

1 For submission to: *Applied and Environmental Microbiology*

2 **TITLE: Diversity of active viral infections within the *Sphagnum* microbiome**

3 Authors:

4 **Joshua M.A. Stough^{1*}, Max Kolton², Joel E. Kostka², David J. Weston^{3,4}, Dale A. Pelletier³,**
5 **and Steven W. Wilhelm^{1#}**

6 Addresses:

7 ¹ Department of Microbiology, University of Tennessee, Knoxville, Tennessee, United States of
8 America 37996

9 ² School of Biology and School of Earth and Atmospheric Sciences, Georgia Institute of
10 Technology, Atlanta, GA, USA

11 ³ Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

12 ⁴ Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, Tennessee,
13 USA

14 #Address correspondence to Steven W. Wilhelm: wilhelm@utk.edu

15

16 *present address: Department of Microbiology & Immunology, University of Michigan, Ann
17 Arbor 48109

18

19

20 Keywords: Viruses, RNA-seq, *Sphagnum*, Peat bogs, microbial ecology

1

21 **Abstract**

22 *Sphagnum*-dominated peatlands play an important role in global carbon storage and represent
23 significant sources of economic and ecological value. While recent efforts to describe microbial
24 diversity and metabolic potential of the *Sphagnum* microbiome have demonstrated the
25 importance of its microbial community, little is known about the viral constituents. We used
26 metatranscriptomics to describe the diversity and activity of viruses infecting microbes within
27 the *Sphagnum* peat bog. The vegetative portions of 6 *Sphagnum* plants were obtained from a
28 peatland in northern Minnesota and total RNA extracted and sequenced. Metatranscriptomes
29 were assembled and contigs screened for the presence of conserved virus marker genes. Using
30 bacteriophage capsid protein, gp23, as a marker for phage diversity, we identified 33 contigs
31 representing undocumented phage s that were active in the community at the time of sampling.
32 Similarly, RNA-dependent RNA polymerase and the Nucleo-Cytoplasmic Large DNA Virus
33 (NCLDV) major capsid protein were used as markers for ssRNA viruses and NCLDV,
34 respectively. In total 114 contigs were identified as originating from undescribed ssRNA viruses,
35 22 of which represent near-complete genomes. An additional 64 contigs were identified as being
36 from NCLDVs. Finally, 7 contigs were identified as putative virophage or polinto-like viruses.
37 We developed co-occurrence networks with these markers in relation to the expression of
38 potential-host housekeeping gene *rpb1* to predict virus-host relationships, identifying 13 groups.
39 Together, our approach offers new tools for the identification of virus diversity and interactions
40 in understudied clades, and suggest viruses may play a considerable role in the ecology of the
41 *Sphagnum* microbiome.

42

43 **Significance**

44 *Sphagnum*-dominated peatlands play an important role in maintaining atmospheric carbon
45 dioxide levels by modifying conditions in the surrounding soil to favor its own growth over other
46 plant species. This slows rates of decomposition and facilitates the accumulation of fixed carbon
47 in the form of partially decomposed biomass. The unique environment produced by *Sphagnum*
48 enriches for the growth of a diverse microbial consortia that benefit from and support the moss's
49 growth, while also maintaining the hostile soil conditions. While a growing body of research has
50 begun to characterize the microbial groups that colonize *Sphagnum*, little is currently known
51 about the ecological factors that constrain community structure and define ecosystem function.
52 Top-down population control by viruses is almost completely undescribed. This study provides
53 insight into the significant viral influence on the *Sphagnum* microbiome, and identifying new
54 potential model systems to study virus-host interactions in the peatland ecosystem.

55

56

57 Introduction

58 Peatlands represent one of the most significant biological carbon sinks on the planet,
59 storing an estimated 25% of terrestrial carbon in the form of partially decomposed organic matter
60 (1-3). This accumulation of carbon is achieved through much slower rates of respiration and
61 decomposition than observed in soil, due in large part to the low pH, nutrient-poor, and
62 anaerobic environments created by the dominant moss population (4, 5), of which the genus
63 *Sphagnum* is most prevalent (6, 7). As these environmental conditions appear to favor the growth
64 of *Sphagnum* over vascular plants, primary production is dominated by the moss, which further
65 retards decomposition due to production of antimicrobial compounds such as sphaginic acid (8-
66 10) and sphagnan (11, 12). Despite this, *Sphagnum* and other peat mosses cultivate a diverse,
67 symbiotic microbiome that appears to abate nutritional gaps for the moss and also contribute to
68 the unique biogeochemical characteristics of the peatland ecosystem (13-15). In addition to their
69 value as reservoirs of microbial diversity, the partially decomposed organic matter, known as
70 *Sphagnum* peat, serves as an important economic resource for use in horticulture. As many peat
71 bogs have begun to experience stress due to anthropogenic disturbances (16-18) and possibly
72 climate change (19), the *Sphagnum* microbiome is of interest in peatland conservation and the
73 ecosystem's services to the surrounding environment.

74 While there is a growing body of research characterizing the microbial groups that
75 colonize *Sphagnum* (15), little is currently known about the ecological factors that define
76 community structure and ecosystem function. Studies suggest that subtle differences in pH and
77 available nutrients, manipulated by different *Sphagnum* species and strains, create distinct
78 microbial consortia (14, 20, 21). Other observations suggest a more homogenous community
79 (22), highlighting a need for further study. Culture-dependent experiments isolating endophytic

80 bacteria indicate *Sphagnum* cultivates symbionts with abilities that include antifungal activity
81 (20, 23) and nitrogen fixation (14), and that these microbiomes may be passed vertically to the
82 moss progeny (21). Yet while examinations of how environmental conditions and host-microbe
83 symbiotic interactions shape the structure and function of microbial communities, the influence
84 of virus populations on the *Sphagnum* microbiome remains unexplored.

85 Viruses are the most abundant biological entities on Earth, and central to global
86 ecosystems as they can drive the host evolution through predator-prey interactions and horizontal
87 gene transfer (24). Moreover, viruses can lyse single-celled primary producers and heterotrophs,
88 releasing nutrient elements from the biomass of prokaryotes and eukaryotic protists (25, 26).
89 Viruses may also act as a top-down control on the composition and evenness of microbial
90 communities, targeting hosts that reach higher cell densities, a phenomenon referred to as the
91 “kill-the-winner” model (27).

92 As lab studies of viruses require hosts that can be grown in culture, many
93 environmentally relevant viruses are poorly understood and their representation in reference
94 databases is often skewed. Previous efforts to describe environmental viromes have focused
95 primarily on the sequencing of shotgun or PCR-targeted metagenomes. While these methods
96 have proven powerful, rapidly expanding available reference material for bacteriophage (28, 29),
97 it leaves the considerable diversity of RNA viruses largely untapped (30). Moreover, the
98 common approach of selecting for viruses based on size-exclusion with filters removes many of
99 the Nucleo-Cytoplasmic Large DNA Viruses (NCLDVs, or commonly “giant viruses”) that are
100 also environmentally relevant and phylogenetically informative (31, 32). Metagenomic
101 sequencing also limits observations to virus particles: from these data inferences on viral activity
102 require tenuous assumptions. The advent of high-throughput RNA sequencing offers viral

103 ecologists the opportunity to study active infections in the environment, as DNA viruses only
104 produce transcripts inside a host. Moreover, this approach also captures fragments of RNA virus
105 genomes. When sequencing is of sufficient depth and multiple samples are collected with spatial
106 and temporal variability, these data present an opportunity to develop hypothetical relationships
107 between virus and host markers (33) for subsequent in lab testing.

108 In this study, we analyzed metatranscriptomes from the microbial community inhabiting
109 the vegetative portion of *Sphagnum fallax* and *S. magellanicum* plants in Northern Minnesota,
110 with the goal of describing active viral infections within the *Sphagnum* microbiome. Using
111 marker genes conserved within several viral taxa, we identified an active and diverse
112 bacteriophage population, largely undescribed in previous studies. We also identified ongoing
113 infections by a diverse consortium of “giant” viruses and potentially corresponding
114 virophage/polinton-like viruses (hereafter referred to as virophage), including several giant
115 viruses closely related to the recently discovered Klosneuviruses (34). Finally, a number of novel
116 positive-sense single-stranded RNA viruses, some of which assembled into near complete
117 genomes, were observed. With this information in hand we developed statistical network
118 analyses, correlating co-expression of viral marker genes with housekeeping transcripts from
119 potential hosts. The resulting observations propose several virus-host pairings that, moving
120 forward, can be tested in a laboratory setting. Together, these results demonstrate new potential
121 model systems to study virus-host interactions in the peat bog ecosystem, and provide insight
122 into the significant viral influence on the *Sphagnum* microbiome.

123

124 **Results**

125 *Identification of resident phage populations*

126 To identify active virus populations in the *Sphagnum* phyllosphere, we obtained *S. fallax*
127 and *S. magellanicum* plant matter samples (3 from each species) from peatland terrariums as a
128 part of the *Spruce and Peatlands Under Changing Environments* (SPRUCE) project for
129 metatranscriptomic sequencing. Across all six *Sphagnum* phyllosphere samples, 33 contigs were
130 identified as transcripts encoding major capsid protein (*gp23*) originating from bacteriophage,
131 while only 6 contigs were identified using three other marker genes. Concurrent with this, more
132 reads mapped to *gp23* contigs than to the other marker genes combined, the most abundant of
133 which were three ribonucleotide reductase contigs.

134 Of the 33 contigs, 18 were assigned to the *Eucampyvirinae* subfamily with
135 *Campylobacter* viruses CP220 and PC18, while the rest were spread amongst the other Myovirus
136 taxa, predominantly the *Tevenvirinae* (Fig 1). SS4 contig 77559 was the most abundant, with
137 consistently high expression across all samples, whereas other contigs dominated just one or two
138 samples. Of the 6 contigs identified using the other 3 viral marker genes, one was identified as a
139 potential *gp20* homologue, originating within *Myoviridae* with *Clostridium* virus phiCD119 as
140 the closest relative (SFig 1). Two contigs were identified as *recA* contigs, likely originating in
141 myovirus and siphovirus relatives (SFig 2), while the remaining three contigs were identified as
142 ribonucleotide reductase transcripts (SFig 3).

143

144 *Single-stranded RNA virus diversity and abundance*

145 Within our samples, 114 contigs originated from RNA viruses, the majority of which
146 belonged to the currently unassigned *Barnaviridae* and Astrovirus-like families (Fig 2).

147 Additionally, a large number of *Picornaviruses* were observed, most of which were closely
148 related to the unclassified marine *Aurantiochytrium* single-stranded RNA virus, and *Secoviridae*
149 plant viruses. Lastly, several contigs were closely related to the *Nidovirales* clade, which
150 generally infect animal species.

151 Among these, 22 contigs were found to be near complete ssRNA virus genomes (based
152 on gene content and size), encoding multiple viral genes in addition to RDRP. Gene regions were
153 identified and annotated using the NCBI conserved domain and PFam HMM search tools, and
154 the full-length RDRP sequence was used to construct a maximum likelihood phylogenetic tree
155 (Fig 3). Of the partial ssRNA genomes that were assembled, 2 were missing the conserved Rhv
156 structural genes, while one was missing a RNA virus Helicase. The majority of these contigs fall
157 under the *Picornavirales* order, which also included the most complete viral genomes. As was
158 observed with the shorter RDRP contigs above, most of the *Picornavirales* contigs were most
159 closely related to the unclassified marine species, or members of the family *Secoviridae* clade,
160 whose membership includes the Parsnip yellow fleck virus. A number of partial *Picornavirus*
161 genomes were also identified as members of the family *Dicistroviridae*. Outside the
162 *Picornavirales*, most contigs clustered closely with the unassigned Astrovirus-like *Phytophthora*
163 *infestans* RNA virus. To determine the relative abundance of different RNA virus genomes in the
164 peat bog samples, we mapped reads back to contigs and calculated transcripts per million (TPM)
165 values to account for contig length and library size. Overall, NCLDV transcripts were the most
166 abundant (range 4,465 – 16,887 reads mapped to MCP contigs, or 46.4 - 200.1 TPM), followed
167 by RNA Viruses (range 13,373 – 166,337 total reads mapped to RNA virus contigs, or 7.11 –
168 97.5 TPM), and bacteriophage (range 287 – 1405 total reads mapped to gp23 contigs, or 3.4 –
169 5.2 TPM). The most abundant contig across all samples was SS4 contig 3964, which was most

170 closely related to the Rotifer birnavirus. All other contigs appear to be abundant prominently in
171 one or two samples, and absent or in low abundance in the others, with no patterns of abundance
172 apparent.

173 *Giant viruses and virophage in Sphagnum microbiome*

174 Of the 10 gene markers tested to identify Nucleo-Cytoplasmic Large DNA Viruses
175 (NCLDVs), only the giant virus major capsid protein (MCP) was detected in the
176 metatranscriptome. 64 contigs were observed with homology to MCP, representing every known
177 group of NCLDVs (Fig 4). Out of the 64 MCP contigs, 46 were placed within the *Mimiviridae*
178 taxa. Most contigs (25) closely aligned with the recently discovered Klosneuviruses, with the
179 Indivirus and Catovirus representing the most diversity in these samples. The next most abundant
180 group were the “extended *Mimiviridae*” (7 contigs), species with known similarity to
181 Mimiviruses but that infect eukaryotic algae. Six contigs phylogenetically were similar to the
182 *Asfarviridae*, here represented by the African swine fever Virus. Potential relatives of the giant
183 virus outliers, Pandoravirus and Pithovirus, were not observed (due to methodological
184 limitations), and the *Iridoviridae* were poorly represented (1 contig). Using the virophage MCP
185 and packaging ATPase as markers, we identified 7 contigs as transcripts originating in putative
186 virophage or polinton-like viruses, all of which were phylogenetically placed amongst isolates
187 identified from freshwater ecosystems (Fig 5).

188 As was observed with the other major viral taxa described, the majority of contigs were
189 most abundantly expressed in one or two samples and present at very low levels in the rest. The
190 most abundant NCLDV-MCP contig in the samples was SS2 contig 73240, most closely related
191 to *Megavirus chilensis*, which was the most highly expressed giant virus contig across all

192 samples. Four other contigs (SS6 contig 110585, SS4 contigs 55722 and 141177, and SS5 contig
193 119519) were highly expressed across all six samples.

194 *Prediction of virus-host pairs*

195 By comparing and correlating expression of virus marker genes to *rpb1* expression from
196 cellular organisms, we endeavored to predict potential virus-host groups in the *Sphagnum*
197 phyllosphere. Fig 6 shows statistically robust networks containing at least one virus and one host,
198 where co-occurrence and correlation were observed in more than one sample. A total of 13 virus-
199 host groups were detected, spread across the major viral taxa detected in this dataset. We note
200 that no networks containing the virophage/polinton-like viruses emerged. Four relationships
201 were predicted from bacteriophage *gp23* abundance, the simplest of which was a *Tevenvirinae*
202 phage-Fungi-Fungi group with moderate correlations (Fig 6a). The other 3 relationships are more
203 complicated, containing multiple potential hosts and, for the largest predicted group, multiple
204 virus transcripts. Some of the potential hosts in these groups were identified as eukaryotic.

205 We observed 4 predicted RNA virus-host clusters, all of which contained multiple hosts
206 grouped with a single virus (Fig 6b). Many of the predicted hosts appear closely related to
207 eukaryotic single-celled protists, including members of the *Cryptophyceae*, *Excavata*, and
208 *Amoebozoa*, as well as a variety of bacterial and archaeal species. Correlation coefficients
209 observed in these relationships are generally higher than observed in the phage-host clusters. The
210 5 predicted NCLDV-host clusters (Fig 6c) were the most highly correlated and complex.
211 Predicted hosts were highly varied, ranging from bacteria to fungi, though all virus members
212 were placed either within *Mimiviridae* or the extended Mimivirus group. MCP contigs
213 originating in close relatives of the recently discovered Klosneuviruses are present in both the 7-
214 and 10-member clusters, in addition to a pair of contigs most closely related to *Aureococcus*

215 *anophagefferens* Virus (AaV). An additional 15 statistically significant clusters across all three
216 viral taxa were observed where the virus and host were present in only one sample (not shown).

217

218 **Discussion**

219 Understanding the virus burden on microbial communities in ecologically-rich
220 ecosystems is an important step forward in resolving their function and predicting how they
221 might respond to various drivers of ecosystem scale change. In the present study we used
222 metatranscriptomes to describe the diversity and activity of the resident virus populations in a
223 peat moss (*Sphagnum*) microbiome. We identified previously undescribed virus activity from
224 multiple taxa, most of which are poorly represented in either the literature or reference sequence
225 databases. We used read mapping to quantify the relative abundance of active viral infections.
226 Lastly, we compared expression of viral transcripts to that of potential hosts, using a correlation
227 co-occurrence networks approach (33) to predict putative hosts for the observed virus
228 populations. Together, our results suggest that the *Sphagnum* phyllosphere represents a
229 significant and largely untapped source of virus diversity and activity. Viruses were highly active
230 across all samples, with some individual viruses exhibiting abundant activity in single samples,
231 while others were more pervasive. Given that our observations were based on RNA sequencing
232 data, they do not represent a full accounting of the virus particles present in the community.
233 However, metatranscriptomic data, allows us to distinguish virus populations active at the time
234 of sampling. In addition, as viruses only transcribe their genes during infection, virus and host
235 transcripts are expected to co-occur, and it is possible that the abundance of transcripts (at least
236 for DNA viruses) could be used to predict natural hosts of viruses observed in the ecosystem

237 which can be tested in a laboratory or field setting. Ultimately, this study identifies from within a
238 complex community a number of candidate virus-host model systems for future study.

239 *Viral diversity and activity in Sphagnum plants*

240 As viruses lack a universal genetic marker like the bacterial 16S rRNA gene, we opted to
241 screen metatranscriptome assemblies for genes previously demonstrated to be largely or wholly
242 conserved amongst individual viral taxa. Within the expanded and diverse genetic potential of
243 giant viruses, only a handful of genes are currently conserved amongst all members (32, 35) and
244 these, in addition to several markers conserved amongst a large portion of giant viruses were
245 used to identify activity in the *Sphagnum* phyllosphere. Out of the 10 genes used to screen the
246 metatranscriptomes, we only MCP transcripts. This is not surprising given the number of capsid
247 proteins needed for viral assembly: indeed this transcriptional pattern was previously observed in
248 both cultures (36) and marine systems by Moniruzzaman *et al.* (2017). It should be noted that the
249 RNA-seq dataset used in those studies was poly-A selected, enriching for eukaryotic transcripts,
250 and thus coverage of eukaryotic virus gene expression would be much higher than in the
251 *Sphagnum* metatranscriptome. That we observed MCP expression in abundance suggests a
252 significant number of infections occurred at the time of sampling. While the magnitude of giant
253 virus diversity in *Sphagnum* dominated ecosystems is, to our knowledge, completely unexplored,
254 the richness observed here is considerably larger than expected compared to better documented
255 systems. 64 distinct MCP genotypes were identified in the *Sphagnum* phyllosphere
256 metatranscriptomes, which is high when compared to one recent survey that identified 30 novel
257 MCP transcripts from multiple environmental datasets (37), and another which observed 107
258 NCLDV sequences in 16 publicly available environmental metagenomes of comparable
259 sequencing depth isolated from different ecosystems (38). Most of the MCP contigs identified

260 were placed in clusters around a small number of virus relatives, highlighting the under-sampled
261 diversity of giant viruses in the literature, poor representation in reference databases, and the
262 considerable diversity present in *Sphagnum* peat bogs. The significant giant virus diversity
263 observed here implies a corresponding eukaryotic richness that is also under-described (39).
264 Additionally, a series of virophage transcripts were detected, indicating a significant response to
265 infections by giant viruses in the system. Many of these are phylogenetically grouped with the
266 polintoviruses, transposable elements that produce virion particles that can exploit the replication
267 machinery of actively infecting giant viruses to reproduce, often at the expense of the giant (40,
268 41). These observations suggest that while an active picoeukaryotic population may persist,
269 mortality mechanisms beyond grazer-driven losses are at play and likely important to carbon
270 flow in the system.

271 The use of RNA-seq presents a unique opportunity to capture the genomic material of
272 RNA viruses that is lost in metagenomic sequencing. As such, RNA virus representation in
273 sequencing databases and the literature is largely constrained to culture-based studies. All known
274 RNA viruses require a functional RNA-dependent RNA polymerase (RDRP) to copy their
275 genome inside the host cell, a function exclusive to viruses, making it a highly specific marker
276 for RNA virus discovery (42, 43). Recent attempts to use metatranscriptomes to describe
277 environmental RNA viruses have proven successful, not only identifying marker gene fragments
278 in datasets, but assembling complete and near-complete genomes (33, 43). The diversity and
279 composition of RNA virus populations in *Sphagnum* peatlands is largely unknown: it is currently
280 limited to the small group of RNA-DNA hybrid chimeric Cruciviruses (44). Here, as was
281 observed with the giant viruses, most RNA virus contigs were placed in clusters with a single
282 represented species, suggesting a significant degree of uncharacterized diversity. This is not

283 entirely surprising, as RNA viruses are expected to make up as much as half of the virus particles
284 in the Earth's oceans, and yet they are almost as poorly understood and represented in
285 sequencing databases as giant viruses (30). Similarly, we assembled and identified 22 near-
286 complete RNA virus genomes, where completeness was determined primarily by size and the
287 presence of the 6 core genes. As there are currently only 265 sequenced genomes within the
288 *Picornavirales*, most of which grouped within the *Picornaviridae*, this represents a sizeable
289 addition to the known diversity of ssRNA viruses. This is especially true for the unassigned and
290 unclassified taxa, and establishes a strong foundation for future efforts to describe RNA virus
291 populations in *Sphagnum*.

292 Description of bacteriophage populations in *Sphagnum* peatlands is currently limited to
293 the ssDNA viruses of the *Microviridae* (45) and *Caudovirales* (46) observed in metagenomics
294 data, though it appears that phage are the most abundant biological entities in the *Sphagnum*
295 phyllosphere (46). Given this, and the dominance of bacteria in the *Sphagnum* microbiome as
296 previously described (15), the relatively low abundance of active bacteriophage in our samples
297 was a surprise. Marker genes to identify bacteriophage were chosen based on their conservation
298 across phage taxa and their success in other environmental datasets. Gp20 (phage portal protein)
299 and Gp23 (major capsid protein) have been shown previously to be highly conserved and
300 effective for phylogenetic assignment of members of the *Myoviridae* (47-49). RecA is conserved
301 across all three bacteriophage taxa and could illuminate lysogeny, and ribonucleotide reductase
302 (RNR) has been used as an effective marker for screening novel viruses from marine sequencing
303 datasets (50). As such, we identified 39 bacteriophage contigs using these markers, 33 of which
304 were from Gp23. This may represent a similar phenomenon as MCP in the giant viruses above,
305 where transcripts encoding structural proteins are much more abundant than other genes and

306 sequencing lacked the depth to detect them. For the purpose of discovering novel phage species,
307 DNA sequencing through metagenomics may prove more successful.

308 *Virus-host predictions*

309 Future study of viral dynamics in peatlands will require the establishment of model
310 virus/host pairs for *in vitro* experimentation and *in situ* tracking. While culture-based techniques
311 can yield model systems, it is not always clear whether the isolated organisms are
312 environmentally relevant. In order to address this, we attempted to use statistical methods to
313 propose virus/host pairs as potential future model systems based on their co-occurrence in
314 samples and the correlation of their abundance. As viruses produce transcripts only when
315 actively infecting a host, positive correlation and co-occurrence between virus and host
316 transcripts is expected, and might be used to predict host-virus relationships, provided an
317 appropriate transcriptional proxy for growth and activity is available (33). In this study, we used
318 the eukaryotic RNA-polymerase gene *rpb1* as a marker for abundance and activity in potential
319 hosts, as it is conserved amongst all eukaryotic organisms, is phylogenetically informative, and
320 has been previously described as one of the more consistently expressed eukaryotic genes in
321 marine systems, scaling well with the activity of the organism (51), though the stability of its
322 expression has not been evaluated in terrestrial ecosystems. We used NCLDV MCP abundance
323 as a proxy for giant virus production, Gp23 for phage production, as transcription is necessary
324 for the assembly of new virus particles and transcript abundance in some appears to be closely
325 linked to viral replication. We also used RDRP as a proxy for RNA virus production,
326 acknowledging the caveat that we cannot distinguish between abundance of free virus particles
327 and active infections (33).

328 Correlation and co-occurrence matrices, clustered into groups by similarity and tested
329 with the SIMPROF permutation test, yielded 13 predicted groups of viruses and hosts. For
330 ssRNA and giant viruses, several of the networks produced in the analysis included multiple
331 bacterial and archaeal sequences picked up in the RNA polymerase screen. As we have no reason
332 to believe bacterial species are infected by NCLDVs or Picornaviruses, it is likely these
333 predictions represent a confounding relationship between prokaryotes and potential eukaryotic
334 hosts, observed in network analyses for all three viral taxa described here, where a beneficial
335 interaction results in an indirect correlation with viral infection. Indeed, previous use of this
336 method in marine systems showed a similar phenomenon, where an algal Mimivirus and a
337 known host were grouped with a fungal species and another virus (33). Even after the
338 consideration of bacterial species within the predicted groups, some remain complicated with
339 multiple viruses and potential eukaryotic hosts, which may be explained by a broader host range
340 amongst giant viruses enabled by the expansion of genetic material and increased independence
341 from host machinery. Similar relationships were observed amongst RNA viruses, though these
342 are more tenuous, as we are unable to distinguish whether sequencing reads originated transcripts
343 or genomic material.

344 All together, we have identified a considerable amount of viral diversity from several
345 major viral taxa active within a poorly understood microbial ecosystem. As they were identified
346 from transcript sequencing data, the viruses described here likely only represent a fraction of the
347 whole virus community, which may be elucidated through further culture-independent work. We
348 have also used transcript abundance within a statistical framework to predict several host-virus
349 relationships which can be sought out and tested in culture. These results establish an important

350 and much needed foundation for future research into the microbial ecology in *Sphagnum* peat
351 bogs.

352

353 **Materials and Methods**

354 *Sample collection and Survey of Environmental Conditions*

355 Triplicate individual plants of *Sphagnum magellanicum* and *Sphagnum fallax* were
356 collected on August 2015 from the SPRUCE experiment site at the S1 bog on the Marcell
357 Experimental Forest (U.S. Forest Service, <http://mnspruce.ornl.gov/>). The S1 Bog is an acidic
358 and nutrient-deficient ombrotrophic *Sphagnum*-dominated peatland bog (surface pH \leq 4.0) located
359 approximately 40 km north of Grand Rapids, Minnesota, USA (47°30.476' N; 93°27.162' W; 418
360 m above mean sea level) (52-54). To characterize the *Sphagnum* virome, *Sphagnum* samples
361 were collected as previously described (54). Only green living plants were sampled: samples
362 focused on the capitulum plus about 2-3 cm of green living stem. B *Sphagnum* stems
363 (phyllosphere) were cleaned from unrelated plant debris, and frozen immediately on dry ice.
364 Frozen samples were overnight shipped to the Georgia Institute of Technology for RNA
365 extraction.

366

367 *RNA Extraction and Sequencing*

368 One gram of *Sphagnum* phyllosphere tissue was ground with a mortar and pestle under liquid
369 nitrogen. The fine powder was transferred to 10 extraction tubes and total RNA isolated using
370 the PowerPlant RNA Isolation Kit with DNase according to the manufacturer's protocol (MoBio

371 Laboratories, Carlsbad, CA, USA). DNA-depleted RNA was quantified using the Qubit RNA HS
372 Assay Kit (Invitrogen, Carlsbad, CA, USA) and quality was assessed on the Agilent 2100
373 BioAnalyzer using the Agilent RNA 6000 Pico Kit (Agilent Technologies). Additionally, the
374 absence of DNA contamination was confirmed by running a polymerase chain reaction using
375 universal bacterial 16S rRNA primers 515F and 806R. Finally, RNA samples without detectible
376 DNA contamination and exhibiting an RNA integrity number (RIN) > 6 were pooled. Extracted
377 total environmental RNA samples were sent on dry ice to the Joint Genome Institute (JGI)
378 facilities for meta-transcriptomes libraries construction and sequencing. All protocols employed
379 were standard JGI protocols Ribosomal RNA subtraction from total environmental RNA was
380 completed using the Ribo-Zero rRNA Removal Kit (Illumina, San Diego, CA). rRNA depleted
381 environmental RNA were used to construct paired end metatranscriptomes libraries using
382 TruSeq kit and sequenced on the Illumina HiSeq2000 platform at the JGI facilities using a
383 single-end 250bp flow cell.

384 *RNA-seq Data Processing*

385 Raw sequences (see Supplementary Table 1, sequencing stats tab) were downloaded from
386 the Department of Energy Joint Genome Institute server and processed using the CLC Genomics
387 Workbench v. 10.0.1 (QIAGEN, Hilden, Germany). Reads below a 0.03 quality score cutoff
388 were removed from subsequent analyses, and the remaining reads were trimmed of any
389 ambiguous and low quality 5' bases. Samples were subjected to a subsequent *in silico* rRNA
390 reduction using the SortmeRNA 2.0 software package (55). Filtered reads were *de novo*
391 assembled with cutoffs of 300 base minimum contig length and average coverage of 2, leaving a
392 total of 705,526 contigs across all samples (Supplementary Table 1, Contig Mappings Tab).

393 *Screening Assemblies for Marker Genes*

394 Marker genes to identify bacteriophage were chosen based on their conservation across
395 phage taxa and their success in other environmental datasets. Gp20 (phage portal protein) and
396 Gp23 (major capsid protein) have been shown previously to be highly conserved and effective
397 for phylogenetic assignment of members of the *Myoviridae* (47-49). RecA is conserved across all
398 three bacteriophage taxa and could illuminate lysogeny, and ribonucleotide reductase (RNR) has
399 been used as an effective marker for screening novel viruses from marine sequencing datasets
400 (50). To identify contigs specific to the NucleoCytoplasmic Large DNA Virus (NCLDV) clade,
401 contig libraries were screened for the presence of 10 genes previously identified as core NCLDV
402 genes as previously described (33). Briefly, contig libraries were queried against Nucleo-
403 Cytoplasmic Virus Orthologous Groups (NCVOG) protein databases for each of the following
404 10 marker genes in a Blastx search with a minimum e-value cutoff of 10^{-3} : A32 virion packaging
405 ATPase (NCVOG0249), VLFT-like transcription factor (NCVOG0262), Superfamily II Helicase
406 II (NCVOG0024), mRNA capping enzyme (NCVOG1117), D5 helicase-primase
407 (NCVOG0023), ribonucleotide reductase small subunit (NCVOG0276), RNA polymerase large
408 subunit (NCVOG0271), RNA polymerase small subunit (NCVOG0274), B-family DNA
409 polymerase (NCVOG0038), and major capsid protein (NCVOG0022). Resulting hits were then
410 queried against the NCBI refseq protein database (56) and only contigs with top hits to virus
411 genes were maintained for subsequent analyses. A similar method was used to identify viroplasm
412 transcripts, where the viroplasm major capsid protein and packaging ATPase genes were used as
413 markers.

414 Contigs derived from ssRNA viruses were identified by screening the contig library for
415 RNA-dependent RNA Polymerase (RDRP), a distinctive and wholly conserved RNA virus gene
416 and a strong phylogenetic marker (57). A BLAST database of RDRP sequences was downloaded

417 from the pfam database (58) under code pf00680. Contigs were aligned using Blastx with a
418 minimum evalue of 10^{-4} . Hits were queried against the NCBI refseq protein database and only
419 hits to viral RDRP genes were retained for downstream analyses. Contigs derived from *rpb1*
420 transcripts were similarly identified using a BLAST database of *rpb1* sequences downloaded
421 from the UniProt database under the K03006 group.

422 To identify RNA virus genome fragments, contig libraries were screened as described
423 above using the following core set of genes observed in RNA viruses: CRPV capsid (Pfam
424 08762), VP4 (Pfam 11492), RDRP (Pfam 00680), Peptidase C3 (Pfam 00548), Peptidase C3G
425 (Pfam 12381), Rhv (Pfam 00073), and RNA Helicase (Pfam 00910). BLAST databases for core
426 RNA virus genes were constructed from reference sequences downloaded from pfam. Query
427 sequences were then cross-referenced to identify contigs with hits to multiple RNA virus core
428 genes. Only contigs > 1000 bases with at least one viral RDRP region were retained for further
429 analysis. ORFs were predicted on these putative partial genomes using the CLC Genomics
430 Workbench. Features on the partial genomes were predicted using the Pfam HMM domain and
431 the NCBI Conserved Domain Database searches (59, 60). Genome architecture was visualized
432 using the Illustrator for Biological Sequences (IBS) software package (61).

433 *Phylogenetic Analysis*

434 Reference sequences for viral marker genes and host Rpb1 were downloaded from the
435 InterPro and RefSeq databases (Supplementary Table 2) (62). Reference sequences were aligned
436 using the MUSCLE alignment algorithm (63) in the MEGA v7.0.26 software package (64).
437 Maximum likelihood phylogenetic trees were constructed in PhyML from reference sequences
438 and contigs containing the respective full length genes (65) with the LG substitution model and
439 the aLRT SH-like likelihood method. Putative viral and Rpb1 contigs assembled from the

440 metatranscriptomes were translated into proteins according to the reading frame of the top
441 BLAST hit. Translated proteins were placed on the reference trees in a maximum likelihood
442 framework in pplacer (66), and contigs were identified based on the most closely related clade in
443 the tree. Trees with abundance data were visualized using the iTOL web interface (67).

444 *Statistical Analysis*

445 Quality filtered and trimmed reads were stringently mapped to the selected contigs (0.97
446 identity fraction, 0.7 length fraction) in CLC Genomics Workbench 10.0.1. Expression values
447 were calculated as a modification of the transcript per million (TPM) metric. Read counts were
448 normalized by contig length in kb to determine the reads per kilobase (RPK) values for every
449 contig within each library. These RPK values were then summed and divided by 1 million, to
450 determine the sequencing depth scaling factor for each library. TPM for a contig was calculated
451 by dividing its RPK value by the scaling factor for the library.

452 Expression values for contigs were imported into the PRIMER7 (68) statistical software
453 package and \log_2 transformed. Expression values from each contig were correlated (Pearson's
454 rho) to one another and statistically grouped by co-occurrence using group average hierarchical
455 clustering. The SIMPROF test (69) was used to determine the statistical significance level of
456 resulting clusters (alpha = 0.05, 1000 permutations). Statistically significant clusters with at least
457 one viral contig, one *rpb1* contig and less than 10 total members were visualized and annotated
458 in Cytoscape 3.5.1 (70).

459 *Accession Numbers*

460 Full RNA-seq libraries have been made publicly available on the JGI website under
461 accession number Gp0146911.

463 *Acknowledgements*

464 Research sponsored by the *Laboratory Directed Research and Development* Program of Oak
465 Ridge National Laboratory and the *Joint Directed Research and Development Program* of the
466 University of Tennessee. Support for the SPRUCE experimental site is from the U.S. Department
467 of Energy, Office of Science, Office of Biological and Environmental Research. Oak Ridge
468 National Laboratory is managed by UT- Battelle, LLC, for the U.S. Department of Energy under
469 contract DE-AC05-00OR22725. Support at UT was received from the *Kenneth & Blaire*
470 *Mossman Endowment* to the University of Tennessee (SWW).

471 **References**

472

- 473 1. **Post WM, Emanuel WR, Zinke PJ, Stangenberger AG.** 1982. Soil carbon pools and
474 world life zones. *Nature* **298**:156-159.
- 475 2. **Gorham E.** 1991. Northern peatlands: role in the carbon cycle and probable responses to
476 climatic warming. *Ecological applications* **1**:182-195.
- 477 3. **Bridgham SD, Patrick Megonigal J, Keller JK, Bliss NB, Trettin C.** 2006. The carbon
478 balance of North American wetlands. *Wetlands* **26**:889-916.
- 479 4. **van Breemen N.** 1995. How Sphagnum bogs down other plants. *Trends in Ecology &*
480 *Evolution* **10**:270-275.
- 481 5. **Lamers LPM, Bobbink R, Roelofs JGM.** 2000. Natural nitrogen filter fails in polluted
482 raised bogs. *Global Change Biology* **6**:583-586.
- 483 6. **Turetsky MR.** 2003. The role of bryophytes in carbon and nitrogen cycling. *Bryologist*
484 **106**:395-409.
- 485 7. **Turetsky MR, Bond-Lamberty B, Euskirchen E, Talbot J, Frohling S, McGuire AD,**
486 **Tuittila ES.** 2012. The resilience and functional role of moss in boreal and arctic
487 ecosystems. *New Phytologist* **196**:49-67.
- 488 8. **Verhoeven JTA, Liefveld WM.** 1997. The ecological significance of organochemical
489 compounds in Sphagnum. *Acta Botanica Neerlandica* **46**:117-130.
- 490 9. **Mellegard H, Stalheim T, Hormazabal V, Granum PE, Hardy SP.** 2009.
491 Antibacterial activity of sphagnum acid and other phenolic compounds found in
492 Sphagnum papillosum against food-borne bacteria. *Letters in Applied Microbiology*
493 **49**:85-90.
- 494 10. **Freeman C, Ostle N, Kang H.** 2001. An enzymic 'latch' on a global carbon store - A
495 shortage of oxygen locks up carbon in peatlands by restraining a single enzyme. *Nature*
496 **409**:149-149.
- 497 11. **Stalheim T, Ballance S, Christensen BE, Granum PE.** 2009. Sphagnum - a pectin-like
498 polymer isolated from Sphagnum moss can inhibit the growth of some typical food
499 spoilage and food poisoning bacteria by lowering the pH. *Journal of Applied*
500 *Microbiology* **106**:967-976.
- 501 12. **Hajek T, Ballance S, Limpens J, Zijlstra M, Verhoeven JTA.** 2011. Cell-wall
502 polysaccharides play an important role in decay resistance of Sphagnum and actively
503 depressed decomposition in vitro. *Biogeochemistry* **103**:45-57.
- 504 13. **Lin X, Tfaily MM, Green SJ, Steinweg JM, Chanton P, Invittaya A, Chanton JP,**
505 **Cooper W, Schadt C, Kostka JE.** 2014. Microbial Metabolic Potential for Carbon
506 Degradation and Nutrient (Nitrogen and Phosphorus) Acquisition in an Ombrotrophic
507 Peatland. *Applied and Environmental Microbiology* **80**:3531-3540.
- 508 14. **Leppanen S, Rissanen A, Tirola M.** 2015. Nitrogen fixation in Sphagnum mosses is
509 affected by moss species and water table level. *Plant and Soil* **389**:185-196.
- 510 15. **Kostka JE, Weston DJ, Glass JB, Lilleskov EA, Shaw AJ, Turetsky MR.** 2016. The
511 Sphagnum microbiome: new insights from an ancient plant lineage. *New Phytologist*
512 **211**:57-64.

24

- 513 16. **Dudova L, Hajkova P, Buchtova H, Opravilova V.** 2013. Formation, succession and
514 landscape history of Central-European summit raised bogs: A multiproxy study from the
515 Hruby Jeseník Mountains. *Holocene* **23**:230-242.
- 516 17. **Ireland AW, Clifford MJ, Booth RK.** 2014. Widespread dust deposition on North
517 American peatlands coincident with European land-clearance. *Vegetation History and*
518 *Archaeobotany* **23**:693-700.
- 519 18. **Swindles GT, Turner TE, Roe HM, Hall VA, Rea HA.** 2015. Testing the cause of the
520 *Sphagnum austinii* (Sull. ex Aust.) decline: Multiproxy evidence from a raised bog in
521 Northern Ireland. *Review of Palaeobotany and Palynology* **213**:17-26.
- 522 19. **Galka M, Tobolski K, Gorska A, Lamentowicz M.** 2017. Resilience of plant and
523 testate amoeba communities after climatic and anthropogenic disturbances in a Baltic bog
524 in Northern Poland: Implications for ecological restoration. *Holocene* **27**:130-141.
- 525 20. **Opelt K, Chobot V, Hadacek F, Schonmann S, Eberl L, Berg G.** 2007. Investigations
526 of the structure and function of bacterial communities associated with *Sphagnum* mosses.
527 *Environmental Microbiology* **9**:2795-2809.
- 528 21. **Bragina A, Cardinale M, Berg C, Berg G.** 2013. Vertical transmission explains the
529 specific Burkholderia pattern in *Sphagnum* mosses at multi-geographic scale. *Frontiers in*
530 *Microbiology* **4**:10.
- 531 22. **Bragina A, Maier S, Berg C, Muller H, Chobot V, Hadacek F, Berg G.** 2012. Similar
532 diversity of Alphaproteobacteria and nitrogenase gene amplicons on two related
533 *Sphagnum* mosses. *Frontiers in Microbiology* **3**:10.
- 534 23. **Opelt K, Berg G.** 2004. Diversity and antagonistic potential of bacteria associated with
535 bryophytes from nutrient-poor habitats of the Baltic Sea coast. *Applied and*
536 *Environmental Microbiology* **70**:6569-6579.
- 537 24. **Brussaard CPD, Wilhelm SW, Thingstad F, Weinbauer MG, Bratbak G, Heldal M,**
538 **Kimmance SA, Middelboe M, Nagasaki K, Paul JH, Schroeder DC, Suttle CA,**
539 **Vaque D, Wommack KE.** 2008. Global-scale processes with a nanoscale drive: the role
540 of marine viruses. *ISME J* **2**:575-578.
- 541 25. **Jover LF, Effler TC, Buchan A, Wilhelm SW, Weitz JS.** 2014. The elemental
542 composition of virus particles: implications for marine biogeochemical cycles. *Nat Rev*
543 *Micro* **12**:519-528.
- 544 26. **Wilhelm SW, Suttle CA.** 1999. Viruses and Nutrient Cycles in the SeaViruses play
545 critical roles in the structure and function of aquatic food webs. *BioScience* **49**:781-788.
- 546 27. **Thingstad TF, Lignell R.** 1997. Theoretical models for the control of bacterial growth
547 rate, abundance, diversity and carbon demand. *Aquatic Microbial Ecology* **13**:19-27.
- 548 28. **Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, Poulos BT,**
549 **Solonenko N, Lara E, Poulain J, Pesant S, Kandels-Lewis S, Dimier C, Picheral M,**
550 **Searson S, Cruaud C, Alberti A, Duarte CM, Gasol JM, Vaque D, Bork P, Acinas**
551 **SG, Wincker P, Sullivan MB, Tara Oceans C.** 2016. Ecogenomics and potential
552 biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**:689-+.
- 553 29. **Simmonds P, Adams MJ, Benko M, Breitbart M, Brister JR, Carstens EB, Davison**
554 **AJ, Delwart E, Gorbalenya AE, Harrach B, Hull R, King AMQ, Koonin EV,**
555 **Krupovic M, Kuhn JH, Lefkowitz EJ, Nibert ML, Orton R, Roossinck MJ,**
556 **Sabanadzovic S, Sullivan MB, Suttle CA, Tesh RB, van der Vlugt RA, Varsani A,**
557 **Zerbini M.** 2017. Virus taxonomy in the age of metagenomics. *Nature Reviews*
558 *Microbiology* **15**:161-168.

- 559 30. **Steward GF, Culley AI, Mueller JA, Wood-Charlson EM, Belcaid M, Poisson G.**
560 2013. Are we missing half of the viruses in the ocean? *The ISME Journal* **7**:672-679.
- 561 31. **Wilhelm SW, Bird JT, Bonifer KS, Calfee BC, Chen T, Coy SR, Gainer PJ, Gann**
562 **ER, Heatherly HT, Lee J, Liang XL, Liu J, Armes AC, Moniruzzaman M, Rice JH,**
563 **Stough JMA, Tams RN, Williams EP, LeCclair GR.** 2017. A Student's Guide to Giant
564 Viruses Infecting Small Eukaryotes: From Acanthamoeba to Zooxanthellae. *Viruses-*
565 *Basel* **9**:18.
- 566 32. **Yutin N, Wolf YI, Raoult D, Koonin EV.** 2009. Eukaryotic large nucleo-cytoplasmic
567 DNA viruses: Clusters of orthologous genes and reconstruction of viral genome
568 evolution. *Virology Journal* **6**:13.
- 569 33. **Moniruzzaman M, Wurch LL, Alexander H, Dyhrman ST, Gobler CJ, Wilhelm**
570 **SW.** 2017. Virus-host relationships of marine single-celled eukaryotes resolved from
571 metatranscriptomics. *Nature Communications* **8**:10.
- 572 34. **Schulz F, Yutin N, Ivanova NN, Ortega DR, Lee TK, Vierheilig J, Daims H, Horn**
573 **M, Wagner M, Jensen GJ, Kyrpides NC, Koonin EV, Woyke T.** 2017. Giant viruses
574 with an expanded complement of translation system components. *Science* **356**:82-+.
- 575 35. **Moniruzzaman M, LeCclair GR, Brown CM, Gobler CJ, Bidle KD, Wilson WH,**
576 **Wilhelm SW.** 2014. Genome of brown tide virus (AaV), the little giant of the
577 Megaviridae, elucidates NCLDV genome expansion and host-virus coevolution.
578 *Virology* **466-467**:60-70.
- 579 36. **Moniruzzaman M, Gann ER, Wilhelm SW.** 2018. Infection by a Giant Virus (AaV)
580 Induces Widespread Physiological Reprogramming in *Aureococcus anophagefferens*
581 CCMP1984 – A Harmful Bloom Algae. *Frontiers in Microbiology* **9**.
- 582 37. **Wilhelm SW, Coy SR, Gann ER, Moniruzzaman M, Stough JMA.** 2016. Standing on
583 the shoulders of giant viruses: five lessons learned about large viruses infecting small
584 eukaryotes and the opportunities they create. *Plos Pathogens* **12**:5.
- 585 38. **Kerepesi C, Grolmusz V.** 2017. The "Giant Virus Finder" discovers an abundance of
586 giant viruses in the Antarctic dry valleys. *Archives of Virology* **162**:1671-1676.
- 587 39. **Rusin LY.** 2016. Metagenomics and biodiversity of sphagnum bogs. *Molecular Biology*
588 **50**:645-648.
- 589 40. **Krupovic M, Koonin EV.** 2014. Evolution of eukaryotic single-stranded DNA viruses of
590 the Bidnaviridae family from genes of four other groups of widely different viruses.
591 *Scientific Reports* **4**:5347.
- 592 41. **Krupovic M, Koonin EV.** 2015. Polintons: a hotbed of eukaryotic virus, transposon and
593 plasmid evolution. *Nature Reviews Microbiology* **13**:105.
- 594 42. **Tomaru Y, Nagasaki K.** 2007. Flow cytometric detection and enumeration of DNA and
595 RNA viruses infecting marine eukaryotic microalgae. *Journal of Oceanography* **63**:215-
596 221.
- 597 43. **Miranda JA, Culley AI, Schvarcz CR, Steward GF.** 2016. RNA viruses as major
598 contributors to Antarctic virioplankton. *Environmental Microbiology* **18**:3714-3727.
- 599 44. **Quaiser A, Krupovic M, Dufresne A, Francez A-J, Roux S.** 2016. Diversity and
600 comparative genomics of chimeric viruses in Sphagnum-dominated peatlands. *Virus*
601 *Evolution* **2**:vew025-vew025.
- 602 45. **Quaiser A, Dufresne A, Ballaud F, Roux S, Zivanovic Y, Colombet J, Sime-Ngando**
603 **T, Francez A-J.** 2015. Diversity and comparative genomics of Microviridae in
604 Sphagnum- dominated peatlands. *Frontiers in Microbiology* **6**.

- 605 46. **Ballaud F, Dufresne A, Francez A-J, Colombet J, Sime-Ngando T, Quaiser A.** 2015.
606 Dynamics of Viral Abundance and Diversity in a Sphagnum-Dominated Peatland:
607 Temporal Fluctuations Prevail Over Habitat. *Frontiers in Microbiology* **6**:1494.
- 608 47. **Dorigo U, Jacquet S, Humbert JF.** 2004. Cyanophage diversity, inferred from g20 gene
609 analyses, in the largest natural lake in France, Lake Bourget. *Applied and Environmental*
610 *Microbiology* **70**:1017-1022.
- 611 48. **Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-**
612 **Ngando T, Debroas D.** 2012. Assessing the Diversity and Specificity of Two Freshwater
613 Viral Communities through Metagenomics. *Plos One* **7**:12.
- 614 49. **Comeau AM, Krisch HM.** 2008. The capsid of the T4 phage superfamily: The
615 evolution, diversity, and structure of some of the most prevalent proteins in the biosphere.
616 *Molecular Biology and Evolution* **25**:1321-1332.
- 617 50. **Sakowski EG, Munsell EV, Hyatt M, Kress W, Williamson SJ, Nasko DJ, Polson**
618 **SW, Wommack KE.** 2014. Ribonucleotide reductases reveal novel viral diversity and
619 predict biological and ecological features of unknown marine viruses. *Proceedings of the*
620 *National Academy of Sciences of the United States of America* **111**:15786-15791.
- 621 51. **Alexander H, Jenkins BD, Rynearson TA, Dyhrman ST.** 2015. Metatranscriptome
622 analyses indicate resource partitioning between diatoms in the field. *Proceedings of the*
623 *National Academy of Sciences* **112**:E2182-E2190.
- 624 52. **Wilson RM, Hopple AM, Tfaily MM, Sebestyen SD, Schadt CW, Pfeifer-Meister L,**
625 **Medvedeff C, McFarlane KJ, Kostka JE, Kolton M, Kolka RK, Kluber LA, Keller**
626 **JK, Guilderson TP, Griffiths NA, Chanton JP, Bridgham SD, Hanson PJ.** 2016.
627 Stability of peatland carbon to rising temperatures. *Nature Communications* **7**:10.
- 628 53. **Hanson PJ, Riggs JS, Nettles WR, Phillips JR, Krassovski MB, Hook LA, Gu L,**
629 **Richardson AD, Aubrecht DM, Ricciuto DM.** 2017. Attaining whole-ecosystem
630 warming using air and deep-soil heating methods with an elevated CO₂ atmosphere.
631 *Biogeosciences* **14**:861.
- 632 54. **Warren MJ, Lin XJ, Gaby JC, Kretz CB, Kolton M, Morton PL, Pett-Ridge J,**
633 **Weston DJ, Schadt CW, Kostka JE, Glass JB.** 2017. Molybdenum-Based Diazotrophy
634 in a Sphagnum Peatland in Northern Minnesota. *Applied and Environmental*
635 *Microbiology* **83**:14.
- 636 55. **Kopylova E, Noe L, Touzet H.** 2012. SortMeRNA: fast and accurate filtering of
637 ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**:3211-3217.
- 638 56. **O'Leary NA, Wright MW, Brister JR, Ciufu S, Haddad D, McVeigh R, Rajput B,**
639 **Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y,**
640 **Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM,**
641 **Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li**
642 **W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S,**
643 **Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H,**
644 **Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W,**
645 **Landrum MJ, Kimchi A, et al.** 2016. Reference sequence (RefSeq) database at NCBI:
646 current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*
647 **44**:D733-745.
- 648 57. **Koonin EV.** 1991. The phylogeny of RNA-dependent RNA polymerases of positive-
649 strand RNA viruses. *Journal of General Virology* **72**:2197-2206.

- 650 58. **Finn RD, Coghill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC,**
651 **Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A.** 2016.
652 The Pfam protein families database: towards a more sustainable future. *Nucleic Acids*
653 *Research* **44**:D279-D285.
- 654 59. **Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A,**
655 **Eddy SR.** 2015. HMMER web server: 2015 update. *Nucleic Acids Research* **43**:W30-
656 W38.
- 657 60. **Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu SN, Chitsaz F, Geer LY, Geer**
658 **RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS,**
659 **Thanki N, Wang ZX, Yamashita RA, Zhang DC, Zheng CJ, Bryant SH.** 2015. CDD:
660 NCBI's conserved domain database. *Nucleic Acids Research* **43**:D222-D226.
- 661 61. **Liu WZ, Xie YB, Ma JY, Luo XT, Nie P, Zuo ZX, Lahrmann U, Zhao Q, Zheng YY,**
662 **Zhao Y, Xue Y, Ren J.** 2015. IBS: an illustrator for the presentation and visualization of
663 biological sequences. *Bioinformatics* **31**:3359-3361.
- 664 62. **Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY,**
665 **Dosztanyi Z, El-Gebali S, Fraser M, Gough J, Haft D, Holliday GL, Huang HZ,**
666 **Huang XS, Letunic I, Lopez R, Lu SN, Marchler-Bauer A, Mi HY, Mistry J, Natale**
667 **DA, Necci M, Nuka G, Orengo CA, Park Y, Pesseat S, Piovesan D, Potter SC,**
668 **Rawlings ND, Radaschi N, Richardson L, Rivoire C, Sangrador-Vegas A, Sigrist C,**
669 **Sillitoe I, Smithers B, Squizzato S, Sutton G, Thanki N, Thomas PD, Tosatto SCE,**
670 **Wu CH, Xenarios I, Yeh LS, Young SY, Mitchell AL.** 2017. InterPro in 2017-beyond
671 protein family and domain annotations. *Nucleic Acids Research* **45**:D190-D199.
- 672 63. **Edgar RC.** 2004. MUSCLE: a multiple sequence alignment method with reduced time
673 and space complexity. *Bmc Bioinformatics* **5**:1-19.
- 674 64. **Kumar S, Stecher G, Tamura K.** 2016. MEGA7: Molecular evolutionary genetics
675 analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* **33**:1870-1874.
- 676 65. **Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O.** 2010.
677 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the
678 performance of PhyML 3.0. *Systematic Biology* **59**:307-321.
- 679 66. **Matsen FA, Kodner RB, Armbrust EV.** 2010. pplacer: linear time maximum-
680 likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree.
681 *Bmc Bioinformatics* **11**:16.
- 682 67. **Letunic I, Bork P.** 2016. Interactive tree of life (iTOL) v3: an online tool for the display
683 and annotation of phylogenetic and other trees. *Nucleic Acids Research* **44**:W242-W245.
- 684 68. **Clark KR, Gorley RN.** 2015. PRIMER v7: User manual/tutorial. PRIMER-E, Plymouth.
- 685 69. **Clarke KR, Somerfield PJ, Gorley RN.** 2008. Testing of null hypotheses in exploratory
686 community analyses: similarity profiles and biota-environment linkage. *Journal of*
687 *Experimental Marine Biology and Ecology* **366**:56-69.
- 688 70. **Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N,**
689 **Schwikowski B, Ideker T.** 2003. Cytoscape: A software environment for integrated
690 models of biomolecular interaction networks. *Genome Research* **13**:2498-2504.

691

692

693 **Figure Legends**

694 Figure 1: Phylogenetic placement of identified phage major capsid protein contigs (red) on a
695 Myovirus *gp23* maximum likelihood reference tree (references in black). Full alignment length
696 477 amino acids. Node support (aLRT-SH statistic) > 50% are shown. Contigs are shown with
697 their abundance (\log_2 transformed TPM) in a heatmap surrounding the tree. Sample order on the
698 heatmap is provided in the inset.

699 Figure 2: Phylogenetic placement of identified ssRNA virus RNA-dependent RNA polymerase
700 contigs on maximum likelihood reference tree. Full alignment length 551 amino acids. Branch
701 width represents the number of contigs placed on the reference branch. Node support (aLRT-SH
702 statistic) >50% are shown.

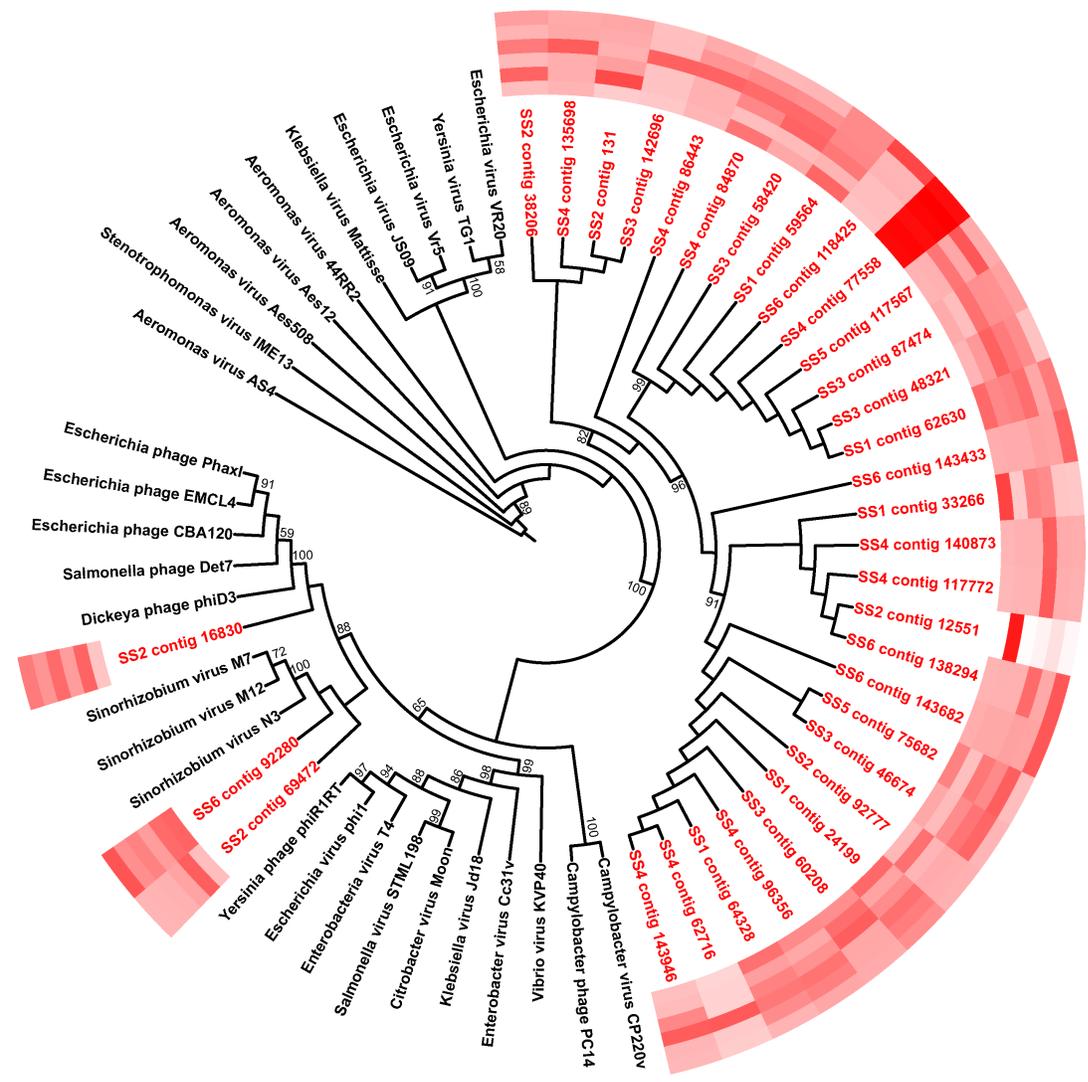
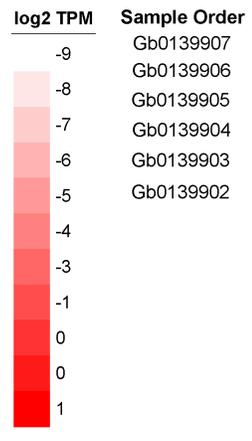
703 Figure 3: Phylogeny, genome architecture, and abundance of partial ssRNA virus genomes. Tree
704 represents phylogenetic placement of RDRP gene regions from partial ssRNA virus genome
705 contigs (red) on a maximum likelihood reference tree (references in black). Full alignment length
706 551 amino acids. Node support (aLRT-SH statistic) >50% are shown. Center panel represents
707 genome architecture determined by conserved domain search and ORF prediction. Length of
708 contigs and gene regions is measured in kb. Heat map in right panel shows abundance of reads
709 mapped to partial genome contigs in \log_2 TPM from each of the 6 metatranscriptome libraries.

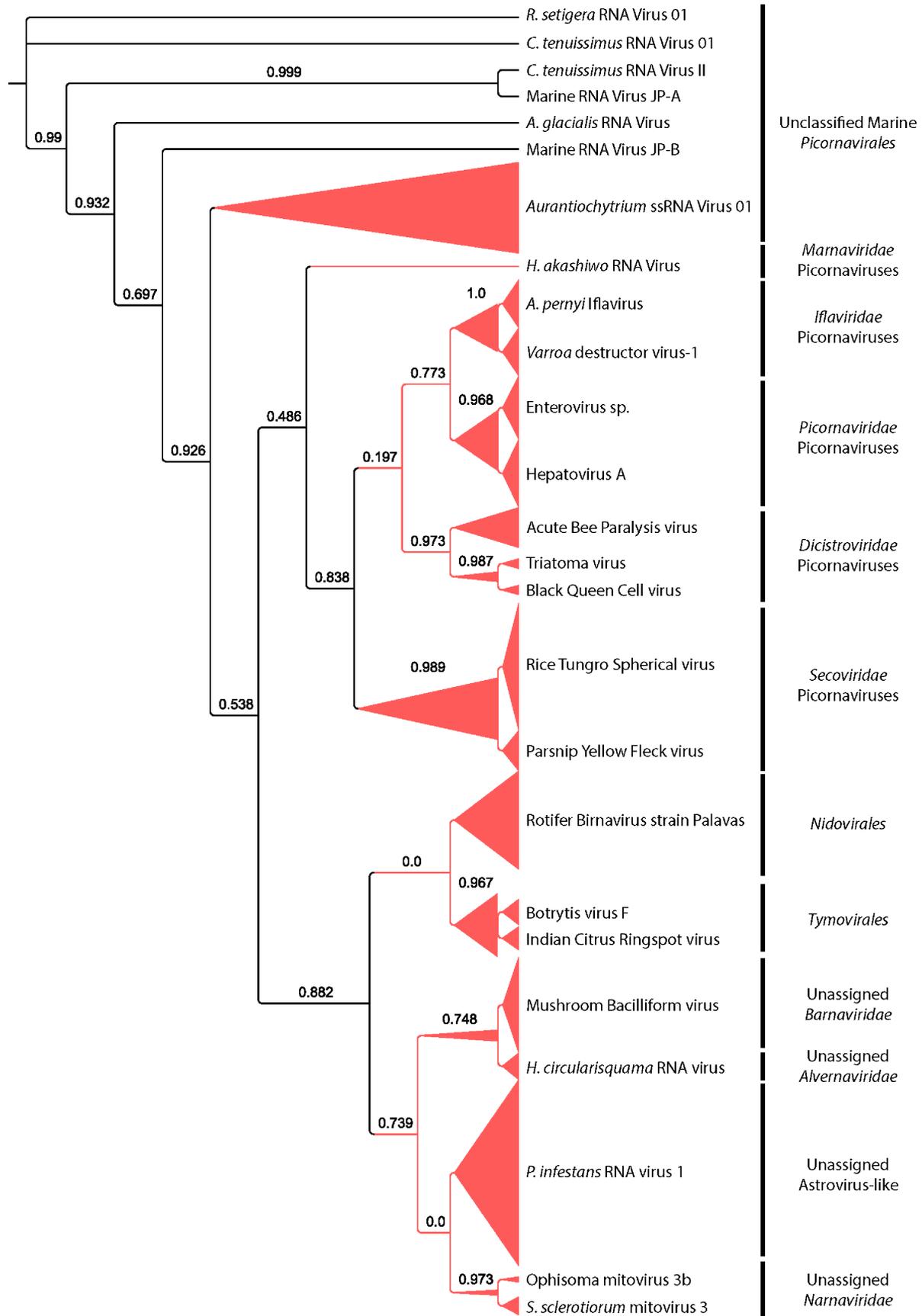
710 Figure 4. Phylogenetic placement of identified NCLDV major capsid protein contigs (red) on a
711 maximum likelihood reference tree (references in black). Full alignment length 477 amino acids.
712 Node support (aLRT-SH statistic) >50% are shown. Contigs are shown with their abundance
713 (\log_2 transformed TPM) in a heatmap surrounding the tree.

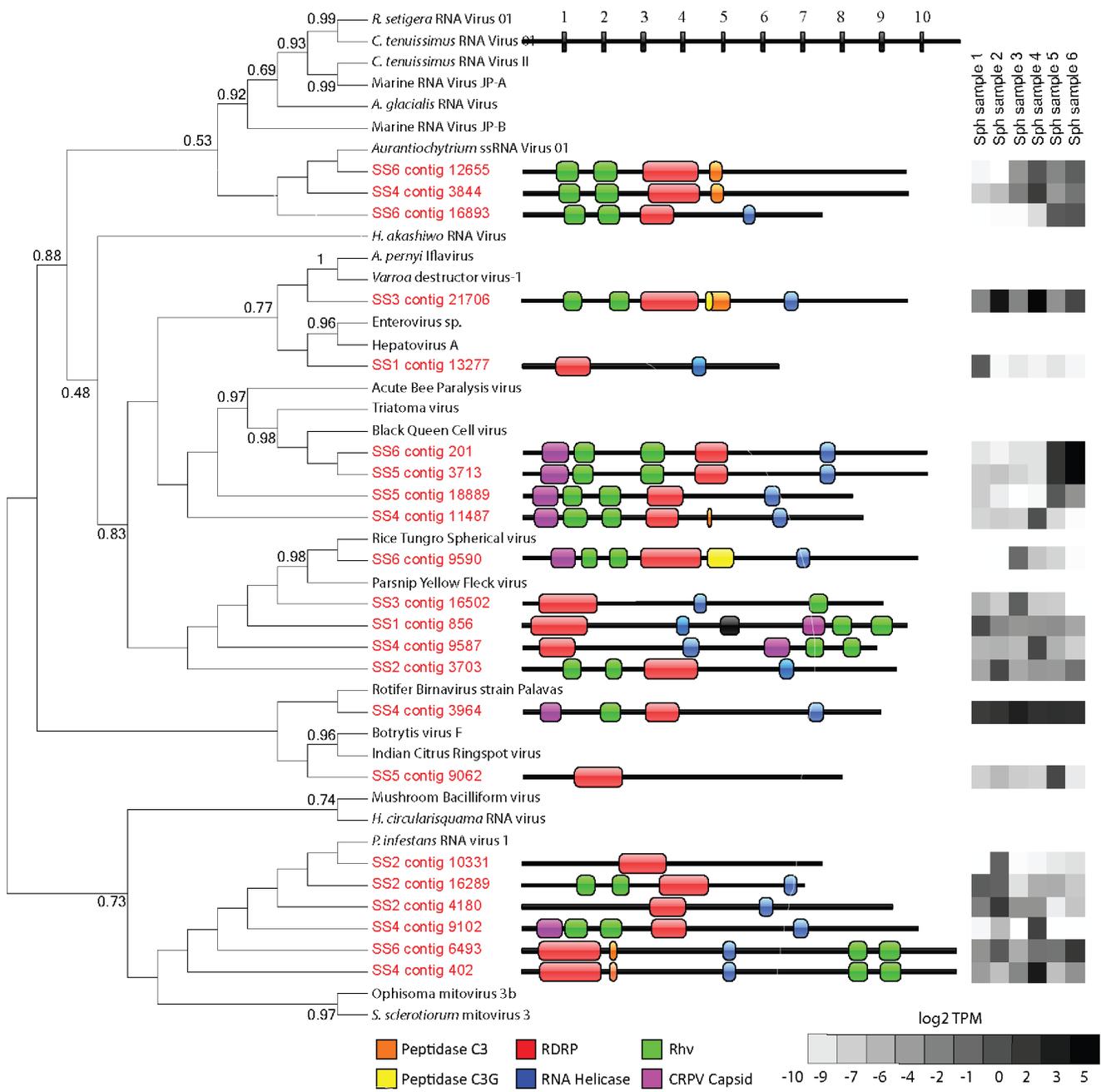
714 Figure 5: Phylogenetic placement of identified virophage A.) major capsid protein (full
715 alignment length 549 amino acids) and B.) ATPase (full alignment length 251 amino acids)
716 contigs (red) on a maximum likelihood reference tree (references in black). Node support (aLRT-
717 SH statistic) >50% are shown.

718 Figure 6: Correlation co-occurrence network analysis of conserved viral gene and host RNA
719 polymerase expression for A.) bacteriophage (Gp23), B.) ssRNA viruses (RDRP), and C.)
720 NCLDV_s (NCLDV MCP). Nodes in red represent virus contigs and blue nodes represent
721 potential hosts. Nodes are connected by edges colored according to the Pearson correlation
722 coefficient values between to contigs. Only relationships with contigs expressed in more than
723 one sample are shown.

724









Sample order

- Gb0139907
- Gb0139906
- Gb0139905
- Gb0139904
- Gb0139903
- Gb0139901

