Assessing the viral content of uncultured picoeukaryotes in the global-ocean by single cell genomics

Yaiza M. Castillo, Jean-François Mangot, Luiz Felipe Benites, Ramiro Logares, Megumi Kuronishi, Hiroyuki Ogata, Olivier Jaillon, Ramon Massana, Marta Sebastian, Dolors Vaque

Assessing the viral content of uncultured picoeukaryotes in the global-ocean by single cell genomics

Yaiza M. Castillo[1,*], Jean-François Mangot[1,*], L. Felipe Benites[2], Ramiro Logares[1], Megumi Kuronishi[3], Hiroyuki Ogata[3], Olivier Jaillon[4], Ramon Massana[1], Marta Sebastián[1,5] and Dolors Vaqué[1]

Author affiliations

[1] Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM), CSIC, Barcelona, Spain.

[2] Integrative Biology of Marine Organisms (BIOM), Sorbonne University, CNRS, Oceanological Observatory of Banyuls, Banyuls-sur-Mer, France.

[3] Bioinformatic Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Japan.

[4] Génomique Métabolique, Genoscope, Institut de biologie François Jacob, CEA, CNRS, Université d'Evry, Université Paris-Saclay, Evry, France.

[5] Institute of Oceanography and Global Change (IOCAG), University of Las Palmas de Gran Canaria, Telde, Spain.

[*] *These authors contributed equally to this work and thus joint first authors.*

Correspondence: Yaiza M. Castillo and Dolors Vaqué. Department of Marine Biology and Oceanography, Institut of Marine Sciences (CSIC), Passeig Marítim de la Barceloneta, 37-49, E-08003 Barcelona, Catalonia, Spain. E-mail: yaiza@icm.csic.es and dolors@icm.csic.es.

Running title: Viral signals in marine protists genomes

Abstract

Viruses are the most abundant biological entities on Earth and have fundamental ecological roles in controlling microbial communities. Yet, although their diversity is being increasingly explored, little is known about the extent of viral interactions with their protist hosts since most studies are limited to a few cultivated species. Here, we exploit the potential of single-cell genomics to unveil viral associations in 65 individual cells of 11 prevalent uncultured stramenopiles lineages sampled during the *Tara* Oceans expedition. We identified viral signals in 57% of the cells covering nearly every lineage and with narrow host specificity signal. Only 7 out of the 64 detected viruses displayed homologies to known viral sequences. A search for our viral sequences in global ocean metagenomes showed that they were preferentially found at the DCM and within the 0.2-3 µm size fraction. Some of the viral signals were widely distributed, while others geographically constrained. Among the viral signals we detected an endogenous mavirus virophage potentially integrated within the nuclear genome of two distant uncultured stramenopiles. Virophages have been previously reported as a cell's defense mechanism against other viruses, and may therefore play an important ecological role in regulating protist populations. Our results point to single-cell genomics as a powerful tool to investigate viral associations in uncultured protists, suggesting a wide distribution of these relationships, and providing new insights into the global viral diversity.

Keywords: Single-cell genomics; viral associations; protists; uncultured stramenopiles; viruses; virophages.

Introduction

Viruses are major players in marine biogeochemical cycles (Jover, Effler, Buchan, Wilhelm, & Weitz, 2014) and constitute the most abundant biological entities in the oceans, ranging from about 10^4 to 10^7 ml^{-1} (Danovaro et al., 2011; Suttle, 2005). They are known to be a major cause of microbial (bacteria, archaea and protists) mortalities (Munn, 2006), leading to approximatively 10^{29} infection events every day (Brussaard et al., 2008) and causing the release of 10^8-10^9 tons of biogenic carbon per day (Brussaard et al., 2008; Suttle, 2005). Furthermore, they are main vectors of gene transfer in the oceans (Middelboe & Brussaard, 2017), impacting microbial community dynamics, diversity and evolution (Breitbart, 2012; Jover et al., 2014; Weitz & Wilhelm, 2012).

Our knowledge of marine viral diversity and biogeography has been constantly expanding during this last decade with the advent of viral metagenomics (e.g., Coutinho et al., 2017; Mizuno, Rodriguez-Valera, Kimes, & Ghai, 2013; Paez-Espino et al., 2016). Multiple studies unveiled a large novel diversity of uncultured viruses, indicating their key roles in nutrient cycling and trophic networks (Brum et al., 2015; Roux et al., 2016). Unfortunately, despite these fruitful advances in viral ecology, our understanding of virus-host interactions is still in its infancy. The question of 'who infects whom' within marine microbial communities has always been central, and the assessment of the true extent of host specificity among marine viruses remains challenging (Brum & Sullivan, 2015). For a long time, studies investigating virus-host interactions were limited to cultured host cells, restricting our knowledge to the 0.1–1% of host cells that are in culture (Rappé & Giovannoni, 2003; Swan et al., 2013), and biasing our knowledge towards virulent lytic viruses (Brüssow & Hendrix, 2002; Swan et al.,

2013). Thus, many viruses are still uncharacterized and novel culture-independent approaches are needed to overcome these methodological limitations.

Several methods have been developed to investigate putative interactions between viruses and uncultured hosts reviewed by Brum & Sullivan (2015) and Breitbart, Bonnain, Malki, & Sawaya (2018). These include analyses by metaviromics (Bolduc, Wirth, Mazurie, & Young, 2015; Brum et al., 2015), matching CRISPR spacers (Anderson, Brazelton, & Baross, 2011; Berg Miller et al., 2012), phageFISH (Fluorescence *In Situ* Hybridization; Allers et al., 2013), viral tagging (Deng et al., 2012, 2014), the polony method (Baran, Goldin, Maidanik, & Lindell, 2018), the use of microfluidic digital PCR (Tadmor, Ottesen, Leadbetter, & Phillips, 2011) and single-cell genomics (SCG) (e.g., Labonté et al., 2015 and Roux et al., 2014). From these, SCG emerged as a powerful complement to cultivation and metagenomics by providing genomic information from individual uncultured cells (Stepanauskas, 2012). Furthermore, it has an incredible potential for cell-specific analyses of organismal interactions, such as parasitism, symbiosis and predation (Krabberød, Bjorbækmo, Shalchian-Tabrizi, & Logares, 2017; Stepanauskas, 2012), giving comprehensive insights of *in situ* virus-host associations. Indeed, this effective approach has revealed new associations between viruses and bacterial (Labonté et al., 2015; Roux et al., 2014) or archaeal cells (Chow, Winget, White, Hallam, & Suttle, 2015; Labonté et al., 2015; Munson-McGee et al., 2018). However, the application of SCG to protist cells is relatively recent and there is still a limited number of Single-cell Amplified Genomes (SAGs) from microeukaryotes (e.g., Bhattacharya et al., 2012; Heywood, Sieracki, Bellows, Poulton, & Stepanauskas, 2011; Mangot et al., 2017; Roy et al., 2015; Troell et al., 2016; Vannier et al., 2016), with only one study that has so far explored virus-host interactions (Yoon et al., 2011).

In the present work, we use SCG to uncover putative interactions between viruses and uncultured protists using 65 SAGs produced during the *Tara* Oceans expedition (Karsenti et al., 2011). These cells were affiliated to 11 stramenopile lineages belonging to MArine STramenopiles (MASTs), Chrysophyceae, Dictyochophyceae and Pelagophyceae, that are known to be important components of marine pico- and nanosized eukaryotic assemblages (1-5 µm, Massana, 2011). Initially detected in molecular diversity surveys, MASTs are formed by at least 18 independent groups of essentially uncultured protists (Massana, del Campo, Sieracki, Audic, & Logares, 2014), some of which display a widespread distribution in sequencing data sets (Lin et al., 2012; Logares et al., 2012; Seeleuthner et al., 2018) and are abundant in microscopy counts (Massana, Terrado, Forn, Lovejoy, & Pedrós-Alió, 2006).Within these MASTs, we analyzed SAGs from three clades, MAST-3, MAST-4 and MAST-7 (Massana et al., 2014). We also report the putative linkages between viruses and SAGs from the uncultured chrysophyte lineages G and H, formed by pigmented and colorless cells respectively, which are abundant in molecular diversity surveys (del Campo & Massana, 2011; Seeleuthner et al., 2018). Finally, we screened for viral signatures in SAGs from a cultured pelagophyte (*Pelagomonas calceolata*) and an uncultured dictyochophyte within the order Pedinellales.

Our results revealed a large diversity of viral sequences associated to protist cells, the vast majority of which correspond to previously unidentified viral lineages. Using global ocean metagenomes from the *TARA* Oceans expedition, we looked at the geographical distribution of the identified viral sequences in epipelagic waters by fragment recruitment analysis, finding that some SAG-associated viruses were widely distributed while others are restricted to certain areas. Finally, special attention was paid to a particular virophage sequence retrieved in two distinct stramenopile lineages that is highly similar to the endogenous

Cafeteriavirus-dependent mavirus, known to be integrated within the nuclear genome of their host (Fischer & Hackl, 2016). Overall, our approach constitutes an initial attempt to determine virus-host associations within protists using culture-independent single-cell genomics.

Materials and methods

Sample collection and single cell sorting

Samples for single-cell sorting were collected during the circumglobal *Tara* Oceans expedition (2009-2013) (Karsenti et al., 2011) and processed as described in Alberti et al., 2017. Flow cytometry cell sorting on cryopreserved samples and genomic DNA amplification by multiple displacement amplification (MDA) were performed at the Single Cell Genomics Center in the Bigelow Laboratory (https://scgc.bigelow.org). SAGs from phototrophic (plastidic) and heterotrophic (aplastidic) cells were screened by PCR using universal eukaryote DNA primers and taxonomically assigned. A total of 65 SAGs affiliated to 11 stramenopiles lineages (Table S1) were selected for sequencing. Sequence data is available at ENA (http://www.ebi.ac.uk/services/tara-oceans-data) under the accession codes listed in Table S2. Main sample-associated environmental data are reported in Table S3, and more details can be found in PANGAEA (Pesant et al., 2015).

SAG sequencing and assembly

After purification of MDA products, 101 bp paired-end libraries were prepared from each single cell as described in Alberti et al., 2017 and cells were independently sequenced on a 1⁄8th Illumina HiSeq lane at the Oregon Health & Science University (US) or at the National

Sequencing Center of Genoscope (France). Reads from SAGs were assembled using SPAdes 3.1 (Nurk et al., 2013). In all assemblies, contigs shorter than 500bp were discarded. Quality profiles and basic statistics (genome size, number of contigs, N50, GC content) of each SAG assemblies were generated with Quast (Gurevich, Saveliev, Vyahhi, & Tesler, 2013). Estimations of genome recovery were done with BUSCO (Benchmarking Universal Single-Copy Orthologs; Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015).

Detection and identification of viral signals in SAGs

Putative viral sequences were retrieved from each assembled SAG using VirSorter v1.0.3 (Roux, Enault, Hurwitz, & Sullivan, 2015) with default parameters and both the *RefSeqABVir* and *Virome* databases through the CyVerse Discovery Environment (Devisetty, Kennedy, Sarando, Merchant, & Lyons, 2016). Contigs identified by VirSorter at all three levels of confidence (from the more to the less confident predictions), categorized as viruses and prophages, were used in subsequent analyses. Sequence similarity between identified full length viral contigs was checked via pairwise BLASTn v2.2.28 (Altschul, Gish, Miller, Myers, & Lipman, 1990). Contigs with sequence similarity >95%, coverage >80% and e-value of $<10^{-5}$ were clustered together. Only one representative contig (i.e., the longest one) for each non-redundant SAG-associated viral sequence of each cluster (here after unique contig) was kept for further analysis.

Taxonomy of SAG-associated viral contigs was inferred using the webserver ViPTree (Nishimura et al., 2017). Proteomic trees of each unique contig were generated based on genome-wide sequence similarities computed by tBLASTx. A measure of genomic similarity based on a normalized bit score of tBLASTx (S_G) was calculated against a set of reference viral genomes database, the GenomeNet Virus–Host Database (Mihara et al., 2016). Since

MDA does not amplify RNA viruses, only ssDNA and dsDNA viruses and virophage/satellites were considered in that analysis. SAG-associated viruses showing a S_G >0.15 with a reference viral genome were assumed to belong to the same viral genus (Nishimura et al., 2017).

Finally, protein-coding genes of each unique SAG-associated viral sequences were predicted using Prodigal v2.6.3 (Hyatt et al., 2010) and annotated with BLASTp v2.7.1 (e-value 0.001, max. 10 hits) using the NCBI's nr database (updated 09 Feb 2019).

Biogeography of SAG-associated viral contigs assessed by fragment recruitment analysis.

The global distribution of each unique SAG-associated virus was estimated by fragment recruitment analysis against metagenomes from the Ocean Microbial Reference Gene Catalog (OM-RGC; Sunagawa et al., 2015) were estimated using an approach similar to Swan et al., 2013. A total of 128 metagenomes from two depths (surface and Deep Chlorophyll Maximum [DCM]) targeting both <0.22 µm (n = 48) and 0.22-3 µm (n= 80) size fractions were analyzed. Metagenomic reads were prior randomly subsampled without replacement to the minimum number of reads within each depth and size fraction using reformat.sh from bbtools suite (https://sourceforge.net/projects/bbmap/). BLAST+ v2.7.1 was then used to recruit reads from the OM-RCG database to each viral sequence (n=64) using default parameter values, except for: -perc_identity 70 -evalue 0.0001. The percentage of unique recruits (~100 bp long and ≥95% identity) from each metagenome matching to each viral sequence was normalized by viral sequence length. SAG-associated viral sequence abundances for each metagenome were calculated from the BLAST output and plotted using custom R scripts.

Identification of virophage contigs and detection/reconstruction of the mavirus integration site

The SAG-associated viral contig SV11, determined by ViPTree in the previous analyses, was highly similar to the virophage genome Maverick-related virus (also referred as mavirus, NC_015230, Fischer & Hackl, 2016), which share an evolutionary origin with a class of self-synthesizing DNA transposons called Maverick/Polinton elements (Fischer & Suttle, 2011). Mavirus was recently found integrated within the nuclear genome of the protist *Cafeteria roenbergensis* in multiple sites, where the endogenous virophage genome (named *Cafeteriavirus*-dependent mavirus and here referred as endogenous mavirus, KU052222) was flanked on either side by terminal inverted repeats (TIRs) (Fischer & Hackl, 2016). However, in comparison with the sequence of the endogenous mavirus, SV11 virophage sequence was partially incomplete. To determine if the incomplete SAG-associated virophages were potentially integrated within their respective host genome, we proceeded as follows. We first identified putative virophage contigs that could have not been detected with VirSorter because this automated tool was only applied on contigs >500bp, and requires a minimum of two predicted genes per contig to identify it as viral. Consequently, for each SAG containing an associated putative virophage sequence (within chrysophyte-G1 and MAST-3A), all contigs (including fragments <500 bp) were searched by a BLASTn analysis against a manually curated sequence of the endogenous mavirus including the TIRs sequences. For each SAG, contigs with a minimum similarity of 95% and maximal e-value of 10^{-4} with the curated endogenous mavirus genome were assumed to belong to the virophage genome and were aligned to the primarily detected virophage contig using ClustalW (Larkin et al., 2007) as implemented in the Geneious package v10.2.2 (Kearse et al., 2012). Then, to increase the completeness of the SAG-associated virophage genome, a fragment recruitment analysis was

performed using BLASTn against all identified virophage contigs in each SAG and reads with at least 99% identity and a maximal e-value of 10^{-4} were kept and assembled to the virophage genome using the Geneious *de novo* assembler with a minimum overlap of 50bp and a minimum identity of 95% (Kearse et al., 2012). Gene prediction of the obtained SAG-associated virophage assemblies was done using Prodigal v2.6.3 and annotated with BLASTp v2.2.28 (e-value 0.001, max. 10 hits) against NCBI's nr database (updated 06 Jun 2017).

Phylogenetic and comparative genomic analysis of the virophages

Phylogenetic analyses of the new SAG-associated virophages were performed with a set of reference virophage sequences from the literature. These include the virophage sequences isolated from cultures such as Sputnik (La Scola et al., 2008), Sputnik 2 (Desnues et al., 2012), Sputnik 3 (Gaia et al., 2013), Zamilon (Gaia et al., 2014) and mavirus (Fischer & Suttle, 2011), combined with sequences assembled from environmental surveys such as Yellowstone Lake (YLSV1-4 (Zhou et al., 2013) and YLSV5-7 (Zhou et al., 2015), Qinghai Lake (QLV, (Oh, Yoo, & Liu, 2016)), Dishui Lake (Dishui, (Gong et al., 2016)), Organic Lake (OLV, (Yau et al., 2011)), Ace Lake (ALM, (Zhou et al., 2013)), Trout Bog epilimnion and hypolimnion (TBE and TBH, Roux et al., 2017) and Mendota (Roux et al., 2017).

As proposed by Roux et al., 2017, phylogenetic trees were built based on a concatenated alignment using four core genes (major capsid protein [MCP], minor capsid protein [mCP], DNA packaging enzyme [ATPase], and cysteine protease [CysProt]) from all virophage genomes, except for the virophage TBE_1002136, which lacked the ATPase. For this last, only 3 genes were included in the multi-marker alignment. For each virophage core gene, individual alignments were generated with MAFFT v7.305b (L-INS-I algorithm, (Katoh & Standley, 2013)), automatically curated to remove all non-informative positions using trimAl

v1.2 (Capella-Gutierrez, Silla-Martinez, & Gabaldon, 2009) and evaluated for optimal amino acid substitution models using ProtTest v3.4.2 (Darriba, Taboada, Doallo, & Posada, 2011). The concatenation of the four core genes alignments was performed using a supermatrix approach with a custom python script (https://github.com/wrf/supermatrix/blob/master/add_taxa_to_align.py). Maximum-likelihood trees of each four individual core genes alignments and the concatenated alignment were constructed with RAxML v. 8.2.9 (Stamatakis, 2014) with 100 trees for both topology and rapid bootstrap analyses, and using the evolutionary models LG+I+G+F (ATPase, CysProt and mCP) and RtREV+I+G+F (MCP).Trees were generated using the ape (Paradis, Claude, & Strimmer, 2004) and ggtree packages (Yu, Smith, Zhu, Guan, & Lam, 2017) in R 3.5.1. (R Development Core Team, 2016), and rooted using QLV. To verify the topology of the trees, bayesian phylogenies on each alignment were also generated with MrBayes v3.2.6 (2,000,000 generations; Ronquist et al., 2012).

Finally, whole-genome synteny comparisons between chrysophyte-G1 and MAST-3A SAG-associated virophages and their closest published relatives (endogenous *Cafeteriavirus*-dependent mavirus and Ace Lake mavirus) were performed with EasyFig v.2.2.2 (Sullivan, Petty, & Beatson, 2011) using tBLASTx and filtering of small hits and annotations option. Since all chrysophyte-G1 SAG-associated virophage are highly similar (mean identity of 99%), only one representative sequence (i.e., longest assembly) per stramenopile lineage are displayed (AB233-L11 for chrysophyte-G1 and AA240-G22 for MAST-3A).

Results

Detection of viral contigs in protist SAGs

We used a total of 65 SAGs from photosynthetic and heterotrophic stramenopiles selected from four *Tara* Oceans stations located in the Mediterranean Sea and Indian Ocean (Table S3): 6 from two lineages of MAST-3 (clades A and F), 27 from three MAST-4 lineages (clades A, C and E), 6 from a lineage of MAST-7 (clade A), 15 from three lineages of Chrysophyceae (clades G1, H1 and H2), 4 from an uncultured clade of Dictyochophyceae and 7 affiliated to *Pelagomonas calceolata* (Table 1, Table S1). Using a relatively similar sequencing depth (mean of 4.99 ± 0.81 Gbp), assembly sizes were variable among the SAGs, averaging from 3.6 (± 2.8) Mbp in Dictyochophyceae to 11.0 (± 8.0) Mbp in MAST-3F (considering contigs >500 bp; Table 1). The variation in assembly completeness was also important, ranging on average from about 1% (in Pelagophyceae and Dictyochophyceae) to 10% (in MAST-3, MAST-4 and chrysophyte-G1; Table 1). Finally, the number of contigs assembled and their respective N50 also varied among SAGs and stramenopile lineages (Table 1).

We first investigated the presence of viral contigs in the 65 stramenopile SAGs assemblies, identifying a total of 79 putative viral sequences in 37 SAGs (~57%) distributed among most analyzed lineages, with the exception of *Pelagomonas calceolata* (Figure. 1a, Table S1). Only two lineages (MAST-4C and MAST-4E) showed less than half of their cells harboring viral contigs (Figure 1a, Table S1). Interestingly, a significant fraction of the SAG-associated viruses (~50%) was found in cells affiliated to chrysophytes (8, 24 and 12 viral contigs in chrysophyte-G1, -H1 and -H2, respectively; Figure 1a) with only one cell without any viral sequence detected out of the 15 analyzed cells (Fig 1a). Furthermore, the isolated chrysophyte cells were very rich in viral sequences, with up to 9 viral contigs retrieved in a single SAG of

chrysophyte-H1 (AA538_K19; Table S1). However, this was an exception, since in general we detected from 1 to 3 viral contigs per cell (Table S1).

We next explored the uniqueness of the detected viral contigs based on a pairwise comparison of their full-length sequences. Of the 79 viral sequences initially identified, we determined 64 non-redundant (i.e., unique) sequences (Table 2), ranging from 1 to 48.5 kbp in length (median = 5.7 kbp; Table 2). From the 64, 61 were associated to a single stramenopile lineage (~95%), and only 3 viral sequences (~5%) were either shared between two (SV11 and SV28) or four lineages (SV2) (Figure 1b): SV2 was found in MAST-4 (clades A and E), MAST-7 and chrysophyte-H1, whereas SV11 was detected in chrysophyte-G1 and MAST-3A, and SV28 in MAST-4 clades A and E (Table 2). With respect to the 61 viral contigs present in only one specific lineage, about 98% of them were reported in only one specific cell (Figure 1b), with the exceptional case of SV51 present in triplicate in the same chrysophyte-H1 cell (AA538_K19, Table S1). Only one viral contig (SV13) was found in two different chrysophyte-H2 cells (Figure 1b, Table 2).

Diversity and distribution of the SAG-associated viral sequences across the sunlit oceans

The 64 unique viral sequences were compared with a set of reference viral genomes (Mihara et al., 2016). On the basis of high genomic sequence similarity ($S_G > 0.15$), 7 of the 64 viruses identified in the SAGs could be putatively assigned to four different viral families. These viruses were two virophages (SV11, SV46), three viruses of *Phycodnaviridae* (SV27, SV35, SV50), one virus of *Myoviridae* (SV48) and one virus of *Podoviridae* (SV64; Table 2). Other viruses showed lower sequence similarities (n = 48; $S_G < 0.15$) or lacked detectable similarity by tBLASTn (n=9) to reference viral genomes, thus being uncertain for their classification at the genus level (Table 2). None of the assigned viral genomes were complete (or circular) but

one particular virus, SV11, which seemed nearly complete based on the similarity to a reference genome. This virus of 15.5 kbp in length was highly similar (98.3%, S_G = 0.96) to the Maverick-related virus genome (mavirus, GenBank accession number: NC_015230) and likely belong to the virophage genus of *Mavirus* (*Lavidaviridae*; Table 2). The remaining identified viral signals includes a set of short genome fragments (from 1 to 6.4 kbp) with intermediate genomic similarities (40-53%) to either an unclassified virophage (SV46 with YLV6), eukaryotic viruses (SV27 and SV50 with *Phycodnaviridae*), or phages (SV48 and SV64) (Table 2). We further predicted protein-coding genes in the 64 unique viral sequences. Of the total of 619 predicted genes (median = 6 predicted genes per SV; Table 2), about ~60% (n= 363) had a close relative in the NCBI's nr database and 103 genes were related to eukaryotic viral functions (Table S4).

In order to address the occurrence of these putative viruses in marine epipelagic waters, we performed a fragment recruitment analysis of the viral signals in the *Tara* Oceans OM-RGC database (Sunagawa et al., 2015). Our findings show that the viral contigs were found preferentially at the DCM, and at the 0.2-3 µm size fraction rather than in the <0.2 µm size fraction (Figure 2 and Figure S2). Regarding their geographic distribution, the SAG-associated viruses displayed some differences. On the one hand, some of them showed a cosmopolitan distribution with different degrees of occurrence. For example, some viral contigs (SV1 and SV2) show a high presence in all oceanic basins and in both size fractions, while others (e.g., SV34) were highly present in the 0.2-3 µm size fraction but absent from the <0.2 µm size fraction (Figure 2). On the other hand, other SAG-associated viruses appeared to be constrained to a lower number of oceanic basins, with some of them showing some biogeography preferences (e.g., SV16 and SV32), whereas others were restricted to few locations with a low presence (e.g., SV53, SV54 and SV55) (Figure 2 and Figure S2).

Genome reconstruction and phylogenetic analysis of SAG-associated virophages

We next focused on five SAG-associated viral contigs (the non-redundant SAG-associated viral contig SV11), retrieved from one MAST-3A (AB240-G22) and from four different chrysophyte-G1 cells (AB233-D06, AB233-L11, AB233-O05 and AB233-P23; Table 2 and Table S1), which were highly similar to mavirus (i.e., Maverick-related virus; Table 2), an endogenous virophage ("provirophage") integrated in the genome of *Cafeteria roenbergensis*. To the best of our knowledge, mavirus constitutes the only case of integration of a *Mavirus* virophage in a protist genome revealed to date by a culture-based approach. However, the virophage genomes identified in each SAG were incomplete compared to mavirus, noting the remarkable lack of two genes coding for an integrase and an helicase, as well as the TIRs, which indicate genome linearity and, therefore, a potential integration into the host genome (Fischer & Hackl, 2016; Roux et al., 2017). After the identification of the putative virophage contigs and a read recruitment analysis within each SAG, to increase the completeness of the SAG-associated virophage genomes (see methods section), we were able to reconstruct the entire SAG-associated virophage genomes of the five stramenopiles cells. This includes the presence of both DNA replication genes and TIRs on either side of all SAG-associated virophage genomes, confirming that SAG-associated mavirus genomes were linear and potentially inserted in the stramenopile host genomes. For the reassembly, from 5 (AB233-O05) to 14 contigs (AB240-G22), ranging from 0.2-0.3 to 7.2-15.5 kbp in length, were necessary to reconstruct the 5 SAG-associated virophage genomes.

To better assess the phylogenetic position of these newly identified SAG-associated mavirus genomes among the virophages, we established a concatenated marker tree using four virophage core genes (mCP, MCP, ATPase and CysProt; Fig 3), including all the available

343 virophage genomes retrieved from culture, metagenomes and the five new SAG-associated
344 virophages. We found that the newly identified virophage sequences form a clade among the
345 genus *Mavirus* together with the mavirus virophage but distinct from the Ace Lake mavirus
346 (ALM (Zhou et al., 2013), a partial *Mavirus* genome retrieved from an environmental
347 sequencing survey) (Figure 3). Similar phylogenetic placements were found when each core
348 gene was analyzed separately (Fig. S1).

349 Finally, we compared the general genome organization of the identified SAG-associated
350 mavirus in MAST-3A (SV11_AB240_G22) and chrysophyte-G1 (SV11_AB233_L11) and
351 their closest published relatives, the endogenous mavirus and Ace Lake mavirus. As expected
352 from the previous analysis, the two SAG-associated mavirus displayed remarkable sequence
353 similarity with the endogenous mavirus integrated within the nuclear genome of *Cafeteria*
354 *roenbergensis* and exhibited clear differences with the Ace Lake mavirus (Figure 4). The
355 main differences between the two SAG-associated mavirus and the endogenous mavirus are
356 the presence of an extra gene coding for an unknown function (gene 11, 71 amino acids) in
357 the two SAG-associated mavirus and the absence in SV11_AB233_L11 mavirus of the gene
358 20 (152 amino acids, unknown function) of the endogenous mavirus genome (Figure 4).
359 Interestingly, we also retrieved an exon structure of one adjacent host gene of unknown
360 function (gene 22, 177 amino acids) in the MAST-3A genome (Figure 4). Although this
361 putative host sequence is relatively short (~ 1kbp), we were able to observe a significant
362 difference in its overall GC content compared with the mavirus sequence (60% *vs* 35%,
363 respectively; Figure 4).

364 **Discussion**

In this study, we used SCG to characterize potential virus-protist interactions. With the exception of *Pelagomonas calceolata*, we found evidence of virus associations in almost all studied protist cells. Indeed, the relatively high frequency of viral associations with protists cells (~57%), retrieved from SAG assemblies with low genome recovery (<10%), suggests that viral association levels are much higher. Same observations were previously made for prokaryotic cells in marine environments (Labonté et al., 2015; Munson-McGee et al., 2018; Roux et al., 2014), implying that (nearly) all microbial cells are susceptible to be infected or to carry viruses. In the case of *Pelagomonas calceolata* cells, the lack of viral signals in their respective SAGs is probably due to the very low assembly coverage (0.5 ± 0.4%) rather than to the absence of any virus. We have not yet observed any significant correlation between genome completeness and the number of viral contigs among SAGs (Table 1). Similar findings were previously reported in bacterioplankton (Labonté et al., 2015), suggesting that the probability to detect viruses among SAGs is independent of the retrieved host genome assembly. The variation in SAG genome coverage may depend on intrinsic properties of selected cells, their DNA integrity, as well as multiple displacement amplification (MDA) biases (Pinard et al., 2006; Stepanauskas, 2012; Woyke et al., 2009). Several methods have been developed to improve genome recovery of uncultured cells such as using partial SAG assemblies to recruit metagenome reads and/or contigs (Saw et al., 2015), sorting multiple natural cells to perform a targeted metagenomic analysis (Cuvelier et al., 2010; Rinke et al., 2013; Vaulot et al., 2012) or co-assembling short reads from multiple SAGs (Mangot et al., 2017; Seeleuthner et al., 2018). However, the application of these approaches to characterize virus–host interactions will miss intraspecific genetic variability of both actors. More recently, several new MDA-like methods, such as WGA-X (Stepanauskas et al., 2017), TruePrime (Picher et al., 2016) or REPLI-g (Ahsanuddin et al., 2017) have been

developed for improving the genome recovery from single environmental cells (bacterial, archaeal and protists) and viral particles with high GC-content genomes. Compared with the conventional MDA, these amplification alternatives may provide a better genome recovery of microbial taxa, including some not amenable to standard MDA (Stepanauskas et al., 2017).

Using VirSorter (Roux et al., 2015) we were able to identify 64 unique viral contigs in 37 stramenopiles cells. We chose VirSorter over VirFinder (Ren, Ahlgren, Lu, Fuhrman, & Sun, 2017) because it has been shown that the later may misclassify eukaryotic sequences as viral. Some other approaches have been recently developed to retrieve viral signals from (meta-) genomic data, such as MARVEL (Amgarten, Braga, da Silva, & Setubal, 2018) and VirMiner (Zheng et al., 2019), but they have been developed to detect viral genomes in prokaryotes. From the 64 viral contigs retrieved in the protist cells, the narrow host range of these viruses was remarkable given that >95% of the detected viral sequences (n=61) were specific to one stramenopile lineage and just a few were shared between lineages (n=3). This is contrary to previous findings on prokaryotic SAGs showing that nearly 50% of the detected viral types were found in more than 2 lineages (Munson-McGee et al., 2018), suggesting that viruses infecting protists are likely more specialist than viruses infecting prokaryotes. Furthermore, while an important fraction (~54%) of cells with viral signals was associated to only one viral sequence, we also retrieved several putative co-infections among the remaining cells, with up to 7 unique (i.e., non-redundant) viral contigs in a single chrysophyte cell. Nonetheless, the risk of a putative accidental co-sorting of a free viral particle with a protist cell during the single-cell sorting process exists. To assess the risk of a possible “viral contamination”, we estimated the frequency of such events based on previous estimates made on prokaryotic cells (Labonté et al., 2015) by adapting the calculations to the cell size range of our studied

stramenopile cells (2-3 µm). We obtained that the frequency of free environmental viral particles present in the cells' shade was less than 1 in 5,000. This reinforces the view that the viruses detected in our study were truly and directly associated to the analyzed protist cells. These associations may consist on i) lytic and/or temperate (i.e. non-lytic) viruses adsorbed in the cell membrane, ii) a temperate virus or a virophage integrated into the host genome, iii) a virus replicating inside the cell, iv) a grazed prokaryote or protist carrying a temperate virus or with an active infection, or v) a predated free virus. A combination of these different scenarios probably explains the high number of viral sequences detected in the chrysophyte cells. For now, our current data set, including mostly fragments of viral sequences rather than complete viral genomes, does not allow us to decipher which mode of virus-host association prevail among the targeted protist cells.

Only 7 (~10%) viral contigs detected in protist cells were taxonomically assigned to known viruses (Table 2), which include some close hits to viruses belonging to the *Phycodnaviridae* family, known as a pathogen of marine eukaryotic algae (e.g., Brussaard, Short, Frederickson, & Suttle, 2004; Derelle et al., 2008), and others to bacteriophages and cyanophages. This suggests that a non-negligible part of the identified viral signals might come from putative infected (bacterial and/or picoeukaryotic) preys grazed by the stramenopiles. Indeed, the analyzed stramenopile lineages are mostly small free-living bacterivorous (Massana et al., 2006; Piwosz, Wiktor, Niemi, Tatarek, & Michel, 2013), with some groups (e.g., MAST-4) showing the ability to also eat picoalgae in grazing experiments (Massana et al., 2009). Nevertheless, previous studies working with a subset of our SAGs (Mangot et al., 2017; Seeleuthner et al., 2018) have shown that genes from bacteria and photosynthetic eukaryotes only represent a very small fraction of the genome assemblies (< 0.3% of fragmented contigs (Mangot et al., 2017)). A search for 16S rDNA genes in the SAGs where bacteriophages were

retrieved was unfruitful (data not shown), making difficult the association of these phages to putative grazed bacteria. It is also possible that some of the detected viral signals come from grazed viruses, since it is well-known that heterotrophic protists can graze on viruses (Fuhrman, 1999; González & Suttle, 1993). However, another plausible explanation for the identification of bacteriophages as closest hits to some SAGs associated virus is the overrepresentation of bacteriophage genomes compared to viruses in reference databases (Klingenberg, Aßhauer, Lingner, & Meinicke, 2013), which is supported by the low sequence similarity between the viral signals and the bacteriophages sequences (Table 2). Although taking all together it is difficult to elucidate which virus-host associations prevail among the targeted protist cells, the geographic distribution of the viral signals supports the view that the detected virus-protists associations reflect in many cases true interactions, because viral signals coming from MAST-4A, MAST-4C and chrysophyte-H1 were ubiquitous (e.g. SV1, SV7 and SV28), while those viral signals coming from MAST-4E, MAST-3A, MAST-3F and chrysophyte-H2 were geographically constrained (e.g. SV12, SV32 and SV54) (Figure 2), in agreement with the biogeography of these stramenopiles (Seeleuthner et al., 2018).

Some of the taxonomically assigned viral contigs were affiliated to known virophages and, more particularly, in the case of SV11 to *Lavidaviridae* (Krupovic, Kuhn, & Fischer, 2016). This virophage family, encompassing the two genera of *Mavirus* and *Sputnikvirus*, comprises obligate parasites of giant DNA viruses of the *Mimiviridae* family (Fischer & Hackl, 2016). Furthermore, virophages encode integrase genes, and provirophages have been reported in the nuclear genome of the marine alga *Bigelowiella natans* (Blanc, Gallot-Lavallée, & Maumus, 2015), and the protozoan *Cafeteria roenbergensis* (Fischer & Hackl, 2016). Provirophages putatively act as a host defense mechanism against giant viruses, in which some cells are sacrificed to protect their kin (Blanc et al., 2015; Fischer & Suttle, 2011). In this study, we

identified the presence of endogenous mavirus virophages in the assembly of five cells affiliated to chrysophyte-G1 and MAST-3A. These SAG-mavirus are highly similar to the *Cafeteriavirus*-dependent mavirus, a parasite of the giant *Cafeteria roenbergensis* virus (CroV) (Fischer, Allen, Wilson, & Suttle, 2010) integrated within the genome of *Cafeteria roenbergensis* (Fischer & Hackl, 2016). The presence of TIRs in the SV11_AB233_L11 and SV11_AB240_G22 virophages, as well as the exon structure of a putative adjacent host gene in the SV11_AB240_G22 sequence (Figure 4), suggests the putative integration of the mavirus in the host genome. This is also confirmed by the lack of any CroV signal in our assembly, whose presence is incompatible with a virophage in its lysogenic stage (Fischer & Hackl, 2016). This finding constitutes the first report of the presence of a putative provirophage isolated from environmental samples using SCG. Only slight differences were observed between the different provirophage genomes, located notably at genomic regions of low conservation (gene 11 in the two SAG-associated mavirus). Little is known about the importance of mavirus provirophage in protist populations as its study is limited to few cases. It is somewhat surprising that the same mavirus virophage was found in three phylogenetically distant lineages (chrysophyte-G1, MAST-3A and *C. roenbergensis*), pointing to a global and important ecological role of virophages in protist populations. Mavirus host cell recognition is carried out through specific receptor interactions, while Sputnik is done through phagocytosis of a composite of the virophage and the giant virus they parasitize (Duponchel & Fischer, 2019). Therefore, a possible explanation for finding mavirus in the different lineages, is that the capsid proteins are evolutionary conserved and have evolved independently of the giant virus infecting the host cell. On the contrary, although the host cells from Sputnik and Zamilon are phylogenetically closer, these virophages may have co-evolved with their corresponding giant virus. This hypothesis is supported by the finding

485 that Sputnik can infect the groups A, B and C of the Mimiviridae group while Zamilon is
486 unable to infect the group A (mimi- and mamavirus) (Gaia et al., 2014). Virophages are
487 repeatedly detected in genomic studies, with different gene content and abundance profiles,
488 likely suggesting that they occupy different ecological niches (Desnues & Raoult, 2012; Roux
489 et al., 2017; Yau et al., 2011). Although the role of virophages in protist populations is still
490 enigmatic, they may play a role in regulating the giant virus population dynamics and virus-
491 host interactions, influencing the ecosystem function and probably the whole microbial food
492 web in aquatic environments (Desnues & Raoult, 2012). Our findings provide new insights
493 into the potential importance of mavirus in the ecology of marine protists, and reinforce the
494 need for more studies to elucidate the role of these fascinating viruses in the environment.
495
496 In summary, this work shows the benefits of single-cell genomics to increase our
497 understanding of virus-host associations in natural protist communities. Although our
498 knowledge of the marine viral diversity is constantly expanding since the development of
499 metagenomics (Coutinho et al., 2017; Mizuno et al., 2013; Paez-Espino et al., 2016), it has
500 been estimated that the majority (63-93%) of viral sequences in marine metagenomes are not
501 represented in public databases (Hurwitz & Sullivan, 2013), emphasizing the need for further
502 isolation, characterization and sequencing of specific marine viruses (Middelboe & Brussaard,
503 2017). A minute fraction of protist viruses is annotated to date (~100 sequenced genomes
504 (~0.6% of all viral genomes) in NCBI Genome database (July 2018), explaining the majority
505 of unassigned viral sequences in our study. Thus, in addition to the ever-increasing
506 knowledge on viral diversity by metagenomic approaches, the incorporation of SAG analysis
507 will allow the specific matching of viruses and their hosts as well as to determine the host
508 range of individual viruses without cultivation. Our findings suggest that protist cells are

susceptible to interact with predominantly specialist viruses and hint to the potential importance of provirophages in protist populations.

Acknowledgments

This work was supported by the Spanish projects MEFISTO (CTM2013-43767-P, MINECO), ALLFLAGS (CTM2016-75083-R, MINECO) and INTERACTOMICS (CTM2015-69936-P, MINECO), and the EU project SINGEK (H2020-MSCA-ITN-2015-675752). YMC was supported by a FPI Spanish fellowship (BES-2014-067849). JFM was beneficiary of a Marie Curie Fellowship (PIEF-GA-2012-331190, EU). LFB was beneficiary of a Marie Curie Fellowship (H2020-MSCA-ITN-2015-675752, EU). RL was supported by a Ramón y Cajal fellowship (RYC-2013-12554, MINECO, Spain). HO was supported by JSPS/KAKENHI (No. 18H02279), and Scientific Research on Innovative Areas from the Ministry of Education, Culture, Science, Sports and Technology (MEXT) of Japan (Nos. 16H06429, 16K21723, 16H06437). OJ was supported by The French Government 'Investissement d'Avenir' programs Oceanomics (ANR-11-BTBR-0008) and FRANCE GENOMIQUE (ANR-10-INBS-09). MS was supported by a Viera y Clavijo contract funded by the ACIISI and the ULPGC. Computing resources were obtained through the MARBITS platform at the ICM-CSIC. We are grateful to Michael E. Sieracki, Patrick Wincker and Colomban de Vargas, members of the *Tara* Oceans consortium, who initiated and designed the sampling and sequencing experiments of protist SAGs. We thank Nigel Grimsley, Matthias Fischer and Simon Roux for their help on the early stage of the analysis of the SAG-associated virophage sequences. We are also grateful to Pablo Sánchez for his advices on computing the fragment recruitment analysis. We finally thank the *Tara* Oceans consortium, people, and sponsors who

supported the *Tara* Oceans Expedition (http://www.embl.de/tara-oceans/) for making the data accessible. This is the contribution number XXX of the *Tara* Oceans Expedition 2009-2013.

References

Ahsanuddin, S., Afshinnekoo, E., Gandara, J., Hakyemezoğlu, M., Bezdan, D., Minot, S., … Mason, C. E. (2017). Assessment of REPLI-g multiple displacement whole genome amplification (WGA) techniques for metagenomic applications. *Journal of Biomolecular Techniques*, *28*(1), 46–55. doi:10.7171/jbt.17-2801-008

Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., Ferrera, I., … Wincker, P. (2017). Viral to metazoan marine plankton nucleotide sequences from the *Tara* Oceans expedition. *Scientific Data*, *4*, 170093. doi:10.1038/sdata.2017.93

Allers, E., Moraru, C., Duhaime, M. B., Beneze, E., Solonenko, N., Barrero-Canosa, J., … Sullivan, M. B. (2013). Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses. *Environmental Microbiology*, *15*(8), 2306–2318. doi:10.1111/1462-2920.12100

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. doi:10.1016/S0022-2836(05)80360-2

Amgarten, D., Braga, L. P. P., da Silva, A. M., & Setubal, J. C. (2018). MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. *Frontiers in Genetics*, *9*, 304. doi:10.3389/fgene.2018.00304

Anderson, R. E., Brazelton, W. J., & Baross, J. A. (2011). Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. *FEMS Microbiology Ecology*, *77*(1), 120–133. doi:10.1111/j.1574-6941.2011.01090.x

Baran, N., Goldin, S., Maidanik, I., & Lindell, D. (2018). Quantification of diverse virus populations in the environment using the polony method. *Nature Microbiology*, *3*(1), 62–72. doi:10.1038/s41564-017-0045-y

Berg Miller, M. E., Yeoman, C. J., Chia, N., Tringe, S. G., Angly, F. E., Edwards, R. A., … White, B. A. (2012). Phage-bacteria relationships and CRISPR elements revealed by a metagenomic survey of the rumen microbiome. *Environmental Microbiology*, *14*(1), 207–227. doi:10.1111/j.1462-2920.2011.02593.x

Bhattacharya, D., Price, D. C., Yoon, H. S., Yang, E. C., Poulton, N. J., Andersen, R. A., & Das, S. P. (2012). Single cell genome analysis supports a link between phagotrophy and primary plastid endosymbiosis. *Scientific Reports*, *2*(1), 356. doi:10.1038/srep00356

Blanc, G., Gallot-Lavallée, L., & Maumus, F. (2015). Provirophages in the *Bigelowiella* genome bear testimony to past encounters with giant viruses. *Proceedings of the National Academy of Sciences*, *112*(38), E5318–E5326. doi:10.1073/pnas.1506469112

Bolduc, B., Wirth, J. F., Mazurie, A., & Young, M. J. (2015). Viral assemblage composition in Yellowstone acidic hot springs assessed by network analysis. *The ISME Journal*, *9*(10), 2162–2177. doi:10.1038/ismej.2015.28

Breitbart, M. (2012). Marine viruses: truth or dare. *Annual Review of Marine Science*, *4*(1), 425–448. doi:10.1146/annurev-marine-120709-142805

573 Breitbart, M., Bonnain, C., Malki, K., & Sawaya, N. A. (2018). Phage puppet masters of the
574 marine microbial realm. *Nature Microbiology*, *3*(7), 754–766. doi:10.1038/s41564-018-
575 0166-y
576 Brum, J. R., Ignacio-Espinoza, J. C., Roux, S., Doulcier, G., Acinas, S. G., Alberti, A., …
577 Sullivan, M. B. (2015). Patterns and ecological drivers of ocean viral communities.
578 *Science*, *348*(6237), 1261498–1261498. doi:10.1126/science.1261498
579 Brum, J. R., & Sullivan, M. B. (2015). Rising to the challenge: accelerated pace of discovery
580 transforms marine virology. *Nature Reviews Microbiology*, *13*(3), 147–159.
581 doi:10.1038/nrmicro3404
582 Brussaard, C P D, Short, S. M., Frederickson, C. M., & Suttle, C. A. (2004). Isolation and
583 phylogenetic analysis of novel viruses infecting the phytoplankton *Phaeocystis globosa*
584 (Prymnesiophyceae). *Applied and Environmental Microbiology*, *70*(6), 3700–3705.
585 doi:10.1128/AEM.70.6.3700-3705.2004
586 Brussaard, Corina P D, Wilhelm, S. W., Thingstad, F., Weinbauer, M. G., Bratbak, G.,
587 Heldal, M., … Wommack, K. E. (2008). Global-scale processes with a nanoscale drive:
588 the role of marine viruses. *The ISME Journal*, *2*(6), 575–578. doi:10.1038/ismej.2008.31
589 Brüssow, H., & Hendrix, R. W. (2002). Phage genomics: small is beautiful. *Cell*, *108*(1), 13–
590 16. doi:10.1016/S0092-8674(01)00637-7
591 Capella-Gutierrez, S., Silla-Martinez, J. M., & Gabaldon, T. (2009). trimAl: a tool for
592 automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*,
593 *25*(15), 1972–1973. doi:10.1093/bioinformatics/btp348
594 Chow, C. E. T., Winget, D. M., White, R. A., Hallam, S. J., & Suttle, C. A. (2015).
595 Combining genomic sequencing methods to explore viral diversity and reveal potential
596 virus-host interactions. *Frontiers in Microbiology*, *6*, 1–15.
597 doi:10.3389/fmicb.2015.00265
598 Coutinho, F. H., Silveira, C. B., Gregoracci, G. B., Thompson, C. C., Edwards, R. A.,
599 Brussaard, C. P. D., … Thompson, F. L. (2017). Marine viruses discovered via
600 metagenomics shed light on viral strategies throughout the oceans. *Nature*
601 *Communications*, *8*, 15955. doi:10.1038/ncomms15955
602 Cuvelier, M. L., Allen, A. E., Monier, A., McCrow, J. P., Messie, M., Tringe, S. G., …
603 Worden, A. Z. (2010). Targeted metagenomics and ecology of globally important
604 uncultured eukaryotic phytoplankton. *Proceedings of the National Academy of Sciences*,
605 *107*(33), 14679–14684. doi:10.1073/pnas.1001665107
606 Danovaro, R., Corinaldesi, C., Dell'Anno, A., Fuhrman, J. A., Middelburg, J. J., Noble, R. T.,
607 & Suttle, C. A. (2011). Marine viruses and global climate change. *FEMS Microbiology*
608 *Reviews*, *35*(6), 993–1034. doi:10.1111/j.1574-6976.2010.00258.x
609 Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2011). ProtTest 3: fast selection of
610 best-fit models of protein evolution. *Bioinformatics*, *27*(8), 1164–1165.
611 doi:10.1093/bioinformatics/btr088
612 del Campo, J., & Massana, R. (2011). Emerging diversity within Chrysophytes,
613 Choanoflagellates and Bicosoecids based on molecular surveys. *Protist*, *162*(3), 435–
614 448. doi:10.1016/j.protis.2010.10.003
615 Deng, L., Gregory, A., Yilmaz, S., Poulos, B. T., Hugenholtz, P., & Sullivan, M. B. (2012).
616 Contrasting life strategies of viruses that infect photo- and heterotrophic bacteria, as
617 revealed by viral tagging. *MBio*, *3*(6), e00373-12. doi:10.1128/mBio.00373-12
618 Deng, L., Ignacio-Espinoza, J. C., Gregory, A. C., Poulos, B. T., Weitz, J. S., Hugenholtz, P.,
619 & Sullivan, M. B. (2014). Viral tagging reveals discrete populations in *Synechococcus*
620 viral genome sequence space. *Nature*, *513*(7517), 242–245. doi:10.1038/nature13459

621 Derelle, E., Ferraz, C., Escande, M.-L., Eychenié, S., Cooke, R., Piganeau, G., … Grimsley,
622 N. (2008). Life-cycle and genome of OtV5, a large DNA virus of the pelagic marine
623 unicellular green alga *Ostreococcus tauri*. *PLoS ONE*, *3*(5), e2250.
624 doi:10.1371/journal.pone.0002250
625 Desnues, C., La Scola, B., Yutin, N., Fournous, G., Robert, C., Azza, S., … Raoult, D.
626 (2012). Provirophages and transpovirons as the diverse mobilome of giant viruses.
627 *Proceedings of the National Academy of Sciences*, *109*(44), 18078–18083.
628 doi:10.1073/pnas.1208835109
629 Desnues, C., & Raoult, D. (2012). Virophages question the existence of satellites. *Nature*
630 *Reviews Microbiology*, *10*(3), 234–234. doi:10.1038/nrmicro2676-c3
631 Devisetty, U. K., Kennedy, K., Sarando, P., Merchant, N., & Lyons, E. (2016). Bringing your
632 tools to CyVerse Discovery Environment using Docker. *F1000Research*, *5*, 1442.
633 doi:10.12688/f1000research.8935.1
634 Duponchel, S., & Fischer, M. G. (2019). Viva lavidaviruses! Five features of virophages that
635 parasitize giant DNA viruses. *PLOS Pathogens*, *15*(3), e1007592.
636 doi:10.1371/journal.ppat.1007592
637 Fischer, M. G., Allen, M. J., Wilson, W. H., & Suttle, C. A. (2010). Giant virus with a
638 remarkable complement of genes infects marine zooplankton. *Proceedings of the*
639 *National Academy of Sciences*, *107*(45), 19508–19513. doi:10.1073/pnas.1007615107
640 Fischer, M. G., & Hackl, T. (2016). Host genome integration and giant virus-induced
641 reactivation of the virophage mavirus. *Nature*, *540*(7632), 288–291.
642 doi:10.1038/nature20593
643 Fischer, M. G., & Suttle, C. A. (2011). A virophage at the origin of large DNA transposons.
644 *Science*, *332*(6026), 231–234. doi:10.1126/science.1199412
645 Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects.
646 *Nature*, *399*(6736), 541–548. doi:10.1038/21119
647 Gaia, M., Benamar, S., Boughalmi, M., Pagnier, I., Croce, O., Colson, P., … La Scola, B.
648 (2014). Zamilon, a novel virophage with Mimiviridae host specificity. *PLoS ONE*, *9*(4),
649 e94923. doi:10.1371/journal.pone.0094923
650 Gaia, M., Pagnier, I., Campocasso, A., Fournous, G., Raoult, D., & La Scola, B. (2013).
651 Broad spectrum of Mimiviridae virophage allows its isolation using a Mimivirus
652 reporter. *PLoS ONE*, *8*(4), e61912. doi:10.1371/journal.pone.0061912
653 Gong, C., Zhang, W., Zhou, X., Wang, H., Sun, G., Xiao, J., … Wang, Y. (2016). Novel
654 virophages discovered in a freshwater lake in China. *Frontiers in Microbiology*, *7*, 5.
655 doi:10.3389/fmicb.2016.00005
656 González, J., & Suttle, C. (1993). Grazing by marine nanofiagellates on viruses and virus-
657 sized particles: ingestion and digestion. *Marine Ecology Progress Series*, *94*, 1–10.
658 doi:10.3354/meps094001
659 Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool
660 for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075.
661 doi:10.1093/bioinformatics/btt086
662 Heywood, J. L., Sieracki, M. E., Bellows, W., Poulton, N. J., & Stepanauskas, R. (2011).
663 Capturing diversity of marine heterotrophic protists: one cell at a time. *The ISME*
664 *Journal*, *5*(4), 674–684. doi:10.1038/ismej.2010.155
665 Hurwitz, B. L., & Sullivan, M. B. (2013). The Pacific Ocean Virome (POV): a marine viral
666 metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS*
667 *ONE*, *8*(2), e57355. doi:10.1371/journal.pone.0057355
668 Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010).

669 Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC*
670 *Bioinformatics*, *11*(1), 119. doi:10.1186/1471-2105-11-119
671 Jover, L. F., Effler, T. C., Buchan, A., Wilhelm, S. W., & Weitz, J. S. (2014). The elemental
672 composition of virus particles: implications for marine biogeochemical cycles. *Nature*
673 *Reviews Microbiology*, *12*(7), 519–528. doi:10.1038/nrmicro3289
674 Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., De Vargas, C., Raes, J., … Wincker, P.
675 (2011). A holistic approach to marine eco-systems biology. *PLoS Biology*, *9*(10),
676 e1001177. doi:10.1371/journal.pbio.1001177
677 Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version
678 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4),
679 772–780. doi:10.1093/molbev/mst010
680 Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., … Drummond,
681 A. (2012). Geneious Basic: an integrated and extendable desktop software platform for
682 the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649.
683 doi:10.1093/bioinformatics/bts199
684 Klingenberg, H., Aßhauer, K. P., Lingner, T., & Meinicke, P. (2013). Protein signature-based
685 estimation of metagenomic abundances including all domains of life and viruses.
686 *Bioinformatics*, *29*(8), 973–980. doi:10.1093/bioinformatics/btt077
687 Krabberød, A., Bjorbækmo, M., Shalchian-Tabrizi, K., & Logares, R. (2017). Exploring the
688 oceanic microeukaryotic interactome with metaomics approaches. *Aquatic Microbial*
689 *Ecology*, *79*(1), 1–12. doi:10.3354/ame01811
690 Krupovic, M., Kuhn, J. H., & Fischer, M. G. (2016). A classification system for virophages
691 and satellite viruses. *Archives of Virology*, *161*(1), 233–247. doi:10.1007/s00705-015-
692 2622-9
693 La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., … Raoult, D.
694 (2008). The virophage as a unique parasite of the giant mimivirus. *Nature*, *455*(7209),
695 100–104. doi:10.1038/nature07218
696 Labonté, J. M., Swan, B. K., Poulos, B., Luo, H., Koren, S., Hallam, S. J., … Stepanauskas,
697 R. (2015). Single-cell genomics-based analysis of virus–host interactions in marine
698 surface bacterioplankton. *The ISME Journal*, *9*(11), 2386–2399.
699 doi:10.1038/ismej.2015.48
700 Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam,
701 H., … Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*,
702 *23*(21), 2947–2948. doi:10.1093/bioinformatics/btm404
703 Lin, Y.-C. C., Campbell, T., Chung, C.-C. C., Gong, G.-C. C., Chiang, K.-P. P., & Worden,
704 A. Z. (2012). Distribution patterns and phylogeny of marine stramenopiles in the North
705 Pacific Ocean. *Applied and Environmental Microbiology*, *78*(9), 3387–3399.
706 doi:10.1128/AEM.06952-11
707 Logares, R., Audic, S., Santini, S., Pernice, M. C., de Vargas, C., & Massana, R. (2012).
708 Diversity patterns and activity of uncultured marine heterotrophic flagellates unveiled
709 with pyrosequencing. *The ISME Journal*, *6*(10), 1823–1833. doi:10.1038/ismej.2012.36
710 Mangot, J.-F., Logares, R., Sánchez, P., Latorre, F., Seeleuthner, Y., Mondy, S., … Massana,
711 R. (2017). Accessing the genomic information of unculturable oceanic picoeukaryotes by
712 combining multiple single cells. *Scientific Reports*, *7*, 41498. doi:10.1038/srep41498
713 Massana, R. (2011). Eukaryotic picoplankton in surface oceans. *Annual Review of*
714 *Microbiology*, *65*(1), 91–110. doi:10.1146/annurev-micro-090110-102903
715 Massana, R., del Campo, J., Sieracki, M. E., Audic, S., & Logares, R. (2014). Exploring the
716 uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within

717 stramenopiles. *The ISME Journal*, *8*(4), 854–866. doi:10.1038/ismej.2013.204
718 Massana, R., Terrado, R., Forn, I., Lovejoy, C., & Pedrós-Alió, C. (2006). Distribution and
719 abundance of uncultured heterotrophic flagellates in the world oceans. *Environmental*
720 *Microbiology*, *8*(9), 1515–1522. doi:10.1111/j.1462-2920.2006.01042.x
721 Massana, R., Unrein, F., Rodríguez-Martínez, R., Forn, I., Lefort, T., Pinhassi, J., & Not, F.
722 (2009). Grazing rates and functional diversity of uncultured heterotrophic flagellates.
723 *The ISME Journal*, *3*(5), 588–596. doi:10.1038/ismej.2008.130
724 Middelboe, M., & Brussaard, C. (2017). Marine viruses: key players in marine ecosystems.
725 *Viruses*, *9*(10), 302. doi:10.3390/v9100302
726 Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., … Ogata,
727 H. (2016). Linking virus genomes with host taxonomy. *Viruses*, *8*(3), 66.
728 doi:10.3390/v8030066
729 Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E., & Ghai, R. (2013). Expanding the marine
730 virosphere using metagenomics. *PLoS Genetics*, *9*(12), e1003987.
731 doi:10.1371/journal.pgen.1003987
732 Munn, C. B. (2006). Viruses as pathogens of marine organisms—from bacteria to whales.
733 *Journal of the Marine Biological Association of the UK*, *86*(03), 453–467.
734 doi:10.1017/S002531540601335X
735 Munson-McGee, J. H., Peng, S., Dewerff, S., Stepanauskas, R., Whitaker, R. J., Weitz, J. S.,
736 & Young, M. J. (2018). A virus or more in (nearly) every cell: ubiquitous networks of
737 virus–host interactions in extreme environments. *The ISME Journal*, *12*(7), 1706–1714.
738 doi:10.1038/s41396-018-0071-7
739 Nishimura, Y., Watai, H., Honda, T., Mihara, T., Omae, K., Roux, S., … Yoshida, T. (2017).
740 Environmental viral genomes shed new light on virus-host interactions in the ocean.
741 *MSphere*, *2*(2). doi:10.1128/mSphere.00359-16
742 Nurk, S., Bankevich, A., Antipov, D., Gurevich, A. A., Korobeynikov, A., Lapidus, A., …
743 Pevzner, P. A. (2013). Assembling single-cell genomes and mini-metagenomes from
744 chimeric MDA products. *Journal of Computational Biology*, *20*(10), 714–737.
745 doi:10.1089/cmb.2013.0084
746 Oh, S., Yoo, D., & Liu, W.-T. (2016). Metagenomics reveals a novel virophage population in
747 a tibetan mountain lake. *Microbes and Environments*, *31*(2), 173–177.
748 doi:10.1264/jsme2.ME16003
749 Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M.,
750 Mikhailova, N., … Kyrpides, N. C. (2016). Uncovering Earth's virome. *Nature*,
751 *536*(7617), 425–430. doi:10.1038/nature19094
752 Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and
753 Evolution in R language. *Bioinformatics*, *20*(2), 289–290.
754 doi:10.1093/bioinformatics/btg412
755 Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., … Searson, S.
756 (2015). Open science resources for the discovery and analysis of *Tara* Oceans data.
757 *Scientific Data*, *2*, 150023. doi:10.1038/sdata.2015.23
758 Picher, Á. J., Budeus, B., Wafzig, O., Krüger, C., García-Gómez, S., Martínez-Jiménez, M. I.,
759 … Schneider, A. (2016). TruePrime is a novel method for whole-genome amplification
760 from single cells based on TthPrimPol. *Nature Communications*, *7*(1), 13296.
761 doi:10.1038/ncomms13296
762 Pinard, R., de Winter, A., Sarkis, G. J., Gerstein, M. B., Tartaro, K. R., Plant, R. N., …
763 Leamon, J. H. (2006). Assessment of whole genome amplification-induced bias through
764 high-throughput, massively parallel whole genome sequencing. *BMC Genomics*, *7*(1),

765 216. doi:10.1186/1471-2164-7-216
766 Piwosz, K., Wiktor, J. M., Niemi, A., Tatarek, A., & Michel, C. (2013). Mesoscale
767 distribution and functional diversity of picoeukaryotes in the first-year sea ice of the
768 Canadian Arctic. *The ISME Journal*, *7*(8), 1461–1471. doi:10.1038/ismej.2013.39
769 R Development Core Team. (2016). R: a language and environment for statistical computing.
770 Retrieved from https://www.r-project.org/
771 Rappé, M. S., & Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual Review*
772 *of Microbiology*, *57*(1), 369–394. doi:10.1146/annurev.micro.57.030502.090759
773 Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2017). VirFinder: a novel k-
774 mer based tool for identifying viral sequences from assembled metagenomic data.
775 *Microbiome*, *5*(1), 69. doi:10.1186/s40168-017-0283-5
776 Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., …
777 Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark
778 matter. *Nature*, *499*(7459), 431–437. doi:10.1038/nature12352
779 Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., …
780 Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and
781 model choice across a large model space. *Systematic Biology*, *61*(3), 539–542.
782 doi:10.1093/sysbio/sys029
783 Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., … Sullivan, M.
784 B. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant
785 ocean viruses. *Nature*, *537*(7622), 689–693. doi:10.1038/nature19366
786 Roux, S., Chan, L.-K., Egan, R., Malmstrom, R. R., McMahon, K. D., & Sullivan, M. B.
787 (2017). Ecogenomics of virophages and their giant virus hosts assessed through time
788 series metagenomics. *Nature Communications*, *8*(1), 858. doi:10.1038/s41467-017-
789 01086-2
790 Roux, S., Enault, F., Hurwitz, B. L., & Sullivan, M. B. (2015). VirSorter: mining viral signal
791 from microbial genomic data. *PeerJ*, *3*, e985. doi:10.7717/peerj.985
792 Roux, S., Hawley, A. K., Torres Beltran, M., Scofield, M., Schwientek, P., Stepanauskas, R.,
793 … Sullivan, M. B. (2014). Ecology and evolution of viruses infecting uncultivated
794 SUP05 bacteria as revealed by single-cell- and meta-genomics. *ELife*, *3*, e03125.
795 doi:10.7554/eLife.03125
796 Roy, R. S., Price, D. C., Schliep, A., Cai, G., Korobeynikov, A., Yoon, H. S., …
797 Bhattacharya, D. (2015). Single cell genome analysis of an uncultured heterotrophic
798 stramenopile. *Scientific Reports*, *4*(1), 4780. doi:10.1038/srep04780
799 Saw, J. H., Spang, A., Zaremba-Niedzwiedzka, K., Juzokaite, L., Dodsworth, J. A.,
800 Murugapiran, S. K., … Ettema, T. J. G. (2015). Exploring microbial dark matter to
801 resolve the deep archaeal ancestry of eukaryotes. *Philosophical Transactions of the*
802 *Royal Society of London. Series B, Biological Sciences*, *370*(1678), 20140328.
803 doi:10.1098/rstb.2014.0328
804 Seeleuthner, Y., Mondy, S., Lombard, V., Carradec, Q., Pelletier, E., Wessner, M., …
805 Wincker, P. (2018). Single-cell genomics of multiple uncultured stramenopiles reveals
806 underestimated functional diversity across oceans. *Nature Communications*, *9*(1), 310.
807 doi:10.1038/s41467-017-02235-3
808 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015).
809 BUSCO: assessing genome assembly and annotation completeness with single-copy
810 orthologs. *Bioinformatics*, *31*(19), 3210–3212. doi:10.1093/bioinformatics/btv351
811 Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis
812 of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313.

doi:10.1093/bioinformatics/btu033
Stepanauskas, R. (2012). Single cell genomics: an individual look at microbes. *Current Opinion in Microbiology*, *15*(5), 613–620. doi:10.1016/j.mib.2012.09.001
Stepanauskas, R., Fergusson, E. A., Brown, J., Poulton, N. J., Tupper, B., Labonté, J. M., … Lubys, A. (2017). Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nature Communications*, *8*(1), 84. doi:10.1038/s41467-017-00128-z
Sullivan, M. J., Petty, N. K., & Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics*, *27*(7), 1009–1010. doi:10.1093/bioinformatics/btr039
Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., … Velayoudon, D. (2015). Structure and function of the global ocean microbiome. *Science*, *348*(6237), 1261359–1261359. doi:10.1126/science.1261359
Suttle, C. A. (2005). Viruses in the sea. *Nature*, *437*(7057), 356–361. doi:10.1038/nature04160
Swan, B. K., Tupper, B., Sczyrba, A., Lauro, F. M., Martinez-Garcia, M., Gonzalez, J. M., … Stepanauskas, R. (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proceedings of the National Academy of Sciences*, *110*(28), 11463–11468. doi:10.1073/pnas.1304246110
Tadmor, A. D., Ottesen, E. A., Leadbetter, J. R., & Phillips, R. (2011). Probing individual environmental bacteria for viruses by using microfluidic digital PCR. *Science*, *333*(6038), 58–62. doi:10.1126/science.1200758
Troell, K., Hallström, B., Divne, A.-M., Alsmark, C., Arrighi, R., Huss, M., … Bertilsson, S. (2016). *Cryptosporidium* as a testbed for single cell genome characterization of unicellular eukaryotes. *BMC Genomics*, *17*(1), 471. doi:10.1186/s12864-016-2815-y
Vannier, T., Leconte, J., Seeleuthner, Y., Mondy, S., Pelletier, E., Aury, J.-M., … Jaillon, O. (2016). Survey of the green picoalga *Bathycoccus* genomes in the global ocean. *Scientific Reports*, *6*(1), 37900. doi:10.1038/srep37900
Vaulot, D., Lepère, C., Toulza, E., De la Iglesia, R., Poulain, J., Gaboyer, F., … Piganeau, G. (2012). Metagenomes of the picoalga *Bathycoccus* from the Chile coastal upwelling. *PLoS ONE*, *7*(6), e39648. doi:10.1371/journal.pone.0039648
Weitz, J., & Wilhelm, S. (2012). Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 Biology Reports*, *4*, 17. doi:10.3410/B4-17
Woyke, T., Xie, G., Copeland, A., González, J. M., Han, C., Kiss, H., … Stepanauskas, R. (2009). Assembling the marine metagenome, one cell at a time. *PLoS ONE*, *4*(4), e5299. doi:10.1371/journal.pone.0005299
Yau, S., Lauro, F. M., DeMaere, M. Z., Brown, M. V., Thomas, T., Raftery, M. J., … Cavicchioli, R. (2011). Virophage control of antarctic algal host-virus dynamics. *Proceedings of the National Academy of Sciences*, *108*(15), 6163–6168. doi:10.1073/pnas.1018221108
Yoon, H. S., Price, D. C., Stepanauskas, R., Rajah, V. D., Sieracki, M. E., Wilson, W. H., … Bhattacharya, D. (2011). Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science*, *332*(6030), 714–717. doi:10.1126/science.1203163
Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36. doi:10.1111/2041-210X.12628
Zheng, T., Li, J., Ni, Y., Kang, K., Misiakou, M.-A., Imamovic, L., … Panagiotou, G. (2019). Mining, analyzing, and integrating viral signals from metagenomic data. *Microbiome*,

7(1), 42. doi:10.1186/s40168-019-0657-y
Zhou, J., Sun, D., Childers, A., McDermott, T. R., Wang, Y., & Liles, M. R. (2015). Three novel virophage genomes discovered from Yellowstone Lake metagenomes. *Journal of Virology*, *89*(2), 1278–1285. doi:10.1128/JVI.03039-14
Zhou, J., Zhang, W., Yan, S., Xiao, J., Zhang, Y., Li, B., … Wang, Y. (2013). Diversity of virophages in metagenomic data sets. *Journal of Virology*, *87*(8), 4225–4236. doi:10.1128/JVI.03398-12

Data accessibility

SAG sequence data are available at ENA under the accession codes listed in Table S2.

Author contribution

DV and RM conceived the study. OJ sequenced the SAGs, RL performed the SAG assemblies, and YMC and JFM performed the SAG bioinformatic and data analysis. MS contributed to the analytic design and LFB, HO and MK contributed with additional data analyses. YMC, JFM, MS, DV and RM interpreted the results, and YMC and JFM wrote the manuscript with inputs from all co-authors. YMC and JFM should be considered joint first author.

ORCID

Yaiza M. Castillo https://orcid.org/0000-0002-1319-2975

Supporting information

Additional supporting information may be found in the online version of this article.

FIGURE LEGEND

FIGURE 1 Occurrence and specificity of viral contigs in 65 marine stramenopiles SAGs. (**a**) Barplots show the number of SAGs with or without any viral contig detected in their assembly. For each lineage, the total number of SAG-associated viral contigs retrieved in

SAGs are indicated on top of each bar. (**b**) Pie charts display the percentage of viral contigs present in only one or shared among 2 or 4 lineages (upper left corner), and the percentage of SAGs that shared a viral contig for those that were lineage-specific (lower right corner). Chryso-G1, Chryso-H1, Chryso-H2, Dictyo and Pelago correspond to the chrysophyte clades G1, H1 and H2, Dictyochophyceae and *Pelagomonas calceolata*, respectively.

FIGURE 2 Biogeographical distribution of SAG-associated viruses, as determined by metagenomic fragment recruitment. Viral contigs are shown in the *y*-axis and epipelagic metagenome stations along the *x*-axis. The scale bar indicates the percentage of read sequences recruited normalized of aligned metagenome sequences with alignments ≥100 bp long and ≥95% identity, normalized by the length of each SAG-associated virus sequence. Results for metagenomes from the <0.2 µm (left panels) and 0.2-3 µm size fractions (right panels) were displayed for both surface (upper panels) and DCM stations (lower panels). Stations where metagenomes were not available are shown in grey (No Data). Color bars represent the different oceanic basins, abbreviations: North Pacific Ocean (NP), South Pacific Ocean (SP), North Atlantic Ocean (NA), South Atlantic Ocean (SA), Southern Ocean (SO), Mediterranean Sea (M), Read Sea (RS) and Indian Ocean (IND).

FIGURE 3 Phylogenetic placement of the new putative SAG-associated mavirus among virophages. The tree topology was inferred from a maximum-likelihood analysis of a concatenated alignment of four core genes (minor [mCP] and major [MCP] capsids proteins, DNA packaging enzyme [ATPase] and Cysteine Protease [CysProt]). Bayesian posterior probabilities (BPP) and bootstrap percentages (BS) are provided at each node (BPP/BS) when support values were higher than 0.7 and 70%, respectively. Black dots indicate maximal support for both posterior probabilities (1.0) and maximum-likelihood bootstraps (100%) at the respective nodes. The five new SAG-associated virophages are highlighted in bold. The origin (culture, metagenome sequencing or SAG) and genome type (linear with TIRs, circular or partial) of each virophage genome are pointed out in the tree. Abbreviated names for virophages are detailed in the Materials and Methods section.

FIGURE 4 Comparison of the SAG-associated mavirus genomes and their closest known relatives. Linear genomic maps show synteny between the mavirus genomes found in chrysophyte-G1 and MAST-3A SAGs (SV11_AB233_L11 and SV11_AB240_G22, respectively) and their closest published relatives, endogenous mavirus and ALM (Zhou et al., 2013). When present, TIRs and exon structures of putative adjacent host genes are displayed to highlight the putative integration of mavirus genomes within their respective host genomes. The main differences between the two SAG-associated mavirus and the endogenous mavirus are indicated with asterisks (†: presence of an extra coding gene [gene 11], ‡: absence of coding gene [gene 20]). Additionally, a GC content plot based on a 100 bp sliding window is shown for SV11_AB240_G22.

TABLE 1 General characteristics (mean (± standard error)) of the 65 draft stramenopiles SAGs obtained by single-cell genomics

Group	Name	Number of cells	Sequencing depth (Gbp)	Assembly size (Mbp)	Total number of contigs	BUSCO completeness (%)	GC content (%)	N50 (kbp)
Chrysophyceae	Chrysophyte-G1	4	5.5 (± 0.5)	9.3 (± 5.2)	3,597 (± 2,009)	11.6 (± 8.2)	40.2 (± 0.2)	5.1 (± 0.4)
	Chrysophyte-H1[†]	8	4 (± 0.7)	4.0 (± 2.1)	1,425 (± 518)	6.1 (± 3.4)	45.1 (± 0.8)	8.6 (± 3.4)
	Chrysophyte-H2[†]	3	4.2 (± 1.0)	4.3 (± 2.4)	1,928 (± 1,073)	3.3 (± 1.9)	47.7 (± 1.7)	4.1 (± 0.2)
Dictyochophyceae	unc. dictyochophyte	4	4.6 (± 0.1)	3.6 (± 2.8)	15,567 (± 948)	1.0 (± 1.0)	46.8 (± 2.5)	4.2 (± 1.4)
MAST-3	MAST-3A	4	5.1 (± 0.5)	7.5 (± 1.9)	2,272 (± 409)	11.4 (± 3.4)	42.5 (± 0.3)	8.6 (± 1.1)
	MAST-3F	2	5.4 (± 1.1)	11.0 (± 8.0)	3,576 (± 2,414)	11.7 (± 9.8)	34.1 (± 0.3)	7.7 (± 0.2)
MAST-4	MAST-4A	14	5 (± 1.8)	10.2 (± 4.9)	3,195 (± 1,393)	12.6 (± 7.5)	33.0 (± 1.0)	9.3 (± 2.7)
	MAST-4C	4	5.4 (± 0.6)	8.3 (± 2.3)	2,389 (± 579)	11.8 (± 3.8)	40.3 (± 0.2)	14.0 (± 1.3)
	MAST-4E	9	4.7 (± 0.8)	6.7 (± 2.6)	1,928 (± 607)	8.6 (± 3.9)	44.0 (± 0.7)	9.3 (± 1.9)
MAST-7	MAST-7A	6	5.5 (± 1.3)	5.6 (± 3.2)	2,002 (± 1,233)	3.8 (± 2.1)	44.7 (± 4.8)	7.0 (± 3.2)
Pelagophyceae	*Pelagomonas calceolata*	7	5.6 (± 0.5)	8.1 (± 0.8)	271 (± 186)	0.5 (± 1.0)	47.5 (± 7.6)	13.0 ±

Abbreviations: SAG, Single Amplified Genome; MAST, Marine Stramenopiles; unc., uncultured; BUSCO, Benchmarking Universal Single-Copy Orthologs; N50, length of the shortest contig from the minimal set of contig representing 50% of the assembly size.

[†] The 18S rRNA genes of these chrysophytes-H SAGs clustered into two distinct lineages (clades -H1 and -H2).

TABLE 2 Summary and taxonomic assigment of the 64 SAG-associated viral contigs

SAG-associated viral contig[†]				Taxonomic assigment on the GenomeNet Virus–Host Database (Mihara et al., 2016)				
Viral contig	*SAG lineage (number of SAGs)*	*Sequence length (kbp)*	*Number of genes*	*Best viral group hit (GenBank accession number)*	*Viral family*	*Known host group*	*SG*[‡]	*Similarity (%)*
SV1	MAST-4A (1)	48.5	44	*Cellulophaga* phage (KC821612)	*Podoviridae*	Bacteroidetes	0.06	40.8
SV2	MAST-4A (1), MAST-4E (1), MAST-7 (4), Chryso-H1 (1)	22.8	48	*Prochlorococcus* phage (NC_006883)	*Myoviridae*	Cyanobacteria	0.07	43.2
SV3	MAST-4A (1)	22.1	25	YSLV5 (NC_028269)	Unclassified virophage	*N/D*	< 0.01	31.9
SV4	Chryso-H2 (1)	21.2	25	*Synechococcus* phage (NC_026928)	*Myoviridae*	Cyanobacteria	< 0.01	42.0
SV5	Chryso-H1 (1)	20.5	20	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.04	42.4
SV6	MAST-3A (1)	18.9	19	*Paramecium bursaria Chlorella* virus (NC_009898)	*Phycodnaviridae*	Ciliophora	0.04	39.2
SV7	Chryso-H1 (1)	17.9	32	*Pseudomonas* phage (NC_028980)	*Siphoviridae*	Gammaproteobacteria	0.06	58.7
SV8	MAST-4E (1)	16.7	19	*Phaeocystis globosa* virus virophage (NC_021333)	Unclassified virophage	Haptophyta	0.01	42.3
SV9	Chryso-H1 (1)	16.7	22	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.07	41.5
SV10	Chryso-H1 (1)	16.2	19	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.08	40.0
SV11	**Chryso-G1 (4), MAST-3A (1)**	15.5	**18**	**Maverick-related virus (mavirus, NC_015230)**	***Lavidaviridae***	**Bicosoecophyceae**	**0.96**	**98.3**
SV12	MAST-4C (1)	15.0	14	*Rhodothermus* phage (NC_004735)	*Myoviridae*	Bacteroidetes	0.02	39.6
SV13	Chryso-H2 (2)	14.3	11	*Chrysochromulina ericina* virus (NC_028094)	*Phycodnaviridae*	Haptophyta	0.04	66.6
SV14	MAST-4A (1)	13.1	14	-	-	-	-	-
SV15	MAST-4A (1)	12.7	11	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.02	37.6
SV16	MAST-3A (1)	12.6	18	Yellowstone lake phycodnavirus (NC_028110)	*Phycodnaviridae*	*N/D*	0.09	52.6
SV17	Chryso-G1 (1)	10.2	4	*Mycobacterium* phage (NC_028662)	*Podoviridae*	Actinobacteria	0.03	38.5
SV18	Chryso-H1 (1)	10.0	10	*Bacillus* phage (NC_006945)	*Tectiviridae*	Firmicutes	< 0.01	36.1
SV19	MAST-4E (1)	9.9	5	*Vibrio* phage (NC_021529)	*Myoviridae*	Gammaproteobacteria	0.06	49.0
SV20	MAST-7 (1)	9.8	8	*Cronobacter* phage (NC_019398)	*Myoviridae*	Gammaproteobacteria	0.02	51.2
SV21	Chryso-H1 (1)	8.6	8	*Phaeocystis globosa* virus virophage (NC_021333)	Unclassified virophage	Haptophyta	0.02	35.3
SV22	Chryso-G1 (1)	7.5	11	*Synechococcus* phage (NC_015286)	*Myoviridae*	Cyanobacteria	0.05	49.7
SV23	MAST-7 (1)	7.4	7	*Acanthocystis turfacea Chlorella* virus (NC_008724)	*Phycodnaviridae*	Chlorophyta	0.08	54.9
SV24	Chryso-H1 (1)	7.2	5	-	-	-	-	-
SV25	Dictyo (1)	6.8	8	*Pseudomonas* phage (NC_026600)	*Myoviridae*	Gammaproteobacteria	0.01	43.1
SV26	Chryso-H1 (1)	6.7	9	-	-	-	-	-
SV27	**Chryso-H2 (1)**	6.4	**8**	**Aureococcus anophagefferens virus (NC_024697)**	***Phycodnaviridae***	**Pelagophyceae**	**1.0**	**52.4**
SV28	MAST-4A (2), MAST-4E (2)	5.9	6	*Cellulophaga* phage (KC821612)	*Podoviridae*	Bacteroidetes	0.07	44.2
SV29	Chryso-H1 (1)	5.8	7	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.14	41.0
SV30	Dictyo (1)	5.8	8	*Ostreococcus tauri* virus (NC_010191)	*Phycodnaviridae*	Chlorophyta	0.02	42.9
SV31	Chryso-H1 (1)	5.8	10	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.08	40.9
SV32	MAST-3A (1)	5.7	6	*Phaeocystis globosa* virus (NC_021312)	*Phycodnaviridae*	Haptophyta	0.07	45.4
SV33	Chryso-G1 (1)	5.6	3	*Anomala cuprea* entomopoxvirus (NC_023426)	*Poxviridae*	Arthropoda	0.1	41.5
SV34	MAST-4A (1)	5.6	2	*Aureococcus anophagefferens* virus (NC_024697)	*Phycodnaviridae*	Pelagophyceae	< 0.01	39.1
SV35	**MAST-3F (1)**	5.4	**7**	**Yellowstone lake phycodnavirus (NC_028110)**	***Phycodnaviridae***	***N/D***	**0.3**	**52.9**
SV36	Chryso-H1 (1)	5.2	5	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.07	42.5
SV37	Chryso-H1 (1)	5.1	10	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.08	39.8
SV38	Chryso-H2 (1)	4.7	7	*Enterobacteria* phage (NC_005066)	*Myoviridae*	Gammaproteobacteria	0.02	39.7
SV39	Chryso-H1 (1)	4.6	3	*Phaeocystis globosa* virus virophage (NC_021333)	Unclassified virophage	Haptophyta	0.04	35.0
SV40	Dictyo (1)	4.6	4	*Enterobacteria* phage (NC_019526)	*Myoviridae*	Gammaproteobacteria	0.13	44.6
SV41	Chryso-H2 (1)	4.4	7	YSLV5 (NC_028269)	Unclassified virophage	*N/D*	0.04	41.3
SV42	MAST-3A (1)	4.3	6	*Campylobacter* phage (NC_027997)	*Myoviridae*	Epsilonproteobacteria	0.02	30.3
SV43	Chryso-H1 (1)	4.1	6	YSLV7 (NC_028257)	Unclassified virophage	*N/D*	0.04	41.1
SV44	MAST-3A (1)	4.1	5	*Aureococcus anophagefferens* virus (NC_024697)	*Phycodnaviridae*	Pelagophyceae	0.03	32.8
SV45	MAST-4A (1)	3.7	3	*Erwinia* phage (HQ728263)	*Myoviridae*	Gammaproteobacteria	0.02	40.4
SV46	**Chryso-H1 (1)**	3.7	**3**	**YSLV6 (NC_028270)**	**Unclassified virophage**	***N/D***	**0.17**	**40.3**
SV47	Chryso-H1 (1)	3.1	5	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.09	42.3
SV48	**Chryso-G1 (1)**	3.0	**2**	***Escherichia* phage (NC_025447)**	***Myoviridae***	**Gammaproteobacteria**	**0.2**	**41.4**
SV49	Chryso-H2 (1)	3.0	4	-	-	-	-	-
SV50	**MAST-7 (1)**	2.9	**3**	***Ectocarpus siliculosus* virus (NC_002687)**	***Phycodnaviridae***	**Phaeophyceae**	**0.2**	**46.1**
SV51	Chryso-H1 (1)	2.9	5	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.1	46.5
SV52	Chryso-H1 (1)	2.9	4	-	-	-	-	-
SV53	Chryso-H2 (1)	2.8	4	-	-	-	-	-
SV54	Chryso-H2 (1)	2.8	4	-	-	-	-	-
SV55	Chryso-H2 (1)	2.7	4	-	-	-	-	-
SV56	Chryso-H1 (1)	2.5	4	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.04	42.3
SV57	MAST-4A (1)	2.5	4	*Synechococcus* phage (NC_015289)	*Myoviridae*	Cyanobacteria	0.03	31.1
SV58	MAST-7 (1)	2.3	3	*Enterobacteria* phage (NC_012740)	*Myoviridae*	Gammaproteobacteria	0.05	44.9
SV59	Chryso-H1 (1)	2.3	4	YSLV7 (NC_028257)	Unclassified virophage	*N/D*	0.09	43.3
SV60	Chryso-H1 (1)	2.1	4	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.06	43.8
SV61	MAST-4A (1)	2.0	4	-	-	-	-	-
SV62	Chryso-H2 (1)	2.0	4	*Microcystis* phage (NC_029002)	*Myoviridae*	Cyanobacteria	0.04	40.4
SV63	Chryso-H1 (1)	1.8	4	YSLV5 (NC_028269)	Unclassified virophage	*N/D*	0.1	47.0
SV64	**MAST-4A (1)**	1.0	**3**	***Planktothrix* phage (NC_016564)**	***Podoviridae***	**Cyanobacteria**	**0.2**	**49.1**

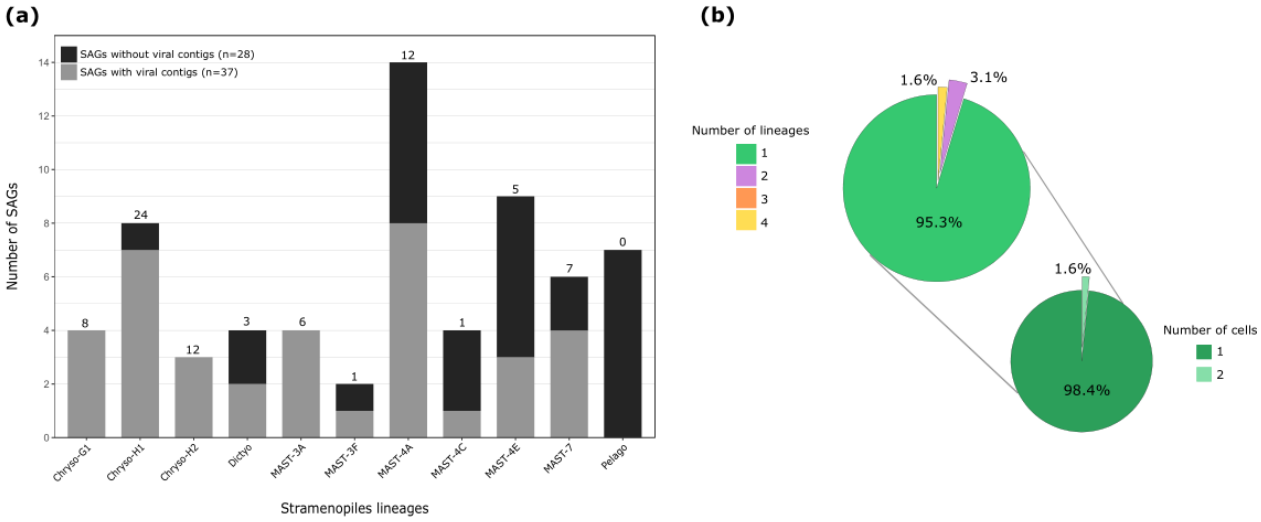
Abbreviations: SAG, Single Amplified Genome; MAST, Marine Stramenopiles; Chryso, Chrysophyte; Dictyo, Dictyochophyceae; YSLV, Yellowstone Lake virophage.

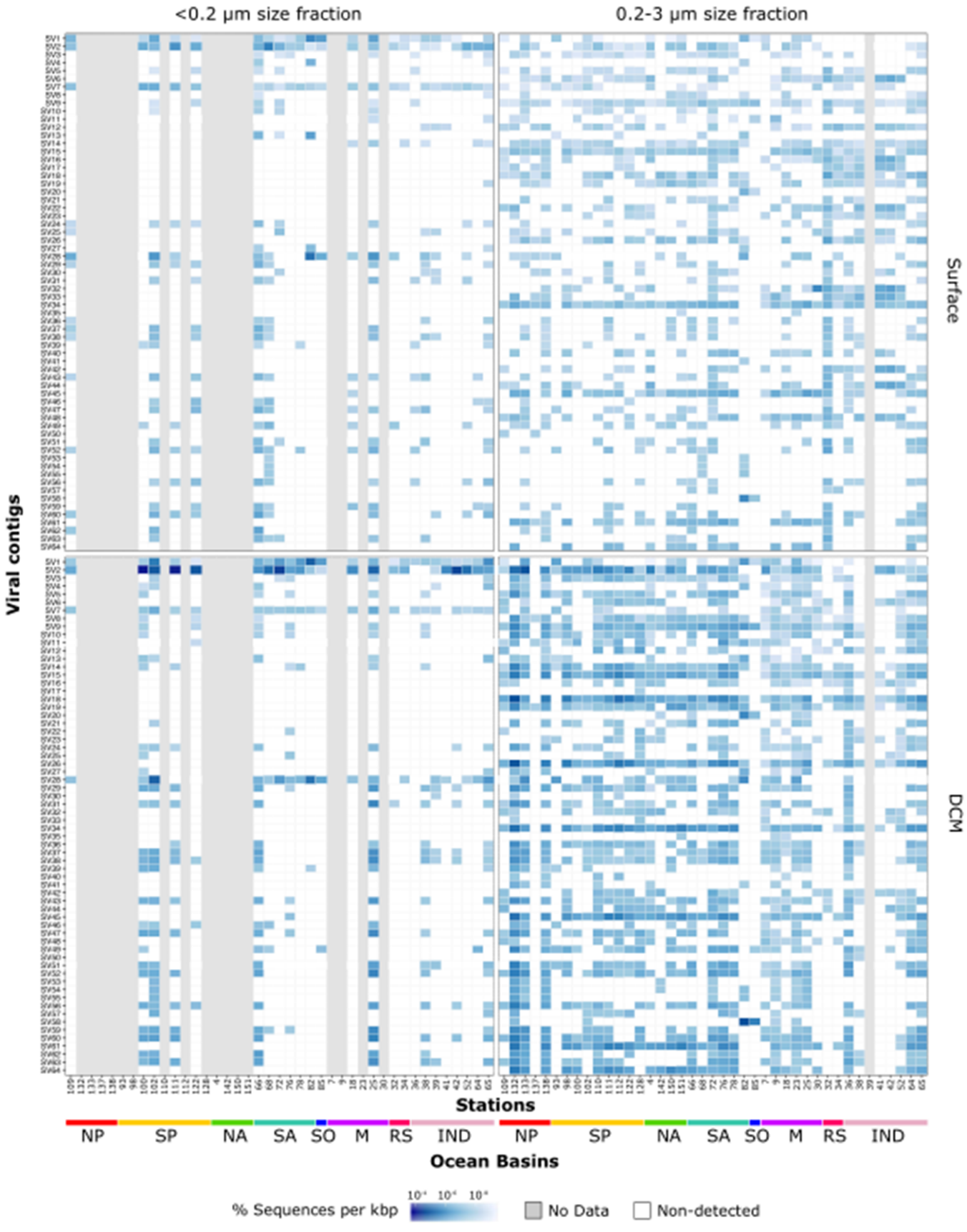
[†] Statistics are computed on the longuest sequence when SAG-associated viral sequences were retrieved in several cell.

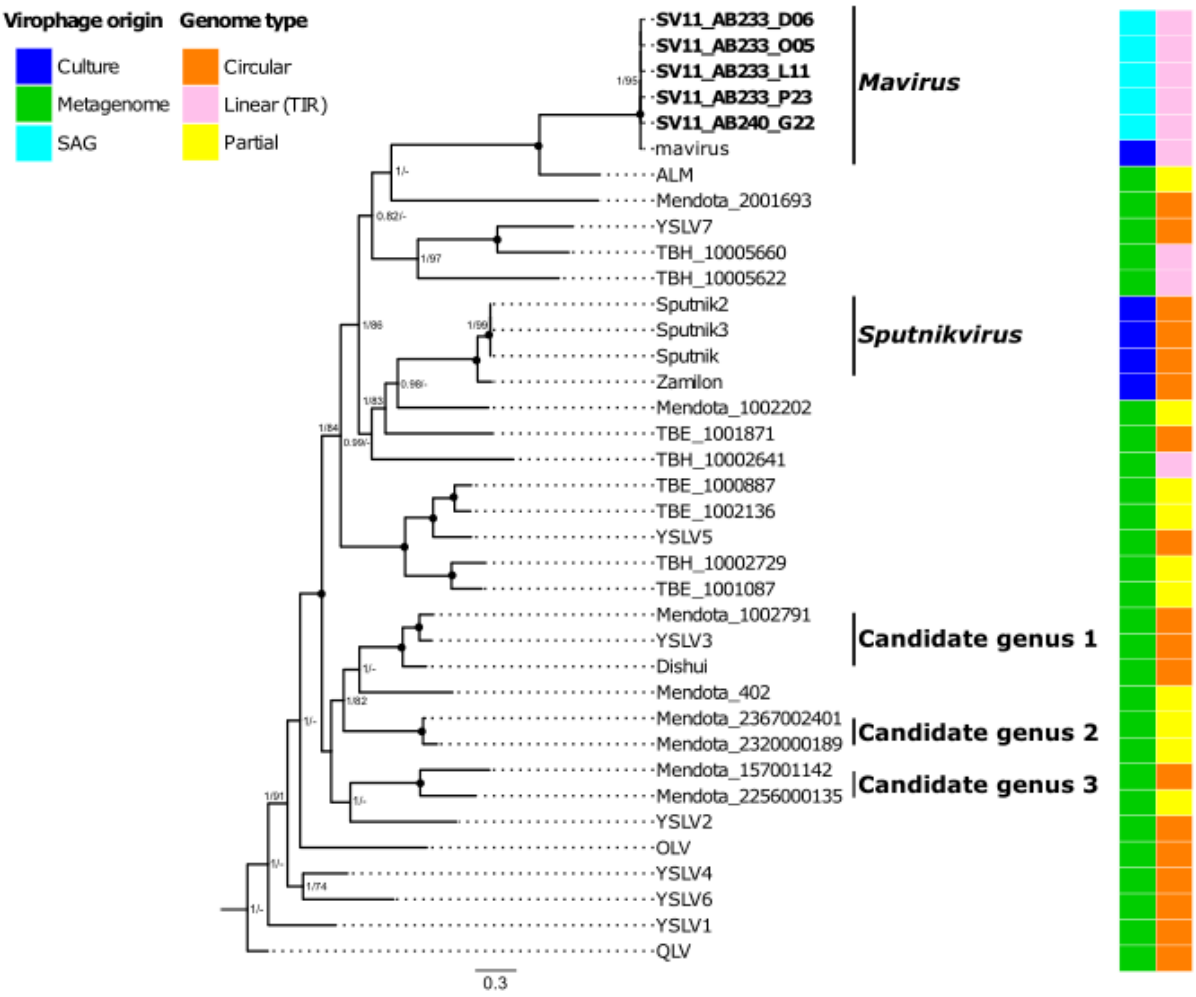
[‡] *SG* tBLASTx score. In bold, SAG-associated viral sequence that can be affiliated to the same genus level than their reference best hit (SG > 0.15).

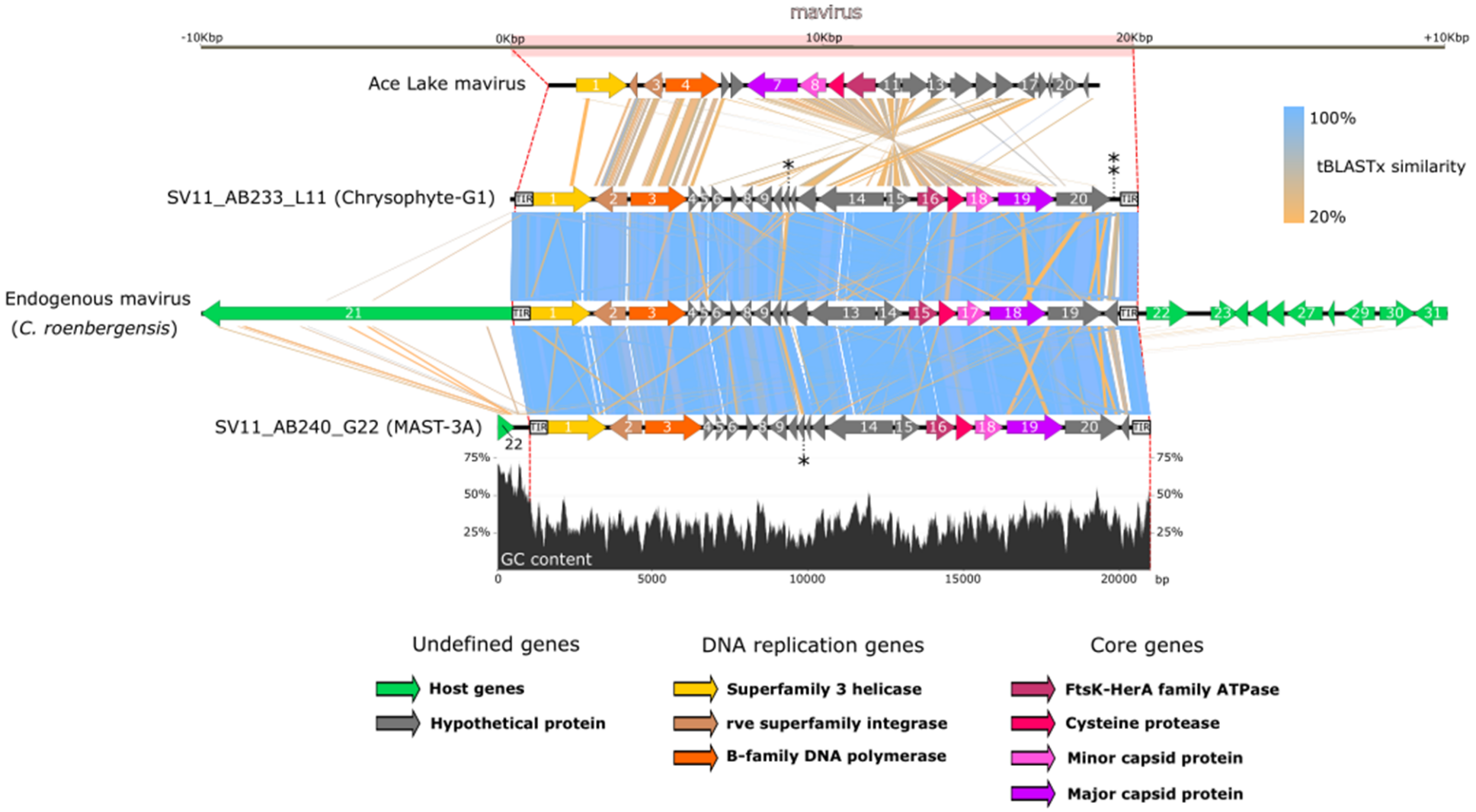
N/D Non Determined. Sequence were isolated from environmental surveys.

Viral signal sequence without taxonomic assignment are shown by the symbol (-)


(a)
SAGs without viral contigs (n=28)
SAGs with viral contigs (n=37)
Number of SAGs
Stramenopiles lineages
Chryso-G1
Chryso-H1
Chryso-H2
Dictyo
MAST-3A
MAST-3F
MAST-4A
MAST-4C
MAST-4E
MAST-7
Pelago
8
24
12
3
6
1
12
1
5
7
0
(b)
Number of lineages
1
2
3
4
1.6%
3.1%
95.3%
1.6%
98.4%
Number of cells
1
2


<0.2 µm size fraction
0.2-3 µm size fraction
Surface
DCM
Viral contigs
Stations
NP SP NA SA SO M RS IND
NP SP NA SA SO M RS IND
Ocean Basins
% Sequences per kbp
No Data
Non-detected


Virophage origin
Culture
Metagenome
SAG
Genome type
Circular
Linear (TIR)
Partial
SV11_AB233_D06
SV11_AB233_O05
SV11_AB233_L11
SV11_AB233_P23
SV11_AB240_G22
mavirus
ALM
Mendota_2001693
YSLV7
TBH_10005660
TBH_10005622
Sputnik2
Sputnik3
Sputnik
Zamilon
Mendota_1002202
TBE_1001871
TBH_10002641
TBE_1000887
TBE_1002136
YSLV5
TBH_10002729
TBE_1001087
Mendota_1002791
YSLV3
Dishui
Mendota_402
Mendota_2367002401
Mendota_2320000189
Mendota_157001142
Mendota_2256000135
YSLV2
OLV
YSLV4
YSLV6
YSLV1
QLV
Mavirus
Sputnikvirus
Candidate genus 1
Candidate genus 2
Candidate genus 3
0.3


mavirus
-10Kbp
0Kbp
10Kbp
20Kbp
+10Kbp
Ace Lake mavirus
SV11_AB233_L11 (Chrysophyte-G1)
Endogenous mavirus
(C. roenbergensis)
SV11_AB240_G22 (MAST-3A)
100%
tBLASTx similarity
20%
75%
50%
25%
GC content
0
5000
10000
15000
20000
bp
Undefined genes
Host genes
Hypothetical protein
DNA replication genes
Superfamily 3 helicase
rve superfamily integrase
B-family DNA polymerase
Core genes
FtsK-HerA family ATPase
Cysteine protease
Minor capsid protein
Major capsid protein