A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators


Edited by W. Ford Doolittle, Dalhousie University, Halifax, Canada, and approved August 8, 2019 (received for review May 27, 2019)

Giant viruses are remarkable for their large genomes, often rivaling those of small bacteria, and for having genes thought exclusive to cellular life. Most isolated to date infect nonmarine protists, leaving their strategies and prevalence in marine environments largely unknown. Using eukaryotic single-cell metagenomics in the Pacific, we discovered a Mimiviridae lineage of giant viruses, which infects choanoflagellates, widespread protist predators related to metazoans. The ChaoanoVirus genomes are the largest yet characterized type-1 rhodopsins, from a virus (PgV) with a known host (a marine alga). Unlike the algal virus, ChaoanoViruses encode the entire pigment biosynthesis pathway and cleavage enzyme for producing the required chromophore, retinal. We demonstrate that the rhodopsin shared by ChaoanoViruses and PgV binds retinal and pumps protons. Moreover, our 1.65-Å resolved VirR₅₇₉ crystal structure and mutational analyses exposed differences from previously characterized type-1 rhodopsins, all of which come from cellular organisms. Multiple VirR types are present in metagenomes from across surface oceans, where they are correlated with and nearly as abundant as a canonical marker gene from Mimiviridae. Our findings indicate that light-dependent energy transfer systems are likely common components of giant viruses of photosynthetic and phagotropic unicellular marine eukaryotes.

Significance

Although viruses are well-characterized regulators of eukaryotic algae, little is known about those infecting unicellular predators in oceans. We report the largest marine virus genome yet discovered, found in a wild predatory choanoflagellate sorted away from other Pacific microbes and pursued using integration of cultivation-independent and laboratory methods. The giant virus encodes nearly 900 proteins, many unlike known proteins, others related to cellular metabolism and organic matter degradation, and 3 type-1 rhodopsins. The viral rhodopsin that is most abundant in marine metagenomes, and also present in an algal virus, pumps protons when illuminated, akin to cellular rhodopsins that generate a proton-motive force. Giant viruses likely provision multiple host species with phototrophic capacities, including predatory unicellular relatives of animals.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

This open access article is distributed under Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Data deposition: Data accession numbers and databases in which they have been deposited are provided in Dataset S2 for all sequence data; the protein biochemical characterization/virology data have been deposited in the Protein Data Bank, www.wwpdb.org (PDB ID code 6JO0).

D.M.N., S.Y., and T.H. contributed equally to this work.

Present address: Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam.

1Station Biologique de Roscoff, Sorbonne Université, CNRS, 29688 Roscoff, France.

2To whom correspondence may be addressed. E-mail: iwasaki@bs.s.u-tokyo.ac.jp or azworden@geomar.de.

This article contains supporting information online at www.pnas.org/cgi/doi/10.1073/pnas.1907517116/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1907517116
A giant virus infects a predatory protist that is considered to be among the closest living unicellular relatives of metazoans. (10) www.pnas.org/cgi/doi/10.1073/pnas.1907517116 Needham et al.

Fig. 1. A giant virus infects a predatory protist that is considered to be among the closest living unicellular relatives of metazoans. (A) Schematic tree of eukaryotes, with supergroups indicated by colors or gray branches in contentious positions. Lineages with giant viruses (pink) known (circles) or discovered here (star) are indicated. (B) Locations of single-cell sorting where ChoanoV1 and its host, B. minor, were recovered (Station M2), where ChoanoV2 (Station 67-70) was found, and where metatranscriptomes were sequenced from unmanipulated seawater (M1, M2, 67-70; Station 67-155, 785 km from shore, not here (star) are indicated. (C) Histogram showing the population (circled) of sorted choanoflagellate cells (blue dots), including the viral-infected cell (pink), based on index sorting and V4 18S rRNA gene amplicon sequencing. Other data points reflect unsorted particles in the stained seawater analyzed. The box (green) indicates the position of YG bead standards run before and after sorting at the same settings. (D) Categorized summary of the top 10 BLASTp matches for 862 ChoanoV1 proteins (e-value < 10^-5) in cellular organisms and NCLDV.

370 to 670 Kb, than many other giant viruses, and all belong to the nucleocytoplasmic large DNA viruses (NCLDV) family, which houses smaller eukaryotic marine viruses as well (24) (Dataset S1). Nevertheless, the marine giants encode a number of AMGs that connect to how they alter host metabolism during infection, such as fermentation-related genes (20) and sphingolipid-biosynthesis genes (6) in algal viruses, essential information for considering downstream biogeochemical processes and modeling the impacts of virus–host interactions on ecosystem processes.

The paucity of giant viruses isolated from marine ecosystems likely results from dependence of classical viral isolation methods on cultured hosts, such as the bacteriophagous stramenopile Cafeteria, for recovering CroV (21). Unfortunately, many marine protists remain uncultured (15, 25) and hence, are not available for use as viral bait. This is especially so for predatory protists, in part because the natural consortia that constitute their food base are outcompeted by a few copiotrophic, relatively large bacterial taxa once in enriched medium in the laboratory (25). In some cases, metagenomics has been used to recover genome-level information while obviating cultivation. In particular, giant virus genomes have been assembled from metagenomic data acquired from low-diversity, simplified ecosystems [e.g., wastewater (12) and a hypersaline lake in Antarctica (26)]. However, these approaches are less successful in high-diversity environments, unless the biological entity has high abundance, and they fail to directly link virus to host (13), an important factor for understanding ecological impacts. To overcome these challenges, we integrated multiple culture-independent and laboratory methods to perform this cross-scale study, in which we first sorted individual wild predatory protists and used single-cell metagenomics to examine these eukaryotes and coassociated entities. With a resulting genome from an uncultured giant virus in hand, we asked how its predicted functional attributes differed from the marine giant virus genomes characterized previously, all of which come from cultivation-based isolation and sequencing, and from the plethora of giant viruses from nonmarine habitats. Furthermore,
we identified conserved attributes and established the distribution and biochemical function of a viral rhodopsin that thus far seems unique to giant viruses in the marine biosphere.

Results and Discussion

A Wild Predatory Prostid in the North Pacific Ocean and Its Virus. To capture uncultured heterotrophic prostids, we used high-purity fluorescence-activated cell sorting (FACS) of single cells with acidic vacuole staining to discriminate protists from prokaryotes and an additional exclusion gate against photosynthetic organisms to select heterotrophic protists only (SI Appendix). In a FACS survey in the eastern North Pacific, we recovered a coherent population of choanoflagellates (Figs. 1B and C), heterotrophic predators belonging to the supergroup Opisthokonta that are considered to be among the closest living unicellular relatives of metazoa (27). Choanoflagellates comprised 99% of the 198 wells for which V4 18S ribosomal RNA (rRNA) gene amplicons were recovered after initial multiple displacement amplification of DNA from single cells, and the remaining 3 wells harbored amplicons with highest identity to uncultured syndiniales (putative parasites) and 2 different uncultured cercozoans (bacterivores), respectively. Choanoflagellates are widespread bacterivorous protists that we expected to be targeted by our staining protocol, because they could contain an acidic food vacuole.

From one choanoflagellate cell, we assembled an 875-Kb viral genome after eukaryotic single-cell metagenomic sequencing (SI Appendix, Figs. S1 and S2). The virus, ChoanoV1, represents the largest pelagic marine giant virus genome sequenced yet; its genomic DNA base composition (GC content) was low (22%), rivaling only nonmarine Hokoviruses (21%) and CroV (23%), whereas other giant viruses range to 64% GC (10, 12, 21) (SI Appendix, Fig. S2 and Dataset S1). The ChoanoV1 genome encoded 862 predicted proteins, and its gene content suggested that it belonged to the NCLDV (Fig. 1D and SI Appendix, Fig. S2), a diverse group of eukaryotic viruses (10, 11).

Presence of a eukaryotic virus associated with a single choanoflagellate cell could reflect several possible ecological interactions: first, that the virus had infected the choanoflagellate and replicated there; second, that the virus had been consumed by the predator as a prey item as reported in 2 prior culture-based field studies (31). Contigs from bacterial prey (and phages) were retrieved in metatranscriptomes that we sequenced from the eastern North Pacific, demonstrating expression (SI Appendix, Fig. S3), at least at this stage in time, of one of the more fragmented ChoanoV2 proteins showing highest identity to uncultured s Syndiniales (putative parasites) (34). Furthermore, the gene content of ChoanoV1 is highly distinct from the many available genome sequences from viruses of picoeukaryotes (35, 36) or other known algal viruses (18–20, 22, 23, 37) (SI Appendix, Fig. S3). Collectively, these results point to us having recovered an actively infected B. minor host cell in which ChoanoV1 had already replicated. After Canararpox virus, which infects birds (38), ChoanoV1 represents just the second giant virus identified with an opisthokont host (Dataset S1).

We next sought to recover a ChoanoVirus genome from another field site. Therefore, we exploited the low GC content observed in ChoanoV1 to sequence and assemble a related virus in an eastern North Pacific sample collected 200 km offshore 7 y before the Bicosta single-cell study (Fig. 1B). This sample was chosen for low %GC DNA enrichment on a density gradient, followed by deep sequencing, because environmental clone libraries showed that the B. minor 18S rRNA gene was present (100% identity) and vintage metagenomic data from the sample (7) contained ChoanoV1-like reads. The resulting ChoanoV2 assembly contained 89% of ChoanoV1 genes (average 94% amino acid identity), despite its fragmented nature resulting from traditional metagenome assembly limitations (SI Appendix, Fig. S44). Our discovery poised us to investigate the evolution, function, and importance of specific metabolic traits in viruses of a key group of opisthokonts or more generally, heterotrophic marine protists and broader ecological implications.

Evolutionary Analyses Establish a Distinct NCLDV Giant Virus Lineage. Preliminary analyses suggested the ChoanoViruses were NCLDVs, with about 20% of the ChoanoV1 predicted proteins and 23% of the more fragmented ChoanoV2 proteins showing highest BLASTp affiliations to NCLDV proteins (Fig. 1D and SI Appendix, Fig. S4B). For proteins that had BLASTp affiliations primarily to cellular life, most of those closest to eukaryotic proteins seemed to be conserved derived secondary acquisitions from hosts in past time (SI Appendix, Fig. S4C). Unfortunately, the paucity of genomic resources for marine eukaryotic viruses and marine protists themselves precludes statistically valid examination of potential horizontal or host-to-virus gene transfer (HGT) at a genome wide scale, and hence, we did not examine questions of origin globally. The other half of the ChoanoVirus proteins have not been seen in cellular organisms or viruses sequenced to date. Overall, these observations, including ~50% of proteins being unknown, are quite typical of newly sequenced NCLDV genomes (13, 39), at least at this stage in time, in which relatively few have been sequenced. Of these ChoanoVirus orphan genes, 70% were detected in metatranscriptomes that we sequenced from the eastern North Pacific, demonstrating expression (SI Appendix, Fig. S2).
To reconstruct evolutionary relationships, we used phylogenomic approaches to analyze proteins considered core to NCLDV genomes (40). We reexamined presence, absence, and copy number patterns for the 47 proteins previously proposed to be core (40). We next excluded, for example, fast-evolving proteins and proteins for which unclear paralogs existed within a single NCLDV genome, and thereby expanded the set of NCLDV proteins suitable for phylogenomics used in recent reconstructions (12) from 5 to 10 (Fig. 2A and Dataset S2). Phylogenomic reconstructions with the 2 protein sets provided similar topologies, with higher statistical node support in the 10 protein phylogeny (SI Appendix, Fig. S5). These reconstructions showed the ChoanoViruses belong to the extended Mimiviridae, comprising a divergent clade from those already established (12, 13). PolB reconstructions highlighted a large group of marine viral PolB, distinct from nonmarine Mimiviridae (Mimiviruses, Tupanviruses, Klosneuviruses) and CroV, when assembled metagenomic sequences from TARA Oceans (41) and Global Ocean Survey (GOS) (42) were searched and included (SI Appendix, Fig. S6). Within this broad marine group, the ChoanoViruses formed a supported clade that incorporated Pacific Ocean, Atlantic Ocean, and Southern Ocean sequences for which the viral hosts remain unknown. These analyses demonstrated the value of recovering viral genomes from uncultured hosts, which exposed here the unique ChoanoVirus lineage and its presence in multiple oceans.

ChoanoVirus Auxiliary Metabolic Genes and Biogeochemical Implications.

AMGs are host-derived genes carried by viruses that are not directly involved in viral replication but rather supplement or augment cellular functions within infected cells (5, 6). An important example in marine bacteriophages is oxygenic photosynthesis proteins that augment cyanobacterial photosynthetic machinery during infection (4). Although oxygenic photosynthesis-related proteins have not been found in eukaryotic viruses sequenced to date, the giant viruses encode a plethora of AMGs that augment cyanobacterial photosynthetic machinery during infection (4). Although oxygenic photosynthesis-related proteins have not been found in eukaryotic viruses sequenced to date, the giant viruses encode a plethora of AMGs that augment cyanobacterial photosynthetic machinery during infection (4). Although oxygenic photosynthesis-related proteins have not been found in eukaryotic viruses sequenced to date, the giant viruses encode a plethora of AMGs that augment cyanobacterial photosynthetic machinery during infection.

Fig. 2. Evolutionary relationships and functional aspects of the ChoanoVirus lineage. (A) Maximum likelihood phylogenomic reconstruction inferred from 10 proteins. Support > 80% (500 bootstrap replicates) is indicated (LG + C20 + F + G-PMSF model) (SI Appendix, Fig. S5), and host group coloring is as in Fig. 1A. ChoanoV1 (star; from M2 single-cell sort) and ChoanoV2 (from Station 67-70; low %GC-selected DNA with metagenomics) branched together in all reconstructions adjacent to an algal stramenopile virus AaV (when included) (SI Appendix, Fig. S5), for which placement appears influenced by long-branch attraction. (B) Total number of tRNAs (Left) and orthogroup functional categorization (heat map; EggNOG categorization) of ChoanoV1 and representative giant NCLDV (Dataset S1). The frequency of each category across the viral genomes determines x-axis ordering. (C) Distribution of functional categories in ChoanoV1 (via EggNOG) for all annotated proteins. (D) ChoanoV1 proteins with no orthologs in the NCLDV representative genome set. Note that, in pies in B to D, we have omitted fractions representing the EggNOG functional category “Unknown function,” but the values are shown as text on panels along with the total number of proteins with no significant database match.
transcription, translation, lipid biosynthesis, and transport of phosphate or ammonium (6–8, 43). Systematic analyses of Chao-
noVirus metabolic potential revealed a broad repertoire of such proteins, several types being enriched or unique in ChaoNoViruses relative to other NCLDV (Fig. 2 B–D and SI Appendix, Figs. S3 and S7A). Like other giant viruses, the ChaoNoVirus genomes encode proteins for augmenting host processes, including aminosyl-
RNA synthetases, photolyases, and proteins involved in signal transduction, replication, recombination and repair, cell wall 
genesis, and posttranslational modifications (Fig. 2B and SI Ap-
pendix, Fig. S7A) (6–13, 18–23). The ChaoNoVirus also encodes 22 tRNAs (Fig. 2B) such that tRNA numbers seem to roughly scale with genome size, with more being found in the larger genome-sized Tupanvirus from deep sea sediment (43) and less in the smaller genome-sized pelagic marine giant viruses TtV, 
Crov, PgV, and CeV (18, 20–22). Furthermore, the ChaoNoVirus tRNAs correspond to amino acid usage, suggesting preferential 
retention of those optimized for amino acid usage of virus over host, and 17 tRNAs are collocated in a single genomic region (SI Appendix, Fig. S7 B and C). Hence, the large ChaoNoVirus ge-
nomes encoded many proteins once considered unique to cel-
lular life, that now seem to be held in common across disparate 
giant viruses (10–13, 18–21).

Clustering based on presence and absence patterns of orthologous protein groups in NCLDV placed ChaoNoV1 ad-
jacent to the only other sequenced marine pelagic virus with a host that is a heterotrophic predator, Crov (SI Appendix, Fig. 
S3) (21). These 2 viruses were part of a broader cluster incor-
porating marine algal giant viruses, which appeared more similar to each other in their orthogroup presence and absence 

patterns than to nonmarine giant viruses or smaller viruses that infect marine algae. Many of the proteins making up these 
	hroghroups lack characterized functions or have only broad 

functional classification. Combined with the limited overall 
representation of giant virus lineages, these findings call for 
a major initiative to expand viral taxonomic sampling so that the 
significance of the presence and absence pattern observations could be estimated.

Comparison of ChaoNoV1 with other genome-sequenced viruses 

shows an enrichment in NCLDV orthologs involved in transport 

and metabolism of nucleotides, amino acids, and carbohydrates (Table S4 and SI Appendix, Table S1). ChaoNoV2 shows the same 
trend, although its more fragmented state precludes robust 

global ortholog comparisons. Even among ChaoNoVirus proteins 
lacking orthologs in other NCLDV, these functional categories 

are prominent (Fig. 2D and SI Appendix, Fig. S7D) and include 
a chitinase new to marine viruses that is present in both Chao-

NoViruses (SI Appendix, Fig. S8). Chitinase degrades the poly-
saccharide chitin, a component of zooplankton, some algae, and 

many other organisms, to labile saccharides readily consumed by 

marine microbes (44). This enzyme has been reported in a virus 
of the freshwater alga Chlorella (45) and viruses that infect in-

sects, specifically Lepidoptera (46). Our phylogenetic analyses 

placed moth virus chitinases in a clade with sequences from their 

Lepidoptera hosts within bacterial chitinases (potentially a com-

plex series of transfer events), while Chlorella virus and fungal 

chitinases grouped together (SI Appendix, Fig. S8). The ChaoNoVirus 

chitinase branched with opisthokont chitinases, suggesting po-
tential acquisition from a host of an ancestral opisthokont virus. 

Collectively, these results suggest that acquisition by each of the 
3 types of viruses occurred in independent events. From a func-
tional perspective, release of viral chitinase in Lepidoptera 

larvae is necessary for liquefaction, but the mechanism and overall 

roles during infection are unclear (46). The Chlorella virus chitinase has hypothesized roles in degrading the chitin-rich host cell wall (45). However, in contrast to moths and Chlorella, which have 

chitin as an abundant structural component, choanoflagellates lack known chitin-based structures, although they possess chitin synthase (47). Thus, ChaoNoVirus chitinase activity, potentially 

on prey material, alongside activities of viral carbohydrate metabo-

lism proteins may supply hosts with nutrition when choano-

flagellate feeding is impacted by the infection or other factors. 

Alternatively, a structural feature of choanoflagellate cells, such 
as the theca, may have an as yet unrecognized chitin-containing 

composition, in which case, the viral chitinase may operate in 

host degradation. Regardless, the organic matter released from 

the lysed host will provide more readily available carbon sources, 
such as labile saccharides, to marine microbes than will hosts 

infected and lysed by viruses that lack these enzymes or other 

forces of mortality. As such, in addition to release of cellular 

substrates on lysis, viral infection may “prime” substrates to be 

acceded more readily, potentially altering the microbial loop (48) 
in terms of rate and fate of the cellular material remineralization in 

the ocean.

Viral Rhodopsin Sequence Characterization. Strikingly, we also 

identified 3 distinct putative rhodopsins in each ChaoNoVirus 
genome (Dataset S4). Rhodopsins are integral membrane proteins 

that capture or sense sunlight using a bound retinal chromophore 
in cellular organisms (49). Microbial (type-I) rhodopsins include a 

variety of light-driven ion pumps (including H$, Cl, Na, K, 

and sensory receptors involved in signal 

transduction (including Sensory Rhodopsins I and II, which have 

been shown to regulate phototaxis in some protists) (50–52). 

Additionally, heliorhodopsins are considered distantly related 

family members and are thought to have light-sensing activities 

(53). Type-1 proton-pumping rhodopsins are widespread in het-

trorophic marine bacteria (54, 55), increasing survival during 

starvation when illuminated (50), and homology-based studies 

postulate that some eukaryotic algae have similar systems (56). 

Phylogenetic analyses show that the ChaoNoVirus rhodopsins split 

into 2 type-1 groups composed primarily of metagenomic se-
quences, which collectively exhibit distinct phylogenetic histories 

from those in cellular organisms (Fig. 3A). Among viruses with 

known hosts, the only other rhodopsin reported is in the giant 

virus PgV, which infects the marine haptophyte alga P. globosa 

(18, 57), and belongs to a clade that includes 1 of the 3 Cho-

anoVirus rhodopsins (Fig. 3F). We term these 2 groups (clades) 

that have this distinct history from those of cellular organisms 

VirR Group-I and VirR Group-II. Importantly, all VirR are highly 

diverged from a microbial rhodopsin clade harboring the fu-

sion protein Rho-PDE that is present in the genome-sequenced 

choanoflagellate Salpingoeca rosetta, wherein it exhibits light-
dependent phosphodiesterase activity (58, 59). While we identi-

fied homologs of Rho-PDE in 2 transcriptome-sequenced 

choanoflagellate species (Fig. 3A), it is absent from genome-

sequenced Monosiga brevicollis and is not found in transcriptome 

assemblies from 17 other choanoflagellate species or in the 

Bicosta 4-well partial genome assembly. Overall, the ChaoNoVirus VirR pro-

tiens do not seem to be derived from extant opisthokonts. Indeed, 

the tree topology and additional testing (SI Appendix) suggest that 
rhodopsin may have been present in an ancestral virus before host-

gene transfer into disparate algae and heterotrophs (Fig. 3A).

Several marine studies have now reported putative viral rhod-

opsins in traditional metagenomic data—for which the viral 

hosts are by default unknown (57, 60–62). The function of these 

is not clear, since often, they lack the amino acid motifs that have 

been shown through biochemical characterization of various 
type-1 rhodopsins to generally confer functional differences. 

Indeed, the function of type-1 rhodopsins can sometimes be 

inferred from 3 key amino acid residues (referred to as motif 

sequences), such as the proton (DTD, DTE) and chloride (TSA, 

NTQ) pump motifs (49). In bacteriorhodopsin (BR), the resi-

dues that make up the motif are at positions 85, 89, and 96. BR 

has been biochemically characterized to function as a proton 
pump, wherein the D85 acts as a proton acceptor. T89 forms a
hydrogen bond with D85, and D96 acts as a proton donor in this DTD motif rhodopsin (49); other motifs have proton pumping or other functions (SI Appendix, Table S1). Previously detected VirR sequences in PgV and GOS were hypothesized to have sensory roles in host phototaxis (57) or to be involved in light sensing in the host (61), because some lack the retinylidine Schiff base proton donor carboxylate, which has been taken to be essential for proton transport, similar to sensory rhodopsins (63). Unlike the observed YML motif, the DTS and DTV motifs have been observed in environmental sequences inferred to come from viruses at Station ALOHA in the North Pacific Gyre (60), in the Red Sea (61), and in coastal sediments (62). Our results provided evidence for VirR proteins being in viruses of heterotrophic protists and for a single virus having both Group-I and Group-II viral rhodopsins. However, the amino acid differences for all VirR from biochemically characterized proteins alongside their long-branch lengths (Fig. 3A) left uncertainty regarding function, as is the case for many proteins identified in marine metagenomic studies.

**Viral Rhodopsin Activity and Structure.** Because of the presence of VirR DTS in the only pelagic marine giant viruses with known hosts (i.e., the uncultured ChoanoVirus and the cultured algal virus PgV), we next turned to laboratory experiments to examine the structure and function of this VirR protein. Heterologous expression in Escherichia coli of the homolog from PgV caused substantial light-induced acidification of retinal-amended medium up on illumination, demonstrating that it has proton-pumping capabilities (Fig. 3B). This clear pH change was abolished by protonophore addition. VirR DTS predominantly possessed all-trans retinal (SI Appendix, Fig. S104). At neutral pH, the Schiff base linkage was protonated \( pK_a = 7.8 \), and a counterion residue was deprotonated \( pK_a = 3.6 \) (SI Appendix, Fig. S10 B and C). We analyzed the photocycle of VirR DTS, demonstrating that time constant of
SI Results and Discussion

β-proton-pumping activity on illumination raises questions regarding the natural source of the carotenoids needed to produce the light-harvesting chromophore, retinal (50, 51), especially in a nonphotosynthetic host, like Bicosta. Most algae, including PyGhVs host Phaeocytis, biosynthesize the required pigment, β-carotene (and related carotenoids), as well as the retinal-producing carotenoid cleavage oxygenase (Blih) (Fig. 4). However, most heterotrophic eukaryotes, including animals, do not bio-synthesize β-carotene, instead acquiring carotenoids through diet. As expected, cultured genome-sequenced chaoanoFLAGellates encode only early steps that overlap between sterol and carotenoid biosynthesis and a final cleavage enzyme (Dataset S5). Likewise, BLASTx searches against the Bicosta 4-well partial genome assembly failed to recover carotenoid biosynthesis enzymes. Remarkably, the ChaoANOvirus genome analyses exposed both the β-carotene biosynthesis pathway and Blih, with 4 proteins being adjacent to one another, similar to the pathway in bacteria (76) (Fig. 4, SI Appendix, Fig. S124, and Dataset S5). Eastern North Pacific metatranscriptomes confirmed expression of all components (Fig. 4). Thus, while the algal virus relies on its host to biosynthesize the pigments used in light-energy transfer, ChaoanoVirus encode the complete rhodopsin-based photosystem.

The evolutionary origins of the retinal biosynthesis proteins in the ChaoanoVirus remain unclear. They seem to derive from

---

**Fig. 4.** Functional attributes of ChaoanoVirus include chromophore biosynthesis. Shown are carotenoid pathway components and final retinal-forming cleavage step in genome data from haptophytes (Phaeocytis antarctica and Chrysochromulina representing P. globosa, which lacks genome data), choanoFLAGellates (M. brevicollis and S. rosetta), and relevant viruses and in metatranscriptomes. The stars indicate the two ChaoanoVirus genomes and a metatranscriptome from the station where ChaoanoV1 was recovered. The circle indicates the only cultured virus with a rhodopsin. *These taxa lack Blih but have RPE65 used for retinal production (e.g., in vertebrates and related). Detection in Pacific metatranscriptomes based on reads recruited to ChaoanoV1 by BLASTx (e-value < 10^{-5}); those that mapped at >95% nucleotide identity are indicated in Dataset S5. OPP, pyrophosphate group; FPP, farnesyl diphosphate; GGPP, geranylgeranyl diphosphate.
archaea (phytoene synthase) or marine bacteria (phytoene desaturase) or are too divergent for robust phylogenetic conclusions (lycopene cyclase, Blh) (SI Appendix, Fig. S12 B–E). In each case, the respective ChoanoV1 and ChoanoV2 proteins clustered together, indicating their common origin. Rhodopsin-bearing bacterial or archaeal lineages with retinal biosynthesis-related genes are each thought to have acquired them together as a unit by HGT (77). However, despite the 4 ChoanoVirus retinal biosynthesis genes being colocated in the genome, long branch lengths and incomplete taxonomic sampling make it unclear whether these proteins were accumulated over time or acquired in a single HGT event, although the latter scenario seems most likely.

Viral Rhodopsins in the Global Ocean. Our studies now provided the structure and function of VirR<sub>DTS</sub>, but the frequency of VirR genes as a whole in nature remained unclear. Prior analyses of viral rhodopsins in traditional metagenomic data focused on individual locations, specifically the Red Sea (61) and Station ALOHA (60), or had relatively shallow sequencing depth, such as GOS (57). It should be noted that one other metagenomic study of coastal sediments reported 30 VirR (62) that were similar to PgV VirR<sub>DTS</sub> and to the VirR metagenomic sequences from Organic Lake that have been suggested to come from another (currently unknown) haptophyte algal virus. These partial metagenomic sequences (62) may well, therefore, represent remnants of a senesced (infected) haptophyte bloom exported to sediments at 11- to 50-m bottom depth. Our searches of TARA metagenomic assemblies greatly expanded the global VirR repertoire (Fig. 5A and Dataset S6). Assembled VirR proteins were recovered at 37 of 39 TARA photic-zone sampling sites examined, and only at photic-zone depths in Station ALOHA profiles that included deep ocean sampling (Fig. 5B), as expected for a sunlight-dependent energy transfer system. Motifs were diverse; however, the DTS motif was the most common vertically and globally (Fig. 5B and C).

![Diagram of Viral Rhodopsins in the Global Ocean](image)

Fig. 5. Viral rhodopsins are distributed across the world oceans. (A) Environmental VirR motifs and cluster analysis of sequences (CLANS)-based relationships between full-length proteins recruited from TARA Oceans and Station ALOHA data. (B) Normalized VirR depth distributions in the North Pacific Gyre determined by mapping metagenomic reads to VirR gene assemblies from ALOHA (60) and VirR motif distributions (pies; colors as in A). (C) VirR motifs in TARA metagenome assemblies having >300,000 contigs from 5 m (304 full-length sequences in total) and samples reflecting a true deep chlorophyll maximum (43 full-length sequences in total), which typically occurs in stratified open ocean water columns between 75 and 130 m. (D) Correlation between Mimiviridae PolB and VirR across analyzed TARA samples. (E) Normalized VirR and Mimiviridae PolB frequencies in TARA assemblies (with >300,000 contigs).
Phylogenetic analyses of the assembled sequences that we recovered from deeply sequenced TARA samples and other metagenomic studies showed 3 statistically supported clades within Group-I VirR proteins: one harboring only DTT (64%) and DTS (36%) motifs, another dominated by DTS (72%) and DTT (17%) and having 4 other motifs represented, and a small clade composed of 5 different motifs (SI Appendix, Fig. S13). Likewise, the Group-II VirR proteins delineated into 5 clades: 2 of which are dominated by the DTT motif, with 100 and 73% DTT, respectively. The latter had 4 additional motifs, including 18% DSV. The smallest Group-II VirR clade had just 4 sequences and 3 motifs. In total, our survey of VirR revealed 8 previously unreported motifs, in addition to the observed DTS, DTV, and DTT, and indicated that motifs generally grouped in a manner connecting to the evolutionary history of these proteins. Functional characterization will be important for understanding the cell biological implications during infection as well identification of the corresponding natural hosts.

Finally, more than 99% of environmental VirR had 1 of 2 amino acids (M, L; Met89 in VirRSTS) that confer increased green light absorption relative to blue, in contrast to the multiple wavelengths used by marine prokaryotic rhodopsins (78). With VirR proteins being in only 2 genome-sequenced viruses with known hosts, we could not parse the metagenomic sequences recovered into percentages coming from viruses of photoplankton vs. heterotrophs. However, our recruitment of assembled PolB genes from TARA (12,684 in total) recovered 6 times more than ancient TARA 454-metagenome analyses (13, 79), and protein similarity networks identified 1,026 PolB as being from Mimiviridae (SI Appendix, Fig. S14). The computed ratio of VirR to Mimiviridae PolB was 0.7, suggesting that VirR is a common component of many giant viruses in sunlit ocean environments (Fig. 5 D and E and Dataset S6).

Conclusions
Predatory protists have important ecosystem roles in the transfer of organic carbon (prey-based) to higher trophic levels in addition to their top-down control of microbial cells (15). Here, using both cultivation-independent and laboratory methods, we performed cross-scale analyses—from the sequencing of giant virus genomes and their host to evolutionary relationships, functional attributes, and presence in broader ocean samples. Our studies reveal a virus of a widespread group of marine predatory protists related to metazoans, the choanoflagellates.

Alongside CroV, which infects a host from different eukaryotic supergroup, the ChaoanoViruses bring the number of genomes available from giant viruses that infect known predatory hosts in pelagic marine environments to a total of 3. While the ChaoanoViruses share cellular life-like proteins observed in nonmarine viruses, overall gene retention patterns seem tuned to habitat, akin to the marine nature of VirR and potentially, to host trophic mode. Additional studies that target uncultured eukaryotes and coassociated entities will more fully expose viral mechanisms for modulating the host environment and the relative strengths of evolutionary history vs. environment on viral gene content.

The AMGIs identified in the ChaoanoVirus genomes are of particular import for the intracellular replication environment (the host) and marine host–environment interactions. During infection, the host can be considered a “virocell,” wherein the viruses modulate intracellular processes and their own replication (39). It is during this time that the ChaoanoVirus has multifaceted roles in utilizing horizontal compounds in a manner that could facilitate rhodopsin-based phototroph, providing mechanisms for reshaping host nutrition, physiology, and the quality of remained organic matter. These findings raise questions on the dynamics of host–virus interactions and potential for transient mutualism. In addition to chitinase and enzymes for transport and metabolism of multiple organic molecules, the ChaoanoViruses encode an unprecedented viral multigene pathway for a rhodopsin-based photosystem. Our characterization of a putative rhodopsin encoded by eukaryotic viruses, crystal structure, and biochemical assays identifies a mechanism for viral-induced light-driven energy transfer. While VirRSTS proton pumping may facilitate energy transfer in connection with host-derived adenosine triphosphate (ATP) synthases, the plethora of other VirR-motif types observed herein remains to be functionally and structurally characterized. This will involve identifying the host membrane in which the rhodopsins localize, be it mitochondrial, plasma, or endomembrane systems, as well as the duration, timing, and precise functional role during infection. Given the variety of viral rhodopsins, it seems that a diverse suite of roles in host manipulation awaits discovery, potentially involving host signaling, photomotility, and action potential for flagellar beating or lysis as well as photoheterotrophy. While type 1 rhodopsins are relatively widespread in marine bacteria (53, 78), they appear to be absent from marine phage genomes. In contrast, our studies, and their extension by analysis of global metagenomes, indicate that VirR is a common aspect of how giant viruses of eukaryotes reshape host physiology and potentially energy transfer in both heterotrophic and photosynthetic marine protists.

Materials and Methods
Detailed materials and methods are provided in the SI Appendix. This includes information on field work, flow sorting, whole-genome amplification, amplicon library construction and sequencing, data processing, assembly and analysis, including gene predictions, as well as phylogenetic, crystallization and phylogenomic analyses. The final ChaoanoV1 genome assembly was assembled from 13,802,665 quality-filtered reads, and viral contigs were differentiated from the cellular assembly by tetrancleotide frequency and GC content. The ChaoanoV1 assembly consisted of 11 contigs with average coverage of 215 ± 157×. For eastern North Pacific Ocean gene expression analyses, reads from metatranscriptomes were mapped to ChaoanoVirus genome at an average coverage of 1.82 ± 1.63. For 18S V4 amplicon sequencing, we had on average 131,385 ± 121,027 amplicons well (the lowest number being 1,037) clustered at 99% and classified via the Prostian Ribosomal Reference Database. Rhodopsin functionality of viral VirRSTS was determined via heterologous expression in E. coli. The VirRSTS crystallization samples were produced by a cell-free system, and crystals were grown using the in mesc apo acceptor. Accession numbers and DOI for alignment and tree files are available in Dataset S2.

Acknowledgments. We thank the captain and crew of the R/V Western Flyer, and we thank M. Ares, L. Gómez-Consarnau, and K. Bergauer for discussions. We are grateful for the availability of Monterey Bay Time Series chlorophyll data through the Monterey Bay Aquarium Research Institute (MBARI). We thank a United States Department of Energy Joint Genome Institute Technology Development Program grant for some initial sequencing. Support came from the Canon Foundation (S.Y., T.H., T.K.-S., M.S., and W.I.), Japan Society for the Promotion of Science KAKENHI Grants 16H06279 (to S.Y. and W.I.) and 18H04136 (to S.Y. and W.I.), Gordon & Betty Moore Foundation Grants GBMF3307 (to T.A.R., P.J.K., A.E.S., and A.Z.W.) and GBMF3788 (to A.Z.W.), MIBARI (A.Z.W.), and GEOMAR Helmholtz Centre for Ocean Research Kiel (A.Z.W.).

References