Original Article

Search for putative complete and near-complete genomes of DNA-containing viruses in metagenomes obtained from the Lake Baikal

Potapov S.A.[1]*, Tupikin A.E.[2], Tikhonova I.V.[1], Zhuchenko N.A.[1], Belykh O.I.[1]

[1] *Limnological Institute Siberian Branch Russian Academy of Sciences, Ulan-Batorskaya St., 3, Irkutsk, 664033, Russia*
[2] *Institute of Chemical Biology and Fundamental Medicine, Siberian Branch of the Russian Academy of Sciences (ICBFM SB RAS), Ac. Lavrentieva ave., 8, Novosibirsk, 630090, Russia*

ABSTRACT. The paper presents the analysis of putatively complete and near-complete genomes of bacteriophages extracted from metagenomic data obtained from DNA samples isolated from Lake Baikal water using modern bioinformatic programs. A total of 73 sequences with lengths ranging from 13.8 kb to 163.7 kb belonging to phages of the Caudoviricetes class were identified. Two contigs belonging putatively to cyanophages with lengths of 36.8 kb and 163.7 kb were detected, and in the latter one an ORF with a length of 159 amino acid residues similar to the small heat shock protein (Hsp20) was identified. Analysis of the amino acid sequences identified in the assembled bacteriophage genomes using the PHROG database revealed that 27.5% of them have an unknown function, while the majority of those with similarity to known ones (23.7%) belong to the category "DNA, RNA and nucleotide metabolism". A number of accessory metabolic genes (AMGs) were also detected in the assembled genomes: *nadM*, *cysC*, *cobS*, *galE*, *cobT*, etc. Most of the sequences with similarity to sequences from the IMG/VR database (89.6%) corresponded to sequences obtained from freshwater bodies.

Keywords: metagenomics, bacteriophages, high-throughput sequencing, Lake Baikal, complete genome

For citation: Potapov S.A., Tupikin A.E., Tikhonova I.V., Zhuchenko N.A., Belykh O.I. Search for putative complete and near-complete genomes of DNA-containing viruses in metagenomes obtained from the Lake Baikal // Limnology and Freshwater Biology. 2024. - № 4. - P. 1050-1065. DOI: 10.31951/2658-3518-2024-A-4-1050

1. Introduction

Currently, metagenomics provides an opportunity to analyze the diversity of viruses in different habitats, particularly in marine and freshwater ecosystems. Virus diversity is extremely high, but only a small fraction of viruses is represented by complete genomes (Paez-Espino et al., 2016).

The assembly of viral genomes from metagenomes is a challenging task (Smits et al., 2014). The identification of viral genomes from metagenomes is negatively affected by several factors, such as contamination by non-viral sequences (Roux et al., 2013), prophages may be flanked by regions belonging to the host, and the presence of many short sequences obtained during assembly. Short contigs are often discarded, and only those longer than 5-10 kb are included in the analysis (Gregory et al., 2019). The lack of universal marker genes and the large variability in viral genome lengths also contribute to the difficulty in virus identification. Circular genomes can be identified by the presence of terminal repeats, and the genome can also be determined by covering the known virus genome (homology search), but due to the small number of cultured viruses and the huge diversity of viruses, this approach currently has limitations.

Nevertheless, due to the difficulty of culturing viruses, the metagenomic approach has probably become the key approach to date, as evidenced by recent work (Gregory et al., 2019; Castro-Nallar et al., 2023), as well as the increase in the number of publicly available viral genomes (fragments) assembled from metagenomic data, from 84 in 2010 to 775,000 in 2018 (Roux et al., 2021). For example, the number of viral sequences from all oceans, including the Arctic Ocean, was recently increased 12-fold. The study identified 5 ecological zones whose formation was primarily driven by temperature (Gregory et al., 2019). A global study of viromes from freshwater ecosystems based on 380 publicly available viral metagenomes enabled the recovery of 549 complete high-quality genomes. The abundance study showed that less than 0.2% of viral contigs occur

*Corresponding author.
E-mail address: poet1988@list.ru (S.A. Potapov)

Received: July 29, 2024; ***Accepted:*** August 14, 2024;
Available online: August 30, 2024

in all biomes studied, i.e. the total pool of contigs is minimal. It is concluded that each water body has a separate virome specific to it. It is shown that 41.4% of viral contigs from freshwater ecosystems are not identified by taxonomy, while the rest mainly belong to tailed phages (Elbehery and Deng, 2022).

The first assembled complete genomes of bacteriophages from Lake Baikal MD8 *Pseudomonas aeruginosa* (*Siphoviridae*) and *P. aeruginosa* PaBG (*Myoviridae*) were obtained by cultivation (Sykilinda et al., 2014; Drucker et al., 2015). Previously, the complete genome of the phage Baikal-20-5m-C28 with a length of 166 kb was assembled from metagenomic data obtained on Lake Baikal, whose host is presumably *Polynucleobacter* spp. (Cabello-Yeves et al., 2018). In addition, 16 putative complete genomes of virophages belonging to the three families *Burtonviroviridae*, *Dishuiviroviridae*, and *Omnilimnoviroviridae* (Potapov and Belykh, 2023) and near-complete genomes of RNA-containing viruses obtained from mollusks (Butina et al., 2023) and the water column (Potapov et al., 2023) were identified.

The aim of this work is to obtain high-quality (i.e., with high confidence) sequences of complete DNA genomes of viruses from metagenomic data obtained from Lake Baikal, annotation of protein sequences, and taxonomic identification of the recovered genomes.

2. Materials and methods

2.1. Sampling, DNA extraction and sequencing

Water samples for analysis were collected from Lake Baikal in its three basins (Southern, Central, Northern), as well as in the Maloye More Strait (Table 1).

Samples were collected from depths of 0 to 50 m (BVP1-8) and 0 to 15 m (RVP4-6) using Niskin bathometers and mixed to obtain an integrated sample, totaling 20 liters per sample. Water samples were sequentially filtered through 0.4 μm and 0.2 μm pore size filters (Sartorius, Germany) to remove detritus, zoo-, phyto- and bacterioplankton. Concentration was then carried out by ultrafiltration in tangential flow with a cutoff of 50 kDa using VivaFlow 200 (Sartorius, Germany) and Vivaspin 15 ultracentrifuge tubes (Sartorius, Germany). This resulted in 100 μl of concentrate per sample. To purify free viral particles from foreign DNA, the sample was treated with DNAase (Thermo Fisher Scientific, MA, USA). The reaction was stopped by adding 50 mM EDTA and incubated at 65°C for 10 min. DNA was extracted using the phenol-chloroform method. DNA concentration was measured on a Qubit 2.0 (Invitrogen, USA). Libraries were prepared and sequenced using Illumina MiSeq (Illumina, USA) and MGIseq 2000 (MGI Tech Co., Ltd., PRC) at the "Genomics Core Facility" (ICBFM SB RAS, Novosibirsk, Russia).

2.2. Bioinformatics analysis

The raw reads were analyzed in FastQC v. 0.12.1 (Andrews, 2010). Adapter removal as well as quality filtering was performed in Trimmomatic v. 0.36 (Bolger et al., 2014). Removal of ribosomal DNA contaminants was performed using SortMeRNA v. 4.3.6 (Kopylova et al., 2012), and Bowtie2 v. 2.4.4 was used to remove human sequences (Langmead and Salzberg, 2012) by mapping reads to a known genome (GRCh38_noalt_as). Combined assembly was performed by combining all samples ("cross-assembly") to achieve longer contigs and increase the number of viral contigs (Bukin et al., 2023), using metaSPAdes v. 4.0.0 (Nurk et al., 2017) with the additional parameter -k 21, 33, 55, 77. Viral sequence extraction was performed using VirSorter2 v. 2.2.4 (Guo et al., 2021) with a score parameter > 0.9, minimum contigs length of 5 kb. The reads were then mapped to the resulting contigs putatively belonging to viruses using Bowtie2, coverage was calculated by the program SAMtools v. 1.13 (Li et al., 2009). Only contigs with coverage depth more than 5 were taken for further analysis. Using the program COBRA v. 1.2.3 (Chen and

Table 1. Water samples collected for virome analysis.

Labeling	Sampling date	Sampling site	Project number in the SRA	Reference
BVP1	22.03.18	7 km from Listvyanka settlement	PRJNA1006167	(Potapov and Belykh, 2023)
BVP2	8.06.18	3 km from Listvyanka settlement		
BVP3	31.05.18	3 km from Turka settlement		
BVP4	4.06.18	3 km from Elokhin Cape		
BVP5	5.08.18	central station Maloye More Strait		
BVP6	27.09.18	central station Listvyanka settlement – Tankhoy settlement		
BVP7	25.09.18	central station Ukhan Cape – Tonkiy Cape		
BVP8	23.09.18	central station Elokhin Cape – Davsha settlement		
RVP4_DNA	29.07.22	central station Listvyanka settlement – Tankhoy settlement	Not deposited	–
RVP5_DNA	07.08.22	central station Ukhan Cape – Tonkiy Cape		
RVP6_DNA	02.08.22	central station Elokhin Cape – Davsha settlement		

Banfield, 2024), we joined the assembled sequences and achieved higher accuracy by increasing the length and completeness of contigs. The contigs were then checked using the program CheckV v. 1.0.3 (Nayfach et al., 2021). Further, only direct terminal repeats (DTR) were used to identify complete genomes, as this is the most established approach. In addition, parameters such as AAI-based completeness > 90%, confidence_ level - high were taken into account. Similar sequences were clustered at the 95% identity level (ANI) using clustering scripts from the CheckV program.

Taxonomic identification of viral genomes was performed using geNomad v. 1.8.0 (Camargo et al., 2023b), Diamond v. 2.1.8.162 with e-value parameters - 10^{-5}, bit score ≥ 50, more sensitive and BlastN v. 2.12.0+ (e-value - 10^{-5}) using the amino acid and nucleotide database of viral genomes RefSeq v. 222. Functional analysis of translated ORFs from contigs was performed applying PHROG v. 4 (Terzian et al., 2021) and VOG v. 219 databases applying HHMER v. 3.2.1 (Eddy, 2011). The IMG/VR v. 4 database (Camargo et al., 2023a) was used to search for similar proteins from uncultured viruses. We also used the automatic phage annotation program VIBRANT v. 1.2.1 (Kieft et al., 2020) to search for auxiliary metabolic genes (AMGs). A proteomic tree was constructed using the on-line service VipTree v. 4.0 (Nishimura et al., 2017).

3. Results

The results of stepwise processing are presented in Table 2. A total of 8288 sequences belonging to viruses were identified from the obtained mix assembly after processing in the VirSorter2 program, which was 39.6% of all assembled contigs of more than 5 kb.

After processing 8033 sequences in COBRA, 49.1% of the sequences managed to increase their length by an average of 37.6%. Quality control of viral contigs in CheckV determined that 73 sequences belonged to high quality over 90% with DTR. At the same time, 60 sequences belonged to complete genomes, and 13 sequences were characterized as near-complete (extend partial group, COBRA). The length of these putative complete and near-complete virus genomes ranged from 13871 to 163727 nucleotides (Supplementary material). The number of ORFs ranged from 20 to 216. The statistical information is presented in Figure 1.

Table 2. Number of reads/contigs after each processing step.

Processing stage	Pair of reads/ contigs
1. Raw reads	176329692
2. Trimmomatic	152536284
3. SortMeRNA	151248958
4. Deletion of human sequences	149200074
5. metaSPAdes, contigs of more than 5 thousand nucleotides	20929
6. VirSorter2, contigs of more than 5 thousand nucleotides	8288
7. Bowtie2, SAMtools, coverage depth more than 5	8033
8. COBRA	3191
9. CheckV	73

Taxonomic identification based on geNomad (virus_score > 0.96) showed that all genomes belonged to the class Caudoviricetes, with 4 sequences (NODE_7814, NODE_610, NODE_598, NODE_665) identified up to the family *Autographiviridae*. It is worth noting that we also identified virophages that were discovered earlier, but we did not include them in this paper because they were described in detail in (Potapov and Belykh, 2023).

Functional analysis using the PHROG database revealed that the category “DNA, RNA, and nucleotide metabolism” was the most represented (23.7%); in addition, 27.5% of amino acid sequences had an unknown function (Fig. 2). Structural proteins accounted for 38.3% (categories “Capsid and Packaging”, “Tail”, and “Connector”). A large terminase subunit was detected in all bacteriophage genomes by PHROG and VOG databases.

In the genome of NODE_40, presumably belonging to a cyanophage, an ORF of 159 amino acid residues in length similar to a small heat shock protein (Hsp20, YP_009134378) was detected. In addition, genes related to microbial resistance to oxidative stress, such as

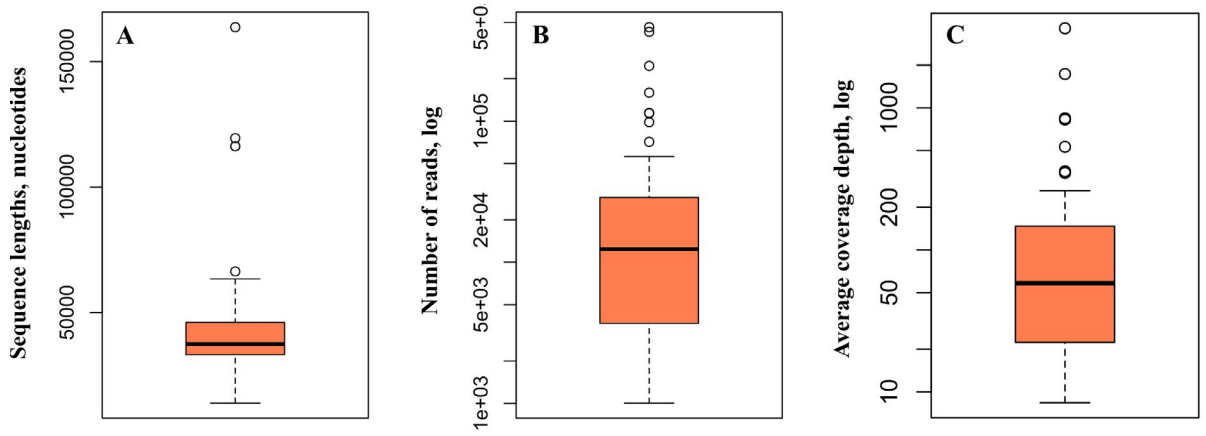

Fig.1. Statistical information of 73 phage sequences. A - range of sequence lengths, nucleotide bases, B - number of reads per genome, C - average depth of genome coverage.

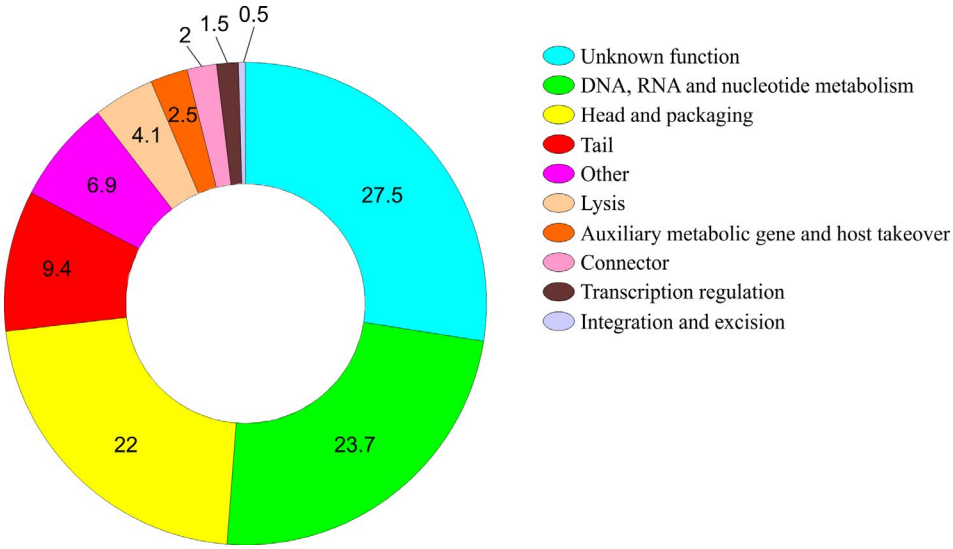

Fig.2. Functional analysis of amino acid sequences of 73 phages. Share is given in percentages.

NAMPT (nicotinamide phosphoribosyltransferase) and *nadM* (bifunctional NMN adenylyltransferase/nudix hydrolase) were identified in this genome. NODE_334 included the *cysC* gene, which is involved in the assimilatory reduction of sulfate. The *cobS* (NODE_4873, NODE_40, NODE_506), *cobT* (NODE_4873) genes are involved in cobalamin biosynthesis. *PurA* (adenylosuccinate synthase) and *purB* (adenylosuccinate lyase) are found in NODE_996 - involved in purine metabolism. Other AMG were also detected: *rfbC*, DNMT1, *galE*, *lpxA*, TSTA3, *pimC*.

Six sequences contained integrase (NODE_506, NODE_926, NODE_7814, NODE_610, NODE_598, NODE_660), with similarities ranging from 28.2% to 39%.

Out of the 4266 ORFs, 4108 (96.3%) had similarities to uncultured virus representatives from the IMG/VR database. Of these, 89.6% corresponded to representatives derived from freshwater ecosystems, with the majority (78.8%) derived from lakes, 16.5% from rivers, and 2.5% from ponds. Only 3.2% of the 4108 ORFs were of marine origin. The rest belonged to representatives obtained from other sources (soil, wastewater, etc.).

According to the RefSeq database, 126 ORFs were similar to phages recently isolated on the basis of the *Flavobacterium* strains from the Baltic Sea (Nilsson et al., 2020; Hoetzinger et al., 2021). Sequence identity at the amino acid level ranged from 21.5 to 87%. Fifty-four ORFs are similar to *Nonlabens* phage (isolates P12024S and P12024L), whose host is the bacterium *Persicivirga* sp. IMCC12024 isolated from the coastal water of the Yellow Sea (Republic of Korea). The amino acid sequence similarity ranged from 22.4 to 68.8 %.

ORFs (203 sequences) similar to proteins of 63 different cyanophages at the amino acid level had similarities ranging from 22.1 to 81.3%, while the lowest e-value (0) was observed with the major capsid protein (YP_004421726) of *Synechococcus* phage S-CBS3 (62.4% similarity, 99.9% coverage) and DNA polymerase (YP_010669768) of *Synechococcus* phage S-SCSM1 (55.7% similarity, 99.6% coverage). *Synechococcus* phage S-CBS3 was isolated from a strain of *Synechococcus* sp. CB0202 (isolation source is Chesapeake Bay), *Synechococcus* phage S-SCSM1 was isolated from *Synechococcus* sp. strain WH 7803 (isolation source is South China Sea). The two sequences NODE_40 (163.7 kb, 216 ORF) and NODE_1081 (36.8 kb, 49 ORF) had 50 and 23 ORFs, respectively, similar to cyanophages from the RefSeq database, indicating with high probability that they belong to cyanophages. Other sequences with similar proteins to cyanophages, but with a smaller number per genome, are difficult to identify due to the lack of known cyanophages in databases.

Seventy-five amino acid sequences were similar to different *Ralstonia* phage isolates (NC_047946, NC_047888, NC_030948 etc.), with similarities at the amino acid level ranging from 25.6 to 75.1%.

Overall, based on the blastp results from the RefSeq database, only 34.7% of ORFs were similar to known proteins. In addition, 80.8% of the sequences similar to proteins from the database showed less than 50% similarity to known proteins, which may indicate the discovery of new phages representing the so-called viral dark matter pool. This is also supported by the low percentage of similarity with known viruses from the RefSeq nucleotide database of complete genomes, such that the maximum similarity and coverage was registered with *Ralstonia* phage RsoP1EGY (13% similarity, 67.4% coverage), *Synechococcus* phage S-CBS3 (25% similarity, 70% coverage), and *Ralstonia* phage RsoP1EGY (18% similarity, 69.6% coverage).

The representation analysis based on TPM (transcripts per kilobase million) showed that 32 sequences (42.5%) were present in all samples. NODE_923 was detected only in sample BVP5 (Maloye More Strait). Two sequences were assembled only from the 2022 reads (NODE_996, NODE_660). Eight sequences were missing from summer 2018 (BVP5, BVP6, BVP7, BVP8) (NODE_598, NODE_660, NODE_547, NODE_996, NODE_5829, NODE_506, NODE_665, NODE_1217). This type of analysis can indicate both the prevalence of phages in the lake across all basins and seasons and

the uniqueness and replication of phages at particular time intervals.

The proteomic tree built on the basis of our assembled genomes shows both formed clusters and individual branches, with the formed clusters containing the closest relatives according to the RefSeq database. This type of analysis also allows us to identify putative hosts, for example, the presence of the closest relative in a cluster for which the host is known and which contains our assembled genomes. Thus, we found 4 phyla of potential hosts Pseudomonodata, Bacteroidota, Cyanobacteriota and Bacillota (Fig. 3).

4. Discussion

In this study, we were able to identify 60 putatively complete and 13 nearly complete phage genomes (vMAG, viral metagenome-assembled genome) belonging to the class Caudoviricetes from the metagenomes of freshwater Lake Baikal.

Due to the mosaic nature of phage genomes, as well as their lack in databases, there is a nontrivial task of determining the closest relative at the species, genus, and even family level arises. It should be noted that we used strict virus identification conditions when using the VirSorter2 program: max_score > 0.9 - high confidence, CheckV: confidence_level - high, aai_completeness more than 90%, checkv_quality - complete, presence of DTR, contamination - 0, etc., as this is the optimal option for obtaining high-quality complete genomes.

A number of auxiliary metabolic genes have been discovered in phage genomes. Changes in productivity at the ecosystem level occur through horizontal transfer of ecologically important genes and expression of virus-encoded AMGs (Hurwitz and U'Ren, 2016). These genes are expressed during infection, increasing and redirecting energy and resources to virus production (Thompson et al., 2011; Hurwitz and U'Ren, 2016; Smet et al., 2016; Howard-Varona et al., 2018). The identified AMGs demonstrate the involvement of viruses in biogeochemical reactions. The *cobS* are involved in cobalamin biosynthesis and may support deoxynucleotide synthesis. The *cysC* gene is involved in the assimilatory reduction of sulfate. The presence of phages with this gene in freshwater ecosystems may influence the sulfur cycle through the process of assimilatory sulfate reduction. Respiratory complex I (NADH: ubiquinone oxidoreductase) uses the energy released by electron transfer from NADH to quinone to pump protons through the plasma membrane (Walker, 1992). The bioavailability of this complex can be altered by the expression of viral auxiliary metabolic genes involved in NAD + biosynthesis (NAMPT, *nadM*). Recently, these genes have been found in contigs assigned to Caudoviricetes derived from the mouse intestine (Ishola et al., 2024) as well as in earthworm intestinal phages (Xia et al., 2023). The *galE* gene encoding UDP-glucose-4-epimerase mediates the conversion of UDP-galactose and UDP-glucose in galactose metabolism and probably allows the virus to participate in carbohydrate metabolism (Heyerhoff et al., 2022). In general, AMGs in phages and their role are still poorly understood, but their importance in phage survival is beyond doubt.

A small heat shock protein (sHSP) was found in the NODE_40 sequence, which is thought to belong to a cyanophage. It was previously shown to be present in some cyanophages (marine and freshwater) that infect the unicellular cyanobacteria *Synechococcus* and *Prochlorococcus* (Dreher et al., 2011; Maaroufi and Tanguay, 2013). Cyanophages have been shown to have acquired the sHSP gene from a class A bacterial ancestor by lateral gene transfer (Maaroufi and Tanguay, 2013).

Detection of assembled phage genomes from Lake Baikal in different seasons and years demonstrates that half of them is constantly present in the time intervals we observed. Inversely, some of them are detected only at a certain time and place. Monthly sampling throughout the year and sequencing will help us to understand how many phages are present in a given season, which we will continue to study.

Compliance with sample preparation requirements to obtain high-quality virus genomes from metagenomic data, such as deliverance from bacterial and eukaryotic DNA, pre-filtering and using DNase, and sequencing with the highest possible read depth in order to get enough reads out of the total number of reads to assemble virus genomes is essential. Besides, attention should be paid to the timely updating of databases in the analysis, as the growth of new known sequences may reveal additional new viruses. In addition, the emergence of new programs simplifies the analysis and also enables to identify new sequences.

5. Conclusions

Bioinformatic analysis of data obtained from sequenced water DNA samples from Lake Baikal (metagenomes of the fraction less than 0.2 μm) was used to recover 60 complete and 13 near-complete genomes of bacteriophages. Taxonomic analysis showed low similarity of the obtained genomes with the available virus genomes in the RefSeq database. Two genomes, putatively belonging to cyanophages, with lengths of 36.8 kb (NODE_1081) and 163.7 kb (NODE_40), were discovered; a small heat shock protein (sHSP) was identified in NODE_40. A number of auxiliary metabolic genes in the genomes of the obtained phages were identified: *nadM*, *cysC*, *cobS*, *galE*, *cobT*, etc. It was shown that half of the assembled phage genomes were present in all seasons and in all three basins and the M. More Strait, while the rest were detected only at a certain time and location.

6. Funding

The work was carried out within the State Assignment of LIN SB RAS No. 0279-2021-0015 "Studies of viral and bacterial communities as a basis for stable functioning of freshwater ecosystems and effective response under anthropogenic impact".

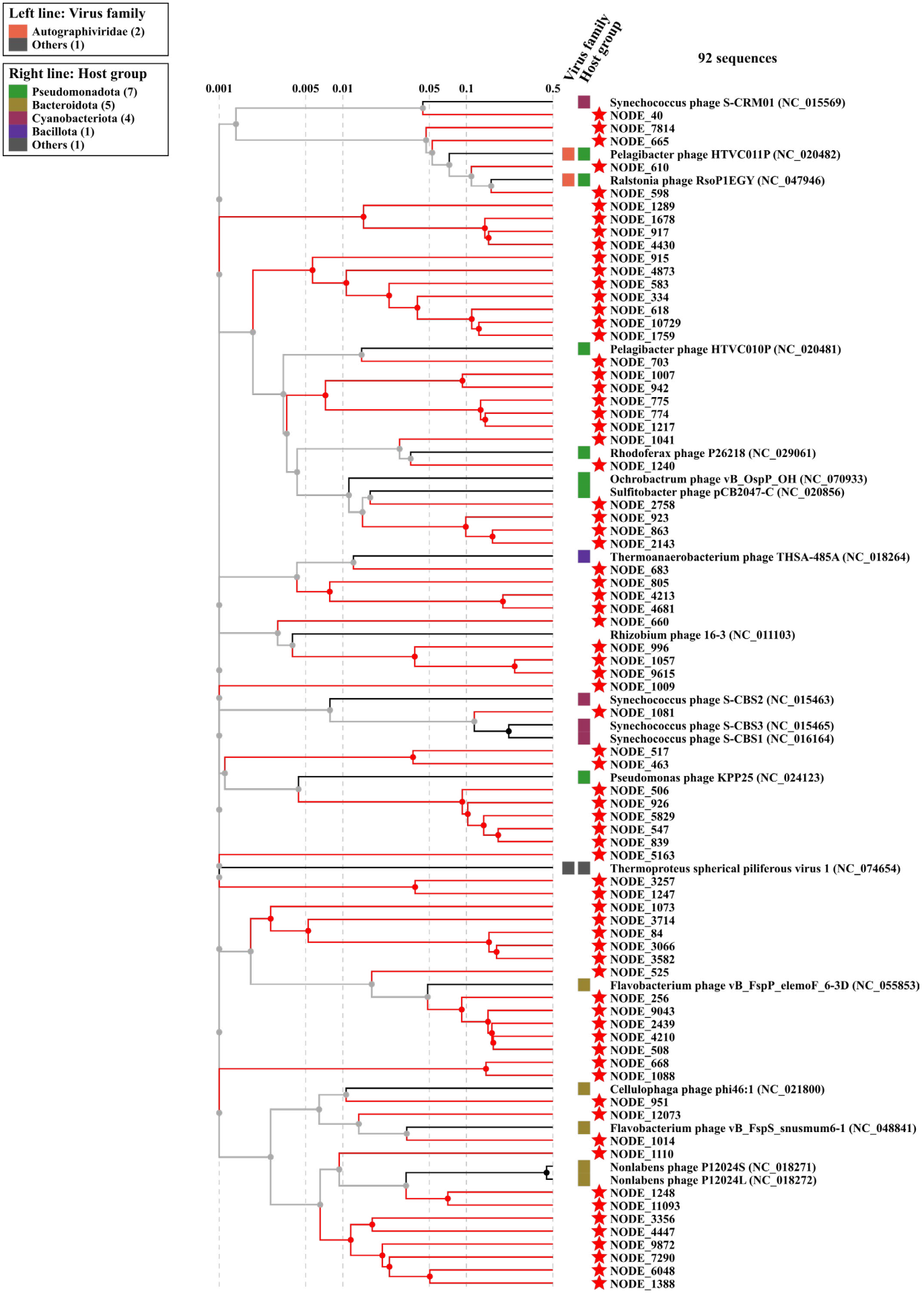

Fig.3. Proteomic tree based on comparison of translated nucleotide sequences of genomes (tBLASTx) with their closest relatives identified using VipTree. Asterisk indicates sequences from this study. Colored squares indicate affiliation with a known reference taxon and its host.

Acknowledgements

The English version of the paper was prepared by Nadezhda Shvedova. The authors are grateful to the crew of R/V "G. Titov" for their assistance in sampling.

Conflict of interest

The authors declare no conflict of interest.

References

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data [Electronic resource]. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bolger A.M., Lohse M., Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15): 2114–2120. DOI: 10.1093/bioinformatics/btu170

Bukin Y.S., Bondaryuk A.N., Butina T.V. 2023. Performance Analysis of Cross-Assembly of Metatranscriptomic Datasets in Viral Community Studies. Mathematical Biology & Bioinformatics 18(2): 418–433. DOI: 10.17537/2023.18.418

Butina T.V., Zemskaya T.I., Bondaryuk A.N. et al. 2023. Viral Diversity in Samples of Freshwater Gastropods Benedictia baicalensis (Caenogastropoda: Benedictiidae) Revealed by Total RNA-Sequencing. International Journal of Molecular Sciences 24(23): 17022. DOI: 10.3390/ijms242317022

Cabello-Yeves P.J., Zemskaya T.I., Rosselli R. et al. 2018. Genomes of novel microbial lineages assembled from the sub-ice waters of Lake Baikal. Applied and Environmental Microbiology 84(1): e02132-17. DOI: 10.1128/AEM.02132-17

Camargo A.P., Nayfach S., Chen I.-M.A. et al. 2023. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. Nucleic Acids Res 51(D1): 733–743. DOI: 10.1093/nar/gkac1037

Camargo A.P., Roux S., Schulz F. et al. 2023. Identification of mobile genetic elements with geNomad. Nature Biotechnology DOI: 10.1038/s41587-023-01953-y

Castro-Nallar E., Berríos-Farías V., Díez B. et al. 2023. Seasonal and Spatially Distributed Viral Metagenomes from Comau Fjord (42°S), Patagonia. Microbiology Resourse Announcements 12(4): 12:e00082-23. DOI: 10.1128/mra.00082-23

Chen L., Banfield J.F. 2024. COBRA improves the completeness and contiguity of viral genomes assembled from metagenomes. Nature Microbiology 9(3): 737–750. DOI: 10.1038/s41564-023-01598-2

Dreher T.W., Brown N., Bozarth C.S. et al. 2011. A freshwater cyanophage whose genome indicates close relationships to photosynthetic marine cyanomyophages. Environmental Microbiology 13(7): 1858–1874. DOI: 10.1111/j.1462-2920.2011.02502.x

Drucker V.V., Bondar A.A., Gorshkova A.S. et al. 2015. Search and studies of autochthonous bacteriophages in different biotopes of Lake Baikal. Contemporary Problems of Ecology 12: 143-154. DOI: 10.1134/S1995425519020045 (in Russian)

Eddy S.R. 2011. Accelerated Profile HMM Searches. PLoS Computational Biology 7(10): e1002195. DOI: 10.1371/journal.pcbi.1002195

Elbehery A.H.A., Deng L. 2022. Insights into the global freshwater virome. Frontiers Microbiology 13. DOI: 10.3389/fmicb.2022.953500

Gregory A.C., Zayed A.A., Conceição-Neto N. et al. 2019. Marine DNA viral macro- and microdiversity from pole to pole. Cell 177(5): 1109-1123.e14. DOI: 10.1016/j.cell.2019.03.040

Guo J., Bolduc B., Zayed A.A. et al. 2021. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome 9(1): 1–13. DOI: 10.1186/s40168-020-00990-y

Heyerhoff B., Engelen B., Bunse C. 2022. Auxiliary Metabolic Gene Functions in Pelagic and Benthic Viruses of the Baltic Sea. Frontiers in Microbiology 13. DOI: 10.3389/fmicb.2022.863620

Hoetzinger M., Nilsson E., Arabi R. et al. 2021. Dynamics of Baltic Sea phages driven by environmental changes. Environmental Microbiology 23(8): 4576–4594. DOI: 10.1111/1462-2920.15651

Howard-Varona C., Hargreaves K.R., Solonenko N.E. et al. 2018. Multiple mechanisms drive phage infection efficiency in nearly identical hosts. ISME J 12(6): 1605–1618. DOI: 10.1038/s41396-018-0099-8

Hurwitz B.L., U'Ren J.M. 2016. Viral metabolic reprogramming in marine ecosystems. Current Opinion in Microbiology 31: 161–168. DOI: 10.1016/j.mib.2016.04.002

Ishola O.A., Kublik S., Durai Raj A.C. et al. 2024. Comparative Metagenomic Analysis of Bacteriophages and Prophages in Gnotobiotic Mouse Models. Microorganisms 12(2): 255. DOI: 10.3390/microorganisms12020255

Kieft K., Zhou Z., Anantharaman K. 2020. VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of virome function from genomic sequences. Microbiome 8 (90) DOI: 10.1186/s40168-020-00867-0

Kopylova E., Noé L., Touzet H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28(24): 3211–3217. DOI: 10.1093/bioinformatics/bts611

Langmead B., Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2 // Nature Methods 9(4): 357–359. DOI: 10.1038/nmeth.1923

Li H., Handsaker B., Wysoker A. et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16): 2078–2079. DOI: 10.1093/bioinformatics/btp352

Maaroufi H., Tanguay R.M. 2013. Analysis and Phylogeny of Small Heat Shock Proteins from Marine Viruses and Their Cyanobacteria Host. PLoS One 8(11): e81207. DOI: 10.1371/journal.pone.0081207

Nayfach S., Camargo A.P., Schulz F. et al. 2021. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nature Biotechnology 39(5): 578–585. DOI: 10.1038/s41587-020-00774-7

Nilsson E., Bayfield O.W., Lundin D. et al. 2020. Diversity and Host Interactions among Virulent and Temperate Baltic Sea Flavobacterium Phages. Viruses 12(2): 158. DOI: 10.3390/v12020158

Nishimura Y., Yoshida T., Kuronishi M. et al. 2017. ViPTree: the viral proteomic tree server. Bioinformatics 33(15): 2379–2380. DOI: 10.1093/bioinformatics/btx157

Nurk S., Meleshko D., Korobeynikov A. et al. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Research 27(5): 824–834. DOI: 10.1101/gr.213959.116

Paez-Espino D., Eloe-Fadrosh E.A., Pavlopoulos G.A. et al. 2016. Uncovering Earth's virome. Nature 536(7617): 425–430. DOI: 10.1038/nature19094

Potapov S., Gorshkova A., Krasnopeev A. et al. 2023. RNA-Seq virus fraction in Lake Baikal and treated wastewaters. International Journal of Molecular Sciences 24(15): 12049. DOI: 10.3390/ijms241512049

Potapov S.A., Belykh O.I. 2023. Virophages Found in Viromes from Lake Baikal. Biomolecules 13(12): 1773. DOI: 10.3390/biom13121773

Roux S., Krupovic M., Debroas D. et al. 2013. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular

sequences. Open Biology 3(12): 130160. DOI: 10.1098/rsob.130160

Roux S., Matthijnssens J., Dutilh B.E. 2021. Metagenomics in Virology. Encyclopedia of Virology. Elsevier, pp. 133–140. DOI: 10.1016/B978-0-12-809633-8.20957-6

Smet J.De., Zimmermann M., Kogadeeva M. et al. 2016. High coverage metabolomics analysis reveals phage-specific alterations to Pseudomonas aeruginosa physiology during infection. Multidisciplinary Journal of Microbial Ecology 10(8): 1823–1835. DOI: 10.1038/ismej.2016.3

Smits S.L., Bodewes R., Ruiz-Gonzalez A. 2014. et al. Assembly of viral genomes from metagenomes. Frontiers in Microbiology 5. DOI: 10.3389/fmicb.2014.00714

Sykilinda N.N., Bondar A.A., Gorshkova A.S. et al. 2014. Complete Genome Sequence of the Novel Giant Pseudomonas Phage PaBG. Genome Announcements 2(1): e00929-13. DOI: 10.1128/genomeA.00929-13

Terzian P., Olo Ndela E., Galiez C. et al. 2021. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genomics and Bioinformatics 3(3): lqab067. DOI: 10.1093/nargab/lqab067

Thompson L.R., Zeng Q., Kelly L. et al. 2011. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proceedings of the National Academy of Sciences 108(39): E757-E764. DOI: 10.1073/pnas.1102164108

Walker J.E. 1992. The NADH: ubiquinone oxidoreductase (complex I) of respiratory chains. Quarterly Reviews of Biophysics 25(3): 253–324

Xia R., Sun M., Balcázar J.L. et al. 2023. Benzo[a]pyrene stress impacts adaptive strategies and ecological functions of earthworm intestinal viromes. Multidisciplinary Journal of Microbial Ecology 17(7): 1004–1014. DOI: 10.1038/s41396-023-01408-x

Оригинальная статья

Поиск предполагаемых полных и почти полных геномов ДНК- содержащих вирусов в метагеномах, полученных из оз. Байкал

Потапов С.А.[1]*, Тупикин А.Е.[2], Тихонова И.В.[1], Жученко Н.А.[1], Белых О.И.[1]

[1] *Лимнологический институт Сибирского Отделения Российской Академии Наук, ул. Улан-Баторская, 3, Иркутск, 664033, Россия*
[2] *Институт химической биологии и фундаментальной медицины Сибирского Отделения Российской Академии Наук, пр. Академика Лаврентьева, 8, Новосибирск, 630090, Россия*

АННОТАЦИЯ. В работе представлен анализ предположительно полных и почти полных геномов бактериофагов, извлечённых из метагеномных данных, полученных из образцов ДНК, выделенной из воды оз. Байкал с применением современных биоинформатических программ. Всего было выявлено 73 последовательности длиной от 13,8 тыс. до 163,7 тыс. оснований, принадлежащие фагам класса Caudoviricetes. Обнаружены два контига, принадлежащие предположительно цианофагам длиной 36,8 тыс. и 163,7 тыс. нуклеотидов, причём в последнем идентифицирована ORF длиной 159 аминокислотных остатков сходная с малым белком теплового шока (Hsp20). Анализ идентифицированных в собранных геномах бактериофагов аминокислотных последовательностей по базе данных PHROG выявил, что 27,5% из них имеют неизвестную функцию, в то время как большая часть из имеющих сходство с известными (23,7%) принадлежит категории «ДНК, РНК и нуклеотидный метаболизм». Также в собранных геномах обнаружен ряд вспомогательных метаболических генов (AMG): *nadM, cysC, cobS, galE, cobT* и др. Большая часть последовательностей, имеющих сходство с последовательностями из базы данных IMG/VR (89,6%) соответствовала последовательностям, полученным из пресноводных водоёмов.

Ключевые слова: метагеномика, бактериофаги, высокопроизводительное секвенирование, оз. Байкал, полный геном

Для цитирования: Потапов С.А., Тупикин А.Е., Тихонова И.В., Жученко Н.А., Белых О.И. Поиск предполагаемых полных и почти полных геномов ДНК- содержащих вирусов в метагеномах, полученных из оз. Байкал // Limnology and Freshwater Biology. 2024. - № 4. - С. 1050-1065. DOI: 10.31951/2658-3518-2024-A-4-1050

1. Введение

В настоящее время метагеномика предоставляет возможность анализировать разнообразие вирусов в различных средах обитания, в частности, в морских и пресноводных экосистемах. Разнообразие вирусов чрезвычайно высоко, однако только небольшая их часть представлена полными геномами (Paez-Espino et al., 2016).

Сборка вирусных геномов из метагеномов является сложной задачей (Smits et al., 2014). На идентификацию вирусных геномов из метагеномов негативно влияют несколько факторов, например, загрязнение невирусными последовательностями, присутствующими в пуле первоначальных данных (Roux et al., 2013), наличие множества коротких последовательностей, полученных при сборке, кроме того, профаги могут быть фланкированы областями, принадлежащими хозяину. При анализе короткие контиги (менее 5-10 тыс. нуклеотидов) чаще всего отбрасывают (Gregory et al., 2019). Сложность в идентификации вирусов так же заключается в отсутствие универсальных маркерных генов и большой вариабельностью длин вирусных геномов. Кольцевые геномы могут быть идентифицированы по наличию терминальных повторов. Вирусный геном также может быть определён путём покрытия ридами известного генома вируса (поиск гомологии), однако ввиду малого количества культивированных вирусов и огромного разнообразия вирусов такой подход в настоящее время имеет ограничения.

Тем не менее из-за сложности культивирования вирусов метагеномный подход на сегодняш-

*Автор для переписки.
Адрес e-mail: poet1988@list.ru (С.А. Потапов)

Поступила: 29 июля 2024; ***Принята:*** 14 августа 2024; ***Опубликована online:*** 30 августа 2024

ний день вероятно стал ключевым, о чём свидетельствуют недавние работы (Gregory et al., 2019; Castro-Nallar et al., 2023), а так же рост количества общедоступных вирусных геномов (фрагментов), собранных из метагеномных данных, с 84 в 2010 г. до 775 тыс. в 2018 г. (Roux et al., 2021). Например, недавно удалось расширить в 12 раз количество вирусных последовательностей из всех океанов, включая Северный Ледовитый океан. В исследовании выявлено 5 экологических зон, формирование которых было обусловлено, в первую очередь, температурой (Gregory et al., 2019). Глобальное исследование виромов из пресноводных экосистем на основе 380 общедоступных вирусных метагеномов позволило восстановить 549 полных высококачественных геномов. Изучение обилия показало, что менее 0,2% вирусных контигов встречаются во всех исследованных биомах, т.е. общий пул контигов минимален. Сделано заключение, что каждый водоём имеет отдельный виром, специфичный для него. Показано, что 41,4% вирусных контигов из пресноводных экосистем не идентифицируются по таксономии, в то время как остальные в основном принадлежат хвостатым фагам (Elbehery and Deng, 2022).

Первые расшифрованные полные геномы бактериофагов из оз. Байкал MD8 *Pseudomonas aeruginosa* (*Siphoviridae*) и *P. aeruginosa* PaBG (*Myoviridae*), получены методом культивирования (Sykilinda et al., 2014; Drucker et al., 2015). При анализе же метагеномных данных из оз. Байкал ранее удалось собрать полный геном фага Baikal-20-5m-C28 длиной 166 тыс. нк., хозяином которого предположительно является *Polynucleobacter* spp. (Cabello-Yeves et al., 2018). Кроме этого, выявлено 16 предполагаемых полных геномов вирофагов, принадлежащие трём семействам *Burtonviroviridae*, *Dishuiviroviridae*, *Omnilimnoviroviridae* (Potapov and Belykh, 2023) и почти полные геномы РНК-содержащих вирусов, полученные из моллюсков (Butina et al., 2023) и водной толщи (Potapov et al., 2023).

Целью работы является получение качественных (т.е. с высокой степенью достоверности) последовательностей полных геномов ДНК вирусов из метагеномных данных, полученных из оз. Байкал, аннотация белковых последовательностей и таксономическая идентификация восстановленных геномов.

2. Материалы и методы

2.1. Отбор образцов, экстракция ДНК и секвенирование

Образцы воды для анализа отобраны из оз. Байкал в трёх его котловинах (Южной, Средней, Северной), а также в проливе Малое Море (Таблица 1).

Отбор проб проводили с глубин от 0 до 50 м (BVP1-8) и от 0 до 15 м (RVP4-6), используя батометры Нискина и смешивали для получения интегральной пробы, всего 20 литров на один образец. Образцы воды последовательно фильтровали через фильтры с размером пор 0,4 мкм и 0,2 мкм (Sartorius, ФРГ) для удаления детрита, зоо-, фито- и бактериопланктона. Далее проводили концентрирование, с помощью ультрафильтрации в тангенциальном потоке с номинальным отсечением по молекулярной массе 50кДа, применяя VivaFlow 200 (Sartorius, ФРГ) и ультрацентрифужные пробирки Vivaspin 15 (Sartorius, ФРГ). Таким образом получали 100 мкл концентрата на образец. Для очистки свободных вирусных частиц от чужеродной ДНК проводили обработку пробы ДНКазой (Thermo Fisher Scientific, MA, США). Реакцию останавливали добавлением 50 mM EDTA и выдерживали при 65˚C 10 минут. Экстрагировали ДНК с помощью фенол-хлороформного метода. Концентрацию ДНК измеряли на Qubit 2.0 (Invitrogen, США). Подготовка библиотек и их секвенирование на Illumina MiSeq (Illumina, США) и MGIseq 2000 (MGI Tech Co., Ltd, КНР) выполнены в ЦКП «Геномика» (ИХБФМ СО РАН, г. Новосибирск, Россия).

Таблица 1. Образцы воды, отобранные для анализа виромов.

Маркировка	Дата отбора	Место отбора	Номер проекта в SRA	Ссылка
BVP1	22.03.18	7 км от п. Листвянка	PRJNA1006167	(Potapov and Belykh, 2023)
BVP2	8.06.18	3 км от п. Листвянка		
BVP3	31.05.18	3 км от п. Турка		
BVP4	4.06.18	3 км от м. Елохин		
BVP5	5.08.18	ц. ст. пролив Малое Море		
BVP6	27.09.18	ц. ст. п. Листвянка – п. Танхой		
BVP7	25.09.18	ц. ст. м. Ухан – м. Тонкий		
BVP8	23.09.18	ц. ст. м. Елохин – п. Давша		
RVP4_DNA	29.07.22	ц. ст. п. Листвянка – п. Танхой	Не депонированы	–
RVP5_DNA	07.08.22	ц. ст. м. Ухан – м. Тонкий		
RVP6_DNA	02.08.22	ц. ст. м. Елохин – п. Давша		

2.2. Биоинформатический анализ

Полученные первоначальные прочтения анализировали в FastQC v. 0.12.1 (Andrews, 2010). Удаление адаптеров, а также фильтрацию по качеству проводили в Trimmomatic v. 0.36 (Bolger et al., 2014). Удаление загрязнений рибосомной ДНК выполняли, используя SortMeRNA v. 4.3.6 (Kopylova et al., 2012); последовательности, принадлежащие человеку, удаляли с помощью программы Bowtie2 v. 2.4.4 (Langmead and Salzberg, 2012), путём картирования ридов на известный геном (GRCh38_noalt_as). Комбинированную сборку осуществляли путём объединения всех образцов («cross-assembly») для получения более длинных контигов и увеличения количества вирусных контигов (Bukin et al., 2023), используя metaSPAdes v. 4.0.0 (Nurk et al., 2017) с дополнительным параметром -k 21, 33, 55, 77. Извлечение вирусных последовательностей проводили с помощью VirSorter2 v. 2.2.4 (Guo et al., 2021) с параметром score > 0,9 и минимальной длине контигов 5000 нк. Затем риды картировали на полученные контиги, предположительно принадлежащие вирусам, с помощью Bowtie2, покрытие считали программой SAMtools v. 1.13 (Li et al., 2009). В дальнейший анализ брали только контиги с глубиной покрытия более 5. Используя программу COBRA v. 1.2.3 (Chen and Banfield, 2024), соединяли собранные последовательности и добивались более высокой точности, посредством увеличения длины и полноты контигов. Далее контиги проверяли с помощью программы CheckV v. 1.0.3 (Nayfach et al., 2021). В дальнейшем использованы исключительно прямые терминальные повторы (direct terminal repeat, DTR) для идентификации полных геномов, поскольку это наиболее устоявшийся подход. Кроме того, учитывали такие параметры как AAI-based completeness > 90%, confidence_level – high. Сходные последовательности сгруппированы на уровне 95% идентичности (ANI), используя скрипты для кластеризации из программы CheckV.

Таксономическую идентификацию вирусных геномов выполняли, применяя geNomad v. 1.8.0 (Camargo et al., 2023b), Diamond v. 2.1.8.162 с параметрами e-value – 10^{-5}, bit score ≥ 50, more sensitive и BlastN v. 2.12.0+ (e-value – 10^{-5}), используя аминокислотную и нуклеотидную базы данных вирусных геномов RefSeq v. 222. Функциональный анализ транслированных ORFs из контигов выполнен, применяя базы данных PHROG v. 4 (Terzian et al., 2021) и VOG v. 219 применяя HHMER v. 3.2.1 (Eddy, 2011). База данных IMG/VR v. 4 (Camargo et al., 2023a) использована для поиска схожих белков из некультивированных вирусов. Также использована программа автоматической аннотации фагов VIBRANT v. 1.2.1 (Kieft et al., 2020) для поиска вспомогательных метаболических генов (auxiliary metabolic gene, AMG). Протеомное дерево строили, используя on-line сервис VipTree v. 4.0 (Nishimura et al., 2017).

Таблица 2. Количество прочтений/контигов после каждого этапа обработки.

Этап обработки	Пар прочтений/ контигов
1. Исходных прочтений	176329692
2. Trimmomatic	152536284
3. SortMeRNA	151248958
4. Удаление последовательностей человека	149200074
5. metaSPAdes, контигов более 5 тыс. нуклеотидов	20929
6. VirSorter2, контигов более 5 тыс. нуклеотидов	8288
7. Bowtie2, SAMtools, глубина чтения более 5	8033
8. COBRA	3191
9. CheckV	73

3. Результаты

Результаты поэтапной обработки представлены в Таблице 2. Всего из полученной микс-сборки после обработки в программе VirSorter2 выявлено 8288 последовательностей, принадлежащих вирусам, что составило 39,6% от всех собранных контигов более 5000 нуклеотидов.

После обработки 8033 последовательностей в программе COBRA, 49,1% последовательностям удалось увеличить длину в среднем на 37,6%. Контроль качества вирусных контигов в CheckV определил, что к высококачественным контигам (high quality > 90%) с DTR принадлежало 73 последовательности. При этом 60 последовательностей принадлежат полным геномам, а 13 последовательностей характеризуются как почти полные (группа extend partial, COBRA). Длина этих предполагаемых полных и почти полных геномов вирусов варьировала от 13871 до 163727 нуклеотидов (дополнительный материал). Количество ORFs было в диапазоне от 20 до 216. Статистическая информация представлена на рисунке 1.

Таксономическая идентификация вирусных геномов на основе geNomad (virus_score > 0,96) показала, что все геномы принадлежали классу Caudoviricetes, при этом до семейства *Autographiviridae* geNomad определено 4 последовательности (NODE_7814, NODE_610, NODE_598, NODE_665). Стоит отметить, что мы идентифицировали также вирофаги, которые обнаружили ранее, но в этой статье их не рассматриваем (исключены на этапе CheckV) т.к. они были подробно описаны в статье (Potapov and Belykh, 2023).

Функциональный анализ, проведённый по базе PHROG, выявил, что категория «ДНК, РНК и нуклеотидный метаболизм» наиболее представлена (23,7%), кроме того, 27,5% аминокислотных последовательностей имели неизвестную функцию (Рис. 2). Структурные белки составили 38,3% (категории

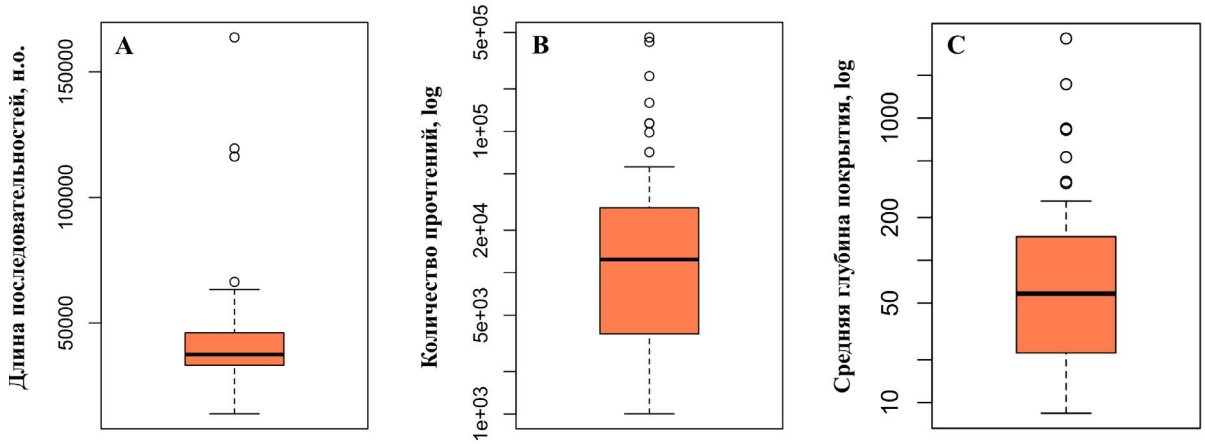

Рис.1. Статистическая информация о 73 последовательностях фагов. А – диапазон длин последовательностей, нуклеотидных оснований, В – количество прочтений на геном, С – средняя глубина покрытия геномов.

«Капсид и упаковка», «Хвост», «Коннектор»). Во всех геномах бактериофагов по базам данных PHROG и VOG обнаружена большая субъединица терминазы.

В геноме NODE_40, предположительно принадлежащего цианофагу, выявлена ORF длиной 159 аминокислотных остатков сходная с малым белком теплового шока (Hsp20, YP_009134378). Кроме того, в этом геноме выявлены гены, связанные с устойчивостью микроорганизмов к окислительному стрессу, такие как NAMPT (nicotinamide phosphoribosyltransferase) и *nadM* (bifunctional NMN adenylyltransferase/nudix hydrolase). NODE_334 включал ген *cysC*, который вовлечен в ассимиляционное восстановление сульфата. Гены *cobS* (NODE_4873, NODE_40, NODE_506), *cobT* (NODE_4873) участвуют в биосинтезе кобаламина. *PurA* (adenylosuccinate synthase) и *purB* (adenylosuccinate lyase), обнаруженные в NODE_996 участвуют в пуриновом метаболизме. Так же выявлены другие вспомогательные метаболические гены: *rfbC*, DNMT1, *galE*, *lpxA*, TSTA3, *pimC*.

Шесть последовательностей имели в составе интегразу (NODE_506, NODE_926, NODE_7814, NODE_610, NODE_598, NODE_660), сходство аминокислотных последовательностей варьировало от 28,2 до 39%.

Из 4266 ORFs 4108 (96,3%) имели сходство с некультивированными представителями вирусов из базы данных IMG/VR. Из них 89,6% соответствовали представителям, полученным из пресноводных экосистем, при этом большая часть из них (78,8%) получена из озёр, 16,5% из рек, 2,5% из прудов. Только 3,2% из 4108 ORFs были морского происхождения. Остальные принадлежали представителям, полученным из других источников (почва, сточные воды и др.).

По базе данных RefSeq 126 ORFs имели сходство с фагами, недавно выделенными на штаммах *Flavobacterium* из Балтийского моря (Nilsson et al., 2020; Hoetzinger et al., 2021). Идентичность последовательностей на аминокислотном уровне варьировала от 21,5 до 87%. 54 ORFs сходны с *Nonlabens* phage (изоляты P12024S и P12024L), хозяином которых является бактерия *Persicivirga* sp. IMCC12024, изолированная из прибрежной воды Жёлтого моря (Южная Корея). Сходство белковых последовательностей варьировало от 22,4 до 68,8 %.

Из выявленных рамок считывания 203 последовательности были сходны с белками 63 различных цианофагов на аминокислотном уровне и имели сходство от 22,1 до 81,3%, при этом наименьшее значение e-value (0) отмечено с основным капсид-

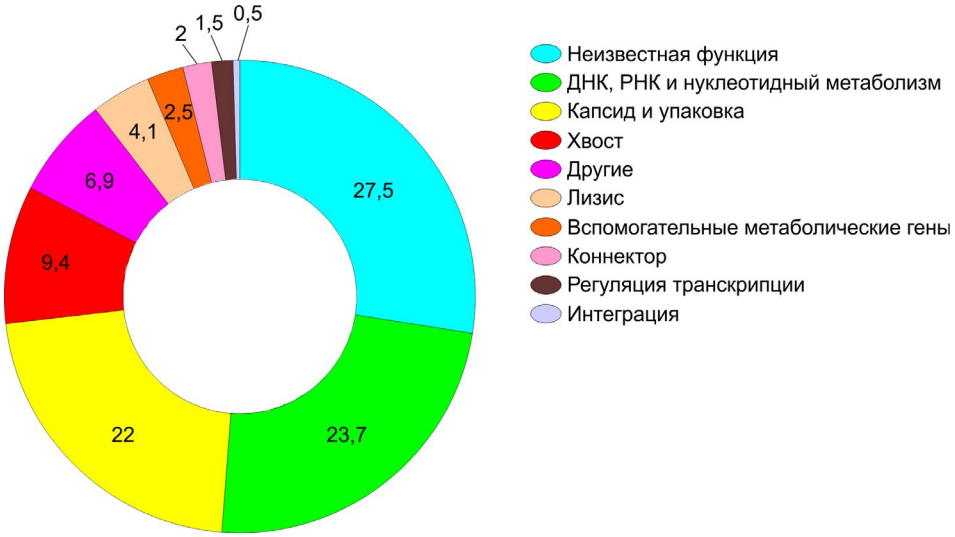

Рис.2. Функциональный анализ аминокислотных последовательностей 73 выявленных фагов. Доли категорий приведены в процентах.

ным белком (YP_004421726) *Synechococcus* phage S-CBS3 (сходство – 62,4%, покрытие – 99,9%) и ДНК полимеразой (YP_010669768) *Synechococcus* phage S-SCSM1 (сходство – 55,7%, покрытие – 99,6%). *Synechococcus* phage S-CBS3 выделен из штамма *Synechococcus* sp. CB0202 (источник изоляции – Чесапикский залив), *Synechococcus* phage S-SCSM1 изолирован из штамма *Synechococcus* sp. WH 7803 (источник изоляции Южно-Китайское море). Две последовательности NODE_40 (163,7 тыс. нукл., 216 ORF) и NODE_1081 (36,8 тыс. нукл., 49 ORF) имели 50 и 23 ORF, сходных с таковыми цианофагов из базы данных RefSeq, что говорит с большой долей вероятности об их принадлежности к цианофагам. Другие последовательности, имеющие сходные белки с цианофагами, но с меньшим количеством на геном сложно определить вследствие недостатка известных цианофагов в базах данных.

75 аминокислотных последовательностей выявленных фагов были сходны с различными изолятами *Ralstonia* phage (NC_047946, NC_047888, NC_030948 и др.), при этом сходство на аминокислотном уровне варьировало от 25,6 до 75,1%.

В целом, основываясь на результатах blastp по базе данных RefSeq можно заключить, что только 34,7% ORFs имели сходство с известными белками. Помимо этого, 80,8% последовательностей из имеющих сходство с белками из базы данных показывали сходство менее 50% с известными, что может указывать на обнаружение новых фагов, представляющих так называемый пул темной вирусной материи (viral dark matter). В поддержку этого так же свидетельствует низкий процент сходства с известными вирусами из нуклеотидной базы полных геномов RefSeq, так максимальное сходство и покрытие зафиксировано с *Ralstonia* phage RsoP1EGY (сходство – 13%, покрытие – 67,4%), *Synechococcus* phage S-CBS3 (сходство – 25%, покрытие – 70%), *Ralstonia* phage RsoP1EGY (сходство – 18%, покрытие – 69,6%).

Анализ представленности на основе TPM (transcripts per kilobase million) показал, что 32 последовательности (42,5 %) присутствуют во всех образцах. NODE_923 выявлен только в образце BVP5 (пролив Малое Море). Две последовательности собраны только из прочтений 2022 г. (NODE_996, NODE_660). В летний период 2018 г. (BVP5, BVP6, BVP7, BVP8) отсутствовали 8 последовательностей (NODE_598, NODE_660, NODE_547, NODE_996, NODE_5829, NODE_506, NODE_665, NODE_1217). Этот вид анализа может свидетельствовать как о распространённости фагов в озере во всех котловинах и сезоны, так и об уникальности и репликации фагов в отдельные временные промежутки.

Протеомное дерево, построенное на основе собранных нами геномов, демонстрирует как сформированные кластеры, так и отдельные ветви, при этом сформированные кластеры содержат ближайших родственников по базе данных RefSeq. Данный тип анализа так же позволяет выявить предполагаемых хозяев, например, наличие ближайшего родственника в кластере, для которого известен хозяин и содержащего собранные нами геномы. Таким образом обнаружены 4 филы потенциальных хозяев Pseudomonodata, Bacteroidota, Cyanobacteriota и Bacillota (Рис. 3).

4. Обсуждение

В данной работе нам удалось выявить 60 предположительно полных и 13 почти полных геномов фагов (vMAG, viral metagenome-assembled genome), принадлежащих классу Caudoviricetes из метагеномов пресноводного оз. Байкал.

Из-за мозаичности фаговых геномов, а также недостатке их в базах данных, возникает нетривиальная задача определения ближайшего родственника на видовом, родовом уровне и даже на уровне семейства. Следует отметить, что мы использовали строгие условия идентификации вирусов при использовании программы VirSorter2: max_score > 0,9 – высокая достоверность, CheckV: confidence_level – high, aai_completeness более 90%, checkv_quality - complete, наличие DTR, contamination – 0 и др. т.к. это оптимальный вариант для получения высококачественных полных геномов.

В геномах фагов обнаружен ряд вспомогательных метаболических генов. Изменение продуктивности на уровне экосистемы происходит посредством горизонтального переноса экологически важных генов и экспрессии кодируемых вирусами AMG (Hurwitz and U'Ren, 2016). Эти гены экспрессируются во время инфекции, увеличивая и перенаправляя энергию и ресурсы на производство вируса (Thompson et al., 2011; Hurwitz and U'Ren, 2016; Smet et al., 2016; Howard-Varona et al., 2018). Выявленные вспомогательные гены демонстрируют участие вирусов в биогеохимических реакциях. Так, *cobS* участвует в биосинтезе кобаламина и может поддерживать синтез дезоксинуклеотидов. Ген *cysC* вовлечен в ассимиляционное восстановление сульфата. Присутствие фагов с этим геном в пресноводных экосистемах, может влиять на цикл серы через процесс ассимиляционной сульфатредукции. Дыхательный комплекс I (NADH: убихинон оксидоредуктаза) использует энергию, высвобождаемую при переносе электронов с NADH на хинон для перекачки протонов через плазматическую мембрану (Walker, 1992). Биодоступность этого комплекса может быть изменена посредством экспрессии вирусных вспомогательных метаболических генов, участвующих в биосинтезе NAD + (NAMPT, *nadM*). Недавно эти гены обнаружены в контигах, полученных из кишечника мышей, отнесенных к Caudoviricetes (Ishola et al., 2024), а также в фагах кишечника дождевых червей (Xia et al., 2023). Ген *galE*, кодирующий UDP-глюкозо-4-эпимеразу, опосредует преобразование UDP-галактозы и UDP-глюкозы в метаболизме галактозы, и вероятно, позволяет вирусу участвовать в метаболизме углеводов (Heyerhoff et al., 2022). В целом AMGs в фагах и их роль до сих пор остаются малоизученными, но их важность в выживании фагов не вызывает сомнений.

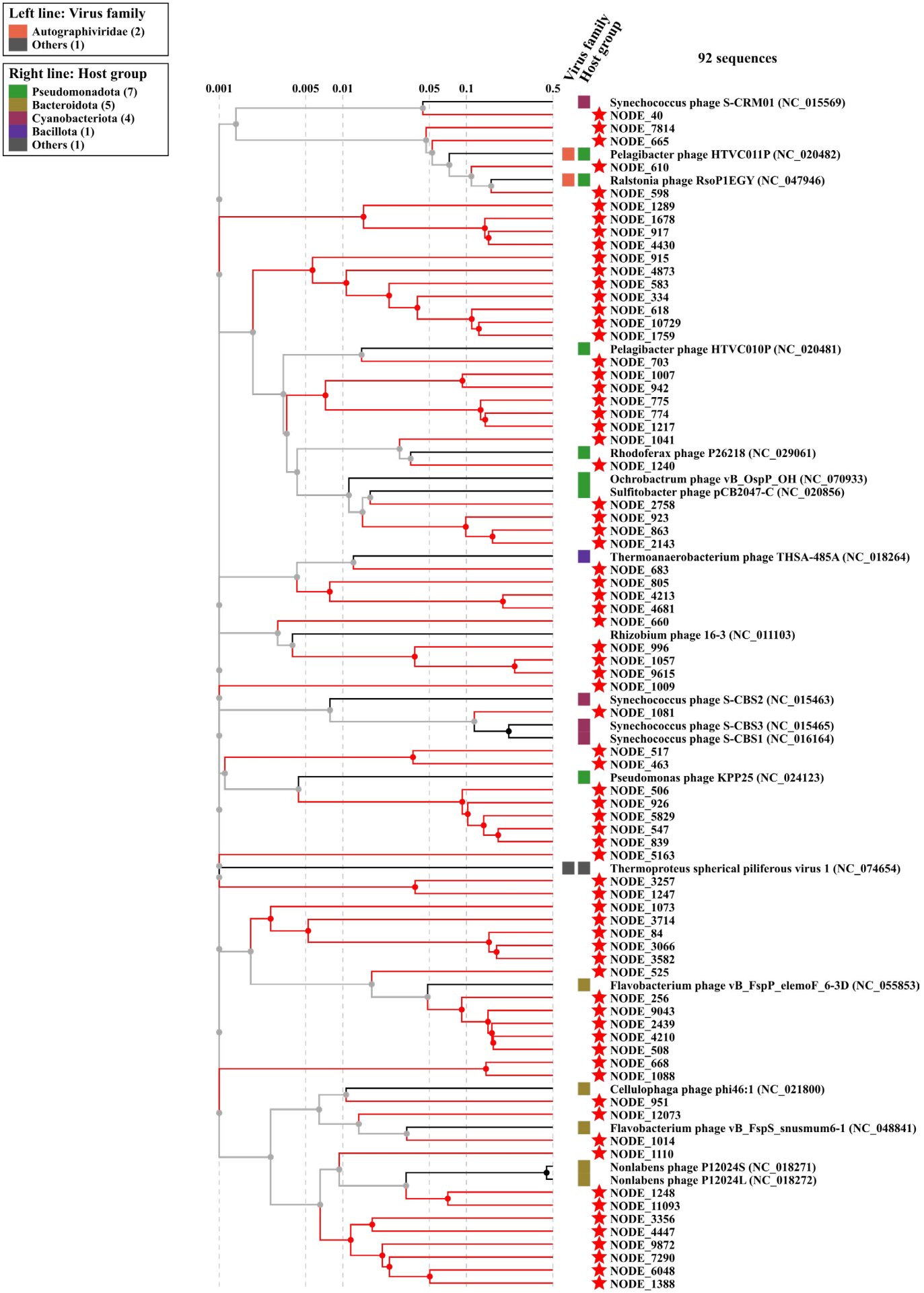

Рис.3. Протеомное дерево, построенное на основе сравнения транслированных нуклеотидных последовательностей геномов (tBLASTx) с ближайшими родственниками, идентифицированных с помощью VipTree. Звездочкой отмечены последовательности из этого исследования. Цветные квадраты означают принадлежность к известному таксону референса и его хозяина.

В последовательности NODE_40, предположительно принадлежащей цианофагу, обнаружен малый белок теплового шока (sHSP). Ранее показано, что он присутствует у некоторых цианофагов (морских и пресноводных), которые инфицируют одноклеточные цианобактерии *Synechococcus* и *Prochlorococcus* (Dreher et al., 2011; Maaroufi and Tanguay, 2013). Показано, что цианофаги приобрели ген, кодирующий sHSP, от бактериального предка класса A с помощью латерального переноса генов (Maaroufi and Tanguay, 2013).

Обнаружение собранных фаговых геномов из оз. Байкал в различные сезоны и годы демонстрирует, что половина из них постоянно присутствует в те временные отрезки, что мы наблюдали. И наоборот, часть из них обнаружена только в определённое время и определённом месте. Пониманию присутствия фагов в те или иные сезоны поможет ежемесячный отбор образцов в течение года и их секвенирование, что является целью нашего дальнейшего исследования.

Соблюдение требований пробоподготовки образцов для получения качественных геномов вирусов из метагеномных данных, в частности, избавления от бактериальной и эукариотической ДНК, предфильтрация и использование ДНКазы, а также секвенирование с максимально возможной глубиной чтения является необходимым условием. Данные этапы позволят из общего количества чтений получить достаточное количество для сборки геномов вирусов. Кроме того, следует уделять внимание своевременному обновлению баз данных в анализе, т.к. рост количества известных последовательностей может способствовать выявлению и идентификации большего количества новых вирусов. Помимо этого, появление и использование новых программ упрощает анализ, а также повышает вероятность идентификации выявленных последовательностей.

5. Выводы

Используя биоинформатический анализ данных, полученных из секвенированных образцов ДНК воды из оз. Байкал (метагеномов фракции менее 0,2 мкм) восстановлены 60 полных и 13 почти полных геномов бактериофагов. Таксономический анализ показал низкое сходство полученных геномов с имеющимися геномами вирусов в базе данных RefSeq. Обнаружено два генома, предположительно принадлежащие цианофагам, длиной 36,8 тыс. (NODE_1081) и 163,7 тыс. (NODE_40) нуклеотидов, в последнем идентифицирован малый белок теплового шока (sHSP). Выявлен ряд вспомогательных метаболических генов в геномах полученных фагов: *nadM*, *cysC*, *cobS*, *galE*, *cobT* и др. Показано, что половина собранных геномов фагов присутствовала во все сезоны и во всех трёх котловинах и проливе М. Море, остальные обнаружены только в определённое время и в определённом месте.

6. Финансирование

Работа выполнена в рамках темы госзадания ЛИН СО РАН № 0279-2021-0015 «Исследования вирусных и бактериальных сообществ как основы стабильного функционирования пресноводных экосистем и эффективного ответа в условиях антропогенного воздействия».

Благодарности

Авторы благодарят команду НИС «Г. Титов» за помощь в отборе образцов.

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов.

Список литературы

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data [Электронный ресурс]. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bolger A.M., Lohse M., Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15): 2114–2120. DOI: 10.1093/bioinformatics/btu170

Bukin Y.S., Bondaryuk A.N., Butina T.V. 2023. Performance Analysis of Cross-Assembly of Metatranscriptomic Datasets in Viral Community Studies. Mathematical Biology & Bioinformatics 18(2): 418–433. DOI: 10.17537/2023.18.418

Butina T.V., Zemskaya T.I., Bondaryuk A.N. et al. 2023. Viral Diversity in Samples of Freshwater Gastropods Benedictia baicalensis (Caenogastropoda: Benedictiidae) Revealed by Total RNA-Sequencing. International Journal of Molecular Sciences 24(23): 17022. DOI: 10.3390/ijms242317022

Cabello-Yeves P.J., Zemskaya T.I., Rosselli R. et al. 2018. Genomes of novel microbial lineages assembled from the sub-ice waters of Lake Baikal. Applied and Environmental Microbiology 84(1): e02132-17. DOI: 10.1128/AEM.02132-17

Camargo A.P., Nayfach S., Chen I.-M.A. et al. 2023. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. Nucleic Acids Res 51(D1): 733–743. DOI: 10.1093/nar/gkac1037

Camargo A.P., Roux S., Schulz F. et al. 2023. Identification of mobile genetic elements with geNomad. Nature Biotechnology DOI: 10.1038/s41587-023-01953-y

Castro-Nallar E., Berríos-Farías V., Díez B. et al. 2023. Seasonal and Spatially Distributed Viral Metagenomes from Comau Fjord (42°S), Patagonia. Microbiology Resourse Announcements 12(4): 12:e00082-23. DOI: 10.1128/mra.00082-23

Chen L., Banfield J.F. 2024. COBRA improves the completeness and contiguity of viral genomes assembled from metagenomes. Nature Microbiology 9(3): 737–750. DOI: 10.1038/s41564-023-01598-2

Dreher T.W., Brown N., Bozarth C.S. et al. 2011. A freshwater cyanophage whose genome indicates close relationships to photosynthetic marine cyanomyophages. Environmental Microbiology 13(7): 1858–1874. DOI: 10.1111/j.1462-2920.2011.02502.x

Drucker V.V., Bondar A.A., Gorshkova A.S. et al. 2015. Search and studies of autochthonous bacteriophages in different biotopes of Lake Baikal. Contemporary Problems of

Ecology 12: 143-154. DOI: 10.1134/S1995425519020045 (in Russian)

Eddy S.R. 2011. Accelerated Profile HMM Searches. PLoS Computational Biology 7(10): e1002195. DOI: 10.1371/journal.pcbi.1002195

Elbehery A.H.A., Deng L. 2022. Insights into the global freshwater virome. Frontiers Microbiology 13. DOI: 10.3389/fmicb.2022.953500

Gregory A.C., Zayed A.A., Conceição-Neto N. et al. 2019. Marine DNA viral macro- and microdiversity from pole to pole. Cell 177(5): 1109-1123.e14. DOI: 10.1016/j.cell.2019.03.040

Guo J., Bolduc B., Zayed A.A. et al. 2021. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome 9(1): 1–13. DOI: 10.1186/s40168-020-00990-y

Heyerhoff B., Engelen B., Bunse C. 2022. Auxiliary Metabolic Gene Functions in Pelagic and Benthic Viruses of the Baltic Sea. Frontiers in Microbiology 13. DOI: 10.3389/fmicb.2022.863620

Hoetzinger M., Nilsson E., Arabi R. et al. 2021. Dynamics of Baltic Sea phages driven by environmental changes. Environmental Microbiology 23(8): 4576–4594. DOI: 10.1111/1462-2920.15651

Howard-Varona C., Hargreaves K.R., Solonenko N.E. et al. 2018. Multiple mechanisms drive phage infection efficiency in nearly identical hosts. ISME J 12(6): 1605–1618. DOI: 10.1038/s41396-018-0099-8

Hurwitz B.L., U'Ren J.M. 2016. Viral metabolic reprogramming in marine ecosystems. Current Opinion in Microbiology 31: 161–168. DOI: 10.1016/j.mib.2016.04.002

Ishola O.A., Kublik S., Durai Raj A.C. et al. 2024. Comparative Metagenomic Analysis of Bacteriophages and Prophages in Gnotobiotic Mouse Models. Microorganisms 12(2): 255. DOI: 10.3390/microorganisms12020255

Kieft K., Zhou Z., Anantharaman K. 2020. VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of virome function from genomic sequences. Microbiome 8 (90) DOI: 10.1186/s40168-020-00867-0

Kopylova E., Noé L., Touzet H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28(24): 3211–3217. DOI: 10.1093/bioinformatics/bts611

Langmead B., Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2 // Nature Methods 9(4): 357–359. DOI: 10.1038/nmeth.1923

Li H., Handsaker B., Wysoker A. et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16): 2078–2079. DOI: 10.1093/bioinformatics/btp352

Maaroufi H., Tanguay R.M. 2013. Analysis and Phylogeny of Small Heat Shock Proteins from Marine Viruses and Their Cyanobacteria Host. PLoS One 8(11): e81207. DOI: 10.1371/journal.pone.0081207

Nayfach S., Camargo A.P., Schulz F. et al. 2021. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nature Biotechnology 39(5): 578–585. DOI: 10.1038/s41587-020-00774-7

Nilsson E., Bayfield O.W., Lundin D. et al. 2020. Diversity and Host Interactions among Virulent and Temperate Baltic Sea Flavobacterium Phages. Viruses 12(2): 158. DOI: 10.3390/v12020158

Nishimura Y., Yoshida T., Kuronishi M. et al. 2017. ViPTree: the viral proteomic tree server. Bioinformatics 33(15): 2379–2380. DOI: 10.1093/bioinformatics/btx157

Nurk S., Meleshko D., Korobeynikov A. et al. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Research 27(5): 824–834. DOI: 10.1101/gr.213959.116

Paez-Espino D., Eloe-Fadrosh E.A., Pavlopoulos G.A. et al. 2016. Uncovering Earth's virome. Nature 536(7617): 425–430. DOI: 10.1038/nature19094

Potapov S., Gorshkova A., Krasnopeev A. et al. 2023. RNA-Seq virus fraction in Lake Baikal and treated wastewaters. International Journal of Molecular Sciences 24(15): 12049. DOI: 10.3390/ijms241512049

Potapov S.A., Belykh O.I. 2023. Virophages Found in Viromes from Lake Baikal. Biomolecules 13(12): 1773. DOI: 10.3390/biom13121773

Roux S., Krupovic M., Debroas D. et al. 2013. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. Open Biology 3(12): 130160. DOI: 10.1098/rsob.130160

Roux S., Matthijnssens J., Dutilh B.E. 2021. Metagenomics in Virology. Encyclopedia of Virology. Elsevier, pp. 133–140. DOI: 10.1016/B978-0-12-809633-8.20957-6

Smet J.De., Zimmermann M., Kogadeeva M. et al. 2016. High coverage metabolomics analysis reveals phage-specific alterations to Pseudomonas aeruginosa physiology during infection. Multidisciplinary Journal of Microbial Ecology 10(8): 1823–1835. DOI: 10.1038/ismej.2016.3

Smits S.L., Bodewes R., Ruiz-Gonzalez A. 2014. et al. Assembly of viral genomes from metagenomes. Frontiers in Microbiology 5. DOI: 10.3389/fmicb.2014.00714

Sykilinda N.N., Bondar A.A., Gorshkova A.S. et al. 2014. Complete Genome Sequence of the Novel Giant Pseudomonas Phage PaBG. Genome Announcements 2(1): e00929-13. DOI: 10.1128/genomeA.00929-13

Terzian P., Olo Ndela E., Galiez C. et al. 2021. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genomics and Bioinformatics 3(3): lqab067. DOI: 10.1093/nargab/lqab067

Thompson L.R., Zeng Q., Kelly L. et al. 2011. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proceedings of the National Academy of Sciences 108(39): E757-E764. DOI: 10.1073/pnas.1102164108

Walker J.E. 1992. The NADH: ubiquinone oxidoreductase (complex I) of respiratory chains. Quarterly Reviews of Biophysics 25(3): 253–324

Xia R., Sun M., Balcázar J.L. et al. 2023. Benzo[a]pyrene stress impacts adaptive strategies and ecological functions of earthworm intestinal viromes. Multidisciplinary Journal of Microbial Ecology 17(7): 1004–1014. DOI: 10.1038/s41396-023-01408-x