Article

https://doi.org/10.1038/s41564-024-01669-y

Single-cell RNA-seq of the rare virosphere reveals the native hosts of giant viruses in the marine environment

Received: 14 July 2023

Accepted: 7 March 2024

Published online: 11 April 2024

Check for updates

Amir Fromm[1,6], Gur Hevroni[1,4,6], Flora Vincent[1,5], Daniella Schatz[1], Carolina A. Martinez-Gutierrez[2], Frank O. Aylward[2,3] & Assaf Vardi[1]

Giant viruses (phylum *Nucleocytoviricota*) are globally distributed in aquatic ecosystems. They play fundamental roles as evolutionary drivers of eukaryotic plankton and regulators of global biogeochemical cycles. However, we lack knowledge about their native hosts, hindering our understanding of their life cycle and ecological importance. In the present study, we applied a single-cell RNA sequencing (scRNA-seq) approach to samples collected during an induced algal bloom, which enabled pairing active giant viruses with their native protist hosts. We detected hundreds of single cells from multiple host lineages infected by diverse giant viruses. These host cells included members of the algal groups Chrysophycae and Prymnesiophycae, as well as heterotrophic flagellates in the class Katablepharidaceae. Katablepharids were infected with a rare Imitervirales-07 giant virus lineage expressing a large repertoire of cell-fate regulation genes. Analysis of the temporal dynamics of these host–virus interactions revealed an important role for the Imitervirales-07 in controlling the population size of the host Katablepharid population. Our results demonstrate that scRNA-seq can be used to identify previously undescribed host–virus interactions and study their ecological importance and impact.

Nucleocytoplasmic large DNA viruses (NCLDVs), commonly known as giant viruses, are a group of double-stranded DNA viruses[1] (phylum *Nucleocytoviricota*). Giant viruses are abundant and have a broad phylogenetic diversity in aquatic ecosystems[2–4]. Infection by giant viruses can have profound metabolic consequences on their host owing to the expression of various viral auxiliary metabolic pathways involved in nutrient uptake, lipid metabolism and even energy production[1,5]. Some giant viruses infect and lyse bloom-forming algae and thereby play an essential role in recycling major nutrients and enhancing the metabolic flux that fuels the ocean microbiome[6]. Moreover, the evolutionary arms race between giant viruses and their hosts can have substantial consequences for gene transfer[7]. It may even lead to an integration of giant virus genomes into those of their hosts, resulting in profound evolutionary consequences that can modulate the host response to a changing environment[8,9].

Considering the key ecological role of giant viruses in the ocean, extensive efforts have been made to map their diversity across various ecosystems worldwide[2,10–12]. Consequently, our current knowledge about the ecological importance of giant viruses stems mainly from metagenomic surveys conducted at the bulk population level. Furthermore, host–giant virus models in the lab mainly consist of protists (that is, amoeba) that can phagocytose giant viruses without necessarily

[1]Department of Plant and Environmental Sciences, Weizmann Institute of Science, Rehovot, Israel. [2]Department of Biological Sciences, Virginia Tech, Blacksburg, VA, USA. [3]Center for Emerging, Zoonotic, and Arthropod-Borne Pathogens, Virginia Tech, Blacksburg, VA, USA. [4]Present address: Google Geo, Tel Aviv, Israel. [5]Present address: Developmental Biology Unit, European Molecular Biological Laboratory, Heidelberg, Germany. [6]These authors contributed equally: Amir Fromm, Gur Hevroni. e-mail: faylward@vt.edu; assaf.vardi@weizmann.ac.il

being their native hosts[13]. Consequently, knowledge about the interactions of giant viruses with their native host is currently limited to only a few model systems, and a deeper understanding of their life cycle and impact on the aquatic environment remains elusive.

Current approaches to predict host–virus pairs include examining correlations between the abundance of viruses and putative hosts[14,15] and identifying genes transferred between viral and host genomes[2]. Despite these efforts, we still lack fundamental knowledge of the native host of most giant viruses, including those highly abundant in the marine environment. Single-cell genomics is an attractive approach for detecting host and virus DNA within the same bacterial or protist cell in diverse environmental samples, especially uncultured or poorly studied organisms[16,17]. However, single-cell genomics also captures viral DNA derived from ingestion by heterotrophic protists rather than infection of native host cells[18–20]. Therefore, it may not be sufficient to link active viral infection to their specific hosts in high confidence. Single-cell transcriptomics is an innovative approach that can capture the high transcriptional heterogeneity in microbial populations[21], and it has recently been used to track host–virus dynamics by detecting the co-expression of a virus and its host transcriptomes within individual cells in the lab[22,23] and the natural environment[24]. This sensitive approach enables the detection of active viral infection at different phases, even for rare viruses that would otherwise be difficult to detect through conventional analysis of bulk metagenomic or metatranscriptomic data. ScRNA-seq targets poly(adenylated) RNA; therefore, when studying DNA viruses, it enables the detection of only active viral infection and eliminates the possibility of detecting viral particles that were taken up as a food source by protists. ScRNA-seq provides the expression profile (of both host and virus genes) within each cell, adding a functional dimension to single-cell genomics. This allows for an in-depth study of host–virus systems of even rare, uncultured species[25]. This approach can also be useful for gaining essential information on the diversity and evolutionary trajectory of uncultured protists without sequenced genomes[26]. The 10x Single-Cell RNA Sequencing (10x Genomics) is a platform for scRNA-seq that was applied to diverse organisms and applications and also used to describe host–virus dynamics in humans, for example, in immune cells infected by cytomegalovirus[27] or COVID-19 (ref. 28).

In the present study, we developed a scRNA-seq approach using the 10x Genomics platform to map infection by giant viruses to their native host cells across tens of thousands of single-cell transcriptomes from samples collected in natural planktonic communities. Using this method, we found dozens of infected cells representing eight distinct pairs of hosts and viruses. We identified the hosts of several giant viruses from multiple lineages, even when the host comprises less than half a per cent of the protist community. Overall, scRNA-seq provides a sensitive tool for identifying the native host of giant viruses and tracking their dynamics in the natural environment.

Results

To identify host–virus interactions in the ocean, we sampled natural plankton communities from an induced *Emiliania huxleyi* bloom during a mesocosm experiment in the Raunefjorden fjord near Bergen, Norway, in May 2018 (ref. 29). During this experiment, seven bags were filled with fjord water and monitored for plankton succession for 24 d. Ten samples of a size fraction of 3–20 μm were obtained from four bags and fixed on-site before being processed in the lab (Fig. 1a,b). Cells were resuspended and partitioned using a 10x Chromium microfluidic device for single-cell partitioning. Partitioned cells were encapsulated in droplets with beads containing cell-specific and sample barcodes (Fig. 1c). Within each droplet, cells were lysed and RNA was reverse transcribed. Each transcript was assigned a unique molecular identifier (UMI; Fig. 1d). Complementary DNA was pooled from all cells and sequenced (Fig. 1e). Cells were computationally demultiplexed by their cell-specific barcodes (Fig. 1f) and their transcripts were aligned to a reference database of giant virus marker genes (Fig. 1g). This database is made up of highly conserved genes that are broadly represented in NCLDVs, such as viral DNA polymerase Family B (PolB), viral type II topoisomerase (TopoII) and major capsid protein (MCP)[4]. The expression of these viral marker genes was quantified (see Methods for details) and cells with high viral expression were selected for further analysis (Fig. 1h). Reads from these selected cells were then assembled to recover longer transcripts (Fig. 1i,j). Despite the 10x RNA-seq method being aimed at sequencing poly(adenylated) messenger RNA, the amount of ribosomal RNA in a cell is high enough (around 80% of cellular RNA[30]) for a considerable amount of rRNA to be sequenced as well. This enabled the assembly of long contigs of 18S rRNA from single cells that were used to identify the native host (Fig. 1k). To identify which viruses are infecting these host cells, reads from selected cells were aligned to the database of viral marker genes. Cells with ambiguous identifiers were discarded (Fig. 1l) and host–virus pairs were determined based on homology to both a virus and a host (Fig. 1m).

After this workflow, 972 cells were defined as infected because they expressed at least 10 viral UMIs, more than 1 viral gene and at least 1 gene with a UMI count >1. Most of these cells ($n = 754$) were infected by *E. huxleyi* virus (EhV), in comparison to 218 cells that were infected by other viruses, confirming the prevalence of infected *E. huxleyi* cells during bloom demise[24,29]. The successful detection of *E. huxleyi*–EhV pairs confirmed that this pipeline could detect authentic hosts infected by a well-characterized giant virus. We have previously analysed in depth the population dynamics of *E. huxleyi* and its virus in this bloom[31]. In the present study, we sought to identify previously undescribed host–virus pairings and hence focused on cells that were not infected by EhV.

Uncovering host–virus interactions at a single-cell resolution

Out of the 218 infected non-*E. huxleyi* cells identified, 71 host–virus pairs were defined at the class or division level for the host and the family level for the virus (Fig. 2 and Source data); 147 cells were omitted because they expressed <10 viral reads confidently aligned to one virus family, could not be identified using 18S rRNA or their 18S rRNA was from two different sources (for example, a chimeric cell; Methods). Such ambiguous cells can stem from a technical error (for example, a doublet formed by the fusion of two individual cells) or after predation or grazing of an infected protist by a different protist. The latter scenario will require an active expression of viral mRNA within the highly acidic microenvironment of the predator's digestive vacuole, which is highly unlikely.

Of the 71 remaining host–virus pairs, viral genes were expressed in protists belonging to diverse and ecologically important taxa such as Chrysophyceae (31%), Prymnesiophyceae (21%) and Dinoflagellata (multiple classes, 10%), as well as the understudied class of Katablepharidaceae (14%) (Fig. 2). In about half (56%, $n = 40$) of infected cells, viral reads matched multiple families rather than a specific match to one virus lineage. This may imply that the specific virus infecting this host has yet to be discovered and is still missing in the reference database based on genomes from isolated viruses. Alternatively, it may suggest that a single cell can be infected by more than one virus lineage, a process known as superinfection[32]. In 44% of the cells ($n = 31$), at least 90% of viral reads matched a specific virus family (Fig. 2). These infected cells represent distinct pairs between eight protist taxa and giant viruses from the order *Imitervirales* (IM): *Mesomimiviridae* (IM_01), a newly defined family of giant viruses[33], *Mimiviridae* (IM_16), IM_09 (recently named *Schizomimiviridae*) and IM_07. This is consistent with the reported dominance of the *Imitervirales* in marine ecosystems[3]. To our knowledge, giant viruses that infect members of the Chrysophyceae have not yet been identified. However, a recent study predicted that giant viruses infect this group based on co-occurrence network analysis of virus–host abundance profiles in metagenomic datasets[3]. Moreover, Chrysophyceae-derived genes in the genomes of giant

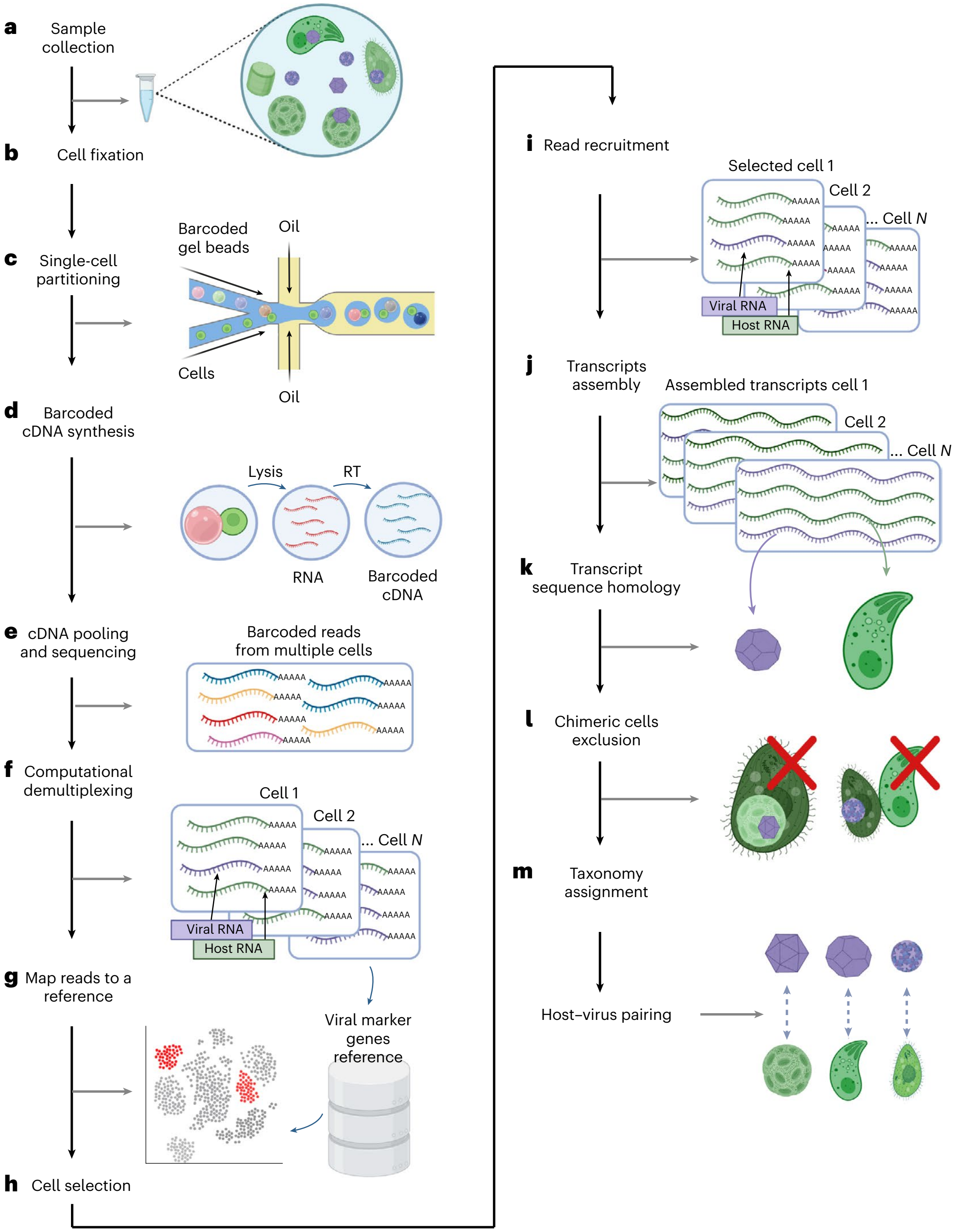


Fig. 1 | A pipeline for detecting host–virus pairs in the natural environment. **a**,**b**, Samples collected from the natural environment (**a**) and fixed in methanol (**b**). **c**, Resuspended cells partitioned by a 10x Chromium microfluidic device. Partitioned cells were combined with beads containing cell-specific barcodes and sample barcodes. **d**, Cells lysed within each droplet and RNA reverse transcribed. Each transcript was assigned a UMI. **e**, The cDNA pooled from all cells and sequenced using Illumina. **f**, Cells computationally demultiplexed by their cell-specific barcodes using Cell Ranger. **g**, Reads from all cells aligned to a reference of giant virus marker genes using 10x Cell Ranger to identify cells expressing viral transcripts. **h**, A subset of cells with high expression of viral transcripts selected for subsequent analysis. **i**, Single-cell transcripts recruited from each selected cell. **j**, Trimmed single-cell reads (60 bp) assembled to generate longer single-cell transcripts (110–2,050 bp). **k,** Prediction of the host encoding for the transcripts determined using assembled sequence homology analysis to 18S rRNA. The virus was identified using the homology of raw reads mapped to core NCLDV genes. **l**, Cells containing 18S rRNA from multiple sources removed. **m**, Taxonomy was assigned to the host and virus using transcripts and reads from each cell and phylogenetic analysis of 18S rRNA genes (host) and NCLDV marker genes (virus). Black arrows indicate the direction of the pipeline. Grey arrows point to the intermediate output of each step. Figure 1 was created with BioRender.com.

viruses suggested that members of the family *Mimiviridae* from the order *Imitervirales* infect Chrysophyceae[3].

Out of the identified virus families, *Mesomimiviridae* was the most prevalent group of viruses that actively infected cells (65% of distinct links, 20 cells) and they infect multiple cells from various families, mostly Chrysophyceae (13 cells) and Prymnesiophyceae (4 cells). These findings suggest that mesomimivirids are important mortality agents for these groups. Several mesomimivirids that infect bloom-forming

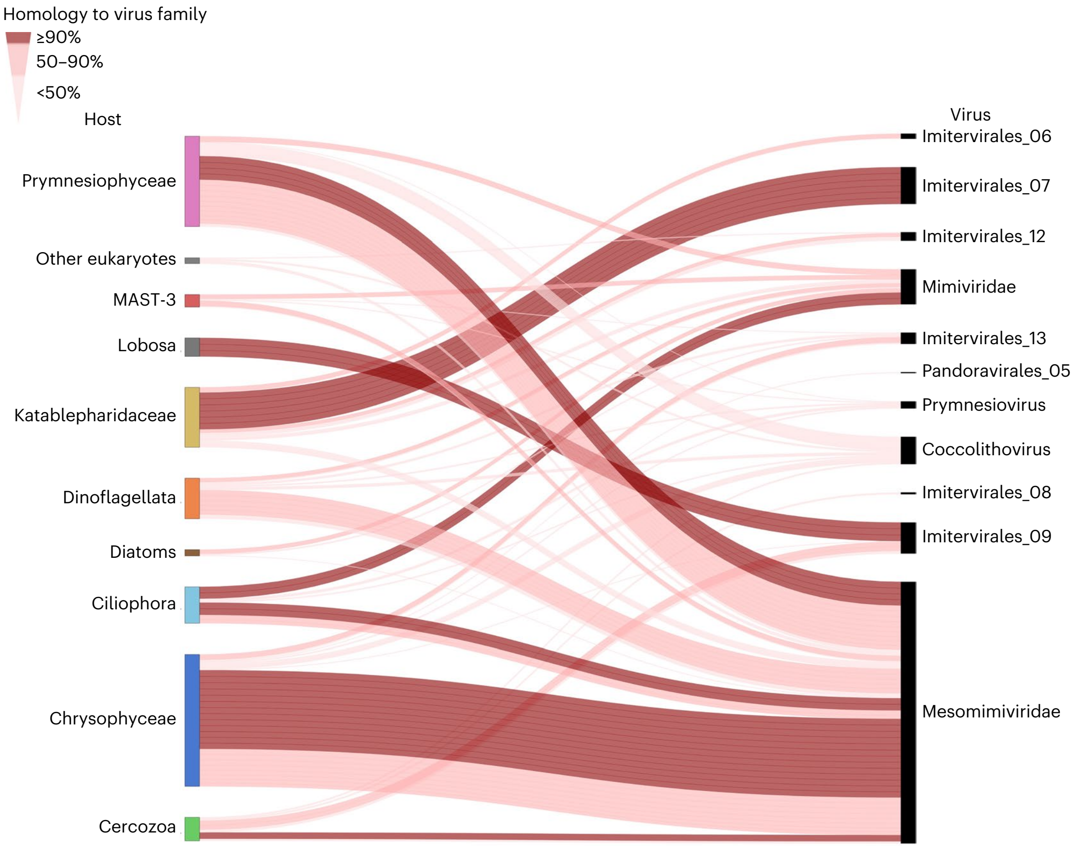


Fig. 2 | Pairing between host cells and their actively infecting giant viruses. Connecting lines represent the predicted pairing of host cells to their infecting viruses ($n = 71$). Dark, thick lines represent an unambiguous pairing of host cells (≥90% of viral reads, 31 cells) to a single virus family, so each line represents one cell. Thinner, lighter lines represent less specific pairing of host cells (<90% of viral reads, 40 cells) to multiple virus families, so each cell can have multiple lines. Cells predominantly infected with EhV were not included in the analysis.

Prymnesiophyceae algae, such as *Chrysochromulina* and *Phaeocystis* spp., have been cultivated[34,35]. These results are consistent with the observation that giant viruses from the family *Mesomimiviridae* are the most abundant and widespread giant viruses in the ocean[4] and are known to infect a wide range of protist hosts. The strongest signal of viral infection that we could detect stems from the virus family IM_07 in six infected cells belonging to the class Katablepharidaceae, a lineage of heterotrophic flagellates related to Cryptophytes[36]. Each of these cells contains between 530 and 3,600 reads aligning to IM_07 viral genes, making them the cells with the strongest signal of viral infection in our analysis (see Source data for Fig. 2). To date, no nuclear genome of any Katablepharidaceae has been sequenced. In general, it is an underexplored protist lineage, although some members of this group are known to be grazers of bacteria in marine ecosystems[37]. To our knowledge, no virus has been described to infect this class and no specific host was reported for a virus in the IM_07 lineage. Only 19 metagenome-assembled genomes from the IM_07 lineage are currently available and all have been found in aquatic ecosystems[4]. These results reveal that heterotrophic flagellates in the class Katablepharidaceae are among the hosts of the cryptic IM_07 lineage of giant viruses.

Our results also predict other links between lineages of giant viruses and their possible hosts. For example, viral transcripts from the IM_09 viral family were found in cells of the Lobosa class (Amoebozoa, 3 cells) and transcripts from the *Mimiviridae* were found in Ciliophorans (Ciliates, 2 cells). However, no NCLDV marker genes could be recovered from these cells, so we could not affirm these links.

To verify the phylogenetic position of the identified host–virus pairs, we constructed phylogenetic trees from host and virus co-expressed transcripts assembled from individual cells (Extended Data Fig. 1). We chose the MCP and PolB as giant virus gene markers because these are conserved NCLDV genes that are typically highly expressed during infection[1]. Assembled 18S rRNA was used to determine host taxa because the databases for this gene span a broad diversity of protists. The phylogenetic trees further affirm the connections between the protist classes Prymnesiophyceae (cells 9 and 10) and Chrysophyceae (cells 1–6) and the virus family *Mesomimiviridae*, the former being consistent with previous studies[34,35]. It also affirmed the connection between Katablepharidaceae (cells 7 and 8) and IM_07. Hence, by using direct single-cell transcript mapping and marker gene analysis, we elucidated multiple virus–host relationships, several of which were previously unknown.

Multiple viral infections co-occurring in a natural population

To examine co-occurring viral infections in different protist populations during bloom succession, all reads were aligned to a customized host–virus reference database, representing the different protist groups in the population. This database was generated based on the single-cell transcriptomes of the selected infected cells (Fig. 2). To this host–virus reference transcriptome, we added genes from EhV and *E. huxleyi*, which dominated the bloom[24]. Data derived from 16,358 RNA sequenced single cells were aligned to the host–virus reference and visualized using Uniform Manifold Approximation and Projection (UMAP) representation. Each cell was assigned taxonomy based on 18S rRNA homology. Cells that expressed at least ten UMIs of newly assembled viral transcripts were considered infected. This analysis revealed active viral infection at a single-cell level, occurring in different protist host cells originating from diverse taxa in the natural environment (Fig. 3). As expected, the largest population is of class Prymnesiophyceae, the class

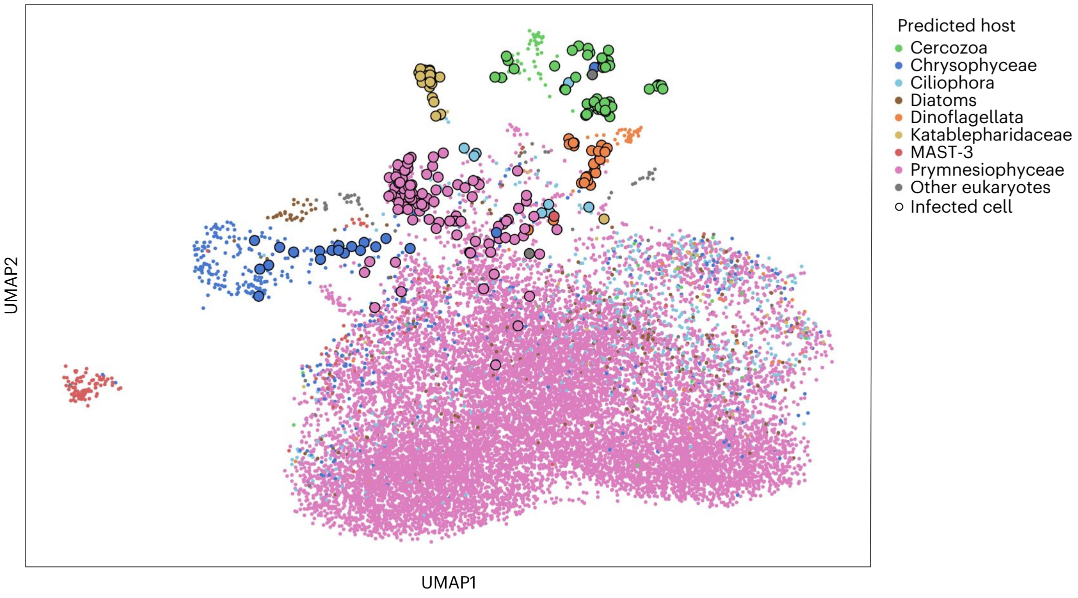


Fig. 3 | Single-cell metatranscriptome of co-occurring viral infections of diverse protist groups. A UMAP projection of cells aligned to a host–virus reference. The reference comprises single-cell transcripts assembled from the previously identified infected cells (Fig. 2) and of EhV–*E. huxleyi* genes. Each cell is coloured according to its predicted taxonomy by 18S rRNA homology (n = 16,358). Cells that express at least ten UMIs of viral transcripts are considered infected and are enlarged and marked with a black edge (n = 239).

of *E. huxleyi*, the bloom-forming species in the mesocosm[29,31]. Smaller yet distinct populations consist of various groups: Dinoflagellata, Diatoms, Chrysophyceae, Cercozoa, Katablepharidaceae and MArine STramenopiles (MAST)-3, a group suggested to be one of the most abundant MAST groups in the ocean[38] (Fig. 3). Infected cells could be identified by high expression of contigs of viral origin (Fig. 3; Methods). These cells belong to previously described taxa (Fig. 2): Prymnesiophyceae, Chrysophyceae, Cercozoa, Dinoflagellata and Katablepharidaceae. These hidden host–virus dynamics and diversity are often entirely masked when the rare virosphere is analysed by bulk metatranscriptomes or metagenomes. Furthermore, these minute subpopulations can be masked when viral infection of the dominated bloom-forming algae occurs. Therefore, scRNA-seq provides an opportunity to detect active viral infection at the cellular level, provides a sensitive lens into host–virus dynamics in the rare virosphere and enables the tracking of fine-scale virus–host interactions and their ecological significance.

Host–virus interactions in the Katablepharidaceae class

Our approach enables mapping active infection at single-cell resolution among diverse protist host cells and can provide a sensitive means to detect rare infected cells. As a case study, we tracked Katablepharidaceae cells for which we detected infection by giant viruses of the IM_07 family (Figs. 2 and 3). No other host was associated with this virus family in our analysis and there is no known host for this group based on previous studies, making it a good case study to explore the dynamics of an undescribed, distinct host–virus interaction (Fig. 4a). Katablepharidaceae represent <0.5% of all detected cells (Fig. 3; n = 67 of 16,358 cells). A distinct subpopulation of infected Katablepharidaceae cells could be observed that makes up about 10% of all infected cells (Fig. 3b, in yellow; n = 26 of 239 infected cells). To explore this infected subpopulation further, we pooled together and assembled the transcriptomes from 26 infected Katablepharidaceae cells from the same sample (bag no. 4, day 20 of the mesocosm experiment). Assembled contigs from these cells matched the 18S rRNA gene of *Leucocryptos marina* (>95% identity, e value < 10^{-10}; Supplementary Data Table 1). The best match for the virus infecting Katablepharidaceae cells is the IM_07 member GVMAG-M-3300020187-27 (identity >99%, e value ≤ 10^{-10}), a virus that was assembled from a metagenomic analysis on samples obtained from Kabeltonne, Helgoland, North Sea, but has not yet been isolated[2]. So far, no virus has been identified to infect the genus *Leucocryptos* (and the class Katablepharidaceae in general). It is also the only definitive host for giant viruses of the IM_07 lineage.

Characterization of the *Leucocryptos* virus

Katablepharidaceae are a class of flagellated heterotrophic plankton that consists of five species, none of which has a published nuclear genome[39]. *Leucocryptos marina*, the closest relative to the predicted host, is abundant in coastal waters with high plankton productivity[37]. The predicted *Leucocryptos* virus has the largest genome recovered from the IM_07 lineage (950 kbp) and encodes for 894 genes (Fig. 4b)[2,4]. Reads from the population of infected cells were pooled together and aligned to the assembled viral genome, and the expression of viral genes was examined (Fig. 4a,b and Source data). The virus encodes a complex repertoire of 13 proteins probably involved in manipulating cellular stress responses and cell-fate regulation, including a predicted Bax-1 apoptosis inhibitor, a metacaspase homologue, a homologue of heat-shock protein 90 (HSP90), two homologues each of HSP70 and eight homologues of DnaJ (HSP40) genes. These proteins are placed inside well-defined viral clades separated from eukaryotic clades and together with other viruses of the order *Imitervirales* (Extended Data Fig. 2), suggesting that these genes were horizontally transferred from host to viral genomes early in their evolution[8]. HSP90 and HSP70 are among the most highly expressed viral genes in the infected Katablepharidaceae population (Fig. 4b and Source data). HSPs play a role in the life cycle of many viruses, mostly in viral replication, and in some cases are encoded by the virus[40]. HSPs of the herpes simplex virus were shown to regulate virus-induced apoptosis and other HSPs[41]. In addition, viral-encoded metacaspases have been hypothesized to regulate host cell death and were identified in diverse giant viruses from the marine environment[42,43]. The high prevalence and expression of these cell-fate regulators encoded by the *Leucocryptos* virus suggest that they have an essential function in its life cycle by controlling its host's cell death.

It is interesting that the virus also encodes for nine predicted MCPs (Fig. 4b), a high number even for giant viruses, which often encode several[2]. Some of these predicted MCPs are co-localized in the genome, suggesting gene duplication[44]. Relative to the anti-apoptotic genes, the

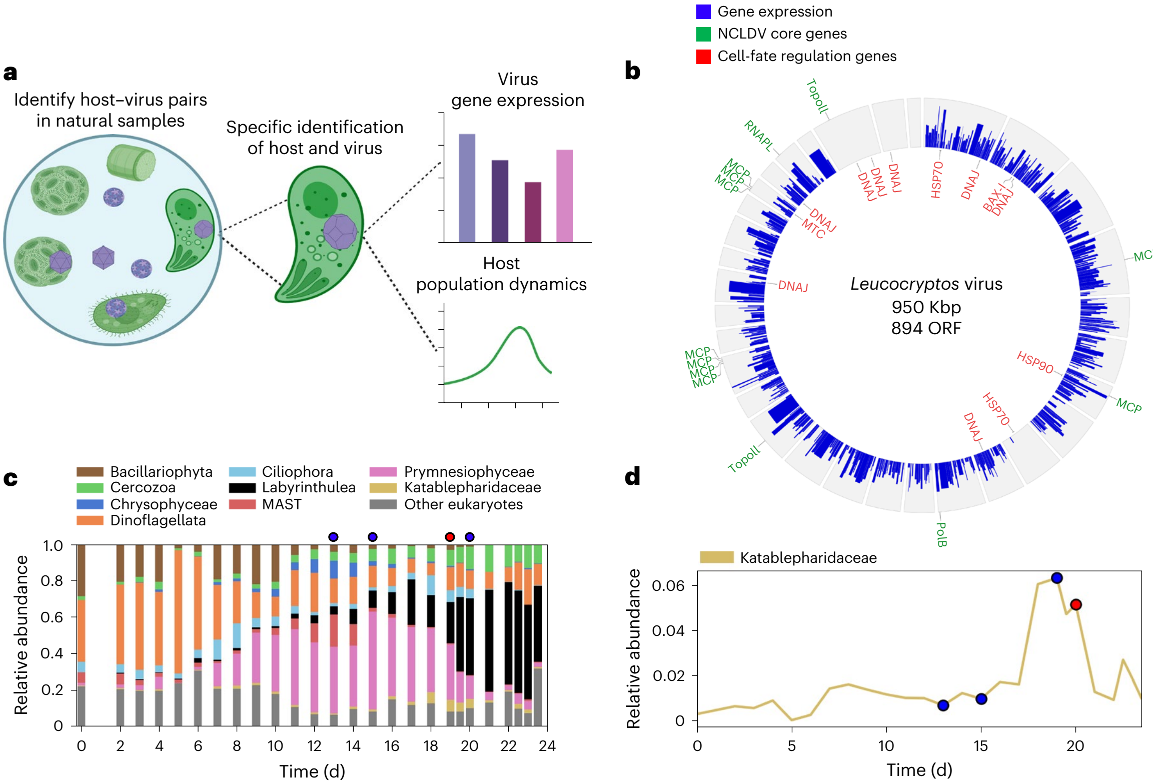


Fig. 4 | Characterization of the predicted *Leucocryptos* virus genome and its suggested impact on the population dynamic of Katablepharidaceae. **a**, A workflow illustration of the predicted *Leucocryptos* virus characterization, from its detection in single cells to finding its putative host and eventually tracking its gene expression and host population dynamics in natural populations. **b**, A Circos plot of the predicted *Leucocryptos* virus. Gene expression (expected read counts, $\log_2$(transformed)) from 26 infected cells identified as *Leucocryptos* is shown in the blue bars. Key predicted genes are shown in their corresponding location. In green are core NCLDV genes: PolB, DNA polymerase family B; TopoII, type II topoisomerase; MCP, Major capsid protein; RNAPL, RNA polymerase. In red are cell-fate regulation genes: HSP, heat-shock protein; BAX-I, apoptosis regulator BAX inhibitor; DnaJ, DnaJ chaperone; MTC, metacaspase. The genome of the virus is shown as circular for convenience. **c**, Relative abundance of different taxonomic groups in bag no. 4 during the mesocosm experiment. Samples for scRNA-seq were collected on marked days 13, 15, 19 and 20. The red dot points to the time when most infected cells were detected. **d**, Relative abundance of the Katablepharidaceae class in bag no. 4 during the time course of the mesocosm experiment[29]. Markers are as in **c**. Panel **a** created with BioRender.com.

MCP genes are lowly expressed in scRNA-seq analysis. This pattern of expression may indicate that infection is at an early phase.

Population dynamics of Katablepharids after viral infection

On pairing of specific giant viruses to their hosts at the single-cell level, we sought to study their host–virus dynamics at the population level (Fig. 4a). The relative abundance of Katablepharidaceae was detected using amplicon sequence variants (ASVs) of 18S rDNA[29] that were sequenced at different time points during the plankton succession in the mesocosm (Fig. 4c). These results confirmed the presence of these protist classes, which we revealed at the single-cell level using scRNA-seq. Moreover, the ASV analysis confirmed the rarity of Katablepharidaceae in the community compared with other taxa analysed here, such as Prymnesiophyceae, Dinoflagellata and Cercozoa. The dominant bloom-forming species of Prymnesiophyceae was *E. huxleyi* until bloom demise on day 18 (ref. 29) (Extended Data Fig. 3). The Katablepharidaceae class increased in abundance from 1% of the community on day 15 to a maximum of 6% on day 19, followed by a population decline back to 1% only 2 d later, on day 21 (Fig. 4d). This sharp demise in the relative abundance of Katablepharidaceae was observed after day 20, the same day on which we detected that 86% ($n = 26$) of the observed Katablepharidaceae single cells were infected. This strongly suggests that the IM_07 lineage virus was responsible for the population's demise.

Discussion

Research in the last decade has revealed that giant viruses are ubiquitous components of ecosystems around the globe[2,3,11]. Extensive metagenomic and single-cell genomic studies have revealed a vast diversity of known giant virus lineages, particularly in the marine environment, and there is a growing interest in their unique infection cycles and ecological roles[2–4]. Still, a major knowledge gap in our understanding of giant virus dynamics and evolution concerns the identity of their native host populations. In the present study, we show that a sensitive scRNA-seq approach can be used to identify the authentic hosts of giant viruses during infection in the marine environment. By applying this scRNA-seq approach, we demonstrate an ability to detect active viral infection in specific protist host cells, including in rare populations. It also enables the study of co-expressed genes of the host and the virus at different phases of the infection dynamic. This approach can provide insights into the life cycle of specific groups of viruses and within their authentic host (including uncultured hosts) in their natural ecosystem, and it can potentially provide a sensitive tool to discover host response to viral infection, including the discovery of anti-viral defence systems. Last, use of scRNA-seq rather than single-cell genomics can help to reduce the probability of capturing ingested free virions. Although the probability of RNA surviving in the digestive vacuole is very low, it does not entirely eliminate the possibility of detecting freshly grazed infected cells. We, therefore, eliminated all cells suspected of being chimeric with ambiguous 18S rRNA contigs. Future analysis could potentially use co-expression of different protists within individual cells to provide insight into grazing rates.

In the present study, we analysed samples from an algal bloom and linked multiple host cells with infecting giant viruses at the single-cell resolution. These findings led us to discover a virus that infects a member of the underexplored protist class Katablepharidaceae, which did not have a known infecting virus. Furthermore, the virus detected belongs to the IM_07 family[4], which has had no predicted hosts so far. This virus was described before only in a metagenomic analysis[2], demonstrating how our approach can be utilized to identify the hosts of viruses previously described in bulk metagenomic data and explain population dynamics in the natural environment[29]. This approach demonstrates how scRNA-seq has the potential to connect the wealth of metagenomic data, which has greatly expanded our knowledge of the diversity of giant viruses, to knowledge gained of their function and ecological roles based on the co-expression of their gene repertoire in their native host cells.

Improved charting of the rare virosphere can deepen our understanding of complex ecosystems. Rare species are often more active than abundant species, have a high per-organism contribution to community activities and enhance the functionality of abundant species[45]. Moreover, active infection by rare viruses can serve as a seed bank population for subsequent infections[46]. This is especially apparent during the phase of post-bloom demise, as in our study, in which the dominance of available host cells rapidly shifts in composition. Recent attempts to quantify viral infection rates have shown how low infection levels are common in marine ecosystems and may have serious consequences for viral persistence over broad geographical areas[47,48]. Tracking active viral infection using single-cell metatranscriptomic approaches may provide insights into the ecological importance of viruses in the marine environment. It will help to bridge the gap between environmental metagenomic analysis and mechanistic studies of virus–host infection dynamics.

Methods

Mesocosm core setup and sampling procedure

Samples were obtained during the AQUACOSM VIMS-Ehux mesocosm experiment in Raunefjorden near Bergen, Norway (60° 16′ 11″ N; 5° 13′ 07″ E), in May 2018. Seven bags were filled with 11 m^3 of water from the fjord, containing natural plankton communities. Algal blooms were induced by nutrient addition and monitored for 24 d, as previously described[29]. Ten samples were collected from four bags, as follows: from bag 3, on days 15 and 20 (named B3T15 and B3T20, respectively); from bag 4, on days 13, 15, 19 and 20 (named B4T13, B4T15, B4T19 and B4T20, respectively); from bag 6, on day 17 (named B6T17); and from bag 7, on days 16, 17 and 18 (named B7T16, B7T17 and B7T18, respectively).

Samples were initially filtered as follows: 2 l of water was filtered with a 20 µm mesh and collected in a glass bottle. The cells were then concentrated through gentle gravity filtration on a 3 µm polycarbonate filter (Whatman), mounted on a reusable bottle top filter holder (Thermo Fisher Scientific). The biomass on the filter was regularly resuspended by gentle pipetting.

For samples B7T16, B7T18, B4T15, B3T15, B6T17, B7T17 and B4T19, the 2 l of seawater was concentrated down to 100 ml, distributed into two 50 ml tubes, which corresponds to a 200× concentration. For B4T13, the concentration factor was 140×. For B4T20 and B3T20, the concentration factor was 100×. The different concentration factors are explained by filter clogging and various field constraints, including processing time. For all samples except B3T20, the 50 ml tubes were centrifuged for 4 min at 2,500*g*, after which the supernatant was discarded. Pellets corresponding to the same day and same bag were pooled and resuspended in a final volume of 200 µl of chilled phosphate-buffered saline (PBS). Then, 1,800 µl of pre-chilled, high-performance liquid chromatography (HPLC)-grade, 100% methanol was added drop by drop to the concentrated biomass. For B3T20, the concentrated biomass was centrifuged for 4 min at 2,500*g*, resuspended in 100 µl of chilled PBS, to which 900 µl of chilled, HPLC-grade, 100% methanol was added. Then, samples were incubated for 15 min on ice and stored at −80 °C until further analysis.

Library preparation and scRNA-seq using 10x Genomics

For analysis by 10x Genomics, tubes were defrosted and gently mixed, and 1.7 ml of the samples was transferred into an Eppendorf Lowbind tube and centrifuged at 4 °C for 3 min at 3,000*g*. The PBS/methanol mix was discarded and replaced by 400 µl of PBS. Cell concentration was measured using an iCyt Eclipse flow cytometer (SONY) based on forward scatter. Cell concentration ranged from 1,044 cells ml^{-1} to 9,855 cells ml^{-1}. All concentrations were brought to 1,000 cells ml^{-1} to target recovery of 7,000 cells, according to the 10x Genomics Cell Suspension Volume Calculator Table provided in the user guide. The cellular suspension was loaded on to Next GEM Chip G targeting 7,000 cells and then ran on a Chromium Controller instrument to generate a GEM emulsion (10x Genomics). 3′-ScRNA-seq libraries were generated according to the manufacturer's protocol (10x Genomics Chromium Single Cell 3′ Reagent Kit User Guide v.3/v.3.1 Chemistry) on different occasions: B4T19 and B7T17 in January 2020 and B3T15, B3T20, B4T13, B4T15, B4T20, B6T17, B7T16 and B7T18 in August 2020, with 12 cycles for cDNA amplification and 15 cycles for library amplification. Library concentrations and quality were measured using the Qubit dsDNA High Sensitivity Assay kit (Life Technologies). Libraries were pooled according to the targeted cell number, aiming for a minimum of 20,000 reads per cell. Pooled libraries were sequenced using the NextSeq 500 High Output kit (75 cycles).

Computational pipeline

A step-by-step description of the computational pipeline from this step onward, including all in-house scripts used, is detailed in the GitHub repository under github.com/vardilab/host-virus-pairing.

Detection of infected cells in the scRNA-seq data using a customized viral genes database

To detect viral transcripts, a reference was built from a database of highly conserved genes[6] from all NCLDVs in the Giant Virus Database[9], such as family PolB, RNA polymerase subunits and the MCP. The genes were clustered using CD-HIT v.4.6.6 at 90% nucleotide identity to remove redundancy[49]. From this database of 34,866 genes, a reference was created using the 10x Genomics Cell Ranger mkref command. The Cell Ranger Software Suite (v.5.0.0) was used to perform barcode processing (demultiplexing) and single-cell UMI counting on the raw reads from 47,391 cells using the count script (default parameters), with the deduplicated NCLDV database as a reference. For downstream analysis, 972 cells that highly expressed multiple NCLDV genes and were considered infected were selected. These infected cells were selected based on the following criteria: (1) cell expresses in total ≥10 viral UMIs[23,24], (2) expression of more than one viral gene (>1) and (3) expression of at least one gene with a UMI count >1. Cell selection was wrapped using an in-house script (choose_cells.py).

Identifying the taxonomy of individual cells by sequence homology to rRNA

Raw reads from each cell were pulled by the cell's unique barcode identifier using seqtk v.1.2. Reads were then trimmed (command: trim_galore --phred33 -j 8 --length 36 -q 5 --stringency 1 --astqc -e 0.1) and poly(A) was removed (command: trim_galore --polyA -j 1 --length 36), using TrimGalore (v.0.6.5), a Cutadapt wrapper[50]. Trimmed reads from each cell were assembled using rnaSPAdes 3.15 (ref. 51) with *k*-mer 21,33. Raw reads pulling, trimming and assembly were wrapped using an in-house script (assemble_cells.sh). To identify the taxonomy of the cells, assembled contigs from each cell were matched against 18S rRNA sequences from the Protist Ribosomal Reference (PR2)[52] and metaPR2 (ref. 53). To remove redundancy, the sequences in each database were clustered using CD-HIT v.4.6.6 at 99% identity[49]. Contigs were filtered using SortMeRNA v.4.3.6 (ref. 54) with default parameters against the PR2 database and then aligned to the PR2 and metaPR2 databases using Blastn[55], at 99% identity, e value ≤ 10^{-10} and alignment length of at least 100 bp. Contigs were ranked by their bitscore and only the best hit was kept for each contig. Each contig was assigned to one of the following taxonomic groups that were prevalent in the sample: the classes Bacillariophyta (diatoms), Prymnesiophyceae, Chrysophyceae, MAST-3 and Katablepharidaceae, and the divisions Pseudofungi, Lobosa (Amoebozoa), Ciliphora (Ciliates), Dinoflagellata and Cercozoa. Contigs that matched other groups were assigned as 'other eukaryotes'. Contigs that matched more than one of these taxonomic groups were considered non-specific and were therefore ignored. Chimeric contigs were determined by different genomic regions matching for different taxonomic groups. Cells with chimeric contigs were also excluded to avoid doublets. This downstream analysis of Blast result was wrapped using an in-house script (Sankey_wrapper_extended.ipynb). To avoid detection of doublets and predators, cells that transcribe 18S rRNA transcripts homologous to more than one taxonomic group were conservatively omitted. Of the 972 infected cells detected, 418 (43%) were omitted because we could not assemble specific 18S rRNA contigs from them or because their identity was ambiguous. None of the cells that were assigned 'other eukaryotes' had contigs with conflicting annotations (contigs matching different classes).

Identifying the infecting virus using a homology search against a customized protein database

To identify transcripts derived from giant viruses, reads from the detected 972 infected cells were compared with a customized protein database using a translated alignment approach. To ensure that as many giant viruses as possible were represented, a database was constructed by combining RefSeq v.207 (ref. 56) with all predicted proteins in the Giant Virus Database[4]. The proteins were then masked with tantan[57] (using the -p option) and generated the database with the lastdb command (using parameters -c, -p). To identify the infecting virus, the raw sequencing reads in each of the 972 single-cell transcriptomes were compared with the constructed database using LASTAL v.959 (ref. 58) (parameters -m 100, -F 15, -u 2) with best matches retained. The same procedure was done for the assembled transcripts from each cell to identify viral transcripts. The results were analysed at different taxonomic levels, consistent with the Giant Virus Database (for giant viruses) or National Center for Biotechnology Information (NCBI) taxonomy[39] (everything else).

Cells ($n = 754$) with best matching virus coccolithovirus, were omitted from the downstream analysis because EhV-infected cells were already reported to be abundant in the algal bloom[31] and our analysis aims to explore other host–virus pairs.

Plotting host–virus pairs in a Sankey plot for host cells and their infecting giant viruses

Of the 218 cells detected as infected by viruses other than EhV, 71 were selected that could be identified using assembled 18S rRNA transcripts and had at least 10 reads aligned to one of the virus families (Fig. 2 and Source data). Only links representing at least 10% of the aligned reads in each cell are shown to highlight the strong links. The Sankey plot was constructed using Holoviews v.1.15.4; see sankey_wrapper.ipynb in the GitHub repository.

Phylogenetic trees of viral and host marker genes

For phylogenetic analysis, 31 cells were chosen based on a strong correlation (≥90% of viral reads matched one virus family) between the host and a virus.

To obtain reference 18S rRNA sequences to include in a phylogeny, all transcripts assembled from these cells were compared with the PR2 database[52] using BLASTN v.2.9.0+ (parameters -perc_identity 95, -evalue 10^{-10}, -max_target_seqs 20, -max_hsps 1). Sequences shorter than 1,000 bp were removed from the reference and the remainder of the sequences were de-replicated with cd-hit v.4.7 (ref. 49) (-c 0.99) to prevent the inclusion of excessive almost identical references. Sequences were aligned with Muscle5 (ref. 59) (default parameters) and diagnostic trees were created with FastTree v.2.1.10 (ref. 60) for quick visualization of trees and pruning long branches. Additional phylogenetic trees were constructed with IQ-TREE v.2.1.2 (ref. 52) to confirm the topology (parameters -m GTR+F+G4 -alrt 1000 -T AUTO --runs 10). To identify MCP sequences in the single-cell transcriptomes, proteins were first predicted using FragGeneScanRs v.1.1.0 (ref. 61) (parameters -t, illumina_10). The resulting protein sequences were compared with MCPs in the Giant Virus Database with BLASTP v.2.12.0+ (parameters -evalue 10^{-3}, -max_target_seqs 20, -max_hsps 1) as well as to a customized MCP hidden Markov model (HMM) that was previously designed[11] using hmmsearch in the HMMER3 v.3.3.2 package[62] (e value $\leq 10^{-3}$). The results of these searches were manually inspected and sequences were subsequently aligned with Muscle5 (default parameters). Similarly, as with the 18S rRNA sequences, diagnostic trees were first made with FastTree v.2.1.10 and pruned long branches before making additional trees with IQ-TREE v.2.1.2 to confirm the overall topology (parameters m LG+F+G4 -alrt 1000 -T AUTO --runs 10). Cells for which transcripts are present in both viral and host trees were denoted (Extended Data Fig. 1 and Source data). All the codes used to produce the trees are wrapped in the folder 'marker_gene_trees' in the GitHub repository.

ScRNA-seq data alignment to a customized reference

A new host–virus reference database was curated from the transcriptome of the infected cells (Fig. 2). Repetitive sequences were removed using BBduk (BBtools 38.90)[63]. An additional long repetitive sequence was removed manually. A database of *E. huxleyi* and EhV genes, which were abundant in the samples[31], was also added to this reference to specifically detect *E. huxleyi* cells and avoid a non-specific alignment of reads from these cells to other contigs. For EhV, the predicted CDSs in the EhVM1 were used as a ref. 64. For the host, an integrated transcriptome reference of *E. huxleyi* was used as a ref. 65. Viral transcripts in the database were identified using a homology search against a customized protein database as described above. A pseudo-GTF file for the combined database was created using Bioawk v.11 (ref. 66). A reference was created from the database using the Cell Ranger mkref command. Raw reads were aligned to this reference database using 10x Genomics Cell Ranger v.5.0.0 count analysis.

Pre-processing of transcript abundance and dimensionality reduction

A total of 28,656 cells from the 10 samples were initially aligned to the reference database. Cells with zero UMIs and cells with the lowest 1% number of UMIs, compared with the distribution of transcripts per cell in the entire dataset, were removed for downstream analyses. To prevent cases of doublet or multiplet cells, which can be biological (cell digestion) or technical (fused cells), cells with the highest 1% number of UMIs were also removed. The raw UMIs of 28,015 cells were further pre-processed using the Python package scprep v.1.0.10: low-expressing genes were filtered with filter.filter_rare_genes and min_cells=2. This number was chosen because we did not want to include genes mapped to only one cell, but we also did not want to exclude low-expressed genes, because they might represent gene expression of low-abundant organisms. Expression was normalized by cell library size with normalize.library_size_normalize and the data were scaled with transform.sqrt. Pre-processing was wrapped in an in-house script (see 00.01.filter_normalize_scale_single_cell_data.py in the GitHub repository).

To represent the cells in two dimensions based on their gene expression profiles, dimensionality reduction was performed using scprep v.1.1.0 package PCA (method = 'svd', eps = 0.1, n_components = 50) and UMAP using the Python library umap-learn v.0.5.1 (minimum distance = 0.4 spread = 2, number of neighbours = 7). Dimensionality reduction was wrapped in an in-house script (00.02.dimentionality_reduction_single_cell_data.py).

Assigning taxonomy to each detected cell using rRNA homology search

To identify the taxonomy of each detected cell, reads from each cell were assembled independently. The taxonomy of the cells was determined by 18S rRNA homology to one of the following groups, which were abundant in the population: the classes Bacillariophyta (diatoms), Prymnesiophyceae, Chrysophyceae, MAST-3 and Katablepharidaceae, and the divisions Ciliphora (Ciliates), Dinoflagellata and Cercozoa. Other taxonomic groups were clustered under 'other eukaryotes'; 16,358 cells were identified in this way and 11,657 cells that could not be identified were excluded from the plot for convenience. Cells with 18S rRNA contigs homologous to more than one taxonomic group were also conservatively omitted. As described above, cells expressing at least ten viral UMIs were considered infected[23,24]. This section was wrapped in a Jupyter notebook (Coexpression_wrapper_extended.ipynb).

Identifying the *Leucocryptos* host and its virus using homology search

To better identify the detected Katablepharidaceae cells and to identify their infecting virus, 26 infected Katablepharidaceae cells from bag no. 4, day 20, were selected. Reads from these cells were retrieved using the UMI and then trimmed using TrimGalore v.0.6.5, a Cutadapt wrapper[50]. Trimming was wrapped in an in-house script (see pull_trim_clean.sh in the GitHub repository). Trimmed read files from all these cells were concatenated into one file and assembled altogether using rnaSPAdes v.3.15 (ref. 51). To identify the specific Katablepharidaceae host, assembled contigs were matched against the PR2 rRNA database using blastn at 90% identity, e value ≤ 10^{-10} and alignment length ≥100 bp. Contigs were best matched to an unknown Katablepharidaceae (>99% nucleotide identity), but, after removal of unidentified genera, these contigs best matched (>95% nucleotide identity) the Katablepharidaceae species *L. marina*. Transcripts that matched classes other than Katablepharidaceae were matched against the entire NCBI database using the NCBI web server[67]. They, too, mostly matched Katablepharidaceae genes, specifically 28S rRNA or internal transcribed spacer sequences (Supplementary Data Table 1). To identify the specific infecting virus, transcripts were matched against an NCLDV gene marker database[11] at 90% identity, e value ≤ 10^{-10} and alignment length ≥100 bp. After finding homology to *Leucocryptos* and the virus GVMAG-M-3300020187-27 (ref. 2), gene expression was calculated using RSEM v.1.3.1 (ref. 68) (rsem-calculate-expression -p10 --bowtie2–fragment-length-mean 58). The genomic features of the virus were taken from Schulz[2] and the viral genome was plotted using ShinyCircos v.2.0 (ref. 69). Gene expression in the plot is measured in expected counts after $\log_2$(transformation). The relative abundance data in Fig. 4 were obtained from an 18S rRNA amplicon sequencing on a size fraction of 2–20 µm in bag no. 4 during the mesocosm experiment[29]. Days 19, 22 and 23 were sampled twice; all other days were sampled once. In Fig. 4c, relative abundance is calculated per taxa as a fraction of all ASVs, excluding metazoans. Figure 4d shows the fraction of Katablepharidaceae out of all ASVs matching Katablepharidaceae (excluding metazoans). *E. huxleyi* abundance was measured by flow cytometry based on high side scatter and high chlorophyll signals. These data were obtained from the source data of the same study[29].

Phylogenetic tree of Katablepharidaceae ASVs and 18S rRNA genes

To verify the taxonomy of the ASVs, a phylogenetic tree was constructed of 89 ASVs identified as Katablepharidaceae, selected 18S rRNA sequences of Katablepharidaceae and other species from the PR2 database, and the longest single-cell assembled contig from the infected Katablepharidaceae cells. Sequences were aligned with ClustalOmega v.1.2.4 (default parameters)[70]. A diagnostic tree was first made with FastTree 2.1.10 (ref. 60) for pruning long branches before making the final tree with IQ-TREE[71]. All but three ASVs and one PR2 sequence clustered together with the assembled *Leucocryptos* transcript, verifying the taxonomy of 97% of the ASVs used in the relative abundance analysis (Extended Data Fig. 4).

Phylogenetic trees of viral HSPs and metacaspase

To examine the evolutionary history of the HSPs encoded in GVMAG-M-3300020187-27, phylogenetic trees of these proteins were constructed together with homologues present in eukaryotes, bacteria, archaea and other giant viruses. For this, a customized database of proteins from reference genomes was compiled from EggNOG v.5.0 (ref. 72) (eukaryotes), bacteria and archaea (the Genome Taxonomy Database (GTDB) v.95)[73] and other giant viruses (the Giant Virus Database[4]). For bacterial and archaeal genomes in the GTDB, proteins were predicted first with Prodigal v.2.6.3 (ref. 74) using default parameters. Proteins were searched against Pfam models for each protein using hmmsearch with the noise cutoff (--cut_nc) and subsequently aligned sequences with ClustalOmega v.1.2.3 (default parameters). Phylogenetic trees were constructed using IQ-TREE v.2.1.2 (ref. 71) (parameters m TEST -bb 1000 -T 6 --runs 10) using ultrafast bootstraps and with the best model determined with ModelFinder[75]. Substation matrixes used for the phylogenetic trees: Bax-1 – VT+F+R7; metacaspase – VT+R7; HSP90 – LG+F+R10; HPS70 – LG+F+R10.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Sequencing data have been deposited under NCBI Bioproject, accession no. PRJNA694552, Biosamples SAMN38317978–SAMN38317987. Additional data used in this paper, including UMI tables generated from 10x Cell Ranger, extended Blast result tables, assembled transcripts and other files that can be used to reproduce our results, are available at Dryad via https://doi.org/10.5061/dryad.s7h44j1c9 (ref. 76). Source data are provided with this paper. Public databases that were used in this manuscript include: the Giant Virus database https://faylward.github.io/GVDB; PR2 database https://pr2-database.org; metaPR2 database https://shiny.metapr2.org/metapr2; RefSeq v.207.

Code availability

All data management and analysis codes are open for review and reuse and archived online at GitHub via https://github.com/vardilab/host-virus-pairing (ref. 77).

References

1. Moniruzzaman, M. et al. Virologs, viral mimicry, and virocell metabolism: the expanding scale of cellular functions encoded in the complex genomes of giant viruses. *FEMS Microbiol. Rev.* **47**, 5 (2023).
2. Schulz, F. et al. Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–436 (2020).
3. Endo, H. et al. Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nat. Ecol. Evol.* **4**, 1639–1649 (2020).
4. Aylward, F. O., Moniruzzaman, M., Ha, A. D. & Koonin, E. V. A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLoS Biol.* **19**, e3001430 (2021).
5. Rosenwasser, S., Ziv, C., Creveld, S. Gvan & Vardi, A. Virocell metabolism: metabolic innovations during host–virus interactions in the ocean. *Trends Microbiol.* **24**, 821–832 (2016).
6. Fuhrman, J. A. Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541–548 (1999).
7. Irwin, N. A. T., Pittis, A. A., Richards, T. A. & Keeling, P. J. Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nat. Microbiol.* **7**, 327–336 (2022).

8. Moniruzzaman, M., Weinheimer, A. R., Martinez-Gutierrez, C. A. & Aylward, F. O. Widespread endogenization of giant viruses shapes genomes of green algae. *Nature* **588**, 1–5 (2020).
9. Nissimov, J. I. et al. Coccolithoviruses: a review of cross-kingdom genomic thievery and metabolic thuggery. *Viruses* **9**, 52 (2017).
10. Ha, A. D., Moniruzzaman, M. & Aylward, F. O. High transcriptional activity and diverse functional repertoires of hundreds of giant viruses in a coastal marine system. *mSystems* **6**, e0029321 (2021).
11. Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R. & Aylward, F. O. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nat. Commun.* **11**, 1710 (2020).
12. Bäckström, D. et al. Virus genomes from deep sea sediments expand the ocean megavirome and support independent origins of viral gigantism. *mBio* **10**, e02497–18 (2019).
13. Queiroz, V. F. et al. Amoebae: hiding in plain sight: unappreciated hosts for the very large viruses. *Annu. Rev. Virol.* **9**, 79–98 (2022).
14. Coutinho, F. H. et al. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nat. Commun.* **8**, 1–12 (2017).
15. Meng, L. et al. Quantitative assessment of nucleocytoplasmic large DNA virus and host interactions predicted by co-occurrence analyses. *mSphere* **6**, e01298–20 (2021).
16. Ciobanu, D. et al. A single-cell genomics pipeline for environmental microbial eukaryotes. *iScience* **24**, 102290 (2021).
17. Stepanauskas, R. et al. Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nat. Commun.* **8**, 84 (2017).
18. Brown, J. M. et al. Single cell genomics reveals viruses consumed by marine protists. *Front. Microbiol.* **11**, 2317 (2020).
19. DeLong, J. P., Van Etten, J. L., Al-Ameeli, Z., Agarkova, I. V. & Dunigan, D. D. The consumption of viruses returns energy to food chains. *Proc. Natl Acad. Sci. USA* **120**, e2215000120 (2023).
20. Gonzalez, J. M. & Suttle, C. A. Grazing by marine nanoflagellates on viruses and virus-sized particles: ingestion and digestion. *Mar. Ecol. Prog. Ser.* **94**, 1–10 (1993).
21. Imdahl, F. & Saliba, A.-E. Advances and challenges in single-cell RNA-seq of microbial communities. *Curr. Opin. Microbiol.* **57**, 102–110 (2020).
22. Mauger, S., Monard, C., Thion, C. & Vandenkoornhuyse, P. Contribution of single-cell omics to microbial ecology. *Trends Ecol. Evol.* **37**, 1–12 (2021).
23. Ku, C. & Sebé-Pedrós, A. Using single-cell transcriptomics to understand functional states and interactions in microbial eukaryotes. *Philos. Trans. R. Soc. B* **374**, 20190098 (2019).
24. Hevroni, G., Vincent, F., Ku, C., Sheyn, U. & Vardi, A. Daily turnover of active giant virus infection during algal blooms revealed by single-cell transcriptomics. *Sci. Adv.* **9**, 41 (2023).
25. Lax, G. et al. Multigene phylogenetics of euglenids based on single-cell transcriptomics of diverse phagotrophs. *Mol. Phylogenet. Evol.* **159**, 107088 (2021).
26. Cooney, E. C. et al. Single-cell transcriptomics of *Abedinium* reveals a new early-branching Dinoflagellate lineage. *Genome Biol. Evol.* **12**, 2417–2428 (2020).
27. Schwartz, M. et al. Molecular characterization of human cytomegalovirus infection with single-cell transcriptomics. *Nat. Microbiol.* **8**, 455–468 (2023).
28. Bost, P. et al. Host-viral infection maps reveal signatures of severe COVID-19 patients. *Cell* **181**, 1475–1488 (2020).
29. Vincent, F. et al. Viral infection switches the balance between bacterial and eukaryotic recyclers of organic matter during coccolithophore blooms. *Nat. Commun.* **14**, 1–17 (2023).
30. Blobel, G. & Potter, V. R. Studies on free and membrane-bound ribosomes in rat liver. I. Distribution as related to total cellular RNA. *J. Mol. Biol.* **26**, 279–292 (1967).
31. Vincent, F., Sheyn, U., Porat, Z., Schatz, D. & Vardi, A. Visualizing active viral infection reveals diverse cell fates in synchronized algal bloom demise. *Proc. Natl Acad. Sci. USA* **118**, e2021586118 (2021).
32. Lawrence, J. E., Brussaard, C. P. D. & Suttle, C. A. Virus-specific responses of *Heterosigma akashiwo* to infection. *Appl. Environ. Microbiol.* **72**, 7829 (2006).
33. Aylward, FrankO. et al. Taxonomic update for giant viruses in the order *Imitervirales* (phylum *Nucleocytoviricota*). *Arch. Virol.* **168**, 11 (2023).
34. Santini, S. et al. Genome of *Phaeocystis globosa* virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. *Proc. Natl Acad. Sci. USA* **110**, 10800–10805 (2013).
35. Gallot-Lavallée, L., Blanc, G. & Claverie, J.-M. Comparative genomics of *Chrysochromulina ericina* virus and other microalga-infecting large DNA viruses highlights their intricate evolutionary relationship with the established *Mimiviridae* family. *J. Virol.* **91**, 230–247 (2017).
36. Okamoto, N. & Inouye, I. The katablepharids are a distant sister group of the Cryptophyta: a proposal for Katablepharidophyta divisio nova/Kathablepharida phylum novum based on SSU rDNA and beta-tubulin phylogeny. *Protist* **156**, 163–179 (2005).
37. Vørs, N. Ultrastructure and autecology of the marine, heterotrophic flagellate *Leucocryptos marina* (Braarud) Butcher 1967 (Katablepharidaceae/Kathablepharidae), with a discussion of the genera *Leucocryptos* and Katablepharis/Kathablepharis. *Eur. J. Protistol.* **28**, 369–389 (1992).
38. Massana, R. et al. Phylogenetic and ecological analysis of novel marine Stramenopiles. *Appl. Environ. Microbiol.* **70**, 3528–3534 (2004).
39. Schoch, C. L. et al. NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database* **2020**, baaa062 (2020).
40. Knox, C., Luke, G. A., Blatch, G. L. & Pesce, E. R. Heat shock protein 40 (Hsp40) plays a key role in the virus life cycle. *Virus Res.* **160**, 15–24 (2011).
41. Gober, M. D. & Wales, S. Q. & Aurelian, L. Herpes simplex virus type 2 encodes a heat shock protein homologue with apoptosis regulatory functions. *Front. Biosci.* **10**, 2788–2803 (2005).
42. Yoshikawa, G. et al. Medusavirus, a novel large DNA virus discovered from hot spring water. *J. Virol.* **93**, 2130–2148 (2019).
43. Wilson, W. H. et al. Genomic exploration of individual giant ocean viruses. *ISME J.* **11**, 1736 (2017).
44. Machado, T. B. et al. Gene duplication as a major force driving the genome expansion in some giant viruses. *J. Virol.* **97**, e01309–e01323 (2023).
45. Jousset, A. et al. Where less may be more: how the rare biosphere pulls ecosystems strings. *ISME J.* **11**, 853–862 (2017).
46. Breitbart, M. & Rohwer, F. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* **13**, 278–284 (2005).
47. Mruwat, N. et al. A single-cell polony method reveals low levels of infected *Prochlorococcus* in oligotrophic waters despite high cyanophage abundances. *ISME J.* **15**, 41–54 (2020).
48. Zhong, K. X., Wirth, J. F., Chan, A. M. & Suttle, C. A. Mortality by ribosomal sequencing (MoRS) provides a window into taxon-specific cell lysis. *ISME J.* **17**, 105–116 (2022).
49. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
50. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetJ* **17**, 10 (2011).
51. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

52. Guillou, L. et al. The protist ribosomal reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* **41**, D597–D604 (2012).
53. Vaulot, D. et al. metaPR2: a database of eukaryotic 18S rRNA metabarcodes with an emphasis on protists. *Mol. Ecol. Resour.* **8**, 3188–3201 (2022).
54. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
55. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 1–9 (2009).
56. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
57. Frith, M. C. A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Res.* **39**, e23 (2011).
58. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).
59. Edgar, RobertC. Muscle5: high-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.* **13**, 1 (2022).
60. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
61. Van der Jeugt, F., Dawyndt, P. & Mesuere, B. FragGeneScanRs: faster gene prediction for short reads. *BMC Bioinform.* **23**, 1–8 (2022).
62. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
63. Bushnell, B. *BBMap: A Fast, Accurate, Splice-aware Aligner* (Lawrence Berkeley National Laboratory, 2014).
64. Fromm, A., Schatz, D., Ben-Dor, S., Feldmesser, E. & Vardi, A. Complete genome sequence of *Emiliania huxleyi* virus strain M1, isolated from an induced *E. huxleyi* bloom in Bergen, Norway. *Microbiol. Resour. Ann.* **11**, e0007122 (2022).
65. Feldmesser, E., Ben-Dor, S. & Vardi, A. An *Emiliania huxleyi* pan-transcriptome reveals basal strain specificity in gene expression patterns. *Sci. Rep.* **11**, 20795 (2021).
66. Li, H. Bioawk: awk modified for biological data. *GitHub* https://github.com/lh3/bioawk (2015).
67. Johnson, M. et al. NCBI BLAST: a better web interface. *Nucleic Acids Res.* **36**, W5–W9 (2008).
68. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinform.* **12**, 323 (2011).
69. Yu, Y., Ouyang, Y. & Yao, W. ShinyCircos: an R/Shiny application for interactive creation of Circos plot. *Bioinformatics* **34**, 1229–1231 (2018).
70. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **27**, 135–145 (2018).
71. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
72. Huerta-Cepas, J. et al. EggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
73. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
74. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).
75. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
76. Fromm, A. et al. Single-cell RNA-seq of the rare virosphere reveals the native hosts of giant viruses in the marine environment [Dataset]. *Dryad* https://doi.org/10.5061/dryad.s7h44j1c9 (2024).
77. Fromm, A., Hevroni, G., Aylward O. F. & Vardi, A. Host-virus pairing. *GitHub* https://github.com/vardilab/host-virus-pairing (2024).

Acknowledgements

We thank Adva Shemi and Roi Avraham for their comments and suggestions for this manuscript. We thank Talia Shaler for helping to explain the technical aspects of 10x sequencing. We thank all AQUACOSM VIMS-Ehux project team members for conducting the mesocosm experiment. The present study was supported by the Simons Foundation grant (no. 735079), 'Untangling the infection outcome of host–virus dynamics in algal blooms in the ocean' awarded to A.V., the National Science Foundation award (no. 2141862) to F.O.A. and National Institutes of Health grant (no. 1R35GM147290-01) awarded to F.O.A.

Author contributions

A.F., G.H., F.O.A. and A.V. designed and conceptualized the project and wrote the paper. A.F. and G.H. designed and wrote the scripts for data analysis. F.V. and D.S. collected the natural samples and prepared the single-cell transcriptomics libraries. F.O.A. and C.A.M.G. conducted phylogenetic analysis and viral homology search. A.F. conducted all other data analyses. All authors read and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41564-024-01669-y.

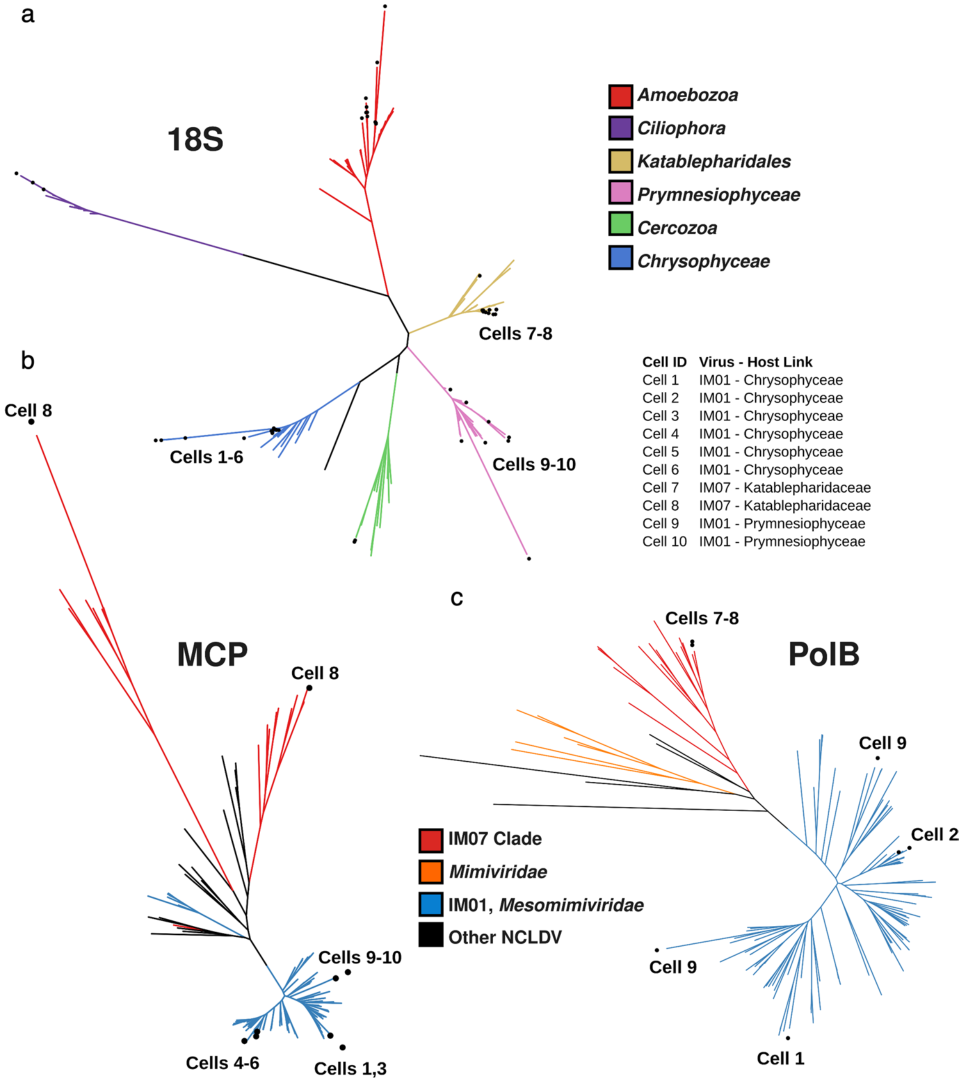
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-024-01669-y.

Correspondence and requests for materials should be addressed to Frank O. Aylward or Assaf Vardi.

Peer review information *Nature Microbiology* thanks Anne-Kristin Kaster, Hiroyuki Ogata and J. Cesar Ignacio-Espinoza for their contribution to the peer review of this work.

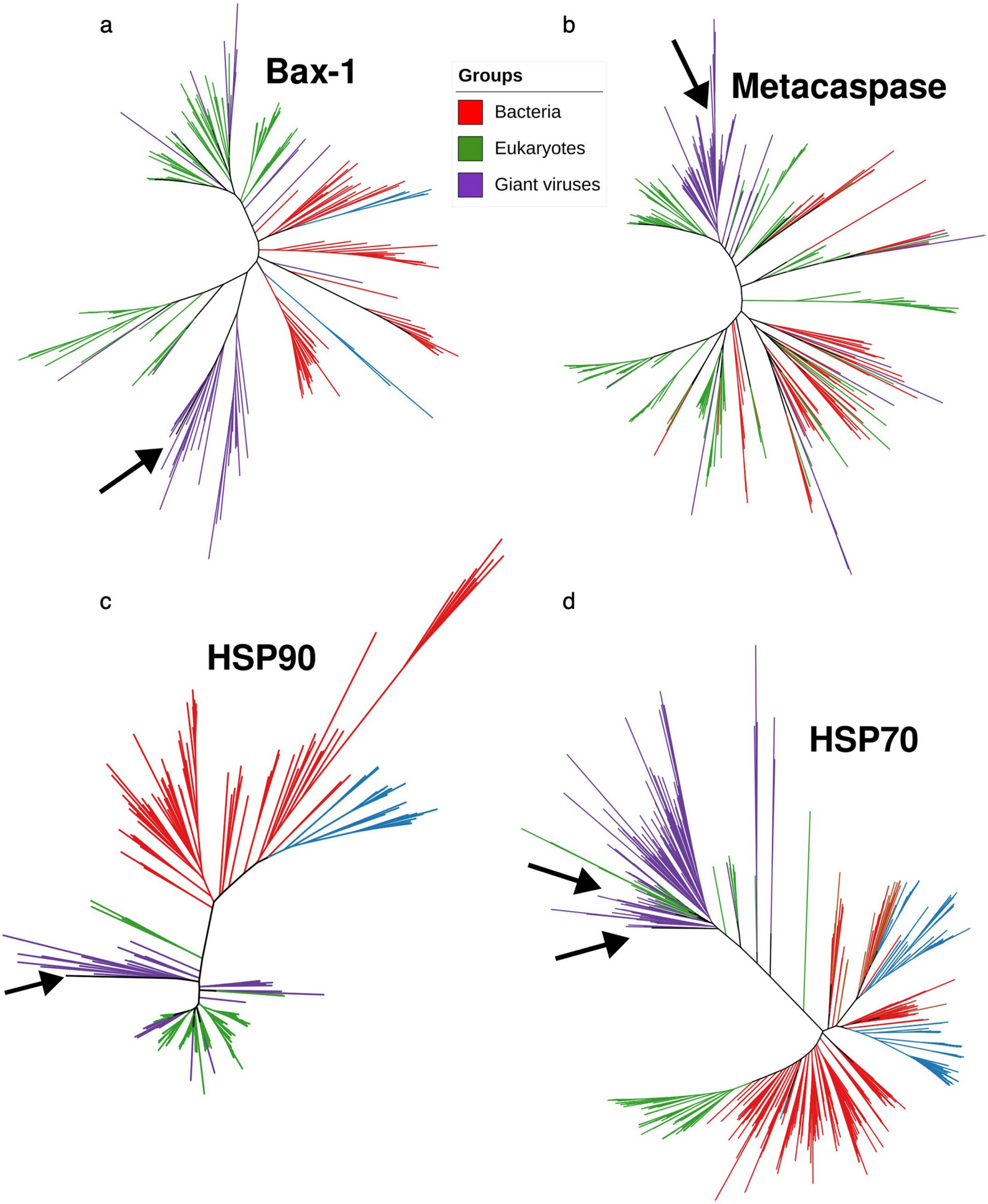
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

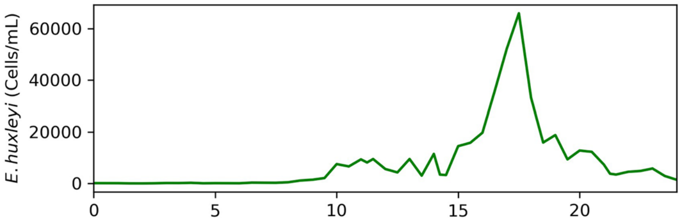


Cell ID	Virus - Host Link
Cell 1	IM01 - Chrysophyceae
Cell 2	IM01 - Chrysophyceae
Cell 3	IM01 - Chrysophyceae
Cell 4	IM01 - Chrysophyceae
Cell 5	IM01 - Chrysophyceae
Cell 6	IM01 - Chrysophyceae
Cell 7	IM07 - Katablepharidaceae
Cell 8	IM07 - Katablepharidaceae
Cell 9	IM01 - Prymnesiophyceae
Cell 10	IM01 - Prymnesiophyceae

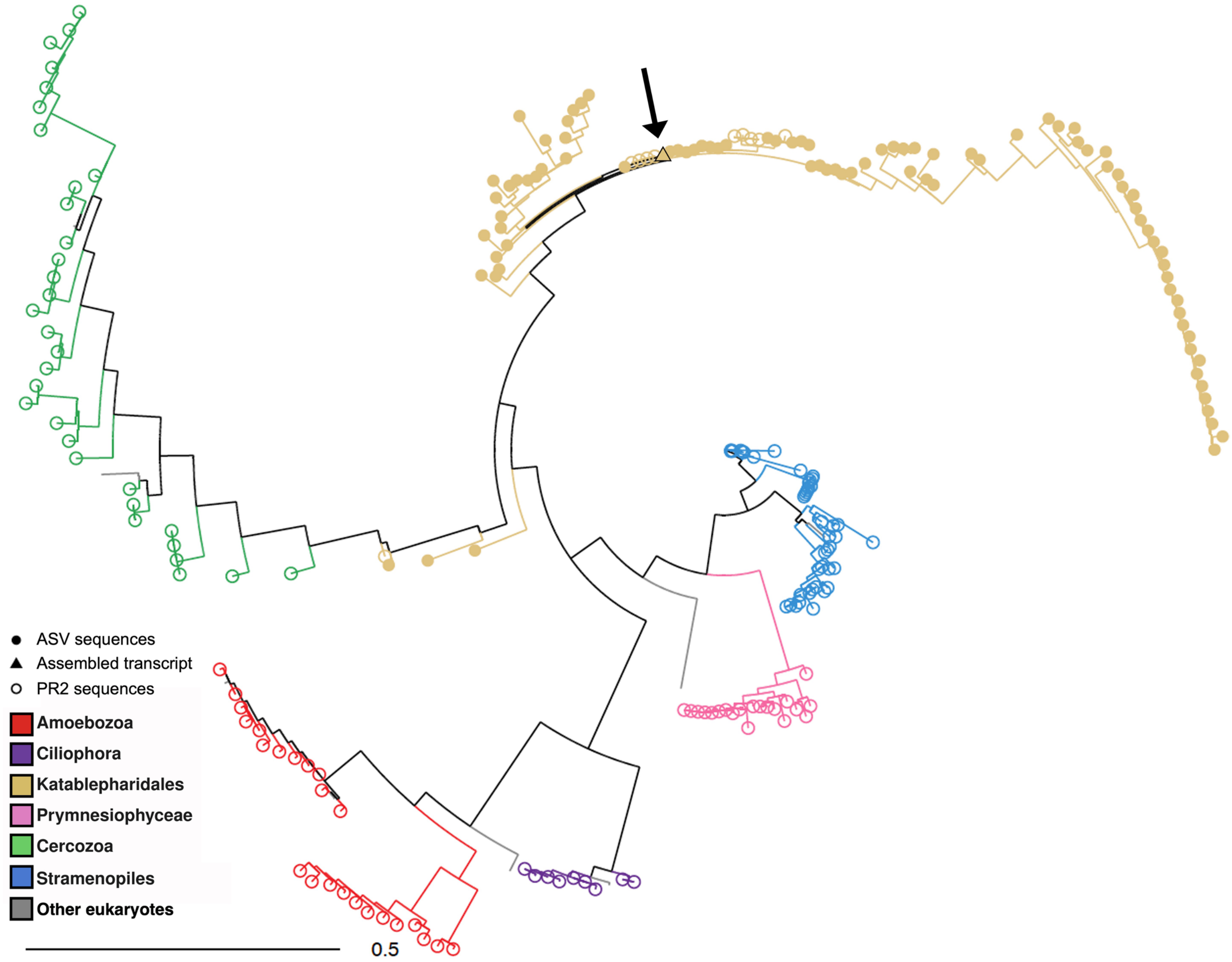
Extended Data Fig. 1 | Phylogenetic trees of giant virus marker genes assembled from the single-cell data. Points denote transcripts assembled from single-cell transcriptomes. Numbers denote cells for which transcripts are present in both viral and host trees. **a**, 18 S rRNA (host). **b**, Major Capsid Protein (virus). **c**, DNA-Polymerase family B (virus).



Extended Data Fig. 2 | Phylogenetic trees of functional genes present in the predicted *Leucocryptos* virus. The different colors represent bacteria (red), eukaryotes (green), or giant viruses (purple). Arrows point at the location of the predicted *Leucocryptos* virus genes. **a**, Bax-1 apoptosis inhibitor. **b**, Metacaspase. **c**, heat-shock protein 90. **d**, heat-shock protein 70.



Extended Data Fig. 3 | Cell abundance of calcified *Emiliania huxleyi* cells during bloom succession in the mesocosm experiment. Calcified *E. huxleyi* cell count in bag no. 4 was measured by flow cytometry based on high side scatter and high chlorophyll signals.



Extended Data Fig. 4 | Phylogenetic tree of Katablepharidaceae ASVs, 18 S rRNA sequences from PR2 database, and single-cell assembled *Leucocryptos* 18 S rRNA gene. The different colors represent the different taxonomic groups analyzed. Filled dots denote ASV sequences, while empty dots denote PR2 sequences. The arrow points at the location of the single-cell assembled *Leucocryptos* 18S rRNA gene (in a triangle).

nature portfolio

Corresponding author(s): Assaf Vardi
Frank O. Aylward

Last updated by author(s): Feb 18, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection: n.a

Data analysis: Bioinformatic pipeline and custom code are available at: https://github.com/vardilab/host-virus-pairing.
Other softwares used:
python v.3.8
10X Genomics Cell Ranger v.5.0.0
seqtk v. 1.2
cutadapt v.4.2
TrimGalore v. 0.6.5
rnaSPAdes 3.15
CD-HIT v. 4.6.6
sortmerna v.4.3.6
bioawk v.1
bcl2fastq v.2.20.0.422
LASTAL v. 959
EggNOG v. 5.0
FastTree 2.1.10
IQ-TREE v. 2.1.2
FragGeneScanRs v. 1.1.0
Holoviews v. 1.15.4
Muscle5

ClustalOmega v. 1.2.4
bowtie2 v.2.3.3
tantan
Prodigal v. 2.6.3
seqtk v.1.2
gcc v.9.2.0
spades v.3.15.0
BLAST+ v.2.11.0
BBtools v.38.90
HMMER3 v. 3.3.2
ModelFinder
scprep v. 1.0.10
RSEM v.1.3.1
ShinyCircos v. 2.0

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Sequencing data has been deposited under NCBI Bioproject PRJNA694552, Biosamples SAMN38317978-SAMN38317987. Additional data used in this paper, including UMI tables generated from 10X Cell Ranger, extended Blast result tables, assembled transcripts, and other files that can be used to reproduce our results, are available at Dryad: https://doi.org/10.5061/dryad.s7h44j1c9.
Public databases that were used in this manuscript include:
The Giant Virus database: https://faylward.github.io/GVDB/
The PR2 database: https://pr2-database.org/
The PR2 database: https://shiny.metapr2.org/metapr2/
RefSeq v. 207

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Not applicable
Reporting on race, ethnicity, or other socially relevant groupings	Not applicable
Population characteristics	Not applicable
Recruitment	Not applicable
Ethics oversight	Not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences □ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We obtained samples from natural plankton communities during a mesocosm experiment (doi.org/10.1038/s41467-023-36049-3) and sequenced them using single-cell RNA-seq. We then analyzed the samples for presence of viral infection. We identified the

taxonomy of the host and linked the host to their infecting virus. We measured the genes expressed by the infecting virus and examined the host-virus dynamics in a temporal resolution.

Research sample

10 samples were obtained during the AQUACOSM VIMS-Ehux mesocosm experiment in Raunefjorden near Bergen, Norway, in May 2018. Seven bags were filled with water from the fjord, containing natural plankton communities.

Sampling strategy

10 samples were collected from four bags for our analysis, representing different stages in the growth and demise of the main protists in the induced algal bloom, Emiliania huxleyi.

Data collection

For each sample, 2 liters of water were filtered by Daniella Schatz and Flora Vincent with a 20 µm mesh and collected in a glass. bottles.

Timing and spatial scale

The experiment was taken place for 24 days between 24th May (day 0) and 16th June (day 23). 10 samples were collected from four bags, as follows: From bag 3, on days 15 and 20. From bag 4, on days 13, 15,19, and 20. From bag 6, on day 17. From bag 7, on days 16, 17, and 18.

Data exclusions

No data was excluded from the analysis.

Reproducibility

The mesocosm experiment is a field experiment, and therefore by definition cannot be strictly reproduced.

Randomization

On the first day, instead of filling each bag completely one by one, the seven bags were gradually filled up to maximize homogeneity across bags and thus random distribution of microbial species across the seven mesocosm enclosures.

Blinding

Blinding is not relevant for this study as the data acquisition was done in the natural environment with clearly identified bags.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions

All the data regarding field conditions has been made available on Dryad doi.org/10.5061/dryad.q573n5tfr

Location

The mesocosm experiment AQUACOSM VIMS-Ehux was carried out in Raunefjorden at the University of Bergen's Marine Biological Station Espegrend, Norway (60°16"11N; 5°13"07E).

Access & import/export

All efforts to access, import, export and use samples or scientific equipments has been conducted in coordination with local scientific partners in absolute compliance with local regulation along with Nagoya Protocols.

Disturbance

The disturbance cause is that at the end of the mesocosm experiement, the content of the bags is released in the natural environment. The disturbance is minimized by the low volumes remaining (after 23 days sampling) and does not introduce any foreign agent in the water as everything in the bags was initiated from a natural community

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals

The study did not involve laboratory animals.

Wild animals	The study did not involve wild animals as all samples were pre-filtered with a 20-micrometer mesh.
Reporting on sex	*Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.*
Field-collected samples	Samples were collected and stored in a freezer at -80C at the mesocosm facility. Samples were exported to the Weizmann Institute in coordination with local scientific partners in absolute compliance with local regulations along with Nagoya Protocols.
Ethics oversight	All efforts to access, import, export and use samples or scientific equipments has been conducted in coordination with local scientific partners in absolute compliance with local regulation along with Nagoya Protocols.

Note that full information on the approval of the study protocol must also be provided in the manuscript.