



**HAL**  
open science

## **Eukaryotic virus composition can predict the efficiency of carbon export in the global ocean**

Hiroto Kaneko, Romain Blanc-Mathieu, Hisashi Endo, Samuel Chaffron, Tom Delmont, Morgan Gaia, Nicolas Henry, Rodrigo Hernández-Velázquez, Canh Hao Nguyen, Hiroshi Mamitsuka, et al.

► **To cite this version:**

Hiroto Kaneko, Romain Blanc-Mathieu, Hisashi Endo, Samuel Chaffron, Tom Delmont, et al.. Eukaryotic virus composition can predict the efficiency of carbon export in the global ocean. *iScience*, 2021, 24 (1), pp.102002. 10.1016/j.isci.2020.102002 . hal-03097258v3

**HAL Id: hal-03097258**

**<https://hal.science/hal-03097258v3>**

Submitted on 10 Nov 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

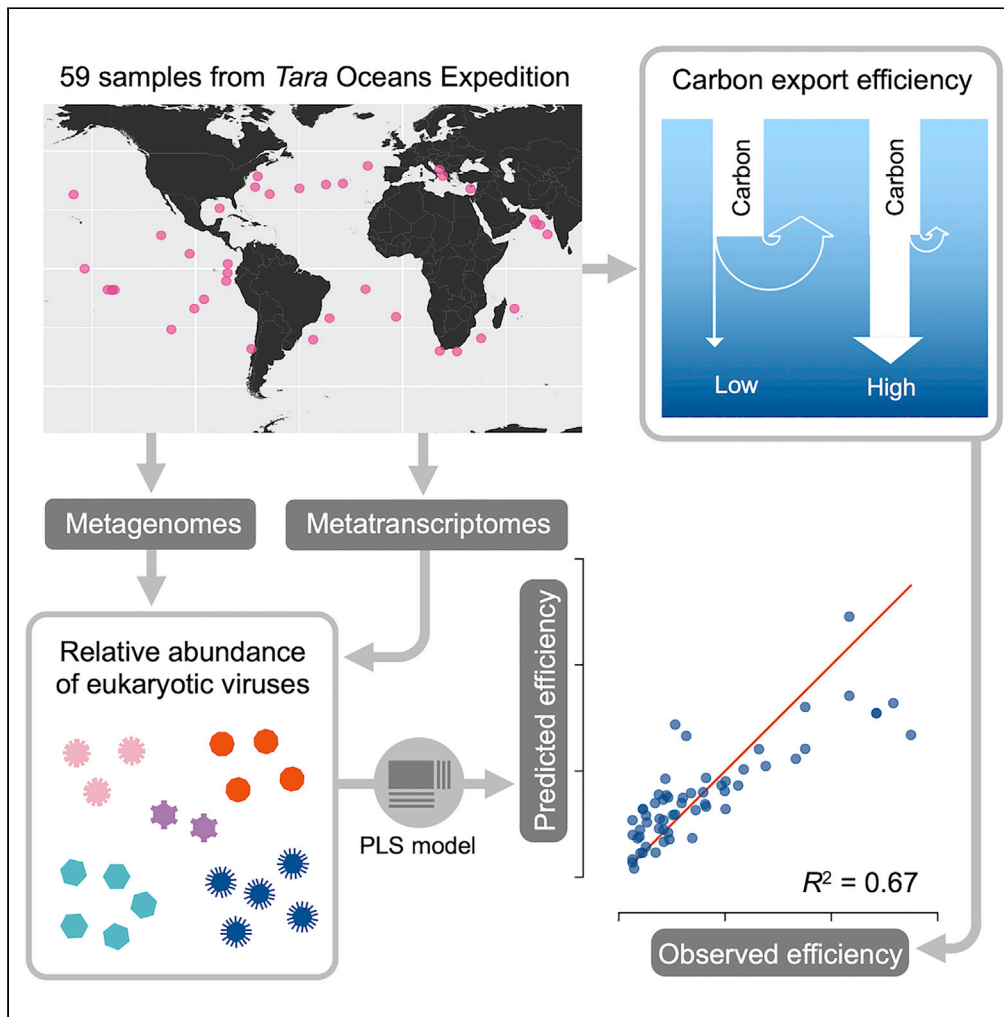
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Article

# Eukaryotic virus composition can predict the efficiency of carbon export in the global ocean



Hiroto Kaneko, Romain Blanc-Mathieu, Hisashi Endo, ..., Curtis A. Suttle, Lionel Guidi, Hiroyuki Ogata

ogata@kuicr.kyoto-u.ac.jp

**HIGHLIGHTS**

Eukaryotic virus community composition is shown to predict carbon export efficiency

Tens of viruses are highly important in the prediction of the efficiency

These viruses are inferred to infect ecologically important hosts



## Article

## Eukaryotic virus composition can predict the efficiency of carbon export in the global ocean

Hiroto Kaneko,<sup>1,11</sup> Romain Blanc-Mathieu,<sup>1,2,11</sup> Hisashi Endo,<sup>1</sup> Samuel Chaffron,<sup>3,4</sup> Tom O. Delmont,<sup>4,5</sup> Morgan Gaia,<sup>4,5</sup> Nicolas Henry,<sup>6</sup> Rodrigo Hernández-Velázquez,<sup>1</sup> Canh Hao Nguyen,<sup>1</sup> Hiroshi Mamitsuka,<sup>1</sup> Patrick Forterre,<sup>7</sup> Olivier Jaillon,<sup>4,5</sup> Colomban de Vargas,<sup>6</sup> Matthew B. Sullivan,<sup>8</sup> Curtis A. Suttle,<sup>9</sup> Lionel Guidi,<sup>10</sup> and Hiroyuki Ogata<sup>1,12,\*</sup>

## SUMMARY

**The biological carbon pump, in which carbon fixed by photosynthesis is exported to the deep ocean through sinking, is a major process in Earth's carbon cycle. The proportion of primary production that is exported is termed the carbon export efficiency (CEE). Based on in-lab or regional scale observations, viruses were previously suggested to affect the CEE (i.e., viral "shunt" and "shuttle"). In this study, we tested associations between viral community composition and CEE measured at a global scale. A regression model based on relative abundance of viral marker genes explained 67% of the variation in CEE. Viruses with high importance in the model were predicted to infect ecologically important hosts. These results are consistent with the view that the viral shunt and shuttle functions at a large scale and further imply that viruses likely act in this process in a way dependent on their hosts and ecosystem dynamics.**

## INTRODUCTION

A major process in the global cycling of carbon is the oceanic biological carbon pump (BCP), an organism-driven process by which atmospheric carbon (i.e., CO<sub>2</sub>) is transferred and sequestered to the ocean interior and seafloor for periods ranging from centuries to hundreds of millions of years. Between 15% and 20% of net primary production (NPP) is exported out of the euphotic zone, with 0.3% of fixed carbon reaching the seafloor annually (Zhang et al., 2018). However, there is wide variation in estimates of the proportion of primary production in the surface ocean that is exported to depth, ranging from 1% in the tropical Pacific to 35%–45% during the North Atlantic bloom (Buesseler and Boyd, 2009). As outlined below, many factors affect the BCP.

Of planktonic organisms living in the upper layer of the ocean, diatoms (Tréguer et al., 2018) and zooplankton (Turner, 2015) have been identified as important contributors to the BCP in nutrient-replete oceanic regions. In the oligotrophic ocean, cyanobacteria, collodarians (Lomas and Moran, 2011), diatoms (Agusti et al., 2015; Karl et al., 2012; Leblanc et al., 2018), and other small (pico- to nano-) plankton (Lomas and Moran, 2011) have been implicated in the BCP. Sediment trap studies suggest that ballasted aggregates of plankton with biogenic minerals contribute to carbon export to the deep sea (Iversen and Ploug, 2010; Klaas and Archer, 2002). The BCP comprises three processes: carbon fixation, export, and remineralization. As these processes are governed by complex interactions between numerous members of planktonic communities (Zhang et al., 2018), the BCP is expected to involve various organisms, including viruses (Zimmerman et al., 2019).

Viruses have been suggested to regulate the efficiency of the BCP. Lysis of host cells by viruses releases cellular material in the form of dissolved organic matter (DOM), which fuels the microbial loop and enhances respiration and secondary production (Gobler et al., 1997; Weitz et al., 2015). This process, coined "viral shunt (Wilhelm and Suttle, 1999)," can reduce the carbon export efficiency (CEE) because it increases the retention of nutrients and carbon in the euphotic zone and prevents their transfer to higher trophic levels as well as their export from the euphotic zone to the deep sea (Fuhrman, 1999; Weitz et al., 2015).

<sup>1</sup>Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

<sup>2</sup>Laboratoire de Physiologie Cellulaire & Végétale, CEA, Univ. Grenoble Alpes, CNRS, INRA, IRIG, Grenoble, France

<sup>3</sup>Université de Nantes, CNRS UMR 6004, LS2N, 44000 Nantes, France

<sup>4</sup>Research Federation (FR2022) Tara Oceans GO-SEE, Paris, France

<sup>5</sup>Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, 91000 Evry, France

<sup>6</sup>Sorbonne Universités, CNRS, Laboratoire Adaptation et Diversité en Milieu Marin, Station Biologique de Roscoff, 29680 Roscoff, France

<sup>7</sup>Institut Pasteur, Department of Microbiology, 25 rue du Docteur Roux, 75015, Paris, France

<sup>8</sup>Department of Microbiology and Department of Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, OH, United States of America

<sup>9</sup>Departments of Earth, Ocean & Atmospheric Sciences, Microbiology & Immunology, and Botany, and the Institute for the Oceans and Fisheries, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

<sup>10</sup>Sorbonne Université, CNRS, Laboratoire d'Océanographie de Villefranche, LOV, 06230 Villefranche-sur-mer, France

<sup>11</sup>These authors contributed equally

<sup>12</sup>Lead contact

\*Correspondence: ogata@kuicr.kyoto-u.ac.jp  
<https://doi.org/10.1016/j.isci.2020.102002>



However, an alternative process is also considered, in which viruses contribute to the vertical carbon export (Weinbauer, 2004). For instance, a theoretical study proposed that the CEE increases if viral lysis augments the ratio of exported carbon relative to the primary production-limiting nutrients (nitrogen and phosphorous) (Suttle, 2007). Laboratory experimental studies reported that cells infected with viruses form larger particles (Peduzzi and Weinbauer, 1993; Yamada et al., 2018), can sink faster (Lawrence and Suttle, 2004), and can lead to preferential grazing by heterotrophic protists (Evans and Wilson, 2008) and/or to higher growth of grazers (Goode et al., 2019). This process termed “viral shuttle” (Sullivan et al., 2017) is supported by several field studies that reported association of viruses with sinking material. Viruses were observed in sinking material in the North Atlantic Ocean (Proctor and Fuhrman, 1991) and sediment of coastal waters where algal blooms occur (Lawrence et al., 2002; Tomaru et al., 2007, 2011). In addition, vertical transport of bacterial viruses between photic and aphotic zones was observed in the Pacific Ocean (Hurwitz et al., 2015) and in Tara Oceans virome data (Brum et al., 2015). A systematic analysis of large-scale omics data from oligotrophic oceanic regions revealed a positive association between the magnitude of carbon flux and bacterial dsDNA viruses (i.e., cyanophages), which were previously unrecognized as possible contributors to the BCP (Guidi et al., 2016).

More recently, viral infection of blooms of the photosynthetic eukaryote *Emiliana huxleyi* in the North Atlantic were found to be accompanied by particle aggregation and greater downward vertical flux of carbon, with the highest export during the early stage of viral infection (Laber et al., 2018; Sheyn et al., 2018). Given the significant contributions of eukaryotic plankton to ocean biomass and net production (Hirata et al., 2011; Li, 1995) and their observed predominance over prokaryotes in sinking materials of Sargasso Sea oligotrophic surface waters (Fawcett et al., 2011; Lomas and Moran, 2011), various lineages of eukaryotic viruses may be responsible for a substantial part of the variation in carbon export across oceanic regions.

If the “viral shunt” and “shuttle” processes function at a global scale and if these involve specific eukaryotic viruses, we expect to detect a statistical association between eukaryotic viral community composition and CEE in a large-scale omics data. To our knowledge, such an association has never been investigated. Although this test per se does not prove that viruses regulate CEE, we consider the association is worth being tested because such an association is a necessary condition for the global model of viral shunt and shuttle and, under its absence, we would have to reconsider the model. Deep sequencing of planktonic community DNA and RNA, as carried out in Tara Oceans, has enabled the identification of marker genes of major viral groups infecting eukaryotes (Hingamp et al., 2013; Carradec et al., 2018; Culley, 2018; Endo et al., 2020). To examine the association between viral community composition and CEE, we thus used the comprehensive organismal dataset from the Tara Oceans expedition (Carradec et al., 2018; Sunagawa et al., 2015), as well as related measurements of carbon export estimated from particle concentrations and size distributions observed in situ (Guidi et al., 2016).

In the present study, we identified several hundred marker-gene sequences of nucleocytoplasmic large DNA viruses (NCLDVs) in metagenomes of 0.2–3  $\mu\text{m}$  size fraction. We also identified RNA and ssDNA viruses in metatranscriptomes of four eukaryotic size fractions spanning 0.8 to 2,000  $\mu\text{m}$ . The resulting profiles of viral distributions were compared with an image-based measure of carbon export efficiency (CEE), which is defined as the ratio of the carbon flux at depth to the carbon flux at surface.

## RESULTS AND DISCUSSION

### Detection of diverse eukaryotic viruses in Tara Oceans gene catalogs

We used profile hidden Markov model-based homology searches to identify marker-gene sequences of eukaryotic viruses in two ocean gene catalogs. These catalogs were previously constructed from environmental shotgun sequence data of samples collected during the Tara Oceans expedition. The first catalog, the Ocean Microbial Reference Gene Catalog (OM-RGC), contains 40 million non-redundant genes predicted from the assemblies of Tara Oceans viral and microbial metagenomes (Sunagawa et al., 2015). We searched this catalog for NCLDV DNA polymerase family B (PolB) genes, as dsDNA viruses may be present in microbial metagenomes because large virions (>0.2  $\mu\text{m}$ ) have been retained on the filter or because viral genomes actively replicating or latent within picoeukaryotic cells have been captured. The second gene catalog, the Marine Atlas of Tara Oceans Unigenes (MATOU), contains 116 million non-redundant genes derived from metatranscriptomes of single-cell microeukaryotes and small multicellular zooplankton (Carradec et al., 2018). We searched this catalog for NCLDV PolB genes, RNA-dependent RNA polymerase

**Table 1. Taxonomic breakdown of viral marker genes**

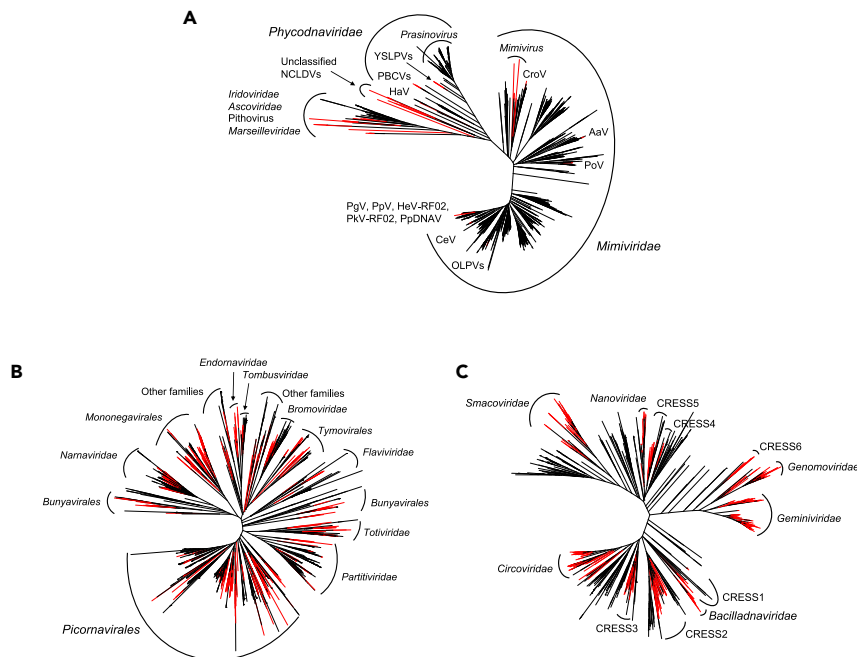
Viruses	Identified	Used in PLS regression <sup>a</sup>
NCLDVs	Mimiviridae	2,923
	Phycodnaviridae	348
	Iridoviridae	198
	Other NCLDVs <sup>b</sup>	17
	Total	3,486
RNA viruses	Picornavirales (ssRNA+)	325
	Partitiviridae (dsRNA)	131
	Narnaviridae (ssRNA+)	95
	Other families	289
	Unclassified	78
	RNA viruses	57
	Total	975
ssDNA viruses	Circoviridae	201
	Geminiviridae	4
	Nanoviridae	4
	Unclassified	39
	ssDNA viruses	51
	Total	299
All	4,760	1,523

<sup>a</sup>The marker genes had to occur in at least five samples and harbor a Spearman correlation coefficient > |0.2| with carbon export efficiency.

<sup>b</sup>There was no unclassified NCLDV.

(RdRP) genes of RNA viruses, and replication-associated protein (Rep) genes of ssDNA viruses, because transcripts of viruses actively infecting their hosts, as well as genomes of RNA viruses, have been captured in this catalog.

We identified 3,874 NCLDV PoIB sequences (3,486 in metagenomes and 388 in metatranscriptomes), 975 RNA virus RdRP sequences, and 299 ssDNA virus Rep sequences (Table 1). These sequences correspond to operational taxonomic units (OTUs) at a 95% identity threshold. All except 17 of the NCLDV PoIBs from metagenomes were assigned to the families *Mimiviridae* ( $n = 2,923$ ), *Phycodnaviridae* ( $n = 348$ ), and *Iridoviridae* ( $n = 198$ ) (Table 1). The larger numbers of PoIB sequences assigned to *Mimiviridae* and *Phycodnaviridae* compared with other NCLDV families are consistent with previous observations (Endo et al., 2020; Hingamp et al., 2013). The divergence between these environmental sequences and reference sequences from known viral genomes was greater in *Mimiviridae* than in *Phycodnaviridae* (Figures 1A, S1A, and S2). Within *Mimiviridae*, 83% of the sequences were most similar to those from algae-infecting *Mimivirus* relatives. Among the sequences classified in *Phycodnaviridae*, 93% were most similar to those in *Prasinovirus*, whereas 6% were closest to *Yellowstone lake phycodnavirus*, which is closely related to *Prasinovirus*. *Prasinovirus* are possibly overrepresented in the metagenomes because the 0.2 to 3  $\mu\text{m}$  size fraction selects their picoeukaryotic hosts. RdRP sequences were assigned mostly to the order *Picornavirales* ( $n = 325$ ), followed by the families *Partitiviridae* ( $n = 131$ ), *Narnaviridae* ( $n = 95$ ), *Tombusviridae* ( $n = 45$ ), and *Virgaviridae* ( $n = 33$ ) (Table 1), with most sequences being distant (30%–40% amino acid identity) from reference viruses (Figures 1B, S1B, and S3). These results are consistent with previous studies on the diversity of marine RNA viruses, in which RNA virus sequences were found to correspond to diverse positive-polarity ssRNA and dsRNA viruses distantly related to well-characterized viruses (Culley, 2018). *Picornavirales* may be overrepresented in the metatranscriptomes because of the polyadenylated RNA selection. The majority ( $n = 201$ ) of Rep sequences were annotated as *Circoviridae*, known to infect animals, which is consistent with a previous report (Wang et al., 2018). Only eight were annotated as plant ssDNA viruses (families *Nanoviridae* and *Geminiviridae*) (Table 1). Most of these environmental sequences are distant (40% to 50% amino acid



**Figure 1. Viruses of eukaryotic plankton identified in Tara Oceans samples are distantly related to characterized viruses**

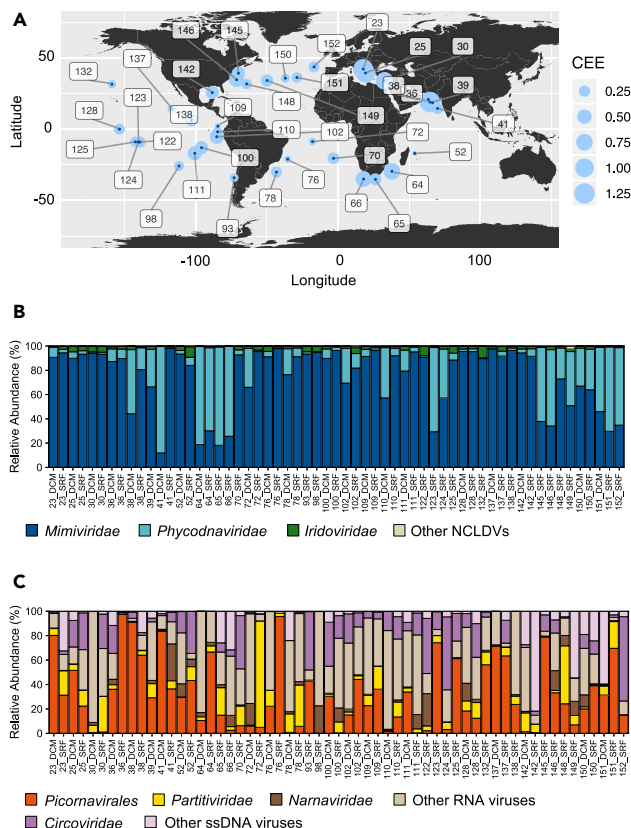
Unrooted maximum likelihood phylogenetic trees containing environmental (black) and reference (red) viral sequences for NCLDV DNA polymerase family B (A), RNA virus RNA-dependent RNA polymerase (B), and ssDNA virus replication-associated protein (C). See also [Figures S1–S4](#)

identity) from reference sequences ([Figures 1C, S1C, and S4](#)). Additional 388 NCLDV PolBs were detected in the metatranscriptomes. The average cosmopolitanism (number of samples where an OTU was observed by at least two reads) for PolBs in metagenomes was 23 samples against 2.9 for metatranscriptome-derived PolB sequences, 5.5 for Reps, and 5.8 for RdRPs. Within metatranscriptomes, the average gene-length normalized read counts for PolBs were respectively ten and three times lower than those of RdRPs and Reps. Therefore, PolBs from metatranscriptomes were not further used in our study.

### Composition of eukaryotic viruses can explain the variation of carbon export efficiency

Among the PolB, RdRP, and Rep sequences identified in the Tara Oceans gene catalogs, 38%, 18%, and 11% (total = 1,523 sequences), respectively, were present in at least five samples and had matching carbon export measurement data ([Table 1](#)). We used the relative abundance (defined as the centered log-ratio transformed gene-length normalized read count) profiles of these 1,523 marker-gene sequences at 59 sampling sites in the photic zone of 39 Tara Oceans stations ([Figure 2](#)) to test for association between their composition and a measure of carbon export efficiency (CEE, see [Transparent Methods, Figure S5](#)). A partial least squares (PLS) regression model explained 67% (coefficient of determination  $R^2 = 67\%$ ) of the variation in CEE with a Pearson correlation coefficient of 0.84 between observed and predicted values. This correlation was confirmed to be statistically significant by permutation test ( $p < 1 \times 10^{-4}$ ) ([Figure 3A](#)).

We also tested for their association with estimates of carbon export flux at 150 meters ( $CE_{150}$ ) and NPP. PLS regressions explained 54% and 64% of the variation in  $CE_{150}$  and NPP with Pearson correlation coefficients between observed and predicted values of 0.74 (permutation test,  $p < 1 \times 10^{-4}$ ) and 0.80 (permutation test,  $p < 1 \times 10^{-4}$ ), respectively ([Figure S6](#)). In these three PLS regression models, 83, 86, and 97 viruses were considered to be key predictors (i.e., Variable Importance in the Projection [VIP] score  $> 2$ ) of CEE,  $CE_{150}$ , and NPP, respectively. PLS models for NPP and  $CE_{150}$  shared a larger number of predictors (52 viruses) compared with the PLS models for NPP and CEE (seven viruses) (two-proportion Z-test,  $p = 4.14 \times 10^{-12}$ ). Consistent with this observation,  $CE_{150}$  was correlated with NPP (Pearson's  $r = 0.77$ ; parametric test,  $p < 1 \times 10^{-12}$ ). This result implies that the magnitude of export in the analyzed samples was partly constrained by primary productivity. However, CEE was not correlated with NPP ( $r = 0.16$ ; parametric



**Figure 2. Carbon export efficiency and relative marker-gene occurrence of eukaryotic plankton viruses along the sampling route**

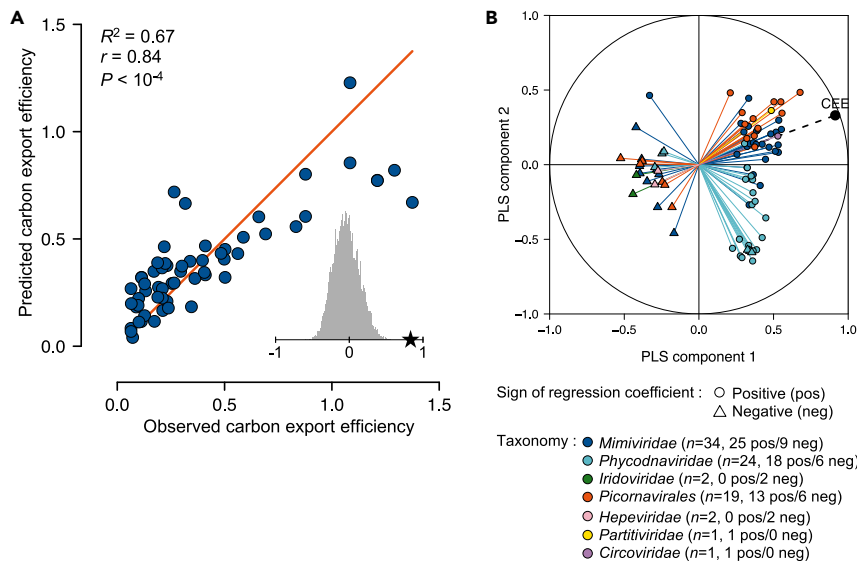
(A) Carbon export efficiency (CEE) estimated at 39 Tara Oceans stations where surface and DCM layers were sampled for prokaryote-enriched metagenomes and eukaryotic metatranscriptomes. See also Figures S5 and S11.

(B and C) Relative marker-gene occurrence of major groups of viruses of eukaryotic plankton for NCLDVs in metagenomes (B) and for RNA and ssDNA viruses in metatranscriptomes (C) at 59 sampling sites.

test,  $p = 0.2$ ) or  $CE_{150}$  ( $r = 0.002$ ; parametric test,  $p = 0.99$ ). Thus, as expected, primary productivity was not a major driver for the efficiency of carbon export.

To assess the sensitivity of the model to the definition of carbon export efficiency, we employed an alternative measure of carbon export efficiency that considers euphotic zone depth ( $T_{100}$ , see Transparent Methods).  $T_{100}$  was correlated with CEE ( $r = 0.66$ ; parametric test,  $p < 1 \times 10^{-8}$ ), and PLS regression explained 44% of the variation in  $T_{100}$  (permutation test,  $p < 1 \times 10^{-4}$ ) (Figure S7). Of 72 predictors of the PLS model for  $T_{100}$ , 30 were shared with that for CEE. This result demonstrates the robustness of the PLS model to definitions of carbon export efficiency.

The 83 viruses (5% of the viruses included in our analysis) that were associated with CEE with a VIP score  $> 2$  are considered to be important predictors of CEE in the PLS regression (Figure 3B, Data S1), and these viruses are hereafter referred to as VIPs (Viruses Important in the Prediction). Fifty-eight VIPs had positive regression coefficient, and 25 had negative regression coefficient in the prediction (Figure 3B). Most of the positively associated VIPs showed high relative abundance in the Mediterranean Sea and in the Indian Ocean where CEE tends to be high compared with other oceanic regions (Figure 4). Among them, 15 (red labels in Figure 4) also had high relative abundance in samples from other oceanic regions, showing that these viruses are associated with CEE at a global scale. In contrast, negatively associated VIPs tend to have higher relative abundance in the Atlantic Ocean and the Southern Pacific Ocean where CEE is comparatively lower. In the following sections, we investigate potential hosts of the VIPs in order to interpret the statistical association between viral community composition and CEE in the light of previous observations in the literature.



**Figure 3. Relative abundance of eukaryotic plankton viruses is associated with carbon export efficiency in the global ocean**

(A) Bivariate plot between predicted and observed values in a leave-one-out cross-validation test for carbon export efficiency. The PLS regression model was constructed using occurrence profiles of 1,523 marker-gene sequences (1,309 PolBs, 180 RdRPs, and 34 Reps) derived from environmental samples.  $r$ , Pearson correlation coefficient;  $R^2$ , the coefficient of determination between measured response values and predicted response values.  $R^2$ , which was calculated as  $1 - \text{SSE} / \text{SST}$  (sum of squares due to error and total) measures how successful the fit is in explaining the variance of the response values. The significance of the association was assessed using a permutation test ( $n = 10,000$ ) (gray histogram in (A)). The red diagonal line shows the theoretical curve for perfect prediction.

(B) Pearson correlation coefficients between CEE and occurrence profiles of 83 viruses that have VIP scores  $>2$  (VIPs) with the first two components in the PLS regression model using all samples. PLS components 1 and 2 explained 83% and 11% of the variance of CEE, respectively. Fifty-eight VIPs had positive regression coefficients in the model (shown with circles), and 25 had negative regression coefficients (shown with triangles). See also [Figures S6, S7, and S12](#), [Table S1](#), and [Data S1](#).

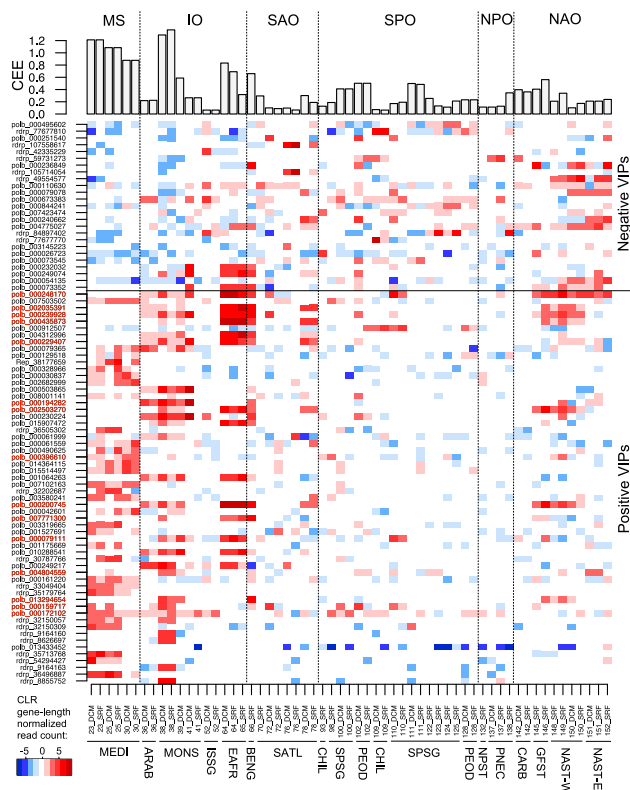
### Viruses correlated with CEE infect ecologically important hosts

Most of the VIPs (77 of 83) belong to *Mimiviridae* ( $n = 34$  with 25 positive VIPs and 9 negative VIPs), *Phycodnaviridae* ( $n = 24$  with 18 positive VIPs and 6 negative VIPs), and ssRNA viruses of the order *Picornavirales* ( $n = 19$  with 13 positive VIPs and 6 negative VIPs) ([Figure 3B](#), [Table S1](#)). All the phycodnavirus VIPs were most closely related to prasinoviruses infecting Mamiellales, with amino acid sequence percent identities to reference sequences ranging between 35% and 95%. The six remaining VIPs were two NCLDVs of the family *Iridoviridae* negatively associated with CEE, three RNA viruses (two ssRNA viruses of the family *Hepeviridae* negatively associated with CEE and one dsRNA virus of the family *Partitiviridae* positively associated with CEE), and one ssDNA virus of the family *Circoviridae* positively associated with CEE. A proportionally larger number of PolBs were included in the model than RdRP and Rep sequences depending on their representations in the input data. Therefore, the larger number of NCLDV VIPs obtained does not necessarily mean that this group of viruses is more important than others regarding their association with CEE.

Host information may help understand the relationship between these VIPs and CEE. We performed genomic context analysis for PolB VIPs and phylogeny-guided network-based host prediction for PolB and RdRP to infer putative relationship between virus and host (see [Transparent Methods](#)).

Taxonomic analysis of genes predicted in 10 metagenome-assembled genomes (MAGs) from the eukaryotic size fractions and 65 genome fragments (contigs) assembled from the prokaryotic size fraction encoding VIP PolBs further confirmed their identity as *Mimiviridae* or *Phycodnaviridae* ([Figure S8](#)). The size of MAGs ranged between 30 kbp and 440 kbp with an average of 210 kbp ([Table S2](#)). The presence of genes with high-sequence similarities to cellular genes in a viral genome is suggestive of a relationship between virus and host ([Monier et al., 2009](#); [Yoshikawa et al., 2019](#)). Two closely related *Mimiviridae* VIPs, PolB 000079111 (positively associated with CEE) and PolB 000079078 (negatively associated with CEE),





**Figure 4. Biogeography of viruses associated with carbon export efficiency**

The upper panel shows carbon export efficiency (CEE =  $CE_{\text{deep}}/CE_{\text{surface}}$ ) for 59 sampling sites. The bottom panel is a map reflecting relative abundances, expressed as centered log-ratio transformed, gene-length normalized read counts of viruses positively and negatively associated with CEE that have VIP scores  $>2$  (VIPs). MS, Mediterranean Sea; IO, Indian Ocean; SAO, South Atlantic Ocean; SPO, South Pacific Ocean; NPO, North Pacific Ocean; NAO, North Atlantic Ocean. The bottom horizontal axis is labeled with Tara Oceans station numbers, sampling depth (SRF, surface; DCM, deep chlorophyll maximum), and abbreviations of biogeographic provinces. Viruses labeled in red correspond to positive VIPs that are highly represented in one or more biogeographic provinces outside MS and IO.

were phylogenetically close to the pelagophyte virus *Aureococcus anophagefferens virus* (AaV). One MAG (268 kbp in size) corresponding to PolB 000079111 encoded seven genes showing high similarities to genes from Pelagophyceae, and another MAG (382 kbp in size), corresponding to PolB 000079078, encoded five genes similar to genes from Pelagophyceae. All but one of these 12 genes were encoded on a genome fragment containing genes annotated as viral, including five NCLDV core genes (Data S2), excluding the possibility of contamination in these MAGs. Two closely related *Phycodnaviridae* VIPs, PolB 001064263 and 010288541, were positively associated with CEE. Both of these PolBs correspond to an MAG (134 kbp in size) encoding one gene likely derived from Mamiellales. The genomic fragment harboring this cellular gene was found to encode 10 genes annotated as viral (Data S2).

We conducted a phylogeny-guided, network-based host prediction analysis for *Mimiviridae*, *Phycodnaviridae*, and *Picornavirales* (Figures S9 and S10). Only a subset of the VIPs was included in this analysis because we kept the most reliable sequences ( $n = 44$ ) to obtain a well-resolved tree topology. Within the *Prasinovirus* clade, which contained thirteen VIPs (nine positive and four negative), seven different eukaryotic orders were detected as predicted host groups for ten nodes in the tree. Mamiellales, the only known host group of prasinoviruses, was detected at eight nodes (five of them had no parent-to-child relationships), whereas the other six eukaryotic orders were found at only one node (or two in the case of Eutreptiales) (Figure S9). The order Mamiellales includes three genera (*Micromonas*, *Ostreococcus*, and *Bathycoccus*), which are bacterial-sized green microalgae common in coastal and oceanic environments and are considered to be influential actors in oceanic systems (Monier et al., 2016). Various prasinoviruses (fourteen with available genome sequences) have been isolated from the three genera.

Within the family *Mimiviridae*, which contains fifteen VIPs (ten positive and five negative), twelve different orders were predicted as putative host groups (Figure S9). Collodaria was detected at fifteen nodes (two of them had no parent-to-child relationships), and Prymnesiales at six nodes (three of them had no parent-to-child relationships), whereas all other orders were present at a maximum of one node each with no parent-to-child relationships. The nodes enriched for Prymnesiales and Collodaria fell within a monophyletic clade (marked by a red arrow in Figure S9) containing four reference haptophyte viruses infecting Prymnesiales and two reference haptophyte viruses infecting Phaeocystales. Therefore, the environmental PolB sequences in this *Mimiviridae* clade (including five positive VIPs and one negative VIP) are predicted to infect Prymnesiales or related haptophytes. The detection of Collodaria may be the result of indirect associations that reflect a symbiotic relationship with Prymnesiales, as some acantharians, evolutionarily related to the Collodaria, are known to host Prymnesiales species (Mars Brisbin et al., 2018). Known species of Prymnesiales and Phaeocystales have organic scales, except one Prymnesiales species, *Prymnesium neolepis*, which bears siliceous scales (Yoshida et al., 2006). Previous studies revealed the existence of diverse and abundant noncalcifying picohaptophytes in open oceans (Endo et al., 2018; Liu et al., 2009). Clear host prediction was not made for the other nine *Mimiviridae* VIPs shown in the phylogenetic tree. Three VIPs (two positive and one negative) in the tree were relatives of AaV. One negatively associated VIP was a relative of *Cafeteria roenbergensis virus* infecting a heterotrophic protist. The five remaining *Mimiviridae* VIPs are very distant from any known *Mimiviridae*.

Sixteen *Picornavirales* VIPs (eleven positive and five negative) were included in the phylogeny-guided, network-based host prediction analysis (Figure S10). Nine (seven positive and two negative) were grouped within *Dicistroviridae* (known to infect insects) and may therefore infect marine arthropods such as copepods, the most ubiquitous and abundant mesozooplankton groups involved in carbon export (Turner, 2015). Three other *Picornavirales* VIPs were placed within a clade containing known bacillarnaviruses. Two of them (35179764 and 33049404) were positively associated with CEE and had diatoms of the order Chaetocerotales as a predicted host group. The third one (107558617) was negatively associated with CEE and distant from other bacillarnaviruses and had no host prediction. Diatoms have been globally observed in the deep sea (Agusti et al., 2015; Leblanc et al., 2018) and identified as important contributors of the biological carbon pump (Tréguer et al., 2018). One positively associated VIP (32150309) was in a clade containing *Aurantiochytrium single-stranded RNA virus* (AsRNAV), infecting a marine fungoid protist thought to be an important decomposer (Takao et al., 2005). The last three *Picornavirales* VIPs (59731273, 49554577, and 36496887) had no predicted host and were too distant from known *Picornavirales* to speculate about their putative host group.

Outside *Picornavirales*, three RNA virus VIPs (two *Hepeviridae*, negatively associated, and one *Partitiviridae*, positively associated) were identified, for which no reliable host inferences were made by sequence similarity. Known *Hepeviridae* infect metazoans, and known *Partitiviridae* infect fungi and plants. The two *Hepeviridae*-like viruses were most closely related to viruses identified in the transcriptomes of mollusks (amino acid identities of 48% for 42335229 and 43% for 77677770) (Shi et al., 2016). The *Partitiviridae*-like VIP (35713768) was most closely related to a fungal virus, *Penicillium stoloniferum virus S* (49% amino acid identity).

One ssDNA virus VIP (38177659) was positively associated with CEE. It was annotated as a *Circoviridae*, although it groups with other environmental sequences as an outgroup of known *Circoviridae*. This VIP was connected with copepod, mollusk, and Collodaria OTUs in the co-occurrence network but no enrichment of predicted host groups was detected for its clade. *Circoviridae*-like viruses are known to infect copepods (Dunlap et al., 2013) and have been reported to associate with mollusks (Dayaram et al., 2015), but none have been reported for Collodaria.

Overall, we could infer hosts for 37 VIPs (Tables 2 and S3). Most of the predicted hosts are known to be ecologically important as primary producers (Mamiellales, Prymnesiales, Pelagophyceae, and diatoms) or grazers (copepods). Of these, diatoms and copepods are well known as important contributors to the BCP but others (i.e., Mamiellales, Prymnesiales, Pelagophyceae) have not been recognized as major contributors to the BCP. Our analysis also revealed that positive and negative VIPs are not separated in either the viral or host phylogenies.

**Table 2. Host predictions per viral and host group for viruses associated with carbon export efficiency**

Virus-Host Relationship	Positive VIPs <sup>a</sup>	Negative VIPs <sup>a</sup>	Total
NCLDV-mamiellales	10	4	15
NCLDV-prymnesiales	5	1	6
NCLDV-pelagophyceae	2	1	3
NCLDV-no prediction	26	11	36
RNA virus-copepoda	7	2	9
RNA virus-chaetocerotales	2	0	2
RNA virus-labyrinthulomycetes	1	0	1
RNA virus-no prediction	4	6	10
ssDNA virus-copepoda	1	0	1
Total	58	25	83

See also [Figures S8–S10](#), [Tables S2](#) and [S3](#), and [Data S2](#).

<sup>a</sup>VIPs refers to viruses having VIP scores > 2. Positive and negative VIPs had positive and negative regression coefficients in the PLS model, respectively.

### Viruses positively correlated with CEE tend to interact with silicified organisms

The phylogeny-guided, network-based host prediction analysis correctly predicted known relationships between virus and host (for viruses infecting Mamiellales, Prymnesiales, and Chaetocerotales) using our large dataset, despite the reported limitations of these co-occurrence network-based approaches (Coenen and Weitz, 2018). This result prompted us to further exploit the species co-occurrence networks (Table S4) to investigate functional differences between the eukaryotic organisms predicted to interact with positive VIPs, negative VIPs, and viruses less important for prediction of CEE (VIP score <2) (non-VIPs). For this purpose, we used literature-based functional trait annotations associated with eukaryotic meta-barcodes (see [Transparent Methods](#)). Positive VIPs had a greater proportion of connections with silicified eukaryotes ( $Q = 0.001$ ) but not with chloroplast-bearing eukaryotes ( $Q = 0.16$ ) nor calcifying eukaryotes ( $Q = 1$ ), compared to non-VIPs (Table 3). No functional differences were observed between negative VIPs and non-VIPs viruses (Table S5) or positive VIPs (Table S6).

### Multifarious ways viruses affect the fate of carbon

Our analysis revealed that eukaryotic virus composition was able to predict CEE in the global sunlit ocean, and 83 out of the 1,523 viruses had a high importance in the predictive model. This association is not a proof that the viruses are the cause of the variation of CEE. Viruses, especially those showing latent/persistent infections (Goic and Saleh, 2012), may be found to be associated with CEE if their host affects CEE regardless of viral infection. Organisms that preferentially grow in marine snow (Bochdansky et al., 2017) may also bring associations between viruses infecting those organisms and CEE. Alternatively, the observed associations between VIPs and CEE may reflect a more direct causal relationship, which we attempt to explore in light of the large body of literature on the mechanisms by which viruses impact the fate of carbon in the oceans.

Among the 83 VIPs, 58 were positively associated with CEE. Such a positive association is expected from the “viral shuttle” model, which states that viral activity could facilitate carbon export to the deep ocean (Fuhrman, 1999; Sullivan et al., 2017; Weinbauer, 2004), because a virus may induce secretion of sticky material that contributes to cell/particle aggregation, such as transparent exopolymeric particles (TEP) (Nissimov et al., 2018). We found that CEE (i.e.,  $CE_{\text{deep}}/CE_{\text{surface}}$ ) increased with the change of particles size from surface to deep ( $\rho = 0.42$ ,  $p = 8 \times 10^{-9}$ ) (Figure S11). This positive correlation may reflect an elevated level of aggregation in places where CEE is high, although it could be also due to the presence of large organisms at depth.

Greater aggregate sinking along with higher particulate carbon fluxes was observed in North Atlantic blooms of *Emiliania huxleyi* that were infected early by the virus EhV, compared with late-infected blooms (Laber et al., 2018). In the same bloom, viral infection stage was found to proceed with water column depth

**Table 3. Functional differences between eukaryotes found to be best connected to positively associated and not associated with carbon export efficiency**

Functional trait	Positive VIPs <sup>a</sup> (n = 50)		Non-VIPs <sup>a</sup> (n = 983)		p value (Fisher's exact test, two sided)	Adjusted p value (BH) (Q)
	Presence	Absence	Presence	Absence		
Chloroplast	20	30	276	690	0.109	0.164
Silicification	11	39	60	920	0.000	0.001
Calcification	1	49	30	950	1.000	1.000

See also Tables S4–S6.

<sup>a</sup>VIPs refer to viruses having VIP scores > 2. Positive VIPs had positive regression coefficients in the PLS model.

(Sheyn et al., 2018). No EhV-like PolB sequences were detected in our dataset, which was probably due to sampled areas and seasons.

Laboratory experiments suggest that viruses closely related to positive VIPs, such as prasinoviruses, have infectious properties that may drive carbon export. Cultures of *Micromonas pusilla* infected with prasinoviruses showed increased TEP production compared with non-infected cultures (Lønborg et al., 2013). The hosts of prasinoviruses (Mamiellales) have been proposed to contribute to carbon export in the western subtropical North Pacific (Shiozaki et al., 2019). Some prasinoviruses encode glycosyltransferases (GTs) of the GT2 family. The expression of GT2 family members during infection possibly leads to the production of a dense fibrous hyaluronan network and may trigger the aggregation of host cells (Van Etten et al., 2017) with an increase in the cell wall C:N ratio. We detected one GT2 in an MAG of two *Phycodnaviridae*-like positive VIPs (000200745 and 002503270) predicted to infect Mamiellales, one in an MAG corresponding to the putative pelagophyte positive VIP 000079111 related to AaV and six in two MAGs (three each) corresponding to two *Mimiviridae*-like positive VIPs (000328966 and 001175669). *Phaeocystis globosa virus* (PgV), closely related to the positive VIP PolB 000912507 (Figure S9), has been linked with increased TEP production and aggregate formation during the termination of a *Phaeocystis* bloom (Brussaard et al., 2007). Two closely related bacillarnavirus VIPs were positively associated with CEE and predicted to infect Chaetocerales. A previous study revealed an increase in abundance of viruses infecting diatoms of *Chaetoceros* in both the water columns and the sediments during the bloom of their hosts in a coastal area (Tomaru et al., 2011), suggesting sinking of cells caused by viruses. Furthermore, the diatom *Chaetoceros tenuissimus* infected with a DNA virus (CtenDNAV type II) has been shown to produce higher levels of large-sized particles (50–400 μm) compared with non-infected cultures (Tomaru et al., 2011; Yamada et al., 2018).

The other 25 VIPs were negatively associated with CEE. This association is compatible with the “viral shunt,” which increases the amount of DOC (Wilhelm and Suttle, 1999) and reduces the transfer of carbon to higher trophic levels and to the deep ocean (Fuhrman, 1999; Weitz et al., 2015). Increased DOC has been observed in culture of Mamiellales lysed by prasinoviruses (Lønborg et al., 2013). A field study reported that PgV, to which the negative VIP PolB 000054135 is closely related (Figure S9), can be responsible for up to 35% of cell lysis per day during bloom of its host (Baudoux et al., 2006), which is likely accompanied by consequent DOC release. Similarly, the decline of a bloom of the pelagophyte *Aureococcus anophagefferens* has been associated with active infection by AaV (to which one negative VIP is closely related) (Moniruzzaman et al., 2017). Among RNA viruses, eight were negative VIPs (six *Picornavirales* and two *Hepeviridae*). The higher representation of *Picornavirales* in the virioplankton (Culley, 2018) than within cells (Urayama et al., 2018) suggests that they are predominantly lytic, although no information exists regarding the effect of *Picornavirales* on DOC release.

It is likely that the “viral shunt” and “shuttle” simultaneously affect and modulate CEE in the global ocean (Zimmerman et al., 2019). The relative importance of these two phenomena must fluctuate considerably depending on the host traits, viral effects on metabolism, stages of infection, and environmental conditions. Reflecting this complexity, viruses of a same host group could be found to be either positively or negatively associated with CEE. We found that even two very closely related *Mimiviridae* viruses (PolBs 000079111 and 000079078 sharing 94% nucleotide identity over their full gene lengths) most likely infecting pelagophyte algae were positively and negatively associated with CEE.

Five percent of the tested viruses were associated with CEE in our study. Similarly, 4% and 2% of bacterial virus populations were found to be associated with the magnitude of carbon export (Guidi et al., 2016) and CEE (Figure S12), respectively. These results suggest that viruses affecting CEE are rather uncommon. It is plausible that such viruses affect CEE by infecting organisms that are functionally important (abundant or keystone species), as we observed in host prediction. The vast majority (95%) of non-VIPs may not have a significant impact on CEE, because they do not strongly impact the host population, for instance, by stably coexisting with their hosts. It is worth noting that experimental studies have reported cultures of algae with viruses that reach a stable co-existence state after a few generations (Yau et al., 2020).

### Conclusions

Eukaryotic virus community composition was able to predict CEE at 59 sampling sites in the photic zone of the world ocean. This statistical association was detected based on a large omics dataset collected throughout the oceans and processed with standardized protocols. The predictability of CEE by viral composition is consistent with the hypothesis that “viral shunt” and “shuttle” are functioning at a global scale. Among 83 viruses with a high importance in the prediction of CEE, 58 viruses were positively and 25 negatively correlated with carbon export efficiency. Most of these viruses belong to *Prasinovirus*, *Mimiviridae*, and *Picornavirales* and are either new to science or with no known roles in carbon export efficiency. Thirty-six of these “select” viruses were predicted to infect ecologically important hosts such as green algae of the order Mamiellales, haptophytes, diatoms, and copepods. Positively associated viruses had more predicted interactions with silicified eukaryotes than non-associated viruses did. Overall, these results imply that the effect of viruses on the “shunt” and “shuttle” processes could be dependent on viral hosts and ecosystem dynamics.

### Limitations of the study

The observed statistical associations between viral compositions and examined parameters (i.e., CEE, CE and NPP) do not convey the information about the direction of their potential causality relationships, and they could even result from indirect relationships as discussed earlier. Certain groups of viruses detected in samples may be over- or underrepresented because of the technical limitations in size fractionation, DNA/RNA extraction, and sequencing.

### Resource availability

#### Lead contact

Further information and requests for resources should be directed to and will be fulfilled by Lead Contact, Hiroyuki Ogata ([ogata@kuicr.kyoto-u.ac.jp](mailto:ogata@kuicr.kyoto-u.ac.jp)).

#### Materials availability

This study did not generate unique reagent.

#### Data and code availability

The authors declare that the data supporting the findings of this study are available within the paper and its supplemental files, as well as at the GenomeNet FTP: [ftp://ftp.genome.jp/pub/db/community/tara/Cpump/Supplementary\\_material/](ftp://ftp.genome.jp/pub/db/community/tara/Cpump/Supplementary_material/).

Our custom R script used to test for association between viruses and environmental variables (CEE, CE<sub>150</sub>, NPP and T<sub>100</sub>) is available along with input data at the GenomeNet FTP: [ftp://ftp.genome.jp/pub/db/community/tara/Cpump/Supplementary\\_material/PLSreg/](ftp://ftp.genome.jp/pub/db/community/tara/Cpump/Supplementary_material/PLSreg/). The Taxon Interaction Mapper (TIM) tool developed for this study and used for virus host prediction is available at <https://github.com/RomainBlancMathieu/TIM>.

## METHODS

All methods can be found in the accompanying [Transparent Methods supplemental file](#).

## SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at <https://doi.org/10.1016/j.isci.2020.102002>.

## ACKNOWLEDGMENTS

We thank the *Tara* Oceans consortium, the projects Oceanomics and France Genomique (grants ANR-11-BTBR-0008 and ANR-10-INBS-09), and the people and sponsors who supported the *Tara* Oceans Expedition (<http://www.embl.de/tara-oceans/>) for making the data accessible. This is contribution number 110 of the *Tara* Oceans Expedition 2009–2013. Computational time was provided by the Super-Computer System, Institute for Chemical Research, Kyoto University. We thank Barbara Goodson, Ph.D. and Sara J. Mason, M.Sc. from Edanz Group (<https://en-author-services.edanzgroup.com/>) for editing a draft of this manuscript. This work was supported by JSPS/KAKENHI (Nos. 26430184, 18H02279, and 19H05667 to H.O. and Nos. 19K15895 and 19H04263 to H.E.); Scientific Research on Innovative Areas from the Ministry of Education, Culture, Science, Sports and Technology (MEXT) of Japan (Nos. 16H06429, 16K21723, and 16H06437 to H.O.); the Collaborative Research Program of the Institute for Chemical Research, Kyoto University (2019-29 to S.C.); the Future Development Funding Program of the Kyoto University Research Coordination Alliance (to R.B.M.); the ICR-KU International Short-term Exchange Program for Young Researchers (to S.C.); and the Research Unit for Development of Global Sustainability (to H.O. and T.O.D.).

## AUTHOR CONTRIBUTIONS

H.O. and R.B.M. conceived the study. H.K. and R.B.M. performed most of the analyses. H.E. and L.G. designed carbon export analysis. R.H.V. and S.C. performed network analysis. N.H. and C.d.V. analyzed eukaryotic sequences. T.O.D., M.G., P.F., and O.J. analyzed viral MAGs. C.H.N. and H.M. contributed to statistical analysis. M.B.S. and C.A.S. contributed to interpretations. All authors edited and approved the final version of the manuscript.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 23, 2020

Revised: November 13, 2020

Accepted: December 23, 2020

Published: January 22, 2021

## REFERENCES

- Agusti, S., González-Gordillo, J.I., Vaqué, D., Estrada, M., Cerezo, M.I., Salazar, G., Gasol, J.M., and Duarte, C.M. (2015). Ubiquitous healthy diatoms in the deep sea confirm deep carbon injection by the biological pump. *Nat. Commun.* 6, 7608.
- Baudoux, A., Noordeloos, A., Veldhuis, M., and Brussaard, C. (2006). Virally induced mortality of *Phaeocystis globosa* during two spring blooms in temperate coastal waters. *Aquat. Microb. Ecol.* 44, 207–217.
- Bochdansky, A.B., Clouse, M.A., and Herndl, G.J. (2017). Eukaryotic microbes, principally fungi and labyrinthulomycetes, dominate biomass on bathypelagic marine snow. *ISME J.* 11, 362–373.
- Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doucier, G., Acinas, S.G., Alberti, A., Chaffron, S., Cruaud, C., Vargas, C.de, Gasol, J.M., et al. (2015). Patterns and ecological drivers of ocean viral communities. *Science* 348, 1261498.
- Brussaard, C.P.D., Bratbak, G., Baudoux, A.-C., and Ruardij, P. (2007). *Phaeocystis* and its interaction with viruses. *Biogeochemistry* 83, 201–215.
- Buesseler, K.O., and Boyd, P.W. (2009). Shedding light on processes that control particle export and flux attenuation in the twilight zone of the open ocean. *Limnol. Oceanogr.* 54, 1210–1232.
- Carradec, Q., Pelletier, E., Silva, C.D., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., et al. (2018). A global ocean atlas of eukaryotic genes. *Nat. Commun.* 9, 373.
- Coenen, A.R., and Weitz, J.S. (2018). Limitations of correlation-based inference in complex virus-microbe communities. *MSystems* 3, e00084–18.
- Culley, A. (2018). New insight into the RNA aquatic virosphere via viromics. *Virus Res.* 244, 84–89.
- Dayaram, A., Goldstien, S., Argüello-Astorga, G.R., Zawar-Reza, P., Gomez, C., Harding, J.S., and Varsani, A. (2015). Diverse small circular DNA viruses circulating amongst estuarine molluscs. *Infect. Genet. Evol.* 31, 284–295.
- Dunlap, D.S., Ng, T.F.F., Rosario, K., Barbosa, J.G., Greco, A.M., Breitbart, M., and Hewson, I. (2013). Molecular and microscopic evidence of viruses in marine copepods. *Proc. Natl. Acad. Sci. U S A* 110, 1375–1380.
- Endo, H., Blanc-Mathieu, R., Li, Y., Salazar, G., Henry, N., Labadie, K., de Vargas, C., Sullivan, M.B., Bowler, C., Wincker, P., et al. (2020). Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nat. Ecol. Evol.* 4, 1639–1649.
- Endo, H., Ogata, H., and Suzuki, K. (2018). Contrasting biogeography and diversity patterns between diatoms and haptophytes in the central Pacific Ocean. *Sci. Rep.* 8, 10916.
- Evans, C., and Wilson, W.H. (2008). Preferential grazing of *Oxyrrhis marina* on virus infected *Emiliania huxleyi*. *Limnol. Oceanogr.* 53, 2035–2040.
- Fawcett, S.E., Lomas, M.W., Casey, J.R., Ward, B.B., and Sigman, D.M. (2011). Assimilation of upwelled nitrate by small eukaryotes in the Sargasso Sea. *Nat. Geosci.* 4, 717–722.
- Fuhrman, J.A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature* 399, 541–548.
- Gobler, C.J., Hutchins, D.A., Fisher, N.S., Cospere, E.M., and Sañudo-Wilhelmy, S.A. (1997). Release and bioavailability of C, N, P, Se, and Fe following viral lysis of a marine chrysophyte. *Limnol. Oceanogr.* 42, 1492–1504.
- Goic, B., and Saleh, M.-C. (2012). Living with the enemy: viral persistent infections from a friendly viewpoint. *Curr. Opin. Microbiol.* 15, 531–537.

- Goode, A.G., Fields, D.M., Archer, S.D., and Martinez, J.M. (2019). Physiological responses of *Oxyrrhis marina* to a diet of virally infected *Emiliania huxleyi*. *PeerJ* 7, e6722.
- Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlmi, A., Roux, S., Darzi, Y., Audic, S., Berline, L., Brum, J.R., et al. (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature* 532, 465.
- Hingamp, P., Grimsley, N., Acinas, S.G., Clerissi, C., Subirana, L., Poulain, J., Ferrera, I., Sarmiento, H., Villar, E., Lima-Mendez, G., et al. (2013). Exploring nucleocytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J.* 7, 1678–1695.
- Hirata, T., Hardman-Mountford, N.J., Brewin, R.J.W., Aiken, J., Barlow, R., Suzuki, K., Isada, T., Howell, E., Hashioka, T., Noguchi-Aita, M., et al. (2011). Synoptic relationships between surface Chlorophyll-*a* and diagnostic pigments specific to phytoplankton functional types. *Biogeosciences* 8, 311–327.
- Hurwitz, B.L., Brum, J.R., and Sullivan, M.B. (2015). Depth-stratified functional and taxonomic niche specialization in the “core” and “flexible” Pacific Ocean Virome. *ISME J.* 9, 472–484.
- Iversen, M.H., and Ploug, H. (2010). Ballast minerals and the sinking carbon flux in the ocean: carbon-specific respiration rates and sinking velocity of marine snow aggregates. *Biogeosciences* 7, 2613–2624.
- Karl, D.M., Church, M.J., Dore, J.E., Letelier, R.M., and Mahaffey, C. (2012). Predictable and efficient carbon sequestration in the North Pacific Ocean supported by symbiotic nitrogen fixation. *Proc. Natl. Acad. Sci. U S A* 109, 1842–1849.
- Klaas, C., and Archer, D.E. (2002). Association of sinking organic matter with various types of mineral ballast in the deep sea: implications for the rain ratio. *Glob. Biogeochem. Cycles* 16, 1116.
- Laber, C.P., Hunter, J.E., Carvalho, F., Collins, J.R., Hunter, E.J., Schieler, B.M., Boss, E., More, K., Frada, M., Thamatrakoln, K., et al. (2018). Coccolithovirus facilitation of carbon export in the North Atlantic. *Nat. Microbiol.* 3, 537–547.
- Lawrence, J.E., and Suttle, C.A. (2004). Effect of viral infection on sinking rates of *Heterosigma akashiwo* and its implications for bloom termination. *Aquat. Microb. Ecol.* 37, 1–7.
- Lawrence, J.E., Chan, A.M., and Suttle, C.A. (2002). Viruses causing lysis of the toxic bloom-forming alga *Heterosigma akashiwo* (Raphidophyceae) are widespread in coastal sediments of British Columbia. *Can. Limnol. Oceanogr.* 47, 545–550.
- Leblanc, K., Quéguiner, B., Diaz, F., Cornet, V., Michel-Rodriguez, M., Durrieu de Madron, X., Bowler, C., Malviya, S., Thyssen, M., Grégori, G., et al. (2018). Nanoplanktonic diatoms are globally overlooked but play a role in spring blooms and carbon export. *Nat. Commun.* 9, 953.
- Li, W. (1995). Composition of ultraphytoplankton in the central north-atlantic. *Mar. Ecol. Prog. Ser.* 122, 1–8.
- Liu, H., Probert, I., Uitz, J., Claustre, H., Aris-Brosou, S., Frada, M., Not, F., and de Vargas, C. (2009). Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans. *Proc. Natl. Acad. Sci. U S A* 106, 12803–12808.
- Lomas, M.W., and Moran, S.B. (2011). Evidence for aggregation and export of cyanobacteria and nano-eukaryotes from the Sargasso Sea euphotic zone. *Biogeosciences* 8, 203–216.
- Lønborg, C., Middelboe, M., and Brussaard, C.P.D. (2013). Viral lysis of *Micromonas pusilla*: impacts on dissolved organic matter production and composition. *Biogeochemistry* 116, 231–240.
- Mars Brisbin, M., Mesrop, L.Y., Grossmann, M.M., and Mitarai, S. (2018). Intra-host symbiont diversity and extended symbiont maintenance in photosymbiotic acantharea (clade F). *Front. Microbiol.* 9, 1998.
- Monier, A., Pagarete, A., de Vargas, C., Allen, M.J., Read, B., Claverie, J.-M., and Ogata, H. (2009). Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. *Genome Res.* 19, 1441–1449.
- Monier, A., Worden, A.Z., and Richards, T.A. (2016). Phylogenetic diversity and biogeography of the Mamiellophyceae lineage of eukaryotic phytoplankton across the oceans. *Environ. Microbiol. Rep.* 8, 461–469.
- Moniruzzaman, M., Wurch, L.L., Alexander, H., Dyhrman, S.T., Gobler, C.J., and Wilhelm, S.W. (2017). Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. *Nat. Commun.* 8, 16054.
- Nissimov, J.I., Vandzura, R., Johns, C.T., Natale, F., Haramaty, L., and Bidle, K.D. (2018). Dynamics of transparent exopolymer particle production and aggregation during viral infection of the coccolithophore, *Emiliania huxleyi*. *Environ. Microbiol.* 20, 2880–2897.
- Peduzzi, P., and Weinbauer, M.G. (1993). Effect of concentrating the virus-rich 2-2nm size fraction of seawater on the formation of algal flocs (marine snow). *Limnol. Oceanogr.* 38, 1562–1565.
- Proctor, L.M., and Fuhrman, J.A. (1991). Roles of viral infection in organic particle flux. *Mar. Ecol. Prog. Ser.* 69, 133–142.
- Sheyn, U., Rosenwasser, S., Lehahn, Y., Barak-Gavish, N., Rotkopf, R., Bidle, K.D., Koren, I., Schatz, D., and Vardi, A. (2018). Expression profiling of host and virus during a coccolithophore bloom provides insights into the role of viral infection in promoting carbon export. *ISME J.* 12, 704–713.
- Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., Qin, X.-C., Li, J., Cao, J.-P., Eden, J.-S., et al. (2016). Redefining the invertebrate RNA virosphere. *Nature* 540, 539–543.
- Shiozaki, T., Hirose, Y., Hamasaki, K., Kaneko, R., Ishikawa, K., and Harada, N. (2019). Eukaryotic phytoplankton contributing to a seasonal bloom and carbon export revealed by tracking sequence variants in the western North Pacific. *Front. Microbiol.* 10, 2722.
- Sullivan, M.B., Weitz, J.S., and Wilhelm, S. (2017). Viral ecology comes of age. *Environ. Microbiol. Rep.* 9, 33–35.
- Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., et al. (2015). Ocean plankton. Structure and function of the global ocean microbiome. *Science* 348, 1261359.
- Suttle, C.A. (2007). Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* 5, 801–812.
- Takao, Y., Nagasaki, K., Mise, K., Okuno, T., and Honda, D. (2005). Isolation and characterization of a novel single-stranded RNA Virus infectious to a marine fungoid protist, *Schizochytrium* sp. (Thraustochytriales, Labyrinthales). *Appl. Environ. Microbiol.* 71, 4516–4522.
- Tomaru, Y., Hata, N., Masuda, T., Tsuji, M., Igata, K., Masuda, Y., Yamatogi, T., Sakaguchi, M., and Nagasaki, K. (2007). Ecological dynamics of the bivalve-killing dinoflagellate *Heterocapsa circularisquama* and its infectious viruses in different locations of western Japan. *Environ. Microbiol.* 9, 1376–1383.
- Tomaru, Y., Fujii, N., Oda, S., Toyoda, K., and Nagasaki, K. (2011). Dynamics of diatom viruses on the western coast of Japan. *Aquat. Microb. Ecol.* 63, 223–230.
- Tréguer, P., Bowler, C., Moriceau, B., Dutkiewicz, S., Gehlen, M., Aumont, O., Bittner, L., Dugdale, R., Finkel, Z., Ludicone, D., et al. (2018). Influence of diatom diversity on the ocean biological carbon pump. *Nat. Geosci.* 11, 27–37.
- Turner, J.T. (2015). Zooplankton fecal pellets, marine snow, phytodetritus and the ocean’s biological pump. *Prog. Oceanogr.* 130, 205–248.
- Urayama, S., Takaki, Y., Nishi, S., Yoshida-Takashima, Y., Deguchi, S., Takai, K., and Nunoura, T. (2018). Unveiling the RNA virosphere associated with marine microorganisms. *Mol. Ecol. Resour.* 18, 1444–1455.
- Van Etten, J., Agarkova, I., Dunigan, D., Tonetti, M., De Castro, C., and Duncan, G. (2017). Chloroviruses have a sweet tooth. *Viruses* 9, 88.
- Wang, H., Wu, S., Li, K., Pan, Y., Yan, S., and Wang, Y. (2018). Metagenomic analysis of ssDNA viruses in surface seawater of Yangshan Deep-Water Harbor, Shanghai, China. *Mar. Genomics* 41, 50–53.
- Weinbauer, M.G. (2004). Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* 28, 127–181.
- Weitz, J.S., Stock, C.A., Wilhelm, S.W., Bourouiba, L., Coleman, M.L., Buchan, A., Follows, M.J., Fuhrman, J.A., Jover, L.F., Lennon, J.T., et al. (2015). A multitrophic model to quantify the effects of marine viruses on microbial food webs and ecosystem processes. *ISME J.* 9, 1352–1364.
- Wilhelm, S.W., and Suttle, C.A. (1999). Viruses and Nutrient Cycles in the SeaViruses play critical roles in the structure and function of aquatic food webs. *BioScience* 49, 781–788.
- Yamada, Y., Tomaru, Y., Fukuda, H., and Nagata, T. (2018). aggregate formation during the viral lysis of a marine diatom. *Front. Mar. Sci.* 5, 167.

Yau, S., Krasovec, M., Benites, L.F., Rombauts, S., Groussin, M., Vancaester, E., Aury, J.-M., Derelle, E., Desdevises, Y., Escande, M.-L., et al. (2020). Virus-host coexistence in phytoplankton through the genomic lens. *Sci. Adv.* *6*, eaay2587.

Yoshida, M., Noël, M.-H., Nakayama, T., Naganuma, T., and Inouye, I. (2006). A haptophyte bearing siliceous scales: ultrastructure and phylogenetic position of *Hyalolithus neolepis* gen. et sp. nov.

(Prymnesiophyceae, Haptophyta). *Protist* *157*, 213–234.

Yoshikawa, G., Blanc-Mathieu, R., Song, C., Kayama, Y., Mochizuki, T., Murata, K., Ogata, H., and Takemura, M. (2019). Medusavirus, a novel large DNA virus discovered from hot spring water. *J. Virol.* *93*, e02130–18.

Zhang, C., Dang, H., Azam, F., Benner, R., Legendre, L., Passow, U., Polimene, L., Robinson,

C., Suttle, C.A., and Jiao, N. (2018). Evolving paradigms in biological carbon cycling in the ocean. *Natl. Sci. Rev.* *5*, 481–499.

Zimmerman, A.E., Howard-Varona, C., Needham, D.M., John, S.G., Worden, A.Z., Sullivan, M.B., Waldbauer, J.R., and Coleman, M.L. (2019). Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. *Nat. Rev. Microbiol.* *18*, 21–34.



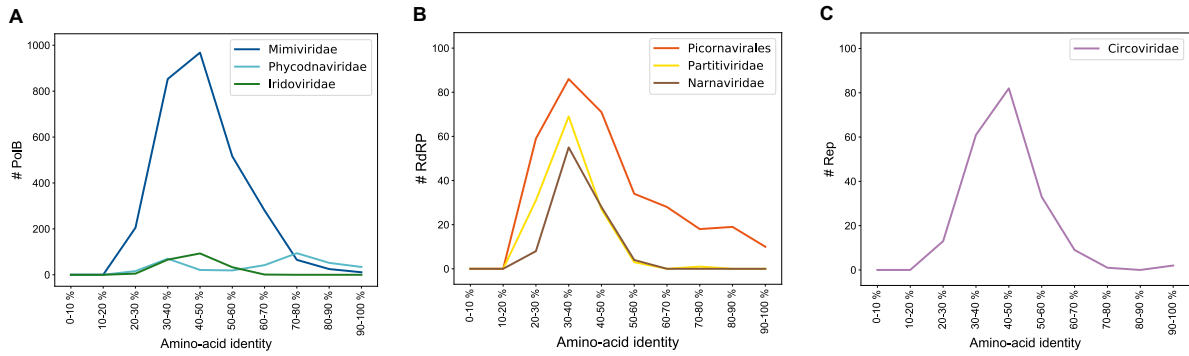
## **Supplemental Information**

### **Eukaryotic virus composition can predict the efficiency of carbon export in the global ocean**

**Hiroto Kaneko, Romain Blanc-Mathieu, Hisashi Endo, Samuel Chaffron, Tom O. Delmont, Morgan Gaia, Nicolas Henry, Rodrigo Hernández-Velázquez, Canh Hao Nguyen, Hiroshi Mamitsuka, Patrick Forterre, Olivier Jaillon, Colombar de Vargas, Matthew B. Sullivan, Curtis A. Suttle, Lionel Guidi, and Hiroyuki Ogata**

1 Supplemental Figures

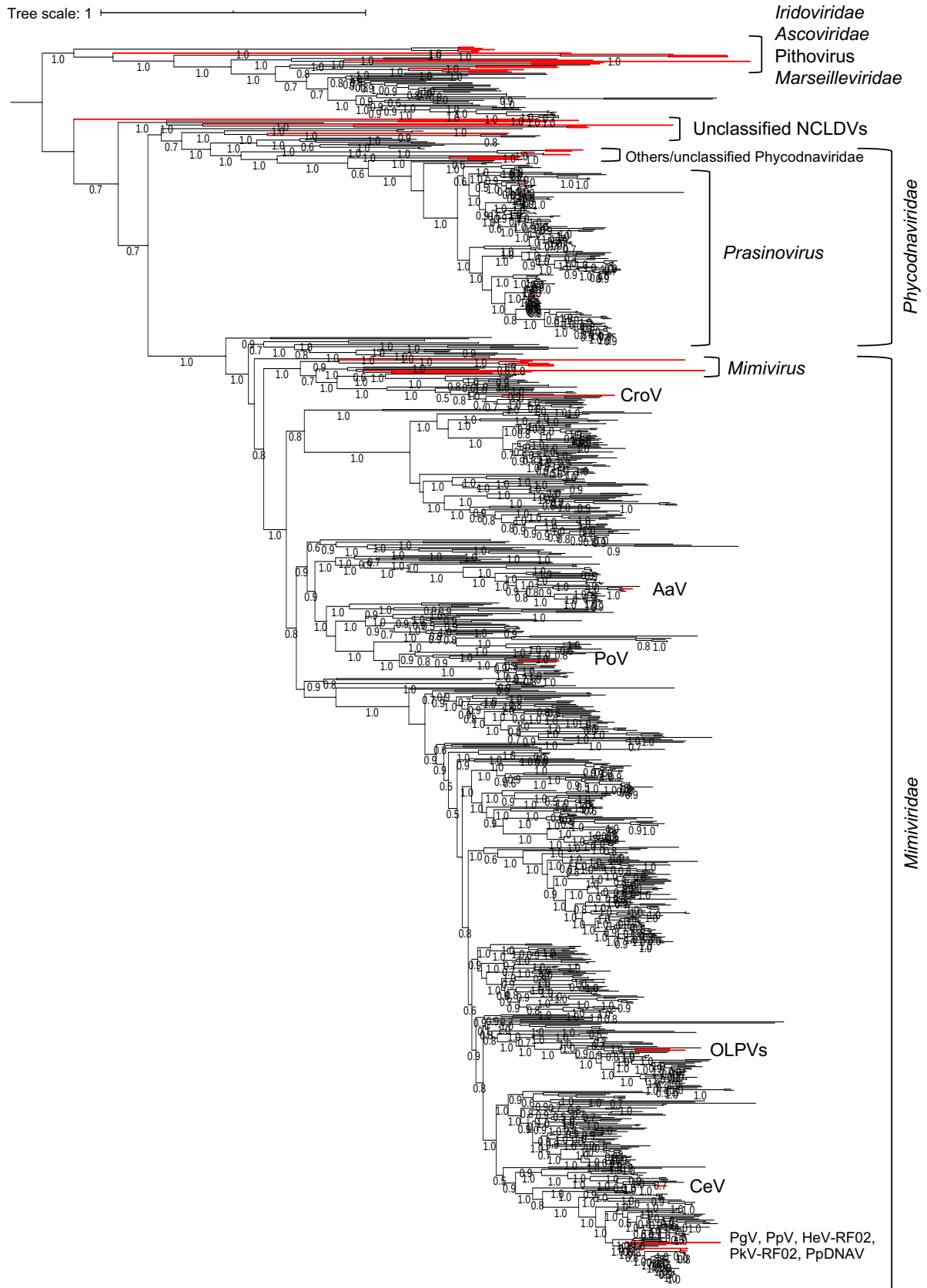
2



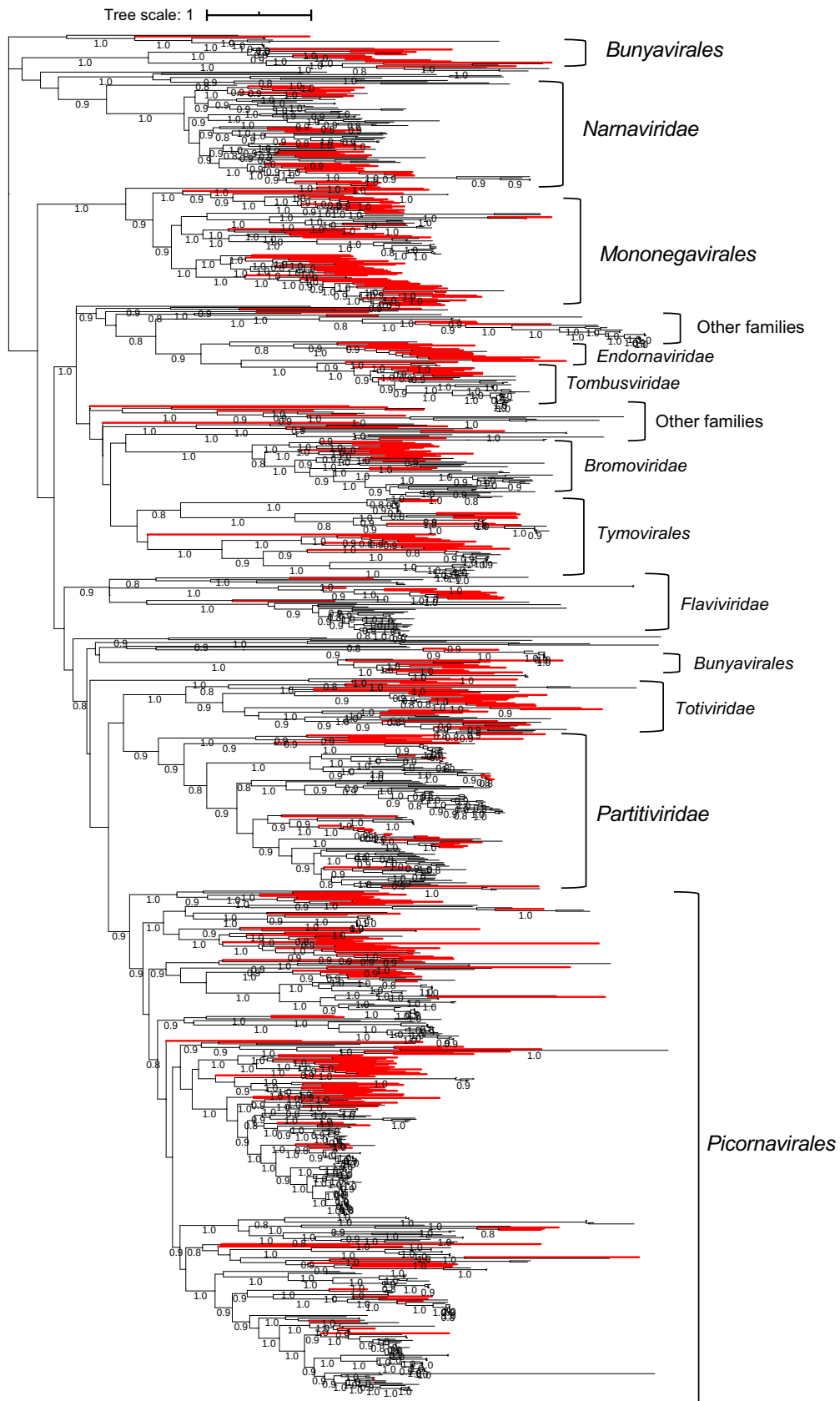
3

4 **Figure S1. Distribution of the degree of amino acid identity between environmental**  
5 **sequences and their best BLAST hits to reference sequences, Related to Figure 1. (A)**  
6 **Nucleocytoplasmic large DNA viruses (NCLDVs). (B) RNA viruses. (C) ssDNA viruses.**

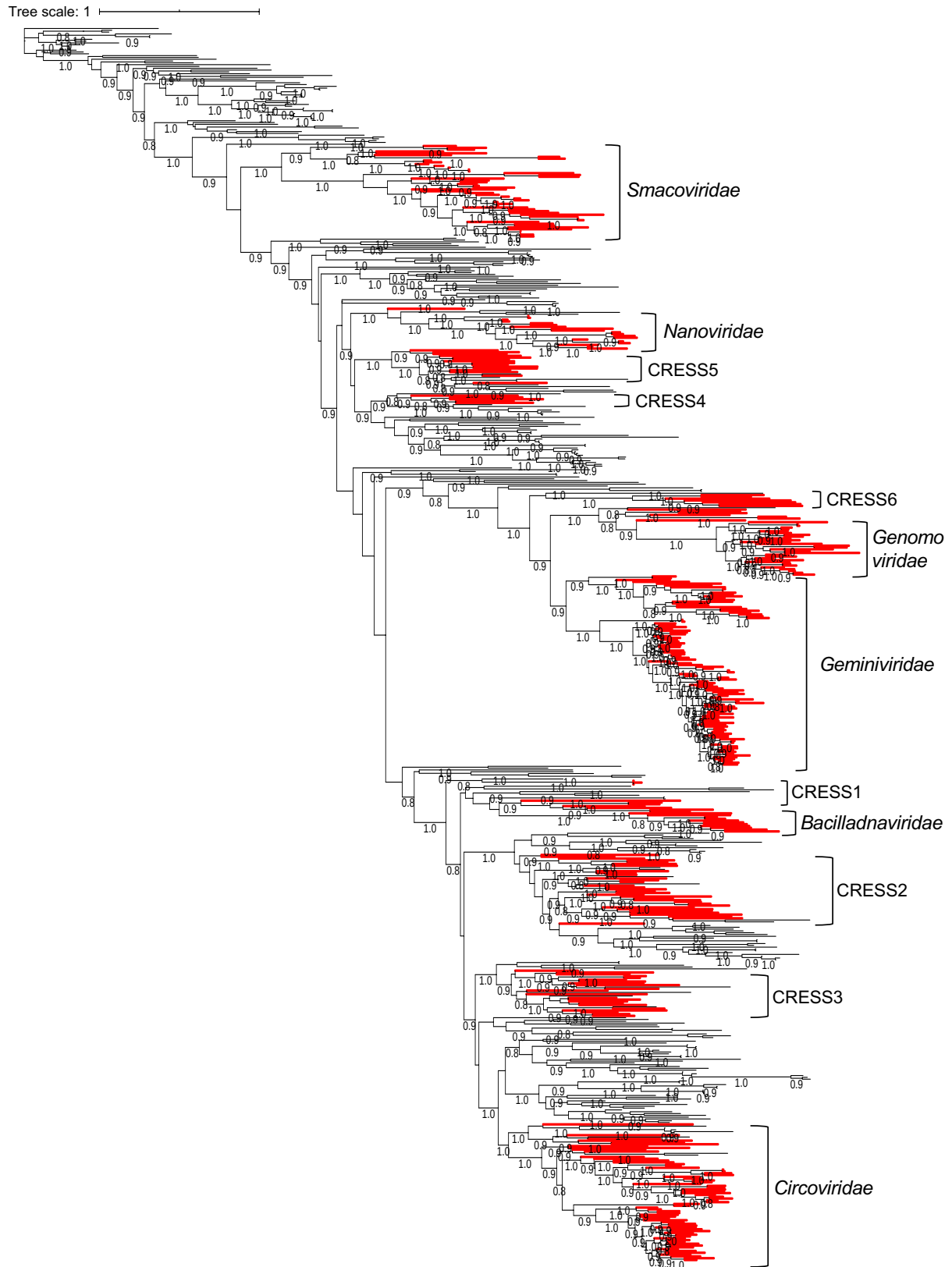
7



8  
9 **Figure S2. Maximum likelihood phylogenetic trees for NCLDV DNA polymerase family**  
10 **B, Related to Figure 1A.** Environmental sequences are shown in black and references in red.  
11 Approximate Shimodaira–Hasegawa (SH)-like local support values greater than 0.8 are  
12 shown. Scale bar indicates one change per site.

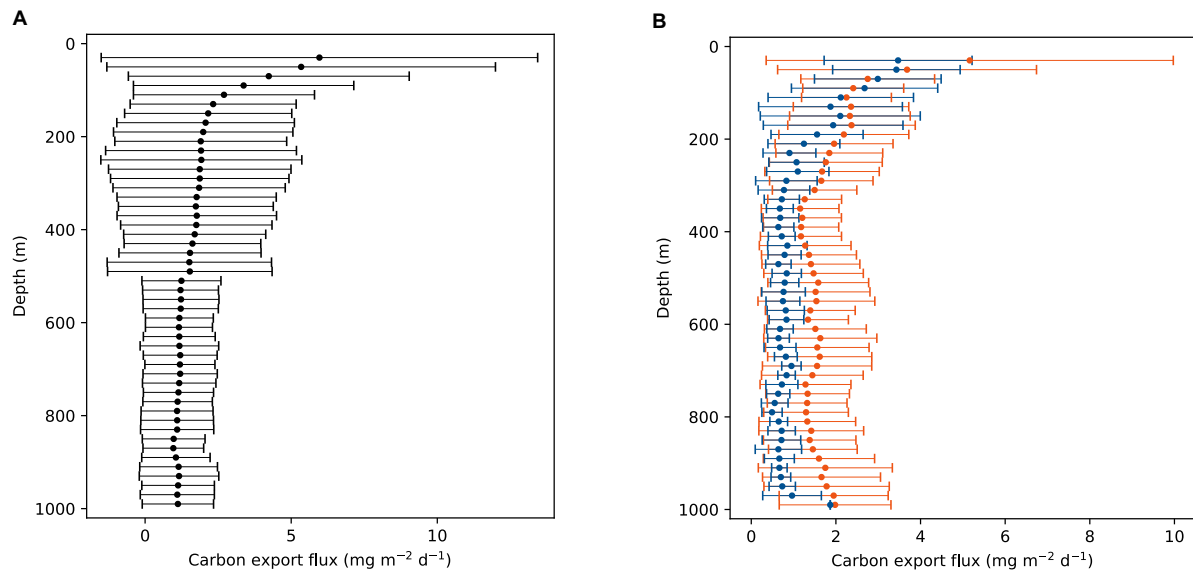


13  
 14 **Figure S3. Unrooted maximum likelihood phylogenetic trees for RNA virus RNA-**  
 15 **dependent RNA polymerase, Related to Figure 1B.** Environmental sequences are shown in  
 16 black and references in red. Approximate SH-like local support values greater than 0.8 are  
 17 shown. Scale bar indicates one change per site.



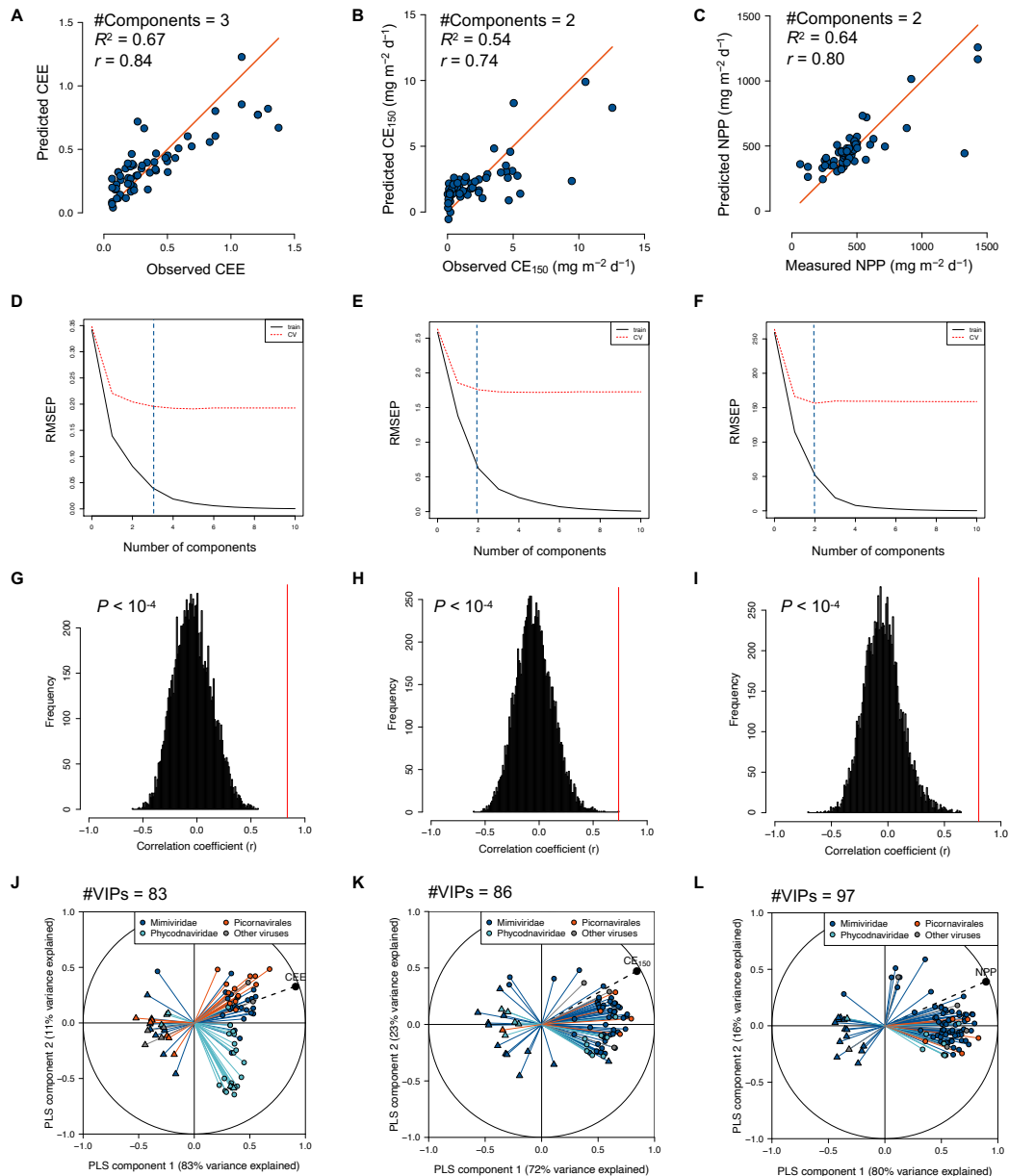
18  
 19  
 20  
 21  
 22  
 23

**Figure S4: Unrooted maximum likelihood phylogenetic trees for ssDNA virus replication-associated protein, Related to Figure 1C.** Environmental sequences are shown in black and references in red. Approximate SH-like local support values greater than 0.8 are shown. Scale bar indicates one change per site.

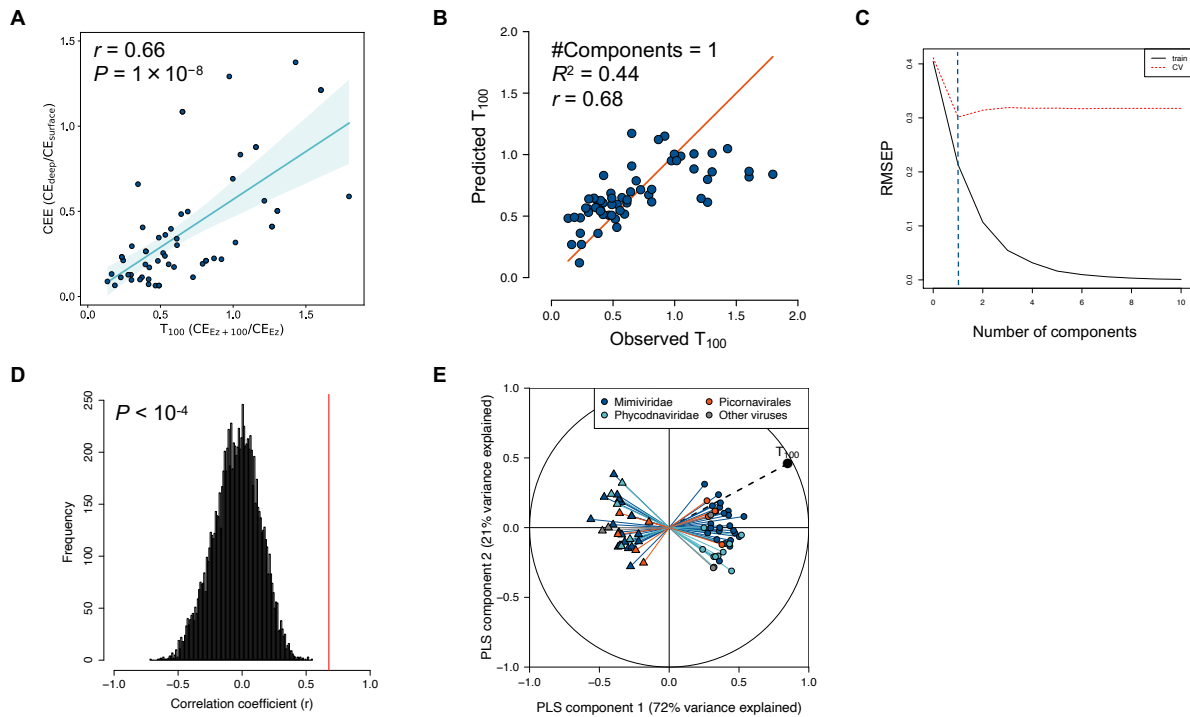


24  
25  
26  
27  
28  
29  
30  
31

**Figure S5. Variation in carbon export flux ( $\text{mg m}^{-2} \text{d}^{-1}$ ) across sampling depths in the water column, Related to Figure 2A; Transparent Methods.** Dots are average values, and horizontal lines represent standard deviation. (A) All sampling sites. (B) Red shows the carbon flux profile of Indian Monsoon Gyres (MONS) where mean CEE is relatively high (0.41) and blue shows that of North Atlantic Subtropical Gyres (West) (NAST-W) where mean CEE is relatively low (0.26).



32  
 33 **Figure S6. The results of PLS regressions using relative abundance profiles of viral**  
 34 **marker-genes to explain the variance of CEE,  $CE_{150}$  and NPP, Related to Figure 3. (A-C)**  
 35 **Bivariate plots between predicted and observed response values in a leave-one-out cross-**  
 36 **validation test. (A) for CEE, (B) for  $CE_{150}$  and (C) for NPP. The red diagonal line shows the**  
 37 **theoretical curve for perfect prediction. (D-F) Variation in root mean squared error of**  
 38 **predictions (RMSEP) for the training set (solid black line) and cross-validation set (red**  
 39 **dashed line) across the number of components. (D) for CEE, (E) for  $CE_{150}$  and (F) for NPP.**  
 40 **Blue dashed line shows the number of components selected for the analysis. (G-I) Results of**  
 41 **the permutation tests ( $n = 10,000$ ) supporting the significance of the association between**  
 42 **viruses and the response variable. (G) for CEE, (H) for  $CE_{150}$  and (I) for NPP. The histograms**  
 43 **show the distribution of Pearson correlation coefficients obtained from PLS models**  
 44 **reconstructed based on the permuted response variable and red line show the non-**  
 45 **permuted response variable. (J-L) Pearson correlation coefficients between the response**  
 46 **variable and abundance profiles of viruses with VIP scores  $> 2$  (VIPs) with the first two**  
 47 **components in the PLS regression model using all samples. (J) for CEE, (K) for  $CE_{150}$  and (L)**  
 48 **for NPP. Viruses with positive regression coefficients are shown with circles, and those with**  
 49 **negative coefficients are shown with triangles.**

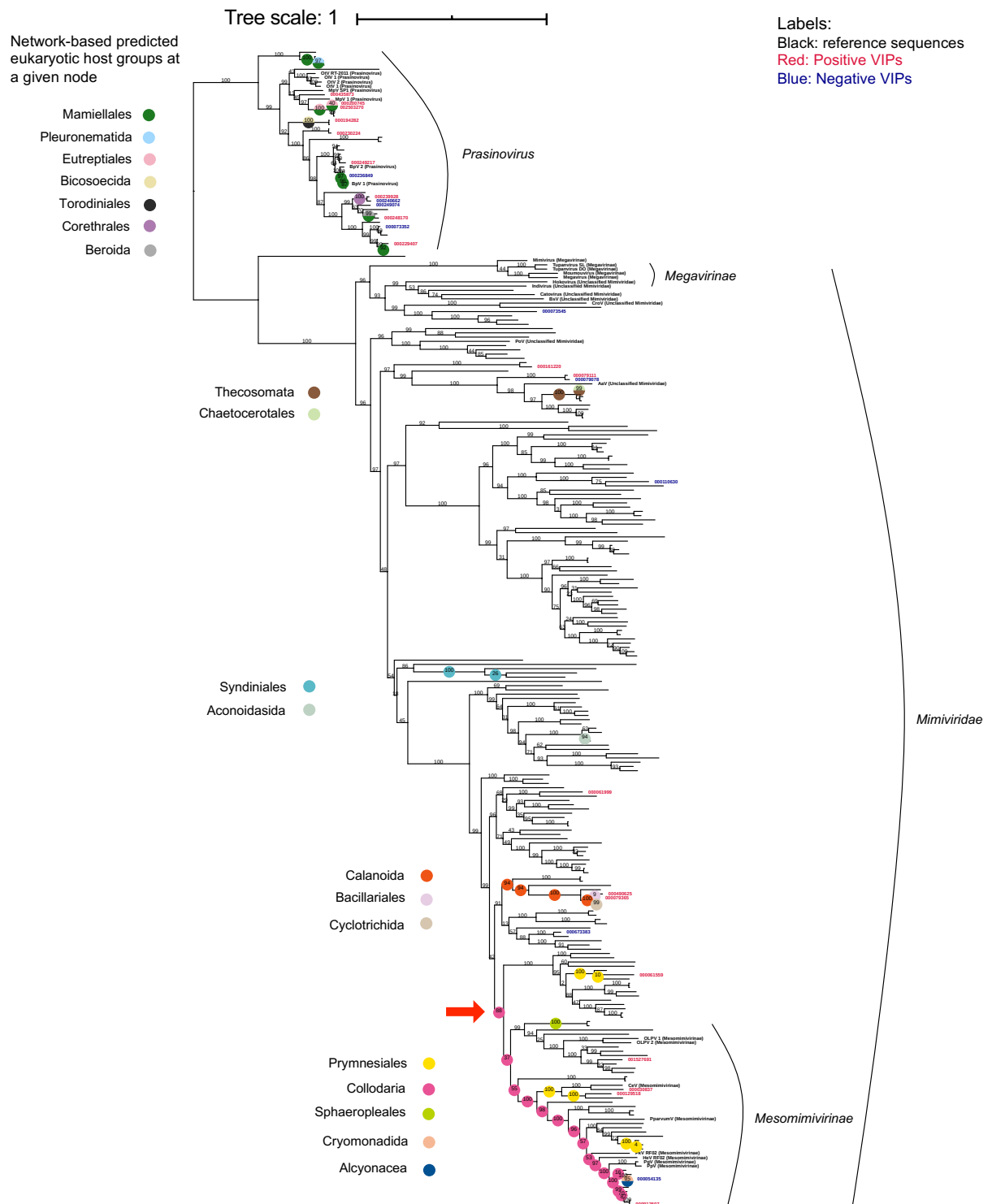


50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62

**Figure S7. The assessment of the sensitivity of the model to the definition of carbon export efficiency, Related to Figure 3.** (A) CEE defined as  $CE_{\text{deep}}/CE_{\text{surface}}$  is well correlated with alternative index of carbon export efficiency defined as  $CE_{Ez+100}/CE_{Ez}$  ( $T_{100}$ ). (B-E) The result of PLS regression using relative abundance profiles of viral marker-genes to explain  $T_{100}$ . (B) Bivariate plots between predicted and observed response values in a leave-one-out cross-validation test. (C) Variation in root mean squared error of predictions (RMSEP) across the number of components. (D) Results of the permutation tests ( $n = 10,000$ ) supporting the significance of the association between viruses and the response variable. (E) Pearson correlation coefficients between the response variable and abundance profiles of viruses with VIP scores  $> 2$  (VIPs) with the first two components in the PLS regression model using all samples. See the legend of Figure S6 for detailed explanation of figures.

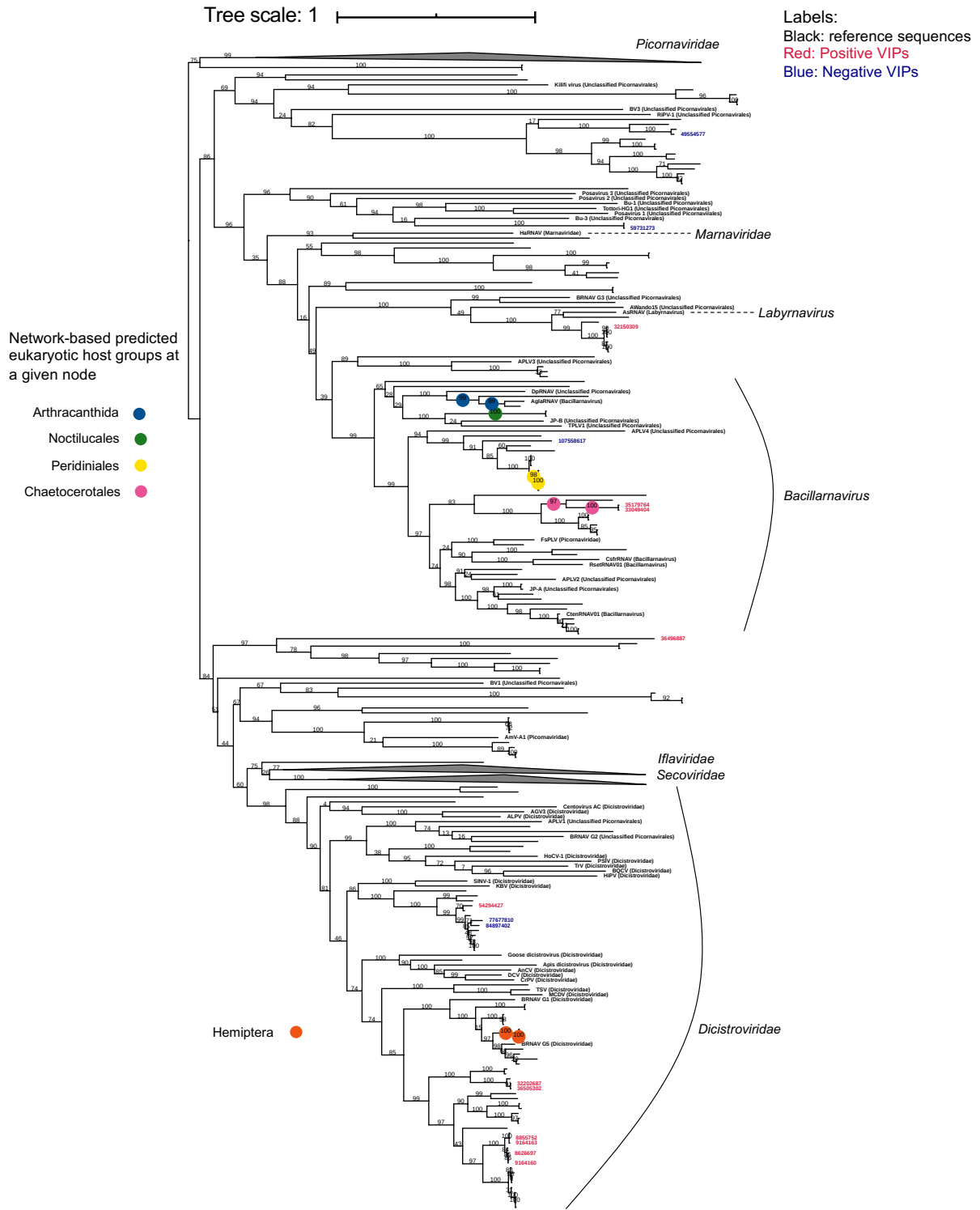




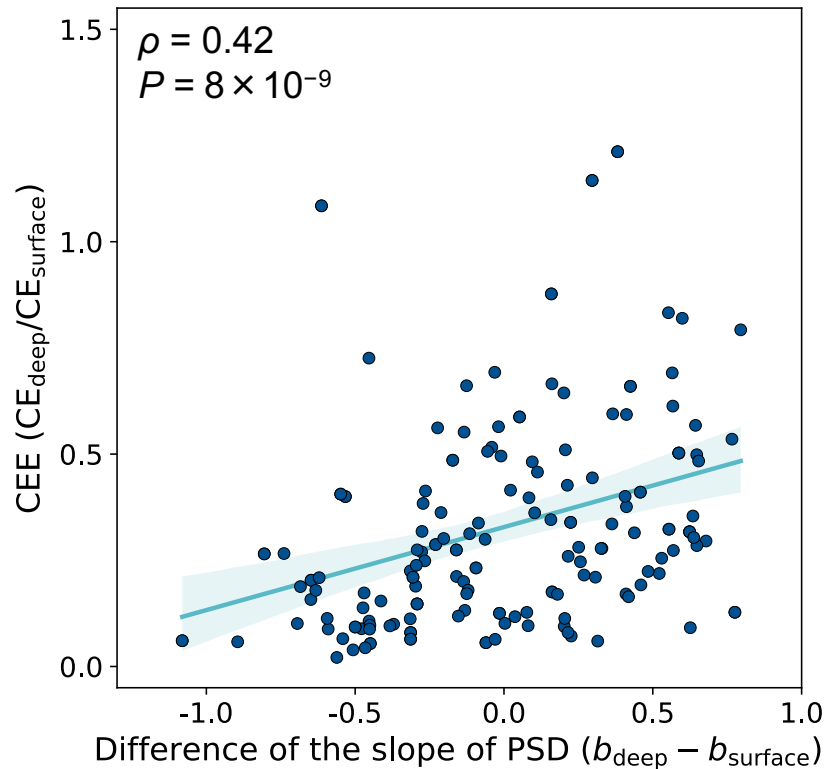


72  
 73  
 74  
 75  
 76  
 77  
 78  
 79  
 80  
 81

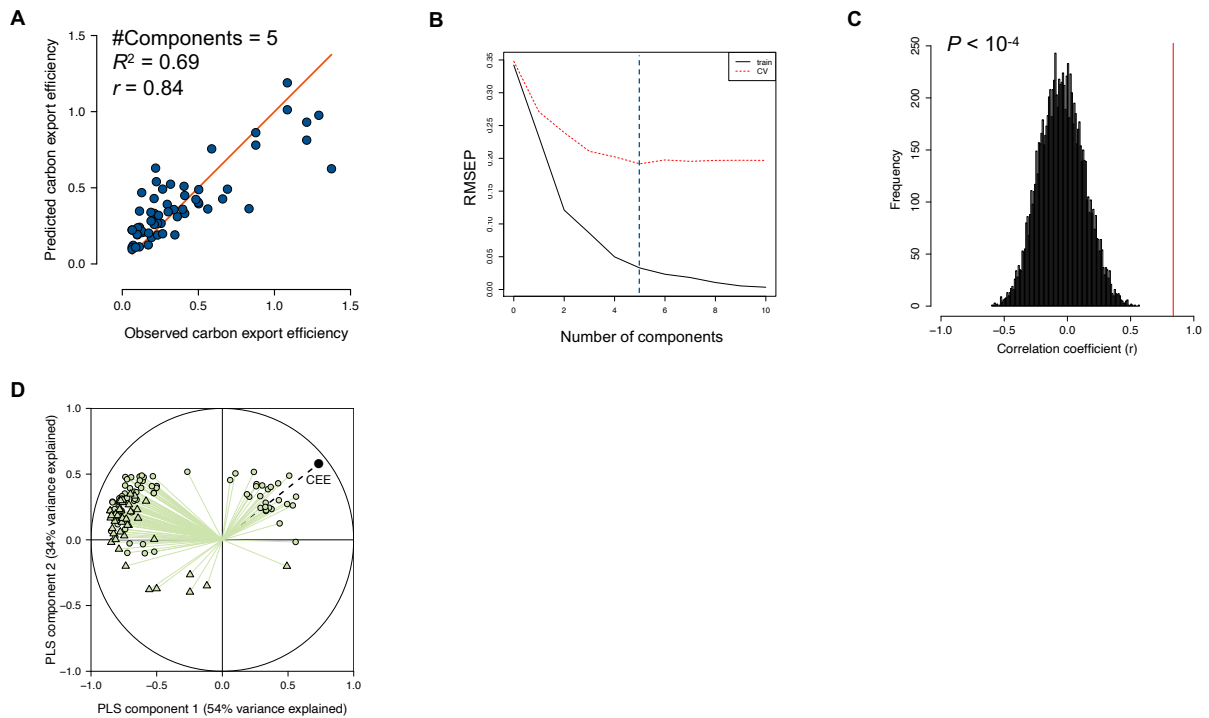
**Figure S9: Phylogenetic positions of NCLDV PolBs associated with CEE and network-based predicted eukaryotic host groups, Related to Table 2; Transparent Methods.** The unrooted maximum likelihood phylogenetic tree contains environmental (labeled in red if VIP score > 2 and the regression coefficient is positive, labeled in blue if negative) and reference (labeled in black) sequences of *Prasinovirus* and *Mimiviridae* PolBs. The approximate SH-like local support values are shown in percentages at nodes, and the scale bar indicates one change per site. Host groups predicted at nodes are shown with colored circles. The red arrow points to a clade of viruses predicted to infect Prymnesiales.



82  
 83 **Figure S10: Phylogenetic position of *Piconavirales* RdRPs associated with CEE and**  
 84 **network-based predicted eukaryotic host groups, Related to Table 2; Transparent**  
 85 **Methods.** The unrooted maximum likelihood phylogenetic tree contains environmental  
 86 (labeled in red if VIP score > 2 and the regression coefficient is positive, labeled in blue if  
 87 negative) and reference (labeled in black) sequences of *Piconavirales* RdRPs. The  
 88 approximate SH-like local support values are shown in percentages at nodes, and the scale bar  
 89 indicates one change per site. Host groups predicted at nodes are shown with colored circles.  
 90



91 **Figure S11. Carbon export efficiency (CEE) is correlated with the change in the slope of**  
 92 **particle size distribution (PSD) that occurred from the surface to deep (below the**  
 93 **euphotic zone), Related to Figure 2A.** Observed PSDs were fitted in the form  $n = ad^b$ , where  
 94  $n$  is the frequency of particles of a given size,  $d$  is the particle diameter, and  $a$  and  $b$  are  
 95 parameters (as described by(Guidi et al., 2008)).  $b$ , the PSD slope, is a proxy for particles size.  
 96 For example,  $b = -5$  indicates presence of a large proportion of smaller particles, whereas  $b =$   
 97  $-3$  indicates a preponderance of larger particles. A higher  $b$  value at deep compared to surface  
 98 is suggestive of aggregation or presence of larger organisms at deep compare to surface. The  
 99 blue line shows the regression line between CEE and the PSD slope difference between  
 100 surface and deep. The shade around the regression line shows the 95% confidence interval.  
 101  
 102



103  
 104 **Figure S12. The result of PLS regression using relative abundance profiles of marker-**  
 105 **genes of T4-like dsDNA bacteriophages to explain CEE, Related to Figure 3. (A)**  
 106 Bivariate plot between predicted and observed response values in a leave-one-out cross-  
 107 validation test. (B) Variation in root mean squared error of predictions (RMSEP) across the  
 108 number of components. (C) Results of the permutation tests ( $n = 10,000$ ) supporting the  
 109 significance of the association between viruses and the response variable. (D) Pearson  
 110 correlation coefficients between the response variable and abundance profiles of viruses with  
 111 VIP scores  $> 2$  (VIPs) with the first two components in the PLS regression model using all  
 112 samples. See the legend of Figure S6 for detailed explanation of figures.  
 113

114 **Supplemental Tables**

115

116

**Table S1. Viral lineages associated with CEE, Related to Figure 3.**

Viruses		VIPs	Positive VIPs	Negative VIPs
NCLDVs	Mimiviridae	34	25	9
	Phycodnaviridae	24	18	6
	Iridoviridae	2	0	2
	Other NCLDVs <sup>a</sup>	0	0	0
	Total	60	43	17
RNA viruses	Picornavirales (ssRNA+)	19	13	6
	Partitiviridae (dsRNA)	1	1	0
	Narnaviridae (ssRNA+)	0	0	0
	Other families	2*	0	2
	Unclassified	0	0	0
	RNA viruses	0	0	0
Total	22	14	8	
ssDNA viruses	Circoviridae	1	1	0
	Geminiviridae	0	0	0
	Nanoviridae	0	0	0
	Unclassified	0	0	0
	ssDNA viruses	0	0	0
Total	1	1	0	
All		83	58	25

117

118

<sup>a</sup>Two Hepeviridae (ssRNA+).

119  
120

**Table S2. Assembly statistics for NCLDV metagenome-assembled genomes and corresponding VIPs, Related to Table 2.**

Metagenome-assembled genome	#contigs	N50 <sup>a</sup>	L50 <sup>b</sup>	Min	Max	Sum	VIPs OTUs (OM-RGC.v1 ID)
TARA_IOS_NCLDV_Bin_127_6	14	21,642	5	8,581	35,822	267,607	PoIB 000079111
TARA_IOS_NCLDV_Bin_173_3	12	12,913	3	2,807	34,517	108,412	PoIB 000248170
TARA_MED_NCLDV_Bin_284_10	34	10,936	10	2,580	29,722	298,760	PoIB 000328966
TARA_MED_NCLDV_Bin_284_14	43	14,837	11	2,756	27,607	439,843	PoIB 001175669
TARA_IOS_NCLDV_Bin_127_4	26	5,734	10	2,560	8,505	133,765	PoIB 001064263 and 010288541
TARA_AON_NCLDV_Bin_289_4	17	9,468	5	3,044	26,201	153,728	PoIB 000200745 and 002503270
TARA_MED_NCLDV_Bin_341_10	5	7,800	2	2,534	7,941	30,478	PoIB 002682999
TARA_PON_NCLDV_Bin_65_10	35	13,866	11	3,781	43,080	382,455	PoIB 000079078
TARA_PON_NCLDV_Bin_102_1	53	4,608	18	2,606	11,485	239,832	PoIB 000495602
TARA_AON_NCLDV_Bin_133_8	8	7,204	3	2,686	10,349	51,009	PoIB 000240662

<sup>a</sup>The length of the contigs for which half of the assembly size is contained in contigs with a length greater than N50.

<sup>b</sup>Number of contigs (or scaffolds) with a size greater or equal to N50.

121  
122  
123

**Table S3. Host prediction per viral OTU for 83 VIPs based on phylogeny, co-occurrence analysis, and genomic context, Related to Table 2.**

Virus types	Virus OTUs	Direction of association with CEE	Classification (LCA annotation)	Clade in the trees used for TIM analysis	TIM-based predicted host	MAGs ID	Genome-based predicted host	Suggested host	Note
NCLDVs	polb_000026723	negative	Mimiviridae	NA	NA	NA	NA	NA	
	polb_000030837	positive	Mimiviridae	Mimiviridae/ Mesomimivirinae	Prymnesiales	NA	NA	Prymnesiales	
	polb_000042601	positive	Mimiviridae	NA	NA	NA	NA	NA	
	polb_000054135	negative	Mimiviridae	Mimiviridae/ Mesomimivirinae	Colodaria	NA	NA	Prymnesiales	
	polb_000061559	positive	Mimiviridae	Mimiviridae/ Mesomimivirinae	Prymnesiales	NA	NA	Prymnesiales	
	polb_000061999	positive	Mimiviridae	Mimiviridae	NA	NA	NA	NA	
	polb_000073352	negative	Phycodnaviridae	Phycodnaviridae/ Prasinovirus	NA	NA	NA	Mamiellales	
	polb_000073545	negative	Mimiviridae	Mimiviridae/ CroV relative	NA	NA	NA	NA	
	polb_000079078	negative	Mimiviridae	Mimiviridae/ AaV relative	NA	NA	PON_NCLDV_Bin_65_10	Pelagophyceae	Pelagophyceae
	polb_000079111	positive	Mimiviridae	Mimiviridae/ AaV relative	NA	NA	IOS_NCLDV_Bin_127_6	Pelagophyceae	Pelagophyceae
	polb_000079365	positive	Mimiviridae	Mimiviridae	NA	NA	NA	NA	NA
	polb_000110630	negative	Mimiviridae	Mimiviridae	NA	NA	NA	NA	NA
	polb_000129518	positive	Mimiviridae	Mimiviridae/ Mesomimivirinae	Prymnesiales	NA	NA	NA	Prymnesiales
	polb_000159717	positive	Phycodnaviridae	NA	NA	NA	NA	NA	NA
	polb_000161220	positive	Mimiviridae	Mimiviridae/ AaV relative	NA	NA	NA	NA	Pelagophyceae
	polb_000172102	positive	Mimiviridae	NA	NA	NA	NA	NA	NA
	polb_000194282	positive	Phycodnaviridae	Phycodnaviridae/ Prasinovirus	Mamiellales	NA	NA	NA	Mamiellales
	polb_000200745	positive	Phycodnaviridae	Phycodnaviridae/ Prasinovirus	Mamiellales	AON_NCLDV_Bin_289_4	NA	NA	Mamiellales
	polb_000229407	positive	Phycodnaviridae	Phycodnaviridae/ Prasinovirus	NA	NA	NA	NA	Mamiellales
	polb_000230224	positive	Phycodnaviridae	Phycodnaviridae/ Prasinovirus	NA	NA	NA	NA	Mamiellales
	polb_000232032	negative	Phycodnaviridae	NA	NA	NA	NA	NA	NA
	polb_000236849	negative	Phycodnaviridae	Phycodnaviridae/ Prasinovirus	Mamiellales	NA	NA	NA	Mamiellales
	polb_000239928	positive	Phycodnaviridae	Phycodnaviridae/ Prasinovirus	NA	NA	NA	NA	Mamiellales
	polb_000240662	negative	Phycodnaviridae	Phycodnaviridae/ Prasinovirus	NA	NA	NA	NA	Mamiellales
	polb_000248170	positive	Phycodnaviridae	Phycodnaviridae/ Prasinovirus	Mamiellales	IOS_NCLDV_Bin_173_3	NA	NA	Mamiellales
	polb_000249074	negative	Phycodnaviridae	Phycodnaviridae/ Prasinovirus	NA	NA	NA	NA	Mamiellales
	polb_000249217	positive	Phycodnaviridae	Phycodnaviridae/ Prasinovirus	NA	NA	NA	NA	Mamiellales
	polb_000251540	negative	Phycodnaviridae	NA	NA	NA	NA	NA	NA
	polb_000328966	positive	Mimiviridae	NA	NA	NA	NCLDV_Bin_284_10	NA	NA
	polb_000396610	positive	Mimiviridae	NA	NA	NA	NA	NA	NA
	polb_000435873	positive	Phycodnaviridae	Phycodnaviridae/ Prasinovirus	NA	NA	NA	NA	Mamiellales
	polb_000490625	positive	Mimiviridae	Mimiviridae	NA	NA	NA	NA	NA
	polb_000495602	negative	Iridoviridae	NA	NA	NA	NCLDV_Bin_102_1	NA	NA
	polb_000503865	positive	Phycodnaviridae	NA	NA	NA	NA	NA	NA
	polb_000673383	negative	Mimiviridae	Mimiviridae	NA	NA	NA	NA	NA
	polb_000844241	negative	Iridoviridae	NA	NA	NA	NA	NA	NA
	polb_000912507	positive	Mimiviridae	Mimiviridae/ Mesomimivirinae	Colodaria	NA	NA	NA	Prymnesiales
	polb_001064263	positive	Phycodnaviridae	NA	NA	NA	IOS_NCLDV_Bin_127_4	Mamiellales	Mamiellales
	polb_001175669	positive	Mimiviridae	NA	NA	NA	MED_NCLDV_Bin_284_14	NA	NA
	polb_001527691	positive	Mimiviridae	Mimiviridae/ Mesomimivirinae	NA	NA	NA	NA	Prymnesiales
	polb_002035391	positive	Phycodnaviridae	NA	NA	NA	NA	NA	NA
	polb_002503270	positive	Phycodnaviridae	Phycodnaviridae/ Prasinovirus	Mamiellales	AON_NCLDV_Bin_289_4	NA	NA	Mamiellales
	polb_002682999	positive	Mimiviridae	NA	NA	NA	NA	NA	NA
	polb_003145223	negative	Mimiviridae	NA	NA	NA	NA	NA	NA
	polb_003319665	positive	Mimiviridae	NA	NA	NA	NA	NA	NA
	polb_003580241	positive	Mimiviridae	NA	NA	NA	NA	NA	NA



	polb_004312996	positive	Phycodnaviridae	NA	NA	NA	NA	NA	
	polb_004775027	negative	Mimiviridae	NA	NA	NA	NA	NA	
	polb_004804559	positive	Mimiviridae	NA	NA	NA	NA	NA	
	polb_007102163	positive	Mimiviridae	NA	NA	NA	NA	NA	
	polb_007423474	negative	Mimiviridae	NA	NA	NA	NA	NA	
	polb_007503502	positive	Mimiviridae	NA	NA	NA	NA	NA	
	polb_007771300	positive	Phycodnaviridae	NA	NA	NA	NA	NA	
	polb_008001141	positive	Mimiviridae	NA	NA	NA	NA	NA	
	polb_010288541	positive	Phycodnaviridae	NA	NA	IOS_NCLDV_Bin_127_4	Mamiellales	Mamiellales	
	polb_013294654	positive	Phycodnaviridae	NA	NA	NA	NA	NA	
	polb_013433452	positive	Mimiviridae	NA	NA	NA	NA	NA	
	polb_014364115	positive	Mimiviridae	NA	NA	NA	NA	NA	
	polb_015514497	positive	Mimiviridae	NA	NA	NA	NA	NA	
	polb_015907472	positive	Phycodnaviridae	NA	NA	NA	NA	NA	
	rdrp_105714054	negative	Picornavirales	NA	NA	NA	NA	NA	
	rdrp_107558617	negative	Picornavirales	Picornavirales/ Bacillamavirus	NA	NA	NA	NA	
	rdrp_30787766	positive	Picornavirales	NA	NA	NA	NA	NA	
	rdrp_32150057	positive	Picornavirales	NA	NA	NA	NA	NA	
	rdrp_32150309	positive	Picornavirales	Picornavirales/ Labyrinthomycetes	NA	NA	NA	Labyrinthomycetes	a
	rdrp_32202687	positive	Picornavirales	Picornavirales/ Dicistroviridae	NA	NA	NA	Copepoda	b
	rdrp_33049404	positive	Picornavirales	Picornavirales/ Bacillamavirus	Chaetocerotal es	NA	NA	Chaetocerotal es	
	rdrp_35179764	positive	Picornavirales	Picornavirales/ Bacillamavirus	Chaetocerotal es	NA	NA	Chaetocerotal es	
	rdrp_35713768	positive	Partitiviridae	NA	NA	NA	NA	NA	
	rdrp_36496887	positive	Picornavirales	Picornavirales	NA	NA	NA	NA	
	rdrp_36505302	positive	Picornavirales	Picornavirales/ Dicistroviridae	NA	NA	NA	Copepoda	b
	rdrp_42335229	negative	Hepeviridae	NA	NA	NA	NA	NA	
	rdrp_49554577	negative	Picornavirales	Picornavirales	NA	NA	NA	NA	
	rdrp_54294427	positive	Picornavirales	Picornavirales/ Dicistroviridae	NA	NA	NA	Copepoda	b
	rdrp_59731273	negative	Picornavirales	Picornavirales	NA	NA	NA	NA	
	rdrp_77677770	negative	Hepeviridae	NA	NA	NA	NA	NA	
	rdrp_77677810	negative	Picornavirales	Picornavirales/ Dicistroviridae	NA	NA	NA	Copepoda	b
	rdrp_84897402	negative	Picornavirales	Picornavirales/ Dicistroviridae	NA	NA	NA	Copepoda	b
	rdrp_8626697	positive	Picornavirales	Picornavirales/ Dicistroviridae	NA	NA	NA	Copepoda	b
	rdrp_8855752	positive	Picornavirales	Picornavirales/ Dicistroviridae	NA	NA	NA	Copepoda	b
	rdrp_9164160	positive	Picornavirales	Picornavirales/ Dicistroviridae	NA	NA	NA	Copepoda	b
	rdrp_9164163	positive	Picornavirales	Picornavirales/ Dicistroviridae	NA	NA	NA	Copepoda	b
ssDNA viruses	rep_38177659	positive	Circoviridae	NA	NA	NA	NA	Copepoda	c

<sup>a</sup>This virus was located in well-separated clade containing Aurantiochytrium single-stranded RNA virus (AsRNAV) which is known to infect Labyrinthulomycetes.

<sup>b</sup>These viruses were grouped within Dicistroviridae (known to infect insects) and may therefore infect marine arthropods such as copepods.

<sup>c</sup>This virus was connected with a copepod, mollusk and Collodaria OTUs in the co-occurrence network reconstructed for the mesoplankton size. Circoviridae-like viruses are known to infect copepod.

126  
127  
128  
129  
130  
131  
132

133  
134

**Table S4. Statistics for the FlashWeave co-occurrence graphs, Related to Table 3; Transparent Methods.**

Viral marker gene	Planktonic size fraction <sup>a</sup>	#Samples	#Viral OTUs	#Eukaryotic OTUs	#Edges in graph	#Virus-to-eukaryote edges	#Viruses connected to a eukaryote (%)
NCLDVs PolB	Piconano	99	2269	4936	20934	3594	1735 (76)
	Nano	51	1775	1872	6704	1027	721 (41)
	Micro	92	2205	2524	12189	2101	1299 (59)
	Meso	95	2238	2250	11624	1796	1126 (50)
RNA viruses RdRP	Piconano	60	125	4484	10754	446	122 (98)
	Nano	36	53	1768	2659	124	46 (87)
	Micro	62	124	2407	5351	367	117 (94)
	Meso	62	48	2100	4329	116	42 (88)
ssDNA viruses Rep	Piconano	60	64	4484	10577	205	63 (98%)
	Nano	36	1	1768	2563	2	1 (100%)
	Micro	62	4	2407	5086	9	4 (100%)
	Meso	62	8	2100	4242	24	8 (100%)

<sup>a</sup>Pico: 0.8 to 5 µm, Nano: 5 to 20 µm, Micro: 20 to 180 µm, Meso: 180 to 2000 µm

135  
136

137  
138

**Table S5: Functional differences between eukaryotes found to be best connected to negative VIPs and non-VIPs, Related to Table 3.**

Functional trait	Negative VIPs ( <i>n</i> = 21)		Non-VIPs ( <i>n</i> = 983)		<i>P</i> -value (Fisher's exact test, two sided)	Adjusted <i>P</i> - value (BH) ( <i>Q</i> )
	Presence	Absence	Presence	Absence		
Chloroplast	3	17	276	690	0.218	0.655
Silicification	0	21	60	920	0.632	0.947
Calcification	0	21	30	950	1.000	1.000

139

140  
141

**Table S6: Functional differences between eukaryotes found to be best connected to positive and negative VIPs, Related to Table 3.**

Functional trait	Positive VIPs (n = 50)		Negative VIPs (n = 21)		<i>P</i> -value (Fisher's exact test, two sided)	Adjusted <i>P</i> - value (BH) ( <i>Q</i> )
	Presence	Absence	Presence	Absence		
Chloroplast	20	30	3	17	0.053	0.079
Silicification	11	39	0	21	0.027	0.080
Calcification	1	49	0	21	1.000	1.000

142

## 143 **Transparent Methods**

### 144 **Data context**

145 We used publicly available data generated in the framework of the *Tara* Oceans expedition.  
146 Single-copy marker-gene sequences for NCLDV and RNA viruses were identified from two  
147 gene catalogs: the Ocean Microbial Reference Gene Catalog (OM-RGC) and the Marine Atlas  
148 of *Tara* Oceans Unigenes (MATOU). The viral marker-gene read count profiles used in our  
149 study are as previously reported for prokaryotic-sized metagenomes (size fraction 0.2–3  $\mu\text{m}$ )  
150 (Sunagawa et al., 2015) and eukaryotic-sized metatranscriptomes (Carradec et al., 2018).  
151 Eukaryotic plankton samples (the same samples were used for metatranscriptomes,  
152 metagenomes and 18S rRNA V9 meta-barcodes) were filtered for categorization into the  
153 following size classes: piconano (0.8–5  $\mu\text{m}$ ), nano (5–20  $\mu\text{m}$ ), micro (20–180  $\mu\text{m}$ ), and meso  
154 (180–2,000  $\mu\text{m}$ ). Eukaryotic 18S rRNA V9 meta-barcodes used in this study (Ibarbalz et al.,  
155 2019) included functional trait annotations (chloroplast-bearing, silicified, and calcified  
156 organisms) based on a literature survey. These functionally annotated sequences are available  
157 from Zenodo (Henry et al., 2019). Indirect measurements of carbon export ( $\text{mg m}^{-2} \text{d}^{-1}$ ) in 5-  
158 m increments from the surface to a 1,000-m depth were taken from Guidi et al. (Guidi et al.,  
159 2016) The original measurements were derived from the distribution of particle sizes and  
160 abundances collected using an underwater vision profiler. These raw data are available from  
161 PANGEA (Picheral et al., 2014). Net primary production (NPP) data were extracted and  
162 averaged from 8-day composites of the vertically generalized production model (VGPM)  
163 (Behrenfeld and Falkowski, 1997) for the week of sampling. Thus, in this study, the  
164 comparisons between NPP and other parameters were not made at the same time point. This  
165 might have affected the results of the regression analysis, especially if there were any short-  
166 term massive bloom events, although there was no bloom signal during the sampling period.

## 167 **Carbon export, carbon export efficiency, and particle size distribution**

168 Carbon flux profiles ( $\text{mg m}^{-2} \text{d}^{-1}$ ) were estimated based on particle size distributions and  
169 abundances. The method used for carbon flux estimation was previously calibrated comparing  
170 sediment trap measurement and data from imaging instruments (Guidi et al., 2008). Carbon  
171 flux values from depths of 30 to 970 meters were divided into 20-m bins, each obtained by  
172 averaging the carbon flux values from the designated 20 m in profiles gathered during  
173 biological sampling within a 25-km radius over 24 h when less than 50% of data were missing  
174 (Figure S5). Carbon export (CE) was defined as the carbon flux at 150 m (Guidi et al., 2016).  
175 Carbon export efficiency was calculated as follows:  $\text{CEE} = \text{CE}_{\text{deep}}/\text{CE}_{\text{surface}}$ . To compare  
176 stations with different water column structures, we defined  $\text{CE}_{\text{surface}}$  as the maximum CE (in a  
177 20 m window) within the first 150 m.  $\text{CE}_{\text{deep}}$  is the average CE (also in a 20 m window) 200  
178 m below this maximum. The 150 m limit serves as a reference point to automatize the  
179 calculation of  $\text{CE}_{\text{surface}}$  and  $\text{CE}_{\text{deep}}$ . The 150m-depth layer was selected because often used as a  
180 reference depth for drifting sediment trap and because most of the deep chlorophyll maximum  
181 (DCM) were shallower except at two (stations 98 (175 m) and 100 (180 m)). The maximum  
182  $\text{CE}_{\text{surface}}$  for these two stations was above 150 m. The sampling strategy of *Tara Oceans* was  
183 designed to study a variety of marine ecosystems and to target well-defined meso- to large-  
184 scale features (based on remote-sensing data). Therefore, this strategy avoided sampling water  
185 with important lateral inputs. Nevertheless, the possibility of having locations with potential  
186 lateral transport cannot be excluded.

187 We also calculated an alternative definition of carbon export efficiency relying on  
188 euphotic zone depth ( $T_{100}$ ), which is often used in the analysis of sediment trap/Thorium field  
189 data.  $T_{100}$  was calculated as CE 100 m below euphotic zone depth (Ez) divided by CE at Ez  
190 (Buesseler et al., 2020). Ez was estimated based on the diffuse attenuation coefficient at 490

191 nm ( $K_d(490)$ ) using the empirical model (Lin et al., 2016).  $K_d(490)$  values were extracted  
192 from GlobColour monthly mapped product (<ftp://ftp.hermes.acri.fr>) built using satellite data.

193 We obtained the particle size distribution (PSD) profiles generated by the *Tara* Oceans  
194 expedition and computed the PSD slope at each depth for all profiles. The slope value  
195 (denoted “ $b$ ”) is used as the descriptor of the particle size distribution as defined in a previous  
196 work (Guidi et al., 2009). For example,  $b = -5$  indicates the presence of a large proportion of  
197 smaller particles, whereas  $b = -3$  indicates a preponderance of larger particles. We averaged  
198 the slope values at each sampling site in the same way as for carbon export flux.

### 199 **Identification of viral marker genes from ocean gene catalogs**

200 Viral genes were collected from two gene catalogs: OM-RGC version 1 and MATOU.  
201 Sequences in these two gene catalogs are representatives of clusters of environmental  
202 sequences (clustered at 95% nucleotide identity). The OM-RGC data were taxonomically re-  
203 annotated, with the NCBI reference tree used to determine the last common ancestor modified  
204 to reflect the current classification of NCLDV (Carradec et al., 2018). We automatically  
205 classified viral gene sequences as eukaryotic or prokaryotic according to their best BLAST  
206 score against viral sequences in the Virus-Host Database (Mihara et al., 2016). DNA  
207 polymerase B (PolB), RNA-dependent RNA polymerase (RdRP), replication-associated  
208 protein (Rep), and major capsid protein (Gp23) genes were used as markers for NCLDVs,  
209 RNA viruses, ssDNA viruses, and T4-like dsDNA bacteriophages, respectively. For PolB,  
210 reference proteins from the NCLDV orthologous gene cluster NCV0G0038 (Yutin et al.,  
211 2009) were aligned using MAFFT-*linsi* (Kato and Standley, 2013). A hidden Markov model  
212 (HMM) profile was constructed from the resulting alignment using *hmmbuild* (Eddy, 2011).  
213 This PolB HMM profile was searched against OM-RGC amino acid sequences and translated  
214 MATOU sequences annotated as NCLDVs, and sequences longer than 200 amino acids that  
215 had hits with  $E$ -values  $< 1 \times 10^{-5}$  were selected as putative PolBs. RdRP sequences were

216 chosen from the MATOU catalog as follows: sequences assigned to Pfam profiles PF00680,  
217 PF00946, PF00972, PF00978, PF00998, PF02123, PF04196, PF04197, or PF05919 and  
218 annotated as RNA viruses were retained as RdRPs. For Rep, we reconstructed an HMM  
219 profile using a comprehensive set of reference sequences (Kazlauskas et al., 2018) and  
220 searched this profile against the translated MATOU sequences annotated as ssDNA viruses.  
221 For Gp23, OM-RGC sequences assigned to Pfam profile PF07068 and annotated as viruses  
222 were retained. We kept sequences that had hits with  $E$ -values  $< 1 \times 10^{-5}$  and removed those  
223 that contained frameshifts.

224         The procedure above identified 3,486 PolB and 6,438 Gp23 sequences in the  
225 metagenomic samples and 975 RdRP, 388 PolB, and 299 Rep sequences in the  
226 metranscriptomes.

## 227 **Testing for associations between viruses with CEE, CE<sub>150</sub>, NPP, and T<sub>100</sub>**

228 To test for associations between occurrence of viral marker genes and CEE, CE<sub>150</sub>, NPP, and  
229 T<sub>100</sub> (response variables), we proceeded as follows. Samples with CEE values greater than  
230 one and with  $Z$ -score greater than two were considered as outliers and removed (this removed  
231 the two samples from station 68). Only marker genes represented by at least two reads in five  
232 or more samples were retained (lowering this minimal number of required samples down to  
233 three or four did not improve the PLS regression model). To cope with the sparsity and  
234 composition of the data, gene-length normalized read count matrices were center log-ratio  
235 transformed, separately for ssDNA viruses, RNA viruses and NCLDVs. We next selected  
236 genes with Spearman correlation coefficients with the response variable greater than 0.2 or  
237 smaller than  $-0.2$  (zero values were removed). To assess the association between these  
238 marker genes and the response variable, we used partial least square (PLS) regression analysis.  
239 The number of components selected for the PLS model was chosen to minimize the root mean  
240 square error of prediction (Figures S6 and S7). We assessed the strength of the association



241 between the response variable and viral marker genes occurrence (the explanatory variables)  
242 by correlating leave-one-out cross-validation predicted values with the measured values of the  
243 response variable. We tested the significance of the correlation by comparing the original  
244 Pearson coefficients between explanatory and response variables with the distribution of  
245 Pearson coefficients obtained from PLS models reconstructed based on permuted data  
246 (10,000 iterations). We estimated the contribution of each gene (predictor) according to its  
247 variable importance in the projection (VIP) score derived from the PLS regression model  
248 using all samples. The VIP score of a predictor estimates its contribution in the PLS  
249 regression. Predictors with high VIP scores ( $> 2$ ) were assumed to be important for the PLS  
250 prediction of the response variable.

## 251 **Phylogenetic analysis**

252 Environmental PolB sequences annotated as NCLDV s were searched against reference  
253 NCLDV PolB sequences using BLAST. Environmental sequences with hits to a reference  
254 sequence that had  $> 40\%$  identity and an alignment length greater than 400 amino acids were  
255 kept and aligned with reference sequences using MAFFT-*linsi*. Environmental RdRP  
256 sequences were translated into six frames of amino acid sequences and combined together  
257 with reference RNA viruses RdRP sequences collected from the Virus-Host Database. They  
258 were searched against the Conserved Domain Database (CDD) using rpsBLAST. The  
259 resulting alignment was used to trim reference and environmental RdRP sequences to the  
260 conserved part corresponding to the domain, CDD: 279070, before alignment with MAFFT-  
261 *linsi*. Rep sequences annotated as ssDNA viruses were treated similarly. PolB, RdRP, and  
262 Rep multiple sequence alignments were manually curated to discard poorly aligned sequences.  
263 Phylogenetic trees were reconstructed using the the *build* function of ETE3 (Huerta-Cepas et  
264 al., 2016) of the GenomeNet TREE tool (<https://www.genome.jp/tools-bin/ete>). Columns

265 were automatically trimmed using *trimAl* (Capella-Gutiérrez et al., 2009), and trees were  
266 constructed using FastTree with default settings (Price et al., 2009).

267 A similar procedure was applied for the trees used in the hosts prediction analysis  
268 albeit selecting sequences for the Phycodnaviridae/Mimiviridae (Figure S9) and the  
269 Picornavirales (Figure S10) and removing the ones occurring in fewer than 10 samples, to  
270 reduce the size of the tree.

### 271 **Virus–eukaryote co-occurrence analysis**

272 We used FlashWeave (Tackmann et al., 2019) with Julia 1.2.0 to predict virus–host  
273 interactions based on their co-occurrence patterns. FlashWeave is a novel approach to  
274 inferring direct co-occurrence associations based on the local-to-global learning. Read count  
275 matrices for the 3,486 PolBs, 975 RdRPs, 299 Repls, and 18S rRNA V9 DNA barcodes  
276 obtained from samples collected at the same location were fed into FlashWeave. The 18S  
277 rRNA V9 data were filtered to retain OTUs with an informative taxonomic annotation. The  
278 18S rRNA V9 OTUs and viral marker sequences were further filtered to conserve only those  
279 present in at least five samples. FlashWeave networks were learned for each of the four  
280 eukaryotic size fractions with the parameters ‘heterogenous’ = false and ‘sensitive’ = true,  
281 and edges receiving a weight > 0.2 and a  $Q$ -value < 0.01 (the default) were retained. The  
282 number of samples per size fraction ranged between 51 and 99 for NCLDV and between 36  
283 and 62 for RNA and ssDNA viruses. The number of retained OTUs per size fraction varied  
284 between 1,775 and 2,269 for NCLDV and between 48 and 125 for RNA viruses (Table S4).

### 285 **Mapping of putative hosts onto viral phylogenies**

286 In our association networks, individual viral sequences were often associated with multiple  
287 18S rRNA V9 OTUs belonging to drastically different eukaryotic groups, a situation that can  
288 reflect interactions among multiple organisms but also noise associated with this type of

289 analysis (Coenen and Weitz, 2018). To extract meaningful information from these networks,  
290 we reasoned as follows. We assumed that evolutionarily related viruses infect evolutionarily  
291 related organisms, similar to the case of phycodnaviruses (Clasen and Suttle, 2009). In the  
292 interaction networks, the number of connections between viruses in a given clade and the  
293 associated eukaryotic host group should accordingly be enriched compared with the number  
294 of connections with non-host organisms arising by chance. Following this reasoning, we  
295 assigned the most likely eukaryotic host group as follows. The tree constructed from viral  
296 marker-gene sequences (PolB, RdRP or Rep) was traversed from root to tips to visit every  
297 node. We counted how many connections existed between leaves of each node and the V9-  
298 OTUs of a given eukaryotic group (order level). We then tested whether the node was  
299 enriched compared with the rest of the tree using Fischer's exact test and applied the  
300 Benjamini–Hochberg procedure to control the false discovery rate among comparisons of  
301 each eukaryotic taxon (order level). To avoid the appearance of significant associations driven  
302 by a few highly connected leaves, we required half of the leaves within a node to be  
303 connected to a given eukaryotic group. Significant enrichment of connections between a virus  
304 clade and a eukaryotic order was considered to be indicative of a possible virus–host  
305 relationship. We refer to the above approach, in which taxon interactions are mapped onto a  
306 phylogenetic tree of a target group using the organism's associations predicted from a species  
307 co-occurrence-based network, as TIM, for Taxon Interaction Mapper. This tool is available at  
308 <https://github.com/RomainBlancMathieu/TIM>. This approach can be extended to interactions  
309 other than virus–host relationships. It has been shown that TIM filtering improves the  
310 performance of network-based host prediction for NCLDVs in a benchmark study (Meng et al.  
311 (2020). bioRxiv <https://doi.org/10.1101/2020.10.16.342030>).

## 312 **Assembly of NCLDV metagenome-assembled genomes (MAGs)**

313 NCLDV metagenome-assembled genomes (MAGs) were assembled from *Tara* Oceans  
314 metagenomes corresponding to size fractions  $> 0.8 \mu\text{m}$ . Metagenomes were first organized  
315 into 11 ‘metagenomic sets’ based upon their geographic coordinates, and each set was co-  
316 assembled using MEGAHIT (Li et al., 2015) v.1.1.1. For each set, scaffolds longer than 2.5  
317 kbp were processed within the bioinformatics platform anvi’o (Eren et al., 2015) v.6.1  
318 following methodology described previously for genome-resolved metagenomics (Delmont et  
319 al., 2018). Briefly, we used the automatic binning algorithm CONCOCT (Alneberg et al.,  
320 2014) to identify large clusters of contigs using both sequence composition and differential  
321 coverage across metagenomes within the set. We then used HMMER (Eddy, 2011) v3.1b2 to  
322 search for a collection of eight NCLDV gene markers (Guglielmini et al., 2019), and  
323 identified NCLDV MAGs by manually binning CONCOCT clusters of interest using the  
324 anvi’o interactive interface. The interface displayed hits for the eight gene markers alongside  
325 coverage values across metagenomes and GC-content. Finally, NCLDV MAGs were  
326 manually curated using the same interface, to minimize contamination as described previously  
327 (Delmont and Eren, 2016).

## 328 **Taxonomic composition of genes predicted in NCLDV genomes of VIPs**

329 VIP’s PolB sequences were searched (using BLAST) against MAGs reconstructed from the  
330 metagenomes of the eukaryotic size fraction ( $> 0.8 \mu\text{m}$ ) and against contigs used to produce  
331 OM-RGCv1. Genome fragments covering 95% of the length of PolB VIPs with  $> 95\%$   
332 nucleotide identity were considered as originating from a same viral OTUs. Genes were  
333 predicted and annotated taxonomically with the same procedure described above  
334 (identification of viral marker genes). Genes contained in viral genome fragments and  
335 annotated as cellular organisms with amino acid identities  $> 60\%$  were manually inspected  
336 ([Supplemental Data 2](#)).

337 **Statistical test**

338 All the statistical significance assessments were performed with two-sided test.

339 **Supplemental References**

- 340 Alneberg, J., Bjarnason, B.S., Bruijn, I. de, Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L.,  
341 Loman, N.J., Andersson, A.F., and Quince, C. (2014). Binning metagenomic contigs by  
342 coverage and composition. *Nat. Methods* *11*, 1144–1146.
- 343 Behrenfeld, M.J., and Falkowski, P.G. (1997). Photosynthetic rates derived from satellite-  
344 based chlorophyll concentration. *Limnol. Oceanogr.* *42*, 1–20.
- 345 Buesseler, K.O., Boyd, P.W., Black, E.E., and Siegel, D.A. (2020). Metrics that matter for  
346 assessing the ocean biological carbon pump. *Proc. Natl. Acad. Sci.* *117*, 9679–9687.
- 347 Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for  
348 automated alignment trimming in large-scale phylogenetic analyses. *Bioinforma. Oxf. Engl.*  
349 *25*, 1972–1973.
- 350 Carradec, Q., Pelletier, E., Silva, C.D., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-  
351 Mendez, G., Rocha, F., Tirichine, L., Labadie, K., et al. (2018). A global ocean atlas of  
352 eukaryotic genes. *Nat. Commun.* *9*, 373.
- 353 Clasen, J.L., and Suttle, C.A. (2009). Identification of freshwater Phycodnaviridae and their  
354 potential phytoplankton hosts, using DNA pol sequence fragments and a genetic-distance  
355 analysis. *Appl. Environ. Microbiol.* *75*, 991–997.
- 356 Coenen, A.R., and Weitz, J.S. (2018). Limitations of Correlation-Based Inference in Complex  
357 Virus-Microbe Communities. *MSystems* *3*, e00084-18.
- 358 Delmont, T.O., and Eren, A.M. (2016). Identifying contamination with advanced visualization  
359 and analysis practices: metagenomic approaches for eukaryotic genome assemblies. *PeerJ* *4*,  
360 e1839.
- 361 Delmont, T.O., Quince, C., Shaiber, A., Esen, Ö.C., Lee, S.T., Rappé, M.S., McLellan, S.L.,  
362 Lückner, S., and Eren, A.M. (2018). Nitrogen-fixing populations of Planctomycetes and  
363 Proteobacteria are abundant in surface ocean metagenomes. *Nat. Microbiol.* *3*, 804–813.
- 364 Eddy, S.R. (2011). Accelerated Profile HMM Searches. *PLOS Comput. Biol.* *7*, e1002195.
- 365 Eren, A.M., Esen, Ö.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., and Delmont,  
366 T.O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* *3*,  
367 e1319.
- 368 Guglielmini, J., Woo, A.C., Krupovic, M., Forterre, P., and Gaia, M. (2019). Diversification  
369 of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc.*  
370 *Natl. Acad. Sci.* *116*, 19585–19592.

371 Guidi, L., Jackson, G.A., Stemmann, L., Miquel, J.C., Picheral, M., and Gorsky, G. (2008).  
372 Relationship between particle size distribution and flux in the mesopelagic zone. *Deep Sea*  
373 *Res. Part Oceanogr. Res. Pap.* 55, 1364–1374.

374 Guidi, L., Stemmann, L., Jackson, G.A., Ibanez, F., Claustre, H., Legendre, L., Picheral, M.,  
375 and Gorsky, G. (2009). Effects of phytoplankton community on production, size, and export  
376 of large aggregates: A world-ocean analysis. *Limnol. Oceanogr.* 54, 1951–1963.

377 Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., Darzi, Y., Audic, S.,  
378 Berline, L., Brum, J.R., et al. (2016). Plankton networks driving carbon export in the  
379 oligotrophic ocean. *Nature* 532, 465.

380 Henry, N., de Vargas, C., Audic, S., Tara Oceans Consortium, C., and Tara Oceans  
381 Expedition, P. (2019). Total V9 rDNA information organized at the OTU level for the Tara  
382 Oceans Expedition (2009-2013), including the Tara Polar Circle Expedition (2013). Zenodo  
383 <https://doi.org/10.5281/zenodo.3768510>

384 Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, Analysis, and  
385 Visualization of Phylogenomic Data. *Mol. Biol. Evol.* 33, 1635–1638.

386 Ibarbalz, F.M., Henry, N., Brandão, M.C., Martini, S., Busseni, G., Byrne, H., Coelho, L.P.,  
387 Endo, H., Gasol, J.M., Gregory, A.C., et al. (2019). Global Trends in Marine Plankton  
388 Diversity across Kingdoms of Life. *Cell* 179, 1084-1097.e21.

389 Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version  
390 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780.

391 Kazlauskas, D., Varsani, A., and Krupovic, M. (2018). Pervasive Chimerism in the  
392 Replication-Associated Proteins of Uncultured Single-Stranded DNA Viruses. *Viruses* 10,  
393 187.

394 Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast  
395 single-node solution for large and complex metagenomics assembly via succinct de Bruijn  
396 graph. *Bioinforma. Oxf. Engl.* 31, 1674–1676.

397 Lin, J., Lee, Z., Ondrusek, M., and Kahru, M. (2016). Attenuation coefficient of usable solar  
398 radiation of the global oceans. *J. Geophys. Res. Oceans* 121, 3228–3236.

399 Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp,  
400 P., Goto, S., and Ogata, H. (2016). Linking Virus Genomes with Host Taxonomy. *Viruses* 8,  
401 66.

402 Picheral, M., Searson, S., Taillandier, V., Bricaud, A., Boss, E., Stemmann, L., Gorsky, G.,  
403 Tara Oceans Consortium, C., and Tara Oceans Expedition, P. (2014). Vertical profiles of  
404 environmental parameters measured from physical, optical and imaging sensors during station  
405 TARA\_080 of the Tara Oceans expedition 2009-2013. PANGAEA  
406 <https://doi.org/10.1594/PANGAEA.836419>

407 Price, M.N., Dehal, P.S., and Arkin, A.P. (2009). FastTree: Computing Large Minimum  
408 Evolution Trees with Profiles instead of a Distance Matrix. *Mol. Biol. Evol.* 26, 1641–1650.

409 Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G.,  
410 Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., et al. (2015). Ocean plankton. Structure  
411 and function of the global ocean microbiome. *Science* 348, 1261359.

412 Tackmann, J., Matias Rodrigues, J.F., and von Mering, C. (2019). Rapid Inference of Direct  
413 Interactions in Large-Scale Ecological Networks from Heterogeneous Microbial Sequencing  
414 Data. *Cell Syst.* 9, 286-296.e8.

415 Yutin, N., Wolf, Y.I., Raoult, D., and Koonin, E.V. (2009). Eukaryotic large nucleo-  
416 cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome  
417 evolution. *Virology* 6, 223.

418