Genomic diversity and global distribution of four new prasinoviruses from the tropical north Pacific

Anamica Bedi de Silva,[1] Shawn W. Polson,[2] Christopher R. Schvarcz,[1,3] Grieg F. Steward,[1,3] Kyle F. Edwards[1]

AUTHOR AFFILIATIONS See affiliation list on p. 19.

ABSTRACT Viruses that infect phytoplankton are an integral part of marine ecosystems, but the vast majority of viral diversity remains uncultivated. Here, we introduce four near-complete genomic assemblies of viruses that infect the widespread marine picoeukaryote *Micromonas commoda*, doubling the number of reported genomes of *Micromonas* dsDNA viruses. All host and virus isolates were obtained from tropical waters of the North Pacific, a first for viruses infecting green algae in the order Mamiellales. Genome length of the new isolates ranges from 205 to 212 kb, and phylogenetic analysis shows that all four are members of the genus *Prasinovirus*. Three of the viruses form a clade that is adjacent to previously sequenced *Micromonas* viruses, while the fourth virus is relatively divergent from previously sequenced prasinoviruses. We identified 61 putative genes not previously found in prasinovirus isolates, including a phosphate transporter and a potential apoptosis inhibitor novel to marine viruses. Forty-eight genes in the new viruses are also found in host genome(s) and may have been acquired through horizontal gene transfer. By analyzing the coding sequences of all published prasinoviruses, we found that ~25% of prasinovirus gene content is significantly correlated with host genus identity (i.e., *Micromonas*, *Ostreococcus*, or *Bathycoccus*), and the functions of these genes suggest that much of the viral life cycle is differentially adapted to the three host genera. Mapping of metagenomic reads from global survey data indicates that one of the new isolates, McV-SA1, is relatively common in multiple ocean basins.

IMPORTANCE The genomes analyzed here represent the first viruses from the tropical North Pacific that infect the abundant phytoplankton order Mamiellales. Comparing isolates from the same location demonstrates high genomic diversity among viruses that co-occur and presumably compete for hosts. Comparing all published prasinovirus genomes highlights gene functions that are likely associated with adaptation to different host genera. Metagenomic data indicate these viruses are globally distributed, and one of the novel isolates may be among the most abundant marine viruses.

KEYWORDS marine virus, Phycodnaviridae, viral ecology, Micromonas

A significant fraction of marine viral diversity is composed of viruses that infect phytoplankton, the diverse unicellular primary producers that perform the majority of marine photosynthesis processes (1). Viruses that infect phytoplankton affect biogeochemical processes by lysing their hosts, thereby shunting nutrients and energy to smaller microbial cells, and by shaping phytoplankton production through mortality, metabolic manipulation, and virus-driven trait evolution (2–5).

In this study, we focus on prasinoviruses, which are double-stranded DNA viruses that infect prasinophytes (6, 7). Prasinophytes are a diverse and paraphyletic group of green algae (8), and the *Prasinovirus* genus is made up of double-stranded DNA viruses in a lineage that falls within the order *Algavirales* (9). The prasinoviruses isolated

to date all infect members of the phytoplankton order Mamiellales, which are included under the generic term prasinophyte (10). This order of algae includes the three cosmopolitan genera *Bathycoccus*, *Ostreococcus*, and *Micromonas* (7, 11). The Mamiellales are ubiquitous in the sunlit ocean and are often major community members in both oligotrophic and eutrophic environments, typically dominating the picoeukaryotic fraction of primary producers under productive conditions (12, 13). These algae are known for their small size, which is exemplified by the species *Ostreococcus tauri*, considered to be the smallest free-living eukaryote at ~0.8 µm cell diameter. *Bathycoccus* and *Micromonas* are somewhat larger, but at ~2 µm in diameter (14–16), they are still small for eukaryotic cells.

The first isolated virus infecting a eukaryotic microalga was a *Micromonas* virus (17), and isolates of prasinoviruses have been used to study many topics, such as viral production and decay (18), marine viral gene content (e.g., [10, 19]), viral alteration of host nutrient uptake (20), consequences of host resistance to lytic infection (e.g., [21, 22]), and diel changes to the dynamics of viral infection (e.g., [23]). Prasinophytes and their viruses are useful model systems for understanding the biology and functional consequences of phytoplankton viruses because they are cosmopolitan, relatively abundant, and amenable to laboratory manipulation (13, 24, 25). In environmental samples from the surface ocean, prasinoviruses are typically among the most abundant members of the *Nucleocytoviricota*, which is a highly diverse phylum of large, eukaryote-infecting viruses with dsDNA genomes (26, 27). As of writing this work, 22 genomes of prasinovirus isolates have been published, including four *Micromonas* viruses, five *Bathycoccus* viruses, and 12 *Ostreococcus* viruses, and some of the key characteristics of prasinovirus gene content have been described (10, 19, 24). Genes shared among all published prasinovirus genomes (i.e., core genes) are largely associated with basic viral functions, such as DNA replication, transcription, and nucleotide metabolism (24). In contrast, non-core genes (i.e., genes present in at least one, but not all, prasinovirus genomes) include many involved in cellular metabolism, such as nutrient acquisition, photosynthesis, and carbohydrate metabolism (19, 24). These genes associated with diverse cellular processes are presumed to alter host metabolism to enhance viral replication, but few prasinovirus genes have been experimentally evaluated for function (20).

The isolation and characterization of virus-host systems is helpful because this provides empirical evidence of virus-host connections and allows for experimental studies of the ecology and (co)evolution of host and virus. Such information provides a valuable context for interpreting metagenomic data. Sequencing additional prasinovirus representatives would facilitate a better understanding of the scope of prasinovirus genetic diversity, as well as how genome content has diverged among prasinovirus clades. In the current work, we sequenced and analyzed the genomes of four *Micromonas commoda* viruses isolated from coastal and open-ocean waters around the island of Oʻahu, Hawaiʻi. We will refer to this suite of virus strains as "HiMcVs," an abbreviation of Hawaiʻi *Micromonas commoda* viruses. These four isolates represent the first prasinovirus genome assemblies from tropical Pacific waters, and they double the number of published *Micromonas* phycodnavirus genomes. Our analysis of these genomes has four aims: (i) establish the phylogenetic relationships of the new isolates to other prasinoviruses (ii) evaluate the gene content of the new isolates in relation to other prasinoviruses and two *Micromonas* hosts to characterize gene novelty, potential gene origins, and genomic diversity among co-occurring prasinoviruses with overlapping host range (iii) quantitatively compare gene content across all prasinovirus genomes to understand what processes might underlie adaptation of viruses to different host genera and (iv) evaluate the distribution of the new isolates in the global ocean using metagenomic surveys.

RESULTS AND DISCUSSION

Virus characteristics

The four HiMcVs had distinct but overlapping host ranges, based on lysis tests with seven putative or confirmed *Micromonas* strains isolated from Kāneʻohe Bay (UHM1060–UHM1065) and Station ALOHA (UHM1080), with each viral strain infecting either two (McV-KB1, McVKB4, and McV-SA1) or as many as six (McV-KB2 and McV-KB3) of the seven algal strains tested (Fig. 1). The most permissive strain, UHM1065, was infected by four of the five viruses tested (all but McV-KB1), and the least permissive, UHM1061, was infected by only one (McV-KB2). Electron micrographs of the HiMcV isolate from Station ALOHA (McV-SA1) revealed an icosahedral morphology and a virion capsid diameter of ~142–160 nm, consistent with other prasinovirus isolates (6; Fig. 2).

Genome analyses

The four HiMcV draft genomes we assembled appear to be complete or near-complete based on nucleotide length, the number of predicted genes, and whole-genome alignments comparing these genomes to each other and to the MpV1 genome. The assemblies range from 205 to 212 kbp, with the largest genome belonging to McV-KB4 (Table 1). The McV-KB2 (GenBank Accession: PP911589) genome G + C content (44.9%) was notably higher than that of the other three genomes (41%–41.3%) and its coding density lower (1.15 vs 1.28–1.32 gene/kbp). To help assess genome completeness, we created a Mauve alignment with the published genome of MpV1 (NC_014767), the virus most closely related to McV-KB3 (PQ109088), McV-KB4 (PQ359806) and McV-SA1 (PQ381123; as described in the next section, phylogeny). These four genomes exhibited a high degree of synteny, although MpV1 is shorter in total by 20 to 26 kbp (Fig. 3A). It appears that MpV1 has one major ~11 kbp inversion, relative to the other three genomes, toward the center of its genome. A Mauve alignment using only McV-KB3, McV-KB4, and McV-SA1 showed that these three genomes have high structural similarity (Fig. 3B). Inclusion of McV-KB2 in Mauve alignments resulted in a large number of Locally Collinear Blocks (LCBs), represented as colored blocks in Fig. 3C, which indicated that there were substantial differences in genome organization between McV-KB2 and the other HiMcVs. Overall, the organization of the HiMcV genomes is consistent with previous findings that prasinoviruses exhibit a high degree of conservation in genome structure (24), with the notable exception of McV-KB2, the divergence of which is also reflected in its distinctive G + C content and coding density.
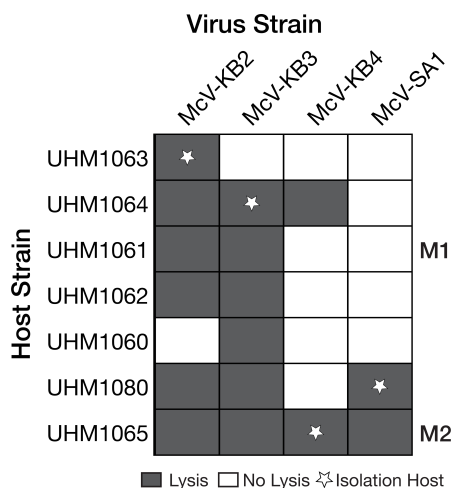

FIG 1 Host range of virus strains used in this study. UHM culture collection IDs for the HiMcVs are listed across the top, with IDs for the seven *Micromonas* host isolates listed on the left-hand side. Shortened names for cell strains UHM1061 and UHM1065 are listed on the right-hand side. Dark gray squares indicate successful lytic infection. Stars represent the host from which each virus strain was isolated.
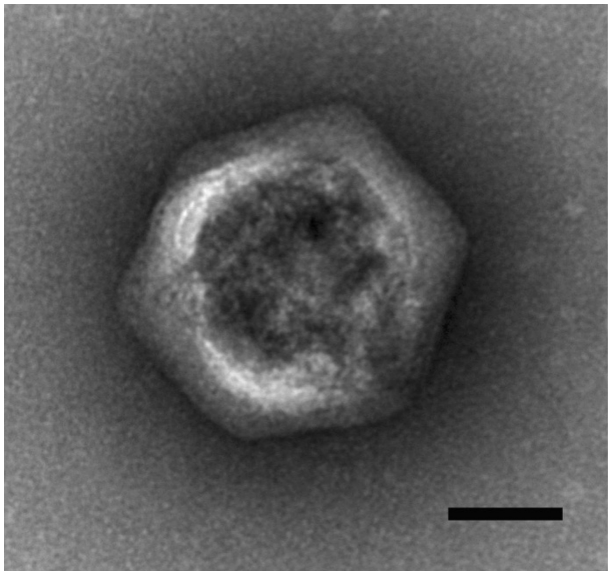
FIG 2 Electron micrograph of the McV-SA1 particle. Scale bar = 50 nm.

Phylogeny

The species tree derived from the concatenation of shared prasinovirus and chlorovirus genes shows that McV-KB3, McV-KB4, and McV-SA1 group into a clade, with McV-KB4 and McV-SA1 most closely related to each other, and the clade of these three HiMcVs lies within a larger clade containing all previously published *Micromonas*- and *Ostreococcus*-infecting virus genomes (Fig. 4). Within this larger clade, the *Ostreococcus*-infecting viruses form a monophyletic group, while the *Micromonas*-infecting viruses are paraphyletic, consistent with previous analyses. McV-KB2 is quite divergent from the other HiMcVs as its closest relative is the clade of *Bathycoccus*-infecting viruses, although it diverged from that group soon after their common ancestor arose from the last common ancestor of all the prasinoviruses. The divergence of McV-KB2 and its grouping with *Bathycoccus* viruses is consistent with the polB phylogeny, as well as the OrthoFinder/STAG-generated species tree (Fig. S2 and S3). Although McV-KB2 is relatively divergent from the other HiMcVs, it should be noted that it, nonetheless, overlaps in host range with each of the other three viruses (Fig. 1). We did not assess whether *Bathycoccus* strains in culture collections can be infected by the HiMcVs or whether known *Bathycoccus* viruses can infect the *Micromonas commoda* strains from Kāneʻohe Bay. As of this writing, no known prasinovirus infects a prasinophyte belonging to a genus other than the one on which it was isolated (see [10] and references therein), although Thomy et al. (28) found evidence of potential host-switching among prasinoviruses in phylogenies created with metagenome-assembled genomes. When considering the evolution of McV-KB2, it should be noted that the *Micromonas*-infecting viruses may be polyphyletic, based on our phylogeny, and some divergences among them are older than the origins of the *Bathycoccus*- and *Ostreococcus*-infecting viruses, both of which are monophyletic. Therefore, the overlap in host range between McV-KB2 and other McVs may simply reflect an older origin of *Micromonas*-infecting viruses, compared to the other prasinoviruses.

TABLE 1 Characteristics of genome assemblies of four Hawaiʻi *Micromonas commoda* virus strains

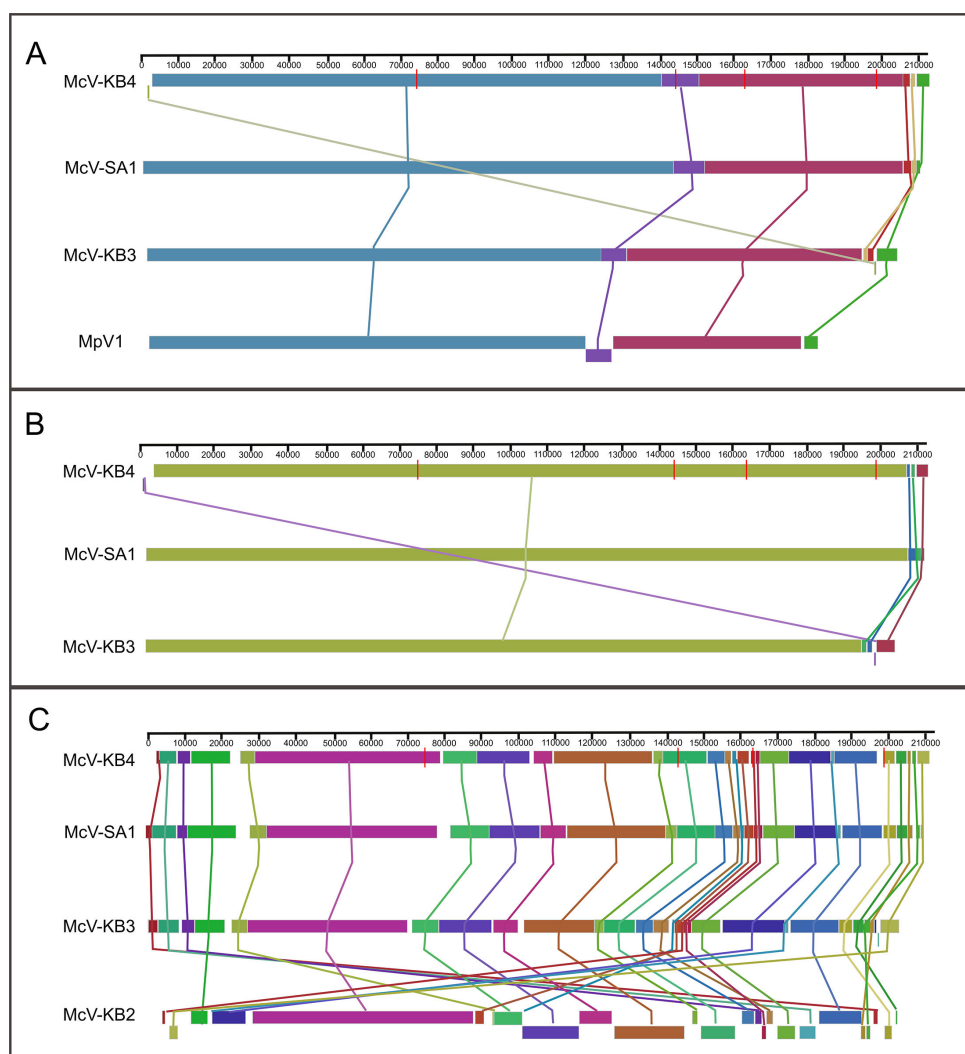| Assembly	Assembly size (bp)	# CDS	GC%	Gene density (gene/kbp)
McV-KB4	212,418	272	41	1.28
McV-SA1	210,087	270	41.2	1.29
McV-KB3	204,582	271	41.3	1.32
McV-KB2	210,100	242	44.9	1.15

FIG 3 Whole-genome alignments created with the progressiveMauve algorithm. Colored blocks represent regions conserved across genomes. (A) An alignment of McV-KB3, McV-KB4, and McV-SA1 with the published genome of *Micromonas pusilla* virus MpV1. (B) An alignment of McV-KB3, McV-KB4, and McV-SA1. (C) An alignment of all four HiMcV assemblies. Note that the inclusion of McV-KB2 generated a greater number of colored blocks because that genome possesses many inversions and rearrangements relative to the other three.

HiMcV core genes

Our OrthoFinder analysis found 344 orthogroups that were present within at least one of the HiMcV genomes (Table 2). The four HiMcVs share 152 orthogroups (i.e., the core HiMcV orthogroups), while the other 192 orthogroups are present in three or fewer genomes (Fig. 5). One hundred and seventeen orthogroups were found in all *Micromonas* virus genomes analyzed, with 56 found in all prasinovirus genomes examined in this study (Table 2). There were 48 orthogroups found in at least one HiMcV that were also found in at least one of the two host genomes, M1 and M2 (Table 2).

Among the HiMcVs, McV-KB2 is the most distinct, with 46 unique orthogroups (Fig. 5). McV-KB3 has the second highest number of unique orthogroups, which is in concordance with the phylogenetic distances between the HiMcVs (Fig. 4). Host range of the HiMcVs (Fig. 1) did not obviously map onto phylogenetic relatedness or similarity of gene content, perhaps not surprising because of the small number of genomes in our data set ($n = 4$). McV-KB2 and McV-KB3 both have relatively broad host ranges and share five host strains out of the six that each can infect (Fig. 1), but these viruses are distantly related

and share only 163 orthogroups (Fig. 4 and 5). The two most closely related HiMcVs, McV-KB4 and McV-SA1 (Fig. 4), share 208 orthogroups (Fig. 5) These viruses are similar in having a relatively narrow host range (2 hosts lysed out of 7 total), although they only share one host strain out of the two that each can infect (Fig. 1).

All 152 core HiMcV orthogroups contain genes that occur in at least one other prasinovirus (Table S5). The core HiMcV orthogroups with functional annotation encompass a diverse suite of viral biology, such as virion structure (major capsid protein), genome replication (DNA polymerase and DNA primase), RNA processing and transcription (mRNA capping enzyme, transcription factors, and RNAse H), breakdown of host polymers (nucleases and proteases), nucleotide metabolism (dUTPase, dCMP deaminase, thymidylate synthase, and ribonucleotide reductase), carbohydrate metabolism (mannitol dehydrogenase), lipid metabolism (glycerophosphoryl diester phosphodiesterase and phospholipase), and glycosylation (glycogen phosphorylase B and nucleotide-diphospho-sugar transferases). Similar to previously sequenced prasinoviruses, our strains do not contain multi-unit RNA polymerases, which are common among other members of *Nucleocytoviricota* (9). All four HiMcVs possess a cAMP-dependent Kef-type K + transporter that is found in M1 and M2, as well as *Micromonas* virus strains SP1 and MpV1. This transporter is known in bacteria to help protect cells from electrophilic compounds (29). Potassium channels are the most common type of membrane transporters encoded by viral genomes, with substantial diversity, and likely multiple
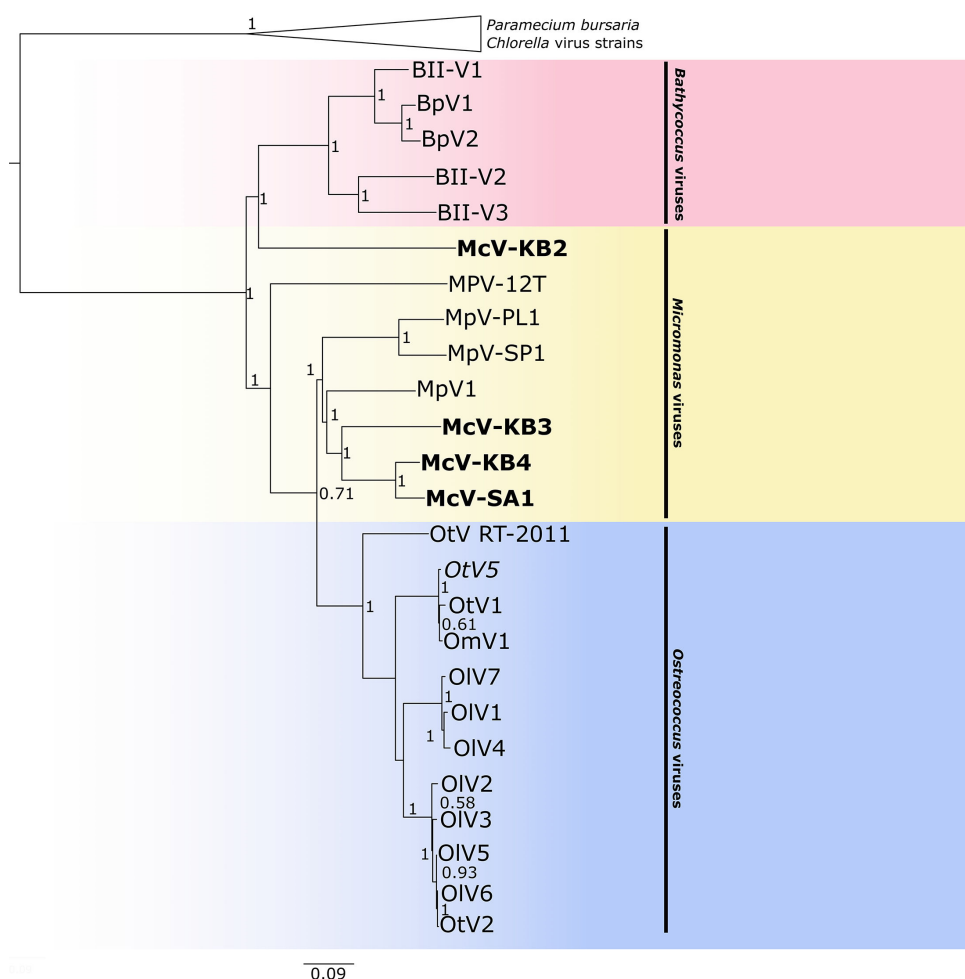

FIG 4 Prasinovirus species tree from concatenated amino acid alignments of the 26 orthogroups shared by all prasinoviruses and chloroviruses in the analysis. Generated with FastTree accessed through Geneious (default settings). Scale bar indicates substitutions per site. The HiMcV isolates are highlighted with bold text.

origins (30). In the Chlorella viruses PBCV-1 and NY-2A, a potassium channel depolarizes the cell membrane during infection, preventing infection of the cell by multiple viruses (31). Thus, the Kef-type potassium transporter may play a significant role in the infection strategy of HiMcVs.

Another noteworthy core HiMcV orthogroup is annotated as a high-light inducible protein in refseq_protein and as chlorophyll a-b binding protein in InterPro. This orthogroup occurs in M1 and M2 hosts and is present in the assembly for *Micromonas* virus MpV1. Chlorophyll a-b binding proteins are found in the light harvesting complex of photosystem II and play an important role in regulating excitation energy under fluctuating light levels (32). The high-light-inducible/chlorophyll a-b binding protein has been found in other prasinoviruses (order *Algavirales*), in Chrysochromulina ericina virus CeV-01B (order *Imitervirales* (27), and a variety of giant virus metagenome-assembled genomes (33, 34). Similar high-light-inducible proteins in cyanobacteria are activated in the presence of excessive photon energy, which suggests that viruses may also use this protein to protect the cell from photodamage during the viral infection cycle (35). In the microalgae *Dunaliella salina,* iron deficiency induces a chlorophyll a-b binding protein, suggesting changes in the expression of this protein may be a general response to stress rather than tied solely to light availability, which supports the hypothesis that such proteins would be involved in virus-induced stress cascades (36).

A putative PhoH-like phosphate starvation-inducible protein is also among the HiMcV core sequences. This protein has been seen previously in some, but not all, prasinoviruses (37) and is common among marine phages (38). Phosphate limitation may pose a particular challenge for viruses because they tend to have a higher stoichiometric P content than their hosts (39). Viruses may therefore benefit from enhanced P supply even if the host is not P-limited. While genes in this family are common in eukaryotic phytoplankton, it appears that our host genomes do not contain orthologs to the PhoH-like sequences found in our HiMcVs. Previously studied prasinovirus versions of this gene appear to be host-derived, which may mean that the HiMcVs obtained it from other *Micromonas* hosts in the environment (37) or possibly from other viruses during coinfection, the mechanism proposed for intein acquisition in prasinoviruses (40).

A final notable gene shared among all four HiMcVs is an alternative oxidase. Within this orthogroup, McV-KB2, McV-KB3, and McV-SA1 had one ortholog each, while McV-KB4 had three orthologous sequences, the host strain M1 had one sequence, and M2 had two sequences. Alternative oxidases have been previously found in cyanophages (41) as well as metagenome-assembled prasinovirus genomes (28), and these enzymes are thought to reduce photodamage to the electron transport chain under stressful conditions, which, in the case of infections, may arise from viral inhibition of photosystem I and ferredoxin NADP +reductase (FNR) (42). Our analysis found alternative oxidases in only one other prasinovirus isolate beyond the HiMcVs, the coastal Mediterranean Ostreococcus tauri virus RT-2011.

TABLE 2 Summary of OrthoFinder results comparing prasinoviruses, chloroviruses, and *Micromonas* hosts M1 and M2[a]

	No. of orthogroups
Total prasinovirus	693
Total HiMcV	344
Core HiMcV	152
Core *Ostreococcus* viruses	129
Core *Micromonas* viruses	117
Core *Bathycoccus* viruses	99
Core prasinovirus	56
Core prasinovirus +chlorovirus	26
HiMcV not found in other prasinoviruses	61
HiMcV shared with host	48

[a]Here, the term "core" refers to orthogroups that are present in 100% of genomes within the indicated taxa.
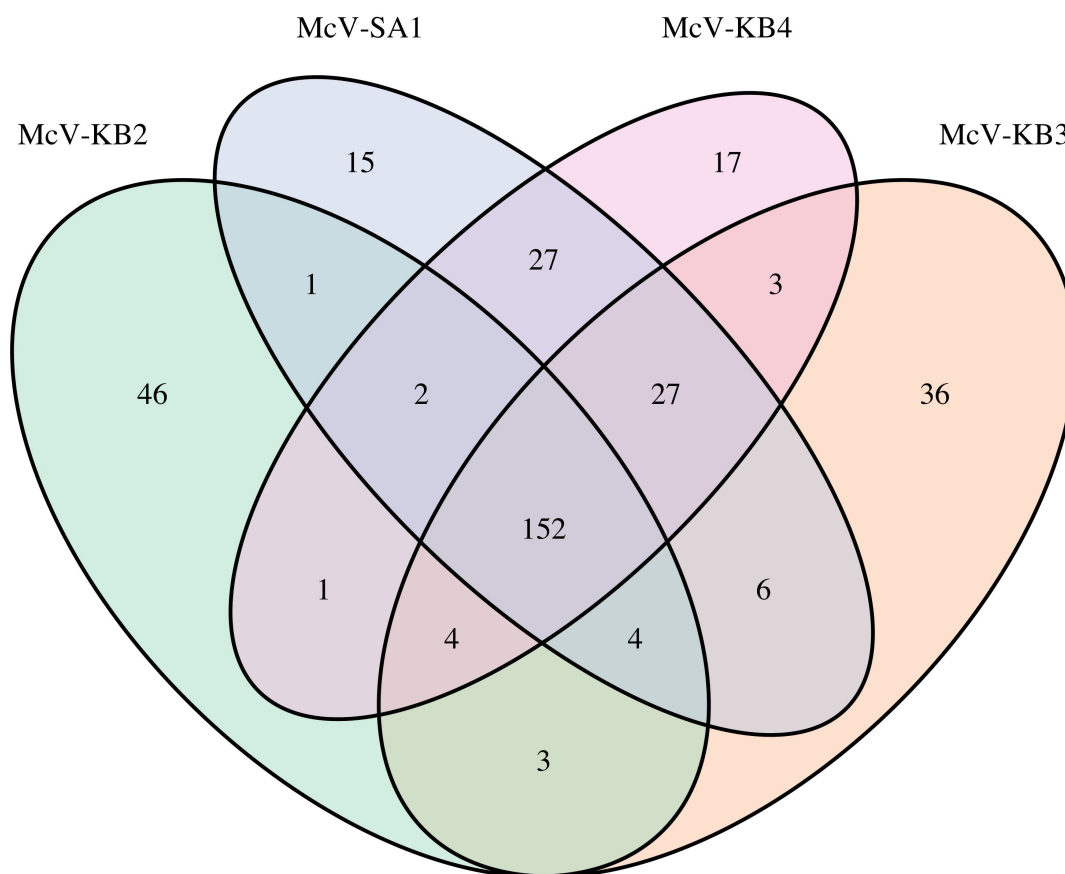
FIG 5 Venn diagram of the number of orthogroups shared by and unique to the four HiMcVs.

In the gene tree resulting from our alignment of diverse alternative oxidases, including ubiquinol oxidases (AOX) and plastoquinol terminal oxidases (PTOX), all prasinovirus sequences appear in the same clade (Fig. 6). The closest relatives of this clade include *Micromonas* isolates, including one of the sequences from M2, although the node joining these clades has only moderate bootstrap support (69%). The *Micromonas*-prasinovirus clade is sister to a clade containing all cyanophage and *Prochlorococcus* sequences, but this node also has moderate support (60%). The next most closely related clade contains *Ostreococcus* and *Bathycoccus* sequences, and its placement has 82% support. The remainder of the tree suggests two salient features of AOX/PTOX evolution. First, AOX and PTOX represent an ancient gene duplication, potentially predating the origin of photosynthetic eukaryotes, which is consistent with a previous analysis using only terrestrial plant sequences (43). Second, the PTOX lineage itself had an ancient duplication event, which is suggested by one algal +plant clade in the top middle portion of the figure and a second algal +plant PTOX clade in the left side of the figure, with a cyanobacterial PTOX clade branching between these two eukaryote clades. Both of the algal +plant clades include Mamiellales sequences, with an M2 *Micromonas commoda* sequence in both clades.

The implications of the phylogeny for viral PTOX evolution are not entirely clear because support for the placement of the cyanophage and prasinovirus clades is not very strong. The topology of Fig. 6 suggests that cyanophages and prasinoviruses both acquired PTOX from a Mamiellales cellular ancestor, and that *Prochlorococcus* acquired its PTOX from its phage. However, we cannot rule out other scenarios, such as cyanophages acquiring their PTOX directly from a prasinovirus source. Interestingly, a previous analysis using only cyanophage and cyanobacteria sequences also found that cyanophage +*Prochlorococcus* sequences formed a clade that was highly divergent from

Synechococcus sequences and that the *Prochlorococcus* sequences formed a clade that lies within a polyphyletic clade of cyanophage sequences (41).

Notable non-core and unique HiMcV genes

There are 61 non-core HiMcV orthogroups that represent proteins not previously found in prasinovirus isolates (verified through a secondary nr BLAST search), which we will refer to as "unique" (Table S6). Limited functional information is available for the 61 unique orthogroups as 40 have no BLAST hits against refseq_protein, eight have BLAST hits to hypothetical proteins, and the remaining 13 include many with low-identity (< 30% amino acid identity) BLAST hits. Results from the InterPro databases are comparable, although many of the proteins are predicted to be membrane-bound. One notable unique orthogroup is a putative phosphate:sodium symporter in McV-SA1 that is also found in both M1 and M2 host genomes. Top BLAST hits for this McV-SA1 CDS in both nr and refseq_protein are from previously published strains of *Micromonas*, including RCC299, a pelagic strain from the equatorial Pacific, and CCMP1445, a coastal strain from the North Atlantic. This CDS has no orthologs among the other prasinovirus isolate genomes, which may indicate that the McV-SA1 symporter gene has a recent cellular origin via horizontal gene transfer. A recent study of diverse metagenome-assembled prasinoviruses from the South China Sea also documented a phosphate:sodium symporter that occurred rarely in prasinoviruses (28). It should be noted that our data set includes a second orthogroup with prasinovirus phosphate transporter sequences that appear distinct from the McV-SA1 symporter as an alignment of the two orthogroups indicated low sequence identity, albeit with scattered matching residues (results not shown). This second orthogroup includes phosphate transporter sequences from BpV1 (HM004431), BII-V1 (MK522034), OlV6 (HQ633059), OlV5 (NC_020851), OtV2 (FN600414), OlV4 (JF974316), OlV2 (NC_028091), and OlV1 (MK514405), as well as sequences from M1 and M2. This orthogroup corresponds to the prasinovirus phosphate transporters
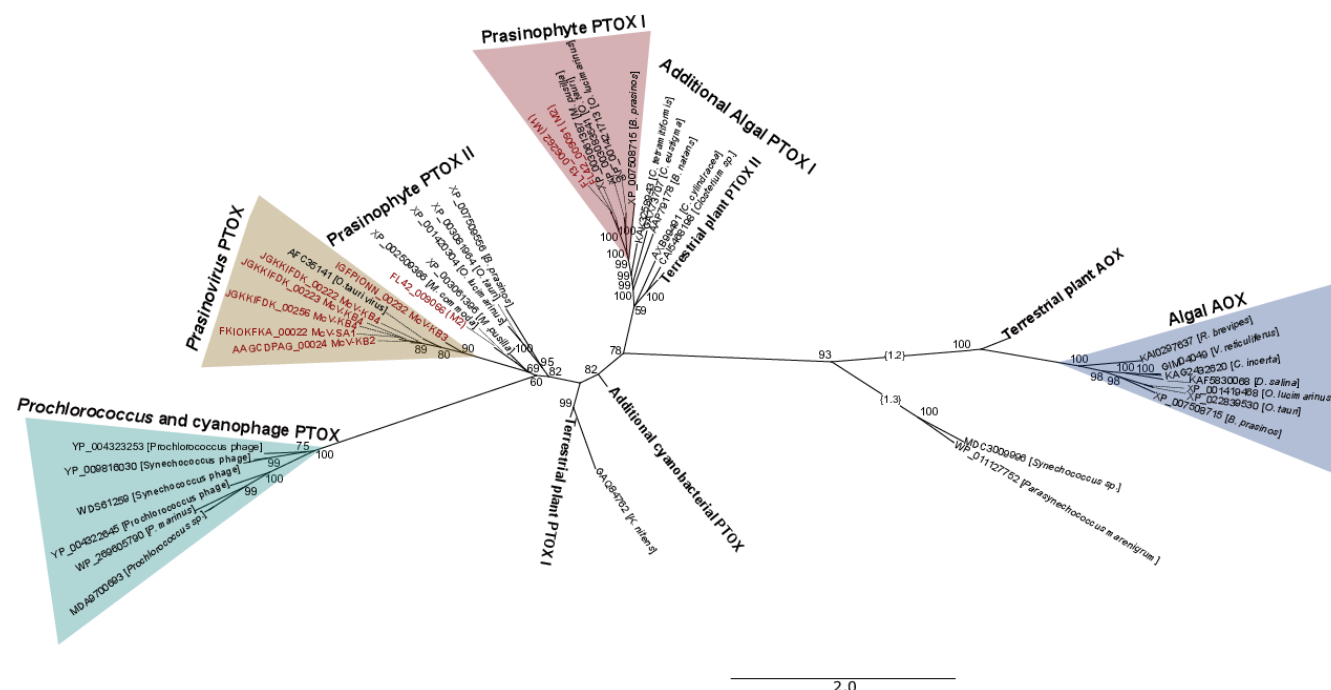

FIG 6 Alternative oxidase/plastoquinol terminal oxidase gene tree, including mitochondrial ubiquinol oxidase (AOX) from plants and eukaryotic phytoplankton and plastoquinol terminal oxidase (PTOX) from plants, eukaryotic phytoplankton, cyanobacteria, prasinoviruses, and cyanophages. Some branches were extended with dotted lines to make labels legible. Two long branches on the right-hand side of the plot were shortened, with numbers in curly brackets indicating total branch length. Gene sequences were aligned in MAFFT and trimmed with GoAlign, and a tree was created in IQ-TREE. Node support values reflect bootstrap values. UHM strains are shown in bolded maroon font. Scale bar indicates substitutions per site.

from the PHO4 superfamily identified by Monier et al. (37) in a comparison of phytoplankton virus phosphate transporters. The PHO4 transporters correspond to InterPro family IPR001204, whereas the novel McV-SA1 transporter corresponds to InterPro family IPR003841. Therefore, the McV-SA1 phosphate:sodium symporter likely indicates a separate acquisition of a phosphate transporter by prasinoviruses, potentially with different uptake affinity or other physiological differences.

Other unique HiMcV orthogroups with predicted functions include a putative bax inhibitor-1 (McV-KB2), N-6 DNA methylase (McV-KB2), polyamine aminopropyltransferase (McV-KB3), adenosylmethionine decarboxylase (McV-KB3), and a glycosyltransferase (McV-SA1). Bax inhibitor-1 is a conserved inhibitor of programmed cell death, and viruses such as deerpox (44) and cytomegalovirus (45) are known to encode other proteins that suppress bax, thereby countering the elimination of infected cells by apoptosis. Top BLAST hits to the McV-KB2 bax inhibitor-1 gene are sequences from fungi, but the amino acid identity is low (30%–35%), making the source of the McV-KB2 gene unclear. The putative adenosylmethionine decarboxylase and polyamine aminopropyltransferase encoded by McV-KB3 may catalyze linked steps in the synthesis of spermidine from putrescine. Spermidine is a polyamine required for cell growth as well as the replication of many viruses. Enzymes related to spermidine synthesis have been found in a variety of phages and eukaryotic viruses (46), including the chlorovirus PBCV-1 (47), but the McV-KB3 genes appear to be the first reported occurrence in a prasinovirus isolate. The two McV-KB3 genes have BLAST hits to genes from various bacteria and archaea, although the relatively low amino acid identity (~30%–40%) provides little information about the proximate origin of the genes. The other unique orthogroups with putative functions (N-6 DNA methylase and a glycosyltransferase) represent categories of enzymes that are commonly encoded by prasinoviruses, although the uniqueness of these orthogroups indicates these specific genes are not closely related to previously documented prasinovirus genes. A final notable unique orthogroup is a gene found only in McV-KB4 that is orthologous to cyanophage genes of unknown function. This putative CDS has not been seen in other prasinoviruses, and the closest database hit is a hypothetical protein from *Prochlorococcus* phage strain P-SSM2, which infects low-light *Prochlorococcus* ecotypes. The orthogroup may be evidence of gene exchange between cyanophage and a eukaryotic virus.

While none of the unique orthogroups have been previously seen in isolates, most (42 of 61) have homologs in the GOEV database. Eight of these genes belong to McV-KB2, ten to McV-KB3, and one to McV-SA1. Seventeen of these genes have limited information from InterPro databases, e.g., indicating a membrane-bound protein and/or a coil. Of the forty-two unique orthogroups with GOEV hits, thirty-five were homologous to MAG-derived GOEV gene sequences classified as belonging to the genus *Prasinovirus,* while the remaining fell into various taxa in the phylum *Nucleocytoviricota,* with two sequences from genomes in the *Imitervirales* order.

HiMcV orthogroups shared with their host genomes

A total of 48 HiMcV orthogroups contain genes also found in the M1 and/or M2 host genomes, 27 of which are shared among all 4 HiMcVs. In total, M1 shared 27 orthogroups with all four HiMcV strains, with 19 additional orthogroups shared between M1 and at least one other virus (Fig. S4). Results from comparison with M2 were similar, with 24 orthogroups shared between M2 and all four viruses, and 22 orthogroups shared between M2 and at least one virus (Fig. S5). The number of shared orthologs is comparable to results from Moreau et al. (24), in which *Micromonas pusilla* virus MpV1 shared 56 CDS with *Micromonas* sp. strain RCC1109. Forty-five of the HiMcV-host shared orthogroups contain genes found in previously published prasinoviruses, as made apparent by refseq_protein and nr BLAST hits. Some of these orthogroups were described in the section "HiMcV core genes" (chlorophyll a-b binding protein, PTOX, cAMP-dependent Kef-type K + transporter). The orthogroups shared with hosts are associated with a variety of cellular processes such as protein

modification/regulation/processing (N-myristoyltransferase, ubiquitin, cysteine protease, and ATP-dependent metalloprotease FtsH), glycosylation (nucleotide-sugar epimerases and transferases), amino acid synthesis (dehydroquinate synthase), nucleotide metabolism (dCMP deaminase and thymidine kinase), transcription regulation (transcription factors), stress response (heat shock protein 70, rhodanese, superoxide dismutase, and mannitol dehydrogenase), nucleic acid processing (exonuclease, ribonucleotide reductase, and DNA polymerase family X), photosynthesis (PTOX and chlorophyll a/b binding protein), and perhaps countering host defenses (methyltransferases) (Table S7).

There are three HiMcV orthogroups not found in other prasinoviruses that are found in both hosts, which include the aforementioned phosphate:sodium symporter found in McV-SA1, as well as two orthogroups shared with McV-KB2. One of the latter contains only hypothetical protein sequences with no hits in NCBI or InterPro databases. The other contains sequences that are annotated as chlorophyllide *a* oxygenase (CAO) for M1 and M2. CAO converts chlorophyll *a* to chlorophyll *b*, an important accessory pigment in green algae (48), which may mean that CAO supports light adsorption during infection by McV-KB2. However, amino acid identity between the McV-KB2 and host sequences is only 17.89%, so it is possible these sequences are not truly orthologous. If McV-KB2 indeed encodes for CAO, it would be the first virus reported to have this gene.

Genes differentiating prasinoviruses that infect different host genera

In total, there were 693 orthogroups in our analysis that occurred in at least one prasinovirus. Linear models relating orthogroup representation for each prasinovirus to host genus found 170 orthogroups that differed significantly between viruses of the three host genera ($P < 0.05$; Table S8). Therefore, 25% of prasinovirus orthogroups were significantly associated with host genus identity, suggesting that a substantial portion of the genome is involved in adapting to infect different host genera in the same taxonomic order. This percentage may be an overestimate because of the shared evolutionary history of the viruses, but it is nonetheless useful to use the *P*-values to rank orthogroups by the strength of differentiation among host genera. Twenty-eight orthogroups that differ strongly between host genera ($P < 0.001$) and that also have functional annotations (Table S3) exemplify the diversity of functions that relate viral gene content to host identity. For example, orthogroups that are absent in *Bathycoccus* viruses but present in most or all *Micromonas* and *Ostreococcus* viruses include asparagine synthetase (nitrogen and amino acid metabolism), dCMP deaminase (nucleotide metabolism), DNA polymerase X (potentially for base excision repair; 49), nucleotide-diphospho-sugar transferases (glycosylation), RNAse H (RNA processing), NTP pyrophosphohydrolase (potentially involved in stress response), and a protein with a rhodanese-like domain (potentially involved in stress response). Orthogroups that are present in most or all *Micromonas* viruses but absent/rare in *Ostreococcus* and *Bathycoccus* viruses include a glycerophosphodiester phosphodiesterase (lipid metabolism), mannitol dehydrogenase (potentially involved in stress response), ubiquitin, a protein with a zinc finger C2H2-type domain (potential transcription factor), a protein with an integrin alpha domain (potentially used for attachment to the host), and a putative tail fiber protein (potentially used for attachment to the host). Therefore, it may be the case that many stages of the viral life cycle, such as attachment to host receptors and manipulation of host metabolism and defenses, are involved in (co)evolution to infect different host genera.

We used unsupervised clustering analysis to further understand how gene content varies among the prasinovirus genomes that we analyzed (Fig. 7). Consistent with the many differences, we found between viruses infecting host genera (Table S3), unsupervised clustering largely groups viruses by host genus, with the exception of one *Ostreococcus tauri* virus that occurs in the *Micromonas* virus cluster. In our phylogenetic analysis, this strain, *Ostreococcus tauri* virus RT-2011 (JN225873), is relatively divergent from the clade containing the other 12 *Ostreococcus* viruses (Fig. 4). It is possible that OtV RT-2011 retained gene content similar to *Micromonas* viruses, while the other *Ostreococcus* viruses evolved more *Ostreococcus*-specific gene content. Finally, the clustering

results again emphasize the uniqueness of McV-KB2 relative to the other HiMcVs as the other three HiMcVs cluster together, while McV-KB2 is grouped with the genome of MpV-12T, a strain isolated from the coast of the Netherlands.

Distribution of HiMcV sequences in the world ocean

Using CoverM, we observed 115 instances of metagenomic reads mapping to our HiMcVs from the combined Station ALOHA, GEOTRACES, and Tara Oceans data sets (Table S9). Although our viruses were isolated at, or relatively near, Station ALOHA, only 8 out of 185 samples at this location included hits to a HiMcV (Fig. 8). The lower relative occurrence of HiMcVs in the Pacific, compared to the Atlantic, was further suggested by the absence of hits from samples in the GEOTRACES transect GP13 in the South Pacific or any of the Pacific Tara Oceans stations (Fig. 8).

The HiMcV strain isolated from the open ocean, McV-SA1, accounts for over half of all HiMcV hits in the metagenomic data sets. This strain was also the only HiMcV found in the North Atlantic and North Pacific basins. McV-KB4, despite being a sister to McV-SA1 in the phylogeny (Fig. 4), is underrepresented, with only three hits. This suggests that relatively small changes in broadly similar genomes can dramatically alter the ecology. HiMcV strain diversity and relative abundance increased on the edges of subtropical gyres, in transition zones, or near islands (Fig. 8), which tend to be more productive compared to the pelagic ocean. This is consistent with the origin of the three of the HiMcV strains from more productive coastal habitat, but more data would be needed to conclusively test the significance of these apparent biogeographical patterns.

In addition to our CoverM search of metagenomes, we compared McV isolate genomes to entire GVMAG assemblies in the GOEV database and found one GVMAG, TARA_AON_NCLDV_00048, that is closely related to McV-SA1. All contigs of this GVMAG could be mapped to the McV-SA1 genome, resulting in 71% coverage with 92% nucleotide identity (Fig. S6). Ha et al. (27) found that this GVMAG was the
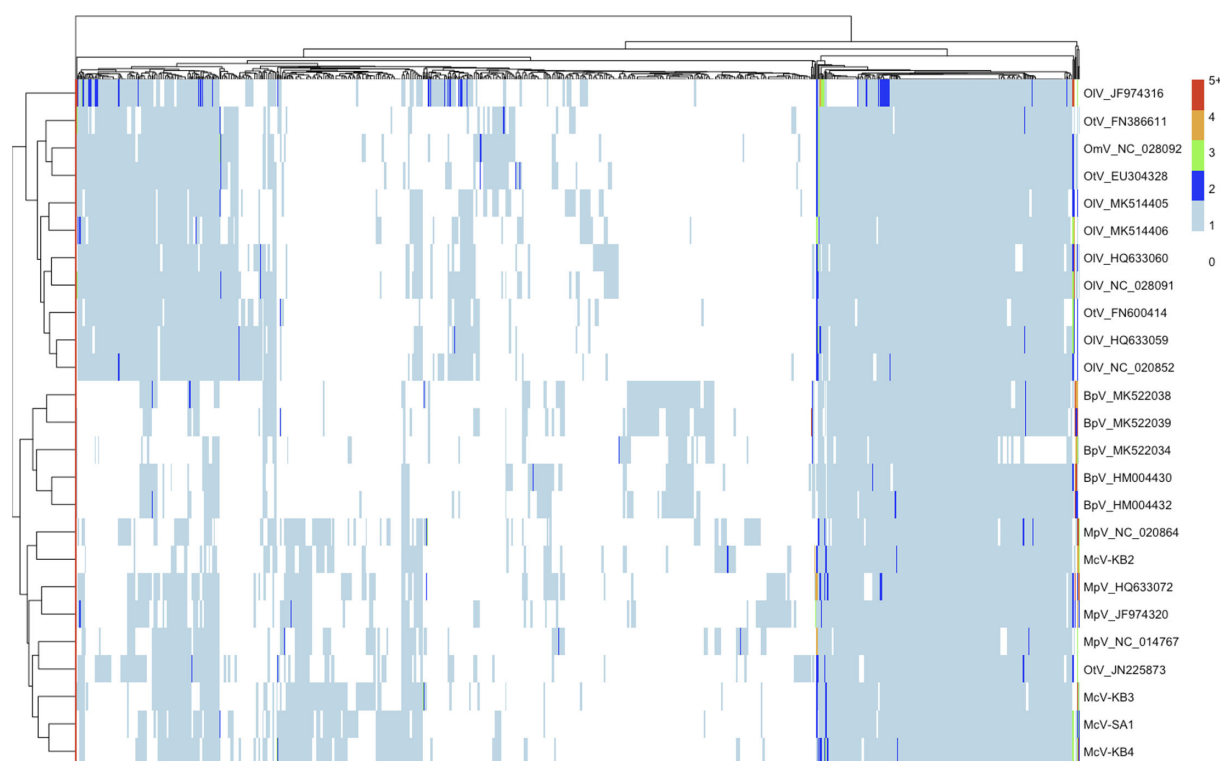

FIG 7 Clustered heatmap of 693 prasinovirus orthogroups. Colors in the heatmap represent the number of gene copies in each orthogroup per genome, with warmer colors being higher. The dendrogram on the left-hand side reflects similarity of virus strains based on orthogroup composition, and the dendrogram at the top clusters orthogroups by similarity in patterns of occurrence across strains.
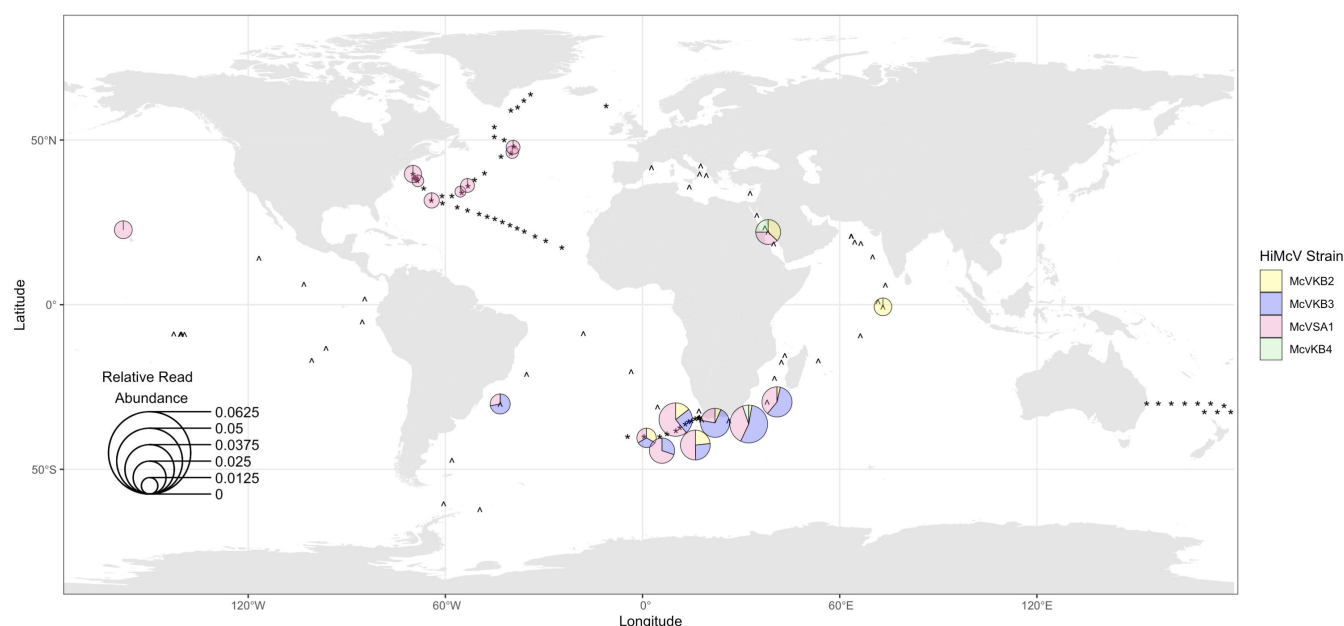
FIG 8 Distribution of HiMcV strains in metagenomic samples from GEOTRACES, Tara Oceans, and Station ALOHA. Pie charts represent strain composition at each location, based on the mean relative abundance of each strain. The size of each pie chart indicates the summed relative abundance of HiMcV transcripts. For HOT data, relative abundance was averaged over eight samples from different time points. Stations where no HiMcV strains were detected are represented with an asterisk for GEOTRACES and a caret for Tara Oceans. The smaller pie charts in the North Atlantic and Station ALOHA all contain reads only from McV-SA1. Pie chart locations are approximate as coordinates were adjusted to prevent overlap. The map was created in R programming software using the packages ggplot2 (50) and rnaturalearth (51).

8th most common *Algavirales* virus and the 15th most common *Nucleocytoviricota* virus in the bioGEOTRACES metagenome survey, out of 696 GVMAGs and 1,382 *Nucleocytoviricota* isolate genomes to which metagenome reads were mapped. The TARA_AON_NCLDV_00048 genome occurred in 42 out of 480 samples. For comparison, 76 of our CoverM hits came from the bioGEOTRACES database, 47 of which came from McV-SA1. Similar to McV-SA1, the Tara GVMAG occurred in the three Atlantic bioGEOTRACES transects, but not the sole Pacific transect.

Conclusions

Our study highlights the genomic diversity of prasinoviruses in several ways. Novel viruses from the same location, with overlapping host range, exhibited substantial variation in gene content, and one of the newly sequenced genomes (McV-KB2) was relatively divergent from all previously sequenced prasinoviruses.

The 61 HiMcV orthogroups that are novel to prasinoviruses also indicate that substantial functional diversity remains to be discovered in this clade, and judging from the novel orthogroups with functional annotations (a phosphate transporter, an apoptosis inhibitor, and enzymes from spermidine synthesis), the prasinovirus pangenome likely includes diverse mechanisms of virocell manipulation. The majority of the 61 novel orthogroups had homologs in metagenome-assembled genomes from the ocean, further suggesting substantial uncultivated genetic diversity. Finally, the various orthogroups that differ among viruses infecting *Micromonas*, *Ostreococcus*, and *Bathycoccus* point toward a better understanding of the functional basis of viral diversification. Viral gene content is strongly tied to host genus identity, and orthogroups unique to each genus suggest these may have diverse functions, such as virion attachment, manipulating host stress responses, and metabolism of components needed for virion construction.

Although the sample size is relatively small, it is noteworthy that all HiMcVs carry one or more orthologs to a host PTOX, and all HiMcVs also encode a high-light-inducible chlorophyll *a-b* binding protein. Among prasinoviruses, a PTOX sequence was previously found in only one other isolate (OtV-RT2011), but PTOX is commonly found in cyanophages (41), and has been documented in metagenome-assembled prasinovirus genomes from the South China Sea (28). The common occurrence of this gene in disparate virus clades may indicate an increased need for protection against photodamage for infected photosynthesizers at low latitudes. The occurrence of the high-light-inducible chlorophyll *a-b* binding protein present in all HiMcVs may also contribute to maintaining host metabolism under light stress, although at this point, it is unknown whether the viral proteins fulfill their annotated functions in infected cells.

Metagenomic data provide evidence that one HiMcV strain, McV-SA1, is relatively common in the global ocean. In the bioGEOTRACES transects this virus occurs at a similar frequency to a closely related uncultivated virus, which was shown to be one of the most common *Nucleocytoviricota* viruses in those transects (27). This suggests that *Micromonas*-infecting viruses are among the most common *Nucleocytoviricota* in the ocean. At the same time, the other three HiMcVs are relatively rare in major ocean basins. Several factors may contribute to the relatively small number of hits against the utilized data sets. These three isolates (McV-KB2, McV-KB3, and McV-KB4) were isolated from a coastal site and may be more abundant in coastal locations compared to the primarily open ocean metagenome stations. The relatively productive region near South Africa, surveyed in the bioGEOTRACES GA10 transect, contained the highest abundance and diversity of HiMcVs, consistent with *Micromonas* being more common under nutrient-rich conditions (12, 13). Therefore, further surveys in productive waters may capture more sequences that map to our HiMcVs.

In conclusion, our study shows that isolating and sequencing new viruses even within a relatively well-studied clade (the prasinoviruses) improves our knowledge of marine viral gene content and genome evolution. Such isolates also provide resources for future functional genomic studies that can resolve questions about putative gene functions and experimental studies to better understand virus contributions to phytoplankton ecology.

MATERIALS AND METHODS

Virus isolation

Four virus strains infecting the marine eukaryote *Micromonas commoda* were examined in this study. Three of the strains, McV-KB2, McV-KB3, and McV-KB4, were isolated from the surface waters (< 2 m) of Kāne'ohe Bay (21°27′N 157°48′W) on the windward side of O'ahu in 2011 (52). The hosts used to isolate these three viruses were algal strains UHM1063, UHM1064, and UHM1065, and the hosts were isolated from the same location 1–2 months prior. The fourth virus strain, McV-SA1, which was previously referred to as MsV-SA1 in (52), was isolated from a depth of 25 m from the pelagic research site Station ALOHA (22°45'N 158°W) in 2015, using the host strain UHM1080, which was isolated from Stn. ALOHA in 2012. Host ranges of the viruses were tested using the four aforementioned algal strains, plus UHM1060-1062. Of the seven host strains used, 18S rRNA gene sequences for UHM1061, UHM1065, and UHM1080 group into a clade with sequences of *Micromonas commoda* (Fig. S1; see [53] for details on phylogenetic analysis), and we therefore identify these strains as *M. commoda*. The four additional strains were tentatively identified as *Micromonas* based on morphological similarity to the sequenced isolates. All virus strains are maintained in the UHM culture collection via propagation on their original hosts, and full isolation methods are described in Schvarcz (52). In brief, whole seawater samples from respective sites were filtered, concentrated, and then added to healthy phytoplankton cultures, which were subsequently monitored for lysis. If lytic effects were confirmed after multiple transfers to healthy culture, then each lytic agent was propagated through several rounds of dilution-to-extinction in an

effort to render it clonal. In the current study, lysate stocks were maintained through fortnightly transfers of the lysate into healthy cultures grown in the f/2 -Si medium (54, 55). Once lysis took place, typically within 4–6 days of the initial challenge, and lysates were stored at 4°C.

Electron microscopy

Virions in the McV-SA1 lysate were purified by banding in a cesium chloride equilibrium buoyant density gradient. The virion-containing fraction was buffer-exchanged into SM using a centrifugal ultrafilter, and then a drop was adsorbed onto a carbon-stabilized formvar on a 200-mesh copper grid. The grid was rendered hydrophilic by glow discharge within a few hours prior to sample deposition. Virions were negatively stained with uranyl acetate and then visualized by transmission electron microscopy (Hitachi HT7700).

Virus genome sequencing

Purification, extraction, and sequencing protocols for McV-SA1 are described in (52). In summary, McV-SA1 virions were purified by banding in a CsCl density gradient, as noted above for electron microscopy. DNA was extracted from purified virions with a MasterPure Complete DNA and RNA purification Kit (Epicenter Biotechnologies, now Biosearch Technologies). PacBio library preparation and sequencing (P6-C4 chemistry) for a pooled sample of DNA from distantly related algal viruses, including McV-SA1, was performed with Sequel II SMRT cells at the University of Washington PacBio Sequencing Services facility. McV-SA1 reads were extracted using high-similarity BLAST searches against reference assemblies. In addition, Illumina NextSeq 150 bp paired-end sequencing of McV-SA1 was conducted at the Georgia Genomics and Bioinformatics Core at the University of Georgia, USA.

The isolates McV-KB2, McV-KB3, and McV-KB4 were prepared without using a density gradient and sequenced using only Illumina short-read technology. Prior to generating lysates for these viruses, we used a combination of filtration and antibiotics to reduce contamination from bacteria and phage present in the host cultures. *Micromonas* host cultures were filtered onto a 1 µm track-etched polycarbonate membrane filter (Nucleopore, Whatman) and then resuspended into sterile f/2 -Si medium containing a cocktail of broad-spectrum antibiotics (Table S1). Cultures were transferred three to four times into fresh antibiotic-containing f/2 -Si at 1:100 inoculum:medium. Flow cytometry counts indicated a tenfold reduction of bacteria in treated *Micromonas* cultures. To obtain McV DNA for sequencing, 5 µL of the 0.2 µm-filtered lysate was added to 50 mL of the antibiotic-treated host culture. After lysis, cultures were filtered through a 0.2 µm pore cellulose acetate syringe filter and subsequently concentrated by centrifugal ultrafiltration in units with 10 kDa MWCO regenerated cellulose membranes (Amicon Ultra, Millipore Sigma). DNA was then extracted from concentrated samples (Wizard Genomic DNA Purification kit, Promega). Library preparation and sequencing for Illumina 151 bp paired-end reads were performed at SeqCenter (formerly Microbial Genome Sequencing Center), located at the University of Pittsburgh, USA. Genome assembly methods varied among the viruses depending on the sequencing method. PacBio reads for Station ALOHA strain McV-SA1, extracted from a multiplexed sample through BLAST, were assembled with Canu v.1.0 (56) and polished with NextSeq data using Quiver v2.0.0, with 100% agreement with Illumina reads after polishing (52). The Kāneʻohe Bay virus strains (McV-KB2, McV-KB3, and McV-KB4) were assembled using two approaches that created relatively complete assemblies. The first approach, used for McV-KB3, used the default assembler in Geneious 11.1.5. Illumina 151 bp paired-end data were trimmed of adapters (kmer = 27), and low-quality reads (minimum = Q20) and short reads (minimum = 20 bp) were excluded using the BBDuk plug-in (v1.0; Biomatters Ltd.) and then normalized with the Geneious built-in tool (default settings) before assembly. We performed a nucleotide sequence similarity search of the resulting contigs with the Basic Local Alignment Search Tool (BLASTn; 52) against the NCBI nonredundant (nr) nucleotide

database (57) and identified two contigs (147 and 60 kbp) with similarity to known prasinoviruses. The reads mapped to these contigs were dissolved and re-assembled. This produced a single 205 kb contig. For McV-KB2, the iterative mapping approach using the Geneious assembler did not produce a single contig, so we instead used the metagenome assembly tool metaviralSPAdes (58, Galaxy Version 3.15.4 + galaxy 2 using SPAdes 3.15.3 accessed through usegalaxy.org), which produced a single 210 kb contig. Finally, for McV-KB4, neither of the above methods resulted in a single contig, so we treated the five putative prasinovirus contigs totaling 212 kb that were obtained from the Geneious assembler as a draft assembly for use in comparative analyses. Illumina raw reads were mapped back to McV-KB2, McV-KB3, and McV-KB4 to check for assembly errors, resulting in 100x to 1,000x continuous coverage. We treat the McV-SA1, McV-KB2, and McV-KB3 assemblies as near-complete based on their lengths, which are comparable to known prasinoviruses (Table 1), and based on whole-genome alignments using progressiveMauve, which indicated high synteny of McV-SA1, McV-KB3, McV-KB4, and the most similar previously sequenced *Micromonas virus* isolate, MpV1 (GenBank NC_014767.1; See *Results*).

Gene prediction was conducted with Prokka v.1.14.5 (59) accessed via Kbase (kbase.us) for all four assemblies. Only CDS with a start and stop codon and a minimum of 65 amino acids (195 nucleotides) were used in downstream analyses. Functional and structural annotation was performed manually by integrating information from EggNOG mapper (v.1.2, evalue ≤0.01, minimum 25% nucleotide identity), the InterProScan web interface (v.5.66-98.0, default settings), and refseq_protein (BLASTp, evalue ≤0.001) (60–62). We identified top BLASTp hits for each predicted gene against a custom database of genes from the Global Ocean Eukaryotic Virus (GOEV; 63). We also searched for giant virus MAGs closely related to our isolates by using the isolate genomes as nucleotide BLAST queries against a custom database comprising the GOEV GVMAG contigs. GOEV includes 581 metagenome-assembled genomes from uncultured marine viruses and 224 *Nucleocytoviricota* isolate reference genomes.

Host genome sequencing

The genomes of two *Micromonas* strains in our collection, UHM1061 and UHM1065 (referred to here as M1 and M2 for simplicity), were sequenced.

We created high-fidelity reference genomes from M1 (GenBank accession: JBHGVY000000000) and M2 (JBHKAF000000000) using the PacBio consensus long read (CLR) technology. As our algal cultures were not axenic, we used serial antibiotic treatments (antibiotic recipe in Table S1), followed by banding in a continuous Percoll density gradient to obtain a partially purified *Micromonas* fraction. Cells were then pelleted, flash-frozen in liquid nitrogen, and shipped to the University of Delaware Sequencing and Genotyping Center. High-molecular-weight DNA was extracted using a CTAB method. Library preparation and sequencing were performed with PacBio Sequel II SMRT cells.

A library of transcript sequences was also generated to aid in gene calling. Total RNA was generated from M1 and M2 cell cultures by sampling 10 mL of the exponential phase culture (~10^6 cell mL^{-1}) every 4 hours for 24 hours to capture diel variation in gene expression. An additional culture sample was given a heat shock treatment to stimulate the stress response, in which 10 mL aliquots were placed in a 30°C water bath for 30 minutes. Samples for RNA extraction were syringe-filtered onto 25 mm diameter, 1 µm pore size polycarbonate filters (Sterlitech), and then frozen immediately in liquid nitrogen and stored at −80°C. Within a week of sampling, filters were thawed over ice and total RNA extracted using the ZymoBIOMICS RNA Miniprep Kit (Zymo Research). Extracted RNA was sent to the University of Delaware Sequencing and Genotyping Center for 81 bp paired-end sequencing on the Illumina NextSeq 550.

PacBio CLR data were assembled using Canu (ver 1.9) in the PacBio-raw CLR assembly mode, providing an estimated genome size parameter of 22 Mbp (56). Resulting contigs were further polished to remove remaining InDel errors by iterative rounds of mapping

CLR reads to reference contigs using BLASR (ver 5.3.3 with default parameters except: maxMatch = 30, minSubreadLength = 750, minAlnLength = 750, minPctSimilarity = 70, minPctAccuracy = 70, hitPolicy = randombest), followed by error correction using Arrow (Pacific Biosciences GenomicConsensus ver 2.3.3 with default parameters except: minCoverage = 5, minConfidence = 40, coverage = 120) until a stable reference was obtained (four iterations; 64). Closing of circular/organellar elements was performed using Circlator (v1.5.5), and additional manual finishing was performed including manual assessment and scaffolding/overlap of adjacent contigs and resolution/dereplication of haplotype bubbles. Chromosome assignments were manually made using alignment to reference genomes with ProgressiveMauve (v2.4.0) (65; GenBank CP001323).

Annotation was performed with Maker (v3.01.03). A custom repeat library was generated using RepeatModeler (v2.0.1) with RepeatScout (v1.0.6) and TRF (v4.0.9). These repeats and repeats for order Mamiellales (CONS-Dfam_3.1-rb20170127) were identified by RMBLAST (v2.10.0) in RepeatMasker (v4.1.0) and masked for gene model annotation. Genome-specific *ab initio* gene calls by GeneMarkES (v2.5p) and SNAP (v2006.07.28) were used to train Augustus (v3.3.3) gene models using e-training scripts (BUSCO v4.0.2).

Illumina RNAseq data were quality-trimmed with TrimGalore (v0.6.5) using Cutadapt (v3.3) and mapped to draft genomes with STAR (v2.7.9a) using the two-pass method and was the basis for a genome-guided transcriptome assembly using Trinity (v2.13.2). Trinity transcripts and primary CDS and protein sequences annotated in *Micromonas pusilla* assemblies RCC299_229_v3.0 and CCMP1545_v3.0, and additional protein sequences extracted from UniProt for order Mamiellales were mapped to the genome as evidence and used to assess support for *ab initio* gene models. Noncoding RNA was identified using tRNAscan-SE (v2.0.5) and RNAmmer (v1.2), functional annotations were made by using BLASTp (v 2.12.0) against a Swiss-Prot database (v2021.04), and additional annotations were applied using InterProScan (v5.53-87.0).

Gene content comparisons

We compared the gene content of the four HiMcVs to determine: (i) how gene content varies among four viruses that overlap in host range and were either isolated from the same coastal location (Kāneʻohe Bay) or an offshore site (Station ALOHA) (ii), which genes are shared between the HiMcVs and the two Hawaiʻi *Micromonas commoda* cell strains with sequenced genomes (iii), whether there are genes in our virus strains not seen in previously published prasinoviruses genomes, and (iv) whether there are genes that consistently distinguish viruses infecting the three Mamiellales genera (*Micromonas*, *Ostreococcus*, and *Bathycoccus*) that may provide insights into the process of adaptation to different host taxa.

To make these comparisons, we used OrthoFinder with default settings (v5.5.; 66) to identify orthologous groups of genes in a data set containing the four HiMcVs, the two *Micromonas commoda* strains isolated from Oʻahu, and all 21 previously known prasinovirus genomes published in GenBank. Additionally, we followed Bachy et al. (10) by including chloroviruses, which form a monophyletic group with prasinoviruses based on marker gene analysis (67) and serve as an outgroup in phylogenetic analyses. Our full genome data set (accession numbers in Table S2) therefore includes strains that infect the Mamiellales genera of *Micromonas* (*n* = 8, including the four HiMcVs), *Ostreococcus* (*n* = 12), and *Bathycoccus* (*n* = 5); *Paramecium bursaria* Chlorella virus strains (*n* = 4); and the *Micromonas commoda* host strains M1 and M2 (*n* = 2).

To determine whether viruses infecting the three Mamiellales genera have consistently different gene content, we used gene count data for all prasinovirus genomes and fit the following linear model for each orthogroup: lm(*orthogroup gene count per genome ~host genus*).

Where "*host genus*" is a categorical predictor. This model quantifies the correlation between the number of genes in an orthogroup and the genus of the host infected by a viral strain, to identify orthogroups that most strongly differentiate viruses infecting

different host genera. Linear models were compared to a null model using chi-squared likelihood ratio tests in R (v.4.30; 68). *P*-values were adjusted for the false discovery rate using the R p.adjust function. The shared evolutionary history of the virus strains used in our models means that each strain is not a completely independent sample, and thus we use *P*-values primarily to rank orthogroups in order to determine orthogroups that differ the most by host genus. Prasinovirus orthogroup count data were also used to create a clustered heatmap using the R pheatmap package (69).

Species and gene tree construction

Phylogenetic relationships of the HiMcV isolates to other members of the order *Algavirales* were assessed by constructing species trees using orthologous genes from the aforementioned published genomes of prasinoviruses and chloroviruses. We used a gene alignment concatenation approach that included all orthogroups shared among all prasinovirus and chlorovirus genomes (i.e., "core" orthogroups; *n* = 26). We separately aligned each orthogroup using MAFFT (v7.450; 70), accessed through Geneious. Gene alignments were trimmed to eliminate non-overlapping sequences with Goalign v0.3.7 (https://github.com/evolbioinfo/goalign). If more than one gene copy was present in a genome, the paralog most closely related to orthologs in other genomes was chosen, and then the 26 ortholog alignments were concatenated in Geneious. A phylogeny was estimated with FastTree (v.1.2; 71) within Geneious. We also constructed a tree using only polB sequences, in order to compare the gene tree of this common prasinovirus marker gene to our core gene-based species trees. We used FigTree (v.1.4.4; http://tree.bio.ed.ac.uk/software/figtree/) to visualize both gene concatenation and polB trees (Fig. 4; Fig. S2). OrthoFinder automatically generates a species tree using shared orthogroups with the STAG algorithm (72), but this tree does not include node support values when < 100 core orthogroups are present, as is the case with our data set. We have provided this tree in the supplement as a point of comparison (Fig. S3).

We identified a core HiMcV alternative oxidase gene that may have recent cellular or phage origins based on the taxonomy of top BLAST hits. We constructed a gene tree to see if the HiMcV orthologs of this gene were more closely related to host or phage sequences, and whether alternative oxidases have been acquired more than once by phytoplankton viruses. We created a MAFFT alignment using alternative oxidase sequences from the four HiMcVs, the two Hawai'i *Micromonas* hosts, OtV RT-2011 (the only other prasinovirus with an alternative oxidase gene), and representative cyanobacteria and cyanophages. This alignment was then trimmed with Goalign, and a phylogeny was constructed in IQ-Tree (v.1.5.12; 73) with the following flags: -m LG + F + R10 wbt -bb 1000. Previous work on plant alternative oxidases has found that the mitochondrial alternative oxidase (AOX, also called ubiquinol oxidase) and the plastid alternative oxidase (PTOX, also called plastoquinol terminal oxidase) are often misannotated or annotated in an inconsistent way (43). Therefore, we also included reference PTOX and AOX plant and algal sequences to aid in interpreting the alignment and phylogeny.

HiMcV detection in metagenomes

To assess the global distribution of the four HiMcVs, we searched publicly available metagenomes, spanning Pacific and Atlantic ocean basins, from the Hawai'i Ocean Time-series (HOT), GEOTRACES, and Tara Oceans (9, 74–77). The HOT data contains samples from 19 timepoints and nine discrete depths above 200 m at Station ALOHA. The other databases contain one sample per station, taken at a depth above 150 m. Sequencing runs from HOT include 20 L samples filtered onto a 0.2 µm filter (*n* = 293) and samples filtered onto a 0.02 µm filter after passing through at 0.2 µm filter (*n* = 185). The GEOTRACES metagenomic data set consists of 100 mL samples of whole seawater from the photic zone filtered onto 0.2 µm filters (*n* = 490), with samples taken from the GA02 and GA03 transects in the North Atlantic, the GA10 transect in the South Atlantic (off the coast of South Africa), and the GP13 transect in the South Pacific (off the coasts of Australia and New Zealand). Metagenome samples from the Tara Oceans

Virome database were collected from 100 L whole seawater obtained from 5 to 100 m depth, depending on the station. Seawater was prefiltered through either a 20 µm or 5 µm net and then filtered with a 0.22 µm pore size membrane filter. Iron flocculation was then used to concentrate viruses in the filtrate. The metagenomic samples used in our analysis thus include some that should be enriched for free virions in the McV size range (i.e., the fraction between 0.02 and 0.2 µm), and others that should contain primarily cells and larger virions (i.e., the >0.2 µm fraction), while potentially including smaller virions that did not pass through the 0.2 µm filter. Samples of this larger size fraction were included because a previous study of the GEOTRACES metagenomes (> 0.2 µm) found that putative prasinoviruses were among the most abundant and diverse *Nucleocytoviricota* representatives in these samples (27).

We used CoverM v0.6.1 (https://github.com/wwood/CoverM) to search the metagenomic data sets for sequences that mapped onto at least one of the four HiMcV assemblies. Requirements of 95% minimum read identity and 20% minimum covered fraction were used (27), indicated with the flags --min-read-percent-identity and --min-covered-fraction. For each successful hit, relative abundances of each virus (i.e., percent of reads from the metagenome sample) derived from CoverM results were then merged with metadata to create a map of hits in R statistical software. CoverM search information, including accession numbers and metadata, is available in Table S4.

During the annotation phase of our genomic analysis, we found a metagenome-assembled genome from the GOEV database that was similar to McV-SA1. This MAG was derived from Tara Ocean data and was the eighth-most common assembly in the GOEV database. To provide some context for the presence of McV-SA1 in the world ocean, we mapped the reads of the GOEV MAG against the McV-SA1 assembly.

AUTHOR AFFILIATIONS

[1]Department of Oceanography, School of Ocean and Earth Science and Technology, University of Hawaiʻi at Mānoa, Honolulu, Hawaiʻi, USA
[2]Center for Bioinformatics and Computational Biology & Department of Computer and Information Sciences, University of Delaware, Newark, Delaware, USA
[3]Center for Microbial Oceanography Research and Education, University of Hawaiʻi at Mānoa, Honolulu, Hawaiʻi, USA

AUTHOR ORCIDs

Anamica Bedi de Silva ⓘ http://orcid.org/0000-0001-6573-7546
Shawn W. Polson ⓘ http://orcid.org/0000-0002-3398-6932
Grieg F. Steward ⓘ http://orcid.org/0000-0001-5988-0522
Kyle F. Edwards ⓘ http://orcid.org/0000-0002-0661-3903

FUNDING

AUTHOR CONTRIBUTIONS

Anamica Bedi de Silva, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review and editing | Shawn W. Polson, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Writing – review and editing | Christopher R. Schvarcz, Conceptualization, Resources, Validation, Writing – review and editing | Grieg F. Steward, Conceptualization, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – review and editing | Kyle F. Edwards, Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – review and editing

DATA AVAILABILITY

All genomic sequences of the HiMcVs have been submitted to GenBank under the following accession numbers: McV-KB2 (PP911589), McV-KB3 (PQ109088), McV-KB4 (PQ359806), and McV-SA1 (PQ381123). Genomics sequences for hosts M1 and M2 were submitted to GenBank under Bioproject accession PRJNA1141538. Additional sequences used in analyses can be found in supplemental materials.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental figures and tables (Spectrum02583-24-s0001.docx). Tables S1 to S3 and Fig. S1 to S6.

Tables S4 to S8 (Spectrum02583-24-s0002.xlsx). Additional data used in analyses.

REFERENCES

1. Suttle CA. 2007. Marine viruses — major players in the global ecosystem. Nat Rev Microbiol 5:801–812. https://doi.org/10.1038/nrmicro1750
2. Wilhelm SW, Suttle CA. 1999. Viruses and nutrient cycles in the sea - viruses play critical roles in the structure and function of aquatic food webs. BioScience 49:8. https://doi.org/10.2307/1313569
3. Weitz JS, Wilhelm SW. 2012. Ocean viruses and their effects on microbial communities and biogeochemical cycles. F1000 Biol Rep 4:17. https://doi.org/10.3410/B4-17
4. Hurwitz BL, Hallam SJ, Sullivan MB. 2013. Metabolic reprogramming by viruses in the sunlit and dark ocean. Genome Biol 14:R123. https://doi.org/10.1186/gb-2013-14-11-r123
5. Våge S, Storesund JE, Thingstad TF. 2013. Adding a cost of resistance description extends the ability of virus-host model to explain observed patterns in structure and function of pelagic microbial communities. Environ Microbiol 15:1842–1852. https://doi.org/10.1111/1462-2920.12077
6. Weynberg KD, Allen MJ, Wilson WH. 2017. Marine prasinoviruses and their tiny plankton hosts: a review. Viruses 9:43. https://doi.org/10.3390/v9030043
7. ICTV. 2022. Virus taxonomy: 2022
8. Leliaert F, Smith DR, Moreau H, Herron MD, Verbruggen H, Delwiche CF, De Clerck O. 2012. Phylogeny and molecular evolution of the green algae. CRC Crit Rev Plant Sci 31:1–46. https://doi.org/10.1080/07352689.2011.615705
9. Aylward FO, Moniruzzaman M, Ha AD, Koonin EV. 2021. A phylogenomic framework for charting the diversity and evolution of giant viruses. PLoS Biol 19:e3001430. https://doi.org/10.1371/journal.pbio.3001430
10. Bachy C, Yung CCM, Needham DM, Gazitúa MC, Roux S, Limardo AJ, Choi CJ, Jorgens DM, Sullivan MB, Worden AZ. 2021. Viruses infecting a warm water picoeukaryote shed light on spatial co-occurrence dynamics of marine viruses and their hosts. ISME J 15:3129–3147. https://doi.org/10.1038/s41396-021-00989-9
11. Yau S, Grimsley N, Moreau H. 2015. Molecular ecology of Mamiellales and their viruses in the marine environment. Perspect Phycol 2:83–89. https://doi.org/10.1127/pip/2015/0026
12. Not F, Latasa M, Marie D, Cariou T, Vaulot D, Simon N. 2004. A single species, *Micromonas pusilla* (Prasinophyceae), dominates the eukaryotic picoplankton in the western English channel. Appl Environ Microbiol 70:4064–4072. https://doi.org/10.1128/AEM.70.7.4064-4072.2004
13. Lopes Dos Santos A, Gourvil P, Tragin M, Noël M-H, Decelle J, Romac S, Vaulot D. 2017. Diversity and oceanic distribution of prasinophytes clade VII, the dominant group of green algae in oceanic waters. ISME J 11:512–528. https://doi.org/10.1038/ismej.2016.120
14. Manton I, Parke M. 1960. Further observations on small green flagellates with special reference to possible relatives of *Chromulina pusilla* Butcher. J Mar Biol Ass 39:275–298. https://doi.org/10.1017/S0025315400013321
15. Eikrem W, Throndsen J. 1990. The ultrastructure of *Bathycoccus* gen. nov. and *B. prasinos* sp. nov., a non-motile picoplanktonic alga (Chlorophyta,

Prasinophyceae) from the Mediterranean and Atlantic. Phycologia 29:344–350. https://doi.org/10.2216/i0031-8884-29-3-344.1

16. Courties C, Vaquer A, Trousseillier M, Lautier J, Chrétiennot-Dinet MJ, Neveux J, Machado C, Claustre H. 1994. Smallest eukaryotic organism. Nature 370:255–255. https://doi.org/10.1038/370255a0

17. Mayer JA, Taylor FJR. 1979. A virus which lyses the marine nanoflagellate *Micromonas pusilla*. Nature 281:299–301. https://doi.org/10.1038/28129 9a0

18. Cottrell MT, Suttle CA. 1995. Genetic diversity of algal viruses which lyse the photosynthetic picoflagellate *Micromonas pusilla* (Prasinophyceae). Appl Environ Microbiol 61:3088–3091. https://doi.org/10.1128/aem.61.8. 3088-3091.1995

19. Finke JF, Winget DM, Chan AM, Suttle CA. 2017. Variation in the genetic repertoire of viruses infecting *Micromonas pusilla* reflects horizontal gene transfer and links to their environmental distribution. Viruses 9:116. https://doi.org/10.3390/v9050116

20. Monier A, Chambouvet A, Milner DS, Attah V, Terrado R, Lovejoy C, Moreau H, Santoro AE, Derelle É, Richards TA. 2017. Host-derived viral transporter protein for nitrogen uptake in infected marine phytoplankton. Proc Natl Acad Sci USA 114:E7489–E7498. https://doi.org/10.1073/p nas.1708097114

21. Thomas R, Jacquet S, Grimsley N, Moreau H. 2012. Strategies and mechanisms of resistance to viruses in photosynthetic aquatic microorganisms. Adv Ocean Limnol 3:1. https://doi.org/10.4081/aiol.201 2.5323

22. Heath SE, Knox K, Vale PF, Collins S. 2017. Virus resistance is not costly in a marine alga evolving under multiple environmental stressors. Viruses PMCID:39. https://doi.org/10.3390/v9030039

23. Derelle E, Yau S, Moreau H, Grimsley NH. 2018. Prasinovirus attack of ostreococcus is furtive by day but savage by night. J Virol PMCID:e01703–17. https://doi.org/10.1128/JVI.01703-17

24. Moreau H, Piganeau G, Desdevises Y, Cooke R, Derelle E, Grimsley N. 2010. Marine prasinovirus genomes show low evolutionary divergence and acquisition of protein metabolism genes by horizontal gene transfer. J Virol 84:12555–12563. https://doi.org/10.1128/JVI.01123-10

25. Clerissi C, Grimsley N, Ogata H, Hingamp P, Poulain J, Desdevises Y. 2014. Unveiling of the diversity of prasinoviruses (Phycodnaviridae) in marine samples by using high-throughput sequencing analyses of PCR-amplified DNA polymerase and major capsid protein genes. Appl Environ Microbiol 80:3150–3160. https://doi.org/10.1128/AEM.00123-14

26. Farzad R, Ha AD, Aylward FO. 2022. Diversity and genomics of giant viruses in the north pacific subtropical gyre. Front Microbiol 13:1021923. https://doi.org/10.3389/fmicb.2022.1021923

27. Ha AD, Moniruzzaman M, Aylward FO. 2023. Assessing the biogeography of marine giant viruses in four oceanic transects. ISME Commun 3:43. https://doi.org/10.1038/s43705-023-00252-6

28. Thomy J, Sanchez F, Prioux C, Yau S, Xu Y, Mak J, Sun R, Piganeau G, Yung CCM. 2024. Unveiling prasinovirus diversity and host specificity through targeted enrichment in the South China Sea. ISME Commun 4:ycae109. h ttps://doi.org/10.1093/ismeco/ycae109

29. Rasmussen T. 2023. The potassium efflux system Kef: bacterial protection against toxic electrophilic compounds. Membranes (Basel) 13:465. https://doi.org/10.3390/membranes13050465

30. Greiner T, Moroni A, Van Etten JL, Thiel G. 2018. Genes for membrane transport proteins: not so rare in viruses. Viruses 10:456. https://doi.org/ 10.3390/v10090456

31. Greiner T, Frohns F, Kang M, Van Etten JL, Käsmann A, Moroni A, Hertel B, Thiel G. 2009. Chlorella viruses prevent multiple infections by depolarizing the host membrane. J Gen Virol 90:2033–2039. https://doi.org/10.10 99/vir.0.010629-0

32. Liu X-D, Shen Y-G. 2004. NaCl-induced phosphorylation of light harvesting chlorophyll *a/b* proteins in thylakoid membranes from the halotolerant green alga, *Dunaliella salina*. FEBS Lett 569:337–340. https:/ /doi.org/10.1016/j.febslet.2004.05.065

33. Gallot-Lavallée L, Blanc G, Claverie J-M. 2017. 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:e00230-17. https://doi.org/10.1 128/JVI.00230-17

34. Moniruzzaman M, Martinez-Gutierrez CA, Weinheimer AR, Aylward FO. 2020. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. Nat Commun 11:1710. https://doi.org/ 10.1038/s41467-020-15507-2

35. Dolganov NA, Bhaya D, Grossman AR. 1995. Cyanobacterial protein with similarity to the chlorophyll a/b binding proteins of higher plants: evolution and regulation. Proc Natl Acad Sci USA 92:636–640. https://doi .org/10.1073/pnas.92.2.636

36. Varsano T, Wolf SG, Pick U. 2006. A chlorophyll *a/b*-binding protein homolog that is induced by iron deficiency is associated with enlarged photosystem I units in the eucaryotic alga *Dunaliella salina*. J Biol Chem 281:10305–10315. https://doi.org/10.1074/jbc.M511057200

37. Monier A, Welsh RM, Gentemann C, Weinstock G, Sodergren E, Armbrust EV, Eisen JA, Worden AZ. 2012. Phosphate transporters in marine phytoplankton and their viruses: cross-domain commonalities in viral-host gene exchanges. Environ Microbiol 14:162–176. https://doi.org/10. 1111/j.1462-2920.2011.02576.x

38. Goldsmith DB, Crosti G, Dwivedi B, McDaniel LD, Varsani A, Suttle CA, Weinbauer MG, Sandaa R-A, Breitbart M. 2011. Development of phoH as a novel signature gene for assessing marine phage diversity. Appl Environ Microbiol 77:7730–7739. https://doi.org/10.1128/AEM.05531-11

39. Jover LF, Effler TC, Buchan A, Wilhelm SW, Weitz JS. 2014. The elemental composition of virus particles: implications for marine biogeochemical cycles. Nat Rev Microbiol 12:519–528. https://doi.org/10.1038/nrmicro32 89

40. Culley AI, Asuncion BF, Steward GF. 2009. Detection of inteins among diverse DNA polymerase genes of uncultivated members of the phycodnaviridae. ISME J 3:409–418. https://doi.org/10.1038/ismej.2008. 120

41. Puxty RJ, Millard AD, Evans DJ, Scanlan DJ. 2015. Shedding new light on viral photosynthesis. Photosynth Res 126:71–97. https://doi.org/10.1007 /s11120-014-0057-x

42. Wang Y, Ferrinho S, Connaris H, Goss RJM. 2023. The impact of viral infection on the chemistries of the earth's most abundant photosynthesizes: metabolically talented aquatic cyanobacteria. Biomolecules 13:1218. https://doi.org/10.3390/biom13081218

43. Nobre T, Campos MD, Lucic-Mercy E, Arnholdt-Schmitt B. 2016. Misannotation awareness: a tale of two gene-groups. Front Plant Sci 7:868. https://doi.org/10.3389/fpls.2016.00868

44. Banadyga L, Lam S-C, Okamoto T, Kvansakul M, Huang DC, Barry M. 2011. Deerpox virus encodes an inhibitor of apoptosis that regulates Bak and Bax. J Virol 85:1922–1934. https://doi.org/10.1128/JVI.01959-10

45. Ma J, Edlich F, Bermejo GA, Norris KL, Youle RJ, Tjandra N. 2012. Structural mechanism of bax inhibition by cytomegalovirus protein vMIA. Proc Natl Acad Sci USA 109:20901–20906. https://doi.org/10.1073/ pnas.1217094110

46. Li B, Liang J, Baniasadi HR, Phillips MA, Michael AJ. 2023. Functional polyamine metabolic enzymes and pathways encoded by the virosphere. Proc Natl Acad Sci USA 120:e2214165120. https://doi.org/10. 1073/pnas.2214165120

47. Baumann S, Sander A, Gurnon JR, Yanai-Balser GM, Van Etten JL, Piotrowski M. 2007. Chlorella viruses contain genes encoding a complete polyamine biosynthetic pathway. Virology (Auckl) 360:209–217. https://doi.org/10.1016/j.virol.2006.10.010

48. Jeffrey SW, Wright SW, Zapata M. 2011. Microalgal classes and their signature pigments, p 3–77. In Roy S, Llewellyn CA, Egeland ES, Johnsen G (ed), Phytoplankton pigments, 1st ed. Cambridge University Press.

49. Fernández-García JL, de Ory A, Brussaard CPD, de Vega M. 2017. *Phaeocystis globosa* virus DNA polymerase X: a "Swiss Army knife", multifunctional DNA polymerase-lyase-ligase for base excision repair. Sci Rep 7:6907. https://doi.org/10.1038/s41598-017-07378-3

50. Wickham H. 2016. Ggplot2: elegant graphics for data analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org.

51. Massicotte P, South A. 2025. Rnaturalearth: World map data from natural earth R package version 1.0.1. https://docs.ropensci.org/rnaturalearth/.

52. Schvarcz CR. 2018. Cultivation and characterization of viruses infecting eukaryotic phytoplankton from the tropical north Pacific Ocean. University of Hawaii at Manoa, Honolulu, HI.

53. Bedi de Silva A, Polson SW, Schvarcz CR, Steward GF, Edwards KF. 2024. Transient, context-dependent fitness costs accompanying viral resistance in isolates of the marine microalga *Micromonas* sp. (class Mamiellophyceae). Environ Microbiol 26:e16686. https://doi.org/10.1111 /1462-2920.16686

54. Guillard RRL. 1975. Culture of phytoplankton for feeding marine invertebrates, p 29–60. In Smith WL, Chanley MH (ed), Culture of marine invertebrate animals. Springer US, Boston, MA.

55. Guillard RR, Ryther JH. 1962. Studies of marine planktonic diatoms. l. Cyclotella nana hustedt, and detonula confervacea (cleve) gran. Can J Microbiol 8:229–239. https://doi.org/10.1139/m62-029

56. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res 27:722–736. https://doi.org/10.1101/gr.215087.116

57. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, Madej T, Marchler-Bauer A, Lanczycki C, Lathrop S, Lu Z, Thibaud-Nissen F, Murphy T, Phan L, Skripchenko Y, Tse T, Wang J, Williams R, Trawick BW, Pruitt KD, Sherry ST. 2022. Database resources of the national center for biotechnology information. Nucleic Acids Res 50:D20–D26. https://doi.org/10.1093/nar/gkab1112

58. Antipov D, Raiko M, Lapidus A, Pevzner PA. 2020. Metaviral SPAdes: assembly of viruses from metagenomic data. Bioinformatics 36:4126–4129. https://doi.org/10.1093/bioinformatics/btaa490

59. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153

60. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

61. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005. InterProScan: protein domains identifier. Nucleic Acids Res 33:W116–W120. https://doi.org/10.1093/nar/gki442

62. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol 38:5825–5829. https://doi.org/10.1093/molbev/msab293

63. Gaïa M, Meng L, Pelletier E, Forterre P, Vanni C, Fernandez-Guerra A, Jaillon O, Wincker P, Ogata H, Krupovic M, Delmont TO. 2023. Mirusviruses link herpesviruses to giant viruses. Nature 616:783–789. https://doi.org/10.1038/s41586-023-05962-4

64. Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics 13:238. https://doi.org/10.1186/1471-2105-13-238

65. Worden AZ, Lee J-H, Mock T, Rouzé P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E, Everett MV, et al. 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. Science 324:268–272. https://doi.org/10.1126/science.1167222

66. Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:238. https://doi.org/10.1186/s13059-019-1832-y

67. Bellec L, Grimsley N, Moreau H, Desdevises Y. 2009. Phylogenetic analysis of new prasinoviruses (*Phycodnaviridae*) that infect the green unicellular algae *Ostreococcus*, *Bathycoccus* and *Micromonas*. Environ Microbiol Rep 1:114–123. https://doi.org/10.1111/j.1758-2229.2009.00015.x

68. Core Team R. 2022. R: a language and environment for statistical computing

69. Kolde R. 2019. pheatmap: pretty heatmaps (. R package version 1.0.12,). R

70. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010

71. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490

72. Emms DM, Kelly S. 2018. STAG: species tree inference from all genes. bioRxiv. https://doi.org/10.1101/267914

73. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/molbev/msaa015

74. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S, Cruaud C, de Vargas C, Gasol JM, et al. 2015. Patterns and ecological drivers of ocean viral communities. Science 348:1261498. https://doi.org/10.1126/science.1261498

75. Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G, Iudicone D, Karsenti E, Speich S, Troublé R, et al. 2015. Open science resources for the discovery and analysis of *Tara* Oceans data. Sci Data 2:150023. https://doi.org/10.1038/sdata.2015.23

76. Mende DR, Bryant JA, Aylward FO, Eppley JM, Nielsen T, Karl DM, DeLong EF. 2017. Environmental drivers of a microbial genomic transition zone in the ocean's interior. Nat Microbiol 2:1367–1373. https://doi.org/10.1038/s41564-017-0008-3

77. Biller SJ, Berube PM, Dooley K, Williams M, Satinsky BM, Hackl T, Hogle SL, Coe A, Bergauer K, Bouman HA, Browning TJ, De Corte D, Hassler C, Hulston D, Jacquot JE, Maas EW, Reinthaler T, Sintes E, Yokokawa T, Chisholm SW. 2018. Marine microbial metagenomes sampled across space and time. Sci Data 5:180176. https://doi.org/10.1038/sdata.2018.176