Mathematical Biology and Bioinformatics
2020. V. 15. № Suppl. P. t88-t108. doi: 10.17537/2020.15.t88

Ttranslation of the original article published in Russian
Zimin A.A., Nikulin N.A., Nazipova N.N. Mathematical Biology and Bioinformatics. 2019; 14(2):683-704. doi: 10.17537/2019.14.683

========== TRANSLATIONS OF PUBLISHED ARTICLES ==========

Homologs of Bacteriophage T4 RNA Ligase 2 in Metagenomes of Ocean Microbiota

Zimin A.A. *[1], Nikulin N.A.[1], Nazipova N.N.[2]

[1]*G.K. Scriabin Institute of Biochemistry and Physiology of Microorganisms Russian Academy of Sciences, Pushchino, Russia*
[2]*Institute of Mathematical Problems of Biology RAS, Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Pushchino, Russia*

Annotation. RNA ligase 2 from a T4 phage is a unique enzyme that is, in contrast to other RNA ligases, functionally similar to DNA ligases, as well as related to editing RNA ligases of parasitic *Trypanosoma* and *Leishmania*. This enzyme is potentially a promising tool in molecular biology and molecular genetics. RNA ligases 2 are present in a limited number of genomes, which, moreover, are strongly scattered throughout the tree of life. In this work, the search for homologues of RNA ligase 2 was performed in the Global Ocean Sampling Expedition (GOS) database of ocean pelagic microbiota genetic data and the database of genetic studies of deep-sea sedimentary microbiota - LCGC14. In the metagenomes of the pelagic and sedimentary deep-sea microbiota, 6 and 12 homologues of RNA ligase 2 of bacteriophage T4 were found, respectively. Phylogenetic analysis of the detected amino acid sequences showed that most of them are similar to the homologs of RNA ligase 2 from bacteria and fungi. On the branch of the phylogenetic tree, common for homologues from T4-type bacteriophages and *Euglenoidea*, five homologues of oceanic origin were found. For two of them, molecular modeling of the structure of proteins was carried out and it was shown with a high reliability the similarity with the structure of RNA ligase 2 of bacteriophage T4 and editing RNA ligase of *Trypanosoma*. This result indicates the presence both in the surface waters of the open ocean and in the bottom deep-sea sediments of new, yet unknown, organisms, whose genomes encode such a rare enzyme.

Key words: *oceanic metagenomes, genetic studies of deep-sea sedimentary and pelagic microbiota, RNA ligase 2, genomics of bacteriophages.*

INTRODUCTION

RNA ligases belong to the superfamily of nucleotidyltransferases. Along with them the family includes enzymes capping mRNA, tRNA ligases and DNA ligases [1]. All enzymes of this superfamily catalyze the formation of a phosphodiester bond by a conservative three-step mechanism, which uses ATP, GTP, or NAD + as a highly active cofactor [1–3].

RNA ligases (EC 6.5.1.3) are enzymes connecting the ends of RNA. They participate in the processes of repair, splicing and editing of RNA. Unlike widespread DNA ligases, RNA ligases have a narrower phylogenetic distribution. The search for homologous sequences revealed RNA ligases in all three fila of cell life forms, but only in a limited subset of species [4]. RNA ligases are divided into two large families [3, 5].

The Rnl1 family includes the RNA ligase 1 (Rnl1) from the bacteriophage T4 and its homologs [2] and tRNA ligase from fungi, yeasts, and plants [3, 5, 6]. These enzymes repair

breaks of the single-stranded RNA made by the site-specific nucleases. The RNA ligase 1 enzyme is found in viruses, mammals, and fungi [6]. The viruses, for example, the bacteriophage T4, use this enzyme to protect themselves against the bacterial antiviral strategies [2], but it is also involved in the splicing of tRNA introns [5] and in non-traditional RNA splicing, initiated by improper assembly of endoplasmic reticulum proteins.

The Rnl2 family includes the RNA ligase 2 of T4 bacteriophage and ligases editing mitochondrial messenger RNAs that can be found in protozoa such as *Trypanosoma* and *Leishmania*. These enzymes repair single-stranded breaks in double-stranded RNA due to the presence of a complementary chain-bridge [4, 7, 8]. RNA ligases 2 have a wide but sparse distribution throughout the tree of life [6]: they are found mainly in viruses with the archetypal example of RNA ligase 2 of the bacteriophage T4 [3], as well as in bacteria, while in archaea and eukaryotes only a few of these proteins are known. The biological role of bacterial RNA ligases 2 is unknown, except of *Trypanosoma* kinetoplast [9–12].

RNA ligases share six conserved nucleotidyltransferase motifs with DNA ligases; however, the similarity level of their consensus is quite low. This makes the classification of RNA ligases difficult and its result contradictory.

For completeness, we shall mention the reports on two non-canonical RNA ligases of archaea *Pyrococcus horikoshii*. The first ligase is the putative 2'-5'-RNA ligase, the structure of which was studied [13]. The second, RtcB, is a very unusual GTP-dependent ligase that attaches the 2', 3'-cyclophosphate or 3'-phosphate end of RNA to the 5'-hydroxyl end of RNA. The structure of this enzyme and the mechanism of its interaction with the new cofactor protein (Archease) have been characterized a short time ago [14, 15].

Like DNA ligases, RNA ligases are widely used in molecular biology. RNA ligases 1 and 2 from the T4 bacteriophage have become indispensable components of many methods for the rapid amplification of cDNA ends, labeling of 3'-RNA, and, most importantly at present, establishing of libraries for miRNA sequencing [16]. ATP-dependent RNA ligases capable of forming phosphodiester bonds between the 5'-phosphate and 3'-hydroxyl tips will be in the focus of attention of experimenters for a long time to come.

RESEARCH OBJECTIVE

Until now, it was believed that RNA ligases 2 are rare and highly scattered throughout the tree of life. Only two members of this family are well studied: mitochondrial RNA-editing ligase from parasitic *Trypanosoma* (*Kinetoplastea*), a promising drug target, and RNA ligase 2 of bacteriophage T4, a working tool of modern molecular biology [17].

Establishing the origin, evolution, and biological role of this unusual enzyme requires a search for its homologs wherever possible. Biota of the oceans is a huge and poorly studied source of genes. The aim of this study is to search for homologs of well-studied phage polynucleotide ligases in metagenomes of oceanic microbiota.

The recent extensive metagenomic studies [18, 19] examined the marine planktonic microbiota from the surface water of the Middle Atlantic and near the Panama Canal in the Pacific Ocean. The total 7.7 million DNA sequences were identified with a total length of 6.3 Gbp. The GOS database contains more than 6 million amino acid sequences translated from DNA.

To analyze the DNA sequences of the pelagic microbiota, a fraction (0.1–0.8 μm) was extracted from 37 surface samples and collected during the first phase of the Sorcerer II global ocean sampling expedition (GOS) [18, 19]. The aim of the research was to characterize the viral sequences from this database in terms of their occurrence and distribution in a variety of aquatic ecosystems. For these reasons, the authors used comparative genomic analysis to find the function of viral sequences by clustering sequence similarity and to elucidate the importance of capturing by ocean viruses host genes encoding environmentally significant metabolic functions in the aquatic environment. This is important in terms of the possibility of horizontal gene transfer.

Another large database of ocean genetic data was created in the course of the study of deep-sea sediments. This work provided the data on the metagenomes of the deep-sea marine microbiota of the bottom sediments of the Arctic Ocean in the region of the underwater Arctic Mid-Ocean Ridge [20]. DNA sequences were obtained by deep metagenomic DNA sequencing of a GC14 sediment sample, resulting in a smaller set of genetic sequences (LCGC14, 8.6 Gb). Then the DNA amplification (MDA) was performed and a larger set of metagenomic data was obtained (LCGC14AMP, 56.6 Gbp) (http://opensource.scilifelab.se/) [20]. At the same time, databases of amino acid sequences from genetic data were created; they were obtained by translation with the standard genetic code.

In this work, the GOS and LCGC14 databases were reviewed for the presence of sequences similar to RNA ligase 2 of the T4 bacteriophage. After the creation of a set of homologs of RNA ligase 2 of bacteriophage T4 in metagenomes of the marine microbiota, a phylogenetic study was carried out. The work is devoted to the search in oceanic metagenomes for previously unknown homologues of RNA ligase 2 and their analysis in the context of the evolutionary process. The sequence selected for comparison from the classical object of biology of bacteriophage T4 is especially interesting due to its viral nature. It was shown in the work that the sequences for the phylogenetic tree of marine homologues of RNA ligase 2 are often encoded by the genomes of giant viruses, rather new objects of biology, which may be interesting for characterizing these new life forms.

MATERIALS AND METHODS

Characterization of RNA ligase 2 sequences from closely related organisms

To characterize the amino acid sequence of T4 bacteriophage RNA ligase 2, we carried out its phylogenetic comparison with a number of homologs from the related bacteriophages. For this purpose, a series of amino acid sequences of *rnlB* gene products from Genbank selected from the well annotated [21] genomes of the related bacteriophages (*Tevenvirinae* subfamily). In some cases, not one, but two proteins were present in the annotations of the genome sequences in the RNA-ligase 2 cluster, in such cases both paralogs were included in the sample. When comparing the phylogenetic tree *Tevenvirinae*, which was built on the base of amino acid sequences of RnlB protein homology, with the currently used classification of viruses by genomic identity [22, 23], it was found that the resulting tree generally corresponds to the current understanding of the evolution of these viruses. Interestingly, the amino acid sequences of RNA ligases of the second genus *Tequatrovirus* were divided into two clades, one of which was supplemented with the proteins of phages from the genus *Mosigvirus*. Since these two genera are fairly close to each other, we can expect the recombination processes that occurred between the ancestors of the representatives of this combined clade in the region of the studied locus, the nucleotide sequences of which encode the homologs of RnlB proteins. Otherwise, there is a clear correlation between the similarity of the genomes and amino acid sequences of RNA ligases 2 within the examined genera of the subfamily *Tevenvirinae*. The possible recombination between genera, one of which includes T4; a wide range of virus hosts of this subfamily; the most diverse ecological niches occupied by these host bacteria; as well as the wide distribution of homologs of the gene encoding RNA ligase 2 in T4 related phages justify the use of this protein as the target sequence for the analysis of marine metagenomes both in terms of phylogeny and ecology. The phylogenetic analysis also tells about the distribution of this enzyme in phages of this subfamily developing in a wide variety of ecological niches, including marine biomes.

Table 1 lists 189 RNA ligases 2 (RnlB proteins) from closely related phage T4 bacteriophages that were found in GenBank [24] by the use of PSI-BLAST tool [25], that were applied to analyze the degree of conservation of the amino acids of this enzyme. The classification of T-even bacteriophages was taken from GenBank [23]; the unclassified part of the family is at the beginning of the table. Colored circles are placed next to the names of the

genera. The last column of the table shows the GenBank identifiers for the amino acid sequences of RNA ligases 2 of the analyzed bacteriophages. In those rows of the last column, where the "*" sign is used instead of the GenBank identifier, the amino acid sequence of RNA ligase 2 was obtained using the RASTtk genome annotation software [21].

Table 1. Bacteriophages, RNA ligases 2 of which were used to analyze the degree of conservatism of the amino acid sequence, grouped by genus

№	Genus	Species	Genbank IDs
1	unclassified *Tevenvirinae*	*Acinetobacter* phage 133	YP_004300608.1
2		*Acinetobacter* phage AM101	AWY10415.1
3		*Acinetobacter* phage Ac42	YP_004009372.1
4		*Acinetobacter* phage Acj61	YP_004009627.1
5		*Acinetobacter* phage Acj9	YP_004010147.1
6		*Acinetobacter* phage KARL-1	AXY82640.1
7		*Acinetobacter* phage ZZ1	YP_006488989.1
8		*Acinetobacter* phage vB_ApiM_fHyAci03	AXF40578.1
9		*Aeromonas* phage 65.2	APU01459.1
10		*Aeromonas* phage 65	YP_004300909.1
11		*Aeromonas* phage AS-szw	ATI17438.1
12		*Aeromonas* phage AS-zj	ASU00157.1
13		*Aeromonas* virus Aeh1	NP_944126.1
14		*Aeromonas* phage AsFcp_2	QAX98490.1
15		*Aeromonas* phage Asswx_1	QAX97878.1
16		*Aeromonas* phage Aswh_1	QAY01272.1
17		*Aeromonas* phage CC2	YP_007010325.1
18		*Aeromonas* phage PX29	YP_009011662.1
19		*Citrobacter* phage IME-CF2	YP_009218766.1
20		*Citrobacter* phage Margaery	YP_009195056.1
21		*Citrobacter* phage Maroon	AYJ73100.1
22		*Citrobacter* phage Miller	YP_009097842.1
23		*Citrobacter* phage vB_CfrM_CfP1	YP_009285773.1
24		*Enterobacteria* phage RB16	YP_003858534.1
25		*Enterobacteria* phage RB43	YP_239225.1
26		*Cronobacter* phage vB_CsaM_GAP161	YP_006986511.1
27		*Cronobacter* phage vB_CsaM_leB	AOG16366.1
28		*Cronobacter* phage vB_CsaM_leN	AOG16651.1
29		*Erwinia* phage Cronus	AWD90329.1
30		*Escherichia* phage Lw1	YP_008060759.1
31		*Klebsiella* phage Marfa	QDB71837.1
32		*Proteus* phage phiP4-3	AUM58455.1
33		*Proteus* phage vB_PmiM_Pm5461	YP_009195583.1
34		*Pseudomonas* phage PspYZU05	ASD52094.1
35		*Klebsiella* phage vB_Kpn_F48	AUO78857.1
36		*Morganella* phage vB_MmoM_MP1	YP_009280031.1
37		*Pectobacterium* bacteriophage PM2	YP_009211620.1
38		*Serratia* phage PS2	AHY25432.1
39		*Vibrio* phage vB_VmeM-32	ALY07226.1.
40	*Moonvirus*	*Citrobacter* phage Merlin	YP_009203914.1
41		*Citrobacter* phage Moon	YP_009146629.1
42	*Karamvirus*	*Cronobacter* phage Pet-CM3-4	SCN45872.1
43		*Enterobacter* phage PG7	YP_009005458.1
44		*Enterobacter* phage myPSH1140	AVR55365.1
45		*Enterobacteria* phage CC31	YP_004010038.1
46	*Slopekvirus*	*Enterobacter* phage phiEap-3	YP_009607157.1
47		*Klebsiella* phage Matisse	YP_009194487.1
48		*Klebsiella* phage Miro	YP_009607435.1

49		*Enterobacteria* phage HX01	*
50		*Enterobacteria* phage RB69	NP_861881.1
51		*Enterobacteria* phage ATK47	ANZ51023.1
52		*Enterobacteria* phage ATK48	ANZ51366.1
53		*Escherichia coli* O157 typing phage 6	YP_009593176.1
54		*Escherichia coli* O157 typing phage 3	YP_009592780.1
55		*Escherichia* phage APCEc01	YP_009225077.1
56		*Escherichia* phage HP3	*
57		*Escherichia* phage OLB35	AYR04095.1
58		*Escherichia* phage SF	AWY07818.1
59		*Escherichia* phage ST0	YP_009608487.1
60		*Escherichia* phage p000v	AYN56237.1
61		*Escherichia* phage p000y	AYN56688.1
62		*Escherichia* phage vB_EcoM_G2285	QBO62608.1
63		*Escherichia* phage vB_EcoM_G2469	QBO62878.1
64	*Mosigvirus*	*Escherichia* phage vB_EcoM_G53	QBO65324.1
65		*Escherichia* phage vB_EcoM_JS09	YP_009037590.1
66		*Escherichia* phage vB_EcoM_KAW3E185	QBQ78739.1
67		*Escherichia* phage vB_EcoM_MM02	QBQ79237.1
68		*Escherichia* phage vB_EcoM_WFK	QBQ77198.1
69		*Escherichia* phage vB_EcoM_WFL6982	QBQ76937.1
70		*Escherichia* phage vB_EcoM_WFbE185	QBQ77690.1
71		*Escherichia* phage vB_EcoM_PhAPEC2	YP_009056764.1
72		*Escherichia* virus vB_Eco_mar005P1	VCU44444.1
73		*Escherichia* virus vB_Eco_mar005P1 strain vB_Eco_mar006P2	VCU44449.1
74		Escherichia virus vB_Eco_mar005P1 strain vB_Eco_mar007P3	VCU43253.1
75		*Escherichia*_virus_vB_Eco_mar005P1_strain_vB_Eco_mar008P4	VCU43961.1
76		*Escherichia*_virus_vB_Eco_mar005P1_strain_vB_Eco_mar009P5	VCU44820.1
77		*Shigella* phage SHSML-52-1	YP_009289105.1
78		*Enterobacteria* phage IME08	YP_003734320.1
79		*Enterobacteria* phage JS10	YP_002922522.1
80		*Escherichia* phage JS98	YP_001595305.1
81		*Escherichia* phage QL01	YP_009202912.1
82	*Dhakavirus*	*Enterobacteria* phage vB_EcoM_VR5	YP_009205871.1
83		*Escherichia* phage AnYang	QAU03652.1
84		*Escherichia* phage EcWhh-1	QAX99933.1
85		*Escherichia* phage MX01	YP_009324073.1
86		*Escherichia* phage WG01	YP_009323380.1
87		Phage NC-G	QBP35520.1
88		*Escherichia* phage vB_EcoM_VR7	YP_004063874.1
89	*Gaprivervirus*	*Escherichia* phage vB_EcoM_VR20	YP_009207369.1
90		*Escherichia* phage vB_EcoM_VR25	YP_009209935.1
91		*Escherichia* phage vB_EcoM_VR26	YP_009214027.1
92		*Shigella* phage SP18	YP_003934814.1
93		Enterobacteria phage Aplg8	ANZ50774.1
94		*Enterobacteria* phage Bp7	YP_007004267.1
95		*Enterobacteria* phage GiZh	*
96		*Escherichia* phage AR1	YP_009167991.1
97	*Tequatrovirus*	*Enterobacteria* phage Kha5h	ANZ51799.1
98		*Enterobacteria* phage RB10	AIT74268.1
99		*Escherichia* virus RB14	YP_002854510.1
100		*Escherichia* virus RB32	YP_803117.1
101		*Enterobacteria* phage RB18	AXF42485.1
102		*Enterobacteria* phage RB27	YP_009102379.1

103		*Enterobacteria* phage RB33	AIT74814.1
104		*Escherichia* phage RB3	YP_009098560.1
105		*Enterobacteria* phage RB51	ACP31095.1
106		*Enterobacteria* phage RB55	AIT75086.1
107		*Enterobacteria* phage RB59	AIT75360.1
108		*Enterobacteria* phage RB5	AIT73181.1
109		*Enterobacteria* phage RB68	AIT75637.1
110		*Enterobacteria* phage RB6	AIT73452.1
111		*Enterobacteria* phage RB7	AIT73723.1
112		*Enterobacteria* phage RB9	AIT73995.1
113		*Enterobacteria* phage T4T	ADJ39898.1
114		*Escherichia* virus T4	NP_049790.1
115		*Enterobacteria* phage T4 strain wild	*
116		*Enterobacteria* phage T6	AXN58211.1
117		*Escherichia* phage vB_EcoM_ACG-C40	YP_006986729.1
118		*Enterobacteria* phage vB_EcoM_IME339	AWD91505.1
119		*Enterobacteria* phage vB_EcoM_IME340	AWD91761.1
120		*Escherichia* phage CF2	*
121		*Escherichia* phage D5505	QBO61184.1
122		*Escherichia* phage ECML-134	YP_009102648.1
123		*Escherichia* phage HY01	YP_009148617.1
124		*Escherichia* phage HY03	YP_009284038.1
125		*Escherichia* phage HY03	YP_009284039.1
126		*Escherichia* phage KIT03	BBG28692.1
127		*Escherichia* phage PE37	ANH49691.1
128		*Escherichia* phage PP01	BBC14517.1
129		*Escherichia* phage T2	BBC14796.1
130		*Escherichia* phage UFV-AREG1	YP_009281510.1
131		*Escherichia* phage vB_EcoM_112	YP_009030783.1
132		*Escherichia* phage slur02	YP_009210095.1
133		*Escherichia* phage slur03	YP_009625074.1
134		*Escherichia* phage slur04	YP_009625392.1
135		*Escherichia* phage slur07	YP_009197280.1
136		*Escherichia* phage slur08	CUL02618.1
137		*Escherichia* phage slur11	CUL02946.1
138		*Escherichia* phage slur13	CUL03732.1
139		*Escherichia* phage slur14	YP_009180676.1
140		*Escherichia* phage vB_EcoM-G28	AVH86005.1
141		*Escherichia* phage vB_EcoM-Sa45lw	QDF15200.1
142		*Escherichia* phage vB_EcoM-UFV13	YP_009290443.1
143		*Escherichia* phage vB_EcoM-fFiEco06	AUV61038.1
144		*Escherichia* phage vB_EcoM-fHoEco02	AUV61311.1
145		*Escherichia* phage vB_EcoM_DalCa	AYP69772.1
146		*Escherichia* phage vB_EcoM_G10400	QBO63692.1
147		*Escherichia* phage vB_EcoM_G2540-3	QBO65583.1
148		*Escherichia* phage vB_EcoM_G2540	QBO63414.1
149		*Escherichia* phage vB_EcoM_G29	QBO64498.1
150		*Escherichia* phage vB_EcoM_G4498	QBO64219.1
151		*Escherichia* phage vB_EcoM_G4500	QBO65870.1
152		*Escherichia* phage vB_EcoM_G4507	QBO66145.1
153		*Escherichia* phage vB_EcoM_G50	QBO65046.1
154		*Escherichia* phage vB_EcoM_G8	QBQ80057.1
155		*Escherichia* phage vB_EcoM_G9062	QBQ77959.1
156		*Escherichia* phage vB_EcoM_JB75	AXC34018.1
157		*Escherichia* phage vB_EcoM_KAW1E185	QBQ78462.1
158		*Escherichia* phage vB_EcoM_OE5505	QBQ79505.1
159		*Escherichia* phage vB_EcoM_R5505	QBQ79789.1
160		*Escherichia* phage wV7	YP_007004917.1
161		*Escherichia* virus VEc20	QDK04303.1
162		*Shigella* phage SH7	APC45045.1

163		*Shigella* phage SHBML-50-1	YP_009288541.1
164		*Shigella* phage SHFML-11	YP_009277549.1
165		*Shigella* phage Sf21	YP_009618985.1
166		*Shigella* phage Sf22	YP_009614828.1
167		*Shigella* phage Sf23	ATE86434.1
168		*Shigella* phage Sf24	YP_009619109.1
169		*Shigella* phage Shfl2	YP_004415069.1
170		*Shigella* phage pSs-1	YP_009110995.1
171		*Yersinia* phage PST	YP_009153775.1
172		*Yersinia* phage fPS-2	VEV89788.1
173		*Yersinia* phage phiD1	YP_009149419.1
174	*Jiaodavirus*	*Klebsiella* phage KP1	AUV57527.1
175		*Klebsiella* phage KPV15	APD20579.1
176		*Klebsiella* phage Mineola	AWY07077.1
177		*Klebsiella* phage PKO111	YP_009289602.1
178		*Klebsiella* phage vB_KpnM_KpV477	YP_009288865.1
179	*Gelderlandvirus*	*Salmonella* phage Melville	YP_009615654.1
180		*Salmonella* phage vB_SenMS16	YP_007501209.1
181		*Salmonella* phage STML-198	YP_009148159.1
182		*Salmonella* phage STP4-a	YP_009126374.1
183		*Salmonella* phage vB_SnwM_CGG4-1	YP_009286529.1
184	*Schizotequatrovirus*	*Vibrio* phage KVP40	NP_899381.1
185		*Vibrio* phage ValB1 HC	QBX05956.1
186		*Vibrio* phage ValKK3	YP_009201230.1
187		*Vibrio* phage phi-Grn1	ALP47010.1
188		*Vibrio* phage phi-ST2	ALP47390.1
189		*Vibrio* phage phi-pp2	AFN37364.1

Figure 1 shows the phylogenetic tree of RNA ligase 2 of the subfamily of T-even bacteriophages constructed using the Maximum Likelihood Method using the model of JTT substitution matrices [26]. A consensus tree was obtained after 1000 iterations of bootstrap analysis undertaken to elucidate the evolution of the analyzed taxa [27]. Branches formed in less than 50% of the repetitions of the statistical analysis were removed. The source tree was obtained using the Maximum Parsimony Method. All positions containing gaps or places where data were missing were removed. The study was performed using the MEGAX software package [28].

On the tree presented in Figure 1, the terminal vertices corresponding to the classified sequences of the subfamily are marked with circles painted with the colors corresponding to their genera. From the Figure, it becomes apparent that the terminal vertices of the same color lie on a tree side by side.

The analysis showed the following. RNA ligases 2 are proteins with a conservative primary structure. The amino acid sequence of RNA ligase 2 of the bacteriophage T4 is a representative sample of the enzyme under examination, and this, in turn, allows an investigator not only to find the reliable orthologues in other genomes, but also to use it for classification.

Preparation of a set of RnlB homolog proteins of bacteriophage T4 from metagenomes of ocean samples and genomes of phylogenetically distant living entities

At the first stage PSI-BLAST algorithm [25] with the reliability level of the *E-value* < $3e^{-21}$ results was used to search for homologs of RNA ligase 2 of the bacteriophage T4 (*rnlB* gene product GenBank ID ADJ39898.1), in the database of protein sequences of oceanic metagenomes on the NCBI server. To obtain a set of amino acid sequences of homologs of RNA ligase 2 of bacteriophage T4, an iterative search for similar sequences was performed until each subsequent iteration revealed no new local similarities, that is, 5 times. Fifteen homologous sequences were found in GOS metagenomes and 21 in LCGC14.

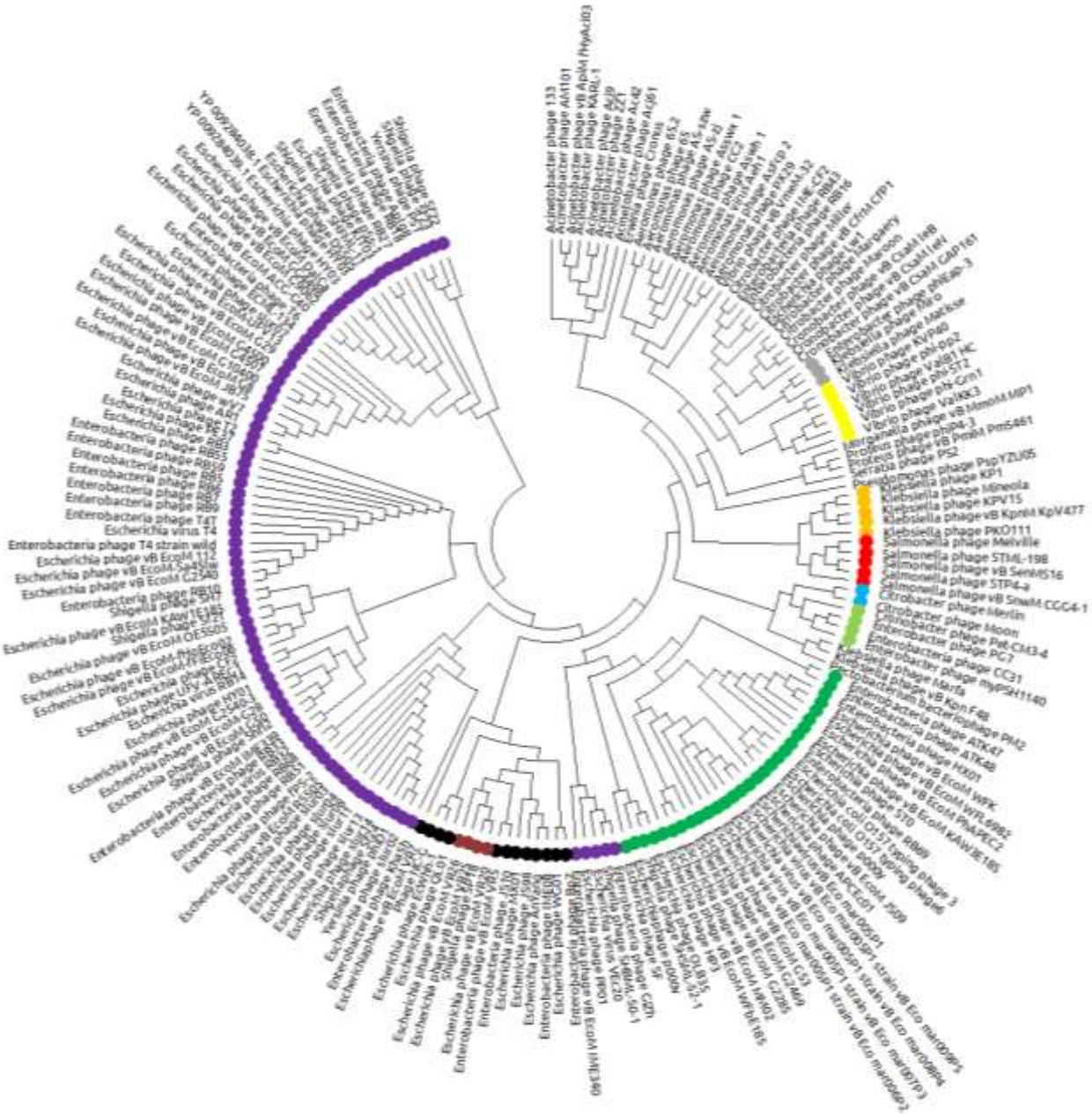

Fig. 1. Phylogenetic tree constructed for 189 RNA ligases 2 subfamilies of T-even bacteriophages.

At the second stage, the data was subjected to additional filtering. Sequences whose length differed from RNA ligase 2 of bacteriophage T4 by more than 20% were removed from the set. This is primarily because up to 70% of the total number of sequences in the metagenome of any sample are short fragments of sequences. The statistical limitations that exist in programs of any phylogenetic analysis package do not allow working with short sequences. The length of the RNA ligase 2 of the bacteriophage T4 is 334 amino acids. As a result of the editing, 21 protein sequences with a length range from 271 to 389 amino acids remained in the data set.

This situation is quite adequate for the genetic sets obtained by metagenomics. In a number of other cases, found amino acid sequences were encoded at the ends of fragments or contigs of a particular metagenome. The process of sample preparation for the metagenomic sequencing of marine microbiota involves several stages of filtering the water or extract of marine benthos. All of them it may affect the intactness of living material. Even at this stage,

the breakdowns occur in various points of the long genomic DNA. Further processing of the DNA extracted and purified by ultrasound creates a lot of additional breaks. Sequencing a wide variety of genomes gives long overlaps suitable for reading. The most significant point is the wide variety of living creatures in the ocean, so the accurate determination of all sequences from one or another organism in readings is difficult.

Our main purpose was to determine location of sequences found in oceanic metagenomes on a phylogenetic tree constructed on the base of the amino acid sequence of RNA ligase 2. So we had to create the most representative taxonomic series of sequences of phage T4 RNA ligase 2 homologs available in the databases. In the cases when all the organisms are screened as a single massive, the abundance of RnlB sequences in bacteriophages and other viruses with a high level of homology masks the results from other taxa with less similarity, especially the simplest ones. Therefore, we decided to make searches for individual taxa.

One could expect the presence of homologs in taxa of protozoa, but with a lower significance level. Each and every control sequence was crucial for determination the taxonomic affiliation of a find from marine metagenomes, so it was important not to miss a single sequence, and was done by painstaking analysis conducted on large taxa of unicellular and multicellular living organisms. It was necessary not to miss any control sequence for determination the taxonomy of marine metagenome homologues. It was done by painstaking analysis conducted on large taxa of unicellular and multicellular living organisms. We carried out a study of the following taxa: *Cercozoa, Ciliophora, Euglenozoa, Amoebozoa, Homidia, Crustacea, Alveolata, Haplosporea, Acantharia, Parabasalia, Mycetozoa, Dinophyta, Fungi, Placozoa, Viridiplantae, Rhodophyta, Rhizaria, Stramenopiles* and *Porifera*, trying to cover the whole variety of genetically studied cellular life forms on the Earth. Hereinafter, we use the classification of living entities of T. Cavallier-Smith [29]. The name *Giant viruses*, includes the giant viruses of the Eukaryotes.

Table 2. Abundance of the RNA-ligase 2 homologs among the protein sequences of the biological entities from various taxa

Taxon name	Number of hits within a significant area	$E\text{-}value_{min}$	$E\text{-}value_{max}$
Cercozoa	0	0.48	
Ciliophora	1		2e-22
Euglenozoa	42	7e-11	
Amoebozoa	4		1e-04
Homidia	0		
Crustacea	0		
Alveolata	2		9e-10
Haplosporea	0	0.47	
Acantharia	0	3.3	
Parabasalia	0	4.8	
Mycetozoa	0	0.78	
Fungi	13		1e-05
Placozoa	0	0.25	
Viridiplantae	1		8e-38
Rhodophyta	0	2.2	
Rhizaria	2		0.003
Stramenopiles	2		0.002
Porifera	1		1e-07
Giant viruses	17	4e-9	

We considered not only unicellular representatives of prokaryotic and eukaryotic microbiota, both of the sea and land, because the pelagic waters contain numerous gametes of

the most diverse multicellular eukaryotes. The concentration of gametes in plankton can reach 10^7 or more particles per 1 ml. Sedimentary rocks capture this significant part of plankton in the course of their formation. Most gametes are quite large, but the products of their destruction can contain DNA of the most diverse representatives of marine biota.

The names of the taxa are given according to the taxonomy currently adopted by the National Center for Biotechnology Information (NCBI, NIH, USA) [23]. The second column shows the number of hits that are in the reliable comparison area, according to the specified comparison parameters (E-value <0.001). The third column shows the value of $E\text{-}value_{\min}$ for the best result from the unreliable comparison area, if any. The fourth column shows $E\text{-}value_{\max}$, the worst values among the results from the significant comparison area. Another information in the insignificant comparison area is not provided.

The results of search for control amino acid sequences are presented in Table 2. Against the background of 42 significant results in *Euglenozoa*, a few homologous sequences were found in *Ciliophora*, *Amoebozoa*, *Alveolata*, *Dinophyta*. Thirteen homologs of RNA ligase 2 of phage T4 were found in fungi in a significant comparison region. Homologs were also reliably found in multicellular organisms: in plants and sponges – one for each, and in *Rhizaria* and *Stramenopiles* – two for each. Some results, most likely, are not only RNA ligase sequences, but those found in marine metagenomes, using the sequence comparison parameters used, can also be similar, but not homologous. For example, for *Rhodophyta*, the best insignificant result was E3 ubiquitin HERC3 ligase protein. The *E-value* for this sequence was 2.2, the coverage was 17%, and the percentage of identical amino acids in pair alignment was 31.67%. In *Haplosporea*, it was actin of *Bonamia ostreae* with a coverage of 7%, *E-value* 0.47, identity 38.46%. Interestingly, *Acantharia*, which was close enough to the class of protozoa, with which the taxonomic group *Haplosporea* was previously combined, the best insignificant result was the finding of contractile protein, *Astrolonche serrata* tubulin with a low coating of 10%, a poor *E-value* of 3.3, and a very high percentage of identical amino acids 40.91% at pair alignment. The low coverage of these two proteins can be explained by their large length. Moreover, sequencing and artificial protein fusion errors cannot be ruled out. We managed to find homologs for these groups of unicellular protozoa. It can be expected that in the near future, more genetic information will appear about these groups and it will be possible to conduct a more reliable search for homologs of RNA ligase 2 phage T4. In *Glaucophyta*, the phytochrome of *Glaucocystis nostochinearum* was found (coating 26%, *E-value* 0.23, identity 29.29%). For a number of taxonomic groups, these were hypothetical proteins, for example, G3 in *Trichomonas vaginalis* in protozoa of the *Parabasalia* type, with a coverage of 39%, insignificant *E-value* 4.8, and a high identity of 23.24%. A hypothetical protein of *Heterostelium album* PN500 was found in *Mycetozoa* with insignificant, but very interesting statistical parameters: 18%, 0.78, and 30.77%. A similar situation was with the *Trichoplax adhaerens* (*Placozoa)* protein: 32%, 0.25 and 26.27%. It can be expected that the process of RNA editing by insertion or deletion of uridines is quite widespread in unicellular eukaryotes and occurs in multicellular protozoa in minor cases as a biochemical “atavism”. For phylogenetic analysis, all sequences from those taxa, where the results were less than 10, were selected. When the number of the cases exceeded 10, sequences with different levels of similarity were selected in accordance with the found *E-values*, evenly distributed in the hit list. Sequences from 17 bacteriophages of the *Tevenvirinae* subfamily were also taken to ensure that the closest sequences were visible in a taxonomic comparative analysis of sequences from metagenomes of oceanic microbiota and sequences from various fila and taxa of cellular forms of life. Using this approach, we tried to provide the most representative sample of various sequences from all taxa for phylogenetic analysis and tree construction. Only a single similar sequence of the representative of the taxon *Viridiplantae* was studied separately.

The control sequences and sequences of the oceanic homologs of RNA ligase 2 were combined into one set, which is shown in Table 3. For analysis, 122 homologs presented in

GenBank (Release 234, 10/05/2019) were used, including the T4 bacteriophage RNA ligase itself, as well as 21 homologs from ocean samples. To construct multiple sequence alignment, the MUSCLE algorithm [30] was used. Based on the obtained alignment, an evolutionary history was deduced, and a phylogenetic tree was constructed using the Maximum Likelihood Method with the model of JTT replacement matrices [26]. The consensus tree was constructed after 3000 repetitions of the statistical bootstrap analysis [27] undertaken to elucidate the evolution of the analyzed taxa. Branches formed in less than 50% of the repetitions of the statistical analysis were removed. The source tree was obtained using the Maximum Parsimony Method. All positions containing gaps or places where data were missing were removed. The study was performed using the MegaX software package [28].

Table 3. Homologs of RNA ligase 2 of the bacteriophage T4 from the genomes of living entities of different taxa, which were used for phylogenetic analysis and location of homologues from oceanic metagenomes on the tree of life

№	Taxon	Protein name	GenBank IDs
1	▲ *Bacteria*	Cand *Thiosymbion oneisti*	WP_089724394.1
2		*Chryseolinea serpens*	SHH97071.1
3		Bacteroidetes bacterium 4484_276	OQX79431.1
4		*Chryseolinea serpens*_II	WP_073143442.1
5		*Desulfamplus magnetovallimortis*	SLM31278.1
6		Cand *Gottesmanbacteria bacterium*	KKW10401.1
7		Cand *Vecturithrix granuli*	GAK56201.1
8		Cand *Propionivibrio aalborgensis*	SBT10709.1
9		Cand *Contendobacter odensis*	WP_051497828.1
10		*Parabacteroides distasonis*	WP_081033086.1
11		*Porphyromonas* sp. 31_2	KEJ87078.1
12		Cand *Contendobacter odensis* Run_B_J11	CDH45942.1
13		*Bacteroides* sp. CAG:633	CDB10874.1
13	● *Stramenopiles*	FNF27_00344 *Cafeteria roenbergensis*	KAA0178495.1
15		FNF31_06539 *Cafeteria roenbergensis*	KAA0153076.1
16		FNF29_01479 *Cafeteria roenbergensis*	KAA0156063.1
17		*Thraustotheca clavata*	OQR97834.1
18		57867_21281 *Aphanomyces stellatus*	VFT97953.1
19	● *Rhizaria*	RNA ed ligase 2 *Reticulomyxa filosa*	ETO29077.1
20		RNA ed lig 1 *Reticulomyxa filosa*	ETO09630.1
21	● *Alveolata*	PPERSA_08025_1 *Pseudocohnilembus persalinus*	KRX08714.1
22		RNA ed ligase 2 mt *Symbiodinium microadriaticum*	OLP91488.1
23	● *Amoebozoa*	RNA lig put *Acanthamoeba castellanii* Neff	XP_004341081.1
24		RNA ligase *Acanthamoeba castellanii* Neff	XP_004333230.1
25		RNA ed lig *Acanthamoeba castellanii* Neff	XP_004349586.1
26		EIN_146330 *Entamoeba invadens* IP1	XP_004254400.1
27	▼ *Archaea*	DRN27_08370 *Thermoplasmata archaeon*	RLF57127.1
28		hp *Thermoplasmata archaeon*	MAH42159.1
29		DRN58_06670 *Thermococci archaeon*	RLF98605.1
30		DRN26_00060 *Thermoplasmata archaeon*	RLF68204.1
31	● *Ciliophora*	PPERSA_08025 *Pseudocohnilembus persalinus*	KRX08714.1
32	● *Euglenozoa*	mt RNA lig2 *Leishmania mexicana* MHOM	XP_003875227.1
33		mt RNA lig2 *Leishmania major* str. Friedlin	XP_001682919.1
34		mt RNA lig2 *Leishmania donovani*	XP_003860510.1
35		mt RNA lig2 *Leishmania tarentolae*	AAN77726.1
36		mt RNA lig2 *Leishmania infantum* JPCM5	XP_001465293.1
37		mt RNA lig2 *Leishmania braziliensis* MHOM	XP_003723137.1
38		mt RNA lig2 *Leishmania panamensis*	XP_010698781.1
39		mt RNA lig2 *Leishmania guyanensis*	CCM15306.1
40		mt RNA lig2 *Trypanosoma congolense* IL3000	CCC92745.1

41		mt RNA lig2 *Leptomonas pyrrhocoris*	XP_015653633.1
42		RNA ed lig *Angomonas deanei*	EPY40852.1
43		mt RNA lig1 *Leishmania major* strain Friedlin	XP_003721581.1
44		RNA lig2 *Trypanosoma theileri*	XP_028882973.1
45		mt RNA lig1 *Leishmania major*	AAR10824.1
46		RNA ed lig *Trypanosoma vivax* Y486	CCC46487.1
47		hp *Phytomonas* sp. isolate EM1	CCW64502.1
48		RNA ed lig *Trypanosoma congolense* IL3000	CCC89377.1
49		mt RNA lig2 *Trypanosoma grayi*	XP_009308197.1
50		mt RNA lig *Trypanosoma brucei gambiense* DAL972	XP_011776426.1
51		*Euglenozoa* RNA ed lig *Trypanosoma conorhini*	XP_029226606.1
52		mt RNA lig1 *Trypanosoma conorhini*	XP_029226428.1
53		RNA ed lig *Trypanosoma rangeli*	XP_029240616.1
54		RNA ed lig *Trypanosoma vivax* Y486	CCC50326.1
55		mt RNA lig1 *Trypanosoma rangeli* SC58	ESL07901.1
56		mt RNA lig1 *Trypanosoma rangeli*	XP_029237953.1
57		KREL2 *Trypanosoma equiperdum*	SCU71059.1
58		mt RNA lig1 *Trypanosoma theileri*	XP_028885257.1
59		mt RNA lig1 *Leishmania panamensis*	XP_010696701.1
60		hp *Phytomonas* sp. isolate EM1	CCW64180.1
61		mt KREL1 *Leptomonas pyrrhocoris*	XP_015652175.1
62		RNA-editing ligase *Angomonas deanei*	EPY33068.1
63		mt RNA lig1 *Trypanosoma grayi*	XP_009311548.1
64		mt RNA lig1 *Leptomonas seymouri*	KPI83474.1
65		RNA ed lig *Bodo saltans*	CUE61637.1
66		REL1 *Trypanosoma cruzi*	RNF20713.1
67		RNA ed lig 1 *Leptomonas seymouri*	KPI86142.1
68		mt RNA lig1 *Trypanosoma cruzi* strain CL Brener	XP_820361.1
69		hp *Phytomonas* sp. isolate Hart1	CCW66362.1
70		RNA ed lig *Bodo saltans*	CUG93424.1
71		RNA ed lig *Perkinsela* sp. CCAP 1560/4	KNH03941.1
72		RNA ed lig2 *Perkinsela* sp. CCAP 1560/4	KNH03941.1
73		RNA ed lig *Angomonas deanei*	EPY29391.1
74		BC937DRAFT_89812 *Endogone* sp. FLAS-F59071	RUS17558.1
75		C2G38_128065 *Gigaspora rosea*	RIB11192.1
76		C1645_817038 *Glomus cerebriforme*	RIA95193.1
77		C1645_795269 *Glomus cerebriforme*	RIA78863.1
78		RclHR1_02510016 *Rhizophagus clarus*	GBB95322.1
79		C2G38_1022383 *Gigaspora rosea*	RIB01290.1
80	● *Fungi*	BC938DRAF_476924 *Jimgerdemannia flammicorona*	RUS31849.1
81		BC936DRAF_147411 *Jimgerdemannia flammicorona*	RUP46053.1
82		DNA lig/mRNA capp *Rhizophagus irregularis*	PKK73150.1
83		GLOIN_2v1881488 *Rhizophagus irregularis* DAOM	XP_025171120.1
84		C1646_745321 *Rhizophagus diaphanus*	RGB29488.1
85		BC938DRAF_470766 *Jimgerdemannia flammicorona*	RUS26442.1
86		CcCBS67573_g06065 *Chytriomyces confervae*	TPX71677.1
87	● *Porifera*	RNA ed lig 2 mtl *Amphimedon queenslandica*	XP_011410182.1
88		RnlB *Escherichia* virus T4	2HVR_A Chain A
89		RnlB *Staphylococcus* phage	ATN93943.1
90		RnlB *Escherichia* phage ECML-134	YP_009102648.1
91		RnlB *Escherichia* phage slur14	YP_009180676.1
92		RnlB *Enterobacteria* phage ATK47	ANZ51023.1
93		RnlB *Escherichia* phage APCEc01	YP_009225077.1
94	◆ *Viruses*	RnlB *Enterobacteria* phage JS10	YP_002922522.1
95		RnlB *Edwardsiella* phage PEi26	BAQ23148.1
96		RnlB *Citrobacter* phage Merlin	YP_009203914.1
97		RnlB *Klebsiella* phage Marfa	QDB71837.1
98		RnlB *Erwinia* phage Cronus	AWD90329.1
99		RnlB *Citrobacter* phage Miller	YP_009097842.1
100		RnlB *Aeromonas* phage Ah1	AUE22774.1

101		RnlB *Vibriophage* phi-pp2	AFN37364.1
102		RNA ligase 2 *Vibrio* phage CHOED	YP_009021711.1
103		RNA ligase 2 *Acinetobacter* phage TAC1	AZF88522.1
104		RNA ligase 2 *Streptomyces* phage Annadreamy	AXG66278.1
105		RNA ligase 2 Tupanvirus deep ocean	AUL79719.1
106		RNA ligase 2 Pithovirus LCPAC401	QBK92682.1
107		RNA ligase 2 Mimivirus sp. SH	AZL89535.1
108		RNA ligase 2 Powai lake megavirus	ANB50426.1
109		RNA ligase 2 Megavirus vitis	AVL93643.1
110		RNA ligase 2 Megavirus chiliensis	YP_004894359.1
111		RNA ligase 2 Bandra megavirus	AUV58249.1
112		RNA ligase 2 Saudi moumouvirus	AQN68128.1
113		RNA ligase 2 Marseillevirus LCMAC101	QBK85972.1
114		RNA ligase 2 Acanthamoeba polyphaga moumouvirus	YP_007354215.1
115		RNA ligase 2 Moumouvirus Monve	AEX63037.1
116		RNA ligase 2 Moumouvirus australiensis	AVL94635.1
117		RNA ligase 2 Harvfovirus sp.	AYV80481.1
118		RNA ligase 2 Moumouvirus goulette	AGF85515.1
119		RNA ligase 2 Tupanvirus2 deep ocean	AUL79055.1
120		RNA ligase 2 Tupanvirus soda lake	AUL77767.1
121		RNA ligase 2 Powai lake megavirus	ANB50967.1
122		RNA ligase 2 *Acanthamoeba polyphaga* mimivirus	AVG46618.1
123		GOS_3224571	ECD49735.1
124	● Homologs from the oceanic planktonic microbiota	GOS_6108359	ECF47459.1
125		GOS_8459675	EBM12135.1
126		GOS_1057497	EDE85953.1
127		GOS_4814015	EBY11901.1
128		GAG86666.1	GAG86666.1
129		GAJ13060.1MSM	GAJ13060.1
130		RNA ligaseMSM	EFK97083.1
131		LCGC14_1419600MSM	KKM72535.1
132		LCGC14_1490430MSM	KKM65524.1
133		LCGC14_0465520MSM	KKN66998.1
134	● Homologs from the microbiota of the oceanic sediments	LCGC14_0236200MSM	KKN89774.1
135		LCGC14_1657970MSM	KKM19213.1
136		LCGC14_2451880MSM	KKL20798.1
137		LCGC14_2597070MSM	KKL06334.1
138		LCGC14_2115790MSM	KKL69358.1
139		LCGC14_0535600MSM	KKN60091.1
140		LCGC14_3030500MSM	KKK59824.1
141		LCGC14_2458620	KKL20123.1
142		LCGC14_1232570MSM	KKM91033.1
143		LCGC14_1282820MSM	KKM86052.1

RESULTS

Phylogenetic analysis of RNA ligases 2 from metagenomes of marine microbiota and living entities of various taxa

A representative set of RNA ligase 2 homologs from various taxa of living entities, the genomic data of which are available to researchers to date, was used to construct a phylogenetic tree, presented in Figure 2.

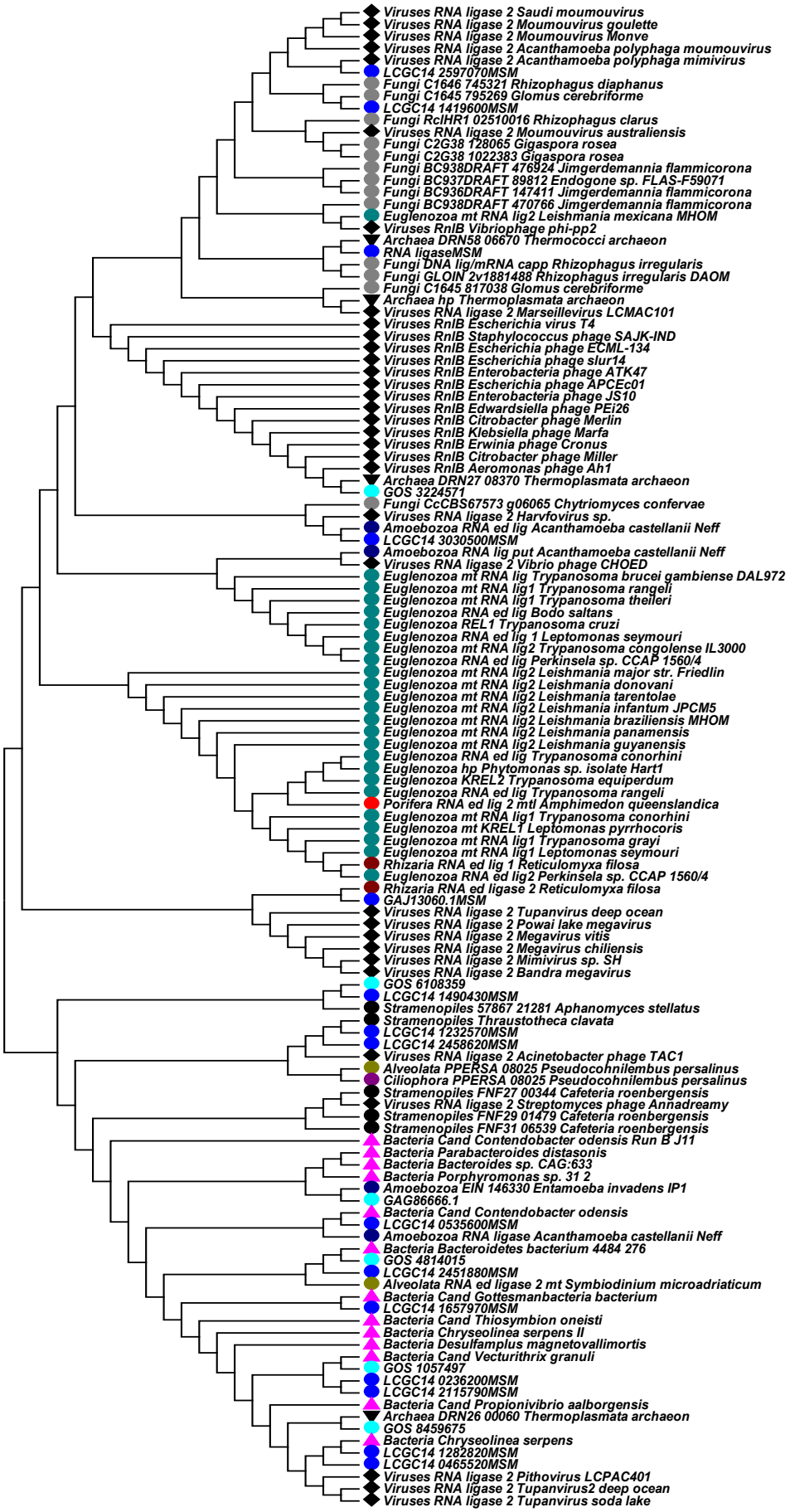

Fig. 2. Molecular phylogenetic analysis of homologs of RNA ligase 2 of bacteriophage T4 from different taxa of living entities.

The phylogenetic tree of RNA ligase 2 homologs, shown in Figure 2, forms two branches, and on its upper branch there are virus proteins, including phages like T4, trypanosome RNA ligases highly homologous to them, fungal proteins, all homologs from *Rhizaria*, *Porifera*, two archaea sequences, and two sequences from the amoeba. On this branch there is only one sequence of planktonic homolog, three deep-water sequences and one of the other aquatic sedimentary rocks. Most sequences of oceanic homologs are represented in another main branch of this tree.

The main sequences presented on the lower branch are proteins of bacteria, a number of bacteriophages and megaviruses, as well as archaea, fungi and some other eukaryotes. All 14 sequences of metagenomes of the ocean floor are also located on this branch. Their distribution on this branch is uneven; there are no associations in large common treasures. The most closely phylogenetically related representatives of marine homologs of T4 bacteriophage RNA ligase are presented in Table 3.

The lower branch of the tree contains three sequences of giant viruses, one bacterial sequence and two homologs from deep-sea ocean sediments. One of the *Tupanovirus* viruses presented in this clade was also isolated from an ocean depth of more than 3,000 m, as were samples for deep-water metagenomic analysis. It should be noted that, as part of this branch, planktonic and deep-water homologs of RNA ligase 2 are often clustered together on the same branches.

Table 3 shows that the most part of the found sequences show a phylogenetic relationship with sequences from bacteria, megaviruses, fungi and archaea. In many cases, the most closely related sequences belonged to bacteria known only from the data of oceanic metagenomes and the determination of which at the present time is not complete. Two sequences were related to the Protozoa of the *Amoebozoa* taxon, one to *Rhizaria* and one to *Stramenopiles*.

A comparative analysis by pairwise alignments (data not shown) showed that the found similarity could tell only about the affinity of these sequences, but not about their coincidence. Thus, the found homologues belong to new species of organisms from these groups.

A position not far from the RNA ligase 2 of *Trypanosoma* is occupied by the sequences of the sponge from the large barrier reef *Amphimedon queenslandica* (*Porifera*) and the unicellular freshwater foraminifera *Reticulomyxa filosa* (*Rhizaria*) belonging to the monospecific genus. In both cases, the origin of this gene in these genomes due to horizontal gene transfer seems most probable. On the other hand, in the unicellular animals, which are currently referred to as *Rhizaria*, due to their little knowledge, the classification and possible common origin of mitochondrial DNA with mitochondrial DNA *Euglenozoa* cannot be ruled out.

Among *Euglenozoa*, homologs of RNA ligase 2 can be found only in trypanosome mitochondria (Fig. 2) and are divided into two related branches. In the upper part of the tree, RNA ligases 2 of the first type are more often represented, while in the lower part the ligases of the second type prevail. Both enzymes are closely related, and are likely to evolve, replacing each other functionally. One of these enzymes is included in the protein complex editing the primary transcripts of trypanosome mitochondria because it deletes uridine, the other – because it inserts this nucleotide. The functions of both RNA ligases are the same in both cases – ligation of a single-stranded break in the RNA that remains after the work of other enzymes.

Viral homologs of RNA ligase 2 are found in various branches of the tree. This is probably due to several reasons. First, these enzymes could initially appear just in the viruses and disperse evolutionally within a variety of viral genomes; later they could be transmitted independently to various cellular life forms. Another explanation is that the nature conducted its experiments using the viruses as vectors of the horizontal gene transfer. These experiments

could lead to the secondary appearance of the genes originating from the genome of one or another cellular life form in viruses.

Table 4. Sequences phylogenetically most closely related to marine homologs

Ocean Metagenome Sequence Name	Phylogenetically closest sequences	
	1*	2**
LCGC14 1419600MSM	Fungi C1645 795269 *Glomus cerebriforme*	Fungi C1646 745321 *Rhizophagus diaphanus*
RNA ligaseMSM	Archaea DRN58 06670 *Thermococci archaeon*	Fungi DNA lig/mRNA capp *Rhizophagus irregularis*
GOS 3224571	Archaea DRN27 08370 *Thermoplasmata archaeon*	Viruses RnlB *Aeromonas phage* Ah1
LCGC14 3030500MSM	Amoebozoa RNA ed lig *Acanthamoeba castellanii* Neff	Viruses RNA ligase 2 *Harvfovirus sp*.
LCGC14 2458620	Stramenopiles *Thraustotheca clavata*	Viruses RNA ligase 2 *Acinetobacter phage* TAC1
LCGC14 1232570MSM	Stramenopiles *Thraustotheca clavata*	Viruses RNA ligase 2 *Acinetobacter phage* TAC1
LCGC14 2597070MSM	Amoebozoa EIN 146330 *Entamoeba invadens* IP1	Fungi BC938DRAFT 470766 *Jimgerdemannia flammicorona*
LCGC14 0535600MSM	Bacteria Cand *Contendobacter odensis*	-**
GOS 1057497	Bacteria Cand *Vecturithrix granuli*	-
LCGC14 0236200MSM LCGC14 2115790MSM	Bacteria Cand *Vecturithrix granuli*	-
LCGC14 1657970MSM	Bacteria Cand *Gottesmanbacteria bacterium*	-
GOS 8459675	Archaea DRN26 00060 *Thermoplasmata archaeon*	Bacteria *Chryseolinea serpens*
LCGC14 0465520MSM	Bacteria *Chryseolinea serpens*	Archaea DRN26 00060 *Thermoplasmata archaeon*
LCGC14 1282820MSM	Bacteria *Chryseolinea serpens*	Archaea DRN26 00060 *Thermoplasmata archaeon*
GAJ13060.1MSM	Rhizaria RNA ed ligase 2 *Reticulomyxa filosa*	Viruses RNA ligase 2 *Acanthamoeba polyphaga mimivirus*
GAG86666.1	Viruses RNA ligase 2 *Tupanvirus2 deep ocean*	Viruses RNA ligase 2 *Tupanvirus soda lake*
GOS 6108359 LCGC14 1490430MSM	Stramenopiles 57867 21281 *Aphanomyces stellatus*	-
LCGC14 2451880MSM GOS 4814015	Bacteria *Bacteroidetes bacterium* 4484 276	Alveolata RNA ed ligase 2 mt *Symbiodinium microadriaticum*

* - closest sequence on phylogenetic tree;
** - slightly more distant on the tree but also a close sequence.

The sequence LCGC14 1419600MSM is located among the fungal homologs and adjacent branches of the protein sequences of giant eukaryotic viruses.

The place next to this branch is occupied by the sequence of RNA ligase MSM from marine sediments, closely adjacent to the homologue from the *Thermococci archaeon* archaea. This branch is also located in the branch of fungal homologs. GOS 3224571 is the only sequence located in the homologous branch of T4-type bacteriophages, although immediately adjacent position is occupied by a homolog from the kingdom of archaea - *Thermoplasmata archaeon*.

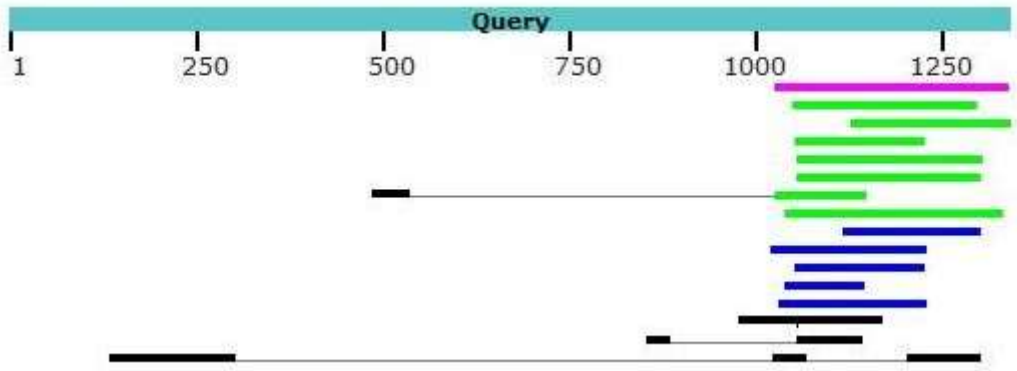

Fig. 3. A graphical representation of the multiple alignment of RnlB protein from *Vigna angularis*, RNA ligase 2 of the T4 phage (lilac line) and oceanic proteins (green, blue and black lines depict descending different levels of reliability of the *E-value* alignment, the values are given in Table 3). The sequence of oceanic proteins in the figure corresponds to the order in the table 3.

In the kingdom of green plants (*Viridiplantae*), a homolog of RNA ligase 2 of T4 phage, 1338 a.a. in length, was found in angular beans, or adzuki. As shown by multiple alignment (Fig. 3) made by the ClustalX program [31], all homologous fragments were concentrated near the C-terminus of this protein. The distribution of the sample of oceanic sequences by similarity to the C-terminus is provided in Table 4.

The oceanic homologs LCGC14_1419600MSM, GAJ13060.1MSM, LCGC14_0465520MSM, LCGC14_0236200MSM, LCGC14_2115790MSM, GOS_3224571, LCGC14_2597070MSM show a significant similarity with a fragment of this bean protein. The presence of a homolog of RNA ligase 2 from phage T4 in only one green plant and the absence of such finds in any to some degree similar species, as well as the large difference in the lengths of the complete amino acid sequences of these proteins did not allow this control to be included in the software package MegaX, and we presented this result separately.

Table 5. The result of alignment of the RnlB protein from *Viridiplantae* (*Vigna angularis*, 1338 a.a.) and T4 phage with oceanic homologs (the notation is the same as in Fig. 2)

№	Sequence name	Percentage of coverage length	*E-value*
	Vigna angularis 1338 a.к.	100%	0
	Enterobacteria phage T4	23%	3e-43
1	LCGC14_1419600MSM	18%	7e-14
2	GAJ13060.1MSM	16%	3e-13
3	LCGC14_0465520MSM	13%	1e-11
4	LCGC14_0236200MSM	18%	2e-11
5	LCGC14_2115790MSM	18%	9e-11
6	GOS_3224571	12%	5e-10
7	LCGC14_2597070MSM	21%	8e-10
8	GAG86666.1	13%	2e-09
9	LCGC14_1657970MSM	15%	3e-09
10	GOS_8459675	13%	8e-08
11	GOS_6108359	7%	1e-07
12	LCGC14_0535600MSM	14%	5e-07
13	RNA ligaseMSM	14%	1e-04
14	LCGC14_2451880MSM	8%	0.005
15	LCGC14_1490430MSM	23%	0.13

DISCUSSION

The Craig Venter scientific group searched for viral sequences in microbial metagenomes by comparing GOS sequences with 27 complete virus genomes [18]. Taxonomic analysis was carried out by the following method. Four best finds were taken from each search with an E-value <1e-10. These data were combined into a common pool of results. The analysis of the metagenomes of seawater samples collected during the Sorcerer II Global Ocean Expedition (GOS) revealed a large number of viral sequences, representing approximately 3% of the total number of predicted proteins. Phylogenetic analysis of these viral sequences revealed a large number of bacterial genes linked to viral sequences. The authors of the study suggested that the obtaining of environmentally significant genes of host bacteria by the virus is more common than previously thought. It was shown that in the microbial fractions viral sequences originating from the genomes of caudate bacteriophages predominated. The global distribution of sequences depending on the virus family was different in the study of water samples taken from different places. A comparison of fragments of viral genomes from the GOS metagenome found in such an analysis with 27 complete genomes of viruses isolated from water revealed that only one reference bacteriophage genome was presented in these samples with a high frequency. This was the P-SSM4 cyanomyovirus gene. This dominance coincided with the dominance in samples of its host, *Prochlorococcus marinus*. The new sequences found did not fully coincide with the genome sequence of the previously sequenced phage P-SSM4, but were only related to it. Based on these observations, the researchers concluded that this bacteriophage, the same as the related phages, can significantly affect the abundance, distribution, and diversity of *Prochlorococcus marinus*, the dominant component of picophytoplankton in ocean oligotrophic waters [18].

Although bacterial viruses are typically much smaller than their host bacteria, there are a number of reasons why viral sequences can be detected in the microbial fraction of seawater. Firstly, viruses whose particle size exceeds a certain filter size (usually from 0.1 to 0.22 microns) are automatically saved due to their geometry; and such viruses have been detected, in increasing numbers, thanks to efforts to take samples from the ocean [18]. However, despite the presence of viruses capable of infecting various groups of eukaryotic hosts, the vast majority of viruses in marine ecosystems are bacteriophages, the size of which is usually less than 0.2 μm [18.19].

C. Venter's group passed the samples through a filter with 0.1 μm pores. The length of particles of the bacteriophage P-SSM4 belonging to the T-even group (*Teevenvirinae*) fluctuates from 110 to over 130 nm. This group of bacteriophages can be effectively retained on 0.1 μm membrane filters (Pall Life Sciences, East Hills, NY), which were used by the authors. The fact that we managed to find only one ocean homolog close to the bacteriophages of this group suggests that both databases contain bacteriophage genomes, including related T4, that do not contain this gene, as insignificant for the bacteriophage T4 and of his relatives [2], or the amino acid sequence of this allele in marine bacteriophages varies greatly. We found a certain number of homologs taxonomically similar to giant phage enzymes, which could be well retained by several types of filtration through membrane filters with pores of 3.0 μm, 0.8 μm, and 0.1 μm [18].

An interesting finding is the detection of RNA ligase 2 homologs in lower invertebrates, namely, representatives of the taxa *Alveolata*, *Stramenopiles*, *Rhizaria*, *Ciliophora*. In these groups of animals, homologs of the RnlB enzyme were found very rarely and could appear only through the horizontal gene transfer. The detected amino acid sequences of RNA ligase 2 homologs can serve for further structural and functional studies of these enzymes.

Most homologs similar to archaea and bacteria proteins belong to the underinvestigated groups of species, including those isolated from the ocean. This means that the detected sequences are new, and the finding these sequences can significantly enrich studies in the field of bacterial and archaeal enzymes.

We conducted the taxonomic affiliation of the homologs of T4 bacteriophage RNA ligase 2 in oceanic metagenomes (6 amino acid sequences from the GOS database and 15 from the LCGC14 database) revealed with the use of the PSI-BLAST program. It showed that most of the findings were related to homologs of RNA ligase 2 of bacterial origin and sequences from giant viruses of Eukaryotes.

A number of oceanic sequences were shown to be very similar to the homologs of RNA ligase 2 from lower invertebrates belonging to the kingdoms of *Alveolata*, *Stramenopiles*, and *Rhizaria*. In these groups of animals, homologs of this enzyme are very rare and may have appeared only in a number of representatives due to the horizontal gene transfer.

The only detected oceanic homolog closest to RNA ligase 2 of the T4 phage and other phages of the *Teevenvirinae* subfamily is also more similar to the representative of the archaea *Thermoplasmata archaeon*.

One of the objectives of the study was examination of the occurrence, evolution, and biodiversity of such rarely found in nature enzymes as RNA ligases 2. The homologous sequences of these proteins of oceanic origin can be used to obtain new enzymes – components of biotechnological tools – through routine gene synthesis and superproduction of the corresponding enzyme in bacterial systems. The discovery and study of RNA-ligase 2-related proteins can add to our knowledge of this group of enzymes and create the fundamental basis for the search for new inhibitors of these enzymes and for the production of drugs to combat the most dangerous diseases: trypanosomiasis and leishmaniasis.

The authors gratefully acknowledge G.V. Mikulinskaya for constructive comments on the text of the article. The work of Nazipova N.N. was supported by the Russian Foundation for Basic Research (RFBR grant No. 19-07-00996). The work of Zimin A.A. and Nikulin N.A. was supported by the Russian Foundation for Basic Research (RFBR grant No. 20-54-53018).

REFERENCES

1. Shuman S., Schwer B. RNA capping enzyme and DNA ligase: A superfamily of covalent nucleotidyl transferases. *Molecular Microbiology*. 1995. V. 17. P. 405–410. doi: 10.1111/j.1365-2958.1995.mmi_17030405.x.
2. Silber R., Malathi V.G., Hurwitz J. Purification and properties of bacteriophage T4-induced RNA ligase. *Proc. Natl. Acad. Sci. U S A*. 1972. V. 69. P. 3009–3013. doi: 10.1073/pnas.69.10.3009.
3. Wang L. K., Shuman S. Structure-function analysis of yeast tRNA ligase. *RNA*. 2005. V. 11. № 6. P. 966–975. doi: 10.1261/rna.2170305.
4. Ho C.K., Shuman S. Bacteriophage T4 RNA ligase 2 (gp24.1) exemplifies a family of RNA ligases found in all phylogenetic domains. *Proc. Natl. Acad. Sci. U S A.* 2002. V. 99. P. 12709–12714. doi: 10.1073/pnas.192184699.
5. Abelson J., Trotta C.R., Li H. tRNA splicing. *The Journal of Biological Chemistry*. 1998. V. 273. P. 12685–12688. doi: 10.1074/jbc.273.21.12685.
6. Englert M., Beier H. Plant tRNA ligases are multifunctional enzymes that have diverged in sequence and substrate specificity from RNA ligases of other phylogenetic origins. *Nucleic Acids Research*. 2005. V. 33. P. 388–399. doi: 10.1093/nar/gki174.
7. Blanc V., Alfonzo J.D., Aphasizhev R., Simpson L. The mitochondrial RNA ligase from *Leishmania tarentolae* can join RNA molecules bridged by a complementary RNA. *Journal of Biological Chemistry*. 1999. V. 274. P. 24289–24296. doi: 10.1074/jbc.274.34.24289.
8. Palazzo S.S., Panigrahi A.K., Igo R.P. Jr., Salavati R., Stuart K. Kinetoplastid RNA editing ligases: complex association, characterization, and substrate requirements.

Molecular and Biochemical Parasitology. 2003. V. 127. P. 161–167. doi: 10.1016/s0166-6851(02)00333-x.

9. Stuart K., Brun R., Croft S., Fairlamb A., Gurtler R.E., McKerrow J., Reed S., Tarleton R. Kinetoplastids: related protozoan pathogens, different diseases. *J. Clin. Invest*. 2008. V. 118. P. 1301–1310. doi: 10.1172/JCI33945.
10. Simpson L., Da Silva A. Isolation and characterization of kinetoplast DNA from *Leishmania tarentolae*. *J. Mol. Biol.* 1971. V. 56. P. 443–473. doi: 10.1016/0022-2836(71)90394-9.
11. Blum B., Bakalara N., Simpson L. A model for RNA editing in kinetoplastid mitochondria: RNA molecules transcribed from maxicircle DNA provide the edited information. *Cell*. 1990. V. 60. P. 89–198. doi: 10.1016/0092-8674(90)90735-W.
12. Sturm N.R., Simpson L. Kinetoplast DNA minicircles encode guide RNAs for editing of cytochrome oxidase subunit III mRNA. *Cell*. 1990. V. 61. P. 879–884. doi: 10.1016/0092-8674(90)90198-N.
13. Rehse P.H., Tahirov T.H. Structure of a putative 2'-5' RNA ligase from *Pyrococcus horikoshii*. *Acta Crystallographica Section D: Biological Crystallography*. 2005. V. 61. P. 1207–1212. doi: 10.1107/s0907444905017841.
14. Desai K.K., Bingman C.A., Phillips G.N. Jr., Raines R.T. Structures of the Noncanonical RNA Ligase RtcB Reveal the Mechanism of Histidine Guanylylation. *Biochemistry*. 2013. V. 52. P. 2518–2525. doi: 10.1021/bi4002375.
15. Desai K.K., Cheng C.L., Bingman C.A., Phillips G.N. Jr., Raines R.T. A tRNA splicing operon: archease endows RtcB with dual GTP/ATP cofactor specificity and accelerates RNA ligation. *Nucleic Acids Research*. 2014. V. 42. P. 3931–3942. doi: 10.1093/nar/gkt1375.
16. Aphasizhev R., Aphasizheva I. Mitochondrial RNA editing in trypanosomes: small RNAs in control. *Biochimie*. 2014. V. 100. P. 125–131. doi: 10.1016/j.biochi.2014.01.003.
17. Moreira S., Noutahi E., Lamoureux G., Burger G. Three-dimensional structure model and predicted ATP interaction rewiring of a deviant RNA ligase 2. *BMC Struct. Biol*. 2015. V. 15. Article No. 20. doi: 10.1186/s12900-015-0046-0.
18. Williamson S.J., Rusch D.B., Yooseph S., Halpern A.L., Heidelberg K.B., Glass J.I., Andrews-Pfannkoch C., Fadrosh D., Miller C.S., Sutton G., Frazier M., Venter J.C.. The Sorcerer II Global Ocean Sampling Expedition: Metagenomic Characterization of Viruses within Aquatic Microbial Samples. *PLoS One*. 2008. V. 3. Article No. e1456. doi: 10.1371/journal.pone.0001456.
19. Yooseph S., Sutton G., Rusch D.B., Halpern A.L., Williamson S.J., Remington K., Eisen J.A., Heidelberg K.B., Manning G., Li W., et al. The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families. *PLoS Biol*. 2007. V. 5. Article No. e16. doi: 10.1371/journal.pbio.0050016.
20. Jorgensen S.L., Hannisdal B., Lanzén A., Baumberger T., Flesland K., Fonseca R., Ovreås L., Steen I.H., Thorseth I.H., Pedersen R.B., Schleper C. Correlating microbial community profiles with geochemical data in highly stratified sediments from the Arctic Mid-Ocean Ridge. *Proc. Natl. Acad. Sci. U S A*. 2012. V. 109. P. E2846–E2855. doi: 10.1073/pnas.1207574109.
21. Brettin T., Davis J.J., Disz T., Edwards R.A., Gerdes S., Olsen G.J., Olson R., Overbeek R., Parrello B., Pusch G.D., et al. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci. Rep*. 2015. V. 5. Article No. 8365. doi: 10.1038/srep08365.
22. King A.M.Q., Lefkowitz E.J., Mushegian A.R., Adams M.J., Dutilh B.E., Gorbalenya A.E., Harrach B., Harrison R.L., Junglen S., Knowles N.J., et al. Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the

International Committee on Taxonomy of Viruses (2018). *Arch. Virol*. 2018. V. 163. P. 2601–2631. doi: 10.1007/s00705-018-3847-1.

23. Federhen S. The NCBI Taxonomy database. *Nucleic Acids Res*. 2012. V. 40. P. D136–D143. doi: 10.1093/nar/gkr1178.
24. Benson D.A., Cavanaugh M., Clark K., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W. GenBank. *Nucleic Acids Res*. 2013. V. 41. P. D36–D42. doi: 10.1093/nar/gks1195.
25. Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997. V. 25. P. 3389–3402. doi: 10.1093/nar/25.17.3389.
26. Jones D.T., Taylor W.R., Thornton J.M. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*. 1992. V. 8. P. 275–282. doi: 10.1093/bioinformatics/8.3.275.
27. Felsenstein J. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*. 1985. V. 39. P. 783–791. doi: 10.1111/j.1558-5646.1985.tb00420.x.
28. Kumar S., Stecher G., Li M., Knyaz C., Tamura K.). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution*. 2018. V. 35. P. 1547–1549. doi: 10.1093/molbev/msy096.
29. Cavalier-Smith T. Kingdom Chromista and its eight phyla: a new synthesis emphasising periplastid protein targeting, cytoskeletal and periplastid evolution, and ancient divergences. *Protoplasma*. 2018. V. 255. P. 297–357. doi: 10.1007/s00709-017-1147-3.
30. Edgar R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004. V. 32. P. 1792–1797. doi; 10.1093/nar/gkh340.
31. Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. ClustalW and ClustalX version 2.0. *Bioinformatics*. 2007. V. 23. P. 2947–2948. doi: 10.1093/bioinformatics/btm404.

Received 23.10.2020. Published 31.12.2020.