobacteria, freshwater
500   SAR11) was the closest relative of the Baikal representatives of T4-phages. Previously, high
501   synteny with this strain of the Baikal sequences was determined (Cabello-Yeves et al., 2018).
502   To assess the similarity of the Baikal sequences with marine phages, we added the *g23*
503   sequences of phages from station ALOHA (Luo et al., 2020), the sequences from viromes
504   obtained from the coastal water of the eutrophic Osaka Bay and viromes (viral and bacterial
505   fraction) of mesotrophic freshwater Lake Biwa. These sequences are absent in the NR and
506   RefSeq databases, which narrows the range of geographic representation of virome communities
507   in aquatic ecosystems. None of these sequences had 100% identity with OTUs obtained from the
508   bacterial fraction of Lake Baikal, although, in general, the maximum identity was higher with the
509   representatives from the ancient Lake Biwa (up to 89.1% identity in the bacterial fraction and up
510   to 98% identity in the viral fraction) than with marine ones from station ALOHA and Osaka Bay
511   (the maximum of 79.8% identity). This was expectable because phages of freshwater ecosystems
512   are more closely related to each other than to marine phages. The sequences from the ancient
513   lakes Biwa and Baikal also did not form separate clusters in the phylogenetic tree but
514   demonstrated high diversity, distributing throughout the tree. Perhaps, this may be due to the

515   method of obtaining the sequences or to their number. Overall, at this stage, it is difficult to
516   identify, with which fraction the studied sequences are more identical: i) the median identity at
517   the amino acid level was slightly higher with the bacterial fraction from Lake Biwa (data not
518   shown); ii) slightly more unique sequences were detected from the bacterial fraction; iii) based
519   on blastp analysis, the highest level of identity was identified with the sequence from the viral
520   fraction. In the future, to elucidate the similarities between the sequences from the ancient lakes,
521   it is necessary to carry out a detailed comparison with other ancient lakes and obtain the *g23*
522   sequences by one method.
523   In addition to the sequences from marine viromes, the cluster with 100% support
524   included one sequence from the uncultivated marine bacteriophage (ABV27494) and, based on
525   blastp analysis, it had the maximum identity of 75% with one OTU from the Turk sample
526   (Turk_OTU-10). Therefore, this can confirm the hypothesis about the common origin, but the
527   conditions of formation of marine and freshwater viral communities differ significantly,
528   primarily in the salinity gradient.
529   Noteworthy is a great number of similar sequences in samples from shallow areas of Lake
530   Baikal (from 18.9 to 25.5%) and wetland sediments (Li et al., 2018), which is much greater than
531   in the samples from the pelagic zone of Lake Baikal and wetland sediments (from 2.9 to 3.3%).
532   This possibly indicates the similar composition and conditions for the existence of bacterial
533   communities from the Baikal bays (shallow water, high productivity and elevated temperature)
534   and the shallow well-warmed productive wetland sediments.
535   A large number of the *g23* sequences from lakes Bourget and Annecy similar to the studied
536   Baikal sequences is likely owing to the identity of hydrophysical and hydrochemical parameters
537   (altitude above sea level, total P, total N, nitrates and pH) of the lake, as we previously indicated
538   in the analysis of the *g23* sequences from the pelagic zone of Lake Baikal (Potapov et al., 2018).
539   Only two sequences had 100% identity with the Baikal sequences (from Lake Annecy (France)
540   and wetland water (China)), which casts doubt on the "everything is everywhere, but the
541   environment selects" theory, probably, because viruses quickly mutate.
542   According to (Chow & Suttle, 2015), four models define biogeography of marine
543   microbiomes and can be applied to viruses: i) local or regional factors do not determine the
544   community composition; ii) the conditions of the local habitat determine the community
545   structure where differences in taxa determine the changes in the community structure (i.e.
546   everything is everywhere but the environment selects for the viability of a community structure);
547   iii) community structure is determined by the extent, to which taxa can migrate to a particular
548   location (i.e., dispersal limitation); iv) dispersal limitation and environmental conditions co-
549   regulate community composition. The UPGMA analysis has revealed that viral diversity does
550   not follow the latitudinal and longitudinal gradients, and the geographical distance does not
551   influence the composition of bacteriophages. In our case, the fourth model, "the limitation of
552   distribution and environmental conditions determine the community composition", is the most
553   applicable. Because Lake Baikal was formed separately for many millions of years, determining
554   the composition and the structure of aquatic organisms, this conclusion confirms the separate

555 formation of the viral *g23* communities in Lake Baikal and, possibly, endemism of
556 bacteriophages in the Earth's oldest lake. Cluster Baikal II (August), irrespective of the fractions
557 (bacterial or viral), includes the samples taken in August, which coincide with the cyanobacterial
558 bloom period and likely testify to the seasonal division of the clusters. Cluster Baikal I (May to
559 June), in addition to neuston, consisted of the samples taken during the spring homothermy when
560 the water column of the lake is actively mixing, and diatoms dominate the plankton (Bondarenko
561 et al., 2019). Neuston is a very specific biotope, and its bacterial and viral composition differs
562 significantly from the water column (Galach'yants et al., 2017; Drucker et al., 2020). Perhaps,
563 for this reason, it was not included in Baikal cluster II. Notably, cluster II consists of the
564 sequences obtained by high throughput sequencing, and cluster I – of the sequences obtained by
565 Sanger sequencing. Possibly, a method of obtaining sequences to some extent affects the results;
566 however, we believe that it does not determine the nature of clustering. The marine sample taken
567 from Kongsfjorden appeared to be in the clade with eutrophic freshwater lakes. Kongsfjorden is
568 exposed to the fresh water influx from several fronts of glaciers of the Svalbard archipelago
569 (Hagen et al., 2003), causing its desalination and likely bringing bacteria and viruses from the
570 land along with meltwater. Previously, it has been shown that some bacterial sequences from
571 glaciers demonstrate a close identity with those from soils of the temperate zone and humic
572 lakes, i.e. represent cosmopolitan taxa (Edwards et al., 2011). The sample was taken from the
573 glacier's surface 2 km from the shore, and the mixing of phage communities between glacial and
574 marine habitats was expectable, taking into account their close relationship (Bellas & Anesio,
575 2013). In general, the main characteristic of the ecosystem (marine, soil or freshwater)
576 substantiates the pattern of distribution of the samples in the UPGMA analysis. We did not
577 include the sequences from viromes in the UPGMA analysis because they were compared to our
578 sequences via blastp analysis and, perhaps, the most distant ones were not detected, i.e. they may
579 not represent the full range of diversity.
580 In the future, to confirm the "uniqueness" of the Baikal viral communities, the diversity of
581 viruses in lakes close in age and origin should be compared, for example, Lake Hovsgol
582 (Mongolia) that is close to Lake Baikal, whose age is ca. 5 million years, or geographically very
583 remote Lake Tanganyika (Africa), the age of which is ca. 12 million years (Padilla, 2007;
584 Hampton et al., 2018). Analysis of viral diversity can elucidate the evolution and formation of
585 viral diversity in the lakes that have been isolated for many millions of years.
586

Conclusions

588   In this study, we obtained the *g23* sequences of planktonic bacteriophages from two
589 shallow bays, as well as coastal and pelagic sites of Lake Baikal, which are part of the bacterial
590 fraction. The *g23* sequences were highly diverse and differed from those in other ecosystems.
591 The bulk of the sequences from the RefSeq database had closest relatives among cyanophages.
592 In the network analysis, the sample from Posolsk Sor occupied the intermediate position between
593 the pelagic and coastal samples as well as the sample from Mukhor Bay, again indicating a close
594 relationship between viruses and hosts because different species of cyanobacteria were present in

595 these bays. The sequences from marine viromes differed more than those from the freshwater
596 lake Biwa, which testifies to the close relationship of freshwater viruses and their difference
597 from marine ones. The formation of clusters in the UPGMA analysis allowed us to conclude that
598 the *g23* sequences of phages from shallow bays are rather similar to each other than to the
599 sequences from the coastal and pelagic sites of Lake Baikal, i.e. they cluster as provided by the
600 productivity of the sites. In the global UPGMA analysis, all sequences from Lake Baikal formed
601 two clusters depending on the season and were different from the sequences obtained from the
602 seas, soil and sediments, which may indicate the separate formation of viral community in the
603 ancient lake.

604

Acknowledgements

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

610

611

612

References

614 Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko
615   SI, Pham S, Prjibelski AD, Pyshkin A V., Sirotkin A V., Vyahhi N, Tesler G, Alekseyev
616   MA, Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its applications to
617   single-cell sequencing. *Journal of Computational Biology* 19:455–477. DOI:
618   10.1089/cmb.2012.0021.
619 Bellas CM, Anesio AM. 2013. High diversity and potential origins of T4-type bacteriophages on
620   the surface of Arctic glaciers. *Extremophiles* 17:861–870. DOI: 10.1007/s00792-013-0569-
621   x.
622 Besemer J. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial
623   genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids*
624   *Research* 29:2607–2618. DOI: 10.1093/nar/29.12.2607.
625 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence
626   data. *Bioinformatics (Oxford, England)* 30:2114–2120. DOI:
627   10.1093/bioinformatics/btu170.
628 Bondarenko NA, Ozersky T, Obolkina LA, Tikhonova IV, Sorokovikova EG, Sakirko MV,
629   Potapov SA, Blinov VV, Zhdanov AA, Belykh OI. 2019. Recent changes in the spring
630   microplankton of Lake Baikal, Russia. *Limnologica* 75. DOI: 10.1016/j.limno.2019.01.002.
631 Buchfink B, Xie C, Huson DH. 2014. Fast and sensitive protein alignment using DIAMOND.
632   *Nature Methods* 12:59–60. DOI: 10.1038/nmeth.3176.
633 Butina TV, Belykh OI, Maksimenko SY, Belikov SI. 2010. Phylogenetic diversity of T4-like
634   bacteriophages in Lake Baikal, East Siberia. *FEMS Microbiology Letters* 309:122–129.
635   DOI: 10.1111/j.1574-6968.2010.02025.x.

636   Butina TV, Belykh OI, Potapov SA, Sorokovikova EG. 2013. Diversity of the major capsid
637   genes (g23) of T4-like bacteriophages in the eutrophic Lake Kotokel in East Siberia, Russia.
638   *Archives of Microbiology* 195:513–520. DOI: 10.1007/s00203-013-0884-8.
639   Danovaro R, Corinaldesi C, Dell'anno A, Fuhrman JA, Middelburg JJ, Noble RT, Suttle CA.
640   2011. Marine viruses and global climate change. *FEMS microbiology reviews* 35:993–1034.
641   DOI: 10.1111/J.1574-6976.2010.00258.X.
642   Drucker VV, Potapov SA, Gorshkova AS, Belykh OI. 2020. *Bacteriophages of Lake Baikal*
643   *(Bakteriofagi ozera Baykal)*. Novosibirsk: Russian Academy of Sciences, Siberian Branch.
644   Edwards A, Anesio AM, Rassner SM, Sattler B, Hubbard B, Perkins WT, Young M, Griffith
645   GW. 2011. Possible interactions between bacterial diversity, microbial activity and
646   supraglacial hydrology of cryoconite holes in Svalbard. *The ISME Journal* 5:150–160. DOI:
647   10.1038/ismej.2010.100.
648   Galach'yants AD, Bel'kova NL, Sukhanova E V, Galach'yants YP, Morozov AA, Parfenova V
649   V. 2017. Taxonomic Composition of Lake Baikal Bacterioneuston Communities. 86:241–
650   249. DOI: 10.1134/S0026261717020084.
651   Galaziy GI (ed.). 1993. *Atlas. Baikal*. (In Russian, Russian Academy of Sciences, Siberian
652   branch. Moscow: Federal Service of geodesy and cartography.
653   Hagen JO, Kohler J, Melvold K, Winther J-G. 2003. Glaciers in Svalbard: mass balance, runoff
654   and freshwater flux. *Polar Research* 22:145–159. DOI: 10.3402/polar.v22i2.6452.
655   Hampton SE, McGowan S, Ozersky T, Virdis SGP, Vu TT, Spanbauer TL, Kraemer BM, Swann
656   G, Mackay AW, Powers SM, Meyer MF, Labou SG, O'Reilly CM, DiCarlo M, Galloway
657   AWE, Fritz SC. 2018. Recent ecological change in ancient lakes. *Limnology and*
658   *Oceanography* 63:2277–2304. DOI: 10.1002/LNO.10938.
659   He M, Cai L, Zhang C, Jiao N, Zhang R. 2017. Phylogenetic diversity of T4-type phages in
660   sediments from the subtropical Pearl River Estuary. *Frontiers in Microbiology* 8. DOI:
661   10.3389/fmicb.2017.00897.
662   Hohmann R, Kipfer R, Peeters F, Piepke G, Imboden DM, Shimaraev MN. 1997. Processes of
663   deep-water renewal in Lake Baikal. *Limnology and Oceanography* 42:841–855. DOI:
664   10.4319/LO.1997.42.5.0841.
665   Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees.
666   *Bioinformatics* 17:754–755.
667   Khodzher TV, Domysheva VM, Sorokovikova LM, Tomberg IV, Sakirko MV. 2018.
668   Hydrochemical studies in Lake Baikal: history and nowadays. *Limnology and Freshwater*
669   *Biology*:2–9. DOI: 10.31951/2658-3518-2018-A-1-2.
670   Kozhova OM, Izmest'eva LR. 1998. *Lake Baikal: Evolutionand Biodiversity*.
671   Li W, Godzik A. 2006. Cd-hit: A fast program for clustering and comparing large sets of protein
672   or nucleotide sequences. *Bioinformatics* 22:1658–1659. DOI:
673   10.1093/bioinformatics/btl158.
674   Luo E, Eppley JM, Romano AE, Mende DR, DeLong EF. 2020. Double-stranded DNA
675   virioplankton dynamics and reproductive strategies in the oligotrophic open ocean water
676   column. *The ISME Journal 2020 14:5* 14:1304–1315. DOI: 10.1038/s41396-020-0604-8.
677   Millard AD, Pearce D, Zwirglmaier K. 2016. Biogeography of bacteriophages at four
678   hydrothermal vent sites in the Antarctic based on g23 sequence diversity. *FEMS*
679   *Microbiology Letters* 363:1–12. DOI: 10.1093/femsle/fnw043.
680   Mruwat N, Carlson MCG, Goldin S, Ribalet F, Kirzner S, Hulata Y, Beckett SJ, Shitrit D, Weitz
681   JS, Armbrust EV, Lindell D. 2021. A single-cell polony method reveals low levels of

682   infected Prochlorococcus in oligotrophic waters despite high cyanophage abundances. *The*
683   *ISME Journal* 15:41–54. DOI: 10.1038/s41396-020-00752-6.
684   Nishimura Y, Watai H, Honda T, Mihara T, Omae K, Roux S, Blanc-Mathieu R, Yamamoto K,
685   Hingamp P, Sako Y, Sullivan MB, Goto S, Ogata H, Yoshida T. 2017. Environmental Viral
686   Genomes Shed New Light on Virus-Host Interactions in the Ocean. *mSphere* 2. DOI:
687   10.1128/MSPHERE.00359-16.
688   Okazaki Y, Nishimura Y, Yoshida T, Ogata H, Nakano S. 2019. Genome - resolved viral and
689   cellular metagenomes revealed potential key virus - host interactions in a deep freshwater
690   lake. *Environmental Microbiology* 21:4740–4754. DOI: 10.1111/1462-2920.14816.
691   Padilla DK. 2007. The Geology, Biodiversity and Ecology of Lake Hövsgöl (Mongolia). Biology
692   of Inland Waters. Edited by Clyde E Goulden, Tatiana Sitnikova, Jon Gelhaus, and ,
693   Bazartseren Boldgiv. Leiden (The Netherlands): Backhuys Publishers . €176.00. xxi + 525
694   p; ill.; s. *The Quarterly Review of Biology* 82:69–70. DOI: 10.1086/513386.
695   Potapov S, Belykh O, Krasnopeev A, Galachyants A, Podlesnaya G, Khanaev I, Tikhonova I.
696   2020. Diversity and biogeography of bacteriophages in biofilms of Lake Baikal based on
697   g23 sequences. *Journal of Great Lakes Research* 46. DOI: 10.1016/j.jglr.2019.06.007.
698   Potapov S, Belykh O, Krasnopeev A, Gladkikh A, Kabilov M, Tupikin A, Butina T. 2018.
699   Assessing the diversity of the g23 gene of T4-like bacteriophages from Lake Baikal with
700   high-throughput sequencing. *FEMS microbiology letters* 365. DOI: 10.1093/femsle/fnx264.
701   Potapov SA, Butina T V., Belykh OI, Belikov SI. 2013. Genetic diversity of T4-like
702   bacteriophages in Lake Baikal. *«The Bulletin of Irkutsk State University». Series «Biology.*
703   *Ecology».* 3:14–19.
704   Shannon P. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular
705   Interaction Networks. *Genome Research* 13:2498–2504. DOI: 10.1101/gr.1239303.
706   Votintsev K. 1961. Water chemistry of Lake Baikal. *Trudy Baikalskoi limnologicheskoi stantsii*
707   *AN SSSR [Proceedings of the Baikal Limnological Station of the Academy of Sciences of the*
708   *USSR] (In Russia)* 20:1–311.
709   Warwick-Dugdale J, Buchholz HH, Allen MJ, Temperton B. 2019. Host-hijacking and
710   planktonic piracy: how phages command the microbial high seas. *Virology Journal* 16:15.
711   DOI: 10.1186/s12985-019-1120-1.
712   Zhang QY, Gui JF. 2018. Diversity, evolutionary contribution and ecological roles of aquatic
713   viruses. *Science China Life Sciences* 61:1486–1502. DOI: 10.1007/s11427-018-9414-7.
714   Zhao Y, Temperton B, Thrash JC, Schwalbach MS, Vergin KL, Landry ZC, Ellisman M,
715   Deerinck T, Sullivan MB, Giovannoni SJ. 2013. Abundant SAR11 viruses in the ocean.
716   *Nature* 494:357–360. DOI: 10.1038/nature11921.
717
718   Ackermann H-W, Krisch HM. 1997. A catalogue of T4-type bacteriophages. Archives of
719   Virology 142:2329–2345. DOI: 10.1007/s007050050246.
720   Adnan M, Khan S, Patel M, Al-Shammari E, Ashankyty IMA. 2013. Agrobacterium: a potent
721   human pathogen. Reviews in Medical Microbiology 24:94–97. DOI:
722   10.1097/MRM.0b013e3283642449.
723   Adriaenssens EM, Cowan DA. 2014. Using signature genes as tools to assess environmental
724   viral ecology and diversity. Applied and Environmental Microbiology 80:4470–4480. DOI:
725   10.1128/AEM.00878-14.

726 Andrews S. 2010.FastQC: a quality control tool for high throughput sequence data. Available at
727 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
728 Auguet JC, Montanié H, Lebaron P. 2006. Structure of virioplankton in the charente estuary
729 (France): Transmission electron microscopy versus pulsed field gel electrophoresis. In:
730 Microbial Ecology. Springer, 197–208. DOI: 10.1007/s00248-005-0043-0.
731 Aylward FO, Boeuf D, Mende DR, Wood-Charlson EM, Vislova A, Eppley JM, Romano AE,
732 DeLong EF. 2017. Diel cycling and long-term persistence of viruses in the ocean's euphotic
733 zone. Proceedings of the National Academy of Sciences of the United States of America
734 114:11446–11451. DOI: 10.1073/pnas.1714821114.
735 Belykh OI, Sorokovikova EG. 2003. Autotrophic picoplankton in Lake Baikal: abundance,
736 dynamics, and distribution. Aquatic Ecosystem Health & Management 6:251–261. DOI:
737 10.1080/14634980301489.
738 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence
739 data. Bioinformatics (Oxford, England) 30:2114–2120. DOI: 10.1093/bioinformatics/btu170.
740 Buerger P, Weynberg KD, Wood-Charlson EM, Sato Y, Willis BL, van Oppen MJH. 2018.
741 Novel T4 bacteriophages associated with black band disease in corals. Environmental
742 Microbiology. DOI: 10.1111/1462-2920.14432.
743 Butina TV, Belykh OI, Maksimenko SY, Belikov SI. 2010. Phylogenetic diversity of T4-like
744 bacteriophages in Lake Baikal, East Siberia. FEMS Microbiology Letters 309:122–129. DOI:
745 10.1111/j.1574-6968.2010.02025.x.
746 Butina TV, Belykh OI, Potapov SA, Sorokovikova EG. 2013. Diversity of the major capsid
747 genes (g23) of T4-like bacteriophages in the eutrophic Lake Kotokel in East Siberia, Russia.
748 Archives of Microbiology 195:513–520. DOI: 10.1007/s00203-013-0884-8.
749 Butina TV, Usova OS, Potapov SA, Belykh OI, Fedotov AP, Belikov SI. 2015. Molecular
750 genetic study of T4-like bacteriophages in plankton of Selenga River. Seriya «Biologiya.
751 Ekologiya» of the «Izvestiya Irkutskogo gosudarstvennogo universiteta» 12:12–22.
752 Cabello-Yeves PJ, Zemskay TI, Rosselli R, Coutinho FH, Zakharenko AS, Blinov V V.,
753 Rodriguez-Valera F. 2018. Genomes of novel microbial lineages assembled from the sub-ice
754 waters of Lake Baikal. Applied and Environmental Microbiology 84. DOI:
755 10.1128/AEM.02132-17.
756 Cahyani VR, Murase J, Ishibashi E, Asakawa S, Kimura M. 2009. T4-type bacteriophage
757 communities estimated from the major capsid genes (g23) in manganese nodules in Japanese
758 paddy fields. Soil Science and Plant Nutrition 55:264–270. DOI: 10.1111/j.1747-
759 0765.2009.00363.x.
760 Cai L, Zhang R, He Y, Feng X, Jiao N. 2016. Metagenomic analysis of Virioplankton of the
761 subtropical Jiulong river estuary, China. Viruses 8:1–13. DOI: 10.3390/v8020035.
762 Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña
763 AG, Goodrich JK, Gordon JI, Huttley G a, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone
764 C a, Mcdonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters W a,
765 Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-

766 throughput community sequencing data Intensity normalization improves color calling in SOLiD
767 sequencing. Nature Publishing Group 7:335–336. DOI: 10.1038/nmeth0510-335.
768 De Cárcer DA, Pedrós-Alió C, Pearce DA, Alcamí A. 2016. Composition and interactions
769 among bacterial, microeukaryotic, and T4-like viral assemblages in lakes from both polar zones.
770 Frontiers in Microbiology 7:1–11. DOI: 10.3389/fmicb.2016.00337.
771 Comeau AM, Krisch HM. 2008. The capsid of the T4 phage superfamily: The evolution,
772 diversity, and structure of some of the most prevalent proteins in the biosphere. Molecular
773 Biology and Evolution 25:1321–1332. DOI: 10.1093/molbev/msn080.
774 Coutinho FH, Cabello-Yeves PJ, Gonzalez-Serrano R, Rosselli R, López-Pérez M, Zemskaya TI,
775 Zakharenko AS, Ivanov VG, Rodriguez-Valera F. 2020. New viral biogeochemical roles
776 revealed through metagenomic analysis of Lake Baikal. Microbiome 8:163. DOI:
777 10.1186/s40168-020-00936-4.
778 Desplats C, Krisch HM. 2003. The diversity and evolution of the T4-type bacteriophages.
779 Research in Microbiology 154:259–267. DOI: http://dx.doi.org/10.1016/S0923-2508(03)00069-
780 X.
781 Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high
782 throughput. Nucleic Acids Research 32:1792–1797. DOI: 10.1093/nar/gkh340.
783 Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics
784 26:2460–2461. DOI: 10.1093/bioinformatics/btq461.
785 Filée J, Tétart F, Suttle CA, Krisch HM. 2005. Marine T4-type bacteriophages, a ubiquitous
786 component of the dark matter of the biosphere. Proceedings of the National Academy of
787 Sciences of the United States of America 102:12471–6. DOI: 10.1073/pnas.0503404102.
788 Fujihara S, Murase J, Tun CC, Matsuyama T, Ikenaga M, Asakawa S, Kimura M. 2010. Low
789 diversity of T4-type bacteriophages in applied rice straw, plant residues and rice roots in
790 Japanese rice soils: Estimation from major capsid gene (g23) composition. Soil Science and
791 Plant Nutrition 56:800–812. DOI: 10.1111/j.1747-0765.2010.00513.x.
792 Garin-fernandez A, Pereira-flores E, Oliver F, Wichels A. 2018. Marine Genomics The North
793 Sea goes viral : Occurrence and distribution of North Sea bacteriophages. Marine Genomics
794 41:31–41. DOI: 10.1016/j.margen.2018.05.004.
795 Goldsmith DB, Brum JR, Hopkins M, Carlson CA, Breitbart M. 2015. Water column
796 stratification structures viral community composition in the Sargasso Sea. Aquatic Microbial
797 Ecology 76:85–94. DOI: 10.3354/ame01768.
798 Gong Z, Liang Y, Wang M, Jiang Y, Yang Q, Xia J, Mcminn A. 2018. Viral diversity and its
799 relationship with environmental factors at the surface and deep sea of Prydz Bay , Antarctica.
800 9:1–17. DOI: 10.3389/fmicb.2018.02981.
801 Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, Ardyna M,
802 Arkhipova K, Carmichael M, Cruaud C, Dimier C, Domínguez-Huerta G, Ferland J, Kandels S,
803 Liu Y, Marec C, Pesant S, Picheral M, Pisarev S, Poulain J, Tremblay JÉ, Vik D, Acinas SG,
804 Babin M, Bork P, Boss E, Bowler C, Cochrane G, de Vargas C, Follows M, Gorsky G, Grimsley
805 N, Guidi L, Hingamp P, Iudicone D, Jaillon O, Kandels-Lewis S, Karp-Boss L, Karsenti E, Not

806 F, Ogata H, Poulton N, Raes J, Sardet C, Speich S, Stemmann L, Sullivan MB, Sunagawa S,
807 Wincker P, Culley AI, Dutilh BE, Roux S. 2019. Marine DNA viral macro- and microdiversity
808 from pole to pole. Cell 177:1109–1123.e14. DOI: 10.1016/j.cell.2019.03.040.

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

812 He M, Cai L, Zhang C, Jiao N, Zhang R. 2017. Phylogenetic diversity of T4-type phages in
813 sediments from the subtropical Pearl River Estuary. Frontiers in Microbiology 8. DOI:
814 10.3389/fmicb.2017.00897.

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

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

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

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

823 Jia Z, Ishihara R, Nakajima Y, Asakawa S, Kimura M. 2007. Molecular characterization of T4-
824 type bacteriophages in a rice field. Environmental Microbiology 9:1091–1096. DOI:
825 10.1111/j.1462-2920.2006.01207.x.

826 Jurczak-Kurek A, Gąsior T, Nejman-Faleńczyk B, Bloch S, Dydecka A, Topka G, Necel A,
827 Jakubowska-Deredas M, Narajczyk M, Richert M, Mieszkowska A, Wróbel B, Węgrzyn G,
828 Węgrzyn A. 2016. Biodiversity of bacteriophages: morphological and biological properties of a
829 large group of phages isolated from urban sewage. Scientific Reports 6:34338. DOI:
830 10.1038/srep34338.

831 Kavagutti VS, Andrei A-Ş, Mehrshad M, Salcher MM, Ghai R. 2019. Phage-centric ecological
832 interactions in aquatic ecosystems revealed through ultra-deep metagenomics. Microbiome
833 7:135. DOI: 10.1186/s40168-019-0752-0.

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

837 Krishnamurthy SR, Wang D. 2017. Origins and challenges of viral dark matter. Virus Research
838 239:136–142. DOI: 10.1016/j.virusres.2017.02.002.

839 Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis
840 version 7.0 for bigger datasets. Molecular biology and evolution 33:msw054. DOI:
841 10.1093/molbev/msw054.

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

844   Li Y, Liu H, Pan H, Zhu X, Liu C, Zhang Q, Luo Y, Di H, Xu J. 2019. T4-type viruses:
845   Important impacts on shaping bacterial community along a chronosequence of 2000-year old
846   paddy soils. Soil Biology and Biochemistry 128:89–99. DOI: 10.1016/j.soilbio.2018.10.007.
847   Li X, Sun Y, Liu J, Yao Q, Wang G. 2018.Myoviridae environmental samples. Available at
848   https://www.ncbi.nlm.nih.gov/popset?LinkName=protein_popset&from_uid=1496533459
849   Liu L, Cai L, Zhang R. 2017. Co-existence of freshwater and marine T4-like myoviruses in a
850   typical subtropical estuary. FEMS Microbiology Ecology 93. DOI: 10.1093/femsec/fix119.
851   Mabizela NB, Litthauer D. 2016.Unculturable T4-like and T7-like phages from South African
852   deep gold mines. Available at
853   https://www.ncbi.nlm.nih.gov/popset?LinkName=protein_popset&from_uid=256033955
854   Millard AD, Pearce D, Zwirglmaier K. 2016. Biogeography of bacteriophages at four
855   hydrothermal vent sites in the Antarctic based on g23 sequence diversity. FEMS Microbiology
856   Letters 363:1–12. DOI: 10.1093/femsle/fnw043.
857   Murtagh F. 1992. Comments on "Parallel algorithms for hierarchical clustering and cluster
858   validity." IEEE Transactions on Pattern Analysis and Machine Intelligence 14:1056–1057. DOI:
859   10.1109/34.159908.
860   Okazaki Y, Nishimura Y, Yoshida T, Ogata H, Nakano S. 2019. Genome - resolved viral and
861   cellular metagenomes revealed potential key virus - host interactions in a deep freshwater lake.
862   Environmental Microbiology 21:4740–4754. DOI: 10.1111/1462-2920.14816.
863   De Paepe M, Taddei F. 2006. Viruses' life history: towards a mechanistic basis of a trade-off
864   between survival and reproduction among phages. PLoS Biology 4:e193. DOI:
865   10.1371/journal.pbio.0040193.
866   Palermo CN, Fulthorpe RR, Saati R, Short SM. 2019. Metagenomic analysis of virus diversity
867   and relative abundance in a eutrophic freshwater Harbour. Viruses 11:792. DOI:
868   10.3390/v11090792.
869   Parvathi A, Zhong X, Jacquet S. 2012. Dynamics of various viral groups infecting autotrophic
870   plankton in Lake Geneva. Advances in Oceanography and Limnology 3:171–191. DOI:
871   10.1080/19475721.2012.738157.
872   Paul JH, Jiang SC, Rose JB. 1991. Concentration of viruses and dissolved DNA from aquatic
873   environments by vortex flow filtration. Applied and Environmental Microbiology 57:2197–2204.
874   DOI: 10.1128/aem.57.8.2197-2204.1991.
875   Popovskaya G. 2000. Ecological monitoring of phytoplankton in Lake Baikal. Aquatic
876   Ecosystem Health and Management 3:215–225. DOI: 10.1016/S1463-4988(00)00021-X.
877   Potapov S, Belykh O, Krasnopeev A, Galachyants A, Podlesnaya G, Khanaev I, Tikhonova I.
878   2020. Diversity and biogeography of bacteriophages in biofilms of Lake Baikal based on g23
879   sequences. Journal of Great Lakes Research 46. DOI: 10.1016/j.jglr.2019.06.007.
880   Potapov S, Belykh O, Krasnopeev A, Gladkikh A, Kabilov M, Tupikin A, Butina T. 2018.
881   Assessing the diversity of the g23 gene of T4-like bacteriophages from Lake Baikal with high-
882   throughput sequencing. FEMS microbiology letters 365. DOI: 10.1093/femsle/fnx264.

883   Pradeep Ram A, Nishimura Y, Tomaru Y, Nagasaki K, Nagata T. 2010. Seasonal variation in
884   viral-induced mortality of bacterioplankton in the water column of a large mesotrophic lake
885   (Lake Biwa, Japan). Aquatic Microbial Ecology 58:249–259. DOI: 10.3354/ame01381.
886   Rahimi-Midani A, Lee YS, Kang S-W, Kim M-K, Choi T-J. 1970. First isolation and molecular
887   characterization of bacteriophages infecting acidovorax citrulli, the causal agent of bacterial fruit
888   blotch. The Plant Pathology Journal 34:59–64. DOI: 10.5423/PPJ.NT.08.2017.0190.
889   Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE,
890   Sánchez-Gracia A. 2017. DnaSP 6: DNA sequence polymorphism analysis of large data sets.
891   Molecular Biology and Evolution 34:3299–3302. DOI: 10.1093/molbev/msx248.
892   Sandaa R-A, Kristiansen H. 2016.T4-type bacteriophages from different marine environments.
893   Available at
894   https://www.ncbi.nlm.nih.gov/popset?LinkName=protein_popset&from_uid=157273680
895   Sime-Ngando T, Bettarel Y, Colombet J, Palesse S, Ram ASP, Charpin M, Amblard C. 2016.
896   Lake Pavin: a pioneer site for ecological studies of freshwater viruses. In: Lake Pavin. Cham:
897   Springer International Publishing, 229–244. DOI: 10.1007/978-3-319-39961-4_14.
898   Straškrábová V, Izmest'yeva LRR, Maksimova EAA, Fietz S, Nedoma J, Borovec J, Kobanova
899   GII, Shchetinina EV V., Pislegina EV V., Stra??kr??bov?? V, Izmest'yeva LRR, Maksimova
900   EAA, Fietz S, Nedoma J, Borovec J, Kobanova GII, Shchetinina EV V., Pislegina EV V. 2005.
901   Primary production and microbial activity in the euphotic zone of Lake Baikal (Southern Basin)
902   during late winter. Global and Planetary Change 46:57–73. DOI:
903   10.1016/j.gloplacha.2004.11.006.
904   Sullivan MB, Waterbury JB, Chisholm SW. 2003. Cyanophages infecting the oceanic
905   cyanobacterium Prochlorococcus. Nature 424:1047–1051. DOI: 10.1038/nature01929.
906   Suttle CA. 2007. Marine viruses — major players in the global ecosystem. 5:801–812. DOI:
907   10.1038/nrmicro1750.
908   Taboada B, Isa P, Gutiérrez-Escolano AL, del Ángel RM, Ludert JE, Vázquez N, Tapia-Palacios
909   MA, Chávez P, Garrido E, Espinosa AC, Eguiarte LE, López S, Souza V, Arias CF. 2018. The
910   Geographic Structure of Viruses in the Cuatro Ciénegas Basin, a Unique Oasis in Northern
911   Mexico, Reveals a Highly Diverse Population on a Small Geographic Scale. Applied and
912   Environmental Microbiology 84. DOI: 10.1128/AEM.00465-18.
913   Tetart F, Desplats C, Kutateladze M, Monod C, Desplats C, Kutateladze M, Monod C,
914   Ackermann H-WW, Krisch HM. 2001. Phylogeny of the major head and tail genes of the wide-
915   ranging T4-type bacteriophages. Journal of bacteriology 183:358–366. DOI:
916   10.1128/JB.183.1.358-366.2001.
917   Vollenweider RA, Kerekes J. 1982. Eutrophication of waters. Monitoring assessment and
918   control. Paris.
919   Wang MN, Ge XY, Wu YQ, Yang X Lou, Tan B, Zhang YJ, Shi ZL. 2015. Genetic diversity and
920   temporal dynamics of phytoplankton viruses in East Lake, China. Virologica Sinica 30:290–300.
921   DOI: 10.1007/s12250-015-3603-6.

922 Watanabe Y, Drucker VV. 1999. Phytoplankton blooms in Lake Baikal, with reference to the
923 lake's present state of eutrophication. In: Kawanabe H, Coulter GG, Roosevelt AC eds. Ancient
924 Lakes: Their Cultural and Biological Diversity. Kenobi Productions: Belgium, 217–225.
925 Wetzel RG, Likens GE. 2000. Limnological Analyses. New York, NY: Springer New York.
926 DOI: 10.1007/978-1-4757-3250-4.
927 Wu S, Zhou L, Zhou Y, Wang H, Xiao J, Yan S, Wang Y. 2020. Diverse and unique viruses
928 discovered in the surface water of the East China Sea. BMC Genomics 21:441. DOI:
929 10.1186/s12864-020-06861-y.
930 Zeigler Allen L, McCrow JP, Ininbergs K, Dupont CL, Badger JH, Hoffman JM, Ekman M,
931 Allen AE, Bergman B, Venter JC. 2017. The Baltic Sea virome: diversity and transcriptional
932 activity of DNA and RNA viruses. mSystems 2:e00125-16. DOI: 10.1128/mSystems.00125-16.
933 Zheng C, Wang G, Liu J, Song C, Gao H, Liu X. 2013. Characterization of the major capsid
934 genes (g23) of T4-type bacteriophages in the wetlands of Northeast China. Microbial Ecology
935 65:616–625. DOI: 10.1007/s00248-012-0158-z.
936 Zhong X, Jacquet S. 2014. Differing assemblage composition and dynamics in T4-like
937 myophages of two neighbouring sub-alpine lakes. Freshwater Biology 59:1577–1595. DOI:
938 10.1111/fwb.12365.
939

Figure 1

Map of the sampling area.

MK - Mukhor Bay near the Kuchelga River, MC – center of Mukhor Bay, Turk – coastal zone near the Turka, Posol_S – Posolsk Sor Bay, LT_05 – central station of the Listvyanka settlement – the Tankhoy settlement section, sampling was carried out in the layer from 0 + 5 m, LT_1015 – central station of the Listvyanka settlement – the Tankhoy settlement section; sampling was carried out in the layer from 10 + 15 m.

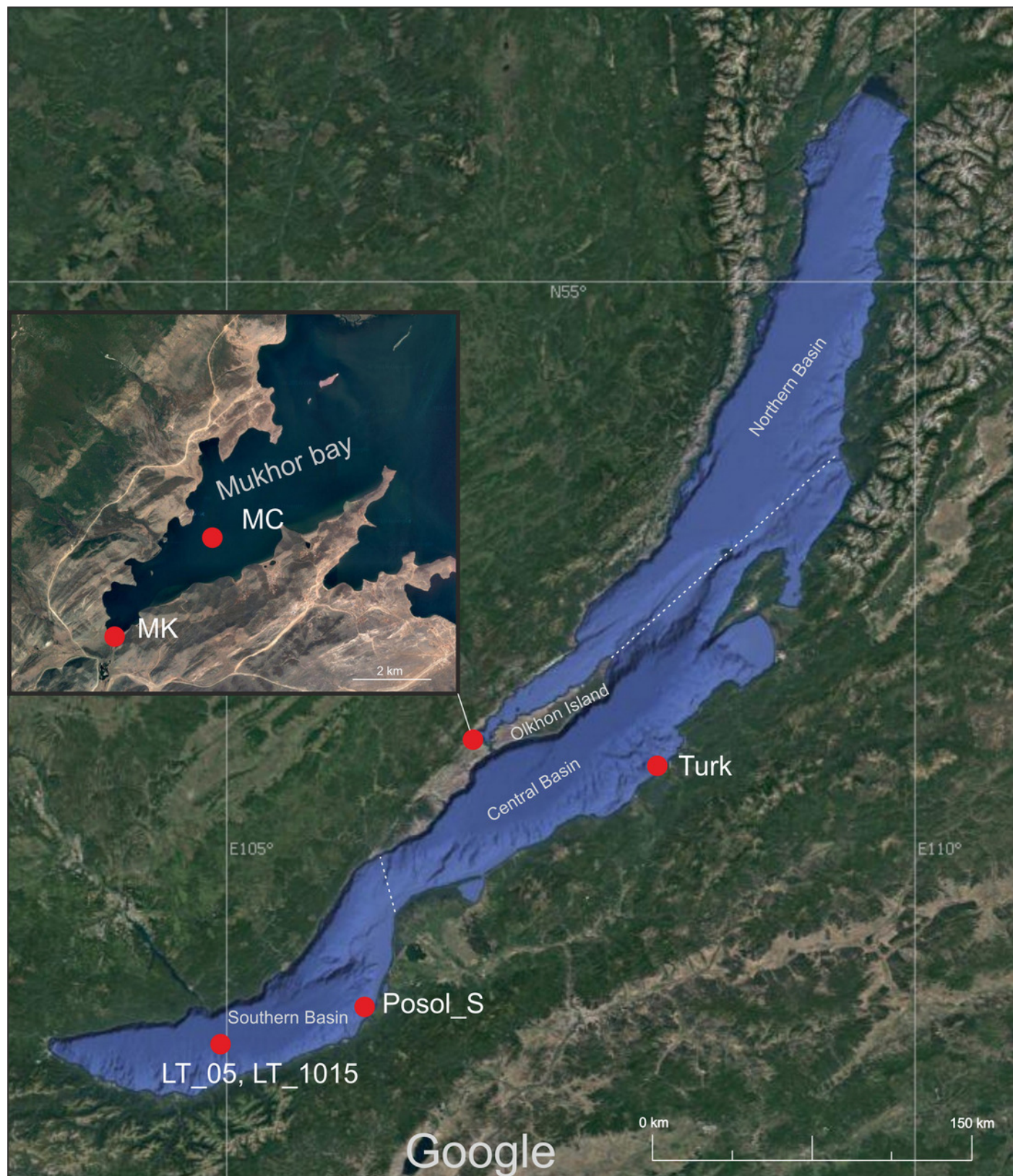



Figure 2

Bayesian phylogenetic tree based on the alignment of 438 *g23* major capsid protein sequences.

LAB – Lake Bourget and Annecy, WS – wetland sediments, EL – East Lake, sPRE – sediments of Perl River estuary, BW – Borehole water, DL – Donghu Lake, WW- wetland water, DW – dairy wastewater, PW – paddy water, RF – rice field, SRF – the soil of rice field, LL – Lake Limnopolar, KL – Lake Kotokel, LS, 2D1, 2D2, 2D4 , 2a – biofilms (Lake Baikal), NB – Neuston Lake Baikal, OBV – Osaka Bay, HOT - sampling site of the Hawaii Ocean Time-series program, BacFR – Lake Biwa (bacterial fraction), VirFR – Lake Biwa (viral fraction).

Figure legend:

- This study (red)
- Freshwater (green)
- Soil and sediments (orange)
- Marine (purple)
- Other (black)
- Cultured (●)
- From virome (✱)

T-, Pseudo T-, Schizo T-evens

Exo T-events and phages of SAR11

0.3

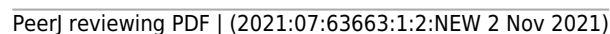



Figure 3

Network demonstrating the relationship of the studied OTUs with other previously obtained sequences: S0508 and N0508 (Butina et al., 2010); 2D1, 1D3Tp, 2D2, 2D4, and Neuston (Potapov et al., 2020) ; Lake Kotokel (Butina et al., 2013).

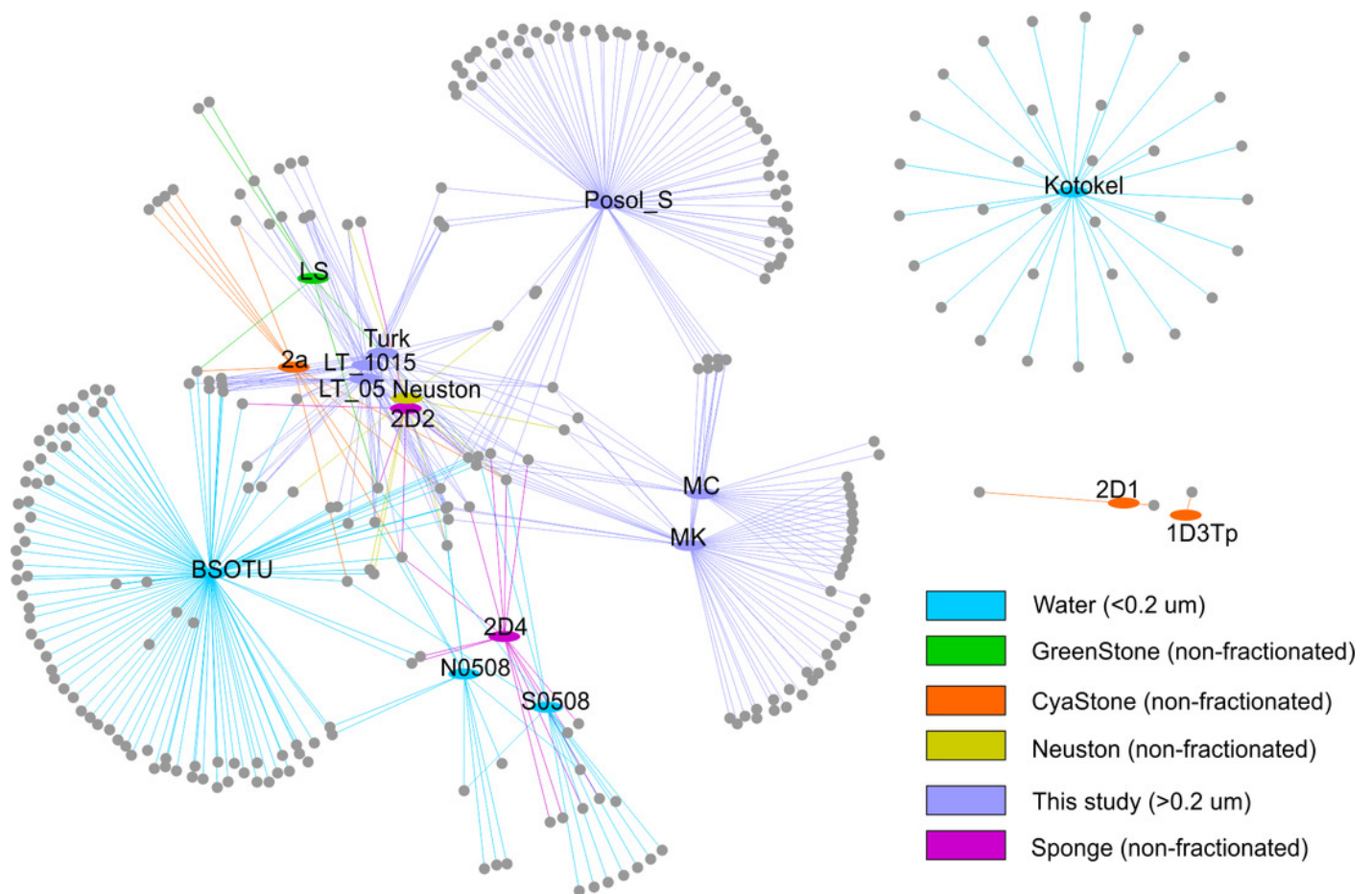



Figure 4

Cluster dendrogram (UPGMA). Samples from this study are marked in bold.

S0508 – southern basin, Lake Baikal, N0508 – northern basin, Lake Baikal, BSOTU – Baikal OTU samples, LAB – lakes Bourget and Annecy. Untagged samples are either unpublished or have no open access articles.

Paddy soil, China ▲
Paddy field, Japan ▲
Wetland sediments, China
Sediments of Perl River Estuary, China ▲

Soil and Sediments

Lake Baikal, LS ▲
Lake Baikal, 2a ▲
Lake Baikal, N0508 ●
Lake Baikal, 2D2 ▲
Lake Baikal, S0508 ●
Lake Baikal, 2D4 ▲
Lake Baikal, Neuston ▲
Lake Baikal, 2D1 ▲
Lake Baikal, 1D3Tp ▲

Baikal cluster I
May-June
(except Neuston)

Lake Baikal, MK ■
Lake Baikal, MC ■
Lake Baikal, Posol_S ■
Lake Baikal, LT_05 ■
Lake Baikal, LT_1015 ■
Lake Baikal, Turk ■
Lake Baikal, BSOTU ●

Baikal cluster II
August

LAB, France ●
Lake Kotokel, Russia ●
Kongsfjorden, Svalbard ●
Lake Donghu, China
Lake East, China ●
Dairy wastewater, Ireland
Lake Limnopolar, Antarctica ●
Proglacial Lake, Svalbard ●
Wetland water, China ●

Hydrothermal vent, East Scotia Ridge ●
Coral colony, Orpheus Island ▲
Chesapeake Bay ●

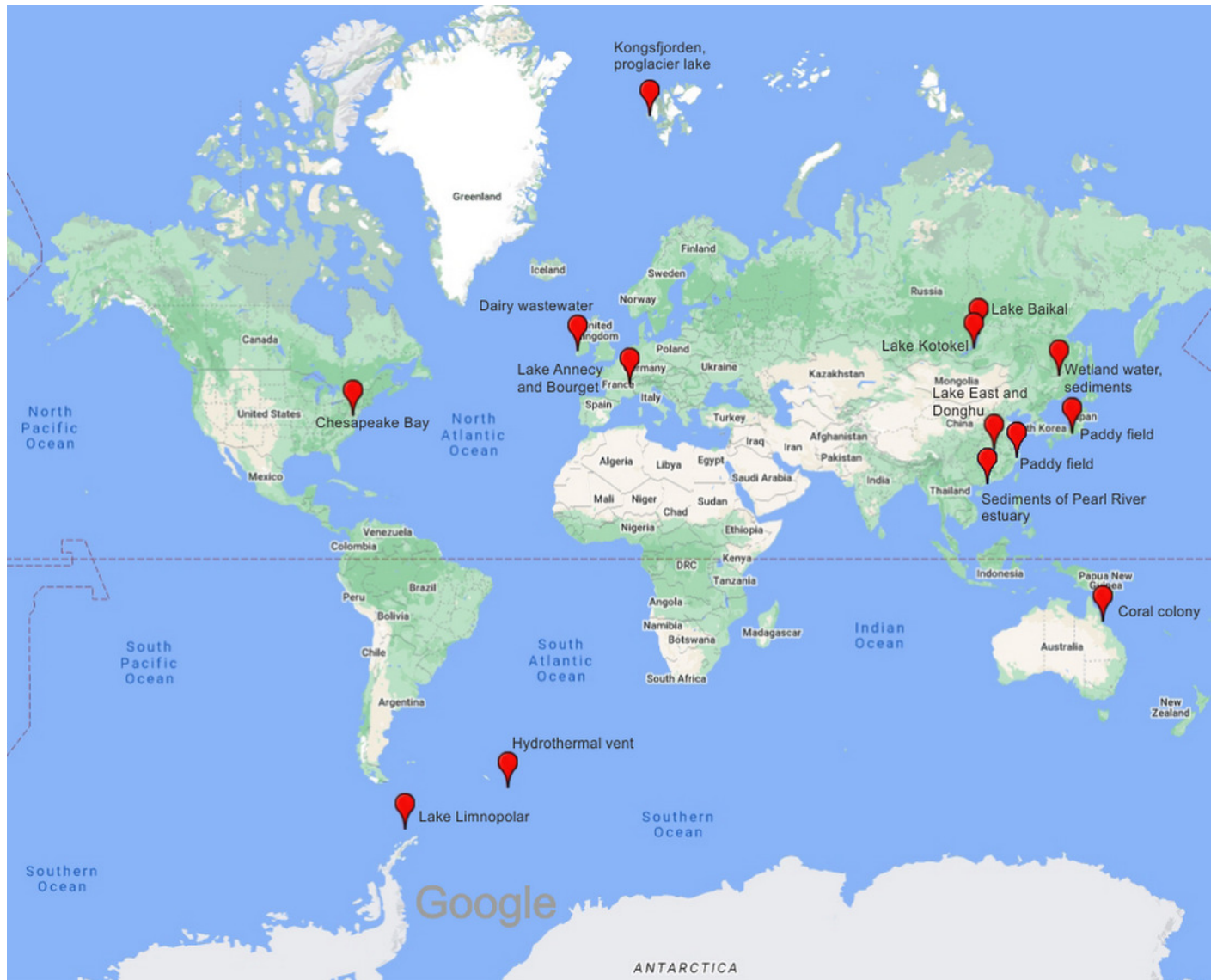
Marine

▲ Non-fractionated
● Fraction < 0.2 μm
● Fraction < 0.4 μm
■ Fraction > 0.2 μm

0.8   0.6   0.4   0.2   0.0

Difference

Figure 5

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

Physical and chemical parameters of water

	MK shallow bay	MC shallow bay	Turk coastal zone	Posol_S shallow bay	LT_05 pelagic zone	LT_1015 pelagic zone
Sampling horizon, m	0	0	0	0	0-5	10-15
Water temperature, °C	19.6	21.5	17.3	19	12 (average)	6.2 (average)
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

Nucleotide diversity of the *g23* gene fragment. Samples from Lake Baikal are marked in bold; samples obtained in this study are highlighted in grey.

1   Table 2

Sample	Nucleotide diversity, π	References
Lake Baikal, LS	**0.38**	**(Potapov et al., 2020)**
Lake Limnopolar (Antarctica)	0.37	(López-Bueno et al., 2009)
Dairy water (Ireland)	0.37	(Knapik & Prentice, 2012)
Lake Baikal, 2D4	**0.37**	**(Potapov et al., 2020)**
MK	**0.36**	This study
MC	**0.35**	This study
Turk	**0.35**	This study
Paddy field (China)	0.35	(Li et al., 2019)
Lake Baikal, 2a	**0.35**	**(Potapov et al., 2020)**
LT_1015	**0.34**	This study
Lake Proglacial (Svalbard)	0.34	(Bellas & Anesio, 2013)
Lake Kotokel (Russia)	0.34	(Butina et al., 2013)
LT_05	**0.33**	This study
Kongsfjorden (Svalbard)	0.33	(Bellas & Anesio, 2013)
Lake Baikal (South basin)	**0.33**	**(Butina et al., 2010)**
Paddy field (Japan)	0.33	(Cahyani et al., 2009)
Wetland, water (China)	0.32	(Zheng et al., 2013)
Posol_S	**0.31**	This study
Lake Donghu (China)	0.31	(Huang, Cheng & Xu, 2011)
Lake Baikal, 2D2	**0.30**	**(Potapov et al., 2020)**
Hydrothermal vent (East Scotia Ridge)	0.30	(Millard, Pearce & Zwirglmaier, 2016)
Lake Baikal (North basin)	**0.30**	**(Butina et al., 2010)**
Chesapeake Bay	0.29	(Jamindar et al., 2012)
Lake East (China)	0.29	(Wang et al., 2015)
Wetland sediments (China)	0.29	(Li et al., 2018)
Lake Baikal, pelagic water (South basin)	**0.28**	**(Potapov et al., 2018)**
Lake Annecy and Bourget (France)	0.28	(Zhong & Jacquet, 2014)
Sediments of Pearl River Estuary (China)	0.28	(He et al., 2017)
Lake Baikal, Neuston	**0.27**	**(Potapov et al., 2020)**
Coral colony (Orpheus Island)	0.23	(Buerger et al., 2018)
Lake Baikal, 2D1	**0.14**	**(Potapov et al., 2020)**
Lake Baikal, 1D3Tp	**0.03**	**(Potapov et al., 2020)**

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

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)

Sets of the *g23* sequences included in analysis

Isolation source	Localization	Reference
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
Lake Baikal, BSOTU	Russia	Potapov et al., 2018
Lake Baikal	Russia	Potapov et al., 2020
Lake Annecy and Bourget, LAB	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
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
Paddy field	Japan	Cahyani et al., 2009