For submission to: *Applied and Environmental Microbiology*

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

Authors:

Joshua M.A. Stough[1*], Max Kolton[2], Joel E. Kostka[2], David J. Weston[3,4], Dale A. Pelletier[3], and Steven W. Wilhelm[1]#

Addresses:

[1] Department of Microbiology, University of Tennessee, Knoxville, Tennessee, United States of America 37996

[2] School of Biology and School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, GA, USA

[3] Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

[4] Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

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

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

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

Abstract

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

Significance

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

Introduction

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

While there is a growing body of research characterizing the microbial groups that colonize *Sphagnum* (15), little is currently known about the ecological factors that define community structure and ecosystem function. Studies suggest that subtle differences in pH and available nutrients, manipulated by different *Sphagnum* species and strains, create distinct microbial consortia (14, 20, 21). Other observations suggest a more homogenous community (22), highlighting a need for further study. Culture-dependent experiments isolating endophytic bacteria indicate *Sphagnum* cultivates symbionts with abilities that include antifungal activity (20, 23) and nitrogen fixation (14), and that these microbiomes may be passed vertically to the moss progeny (21). Yet while examinations of how environmental conditions and host-microbe symbiotic interactions shape the structure and function of microbial communities, the influence of virus populations on the *Sphagnum* microbiome remains unexplored.

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

As lab studies of viruses require hosts that can be grown in culture, many environmentally relevant viruses are poorly understood and their representation in reference databases is often skewed. Previous efforts to describe environmental viromes have focused primarily on the sequencing of shotgun or PCR-targeted metagenomes. While these methods have proven powerful, rapidly expanding available reference material for bacteriophage (28, 29), it leaves the considerable diversity of RNA viruses largely untapped (30). Moreover, the common approach of selecting for viruses based on size-exclusion with filters removes many of the Nucleo-Cytoplasmic Large DNA Viruses (NCLDVs, or commonly "giant viruses") that are also environmentally relevant and phylogenetically informative (31, 32). Metagenomic sequencing also limits observations to virus particles: from these data inferences on viral activity require tenuous assumptions. The advent of high-throughput RNA sequencing offers viral ecologists the opportunity to study active infections in the environment, as DNA viruses only produce transcripts inside a host. Moreover, this approach also captures fragments of RNA virus genomes. When sequencing is of sufficient depth and multiple samples are collected with spatial and temporal variability, these data present an opportunity to develop hypothetical relationships between virus and host markers (33) for subsequent in lab testing.

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

Results

Identification of resident phage populations

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

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

Single-stranded RNA virus diversity and abundance

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

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

Among these, 22 contigs were found to be near complete ssRNA virus genomes (based on gene content and size), encoding multiple viral genes in addition to RDRP. Gene regions were identified and annotated using the NCBI conserved domain and PFam HMM search tools, and the full-length RDRP sequence was used to construct a maximum likelihood phylogenetic tree (Fig 3). Of the partial ssRNA genomes that were assembled, 2 were missing the conserved Rhv structural genes, while one was missing a RNA virus Helicase. The majority of these contigs fall under the *Picornavirales* order, which also included the most complete viral genomes. As was observed with the shorter RDRP contigs above, most of the *Picornavirales* contigs were most closely related to the unclassified marine species, or members of the family *Secoviridae* clade, whose membership includes the Parsnip yellow fleck virus. A number of partial *Picornavirus* genomes were also identified as members of the family *Dicistroviridae*. Outside the *Picornavirales,* most contigs clustered closely with the unassigned Astrovirus-like *Phytophthora infestans* RNA virus. To determine the relative abundance of different RNA virus genomes in the peat bog samples, we mapped reads back to contigs and calculated transcripts per million (TPM) values to account for contig length and library size. The most abundant contig across all samples was SS4 contig 3964, which was most closely related to the Rotifer birnavirus. All other contigs appear to be abundant prominently in one or two samples, and absent or in low abundance in the others, with no patterns of abundance apparent.

Giant viruses and virophage in Sphagnum microbiome

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

As was observed with the other major viral taxa described, the majority of contigs were most abundantly expressed in one or two samples and present at very low levels in the rest. The most abundant NCLDV-MCP contig in the samples was SS2 contig 73240, most closely related to *Megavirus chilensis*, which was the most highly expressed giant virus contig across all samples. Four other contigs (SS6 contig 110585, SS4 contigs 55722 and 141177, and SS5 contig 119519) were highly expressed across all six samples.

Prediction of virus-host pairs

By comparing and correlating expression of virus marker genes to *rpb1* expression from cellular organisms, we endeavored to predict potential virus-host groups in the *Sphagnum* phyllosphere. Fig 6 shows statistically robust networks containing at least one virus and one host, where co-occurrence and correlation were observed in more than one sample. A total of 13 virus-host groups were detected, spread across the major viral taxa detected in this dataset. We note that no networks containing the virophage/polinton-like viruses emerged. Four relationships were predicted from bacteriophage *gp23* abundance, the simplest of which was a *Tevenvirinae* phage-*Metazoa*-*Rhizaria* group with moderate correlations (Fig 6a). The other 3 relationships are more complicated, containing multiple potential hosts and, for the largest predicted group, multiple virus transcripts. The majority of potential hosts in these groups were identified as eukaryotic, with only one putative bacterium and two archaea. Correlation coefficients for the phage-prokaryote clusters were lower than was observed in the other major viral taxa, with low to moderate correlations between viruses and bacteria.

We observed 4 predicted RNA virus-host clusters, all of which contained multiple hosts grouped with a single virus (Fig 6b). Most of the predicted hosts appear closely related to eukaryotic single-celled protists, within the *Excavata* and *Rhizaria* supergroups. Correlation coefficients observed in these relationships are generally higher than observed in the phage-host clusters. The 5 predicted NCLDV-host clusters (Fig 6c) were the most highly correlated and complex. Predicted hosts were highly varied, ranging from diatoms to animals, though all virus members were placed either within *Mimiviridae* or the extended Mimivirus group. MCP contigs originating in close relatives of the recently discovered Klosneuviruses are present in both the 7- and 10-member clusters, in addition to a pair of contigs most closely related to *Aureococcus anophagefferens* Virus (AaV). An additional 15 statistically significant clusters across all three viral taxa were observed where the virus and host were present in only one sample (not shown).

Discussion

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

Viral diversity and activity in Sphagnum plants

As viruses lack a universal genetic marker like the bacterial 16S rRNA gene, we opted to screen metatranscriptome assemblies for genes previously demonstrated to be largely or wholly conserved amongst individual viral taxa. Within the expanded and diverse genetic potential of giant viruses, only a handful of genes are currently conserved amongst all members (32, 35) and these, in addition to several markers conserved amongst a large portion of giant viruses were used to identify activity in the *Sphagnum* phyllosphere. Out of the 10 genes used to screen the metatranscriptomes, we only MCP transcripts. This is not surprising given the number of capsid proteins needed for viral assembly: indeed this transcriptional pattern was previously observed in both cultures (36) and marine systems by Moniruzzaman *et al.* (2017). It should be noted that the RNA-seq dataset used in those studies was poly-A selected, enriching for eukaryotic transcripts, and thus coverage of eukaryotic virus gene expression would be much higher than in the *Sphagnum* metatranscriptome. That we observed MCP expression in abundance suggests a significant number of infections occurred at the time of sampling. While the magnitude of giant virus diversity in *Sphagnum* dominated ecosystems is, to our knowledge, completely unexplored, the richness observed here is considerably larger than expected compared to better documented systems. 64 distinct MCP genotypes were identified in the *Sphagnum* phyllosphere metatranscriptomes, which is high when compared to one recent survey that identified 30 novel MCP transcripts from multiple environmental datasets (37), and another which observed 107 NCLDV sequences in 16 publicly available environmental metagenomes of comparable sequencing depth isolated from different ecosystems (38). Most of the MCP contigs identified were placed in clusters around a small number of virus relatives, highlighting the under-sampled diversity of giant viruses in the literature, poor representation in reference databases, and the considerable diversity present in *Sphagnum* peat bogs. The significant giant virus diversity observed here implies a corresponding eukaryotic richness that is also under-described (39). Additionally, a series of virophage transcripts were detected, indicating a significant response to infections by giant viruses in the system. Many of these are phylogenetically grouped with the polintoviruses, transposable elements that produce virion particles that can exploit the replication machinery of actively infecting giant viruses to reproduce, often at the expense of the giant (40, 41). These observations suggest that while an active picoeukaryotic population may persist, mortality mechanisms beyond grazer-driven losses are at play and likely important to carbon flow in the system.

The use of RNA-seq presents a unique opportunity to capture the genomic material of RNA viruses that is lost in metagenomic sequencing. As such, RNA virus representation in sequencing databases and the literature is largely constrained to culture-based studies. All known RNA viruses require a functional RNA-dependent RNA polymerase (RdRP) to copy their genome inside the host cell, a function exclusive to viruses, making it a highly specific marker for RNA virus discovery (42, 43). Recent attempts to use metatranscriptomes to describe environmental RNA viruses have proven successful, not only identifying marker gene fragments in datasets, but assembling complete and near-complete genomes (33, 43). The diversity and composition of RNA virus populations in *Sphagnum* peatlands is largely unknown: it is currently limited to the small group of RNA-DNA hybrid chimeric Cruciviruses (44). Here, as was observed with the giant viruses, most RNA virus contigs were placed in clusters with a single represented species, suggesting a significant degree of uncharacterized diversity. This is not entirely surprising, as RNA viruses are expected to make up as much as half of the virus particles in the Earth's oceans, and yet they are almost as poorly understood and represented in sequencing databases as giant viruses (30). Similarly, we assembled and identified 22 near-complete RNA virus genomes, where completeness was determined primarily by size and the presence of the 6 core genes. As there are currently only 265 sequenced genomes within the *Picornavirales*, most of which grouped within the *Picornaviridae*, this represents a sizeable addition to the known diversity of ssRNA viruses. This is especially true for the unassigned and unclassified taxa, and establishes a strong foundation for future efforts to describe RNA virus populations in *Sphagnum*.

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

Virus-host predictions

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

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

All together, we have identified a considerable amount of viral diversity from several major viral taxa active within a poorly understood microbial ecosystem. As they were identified from transcript sequencing data, the viruses described here likely only represent a fraction of the whole virus community, which may be elucidated through further culture-independent work. We have also used transcript abundance within a statistical framework to predict several host-virus relationships which can be sought out and tested in culture. These results establish an important and much needed foundation for future research into the microbial ecology in *Sphagnum* peat bogs.

Materials and Methods

Sample collection and Survey of Environmental Conditions

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

RNA Extraction and Sequencing

One gram of *Sphagnum* phyllosphere tissue was ground with a mortar and pestle under liquid nitrogen. The fine powder was transferred to 10 extraction tubes and total RNA isolated using the PowerPlant RNA Isolation Kit with DNase according to the manufacturer's protocol (MoBio Laboratories, Carlsbad, CA, USA). DNA-depleted RNA was quantified using the Qubit RNA HS Assay Kit (Invitrogen, Carlsbad, CA, USA) and quality was assessed on the Agilent 2100 BioAnalyzer using the Agilent RNA 6000 Pico Kit (Agilent Technologies). Additionally, the absence of DNA contamination was confirmed by running a polymerase chain reaction using universal bacterial 16S rRNA primers 515F and 806R. Finally, RNA samples without detectible

DNA contamination and exhibiting an RNA integrity number (RIN) > 6 were pooled. Extracted total environmental RNA samples were was sent on dry ice to the Joint Genome Institute (JGI) facilities for meta-transcriptomes libraries construction and sequencing. All protocols employed were standard JGI protocols Ribosomal RNA subtraction from total environmental RNA was completed using the Ribo-Zero rRNA Removal Kit (Illumina, San Diego, CA). rRNA depleted environmental RNA were used to construct paired end metatranscriptomes libraries using TruSeq kit and sequenced on the Illumina HiSeq2000 platform at the JGI facilities using a single-end 250bp flow cell.

RNA-seq Data Processing

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

Screening Assemblies for Marker Genes

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

Contigs derived from ssRNA viruses were identified by screening the contig library for RNA-dependent RNA Polymerase (RDRP), a distinctive and wholly conserved RNA virus gene and a strong phylogenetic marker (57). A BLAST database of RDRP sequences was downloaded from the pfam database (58) under code pf00680. Contigs were aligned using Blastx with a minimum evalue of 10^{-4}. Hits were queried against the NCBI refseq protein database and only hits to viral RDRP genes were retained for downstream analyses.

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

Phylogenetic Analysis

Reference sequences for viral marker genes and Rpb1 were downloaded from the InterPro and RefSeq databases (STable 1) (62). Reference sequences were aligned using the MUSCLE alignment algorithm (63) in the MEGA v7.0.26 software package (64). Maximum likelihood phylogenetic trees were constructed in PhyML (65) with the LG substitution model and the aLRT SH-like likelihood method. Putative viral and Rpb1 contigs assembled from the metatranscriptomes were translated into proteins according to the reading frame of the top BLAST hit. Translated proteins were placed on the reference trees in a maximum likelihood framework in pplacer (66). Trees with abundance data were visualized using the iToL web interface (67).

Statistical Analysis

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

Expression values for contigs were imported into the PRIMER7 (68) statistical software package and $\log_2$ transformed. Expression values from each contig were correlated (Pearson's rho) to one another and statistically grouped by co-occurrence using group average hierarchical clustering. The SIMPROF test (69) was used to determine the statistical significance level of resulting clusters (alpha = 0.05, 1000 permutations). Statistically significant clusters with at least one viral contig, one *rpb*1 contig and less than 10 total members were visualized and annotated in Cytoscape 3.5.1 (70).

Accession Numbers

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

Acknowledgements

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

References

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

16. **Dudova L, Hajkova P, Buchtova H, Opravilova V.** 2013. Formation, succession and landscape history of Central-European summit raised bogs: A multiproxy study from the Hruby Jesenik Mountains. Holocene **23:**230-242.

17. **Ireland AW, Clifford MJ, Booth RK.** 2014. Widespread dust deposition on North American peatlands coincident with European land-clearance. Vegetation History and Archaeobotany **23:**693-700.

18. **Swindles GT, Turner TE, Roe HM, Hall VA, Rea HA.** 2015. Testing the cause of the Sphagnum austinii (Sull. ex Aust.) decline: Multiproxy evidence from a raised bog in Northern Ireland. Review of Palaeobotany and Palynology **213:**17-26.

19. **Galka M, Tobolski K, Gorska A, Lamentowicz M.** 2017. Resilience of plant and testate amoeba communities after climatic and anthropogenic disturbances in a Baltic bog in Northern Poland: Implications for ecological restoration. Holocene **27:**130-141.

20. **Opelt K, Chobot V, Hadacek F, Schonmann S, Eberl L, Berg G.** 2007. Investigations of the structure and function of bacterial communities associated with Sphagnum mosses. Environmental Microbiology **9:**2795-2809.

21. **Bragina A, Cardinale M, Berg C, Berg G.** 2013. Vertical transmission explains the specific Burkholderia pattern in Sphagnum mosses at multi-geographic scale. Frontiers in Microbiology **4:**10.

22. **Bragina A, Maier S, Berg C, Muller H, Chobot V, Hadacek F, Berg G.** 2012. Similar diversity of Alphaproteobacteria and nitrogenase gene amplicons on two related Sphagnum mosses. Frontiers in Microbiology **3:**10.

23. **Opelt K, Berg G.** 2004. Diversity and antagonistic potential of bacteria associated with bryophytes from nutrient-poor habitats of the Baltic Sea coast. Applied and Environmental Microbiology **70:**6569-6579.

24. **Brussaard CPD, Wilhelm SW, Thingstad F, Weinbauer MG, Bratbak G, Heldal M, Kimmance SA, Middelboe M, Nagasaki K, Paul JH, Schroeder DC, Suttle CA, Vaque D, Wommack KE.** 2008. Global-scale processes with a nanoscale drive: the role of marine viruses. ISME J **2:**575-578.

25. **Jover LF, Effler TC, Buchan A, Wilhelm SW, Weitz JS.** 2014. The elemental composition of virus particles: implications for marine biogeochemical cycles. Nat Rev Micro **12:**519-528.

26. **Wilhelm SW, Suttle CA.** 1999. Viruses and Nutrient Cycles in the SeaViruses play critical roles in the structure and function of aquatic food webs. BioScience **49:**781-788.

27. **Thingstad TF, Lignell R.** 1997. Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. Aquatic Microbial Ecology **13:**19-27.

28. **Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, Poulos BT, Solonenko N, Lara E, Poulain J, Pesant S, Kandels-Lewis S, Dimier C, Picheral M, Searson S, Cruaud C, Alberti A, Duarte CM, Gasol JM, Vaque D, Bork P, Acinas SG, Wincker P, Sullivan MB, Tara Oceans C.** 2016. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature **537:**689-+.

29. **Simmonds P, Adams MJ, Benko M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B, Hull R, King AMQ, Koonin EV, Krupovic M, Kuhn JH, Lefkowitz EJ, Nibert ML, Orton R, Roossinck MJ, Sabanadzovic S, Sullivan MB, Suttle CA, Tesh RB, van der Vlugt RA, Varsani A, Zerbini M.** 2017. Virus taxonomy in the age of metagenomics. Nature Reviews Microbiology **15:**161-168.

30. **Steward GF, Culley AI, Mueller JA, Wood-Charlson EM, Belcaid M, Poisson G.** 2013. Are we missing half of the viruses in the ocean? The ISME Journal **7:**672-679.

31. **Wilhelm SW, Bird JT, Bonifer KS, Calfee BC, Chen T, Coy SR, Gainer PJ, Gann ER, Heatherly HT, Lee J, Liang XL, Liu J, Armes AC, Moniruzzaman M, Rice JH, Stough JMA, Tams RN, Williams EP, LeCleir GR.** 2017. A Student's Guide to Giant Viruses Infecting Small Eukaryotes: From Acanthamoeba to Zooxanthellae. Viruses-Basel **9:**18.

32. **Yutin N, Wolf YI, Raoult D, Koonin EV.** 2009. Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. Virology Journal **6:**13.

33. **Moniruzzaman M, Wurch LL, Alexander H, Dyhrman ST, Gobler CJ, Wilhelm SW.** 2017. Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. Nature Communications **8:**10.

34. **Schulz F, Yutin N, Ivanova NN, Ortega DR, Lee TK, Vierheilig J, Daims H, Horn M, Wagner M, Jensen GJ, Kyrpides NC, Koonin EV, Woyke T.** 2017. Giant viruses with an expanded complement of translation system components. Science **356:**82-+.

35. **Moniruzzaman M, LeCleir GR, Brown CM, Gobler CJ, Bidle KD, Wilson WH, Wilhelm SW.** 2014. Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host–virus coevolution. Virology **466-467:**60-70.

36. **Moniruzzaman M, Gann ER, Wilhelm SW.** 2018. Infection by a Giant Virus (AaV) Induces Widespread Physiological Reprogramming in Aureococcus anophagefferens CCMP1984 – A Harmful Bloom Algae. Frontiers in Microbiology **9**.

37. **Wilhelm SW, Coy SR, Gann ER, Moniruzzaman M, Stough JMA.** 2016. Standing on the shoulders of giant viruses: five lessons learned about large viruses infecting small eukaryotes and the opportunities they create. Plos Pathogens **12:**5.

38. **Kerepesi C, Grolmusz V.** 2017. The "Giant Virus Finder" discovers an abundance of giant viruses in the Antarctic dry valleys. Archives of Virology **162:**1671-1676.

39. **Rusin LY.** 2016. Metagenomics and biodiversity of sphagnum bogs. Molecular Biology **50:**645-648.

40. **Krupovic M, Koonin EV.** 2014. Evolution of eukaryotic single-stranded DNA viruses of the Bidnaviridae family from genes of four other groups of widely different viruses. Scientific Reports **4:**5347.

41. **Krupovic M, Koonin EV.** 2015. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. Nature Reviews Microbiology **13:**105.

42. **Tomaru Y, Nagasaki K.** 2007. Flow cytometric detection and enumeration of DNA and RNA viruses infecting marine eukaryotic microalgae. Journal of Oceanography **63:**215-221.

43. **Miranda JA, Culley AI, Schvarcz CR, Steward GF.** 2016. RNA viruses as major contributors to Antarctic virioplankton. Environmental Microbiology **18:**3714-3727.

44. **Quaiser A, Krupovic M, Dufresne A, Francez A-J, Roux S.** 2016. Diversity and comparative genomics of chimeric viruses in Sphagnum-dominated peatlands. Virus Evolution **2:**vew025-vew025.

45. **Quaiser A, Dufresne A, Ballaud F, Roux S, Zivanovic Y, Colombet J, Sime-Ngando T, Francez A-J.** 2015. Diversity and comparative genomics of Microviridae in Sphagnum- dominated peatlands. Frontiers in Microbiology **6**.

46. **Ballaud F, Dufresne A, Francez A-J, Colombet J, Sime-Ngando T, Quaiser A.** 2015. Dynamics of Viral Abundance and Diversity in a Sphagnum-Dominated Peatland: Temporal Fluctuations Prevail Over Habitat. Frontiers in Microbiology **6:**1494.

47. **Dorigo U, Jacquet S, Humbert JF.** 2004. Cyanophage diversity, inferred from g20 gene analyses, in the largest natural lake in France, Lake Bourget. Applied and Environmental Microbiology **70:**1017-1022.

48. **Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-Ngando T, Debroas D.** 2012. Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. Plos One **7:**12.

49. **Comeau AM, Krisch HM.** 2008. The capsid of the T4 phage superfamily: The evolution, diversity, and structure of some of the most prevalent proteins in the biosphere. Molecular Biology and Evolution **25:**1321-1332.

50. **Sakowski EG, Munsell EV, Hyatt M, Kress W, Williamson SJ, Nasko DJ, Polson SW, Wommack KE.** 2014. Ribonucleotide reductases reveal novel viral diversity and predict biological and ecological features of unknown marine viruses. Proceedings of the National Academy of Sciences of the United States of America **111:**15786-15791.

51. **Alexander H, Jenkins BD, Rynearson TA, Dyhrman ST.** 2015. Metatranscriptome analyses indicate resource partitioning between diatoms in the field. Proceedings of the National Academy of Sciences **112:**E2182-E2190.

52. **Wilson RM, Hopple AM, Tfaily MM, Sebestyen SD, Schadt CW, Pfeifer-Meister L, Medvedeff C, McFarlane KJ, Kostka JE, Kolton M, Kolka RK, Kluber LA, Keller JK, Guilderson TP, Griffiths NA, Chanton JP, Bridgham SD, Hanson PJ.** 2016. Stability of peatland carbon to rising temperatures. Nature Communications **7:**10.

53. **Hanson PJ, Riggs JS, Nettles WR, Phillips JR, Krassovski MB, Hook LA, Gu L, Richardson AD, Aubrecht DM, Ricciuto DM.** 2017. Attaining whole-ecosystem warming using air and deep-soil heating methods with an elevated CO 2 atmosphere. Biogeosciences **14:**861.

54. **Warren MJ, Lin XJ, Gaby JC, Kretz CB, Kolton M, Morton PL, Pett-Ridge J, Weston DJ, Schadt CW, Kostka JE, Glass JB.** 2017. Molybdenum-Based Diazotrophy in a Sphagnum Peatland in Northern Minnesota. Applied and Environmental Microbiology **83:**14.

55. **Kopylova E, Noe L, Touzet H.** 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics **28:**3211-3217.

56. **O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, et al.** 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res **44:**D733-745.

57. **Koonin EV.** 1991. The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. Journal of General Virology **72:**2197-2206.

58. **Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A.** 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research **44:**D279-D285.

59. **Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR.** 2015. HMMER web server: 2015 update. Nucleic Acids Research **43:**W30-W38.

60. **Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu SN, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang ZX, Yamashita RA, Zhang DC, Zheng CJ, Bryant SH.** 2015. CDD: NCBI's conserved domain database. Nucleic Acids Research **43:**D222-D226.

61. **Liu WZ, Xie YB, Ma JY, Luo XT, Nie P, Zuo ZX, Lahrmann U, Zhao Q, Zheng YY, Zhao Y, Xue Y, Ren J.** 2015. IBS: an illustrator for the presentation and visualization of biological sequences. Bioinformatics **31:**3359-3361.

62. **Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztanyi Z, El-Gebali S, Fraser M, Gough J, Haft D, Holliday GL, Huang HZ, Huang XS, Letunic I, Lopez R, Lu SN, Marchler-Bauer A, Mi HY, Mistry J, Natale DA, Necci M, Nuka G, Orengo CA, Park Y, Pesseat S, Piovesan D, Potter SC, Rawlings ND, Redaschi N, Richardson L, Rivoire C, Sangrador-Vegas A, Sigrist C, Sillitoe I, Smithers B, Squizzato S, Sutton G, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Xenarios I, Yeh LS, Young SY, Mitchell AL.** 2017. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Research **45:**D190-D199.

63. **Edgar RC.** 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. Bmc Bioinformatics **5:**1-19.

64. **Kumar S, Stecher G, Tamura K.** 2016. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular Biology and Evolution **33:**1870-1874.

65. **Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O.** 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology **59:**307-321.

66. **Matsen FA, Kodner RB, Armbrust EV.** 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. Bmc Bioinformatics **11:**16.

67. **Letunic I, Bork P.** 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Research **44:**W242-W245.

68. **Clark KR, Gorley RN.** 2015. PRIMER v7: User manual/tutorial. PRIMER-E, Plymouth.

69. **Clarke KR, Somerfield PJ, Gorley RN.** 2008. Testing of null hypotheses in exploratory community analyses: similarity profiles and biota-environment linkage. Journal of Experimental Marine Biology and Ecology **366:**56-69.

70. **Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T.** 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research **13:**2498-2504.

Figure Legends

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

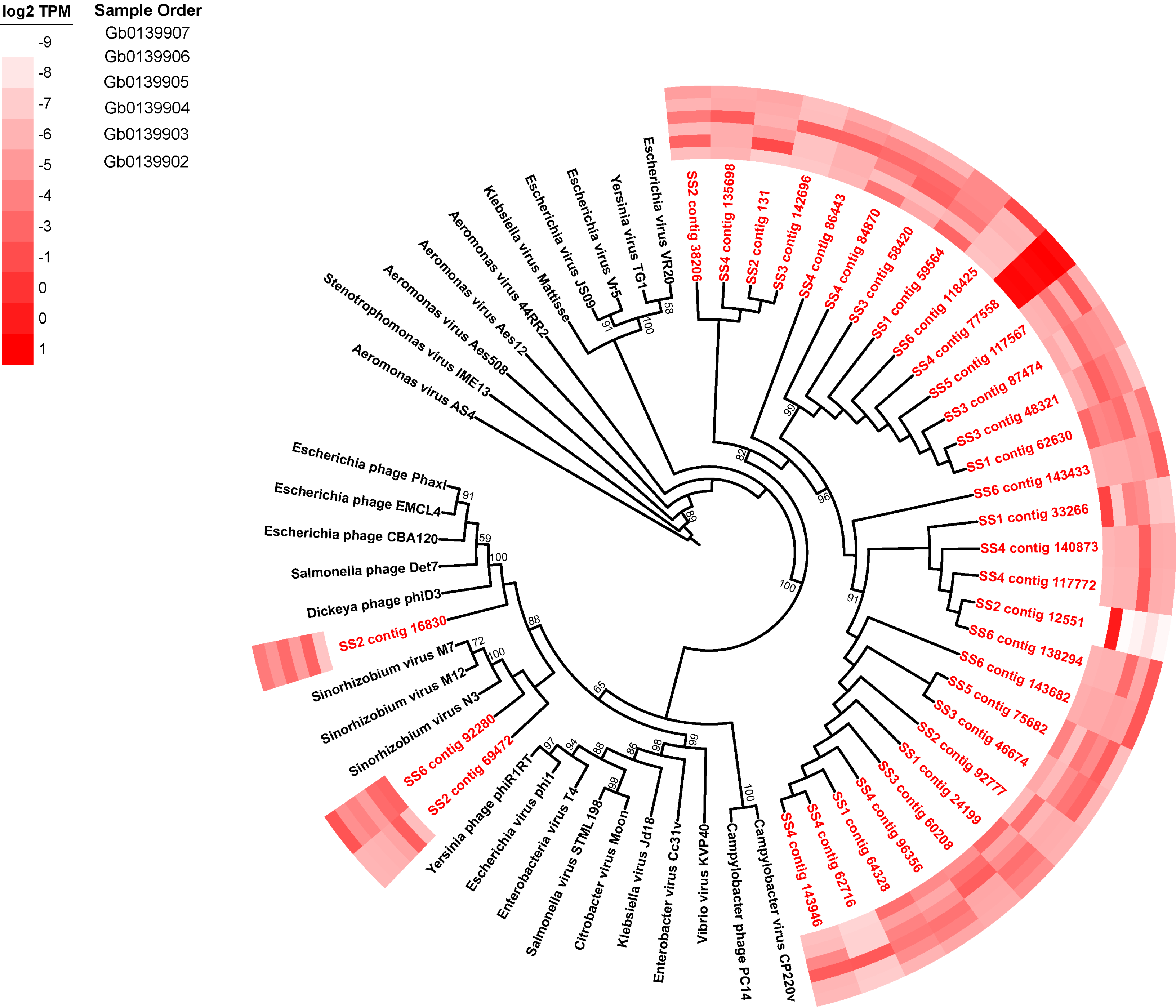
Figure 2: Phylogenetic placement of identified ssRNA virus RNA-dependent RNA polymerase contigs on maximum likelihood reference tree. Branch width represents the number of contigs placed on the reference branch. Node support (aLRT-SH statistic) >50% are shown.

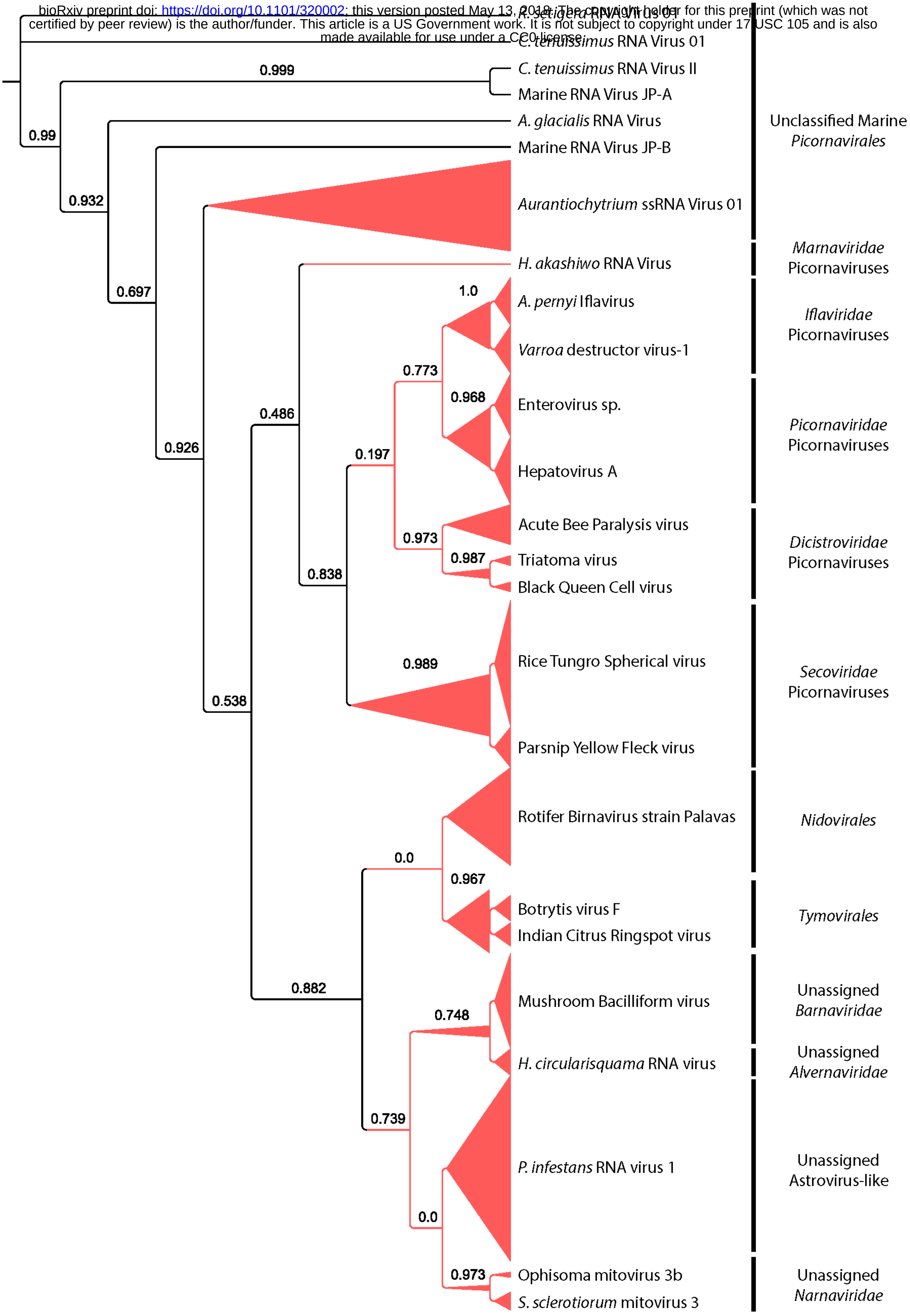
Figure 3: Phylogeny, genome architecture, and abundance of partial ssRNA virus genomes. Tree represents phylogenetic placement of RDRP gene regions from partial ssRNA virus genome contigs (red) on a maximum likelihood reference tree (references in black). Node support (aLRT-SH statistic) >50% are shown.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

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

Figure 5: Phylogenetic placement of identified virophage A.) major capsid protein and B.) ATPase contigs (red) on a maximum likelihood reference tree (references in black). Node support (aLRT-SH statistic) >50% are shown.

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


log2 TPM
-9
-8
-7
-6
-5
-4
-3
-1
0
0
1
Sample Order
Gb0139907
Gb0139906
Gb0139905
Gb0139904
Gb0139903
Gb0139902
Escherichia virus VR20
Yersinia virus TG1
Escherichia virus Vr5
Escherichia virus JS09
Klebsiella virus Mattisse
Aeromonas virus 44RR2
Aeromonas virus Aes12
Aeromonas virus Aes508
Stenotrophomonas virus IME13
Aeromonas virus AS4
Escherichia phage PhaxI
Escherichia phage EMCL4
Escherichia phage CBA120
Salmonella phage Det7
Dickeya phage phiD3
SS2 contig 16830
Sinorhizobium virus M7
Sinorhizobium virus M12
Sinorhizobium virus N3
SS6 contig 92280
SS2 contig 69472
Yersinia phage phiR1RT
Escherichia virus phi1
Enterobacteria virus T4
Salmonella virus STML198
Citrobacter virus Moon
Klebsiella virus Jd18
Enterobacter virus Cc31v
Vibrio virus KVP40
Campylobacter phage PC14
Campylobacter virus CP220v
SS4 contig 143946
SS4 contig 62716
SS1 contig 64328
SS4 contig 96356
SS3 contig 60208
SS1 contig 24199
SS2 contig 92777
SS3 contig 46674
SS5 contig 75682
SS6 contig 143682
SS6 contig 138294
SS2 contig 12551
SS4 contig 117772
SS4 contig 140873
SS1 contig 33266
SS6 contig 143433
SS1 contig 62630
SS3 contig 48321
SS3 contig 87474
SS5 contig 117567
SS4 contig 77558
SS6 contig 118425
SS1 contig 59564
SS3 contig 58420
SS4 contig 84870
SS4 contig 86443
SS3 contig 142696
SS2 contig 131
SS4 contig 135698
SS2 contig 38206
58
100
91
82
99
96
89
100
91
91
59
100
88
72
100
65
97
94
88
86
98
99
99
100


R. setigera RNA Virus 01
C. tenuissimus RNA Virus 01
0.999
C. tenuissimus RNA Virus II
Marine RNA Virus JP-A
0.99
A. glacialis RNA Virus
Marine RNA Virus JP-B
0.932
Unclassified Marine *Picornavirales*
Aurantiochytrium ssRNA Virus 01
0.697
Marnaviridae Picornaviruses
H. akashiwo RNA Virus
1.0
A. pernyi Iflavirus
Iflaviridae Picornaviruses
Varroa destructor virus-1
0.773
0.486
0.968
Enterovirus sp.
Picornaviridae Picornaviruses
0.926
0.197
Hepatovirus A
Acute Bee Paralysis virus
0.973
Dicistroviridae Picornaviruses
0.987
Triatoma virus
0.838
Black Queen Cell virus
0.989
Rice Tungro Spherical virus
Secoviridae Picornaviruses
0.538
Parsnip Yellow Fleck virus
Rotifer Birnavirus strain Palavas
Nidovirales
0.0
0.967
Botrytis virus F
Tymovirales
Indian Citrus Ringspot virus
0.882
Mushroom Bacilliform virus
Unassigned *Barnaviridae*
0.748
H. circularisquama RNA virus
Unassigned *Alvernaviridae*
0.739
P. infestans RNA virus 1
Unassigned Astrovirus-like
0.0
0.973
Ophisoma mitovirus 3b
Unassigned *Narnaviridae*
S. sclerotiorum mitovirus 3

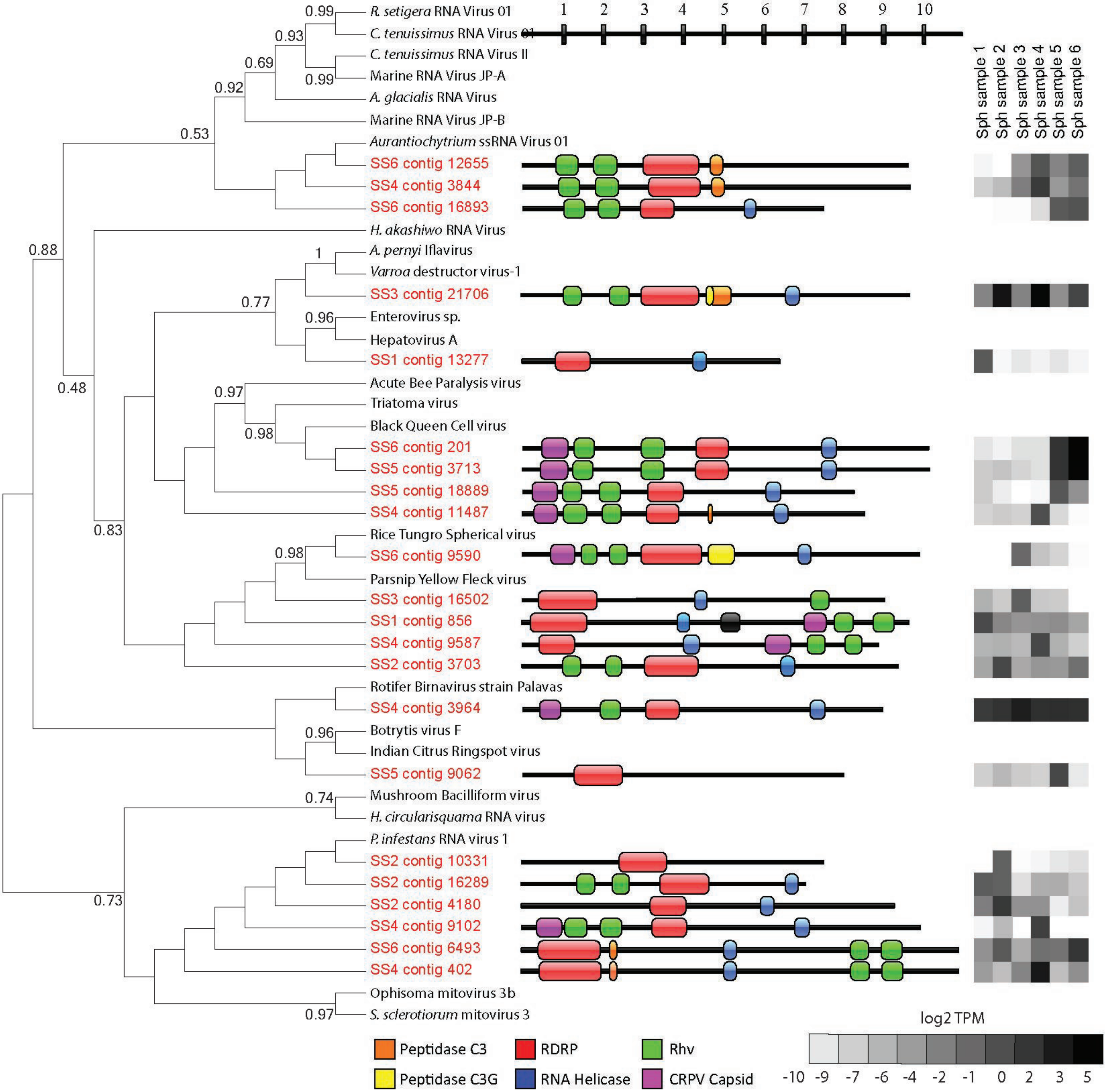
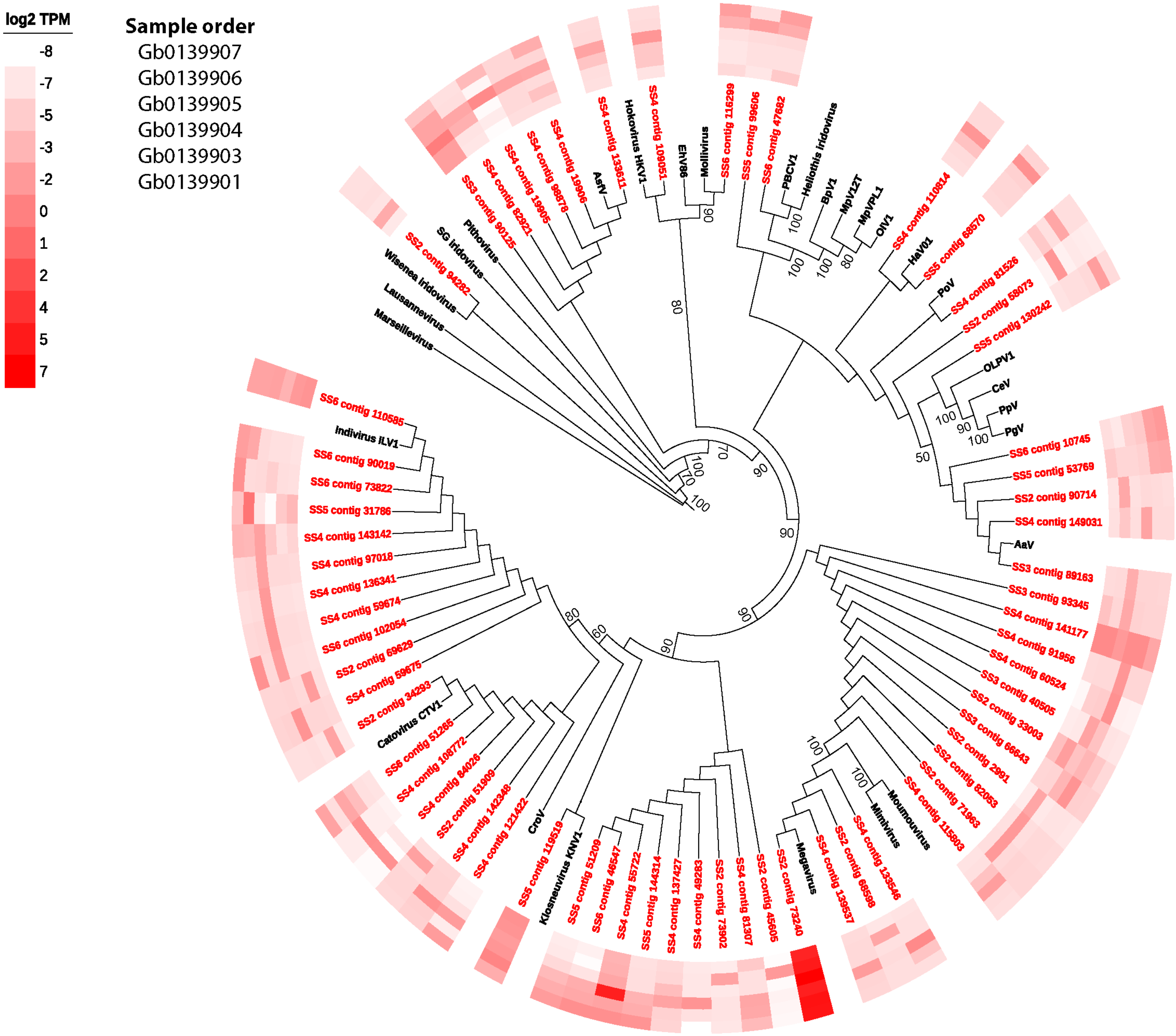
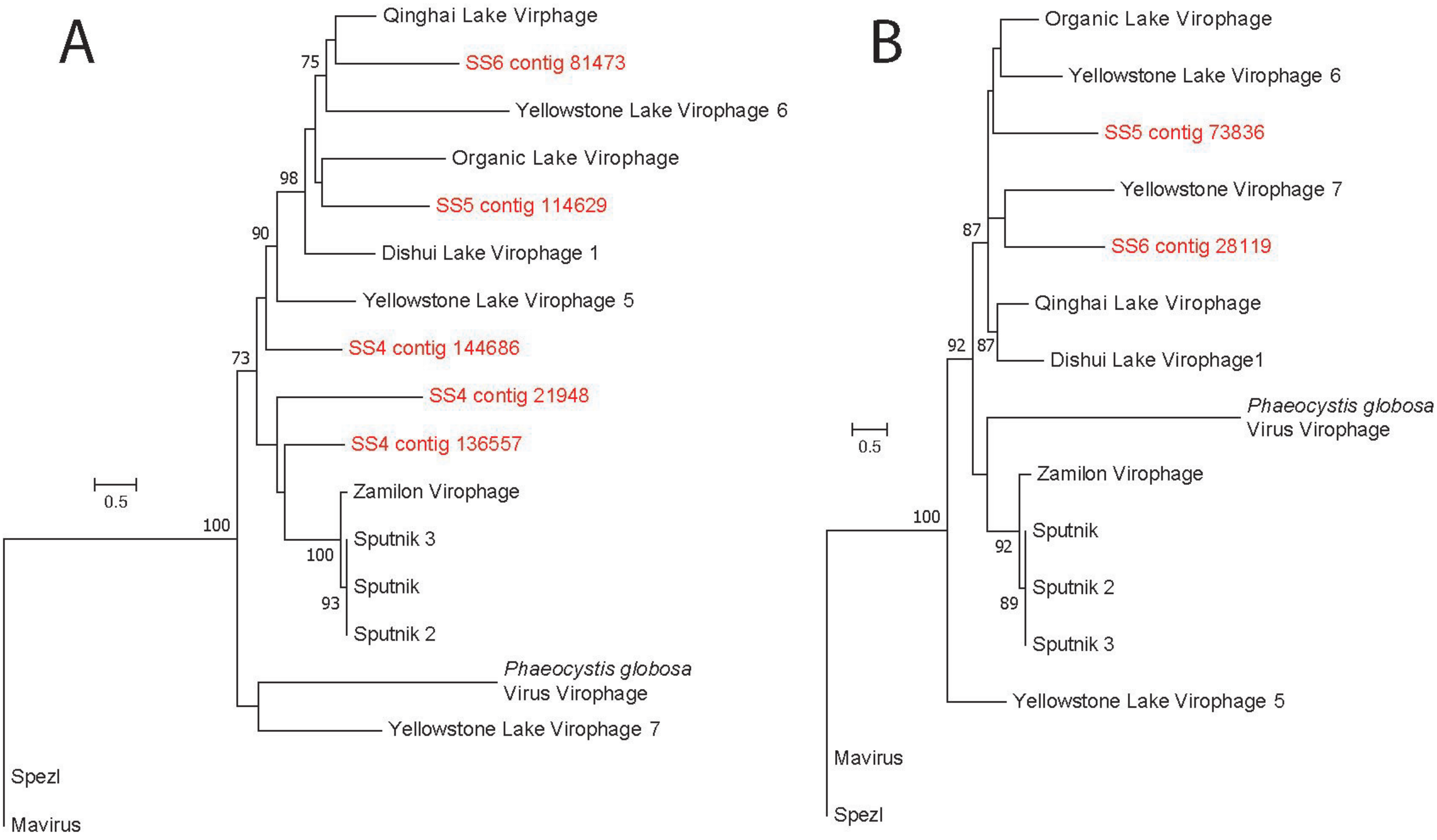
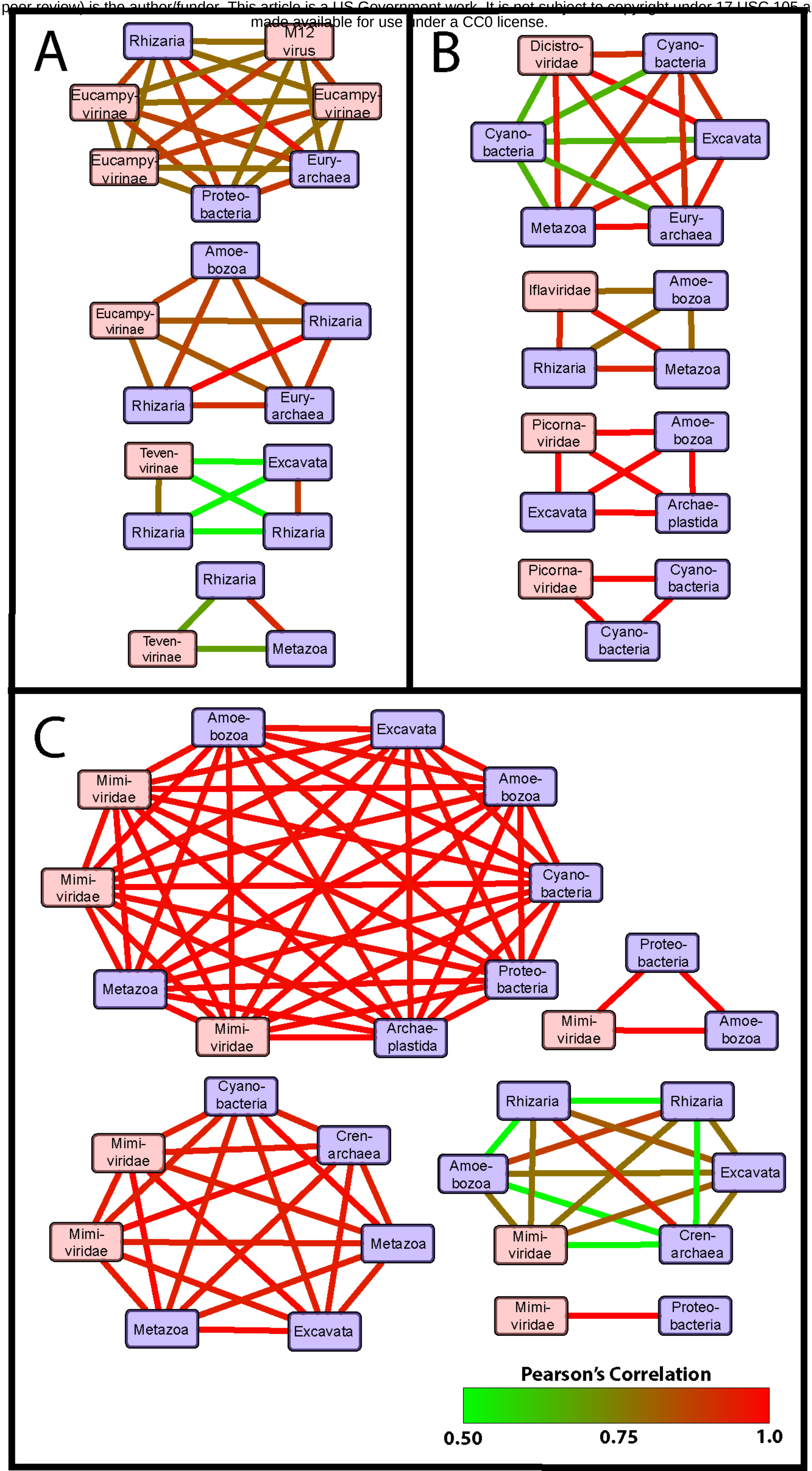


Figure 3: Phylogeny, genome architecture, and abundance of partial ssRNA virus genomes. Tree represents phylogenetic placement of RDRP gene regions from partial ssRNA virus genome contigs on a maximum likelihood reference tree. Node support (aLRT-SH statistic) >50% are shown. Center panel represents genome architecture determined by conserved domain search and ORF prediction. Length of contigs and gene regions is measured in kb. Heat map in right panel shows abundance of reads mapped to partial genome contigs in log2 TPM from each of the 6 metatranscriptome libraries.


log2 TPM
-8
-7
-5
-3
-2
0
1
2
4
5
7
Sample order
Gb0139907
Gb0139906
Gb0139905
Gb0139904
Gb0139903
Gb0139901
Marseillevirus
Lausannevirus
Wisenea iridovirus
SS2 contig 94282
SG iridovirus
Pithovirus
SS3 contig 90125
SS4 contig 82921
SS4 contig 19905
SS4 contig 98878
SS4 contig 19906
AsfV
SS4 contig 133611
Hokovirus HKV1
SS4 contig 109051
EhV86
Mollivirus
SS6 contig 116299
SS5 contig 99606
SS6 contig 47682
PBCV1
Heliothis iridovirus
BpV1
MpV12T
MpVPL1
OlV1
SS4 contig 110814
HaV01
SS5 contig 68570
PoV
SS4 contig 81526
SS2 contig 58073
SS5 contig 130242
OLPV1
CeV
PpV
PgV
SS6 contig 10745
SS5 contig 53769
SS2 contig 90714
SS4 contig 149031
AaV
SS3 contig 89163
SS3 contig 93345
SS4 contig 141177
SS4 contig 91956
SS4 contig 60524
SS3 contig 40505
SS2 contig 33003
SS3 contig 66643
SS2 contig 2991
SS2 contig 82053
SS2 contig 71963
SS4 contig 115803
Moumouvirus
Mimivirus
SS4 contig 133546
SS2 contig 68598
SS4 contig 139537
Megavirus
SS2 contig 73240
SS2 contig 45605
SS4 contig 81307
SS2 contig 73902
SS4 contig 49283
SS4 contig 137427
SS5 contig 144314
SS4 contig 55722
SS6 contig 46547
SS5 contig 51209
Klosneuvirus KNV1
SS5 contig 119519
CroV
SS4 contig 121422
SS4 contig 142348
SS2 contig 51909
SS4 contig 84026
SS4 contig 108772
SS6 contig 51265
Catovirus CTV1
SS2 contig 34293
SS4 contig 59675
SS2 contig 69629
SS6 contig 102054
SS4 contig 59674
SS4 contig 136341
SS4 contig 97018
SS4 contig 143142
SS5 contig 31786
SS6 contig 73822
SS6 contig 90019
Indivirus ILV1
SS6 contig 110585
80
90
100
100
100
80
100
100
90
100
50
90
70
100
70
100
90
90
80
60
90
100
100


A
Qinghai Lake Virphage
SS6 contig 81473
75
Yellowstone Lake Virophage 6
Organic Lake Virophage
98
SS5 contig 114629
90
Dishui Lake Virophage 1
Yellowstone Lake Virophage 5
SS4 contig 144686
73
SS4 contig 21948
SS4 contig 136557
Zamilon Virophage
0.5
100
Sputnik 3
100
Sputnik
93
Sputnik 2
Phaeocystis globosa
Virus Virophage
Yellowstone Lake Virophage 7
Spezl
Mavirus
B
Organic Lake Virophage
Yellowstone Lake Virophage 6
SS5 contig 73836
Yellowstone Virophage 7
87
SS6 contig 28119
Qinghai Lake Virophage
92
87
Dishui Lake Virophage1
Phaeocystis globosa
Virus Virophage
0.5
Zamilon Virophage
100
Sputnik
92
Sputnik 2
89
Sputnik 3
Yellowstone Lake Virophage 5
Mavirus
Spezl


A
Rhizaria
M12 virus
Eucampy-virinae
Eucampy-virinae
Eucampy-virinae
Eury-archaea
Proteo-bacteria
Amoe-bozoa
Eucampy-virinae
Rhizaria
Rhizaria
Eury-archaea
Teven-virinae
Excavata
Rhizaria
Rhizaria
Rhizaria
Teven-virinae
Metazoa
B
Dicistro-viridae
Cyano-bacteria
Cyano-bacteria
Excavata
Metazoa
Eury-archaea
Iflaviridae
Amoe-bozoa
Rhizaria
Metazoa
Picorna-viridae
Amoe-bozoa
Excavata
Archae-plastida
Picorna-viridae
Cyano-bacteria
Cyano-bacteria
C
Amoe-bozoa
Excavata
Mimi-viridae
Amoe-bozoa
Mimi-viridae
Cyano-bacteria
Metazoa
Proteo-bacteria
Mimi-viridae
Archae-plastida
Proteo-bacteria
Mimi-viridae
Amoe-bozoa
Cyano-bacteria
Mimi-viridae
Cren-archaea
Mimi-viridae
Metazoa
Metazoa
Excavata
Rhizaria
Rhizaria
Amoe-bozoa
Excavata
Mimi-viridae
Cren-archaea
Mimi-viridae
Proteo-bacteria
Pearson's Correlation
0.50
0.75
1.0