

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/351462799>

# Revealing RNA virus diversity and evolution in unicellular algae transcriptomes

Preprint · May 2021

DOI: 10.1101/2021.05.09.443335

CITATIONS

0

READ

1

3 authors, including:



**Justine Charon**

The University of Sydney

34 PUBLICATIONS 206 CITATIONS

[SEE PROFILE](#)



**Edward C Holmes**

The University of Sydney

2,520 PUBLICATIONS 75,602 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Molecular biology in Trypanosoma species [View project](#)



Protein Intrinsic Disorder and Virus Molecular Evolution [View project](#)

# 1 **Revealing RNA virus diversity and evolution in unicellular** 2 **algae transcriptomes**

3

4

5 Justine Charon<sup>1</sup>, Shauna Murray<sup>2</sup>, Edward C. Holmes<sup>1\*</sup>

6

7

8 <sup>1</sup>Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and  
9 Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney  
10 NSW 2006, Australia.

11 <sup>2</sup>University of Technology Sydney, School of Life Sciences, Sydney NSW 2007, Australia.

12

13

14 \* Corresponding author:

15 Prof. Edward C. Holmes,

16 Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and

17 Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney

18 NSW 2006, Australia.

19 Email: [edward.holmes@sydney.edu.au](mailto:edward.holmes@sydney.edu.au)

## 20 **Abstract**

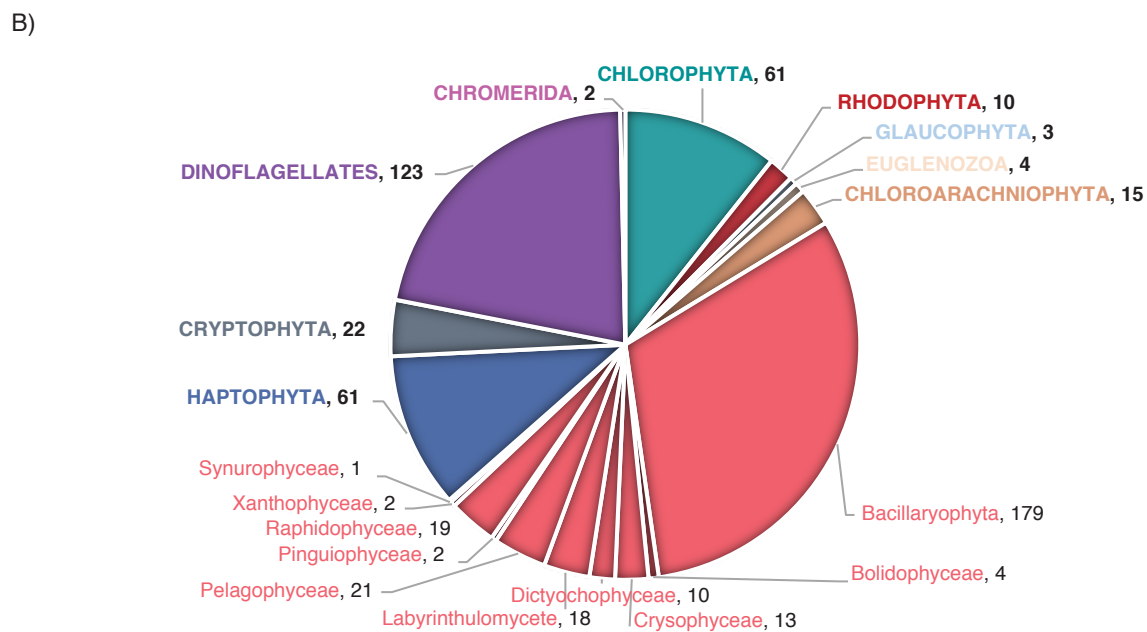
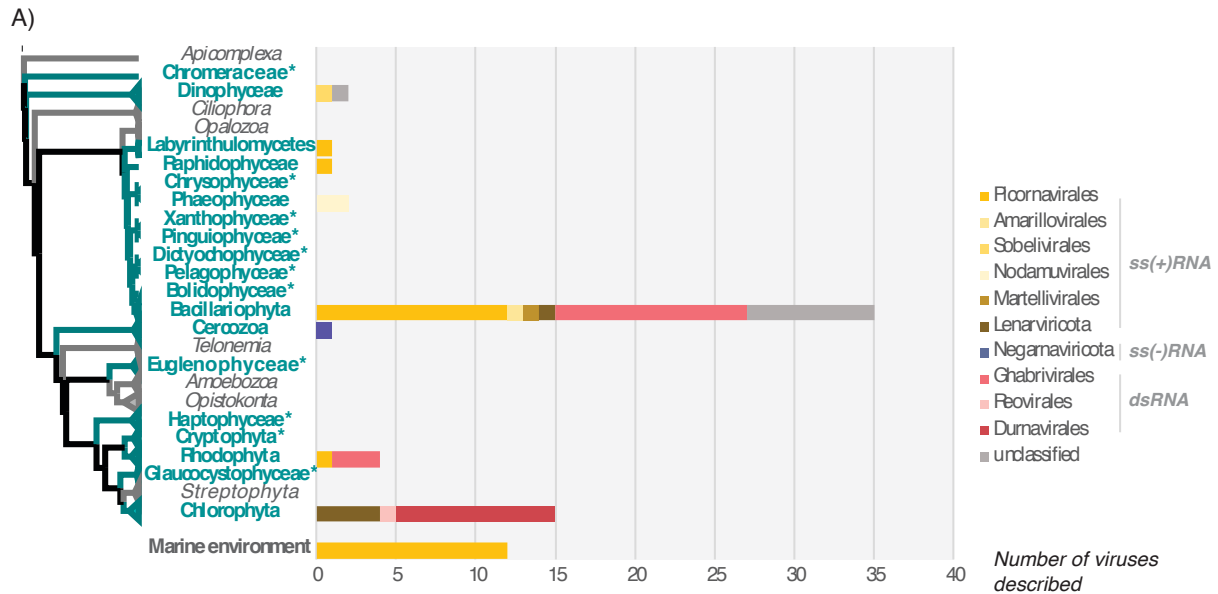
21 Remarkably little is known about the diversity and evolution of RNA viruses in unicellular  
22 eukaryotes. We screened a total of 570 transcriptomes from the Marine Microbial Eukaryote  
23 Transcriptome Sequencing Project (MMETSP) project that encompasses a wide diversity of  
24 microbial eukaryotes, including most major photosynthetic lineages (i.e. the microalgae).  
25 From this, we identified 30 new and divergent RNA virus species, occupying a range of  
26 phylogenetic positions within the overall diversity of RNA viruses. Approximately one-third  
27 of the newly described viruses comprised single-stranded positive-sense RNA viruses from  
28 the order *Lenarviricota* associated with fungi, plants and protists, while another third were  
29 related to the order *Ghabrivirales*, including members of the protist and fungi-associated  
30 *Totiviridae*. Other viral species showed sequence similarity to positive-sense RNA viruses  
31 from the algae-associated *Marnaviridae*, the double-stranded RNA *Partitiviridae*, as well as a  
32 single negative-sense RNA virus related to the *Qinviridae*. Importantly, we were able to  
33 identify divergent RNA viruses from distant host taxa, revealing the ancestry of these viral  
34 families and greatly extending our knowledge of the RNA viromes of microalgal cultures.  
35 Both the limited number of viruses detected per sample and the low sequence identity to  
36 known RNA viruses imply that additional microalgal viruses exist that could not be detected  
37 at the current sequencing depth or were too divergent to be identified using sequence  
38 similarity. Together, these results highlight the need for further investigation of algal-  
39 associated RNA viruses as well as the development of new tools to identify RNA viruses that  
40 exhibit very high levels of sequence divergence.

## 41 **1. Introduction**

42 Viruses likely infect most, if not all, cellular species. For example, metagenomic studies of  
43 marine environments have revealed an enormous abundance and diversity of both DNA and  
44 RNA viruses (up to  $10^8$  viruses/ ml)<sup>1</sup> as well as their key role in biogeochemical processes<sup>2</sup>.  
45 Such ubiquity highlights the importance of obtaining a comprehensive picture of global virus  
46 diversity, including in host taxa that have only been poorly sampled to date<sup>3</sup>. Viruses of  
47 protists are a major exemplar of this untapped diversity.

48 Protists, defined as eukaryotic organisms that are not animal, plant, or fungi<sup>4</sup>, are  
49 highly diverse and include the algae. Some protists play a critical role in ecosystems as  
50 primary producers as well as being involved in nutrient cycling. Next generation sequencing  
51 (NGS) of protists has shown that their diversity is far greater than previously thought, with  
52 species numbers likely exceeding one million, although only a tiny fraction have been  
53 described to date<sup>5</sup>. In addition, protists have already proven to be an important source of virus  
54 diversity, with the giant *Mimiviridae* from the Amoebozoa a notable case in point<sup>6</sup>. Despite  
55 this, protist viruses remain largely overlooked, especially those associated with the  
56 unicellular microalgae. This is particularly striking in the case of RNA viruses: although  
57 RNA viruses were first described in unicellular algae in 2003<sup>7</sup>, they still comprise only 73  
58 species from a very small number of algal lineages (Figure 1A)<sup>8</sup>.

59 There have been several metagenomic studies of viruses in aquatic microbial  
60 eukaryotes<sup>9,10</sup>. These have identified many thousands of virus sequences, with at least half  
61 predicted to have RNA genomes<sup>11,12</sup>. Similarly, metagenomics is proving a valuable means to  
62 mine viral diversity in uncultivable organisms<sup>13</sup>. However, because these studies have been  
63 conducted with environmental samples they cannot identify the specific host taxon with  
64 certainty.



65

66 **Figure 1. Currently reported RNA virus diversity in microalgae and the taxa studied**  
 67 **here. (A)** Left, Eukaryote phylogeny. The microalgae-containing eukaryotic lineages  
 68 investigated here are highlighted in bold green. \*Microalgae lineages for which no RNA  
 69 viruses have been reported to date. Right, number of total viruses formally or likely  
 70 associated with microalgae reported at NCBI (<https://www.ncbi.nlm.nih.gov/labs/virus/vssi/>),  
 71 VirusHostdb (<https://www.genome.jp/virushostdb/>) and the literature. Viruses are coloured  
 72 based on their taxonomy and genome composition. **(B)** Representative taxa from major algal  
 73 lineages used in this study and the total number of transcriptomes analysed for each lineage.  
 74

75 This illustrates the inference gap between broad scale metagenomic surveys that identify

76 huge numbers of new viral sequences, creating a large but unassigned depiction of the

77 virosphere, and those studies based on virus isolation and detailed particle characterization,

78 including cell culture, that are conducted on a very limited of number of viruses and create a  
79 highly accurate, but very narrow, vision of the virosphere<sup>14</sup>. However, establishing strong  
80 links between viruses and their specific hosts provides a firmer understanding of virus  
81 ecology and evolution, as well as virus-host interactions. Hence, the NGS-based investigation  
82 of RNA virus diversity from individual host species serves as a good compromise to fill the  
83 gap between large-scale virus detection through metagenomics and the detailed assignment of  
84 hosts through virus isolation and cell culture.

85         To better understand diversity of RNA viruses associated with microalgae, we  
86 performed viral metatranscriptomic analyses of data obtained from the Marine Microbial  
87 Eukaryote Transcriptome Sequencing Project (MMETSP)<sup>15</sup>. With 210 unique genera  
88 covering most unicellular algal-comprising lineages, the MMETSP constitutes the largest  
89 collection of transcriptome data collected from microbial eukaryote cultures, including axenic  
90 ones, and hence depicts a large component of eukaryotic diversity<sup>15</sup> (Figure 1). Accordingly,  
91 we used both sequence and structural-based approaches to screen 570 transcriptomes from 19  
92 major microalgae-containing lineages for the most conserved “hallmark” protein of RNA  
93 viruses – the RNA-dependent RNA polymerase (RdRp). To the best of our knowledge, this is  
94 the broadest exploration of RNA viruses conducted at the single host species level in  
95 microbial eukaryotes and the first attempt to identify RNA viruses in most of the microalgal  
96 lineages investigated here (Figure 1).

## 97 **2. Methods**

### 98 **2.1 MMETSP contig retrieval**

99 In total, 570 MMETSP accessions, corresponding to the microalgal-containing lineages, were  
100 included in this study. Contig data sets corresponding to each accession were retrieved from a  
101 Trinity re-assembly performed on the RNA-Seq data sets from MMETSP and available at

102 <https://doi.org/10.5281/zenodo.740440><sup>16</sup>. A description of all the transcriptome accessions  
103 and samples analysed here is available in Table S1.

## 104 **2.2 ORF annotation**

105 To optimize our computational analysis of the 570 contig data sets, we focused on those  
106 predicted to encode ORFs with a minimum length of 200 amino acids (assuming that shorter  
107 contigs would be too short to be included in a robust phylogenetic analyses). Accordingly,  
108 ORFs >200 amino acids in length were predicted using the GetORF tool from the EMBOSS  
109 package (v6.6.0). ORFs were predicted using the standard genetic code (with alternative  
110 initiation codons) as alternative genetic codes are not used in the microalgae analysed here<sup>17</sup>.  
111 The option -find 0 (translation of regions between STOP codons) was used to enable the  
112 detection of partial genomes, in which START codons could be missing due to partial virus  
113 genome recovery.

## 114 **2.3 RNA virus sequence detection using sequence similarity**

115 All predicted ORFs were compared to the entire non-redundant protein database (nr) (release  
116 April 2020) using DIAMOND BLASTp (v0.9.32)<sup>18</sup> with the following options: --max-target-  
117 seqs 1 (top hit with best score retained) and an e-value cut-off of 1e-03. Additional sequence  
118 comparisons with identical BLASTp parameters were performed using either the newly-  
119 detected RdRp sequences or the RdRps from a previous large-scale analysis<sup>12</sup> (available at  
120 [ftp://ftp.ncbi.nih.gov/pub/wolf/\\_suppl/yangshan/rdrp.ya.fa](ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/yangshan/rdrp.ya.fa)).

121 To limit false-negative detection due to a bias in ORF prediction (in particular, partial  
122 genomes may not be detected due to their short length), all the contig nucleotide sequences  
123 were submitted to a RdRp protein database using DIAMOND BLASTx (v0.9.32, more  
124 sensitive option and 1e-03 e-value cut-off)<sup>18</sup> to identify any additional RNA viruses. Top hits  
125 were retained and re-submitted against the entire nr protein database (April 2020 release) to

126 remove false-positive hits (queries with a greater match to non-viral hits). All sequences  
127 retained from both the BLASTp and RdRp BLASTx analysis were manually checked to  
128 remove non-RNA virus sequences based on their taxonomy (predicted using the TaxonKit  
129 tool from NCBI; <https://github.com/shenwei356/taxonkit>).

130 All RNA virus-like sequences detected were functionally annotated using  
131 InterProscan (v5.39-77.0, default parameters) and non-RdRp sequences were filtered out.  
132 One sequence, sharing homology with the QDH87844.1 hypothetical protein  
133 H3RhizoLitter144407\_000001, partial [Mitovirus sp.], was observed in 86 of the 570 data  
134 sets, including multiple species from multiple sampling locations. Considering the prevalence  
135 of this hit and the 100% identity between samples, we assumed this originates from  
136 environmental or sequencing-associated contamination. In addition, a small number of RNA  
137 virus-like sequences were identified based on their similarity to the RdRp from bovine viral  
138 diarrhea viruses 1 and 2 and considered biological product contaminants<sup>19</sup>. These were also  
139 discarded.

#### 140 **2.4 RNA virus sequence detection using protein profiles and 3D structures**

141 In an attempt to detect more divergent viral RdRps we compared all the “orphan” ORFs (i.e.  
142 ORFs without any BLASTp hits at the 1e-03 e-value cut-off) against the viral RdRp-related  
143 profiles from the PFAM<sup>20</sup> and PROSITE databases (Table S2) using the HMMer3 program<sup>21</sup>  
144 (v3.3, default parameters, e-value<1e-05). An additional attempt to annotate orphan  
145 translated-ORFs was performed on the remaining sequences using the InterProscan software  
146 package from EMBL-EBI (v5.39-77.0, default parameters) ([https://github.com/ebi-pf-](https://github.com/ebi-pf-team/interproscan)  
147 [team/interproscan](https://github.com/ebi-pf-team/interproscan)).

148 The RdRp-like candidates identified in both the HMMer3 and InterProscan analysis  
149 were submitted to the Protein Homology/analogy Recognition Engine v 2.0 (Phyre2) web  
150 portal<sup>22</sup> to confirm the presence of a RdRp signature (Table S3). Non-viral proteins (i.e. non-



151 viral Phyre2 hit >90% confidence) were discarded, as were sequences with low HMM (e-  
152 value >1e-03) and Phyre2 scores (confidence level > 90%). Sequences that matched either the  
153 HMM RdRp (>1e-05) and/or Phyre2 RdRp (>90% confidence) were retained for further  
154 characterization as potential RNA viruses. In total, 80 RdRp-like candidates were quality-  
155 assessed by coverage analysis and manual checked for the presence of the standard A, B and  
156 C catalytic viral RdRp sequence motifs<sup>23</sup> using Geneious (v11.1.4)<sup>24</sup>. Only those displaying  
157 related RdRp-like motifs were retained as potential RdRp protein candidates (Table S3).

## 158 **2.5 Contig manual extension and genome annotation**

159 Full-length nucleotide sequences encoding the protein retained from the sequence-based and  
160 structure-based detection approaches were retrieved and used as references for mapping SRA  
161 reads corresponding to each sample (BioProject PRJNA231566) using the SRA extension  
162 package of Bowtie2 (v2.3.5.1-sra)<sup>25</sup>. Read coverages of each contig were checked using  
163 Geneious (v11.1.4) and, when needed, extremities were manually extended and contigs re-  
164 submitted to read mapping, until no overhanging extremities were observed.

165 The relative abundance of each putative viral sequence was reported as the number of  
166 reads per million: that is, the number of reads mapping to the contig divided by the total  
167 number of reads of the corresponding SRA library multiplied by one million. Poorly-  
168 represented viral sequences were considered as potential cross-library contaminants derived  
169 from index-hopping and discarded when they accounted for less than 0.1% of the highest  
170 abundance of the same sequence in another library<sup>26</sup>.

171 Genomic organizations were constructed using Geneious (v11.1.4). ORFs were  
172 predicted using the standard genetic code or, when suitable, using alternative mitochondrial  
173 or plastid-associated genetic codes. Tentative virus names were taken from Greek mythology.

## 174 **2.6 Host *rbcL* gene abundance estimation**

175 To estimate levels of virus abundance in comparison to those from their putative hosts, the  
176 abundance of the host Ribulose biphosphate carboxylase large chain (*rbcL*) gene was  
177 assessed using the Bowtie2 SRA package (v2.3.5.1-sra) and mapped to SRA reads from the  
178 *rbcL* gene of each corresponding species (whenever available)<sup>25</sup>. The SRA and *rbcL* gene  
179 accessions used are reported in Table S4.

## 180 **2.7 Secondary host profiling**

181 According to the MMETSP sample requirements, all cultures were subjected to SSU rRNA  
182 sequencing to ensure they were mono-strain and not contaminated with additional microbial  
183 eukaryotes. Nevertheless, the presence of other microbial contaminants was possible. As we  
184 expect most of the potential Archaea and Bacteria contaminants will not have an available  
185 genome sequence, their profiling in the samples was performed by analysing the closest  
186 homologs of each contig using both BLASTn (BLAST+ package, v2.9.0) and BLASTp  
187 (DIAMOND, v2.0.4) against the nt and nr databases, respectively. Contigs were grouped at  
188 the kingdom level based on the taxonomic affiliation of their closest homologs in the  
189 databases, with the abundance of each kingdom defined as the sum of each contig abundance  
190 value (transcripts per million)<sup>16</sup>.

## 191 **2.8 Phylogenetic analysis**

192 For each virus phylum and order, the RefSeq and most closely related RdRp sequences were  
193 retrieved from GenBank and aligned with newly identified RdRp sequences using the L-INS-  
194 I algorithm in the MAFFT program (v7.402)<sup>27</sup>. Resulting sequence alignments were trimmed  
195 using TrimAl to remove ambiguously aligned regions with different levels of stringency,  
196 optimized for each alignment (v1.4.1, “automated1” mode). Maximum likelihood  
197 phylogenies based on amino acid alignments were inferred using IQ-TREE (v2.0-rc1)<sup>28</sup>, with

198 ModelFinder used to find the best-fit substitution model in each case (see figure legends)<sup>29</sup>  
199 and both the SH-like approximate likelihood ratio test and ultrafast nonparametric bootstrap  
200 (1000 replicates) used to assign support to individual nodes<sup>30</sup>. All phylogenies were  
201 visualized, and mid-point rooted (for clarity only) using the Figtree software (v1.4.4).

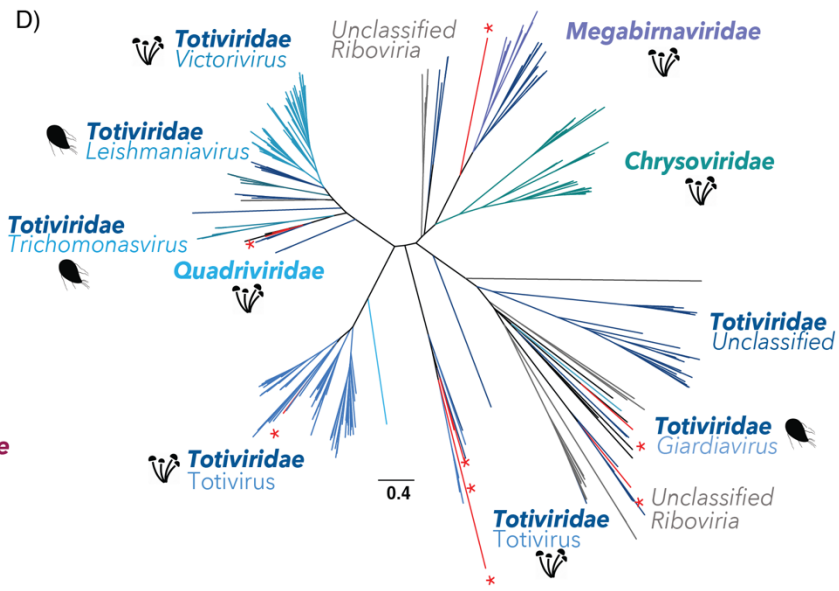
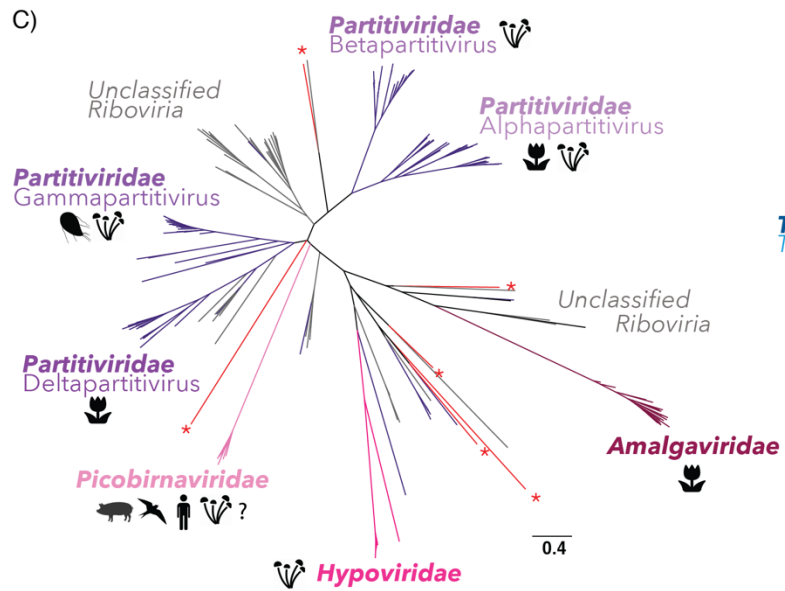
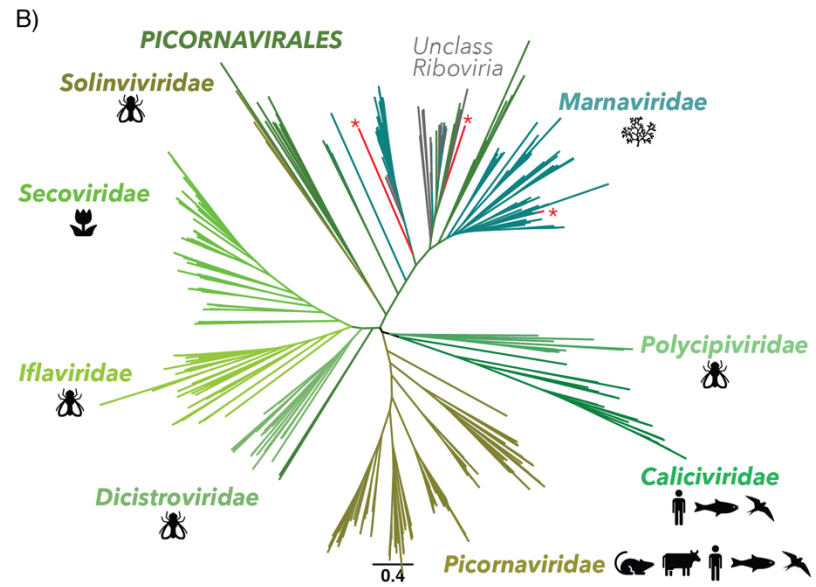
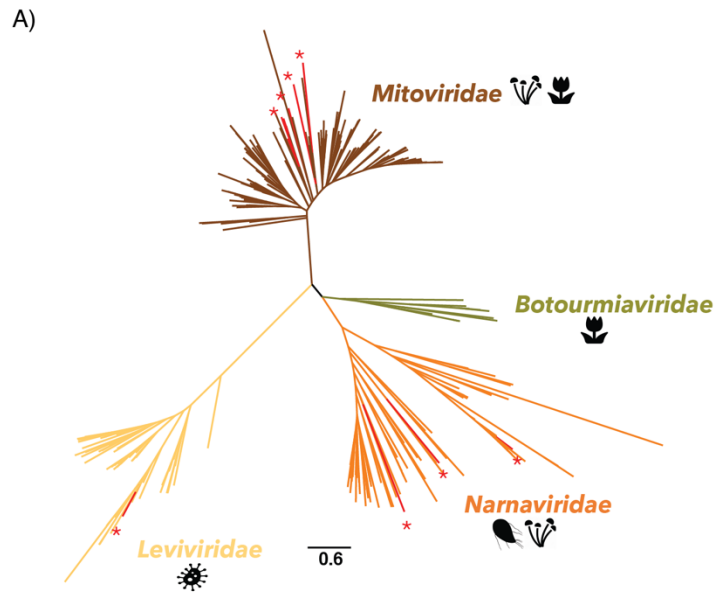
## 202 **2.9 Detection of endogenous viral elements**

203 To determine whether any of the newly detected viral sequences were endogenous viral  
204 elements (EVEs) rather than true exogenous viruses, the nucleotide sequences of viral  
205 candidates were used as a query for BLASTn (online version, default algorithm parameters)  
206 against corresponding host genome sequence, whenever available.

## 207 **3. Results**

### 208 **3.1 Overall virus diversity**

209 Our analysis of the 570 MMETSP transcriptomes obtained from 247 total microalgal species  
210 spread over 10 major groups of algae (Table 1B) identified 30 new RNA viral species. These  
211 newly identified viruses largely represented the single-stranded positive-sense RNA  
212 (ssRNA+) virus phylum *Lenarviricota* and the order *Picornavirales* (Figure 2A and B), as  
213 well as the double-stranded (dsRNA) RNA virus orders *Durnavirales* and *Ghabrivirales*  
214 (Figure 2C and D). A single negative-sense RNA (ss-RNA) virus was also identified in  
215 *Pseudo-nitzchia heimii* that fell within the *Qinviridae* (order *Muvirales*).



217 **Figure 2. Newly described RNA virus sequences within the diversity of RNA viruses using RdRp phylogenies.** Newly described sequences  
218 are indicated in red with “\*” symbols. Phylogenies of: (A) the phylum *Lenarnaviricota* (ssRNA+); (B) the order *Picornavirales* (ssRNA+); (C)  
219 the order *Durnavirales* (dsRNA); (D) the order *Ghabrivirales* (dsRNA). For each viral family, the host range was retrieved from VirusHostdb  
220 and the ICTV report<sup>31,32</sup>.

221 **Table 1. List of new RNA viruses discovered in this study.** Read abundances are indicated as the number of reads per million. Likely hosts  
 222 correspond to eukaryotic lineages detected at levels using BLASTn/BLASTp analysis and phylogenies.  
 223

Virus name	MMETSP sample (Phylum/class)	Genome status	Reads/ million	BLASTp best hits (GenBank acc./Organism)	%ID	E-value	Likely host(s) (BLAST)	Likely host(s) (Phylogenies)	Proposed host
Amphitrite narna-like virus	MMETSP1061 <i>P. pungens</i> (Bacillariophyta)	Full-length	48	QIR30281.1 RdRp [Plasmopara viticola associated narnavirus 2]	41	5E-144	Bacillariophyta	Fungi/Protist	Bacillariophyta
Poseidon narna-like virus	MMETSP0418 <i>A. radiata</i> (Bacillariophyta)	Partial	8	QDH89392.1 RdRp, partial [Mitovirus sp.]	34	4E-17	Bacillariophyta	Marine arthropod	Bacillariophyta
Halia narna-like virus	MMETSP0418 <i>A. radiata</i> (Bacillariophyta)	Full-length	108	QBC65281.1 RdRp, partial [Rhizopus microsporus 23S narnavirus]	32	4E-17	Bacillariophyta	Protist	Bacillariophyta
Triton levi-like virus	MMETSP1471 <i>P. provasolii</i> (Chlorophyta)	Partial	64	APG76993.1 hypothetical protein [Beihai levi-like virus 20]	46	3E-65	Chlorophyta; Bacteria	Bacteria	Bacteria
Aiolos mito-like virus	MMETSP0286 <i>P. polylepis</i> (Haptophyta)	Full-length	54	YP_009272901.1 RdRp [Fusarium poae mitovirus 4]	35	3E-38	Haptophyta	Sea sponge	Haptophyta
Asopus mito-like virus	MMETSP0164 <i>C. braarudii</i> (Haptophyta)	Partial	12	QDM55307.1 RdRp [Geopora sumneriana mitovirus 1]	34	2E-35	Haptophyta	Sea sponge	Haptophyta
Athena mito-like virus	MMETSP0719 <i>C. curvisetus</i> (Bacillariophyta)	Partial	54	ASM94070.1 putative RdRp, partial [Barns Ness breadcrumb sponge narna-like virus 5]	65	6E-72	Bacillariophyta; Bacteria	Sea sponge	Bacillariophyta
Daimones mito-like virus	MMETSP0286	Full-length	104	YP_009552787.1 RNA-directed RNA polymerase	26	4E-16	Haptophyta	Freshwater arthropods	Haptophyta

	<i>P. polylepis</i> (Haptophyta)			[Rhizophagus sp. RF1 mitovirus]					
Despoena mito-like virus	MMETSP0167 <i>R. maculata</i> (Rhodophyta)	Full- length	115	ALM62241.1 RdRp [Soybean leaf-associated mitovirus 1]	34	6E-32	Rhodophyta; Bacteria	Freshwater arthropods	Rhodophyta
Proteus mito-like virus	MMETSP1081 <i>P. amyliifera</i> (Chlorophyta)	Full- length	388	ALM62242.1 RdRp [Soybean leaf-associated mitovirus 2]	32	7E-46	Chlorophyta	Fungi/Protist	Chlorophyta
Telchines mito-like virus	MMETSP0725 <i>Amphiprora</i> (Bacillariophyta)	Partial	15	QDA33961.1 RdRp [Mitovirus 1 BEG47]	25	5E-21	Bacillariophyta	Algae	Bacillariophyta
	MMETSP0724 <i>Amphiprora</i> (Bacillariophyta)	Partial	26						
Susy yue-like virus	MMETSP1423 <i>P. heimii</i> (Bacillariophyta)	Partial	5	QDH86724.1 RdRp, partial [Qinviridae sp.]	42	1E-21	Bacillariophyta	Soil samples/ Marine arthropod	Bacillariophyta
Aethusa amalga- like virus	MMETSP0011 <i>R. marinus</i> (Rhodophyta)	Partial	83	ANN12897.1 putative CP/RdRp [Zygosaccharomyces bailii virus Z]	43	2E-12	Rhodophyta; Bacteria	Marine arthropod	Rhodophyta
Benthesicyme durna-like virus	MMETSP1319 <i>T. pacifica</i> (Bolidophyceae)	Partial	404	QDH90748.1 RdRp, partial [Partitiviridae sp.]	29	1E-17	Bolidophyceae	Protist	Bolidophyceae
Herophile durna- like virus	MMETSP0140 <i>P. australis</i> (Bacillariophyta)	Partial	10	QOW97238.1 RdRp [Amalga-like lacheneavirus]	27	2E-19	Bacillariophyta	Chlorophyta	Bacillariophyta
Cymopoleia durna- like virus	MMETSP1081 <i>P. amyliifera</i> (Chlorophyta)	Partial	10	YP_009551448.1 RdRp [Diatom colony associated dsRNA virus 2]	31	2E-34	Chlorophyta	Fungi	Chlorophyta

Ourea durna-like virus	MMETSP0797 <i>D. acuminata</i> (Dinophyceae)	Partial	4	ARO72610.1 RdRp [Spinach deltapartitivirus 1]	27	4E-11	Dinophyceae; Bacteria	Land plant	Dinophyceae
Aegean partiti-like virus	MMETSP0491 <i>T. chuii</i> (Chlorophyta)	Full-length	3296	QOW97235.1 RdRp [Partiti-like lacotivirus]	29	6E-62	Chlorophyta	Chlorophyta	Chlorophyta
Pelias marna-like virus	MMETSP1377 <i>Symbiodinium sp.</i> (Dinophyceae)	Full-length	60553	YP_009337401.1 hypothetical protein 2 [Wenzhou picorna-like virus 4]	26	8E-98	Dinophyceae	Algae	Xanthophyceae
Neleus marna-like virus, 1	MMETSP0946 <i>V. litorea</i> (Xanthophyceae)	Full-length	806763	YP_009336927.1 hypothetical protein 1 [Shahe picorna-like virus 3]	33	3E-180	Vaucheriaceae	Algae	Xanthophyceae
Neleus marna-like virus, 2	MMETSP0945 <i>V. litorea</i> (Xanthophyceae)	Full-length	711119	YP_009336927.1 hypothetical protein 1 [Shahe picorna-like virus 3]	33	4E-180	Vaucheriaceae	Algae	Xanthophyceae
Tyro marna-like virus	MMETSP0905 <i>T. antarctica</i> (Bacillariophyta)	Partial	126	YP_001429582.1 hypothetical protein JP-A_gp2 [Marine RNA virus JP-A]	75	3E-272	Bacillariophyta; Bacteria	Algae	Bacillariophyta
	MMETSP0903 <i>T. antarctica</i> (Bacillariophyta)	Partial	2034						
	MMETSP0902 <i>T. antarctica</i> (Bacillariophyta)	Partial	237						
Aloadae toti-like virus,1	MMETSP1388 <i>Isochrysis</i> (Haptophyta)	Partial	39	QIJ70132.1 RdRp [Keenan toti-like virus]	33	2E-109	Haptophyta	Fungi /Invertebrates	Haptophyta
Aloadae toti-like virus,2	MMETSP1090	Partial	11	QIJ70132.1 RdRp [Keenan toti-like virus]	33	2E-109	Haptophyta	Fungi /Invertebrates	Haptophyta



	<i>Isochrysis</i> (Haptophyta)									
Antaeus toti-like virus,1	MMETSP0154 <i>T. antarctica</i> (Bacillariophyta)	Full-length	27	QGY72637.1 putative coat protein [Plasmopara viticola associated totivirus-like 2]	22	1E-10	Bacillariophyta	Protist	Bacillariophyta	
Antaeus toti-like virus,2	MMETSP0152 <i>T. antarctica</i> (Bacillariophyta)	Full-length	7	BBJ21451.1 CP-RdRp fusion protein [Pythium splendens RNA virus 1]	40	5E-53	Bacillariophyta	Protist	Bacillariophyta	
Charybdis toti-like virus	MMETSP0853 <i>P. fraudulenta</i> (Bacillariophyta)	Partial	38	YP_003288763.1 RdRp [Rosellinia necatrix megabirnavirus 1/W779]	30	4E-24	Bacillariophyta; Bacteria	Fungi	Bacillariophyta	
	MMETSP0851 <i>P. fraudulenta</i> (Bacillariophyta)	Partial	44							
	MMETSP0850 <i>P. fraudulenta</i> (Bacillariophyta)	Partial	41							
	MMETSP0852 <i>P. fraudulenta</i> (Bacillariophyta)	Partial	14							
Chrysaor toti-like virus	MMETSP0418 <i>A. radiata</i> (Bacillariophyta)	Partial	40	YP_009551502.1 RdRp [Diatom colony associated dsRNA virus 17 genome type B]	27	9E-95	Bacillariophyta; Bacteria	Soil	Bacillariophyta	
Laestrygon toti-like virus	MMETSP1451 <i>V. brassicaformis</i> (Chromeraceae)	Partial	29	YP_009551504.1 RdRp [Diatom colony associated dsRNA virus 17 genome type A]	34	4E-112	Chromeraceae	Soil	Chromeraceae	

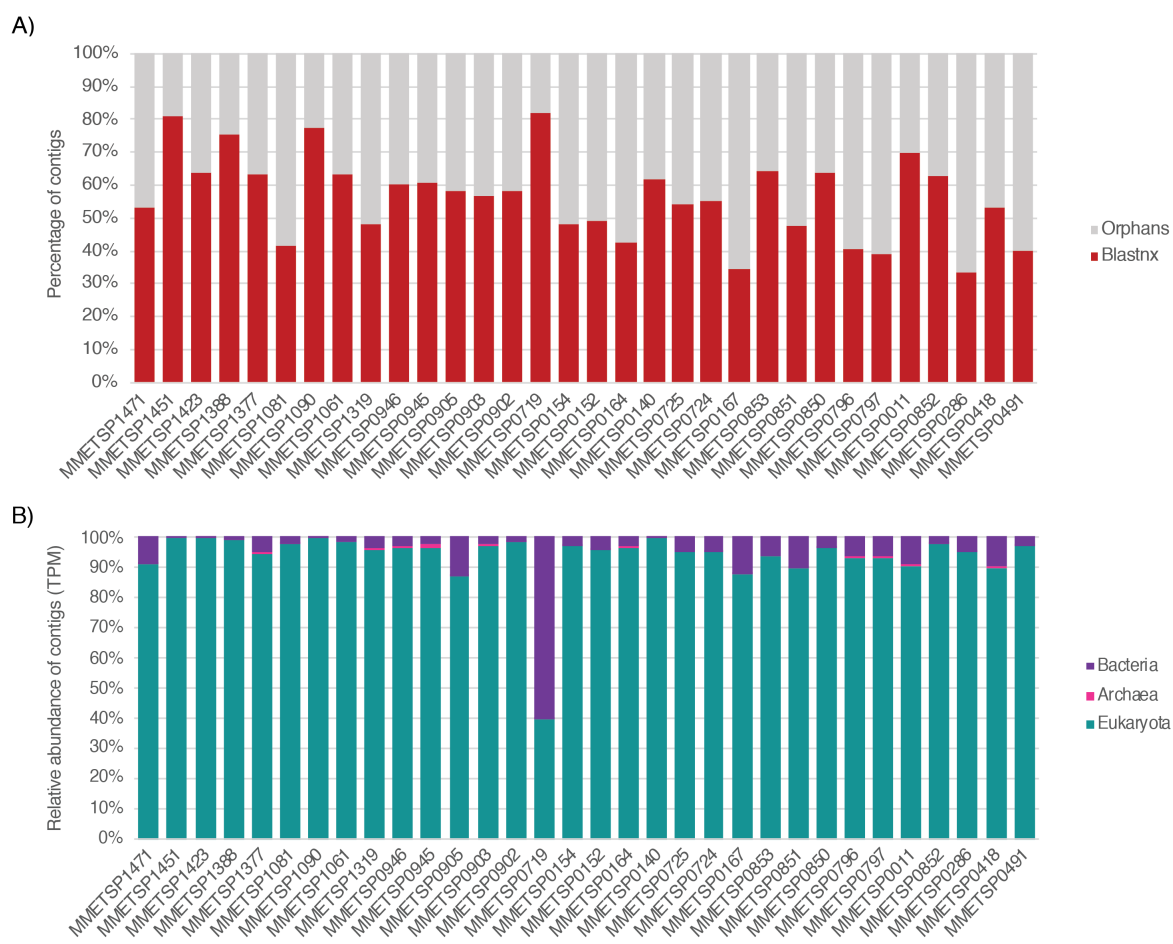
Arion toti-like virus	MMETSP0796 <i>P. bahamense</i> (Dinophyceae)	Partial	31	QGA70930.1 RdRp [Ahus virus]	25	3E-18	Dinophyceae; Bacteria	Protist/ Marine host	Dinophyceae
Otus toti-like virus	MMETSP0011 <i>R. marinus</i> (Rhodophyta)	Full-length	31	AMB17469.1 RdRp, partial [Delisea pulchra totivirus IndA]	51	3E-120	Rhodophyta; Bacteria	Fungi	Rhodophyta
Polyphemus toti-like virus	MMETSP0418 <i>A. radiata</i> (Bacillariophyta)	Partial	10	YP_009552789.1 RdRp [Diatom colony associated dsRNA virus 5]	59	3E-79	Bacillariophyta; Bacteria	Algae/ Protist	Bacillariophyta
Ephialtes toti-like virus	MMETSP0418 <i>A. radiata</i> (Bacillariophyta)	Partial	13	YP_009552789.1 RdRp [Diatom colony associated dsRNA virus 5]	63	1E-200	Bacillariophyta; Bacteria	Algae/ Protist	Bacillariophyta

225 Notably, all the RdRps identified in the BLAST analysis exhibited very high levels of  
226 sequence divergence, with median pairwise identity values of only ~35% to the closest  
227 known virus homolog (Table 1). In addition, with the exceptions of Pelias marna-like virus  
228 and Neleus marna-like virus, the newly described viral sequences were at relatively low  
229 abundance all (Table 1). This may reflect the lack of an rRNA depletion step used in the  
230 MMETSP library preparation, such that any RNA viruses would necessarily only represent a  
231 small proportion of reads. To shed more light on this issue, we compared levels of virus  
232 abundance with the expression levels of a host gene – that encoding the large subunit of the  
233 Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase (*rbcL*) (Figure S1, Table S4). The *rbcL*  
234 gene is commonly used as a diversity marker in algae<sup>33</sup>, and sequences are available for all  
235 the microalgal species used here. Overall, the number of reads mapping to putative RNA  
236 viruses are in the same order of magnitude or higher than those reported for the host *rbcL*  
237 gene (Figure S1), compatible with their designation as replicating viruses.

### 238 **3.2 Additional cellular organisms in the transcriptome data**

239 We used mono-strain cultures of microbial eukaryotes to investigate the relationship among  
240 RNA viruses and their hosts. While the lack of additional eukaryotic organisms (fungi, other  
241 protists) was supposedly ensured under the MMETSP project guidelines, with 18S rRNA  
242 sequencing of each culture<sup>15</sup>, some caveats remain for non-axenic cultures (Table S5).  
243 Indeed, some cultures likely contain contaminating Bacteria or Archaea, sometimes as  
244 intracellular parasites or as obligate mutualists in the culture media<sup>5</sup>. To assess this, contigs  
245 from libraries positive for RNA viruses were submitted to BLASTn and BLASTx. The ratio  
246 of assigned contigs and their kingdom assignments are summarized Figure 3 and used to infer  
247 the likely host organisms (Table 1).

248



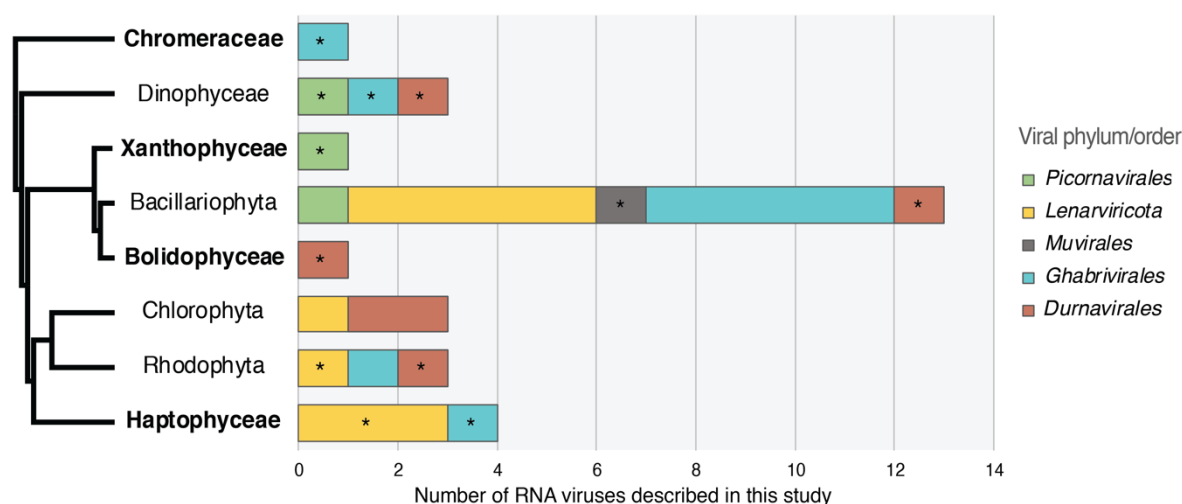
249

250 **Figure 3. Taxonomic assignment of contigs in RNA virus positive MMETSP libraries.**  
 251 (A) Ratio of contigs with hits to the nt and nr databases (red) versus orphans contigs (grey).  
 252 (B) Relative abundance of cellular organism-like contigs based on the taxonomic assignment  
 253 of their closest homologs in the nr and nt databases at the kingdom level. Contig abundances  
 254 are calculated as transcripts per million (TPM).  
 255

256 Approximately half of the total contigs identified here could not be assigned using BLAST  
 257 approaches (Figure 3A), with prokaryotic organisms on average representing less than 10%  
 258 of assigned contigs (Figure 3B). However, the MMETSP0719 containing *C. curvisetus*  
 259 (Bacillariophyta) is enriched with co-infecting bacteria. According to the BLASTn/BLASTp  
 260 entries obtained for this sample, this seems largely due to the presence of the marine  
 261 alphaproteobacteria *Jannaschia*. This is to be expected as some algal species require the  
 262 presence of particular bacterial species to obtain essential nutrients<sup>34</sup>.

### 263 3.3 Distribution and prevalence of RNA viruses in MMETSP cultured strains

264 We found evidence for RNA viruses – that is, hits to the viral RdRp – in eight of the 19 major  
 265 groups of microalgae, without detectable virus/algal taxon specificity (Figure 4).



266

267 **Figure 4. Distribution of RNA virus groups identified in algae.** Only algal lineages  
 268 containing RNA virus RdRps are shown. Left, cladogram of the algal host lineages positive  
 269 for RNA viruses. Taxa for which no RNA viruses have previously been reported are  
 270 indicated in bold. Right, total counts of newly described RNA viral sequences in each algal  
 271 taxon (including viruses observed in several samples from the same taxa). \*First observation  
 272 of this virus taxon in the corresponding algal clade. The levi-like sequence that likely infects  
 273 a bacterial host was excluded.

274

275 The distribution of RNA viruses is highly heterogeneous among the microalgae studied here,  
 276 with a large representation in the Bacillariophyta, Dinophyceae and Haptophyceae, with only  
 277 few or no viruses in the other taxa analyzed here (Figure 4). It is important to note that the  
 278 number of viruses is strongly associated with the number of libraries analysed and thus likely  
 279 depicts a limit of detection imposed by small sample sizes in some groups (i.e. large numbers  
 280 of transcriptomes are available for the Bacillariophyta, Dinophyceae and Haptophyceae).

### 281 3.4 Positive-sense RNA viruses (ssRNA+)

282 Eleven of the 30 viruses discovered in this study show clear homology to three of the four  
 283 families that comprise the recently classified phylum *Lenarviricota* of ssRNA+ viruses: the

284 *Leviviridae*, the *Narnaviridae* and the *Mitoviridae* (Table 1). In all cases, levels of RdRp  
285 identities to the closest homologs were <60%, reflecting high levels of sequence divergence  
286 and leading us to propose that these 11 sequences are novel viral species (**Table 1**).

287

### 288 **3.4.1 *Narnaviridae*-like sequences**

289 Three RdRp-containing contigs – denoted Amphitrite narna-like virus, Poseidon narna-like  
290 virus and Halia narna-like virus – were related to the *Narnaviridae*, occupying diverse  
291 positions in a phylogeny of this virus family (Figure 2 and Figure 5).

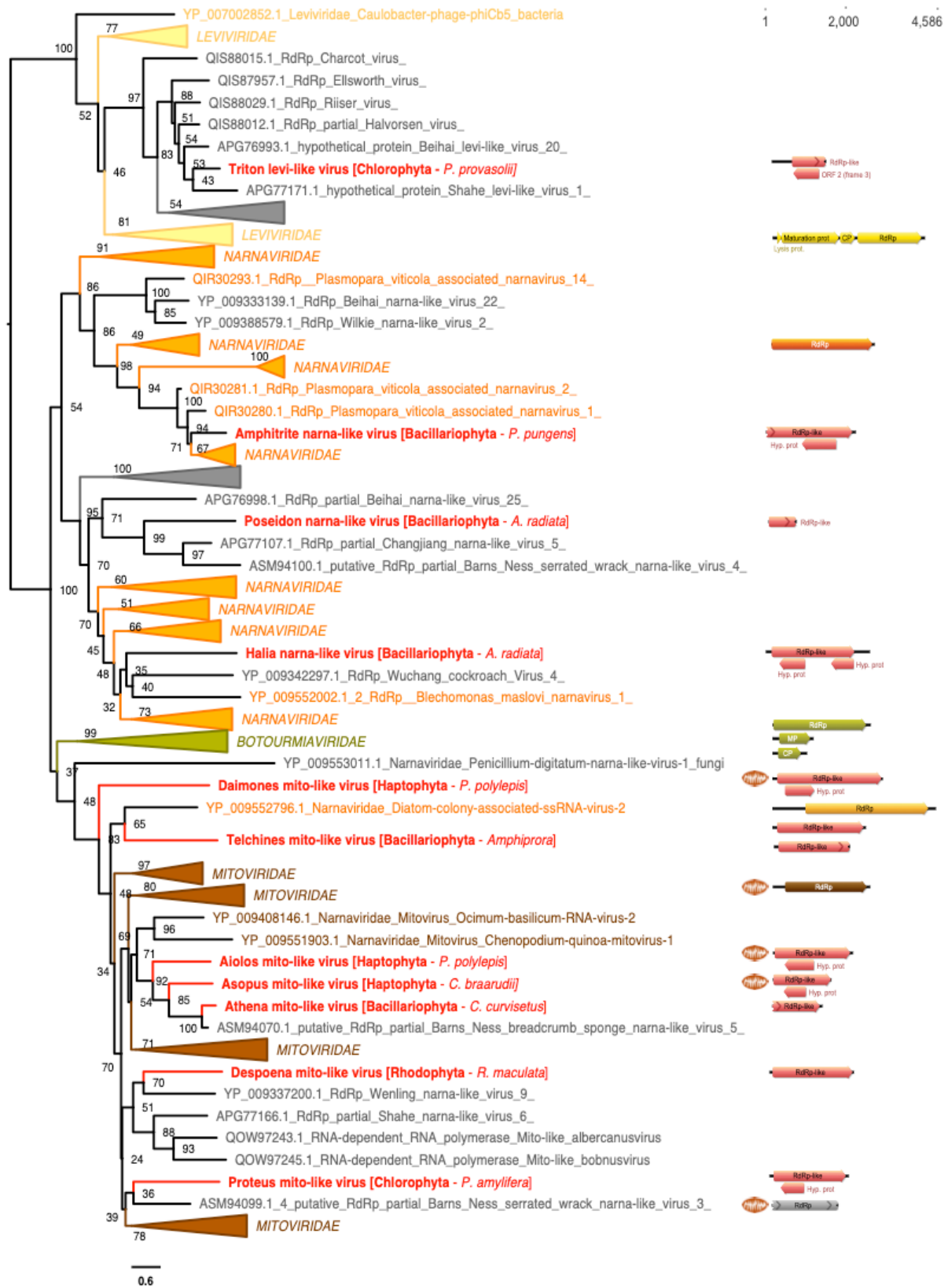
292 While the closest homologs of these narna-like viruses were identified in fungi,  
293 oomycete (protist) and marine arthropod samples, all three samples that contain these viruses  
294 are Bacillariophyta species (*A. radiata* and *P. pungens*) (Table 1, Figure 5). As their genome  
295 sequences share ~12% pairwise identity with other *Narnaviridae* we propose that Amphitrite  
296 narna-like virus, Poseidon narna-like virus and Halia narna-like virus represent novel species  
297 within the genus *Narnavirus*.

298

### 299 **3.4.2 *Mitoviridae*-like sequences**

300 Seven RdRp protein sequences, retrieved from diverse algae host lineages – Rhodophyta,  
301 Haptophyta, Chlorophyta and Bacillariophyta – are related to members of the *Mitoviridae*  
302 (Figure 5). According to their placement in the *Mitoviridae* phylogeny as well as their level  
303 of divergence to existing mitoviruses (Figure 5, Table 1), these seven new viruses are  
304 potential members of the genus *Mitovirus*. All these mitovirus-like sequences have similar  
305 genome organizations, with the exception of one putative mitovirus with a genome that  
306 seemingly encodes a single RdRp-containing ORF (Figure 5). It is also notable that the  
307 RdRp-encoding ORFs from Aiolos mito-like virus, Asopus mito-like virus and Daimones  
308 mito-like virus can only be predicted using the mitochondrial code (Figure 5).

309



310

311 **Figure 5. Phylogenetic position of the newly described RNA virus sequences in the**  
 312 **phylum *Lenarviricota*.** Left: ML phylogeny of the *Lenarviricota* RdRp (LG+F+R8 amino

313 acid substitution model). Newly described viruses are shown in red. Algal host taxa are  
314 specified in brackets. Branch labels = bootstrap support (%). The tree is mid-point rooted for  
315 clarity only. Right: genomic organisation of the newly described viruses (red), closest  
316 homologs and *Lenarviricota* RefSeq representatives: Cassava virus C (NC\_013111;  
317 *Botourmiaviridae*), *Saccharomyces* 23S RNA (NC\_004050; *Narnaviridae*), *Acinetobacter*  
318 phage AP205 (NC\_002700; *Leviviridae*), *Chenopodium quinoa* mitovirus 1 (NC\_040543;  
319 *Mitoviridae*). ORFs translated with the mitochondrial genetic code are marked a  
320 mitochondria icon. For clarity, some lineages were collapsed (a non-collapsed version of the  
321 tree is available as Supplementary Information).  
322

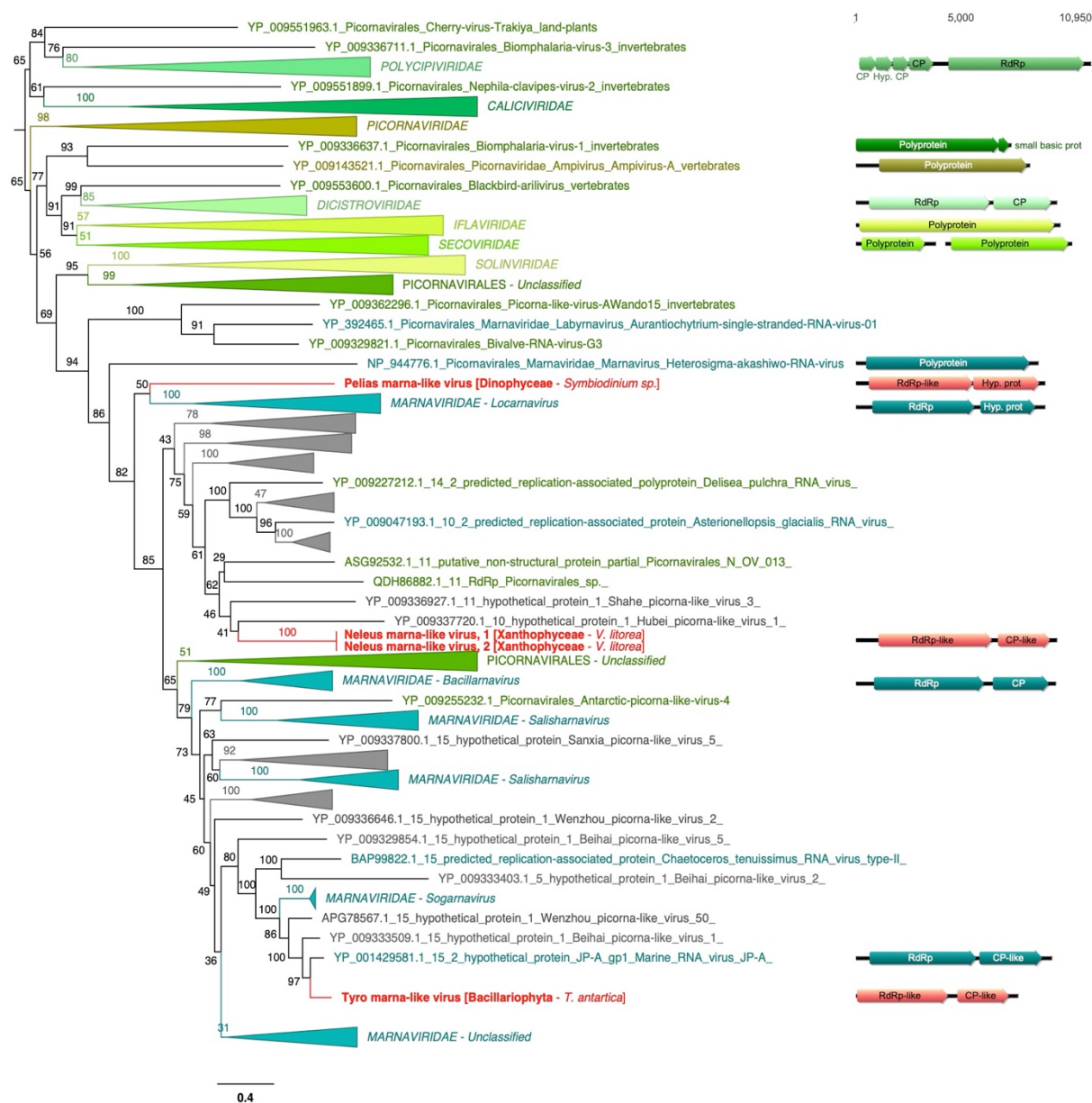
### 323 **3.4.3 *Leviviridae*-like sequences**

324 One viral RdRp-like hit, in the Chlorophyta species *Pycnococcus provasolii*, is related to  
325 some bacteria-infecting *Leviviridae* and based on the levels of sequence identity this likely  
326 constitutes a new genus in this family (Table 1). As there were some bacterial reads in the  
327 *Pycnococcus provasolii* samples (MMETSP1471) (Figure 3B), it is likely that this Triton  
328 levi-like virus sequence infects bacteria (Actinobacteria or Proteobacteria-like) also present in  
329 the culture rather than *Pycnococcus provasolii*.  
330

### 331 **3.4.4 *Picornavirales*-like sequences**

332 Three sequences – denoted Pelias marna-like virus, Neleus marna-like virus and Tyro marna-  
333 like virus – were identified in diverse cultures belonging to various taxa (Figure 4):  
334 *Symbiodinium sp.* (Dinophyceae), *V. litorea* (Xanthophyceae) and *T. antarctica*  
335 (Bacillariophyta). These viruses exhibit sequence similarity with ssRNA<sup>+</sup> viruses from the  
336 order *Picornavirales*. Specifically, they fell within the large algal associated family  
337 *Marnaviridae* (Figure 2C) and based on their respective positions in the phylogeny and the  
338 level of sequence divergence, Pelias marna-like virus could constitute a new genus in the  
339 *Marnaviridae*, while Neleus marna-like virus and Tyro marna-like virus are likely members  
340 of the genera *Kusarnavirus* and *Sogarnavirus*, respectively (Figure 6, Table 1). They also  
341 seem to share similar genome lengths and organizations as their closest relatives (Figure 6).





342

343 **Figure 6. Phylogenetic placement of the newly described RNA virus sequences in the**  
 344 **order *Picornavirales*.** Left, ML phylogeny of the *Picornaviruses* RdRp (assuming the  
 345 LG+F+R10 amino acid substitution model). Newly described viruses are indicated in red.  
 346 Algae host taxon and species are specified in brackets. Branch labels = bootstrap support (%).  
 347 The tree is mid-point rooted for clarity only. Right, genomic organisation of newly described  
 348 viruses (red), closest homologs and the following *Picornavirales* order RefSeq  
 349 representatives: *Solenopsis invicta* virus 2 (NC\_039236; *Polycipiviridae*), Porcine enteric  
 350 sapovirus (NC\_000940; *Caliciviridae*), Foot-and-mouth disease virus - type O (NC\_039210;  
 351 *Picornaviridae*), Acute bee paralysis virus (NC\_002548; *Dicistroviridae*), Infectious  
 352 flacherie virus (NC\_003781; *Iflaviridae*), Cowpea severe mosaic virus  
 353 (NC\_003544/NC\_003545; *Secoviridae*). For clarity, some lineages were collapsed (a non-  
 354 collapsed version of the tree is available as Supplementary Information).

### 355 3.5 Double-stranded (dsRNA) viruses

356 Almost a third of the RNA viruses newly reported here were related to dsRNA viruses of the  
357 family *Totiviridae* (Figure 2D). The single exception was the more divergent Charybdis toti-  
358 like virus, the exact placement of which within the order *Ghabrivirales* was unclear as it  
359 occupied a basal position in the phylogenetic tree and showed only low levels of sequence  
360 similarity to related viruses (~30% at RdRp protein level) (Figure 7, Table 1).

361 Aloadae toti-like virus, found in Haptophyta *Isochrysis sp.*, groups with the protist-  
362 associated *Giardiavirus* genus of the *Totiviridae*, and more surprisingly with Keenan toti-like  
363 virus recently identified in ectoparasitic flies (Figure 7), although with very high levels of  
364 sequence divergence (Table 1). Similarly, Chrysaor toti-like virus, Laestrygon toti-like virus  
365 and Arion toti-like virus, retrieved from Bacillariophyta, Chromerid and Dinophyceae,  
366 respectively, form a clade with *Totiviridae*-like sequences identified in either marine  
367 arthropods or oomycete protists (Figure 7). While these likely constitute a newly genus  
368 within the *Totiviridae*, their host remains uncertain. Antaeus toti-like virus, retrieved from the  
369 Bacillariophyta *T. antarctica*, groups with *Pythium polare* RNA virus 1 that infects the  
370 oomycete *Pythium polare*, confirming the presence of a polar stramenopile clade in the  
371 *Totiviridae*. Otus toti-like virus, identified in the Rhodophyta *R. marinus*, clusters (51%  
372 sequence identity) with the *Delisea pulchra totivirus* identified in the Rhodophyta (Figure 7).

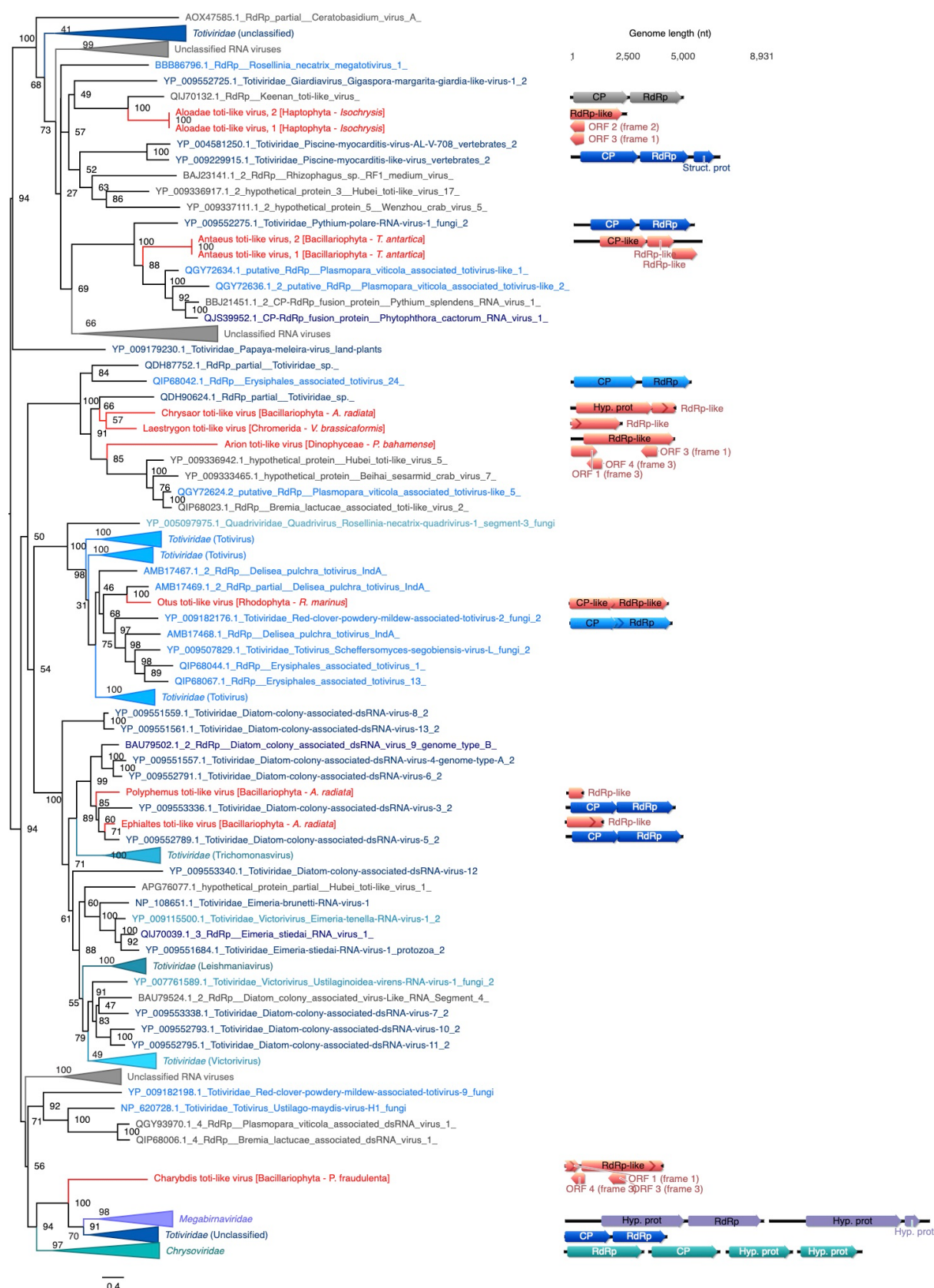
373 Two additional toti-like viruses – Polyphemus toti-like virus and Ephialtes toti-like  
374 virus – were identified in *A. radiata* (Bacillariophyta) and, together with the diatom colony  
375 associated dsRNA viruses, form a new dsRNA viral clade, and likely genus, specifically  
376 associated with Bacillariophyta (diatoms) (Figure 7).

377 Strong similarities in genome organization were observed between the Otus toti-like  
378 virus and Antaeus toti-like virus and their toti-like homologs, with a potential single segment  
379 encoding a coat protein (CP) in 5' and a RdRp in 3' (Figure 7). As Charybdis toti-like virus,

380 Chrysaor toti-like virus, Laestrygon toti-like virus, Arion toti-like virus, Polyphemus toti-like  
381 virus and Ephialtes toti-like virus all had partial genomes we were unable to determine their  
382 genomic organization, aside from the observation that they are likely unsegmented as they  
383 fall within the unsegmented *Totiviridae*. Unfortunately, such an assumption cannot be made  
384 for Charybdis toti-like virus, because of its basal position within the *Ghabrivirales*.

385 We identified six RdRp hits to members of the *Durnavirales* order of dsRNA virus  
386 (Figure 2C). With the exception of Aethusa amalga-like virus and Aegean partiti-like virus,  
387 their exact position within the six families that comprise this order (*Partitiviridae*,  
388 *Hypoviridae*, *Picobirnaviridae* and *Amalgaviridae*) is unclear due to their basal phylogenetic  
389 position (Figure 8). Moreover, these sequences seemingly have no association with specific  
390 microalgal groups, being observed in species of Rhodophyta, Bolidophyceae,  
391 Bacillariophyta, Chlorophyta and Dinophyceae (Figure 4). Aethusa amalga-like virus,  
392 retrieved from the Rhodophyta *R. marinus*, is clearly related to the *Amalgaviridae* (Figure 2  
393 and Figure 8) and displays a moderate level of sequence divergence (43% identity in the  
394 RdRp) with *Zygosaccharomyces bailii virus Z* identified in fungi (Table 1). Whether this  
395 constitutes a new genus within the *Amalgaviridae* remains to be determined.

396 Three other viruses, Benthescyme durna-like virus, Herophile durna-like virus and  
397 Cymopoleia durna-like virus, were related to the Amalga-like lacheneauvirus and Amalga-  
398 like chassivirus, both previously identified in cultures of *Ostreobium* sp. (Chlorophyta), and  
399 that fell between the *Amalgaviridae* and *Partitiviridae* families in our phylogenetic analysis  
400 (Figure 8). The genomic sequences for Benthescyme durna-like virus, Herophile durna-like  
401 virus and Cymopoleia durna-like virus were likely partial such that their organization,  
402 particularly whether they comprise one of two segments, could not be established (Figure 8).



403

404 **Figure 7. Phylogenetic position of the newly described RNA virus sequences among the**  
 405 **Ghabrivirales.** Left, ML phylogeny of the *Ghabrivirales* RdRp (assuming the LG+F+R10  
 406 amino acid substitution model). Newly described viruses are indicated in red. Algae host  
 407 taxon and species are specified in brackets. Branch labels = bootstrap support (%). The tree is



408 mid-point rooted for clarity only. Right, genomic organisation of the newly described viruses  
409 (red), closest homologs and the following representative *Ghabrivirales*: Rosellinia necatrix  
410 megabirnavirus 1/W779 (NC\_013462/NC\_013463; *Megabirnaviridae*), Tuber aestivum virus  
411 1 (NC\_038698; *Totiviridae*), Penicillium chrysogenum virus  
412 (NC\_007539/NC\_007540/NC\_007541/NC\_007542; *Chrysoviridae*). For clarity, some  
413 lineages were collapsed (a non-collapsed version of the tree is available as Supplementary  
414 Material).  
415

416 Aegean partiti-like virus falls in the *Partitiviridae*, grouping with the Partiti-like  
417 lacotivirus, Partiti-like allasinovirus, Partiti-like Adriusvirus and Bryopsis cinicola  
418 chloroplast dsRNA (BDRC): these are all *Partitiviridae* and associated with Ulvophyceae  
419 algae (Figure 8). The presence of Aegean partiti-like virus in *Tetraselmis chuii* (Chlorophyta)  
420 strongly supports the existence of a Chlorophyta-infecting partiti-like viral genus. Assuming  
421 a homologous genome organization, the genome of Aegean partiti-like virus would comprise  
422 a single segment encoding a RdRp in its 5' region as well as a hypothetical protein,  
423 potentially a coat protein, in the 3' region. Whether Aegean partiti-like virus is associated  
424 with the host chloroplast remains uncertain. Finally, Ourea durna-like virus is highly  
425 divergent and falls basal to the bi-segmented *Partitiviridae* (Figure 8). However, considering  
426 the length and the single ORF organization of the partial genomic sequence retrieved, it is  
427 likely that a second segment encoding a CP may not have been detected by BLAST due to  
428 very high levels of sequence divergence.



429

430 **Figure 8. Phylogenetic positions of the newly described RNA viruses among the**  
 431 ***Durnavirales*.** Left, ML phylogeny of the *Durnavirales* RdRp (assuming the LG+F+R8  
 432 amino acid substitution model). Newly described viruses are indicated in red. Algae host  
 433 taxon and species are specified in brackets. Branch labels = bootstrap support (%). The trees  
 434 are mid-point rooted for clarity only. Right, genomic organisation of newly-discovered  
 435 viruses (red), closest homologs and the following Partiti-picobirna super-clade  
 436 representatives: *Zygosaccharomyces bailii* virus Z (NC\_003874; *Amalgaviridae*),  
 437 *Cryphonectria hypovirus 2* (NC\_003534; *Hypoviridae*), *Chicken picornavirus* (NC\_003534/  
 438 NC\_040438; *Picobirnaviridae*), *Fig cryptic virus* (NC\_015494/NC\_015495;  
 439 *Deltapartitivirus*), *Discula destructiva virus 1* (NC\_002797/NC\_002800;  
 440 *Gammapartitivirus*), *Ceratocystis resinifera virus 1* (NC\_010755/NC\_010754;  
 441 *Betapartitivirus*), *White clover cryptic virus 1* (NC\_006275/NC\_006276; *Alphapartitivirus*).  
 442 ORFs translated with the plastid genetic code are labelled with a green plastid. For clarity,  
 443 some lineages were collapsed (a non-collapsed version of the tree is available as  
 444 Supplementary Information).

#### 445 Negative-sense viruses (ssRNA-)

446 A novel RdRp sequence, *Susy yue*-like virus, was identified in the *Pseudo-nitzschia heimii*  
 447 (*Bacillariophyta*) culture. This virus clusters among the ssRNA- *Haploviricotina*, falling

448 between the *Qinviridae* and the *Yueviridae* families (Figure 9). Considering the length of the  
 449 RdRp segment and the bi-segmented genome organization of related members of the  
 450 *Qinviridae* and *Yueviridae* (Figure 9), it is highly likely that the Susy yue-like virus genome  
 451 is partial. In similar manner to the *Qinviridae*, Susy yue-like virus has an IDD sequence motif  
 452 instead of the common GDD triad in the catalytic core of its RNA virus replicase (RdRp),  
 453 although the functional implications of this alternative motif are unclear.



454  
 455 **Figure 9. Position of the newly described RNA virus in the phylum *Haploviricotina*.** Left,  
 456 ML phylogeny of the *Haploviricotinia* RdRp (employing the LG+F+R10 amino acid  
 457 substitution model). The virus newly described here is shown in red. Algae host taxon and  
 458 species are specified in brackets. Branch labels = bootstrap support (%). The tree is mid-point  
 459 rooted for clarity only. Right, genomic organisation of the newly described virus (red) and the  
 460 following homologs representatives: Shahe yuevirus-like virus 1 (NC\_033289/NC\_033290;  
 461 *Yueviridae*), Beihai sesarmid crab virus 4 (NC\_032274/NC\_032272; *Qinviridae*), Blueberry  
 462 mosaic associated virus (NC\_033754/NC\_036634/NC\_036635; *Aspiviridae*). For clarity,  
 463 some lineages were collapsed (a non-collapsed version of the tree is available as  
 464 Supplementary Information).

#### 465 **Detection of divergent RNA viruses based on RdRp motifs and structural features**

466 The microalgal transcriptomes sequenced as part of the MMETSP likely contain viruses that  
 467 are highly divergent in sequence, sharing only limited sequence similarity to those currently  
 468 available and hence challenging to detect using BLAST-based methods. To identify RNA  
 469 viruses at lower levels of homology, we conducted an extensive analysis utilising RdRp

470 protein functional motifs and structural features on all the BLAST-unannotated sequences:  
471 this accounted for 10-34% of the total predicted ORFs of at least 200 amino acid residues in  
472 length (Figure S2).

473 A very large proportion of the sequences retained from our combined RdRp-based  
474 HMM, InterproScan analysis were false-positive hits as they were either confidently detected  
475 as eukaryotic-like sequences using Phyre2 or were too distant to be safely considered as an  
476 RdRp (i.e. unreliable alignment and no detection of RdRp catalytic motifs) (Table S3).  
477 However, five RdRp-like candidates were retained from the manual curation steps. While no  
478 robust RdRp-like signal could be detected using Phyre2 (i.e. prediction confidence scores  
479 below 90%) (Table S3), the presence of a significant HMM-detected homology with the  
480 PROSITE PS50507 profile (i.e. RdRp of ssRNA+ virus catalytic domain profile; Table S2)  
481 enabled us to further analyze these candidates as potential RdRp sequences.

482 Four of these five RdRps came from the genus *Bigelowiella*, and three  
483 (MMETSP0045\_DN12861, MMETSP1054\_DN18666 and MMETSP1052\_DN19445)  
484 shared high identity levels (>90% at both protein and nucleotide levels), while  
485 MMETSP1359\_DN14104 shared only 70% identity (Table 2). Although the PROSITE  
486 PS50507 profiles were built from ssRNA+ RdRp sequences, the IDD C-motif exhibited by  
487 these four RdRp-like candidates is found in the ssRNA- *Qinviridae*-like viruses as well as the  
488 new Susy yue-like virus found in *Pseudo-nitzschia heimii* (MMETSP1423). However, the  
489 nucleotide sequences of these RdRp-encoding contig candidates exhibited a strong match (e-  
490 value < 1E-90) with a genome contig (BIGNA scaffold\_41\_Cont1731) from the *Bigelowiella*  
491 *natans* genome (GCA\_000320545.1). Hence, rather than representing an exogenous RNA  
492 virus, the distant RdRp hit in this case most likely constitutes an endogenous viral element  
493 (EVEs) indicative of a past, and likely ancient, infection event.

494



495 **Table 2. RdRp-like hits retrieved from the HMM-profile and Phyre2 analyses.** Presence  
 496 of the A, B and C motifs are noted along with the sequence of the C-motif.  
 497

Contig ID	Taxon	RdRp profile	E-value	A	B	C	Phyre2 confid %	%ID	Hit info
MMETSP1359_DN14104_c0_g1_i1_len843_1	<i>Bigelowiella longifila</i> (Cercozoa)	PS50507	6.0E-07	Yes	?	IDD	64.2	16	PDB header: transferase
MMETSP0045_DN12861_c0_g1_i1_len664_1	<i>Bigelowiella natans</i> (Cercozoa)	PS50507	8.7E-06	Yes	?	IDD	40.7	24	DNA/RNA polymerases
MMETSP1054_DN18666_c0_g1_i1_len657_1	<i>Bigelowiella natans</i> (Cercozoa)	PS50507	8.9E-06	Yes	?	IDD	41.6	24	DNA/RNA polymerases
MMETSP1052_DN19445_c0_g1_i1_len738_1	<i>Bigelowiella natans</i> (Cercozoa)	PS50507	1.0E-05	Yes	?	IDD	40.4	24	DNA/RNA polymerases
MMETSP0202_DN4292_c0_g1_i1_len814_1	<i>Karenia brevis</i> (Dinophyceae)	PS50507	4.6E-05	Yes	?	GDT	56.7	17	PDB header: hydrolase

498  
 499 In the case of the remote RdRp-like signal in MMETSP0202\_DN4292, no GDT sequence at  
 500 motif C could be identified in an expansive RdRp data set<sup>35</sup>. Hence, it is unclear if  
 501 MMETSP0202\_DN4292 is a true viral RdRp or a false-positive hit.

## 502 4. Discussion

503 To the best of our knowledge we report the largest survey of RNA viruses in microalgal  
 504 curated cultures. With the discovery of 30 new and divergent viruses, 29 of which are likely  
 505 to infect algae species in which no viruses have previously been reported, this study greatly  
 506 extends our knowledge of the microalgae RNA virosphere. More broadly, this work  
 507 demonstrates the potential of protists to be major reservoirs of novel RNA viruses.

508 Despite the viral diversity documented, only 6% (33 of 570) of the transcriptomes  
 509 analysed here contained evidence of an RNA virus, far lower than equivalent meta-  
 510 transcriptomic studies of single organisms<sup>36-38</sup>. The use of purified cultures is expected to

511 reduce the number of viruses compared to direct environmental samples, preventing the  
512 sequencing of co-circulating viruses as well as those infecting other microorganisms in the  
513 environment. However, this relative paucity of RNA viruses could also reflect  
514 methodological limitations. First, the lack of rRNA depletion in the library processing leads a  
515 concomitant reduction in the number of non-rRNA transcripts, including those from viruses.  
516 Indeed, most of the viruses reported here display very low transcript abundance, suggesting  
517 that additional RNA viruses may have been undetected due to poor sequencing coverage.  
518 Second, the limited number of viruses identified is likely to reflect the high levels of  
519 sequence divergence expected for protist viruses compared to those currently available in  
520 sequence databases. Indeed, many of the viruses identified in this study share less than 30-  
521 40% sequence identity, toward what might be the limit of a viable BLAST-based analysis.  
522 Hence, this study has been conducted at the boundaries of the detectable virosphere, with a  
523 myriad of more divergent viruses yet to be discovered.

524

#### 525 **4.1 RNA virus are widespread among lineages of unicellular algae**

526 Our knowledge of RNA viruses associated with microalgae is scarce. The small number  
527 reported so far are mostly associated with a specific subset of algal species from the  
528 Bacillariophyta and Chlorophyta, ignoring the wide diversity of microalgae (Figure 1A). We  
529 extend this diversity by revealing, for the first time, RNA viruses (i.e. RdRp sequences) in the  
530 Haptophyta, Chromeraceae (Alveolates), as well as in the Stramenopiles Xanthophyceae and  
531 Bolidophyceae. We also identified new virus-algae clade associations. For example, we  
532 present the first observation of *Picornavirales*, *Ghabrivirales* (*Totiviridae*) and *Durnavirales*  
533 (*Partitiviridae*) in Dinophyceae cultures, *Lenarviricota* and *Durnavirales* in Rhodophyta  
534 cultures, and *Durnavirales* in Bacillariophyta cultures. Importantly, our study also constitutes

535 the first observation of a *Muvirales*-like ssRNA- virus in a Bacillariophyta sample, perhaps  
536 only the second negative-sense RNA virus identified in microalgae.

537 With the exception of *Symbiodinium* sp. for which a ssRNA+ virus was previously  
538 reported<sup>39,40</sup>, all the viruses described in this study represent the first observation of an RNA  
539 virus in each respective host species. In addition, none of the 73 microalgal viruses reported  
540 previously were identified here. More generally, the distribution of RNA viruses obtained in  
541 this study, comprising ssRNA+, ssRNA- and dsRNA viruses, varies considerably between  
542 taxa and likely reflects sampling bias rather than a host specificity of RNA virus infection.  
543 These factors might have contributed to the lack of viral identification in poorly investigated  
544 and divergent taxa such as Euglena, Glaucophytes and Cryptophytes. Further studies with  
545 particular emphasis on these taxa are clearly required.

546 The first observation of an ssRNA- virus in a Bacillariophyta, together with the  
547 previous observation of a bunya-like virus reported in the distantly-related  
548 Chloroarachniophyte *C. reptans* (Cercozoa) and bunya-like siRNAs in brown algae  
549 (Phaeophyta)<sup>41</sup>, again demonstrates that microalgae can be infected with negative-sense RNA  
550 viruses. Interestingly, the related *Qinviridae* and *Yueviridae* have been exclusively identified  
551 from meta-transcriptomic studies conducted on marine arthropods holobionts, such that algae  
552 could constitute the true hosts for most of these viruses<sup>42,43</sup>. Undoubtedly, the presence of  
553 ssRNA- viruses in microbial eukaryotes needs to be further characterized.

554

#### 555 **4.2 *Narnaviridae*-like and *Mitoviridae*-like viruses are common in microalgal cultures**

556 A third of the viruses reported here were from the order *Lenarviricota* that includes the  
557 *Narnaviridae* and *Mitoviridae* and often characterised by a single RdRp ORF<sup>44</sup>. Although  
558 they were initially thought to be restricted to fungi, these seemingly simple RNA viruses  
559 appear to be more widespread than initially thought. Indeed, *Narnaviridae*-like viruses have

560 recently been associated with a wide range of protist organisms, including protozoan  
561 parasites like *Plasmodium vivax*<sup>45–48</sup> and the oomycete *Phytophthora infestans*<sup>49</sup>, while narna-  
562 like viruses have also been detected in diatoms<sup>50</sup>. Similarly, the *Mitoviridae* were considered  
563 as exclusively infecting fungi, until the recent discovery of the *Chenopodium quinoa*  
564 mitovirus 1 in a plant<sup>51</sup> and mito-like viruses in the Chlorophyta *Osteobium* sp.<sup>52</sup> led their  
565 host range to be re-evaluated. The three new narna-like viruses in Bacillariophyta discovered  
566 here, as well as the proposal of seven new mitovirus-like species in algal lineages as diverse  
567 as Haptophyta, Bacillariophyta, Rhodophyta and Chlorophyta, provides further evidence for  
568 the ubiquity of these viruses in protists.

569       Whether all the mitoviruses documented here are associated with the mitochondria, as  
570 typical of the *Mitoviridae*, remains to be determined. In addition, while the unique RdRp-  
571 encoding segment has already been demonstrated as sufficient for virus infectivity, recent  
572 studies have suggested the presence of an additional segment, without an assigned function,  
573 in both *Leptomonas seymouri* and *Plasmodium vivax*<sup>45,48</sup>. Whether the viruses newly  
574 described here have unsegmented or bipartite genomes remains to be determined. Most of the  
575 *Lenarviricota*-like sequences described here display ambigrammatic ORFs, with their reverse  
576 strand encoding additional ORFs. This feature has already been reported in narnaviruses and  
577 could represent a potential solution to extreme genome compaction<sup>53–55</sup>.

578       The growing evidence for the extended host range of both *Narnaviridae* and  
579 *Mitoviridae* beyond the fungal clades has important consequences in our knowledge of the  
580 early events in the evolution of eukaryotic RNA viruses. Indeed, the ubiquity of *Mitoviridae*  
581 and *Narnaviridae* in eukaryotes is compatible with the protoeukaryotic origins of these  
582 viruses and the bacterial *Leviviridae*, such that they are relics of a past endosymbiont  
583 infection of a eukaryotic ancestor. Accordingly, cytoplasmic *Narnaviridae* would have  
584 escaped from mitochondria to the more RNA hospitable cytosol<sup>3</sup>. In addition, *Narnaviridae*

585 and *Mitoviridae* are not associated with cellular membranes<sup>56</sup>, which could also reflect their  
586 ancient origin from a protoeukaryote ancestor without cellular compartments.

587

#### 588 **4.3 The extension of the *Marnaviridae* to new algal taxa**

589 Most of the algal RNA viruses described to date belong to the order *Picornavirales*<sup>8</sup>,  
590 including the *Marnaviridae*. Currently, the *Marnaviridae* comprise 20 species, distributed  
591 among seven genera based on their capsid similarities. Notably, all these viral species are  
592 associated with marine samples or algae cultures<sup>57</sup>. The three picorna-like viruses newly  
593 identified in this study fell within the *Marnaviridae*. Despite similar genome organizations,  
594 these three viruses have relatively high levels of divergence from known *Marnaviridae*, in  
595 turn suggesting that the *Marnaviridae* diversity has only been sparsely sampled. This  
596 diversity will very likely increase with the sequencing of phytoplankton cells. While the  
597 detection of *Neleus marna*-like virus and *Tyro marna*-like virus in Bacillariophyta and  
598 Xanthophyceae could reflect the specificity of *Sogarnavirus* and *Kusarnavirus* to  
599 Stramenopile algae, the first detection of a *Marnaviridae*-like virus in the Dinophyceae  
600 species *Symbiodinium* sp. suggests that the host range of this algal-infecting viral family is  
601 not restricted to Stramenopile eukaryotes.

602

#### 603 **4.4 The ancestry of the *Durnavirales* and *Ghabrivirales* dsRNA viruses**

604 Approximately half of the RNA viruses identified in this study are related to the *Totiviridae*  
605 (*Ghabrivirales*) and *Partitiviridae* (*Durnavirales*) families of dsRNA virus. The *Totiviridae*  
606 currently comprises 28 formally-assigned species divided into five genera<sup>32,58</sup>. Interestingly,  
607 *Totiviridae* are exclusively associated with unicellular eukaryotes, with two of the five  
608 *Totiviridae* genera associated with latent fungal infections (*Totivirus* and *Victorivirus*), while

609 *Trichomonasvirus*, *Giardiavirus* and *Leishmanivirus* have been associated with protozoan  
610 parasite infections<sup>32</sup>.

611 Each of the new *Totiviridae*-like sequences identified here were retrieved from a  
612 range of algal hosts spread among diverse branches of the microbial eukaryote tree  
613 (Bacillariophyta, Dinophyceae, Haptophyceae, Rhodophyta and Chromeraceae). Hence, as  
614 with the *Marnaviridae*, the diversity of the *Totiviridae* has likely been greatly  
615 underestimated. In addition, some of the novel viruses identified cluster with totiviruses  
616 previously reported in Bacillariophyta diatoms<sup>59,60</sup> and the Rhodophyta *Delisea pulchra*<sup>61</sup>.  
617 These observations support the existence of a Bacillariophyta and a Rhodophyta-infecting  
618 clade in the genus *Totivirus* that will need to be confirmed with studies of additional species.  
619 It was also notable that other toti-like viruses identified here cluster with viruses found in  
620 non-algal hosts, such as invertebrates (ticks, crustaceans), fungi and protozoan parasites.  
621 While host mis-annotations cannot be formally excluded, the presence of *Totiviridae* in  
622 protozoan parasites, fungi and algae could signify that the host range of the *Totiviridae* is far  
623 larger than appreciated.

624 Six dsRNA-like new viruses identified here show clear homology with those of the  
625 order *Durnavirales*, including the *Partitiviridae* and the *Amalgaviridae* that comprise bi-  
626 segmented and unsegmented dsRNA viruses, respectively. The *Partitiviridae* are classified  
627 into five genera and mainly associated with plants and fungi, although more recently with  
628 oomycetes<sup>62</sup> and to Apicomplexa<sup>63</sup>. The *Amalgaviridae* comprise two genera associated with  
629 either fungi (*Zybavirus* genus) or land plants (*Amalgavirus* genus)<sup>58,64</sup>. In addition to the  
630 recent association of newly described partiti- and amalgavirus-like viruses in the microalgae  
631 *Ostreobium* sp. (Chlorophyta)<sup>52</sup>, our identification of these novel and divergent  
632 *Durnavirales*-like viruses in several distant algae taxa again suggests that host range for this  
633 viral order has been underestimated.

634

#### 635 **4.5 Are cryptic viruses a common feature of unicellular eukaryotes?**

636 RNA viruses causing host cell lysis and hence mortality are commonly reported<sup>65</sup>, with an  
637 emblematic example being the lysis of the harmful algal bloom-forming diatoms, haptophytes  
638 and dinoflagellates, leading to bloom collapse<sup>66,67</sup>. Although we did not aim to assess the  
639 phenotypic effects of viral infection on algal hosts, it is noticeable that most of the viruses  
640 identified here were related to the *Totiviridae*, *Partitiviridae*, *Mitoviridae* and *Narnaviridae*,  
641 all previously reported as associated with cryptic and persistent infections<sup>32</sup>. This is  
642 consistent with the design of the MMETSP study that would tend to identify non-pathogenic  
643 viruses. It is also in accordance with the growing evidence that a non-neglectable component  
644 of RNA virus-host associations are symptomless or even beneficial to their host, with  
645 potentially important evolutionary implications<sup>68,69</sup>.

646

#### 647 **4.6 Limitations to virus discovery and inferring virus-host relationships**

648 A key element of this study was use of mono-strain cultures, which were axenic whenever  
649 possible, enabling more accurate virus-host assignments. While Bacteria, and to a lesser  
650 extent, Archaea, were present in the non-axenic cultures, the placement of most of the newly  
651 described viruses within eukaryotic-infecting viral families clearly supports their association  
652 with algae. Despite this, some of the newly-described viruses were associated with viral  
653 lineages traditionally associated with fungal or metazoan hosts. This likely reflects the lack of  
654 representation of microalgal viruses in current sequence databases or a mis-annotation to  
655 secondary metazoan host, particularly given the recent efforts to describe the fungal  
656 virome<sup>70-73</sup>. Similarly, many of the newly identified viruses share homology with viruses  
657 identified in metagenomics studies on marine invertebrates<sup>36</sup>. It is widely established that  
658 such similarities to holobiont virome studies should be treated with caution, as the viruses

659 reported could in fact be infecting symbionts, eukaryotic parasites, or bacteria that are also  
660 present in these samples<sup>3</sup>. Marine invertebrate organisms are also important ocean filters and  
661 virus removers<sup>74</sup>, again compatible with the idea that at least some of the viruses identified  
662 here may infect other marine organisms.

663 We also attempted to identify more distant RNA viruses using a protein profile and  
664 structural-based approach. However, no remote RNA virus signals could be confidently  
665 detected using this method, although a distant endogenous viral element in *Bigelowiella* was  
666 identified. While the *de novo* prediction of protein 3D structures has experienced major  
667 improvements over the last decade<sup>75</sup>, revealing robust homology strongly relies on structural  
668 comparisons and modelling based on pre-existing structures<sup>22</sup>. Critically, however, only a  
669 very limited number of non-human viruses are available among the viral proteins deposited in  
670 the Protein Data Bank. This poor representativeness of protein structures is a major roadblock  
671 in the ability to detect highly divergent RdRps. Indeed, a better characterization of RdRp  
672 structures combined with the enrichment of RdRp motif and profile databases will help  
673 counter the challenge posed by the high levels of sequence divergence in protist samples and  
674 the concomitant loss of detectable evolutionary signals. In addition, the high percentage of  
675 false positives in the HMM analysis highlights the need to increase and optimize the  
676 sensitivity and stringency of such methods.

677 While our study significantly extends our knowledge of RNA virus diversity among  
678 unicellular eukaryotes, experimental confirmation is needed to formally assign such viruses  
679 to their specific microalgae hosts and to assess the impact of viral infection on host biology.  
680 Perhaps more importantly, additional effort is needed to detect the signal of remote sequence  
681 homology in the highly divergent RNA viruses that are likely commonplace in protists.

682

683 **Acknowledgments**



684 SM thanks the Moore Foundation for funding her involvement in the MMETSP project. ECH  
685 is funded by an Australian Research Council Australian Laureate Fellowship (FL170100022).

686

687 **Data availability**

688 All viral genomes and corresponding sequences detected in this study will be deposited in the

689 NCBI GenBank and SRA upon the acceptance. The accessions ID will be listed in Table 1.

690 **References**

- 691 1. Wigington CH, Sonderegger D, Brussaard CPD, Buchan A, Finke JF, Fuhrman JA *et al.*  
692 Re-examination of the relationship between marine virus and microbial cell abundances.  
693 *Nat Microbiol* 2016; **1**: 15024.
- 694 2. Suttle CA. Marine viruses-major players in the global ecosystem. *Nat Rev Microbiol*  
695 2007; **5**: 801–812.
- 696 3. Dolja V V., Koonin E V. Metagenomics reshapes the concepts of RNA virus evolution  
697 by revealing extensive horizontal virus transfer. *Virus Res* 2018; **244**: 36–52.
- 698 4. Burki F, Roger AJ, Brown MW, Simpson AGB. The new tree of eukaryotes. *Trends Ecol*  
699 *Evol* 2020; **35**: 43–55.
- 700 5. Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, Berney C *et al.* CBOL protist working  
701 group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms.  
702 *PLoS Biol* 2012; **10**: e1001419.
- 703 6. Raoult D, Forterre P. Redefining viruses : lessons from Mimivirus. *Nat Rev Microbiol*  
704 2008; **6**: 315–319.
- 705 7. Tai V, Lawrence JE, Lang AS, Chan AM, Culley AI, Suttle CA. Characterization of  
706 Harnav, a single-stranded RNA virus causing lysis of *Heterosigma akashiwo*  
707 (*Raphidophyceae*). *J Phycol* 2003; **39**: 343–352.
- 708 8. Short SM, Staniewski MA, Chaban Y V, Long AM, Wang D. Diversity of viruses  
709 infecting eukaryotic algae. *Curr Issues Mol Biol* 2020; **39**: 29–62.
- 710 9. Brum JR, Cesar Ignacio-Espinoza J, Roux S, Doulcier G, Acinas SG, Alberti A *et al.*  
711 Patterns and ecological drivers of ocean viral communities. *Science* 2015; **348**: 25.
- 712 10. Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A *et al.*  
713 Marine DNA viral macro- and microdiversity from pole to pole. *Cell* 2019; **177**: 1109–  
714 1123.

- 715 11. Steward GF, Culley AI, Mueller JA, Wood-Charlson EM, Belcaid M, Poisson G. Are we  
716 missing half of the viruses in the ocean? *ISME J* 2013; **7**: 672–679.
- 717 12. Wolf YI, Silas S, Wang Y, Wu S, Bocek M, Kazlauskas D *et al.* Doubling of the known  
718 set of RNA viruses by metagenomic analysis of an aquatic virome. *Nat Microbiol* 2020;  
719 **5**: 1–9.
- 720 13. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB *et al.* Virus  
721 taxonomy in the age of metagenomics. *Nat Rev Microbiol* 2017; **15**: 161–168.
- 722 14. Nissimov JI, Campbell CN, Probert I, Wilson WH. Aquatic virus culture collection: an  
723 absent (but necessary) safety net for environmental microbiologists. *Appl Phycol* 2020;  
724 doi: 10.1080/26388081.2020.1770123
- 725 15. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA *et al.* The  
726 marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating  
727 the functional diversity of eukaryotic life in the oceans through transcriptome  
728 sequencing. *PLoS Biol* 2014; **12**: e1001889.
- 729 16. Johnson LK, Alexander H, Brown CT. Re-assembly, quality evaluation, and annotation  
730 of 678 microbial eukaryotic reference transcriptomes. *Gigascience* 2019; **8**: 1–12.
- 731 17. Swart EC, Serra V, Petroni G, Nowacki M. Genetic codes with no dedicated stop codon:  
732 context-dependent translation termination. *Cell* 2016; **166**: 691–702.
- 733 18. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND.  
734 *Nat Methods* 2015; **12**: 59–60.
- 735 19. Giangaspero M. Pestivirus species potential adventitious contaminants of biological  
736 products. *Trop Med Surg* 2013; **1**:1000153.
- 737 20. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC *et al.* The Pfam  
738 protein families database in 2019. *Nucleic Acids Res* 2019; **47**: D427–D432.
- 739 21. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol* 2011; **7**: e1002195.

- 740 22. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for  
741 protein modeling, prediction and analysis. *Nat Protoc* 2015; **10**: 845–858.
- 742 23. Venkataraman S, Prasad BVLS, Selvarajan R. RNA Dependent RNA polymerases:  
743 insights from structure, function and evolution. *Viruses* 2018; **10**: 76.
- 744 24. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S *et al*. Geneious  
745 Basic: An integrated and extendable desktop software platform for the organization and  
746 analysis of sequence data. *Bioinformatics* 2012; **28**: 1647–1649.
- 747 25. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Meth* 2012;  
748 **9**: 357–359.
- 749 26. Pettersson JH-O, Ellström P, Ling J, Nilsson I, Bergström S, González-Acuña D *et al*.  
750 Circumpolar diversification of the *Ixodes uriae* tick virome. *PLOS Pathog* 2020; **16**:  
751 e1008759.
- 752 27. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7:  
753 improvements in performance and usability. *Mol Biol Evol* 2013; **30**: 772–780.
- 754 28. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A fast and effective  
755 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*  
756 2015; **32**: 268–274.
- 757 29. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermin LS. ModelFinder:  
758 fast model selection for accurate phylogenetic estimates. *Nat Methods* 2017; **14**: 587–  
759 589.
- 760 30. Minh BQ, Nguyen MAT, Von Haeseler A. Ultrafast approximation for phylogenetic  
761 bootstrap. *Mol Biol Evol* 2013; **30**: 1188–1195.
- 762 31. Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H *et al*.  
763 Linking virus genomes with host taxonomy. *Viruses* 2016; **8**: 66.
- 764 32. Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB. Virus

- 765 taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV).  
766 *Nucleic Acids Res* 2018; **46**: D708–D717.
- 767 33. John DE, Patterson SS, Paul JH. Phytoplankton-group specific quantitative polymerase  
768 chain reaction assays for RuBisCO mRNA transcripts in seawater. *Mar Biotechnol* 2007;  
769 **9**: 747–759.
- 770 34. Bolch CJS, Subramanian TA, Green DH. The toxic dinoflagellate gymnodinium  
771 *catenatum* (Dinophyceae) Requires marine bacteria for growth. *J Phycol* 2011; **47**: 1009–  
772 1022.
- 773 35. Wolf YI, Kazlauskas D, Iranzo J, Lucía-Sanz A, Kuhn JH, Krupovic M *et al.* Origins and  
774 evolution of the global RNA virome. *MBio* 2018; **9**: e02329-18.
- 775 36. Shi M, Lin X-D, Tian J-H, Chen L-J, Chen X, Li C-X *et al.* Redefining the invertebrate  
776 RNA virosphere. *Nature* 2016; **540**: 539–543.
- 777 37. Shi M, Lin X-D, Chen X, Tian J-H, Chen L-J, Li K *et al.* The evolutionary history of  
778 vertebrate RNA viruses. *Nature* 2018; **556**: 197–202.
- 779 38. Geoghegan JL, Di Giallonardo F, Cousins K, Shi M, Williamson JE, Holmes EC. Hidden  
780 diversity and evolution of viruses in market fish. *Virus Evol* 2018; **4**: vey031.
- 781 39. Correa AMS, Welsh RM, Vega Thurber RL. Unique nucleocytoplasmic dsDNA and  
782 +ssRNA viruses are associated with the dinoflagellate endosymbionts of corals. *ISME J*  
783 2013; **7**: 13–27.
- 784 40. Levin RA, Voolstra CR, Weynberg KD, Van Oppen MJH. Evidence for a role of viruses  
785 in the thermal sensitivity of coral photosymbionts. *ISME J* 2017; **11**: 808–812.
- 786 41. Waldron FM, Stone GN, Obbard DJ. Metagenomic sequencing suggests a diversity of  
787 RNA interference-like responses to viruses across multicellular eukaryotes. *PLoS Genet*  
788 2018; **14**: e1007533.
- 789 42. Käfer S, Paraskevopoulou S, Zirkel F, Wieseke N, Donath A, Petersen M *et al.* Re-

- 790 assessing the diversity of negative strand RNA viruses in insects. *PLoS Pathog* 2019; **15**:  
791 e1008224.
- 792 43. Wu H, Pang R, Cheng T, Xue L, Zeng H, Lei T *et al.* Abundant and diverse RNA viruses  
793 in insects revealed by RNA-Seq analysis: ecological and evolutionary implications.  
794 *mSystems* 2020; **5** :e00039-20
- 795 44. Hillman BI, Cai G. The family *Narnaviridae*: simplest of RNA viruses. *Adv Virus Res*  
796 2013; **86**: 149-176.
- 797 45. Lye L-F, Akopyants NS, Dobson DE, Beverley SM. A narnavirus-like element from the  
798 trypanosomatid protozoan parasite *Leptomonas seymouri*. *Genome Announc* 2016; **4**:  
799 713–729.
- 800 46. Grybchuk D, Kostygov AY, Macedo DH, d’Avila-Levy CM, Yurchenko V. RNA viruses  
801 in trypanosomatid parasites: a historical overview. *Mem Inst Oswaldo Cruz* 2018; **113**:  
802 e170487.
- 803 47. Akopyants NS, Lye L-F, Dobson DE, Lukeš J, Beverley SM. A narnavirus in the  
804 trypanosomatid protist plant pathogen *Phytomonas serpens*. *Genome Announc* 2016; **4**:  
805 e00711-16.
- 806 48. Charon J, Grigg MJ, Eden JS, Piera KA, Rana H, William T *et al.* Novel RNA viruses  
807 associated with *Plasmodium vivax* in human malaria and *Leucocytozoon* parasites in  
808 avian disease. *PLoS Pathog* 2019; **15**: e1008216.
- 809 49. Cai G, Myers K, Fry WE, Hillman BI. A member of the virus family *Narnaviridae* from  
810 the plant pathogenic oomycete *Phytophthora infestans*. *Arch Virol* 2012; **157**: 165–169.
- 811 50. Urayama SI, Takaki Y, Nunoura T. FLDS: A comprehensive DSRNA sequencing  
812 method for intracellular RNA virus surveillance. *Microbes Environ* 2016; **31**: 33–40.
- 813 51. Nerva L, Vigani G, Di Silvestre D, Ciuffo M, Forgia M, Chitarra W *et al.* Biological and  
814 molecular characterization of *Chenopodium quinoa* Mitovirus 1 reveals a distinct small

- 815 RNA response compared to those of cytoplasmic RNA viruses. *J Virol* 2019; **93**:  
816 e01998-18.
- 817 52. Charon J, Marcelino VR, Wetherbee R, Verbruggen H, Holmes EC. Metatranscriptomic  
818 identification of diverse and divergent RNA viruses in green and chlorarachniophyte  
819 algae cultures. *Viruses* 2020; **12**: 1180.
- 820 53. DeRisi JL, Huber G, Kistler A, Retallack H, Wilkinson M, Yllanes D. An exploration of  
821 ambigrammatic sequences in narnaviruses. *Sci Rep* 2019; **9**: 17982.
- 822 54. Belshaw R, Pybus OG, Rambaut A. The evolution of genome compression and genomic  
823 novelty in RNA viruses. *Genome Res* 2007; **17**: 1496–1504.
- 824 55. Dinan AM, Lukhovitskaya NI, Olendraite I, Firth AE. A case for a negative-strand  
825 coding sequence in a group of positive-sense RNA viruses. *Virus Evol* 2020; **6**: veaa007.
- 826 56. Solórzano A, Rodríguez-Cousiño N, Esteban R, Fujimura T. Persistent yeast single-  
827 stranded RNA viruses exist *in vivo* as genomic RNA·RNA polymerase complexes in 1:1  
828 stoichiometry. *J Biol Chem* 2000; **275**: 26428–26435.
- 829 57. Vlok M, Lang AS, Suttle CA. Application of a sequence-based taxonomic classification  
830 method to uncultured and unclassified marine single-stranded RNA viruses in the order  
831 Picornavirales. *Virus Evol* 2019; **5**: vez056.
- 832 58. Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Adriaenssens EM, Dempsey DM  
833 *et al.* Changes to virus taxonomy and the statutes ratified by the International Committee  
834 on Taxonomy of Viruses (2020). *Arch Virol* 2020; **165**: 2737–2748.
- 835 59. Chiba Y, Tomaru Y, Shimabukuro H, Kimura K, Hirai M, Takaki Y *et al.* Viral RNA  
836 genomes identified from marine macroalgae and a diatom. *Microbes Environ* 2020; **35**:  
837 ME20016.
- 838 60. Sasai S, Tamura K, Tojo M, Herrero ML, Hoshino T, Ohki ST *et al.* A novel non-  
839 segmented double-stranded RNA virus from an Arctic isolate of *Pythium polare*.

- 840 *Virology* 2018; **522**: 234–243.
- 841 61. Lachnit T, Thomas T, Steinberg P. Expanding our understanding of the seaweed  
842 holobiont: RNA viruses of the red alga *Delisea pulchra*. *Front Microbiol* 2016; **6**: 1489.
- 843 62. Shiba K, Hatta C, Sasai S, Tojo M, T. Ohki S, Mochizuki T. Genome sequence of a  
844 novel partitivirus identified from the oomycete *Pythium nunn*. *Arch Virol* 2018; **163**:  
845 2561–2563.
- 846 63. Nibert ML, Woods KM, Upton SJ, Ghabrial SA. Cryspovirus: a new genus of protozoan  
847 viruses in the family *Partitiviridae*. *Arch Virol* 2009; **154**: 1959–1965.
- 848 64. Park D, Goh CJ, Kim H, Hahn Y. Identification of two novel amalgaviruses in the  
849 common eelgrass (*Zostera marina*) and in silico analysis of the amalgavirus +1  
850 programmed ribosomal frameshifting sites. *Plant Pathol J* 2018; **34**: 150–156.
- 851 65. Middelboe M, Brussaard C. Marine viruses: key players in marine ecosystems. *Viruses*  
852 2017; **9**: 302.
- 853 66. Nagasaki K. Dinoflagellates, diatoms, and their viruses. *J Microbiol* 2008; **46**: 235–243.
- 854 67. Brussaard CPD, Martínez J. Algal bloom viruses. *Plant Viruses* 2008; **2**: 1–10.
- 855 68. Takahashi H, Fukuhara T, Kitazawa H, Kormelink R. Virus latency and the impact on  
856 plants. *Front Microbiol* 2019; **10**: 2764.
- 857 69. Roossinck MJ. The good viruses: Viral mutualistic symbioses. *Nat Rev Microbiol* 2011;  
858 **9**: 99–108.
- 859 70. Deakin G, Dobbs E, Bennett JM, Jones IM, Grogan HM, Burton KS. Multiple viral  
860 infections in *Agaricus bisporus* - Characterisation of 18 unique RNA viruses and 8  
861 ORFans identified by deep sequencing. *Sci Rep* 2017; **7**: 1–13.
- 862 71. Ghabrial SA, Castón JR, Jiang D, Nibert ML, Suzuki N. 50-plus years of fungal viruses.  
863 *Virology* 2015; **479–480**: 356–368.
- 864 72. Marzano S-YL, Nelson BD, Ajayi-Oyetunde O, Bradley CA, Hughes TJ, Hartman GL *et*



- 865 *al.* Identification of diverse mycoviruses through metatranscriptomics characterization of  
866 the viromes of five major fungal plant pathogens. *J Virol* 2016; **90**: 6846–6863.
- 867 73. Xie J, Jiang D. New insights into mycoviruses and exploration for the biological control  
868 of crop fungal diseases. *Annu Rev Phytopathol* 2014; **52**: 45–68.
- 869 74. Welsh JE, Steenhuis P, de Moraes KR, van der Meer J, Thieltges DW, Brussaard CPD.  
870 Marine virus predation by non-host organisms. *Sci Rep* 2020; **10**: 1–9.
- 871 75. Callaway E. ‘It will change everything’: DeepMind’s AI makes gigantic leap in solving  
872 protein structures. *Nature* 2020; **588**: 203–204.
- 873