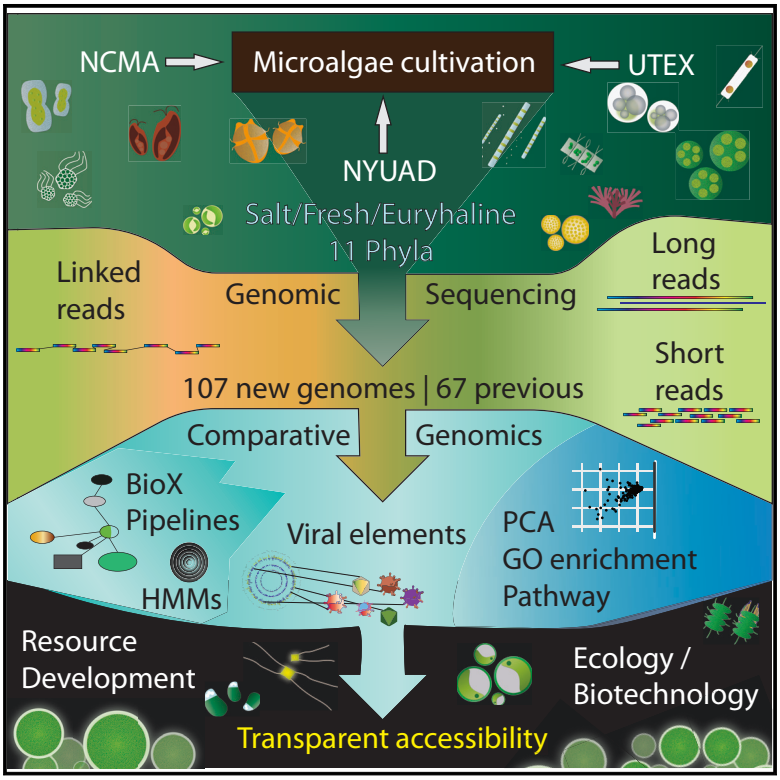
Article

Cell Host & Microbe

Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution

Graphical Abstract



Highlights

- 107 microalgae from diverse phlya and environments were cultured and sequenced
- Membrane-related protein families were significantly enriched in saltwater species
- Nuclear-related protein families were significantly enriched in freshwater species
- More than 90,000 viral-origin sequences in 184 algal genomes were discovered

Authors

David R. Nelson, Khaled M. Hazzouri, Kyle J. Lauersen, ..., Michael W. Lomas, Khaled M.A. Amiri, Kourosh Salehi-Ashtiani

Correspondence

drn2@nyu.edu (D.R.N.), ksa3@nyu.edu (K.S.-A.)

In Brief

Nelson et al. combine high-throughput cultivation, genome sequencing, and bioinformatics strategies to generate resources for gaining new perspectives on microalgae from diverse lineages and environments. This resource greatly expands the available collection of algal genomes and provides insight into how viral sequence integration shaped major algal clades.

Nelson et al., 2021, Cell Host & Microbe *29*, 1–17
February 10, 2021
https://doi.org/10.1016/j.chom.2020.12.005

CellPress

Cell Host & Microbe

CellPress

Article

Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution

David R. Nelson,[1,*] Khaled M. Hazzouri,[2,7] Kyle J. Lauersen,[3] Ashish Jaiswal,[4] Amphun Chaiboonchoe,[4] Alexandra Mystikou,[1] Weiqi Fu,[4] Sarah Daakour,[1] Bushra Dohai,[4] Amnah Alzahmi,[1] David Nobles,[5] Mark Hurd,[6] Julie Sexton,[6] Michael J. Preston,[6] Joan Blanchette,[6] Michael W. Lomas,[6] Khaled M.A. Amiri,[2,7] and Kourosh Salehi-Ashtiani[1,4,8,*]

[1]Center for Genomics and Systems Biology, New York University Abu Dhabi, Abu Dhabi, UAE
[2]Khalifa Center for Genetic Engineering and Biotechnology (KCGEB), UAE University, Al Ain, Abu Dhabi, UAE
[3]Biological and Environmental Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia
[4]Division of Science and Math, New York University Abu Dhabi, Abu Dhabi, UAE
[5]UTEX Culture Collection of Algae at the University of Texas at Austin, Austin, TX, USA
[6]National Center for Marine Algae and Microbiota, East Boothbay, ME, USA
[7]Biology Department, College of Science, UAE University, Al Ain, Abu Dhabi, UAE
[8]Lead Contact
*Correspondence: drn2@nyu.edu (D.R.N.), ksa3@nyu.edu (K.S.-A.)
https://doi.org/10.1016/j.chom.2020.12.005

Summary

Being integral primary producers in diverse ecosystems, microalgal genomes could be mined for ecological insights, but representative genome sequences are lacking for many phyla. We cultured and sequenced 107 microalgae species from 11 different phyla indigenous to varied geographies and climates. This collection was used to resolve genomic differences between saltwater and freshwater microalgae. Freshwater species showed domain-centric ontology enrichment for nuclear and nuclear membrane functions, while saltwater species were enriched in organellar and cellular membrane functions. Further, marine species contained significantly more viral families in their genomes ($p = 8e{-}4$). Sequences from Chlorovirus, Coccolithovirus, Pandoravirus, Marseillevirus, Tupanvirus, and other viruses were found integrated into the genomes of algal from marine environments. These viral-origin sequences were found to be expressed and code for a wide variety of functions. Together, this study comprehensively defines the expanse of protein-coding and viral elements in microalgal genomes and posits a unified adaptive strategy for algal halotolerance.

Introduction

Viruses of microbial eukaryotes have extreme diversity and considerable influence in local (Correa et al., 2016) and global (Gregory et al., 2019) ecosystems. Nearly 200,000 marine viral populations have been found to be active and segregate through five distinct zones throughout global oceans (Gregory et al., 2019). Viral infection of microalgae can influence oceanic aerosol release, including dimethyl sulfide (DMS) (Trainic et al., 2018) and atmospheric gas composition (Bonetti et al., 2019), via rapid destruction of microalgal populations (Sorensen et al., 2009). Viral-mediated cell death in coccolithophore microalgae leads to massive carbonate depositions from skeleton sedimentation, which significantly contributes to CO_2 sequestration (Jover et al., 2014; Nilsson, 2019; Ruiz et al., 2017; Villain et al., 2016). Finally, life support roles of microalgae, including sustaining coral reef species (Correa et al., 2016) and producing atmospheric oxygen are modulated and occasionally threatened by their viral predators. Despite these roles, pan genomic-level effects of viruses on microalgae from different lineages and environments are unknown. The information on the presence of viral elements in algae is currently tenuous, particularly regarding their distribution in various habitats, including marine and terrestrial ecosystems.

Algae are a polyphyletic group of photosynthetic microorganisms that provide foundational nutrients for every biome on Earth and play essential roles in the exchange of molecules, including O_2, CO_2, and DMS, between the atmosphere and biosphere. Although microalgae are fundamental to global ecosystems and have the potential for sustainable biotechnological development, they have received far less research attention than other microbes. For example, more than 30,000 bacterial genomes, but only 62 algal genomes, have been sequenced (ncbi.nlm.nih.gov). Genome sequences of most algal species are unknown, with sparse representatives from some clades, such as the Chromeridia and Myzozoa, and no representatives from other clades, including Euglenozoa. Algal culture collection centers, including the national center for algae and microbiota (NCMA; ncma.bigelow.org) and the culture for collection of algae at the University of Texas, Austin (UTEX; utex.org), among other centers,

Please cite this article in press as: Nelson et al., Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host & Microbe (2020), https://doi.org/10.1016/j.chom.2020.12.005

host thousands of microalgal species from around the world. The disparity between the availability of algal species in collections and the understanding of their genomic contents can, presently, be resolved through high-throughput sequencing.

Expanded genomic sequence databases would allow for statistical resolution of essential questions of microalgal evolution and niche habitation. The viral contribution to algal genomes has not been studied on a large scale, but evidence suggests that viruses have contributed to their hosts' adaptation to different environments. Gene shuffling between algal hosts and viruses has led to the emergence of giant viruses that incorporate entire biosynthetic pathways, sourced from their algal hosts, into their enormous genomes (Abrahão et al., 2018; Filée, 2015; Filée and Chandler, 2010; Moniruzzaman et al., 2020; Schulz et al., 2020; Van Etten, 2003; Van Etten and Meints, 1999; Xiao and Rossmann, 2011; Yamada, 2011; Yoosuf et al., 2012). As examples, the *Klosneuvirus* (1.57 Mb), *Bodo saltans* virus (1.39 Mb), and the *Megavirus chilensis* (1.23 Mb) genomes encode for 1,000–1,300 proteins each, many that are host-derived. When host specificity expands, viral genes can be transferred to distantly related organisms and confer specific evolutionary adaptations, such as the introduction of new metabolic pathways that facilitate the assimilation of fresh nutrients (Piacente et al., 2014) or abiotic stress-resistance genes that promote survival in niche habitats.

The host specificity of microbial, eukaryotic viruses displays erratic patterns; algal viruses can cross-kingdom boundaries and infect mammals. For example, the Chlorella virus ACTV-1 infects human oropharynx tissues and causes reduced mental performance in mice (Yolken et al., 2014). A comprehensive survey of viral elements in algal genomes would provide foundational data that are necessary to detect other algal viruses to human transmissions. The recent COVID-19 pandemic is believed to have zoonotic origins, but in many cases, the mechanisms underlying cross-species host specificity are not well understood (Cohen, 2020; Cohen and Kupferschmidt, 2020a, 2020b; Service, 2020). Genomic analysis is essential to discover host specificity (Kupferschmidt, 2020). Additional genomic sequences from new algal species and characterization of extant viral elements in their genomes, as records of past infections, would resolve urgent questions about other possible environmental sources of viruses. Here, we sequenced more than 100 new microalgal genomes from 11 phyla to comprehensively define the expanse of protein-coding and viral elements in microalgal genomes. This study sheds light on microalgal evolution by defining clade- and environment-specific protein-coding and viral genomic elements present in microalgae.

Results and Discussion

Resource selection and generation

To facilitate fundamental and applied research projects using broad subsets of representative microalgal species, we selected and performed whole-genome sequencing on 107 different species of microalgae from the UTEX and NCMA culture collection centers and our New York University Abu Dhabi (NYUAD) isolate collection (Figure 1; Table S1). All genomes assembled *de novo* in this study from our monocultures are hosted at National Center for Biotechnology Information (NCBI) under the Bioproject accession PRJNA517804 (see Data S1 for assemblies, and Figures 2 and S1, Table S1, and Data S2 for quality controls and reference species validations). A master dataset, "Data S1," contains sub-Data S1–S13, including all of the data and results produced in this manuscript. All of the data produced in this project are available to the reader as bulk archive downloads, including genome assemblies, predicted CDSs and proteins, protein families (PFAMs) and viral family (VFAM) predictions, as well as VFAM-coding sequence (CDS), endogenous viral-origin Pfams (EVOPs), and other data described in the manuscript.

Our selection of species was aimed at broad representation across microalgal phyla, with representatives from the Rhodophytes, Chlorophytes, Haptophytes, Cercozoans, Ochrophytes, Dinophyta, Euglenophyta, Heterokonta, Streptophyta, and Chromerida. Microalgae from these phyla vary considerably in size and content (see also Table S1). For example, the symbiont dinoflagellate species that support coral reefs have large genomes (1 Gbps+; Lin et al., 2015; Shoguchi et al., 2013); the free-living picoeukaryotes, which live in the same aquatic environment, have highly condensed genomes (10–13 Mbp, Blanc-Mathieu et al., 2017; Nelson et al., 2019). Thus, the genomic diversity of microalgae in this work, combined with a shared phenotype (i.e., photosynthetic, eukaryotic, microbial), will provide the needed information to address questions about convergent evolution in photosynthetic microbes.

In addition to these newly sequenced algal genomes (n = 107), we used available microalgal genomes from the NCBI and Phytozome to investigate genomic differences between microalgae from different habitats and clades (Table S1). The increased genomic assembly sample size (n = 174) allowed for a statistically significant resolution of comparative genomics questions, such as the impact of viral sequence contribution on microalgal evolution. We find distinct differences in the occurrences of multiple viral elements between salt and freshwater microalgae. Viral sequences correlated with functional domains and differed with phylogeny and environmental conditions, suggesting that they are core, defining features of microalgal lineages. Here, we outline how viral sequence acquisition in diverse microalgal groups underpins niche-specific biological processes, including core saltwater-specific proteins involved in membrane reinforcement.

Protein family correlation by niche and lineage

A Pearson's correlation of protein family (PFAM) domain counts by species demonstrated that microalgae from different lineages cluster by their environment irrespective of their phylogenetic affiliation (Figure 1; Data S4 and S5). This result indicated that the convergent evolution of hidden Markov model (HMM)-inferred functions occurred across widely divergent lineages of microalgae. Bi-clustering of PFAM count arrays revealed a subset of genes with high domain copy numbers (dCNs) in saltwater species (Figure 1C). Freshwater species were more diverse with regard to PFAM count distributions and types (Figure 1D). These characteristics were also indicated by the ranges in Pearson's coefficient values and more dispersed clustering in t-distributed stochastic neighbor embedding (tSNE, Figure S2). Freshwater Chlorophytes (FC), including Chlamydomonas, have higher basal nucleotide diversity (Flowers et al., 2015). We found that all freshwater lineages sampled had higher functional diversity. This result is consistent with relaxed selection in freshwater habitats. A similar divergence has been seen in the

Please cite this article in press as: Nelson et al., Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host & Microbe (2020), https://doi.org/10.1016/j.chom.2020.12.005

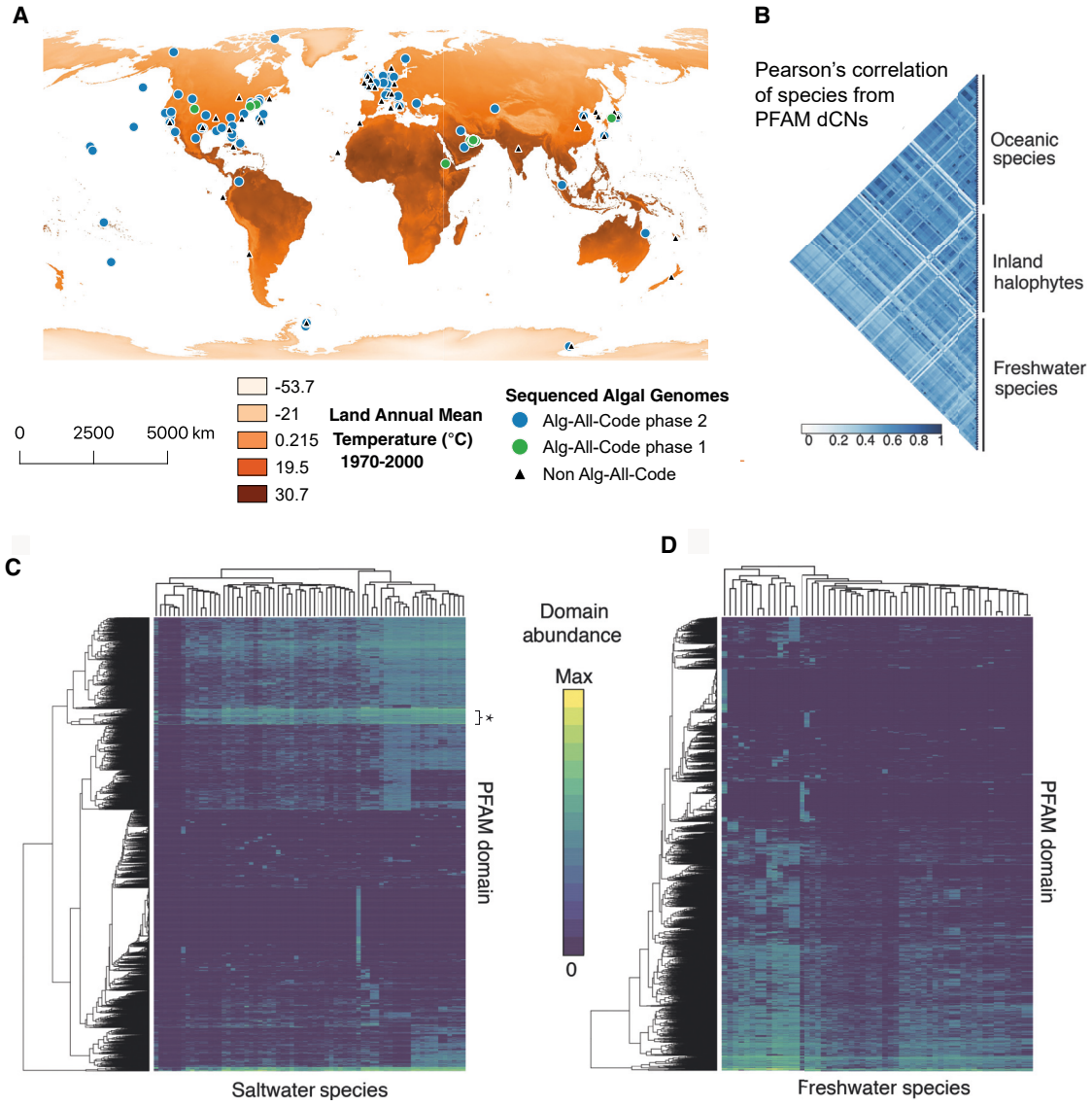


Figure 1. Geographical sources of species and hidden Markov-model-based analysis for sequenced microalgal genomes

Figure360▷ For a Figure360 author presentation of this figure, see https://doi.org/10.1016/j.chom.2020.12.005

(A) Isolation origin sites for microalgae (genomes from NCBI, triangles; our previous work [green circles, algalICODE phase I; Nelson et al., 2019, 2017]; sequenced for this project [blue circles, algalICODE phase II; see Table S1]). Eleven species were re-sequenced as internal controls for validation (see Table S1; Data S1).

(B) Similarity matrix based on Pearson's correlation scores for PFAM counts from microalgal genomes. Species from similar habitats clustered together despite originating from distant clades (e.g., marine chlorophytes and Ochrophytes). PFAM count arrays from saltwater species were more highly correlated than arrays from freshwater species; this functional divergence is also in (D).

(C) Hierarchical bi-clustering of PFAM counts for saltwater species. Saltwater species had high dCNs for a subset of PFAMs (indicated by the asterisk). In contrast, in freshwater species, a subset of species had higher dCN of several PFAMs (Data S5).

(D) Heatmap created using the same algorithm as for freshwater species. Groups of PFAMs and species distinctly clustered in saltwater and freshwater species, respectively. Freshwater species were more divergent based on function and did not share a distinctive cluster of PFAMs.

terrestrial expansion of diatoms from marine habitats into freshwater lakes, rivers, and estuaries (Alverson et al., 2011; Alverson and Theriot, 2005; Nakov et al., 2018b; Onyshchenko et al., 2019). For the saltwater species, our results suggested that constriction by ionic conditions forced convergence toward the PFAMs with increased dCNs shown in Figure 1C. Every member of this subset of PFAMs examined in a Uniform Manifold Approximation Projection (UMAP [McInnes et al., 2018]) had at least one corresponding viral-related neighbor (Figure S3).

Algal genomes contain lineage-specific viral sequences

We used HMMs to survey CDSs for VFAM domains that may be independent proteins or partial internal domains of proteins with incorporated cellular roles with an evolutionary relationship to

Please cite this article in press as: Nelson et al., Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host & Microbe (2020), https://doi.org/10.1016/j.chom.2020.12.005

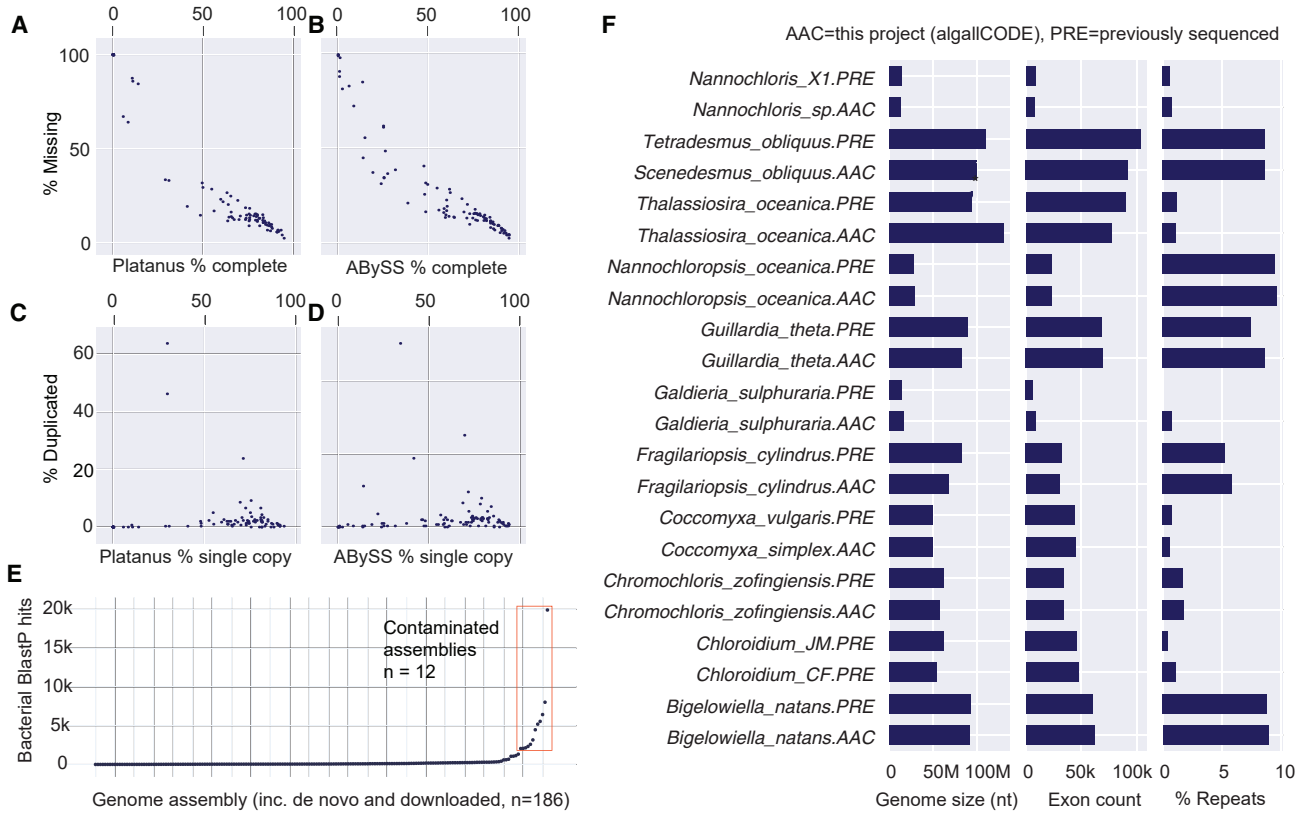


Figure 2. Assessment of *de novo* assemblies and contamination screening by alignment with bacterial proteins

(A–D) Based on evolutionary-informed expectations of the gene content of near-universal single-copy orthologs (BUSCO) metrics for the *de novo*-assembled genomes from short reads (n = 101). The BUSCO metrics are complementary to technical metrics included in the genome QUality ASsessment Tool (QUAST) results (Table S1). Percent complete BUSCO orthologs (x axis) versus percent missing or incomplete genes (y axis) in (A) Platanus and (B) ABySS assemblies. For Platanus, seven assemblies had evidence of poor quality based on BUSCO results and assembly size compared to expected sizes from close relatives. Percent single copy versus BUSCO genes with detected duplicates in (C) Platanus and (D) ABySS *de novo* assemblies.

(E) BLASTP hits of predicted microalgal proteins with bacterial species were used as markers for determining contamination levels in assemblies. We observed a wide range of bacterial BLASTP hit counts, with one assembly, *Entomoneis* spp., yielding more than 20,000 hits at an E value of <1e-9. Eleven other species had more than 3,000 bacterial hits at this E value cutoff, and this hit count was used as the threshold to remove assemblies that were highly contaminated from downstream comparative analyses.

(F) Examples of resequencing validations. Species with reference genomes (Table S1) were independently cultured and sequenced; sequencing reads were assembled *de novo*, and assembly statistics comparisons are shown. All whole-genome sequencing reads in this study are hosted at NCBI, Bioproject: PRJNA517804.

origin (Skewes-Cox et al., 2014) (Figures 3 and 4; Table S2). This approach was used to describe new viral ecology dynamics for humans (Bzhalava et al., 2018) and may potentially be used to address new viral threats without established solutions (Cohen and Kupferschmidt, 2020a). In this study, we used it to find differences in viral element content among microalgae from different lineages and habitats. Multiple lineages of microalgae were found to contain remnants of infections from virus families (Figure 3; Data S6 and S7). These sequence records show that the host range for these virus families may be broader than previously appreciated.

Each algal phylum had distinct collections of VFAMs that clustered according to Pearson's correlation scores from domain counts (Figure 3). Clades whose members share environmental niches, such as the open ocean (Chlorophyts and Ochrophytes [e.g., picoeukaryotes and diatoms]) or within corals (Myzozoa and Chromerida [e.g., Dinoflagellates and Chromerids]), clustered together according to their VFAM domain counts. Genes containing shared VFAMs (VFAM-CDSs) often had high-confidence BLAST matches to their counterparts in unrelated clades. For example, B-type asparagine synthetase chloroplast precursor (Micromonas pusilla CCMP1545) included homologs from Porphyridium purpureum (Rhodophyte, E value = 3.00E-120), *Porphyra umbilicalis* (Rhodophyte, Evalue=5.00E-113), *Emiliania huxleyi* CCMP1516 (Coccolithophore, E value = 9.00E-102), and *Fragilariopsis cylindrus* CCMP1102 (Diatom, E value = 5.00E-101). This CDS had vFam_1034 (E value = 6.7E-77), which includes highly similar sequences from *Micromonas sp.* RCC1109 virus MpV1, *Cafeteria roenbergensis* virus BV-PW1, and *Phaeocystis globosa* virus. Our data suggest that unrelated species have acquired virus-sourced genes as a result of competition in shared niches.

Our set of 91,757 VFAM-CDS revealed only 720 proteins with putative roles in DNA transposition (0.78%). Many transposons

Please cite this article in press as: Nelson et al., Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host & Microbe (2020), https://doi.org/10.1016/j.chom.2020.12.005

CellPress

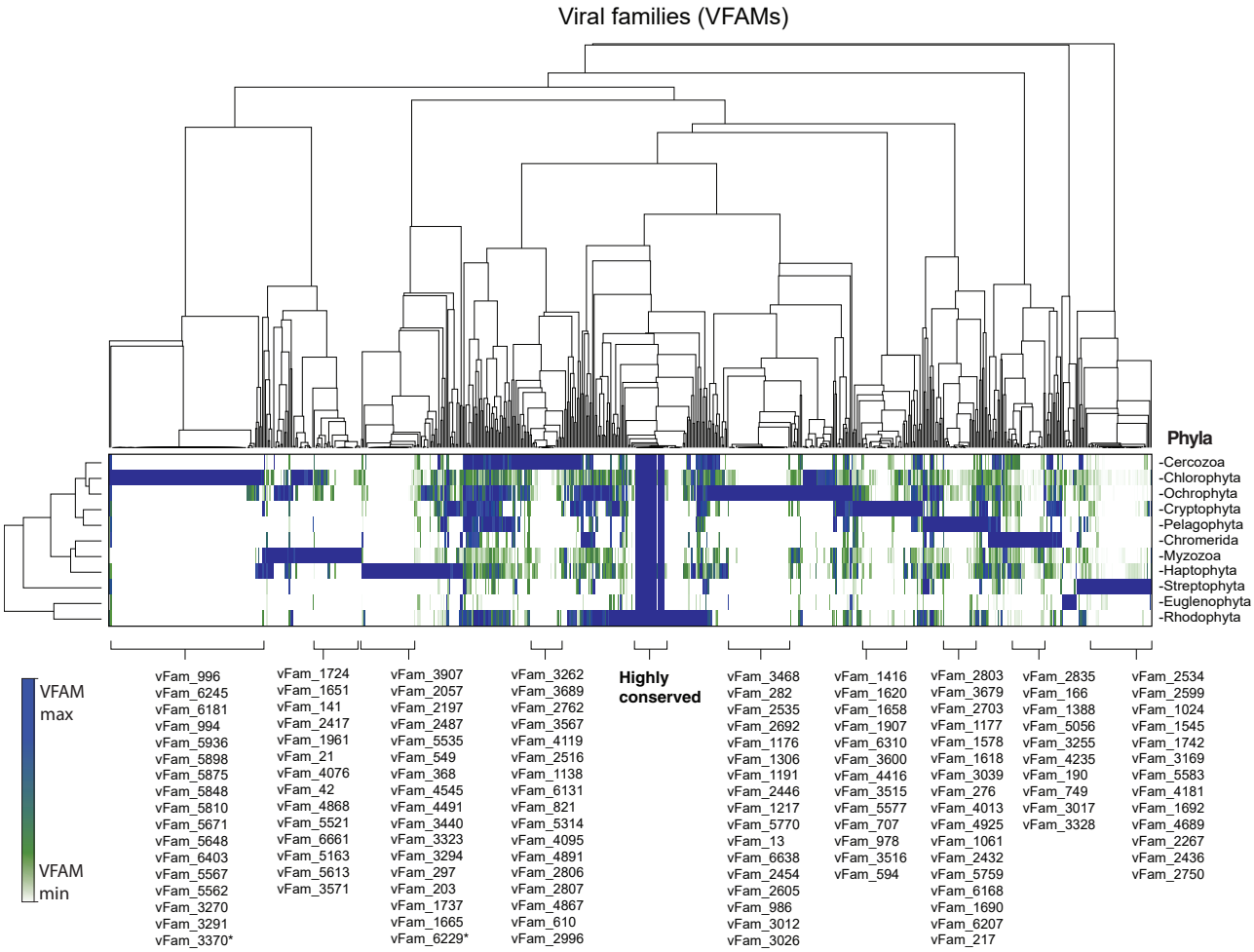


Figure 3. Phylogeny of VFAMs in algal phyla

Each microalgal phylum has a unique set of viral-origin sequences. Species were bi-clustered by VFAM domain counts with Pearson's correlation scores. Domain counts represented by the heatmap are averages from all individuals in the phylum with genome sequence available (AAC and PRE). Phyla-specific VFAMs formed distinct clusters, and many individual VFAMs could be identified from viruses with known virus/host interactions in species from each phylum. For example, Coccolithophores had VFAMs from Coccolithoviruses, chlorophytes had VFAMs from Chloroviruses and Mimiviruses, etc. Chlorophytes clustered most closely with Ochrophytes. Myzozoa most closely with Chromerids. Remarkably, the Rhodophyta VFAM cluster centers on VFAMs conserved in every other microalgal lineage. These results indicate that genetic donations from viruses shaped early- and later-branching photosynthetic microbial eukaryote lineages (i.e., microalgae). Asterisks signify abbreviated lists.

in extant multicellular organisms, especially in plants, are thought to have viral origins (Galindo-González et al., 2017; Gao et al., 2018; Krupovic and Koonin, 2015; Ochoa Cruz et al., 2016; Ustyantsev et al., 2017; Woodrow et al., 2012). Phylogenies created from these proteins and high-confidence homologs in NCBI (NR and Viral databases) revealed genetic exchange among species from unrelated clades. For example, a *Chlorella autotrophica* protein (scaffold1110_cov163-*Chlorella_autotrophica*.AAC.7) coding for a putative multifunctional polymerase/reverse transcriptase aligned with a homolog from *Chloroflexi* (a green non-sulfur bacterium prone to horizontal gene transfers [HGT]). A homolog of this gene was also found in the genome of *Chlamydomonas nivalis* (scaffold12649_cov40-Chlamydomonas_nivalis.AAC.1), an alpine snow alga, among other green algae. Both of these predicted proteins had VFAMs 3,011, 37, and 3,987, indicating that this putative multifunctional protein is conserved in unrelated viral lineages.

Viral family domain-containing genes are expressed and likely functional

We used transcriptomic data from the marine microbial eukaryote transcriptome sequencing project (MMETSP [Keeling et al., 2014]) to validate the presence of VFAMs in expressed CDSs (Data S7). The analysis of the expression data from the MMETSP (Keeling et al., 2014) supported the hypothesis that an estimated majority of VFAMs in CDSs are expressed in natural conditions. Genes that code for nutrient uptake and other metabolic functions had VFAMs. Coding sequences containing VFAMs (VFAM-CDSs) had high-confidence HMM matches to known functional domains (i.e., EVOPs). Freshwater EVOPs

Please cite this article in press as: Nelson et al., Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host & Microbe (2020), https://doi.org/10.1016/j.chom.2020.12.005

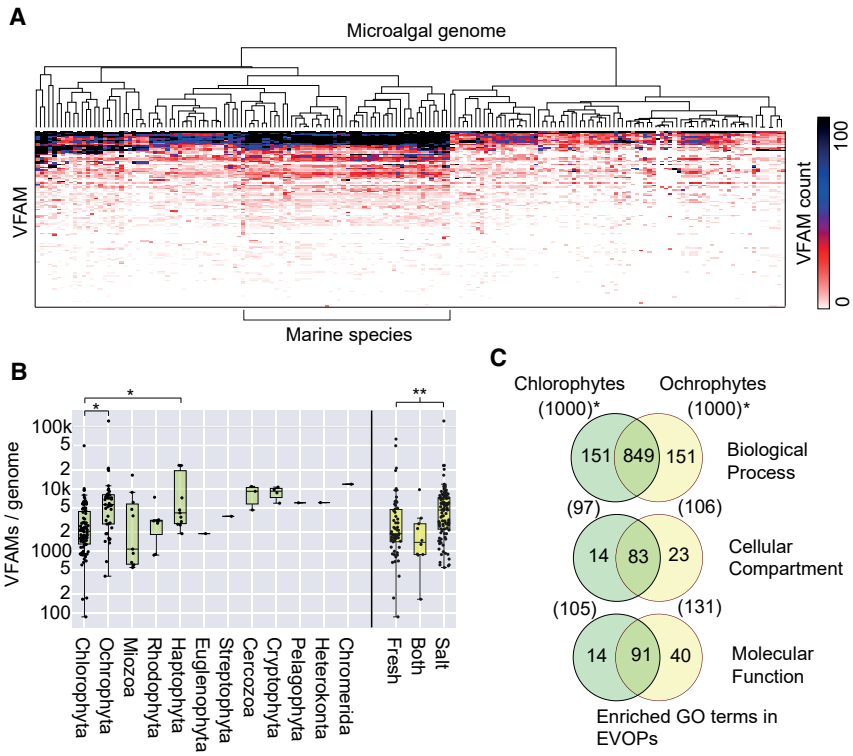


Figure 4. VFAMs are differentially distributed across microalgal clades and environments and encode for a broad range of functions

(A) Hierarchical bi-clustering, based on Euclidean distance of VFAM counts from microalgal genomes, including those we sequenced (n = 107) and those previously sequenced (n = 67). Species of Cercozoa, Cryptophyta, Haptophyta, and Ochrophyta had significantly higher numbers of genomically integrated viral sequences than species from the Chlorophyta and Myzozoa phyla. Larger-genome microalgae, including *Amphidinium*, *Symbiodinium*, *Klebsormidium*, and *Heterococcus*, had VFAMs from crop viruses, including *Hordeivirus* and *Tombusvirus* (Fujisaki et al., 2018; Gunawardene et al., 2017; Jackson et al., 2009; Yergaliyev et al., 2016). The presence of VFAMs in the microalgal genomes was validated in high-coverage, long-read assemblies (Figure 5; Data S6).

(B) Count of VFAMs significantly differed between microalgae phyla; mean values compared using Student's t tests, p values for significant differences indicated by brackets (*$p < 0.01$, **$p < 0.001$). Chlorophytes have fewer VFAMs than Ochrophytes and haptophytes (C, $p = 0.00005$ and $p = 0.06082$, respectively), and freshwater species had significantly fewer VFAMs than their marine counterparts ($p = 0.0008$).

(C) Enrichment for GO terms EVOPs from chlorophytes (n = 36 genomes), compared with Ochrophytes (n = 30 genomes, asterisk, top 1,000 GO terms used; Table S4).

had functions associated with sugar metabolism (galactose oxidase, phosphofructokinase, and ATPases) and amino acid metabolism (asparagine synthase, a serine-threonine kinase, and amidophosphoribosyltransferase) (Data S8, also see Table S3). In saltwater species, VFAMs in CDSs for transporter and other membrane proteins were uniquely present at higher copy numbers or with strong selection signatures, compared with their freshwater counterparts (see Data S9 for gene ontology [GO] lists and Figure S4 for examples of potential EVOP acquisition). Pressures associated with survival in the open ocean may encourage selection for novel viral genes (especially transporters) that can provide access to new resources, including carbon, nitrogen, and growth-promoting cofactors (Rosenwasser et al., 2016).

The comparison of microalgae grouped by the environment (e.g., freshwater versus saltwater) or clade (e.g., Chlorophyta versus Ochrophyta) revealed niche- and clade-specific PFAM distributions and GO term enrichment. Although saltwater species had a higher total VFAM dCN (Figure 4), the CDSs containing these VFAMs had less unique PFAM domains than the freshwater group; there were 331 shared EVOPs, saltwater species had 63 unique EVOPs, and freshwater species had 348 EVOPs. Our data clearly show a drastic difference in the redundancy of EVOP functions for the two environmental groups. Freshwater microalgae had EVOPs uniquely enriched for sugar- and metal-assimilation functions. Marine species were enriched for ion transporters and photosynthetic machinery, nuclear and organellar membrane "cellular compartment" GO terms, including included "intracellular organelle part," "intracellular organelle lumen," "endomembrane system," and "membrane" (FDR-corrected p values < 0.001; 40 GO terms enriched, all p values < 0.001).

Some EVOP functions were tightly conserved and widely shared among diverse lineages (e.g., protein phosphatases and ankyrins). Others, including sugar metabolism genes, were distributed in specific families. For example, the Trebouxiophyceae had various, unique sets of carbohydrate-acting, viral-associated genes (Tables S4 and S5; Data S6). In Chlorophytes and Ochrophytes, >1,500 GO terms were enriched in EVOPs ($p < 0.001$); 1,219 of these were shared between the two phyla (Data S5). Chlorophytes (308 GO terms) had more uniquely enriched GO terms than Ochrophytes (279 GO terms), but Ochrophytes had more VFAMs per genome ($p = 0.008$ for the difference between the two groups, Figure 4B; Table S2; Data S6). Uniquely enriched GO terms for Ochrophytes (biological processes, BP) included "sulfur compound transport," "organophosphate biosynthetic process," "ion transmembrane transport," and "L-amino acid transport." Uniquely enriched GO terms for chlorophytes (BP) included "sporulation," "regulation of Wnt signaling pathway," "regulation of GTPase activity," and "regulation of mitotic sister chromatid segregation."

Core viral genes, including polymerases, ribonucleotide reductases, helicases, topoisomerases, and nucleases, were broadly distributed among microalgal lineages (Data S3–S5). For example, protein phosphatases, which are essential for cell-cycle regulation (Nilsson, 2019), were found in the VFAM-tagged gene sets of >90% of the analyzed algal species. Donated genes also included arrays of adaptive and essential metabolic, nucleic-acid-acting, amino-acid-forming, structural, signaling, and stress-resistance genes (Figure 5; Table S3; Data S5). In Chlorophytes, EVOPs

Please cite this article in press as: Nelson et al., Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host & Microbe (2020), https://doi.org/10.1016/j.chom.2020.12.005

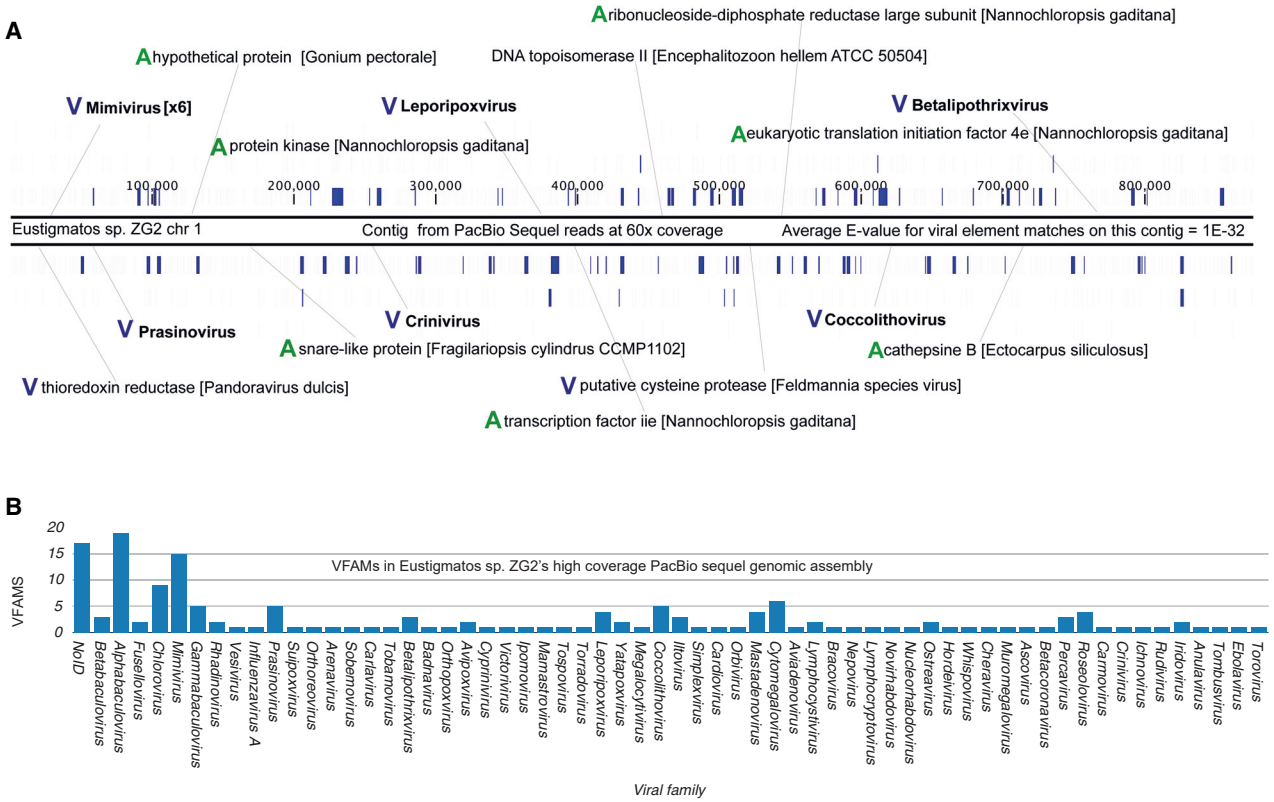


Figure 5. VFAM integration in *Eustigmatos*

(A) Contig from *Eustigmatos sp. ZG2* sequenced with PacBio Sequel long reads at 60× coverage. This sequence shows (A) algal genes interspersed with viral (V) elements. The cluster of Mimivirus-related sequences near the contig end shows "gene-gang" synteny characteristic of Chloroviruses (Seitzer et al., 2018). This intact synteny suggests a single insertion of a continuous track of DNA in the transfer even and later, sectional, replacement of genomic regions by host sequences.

(B) Viral families (VFAMs) detected in the isogenic PacBio *Eustigmatos sp. ZG2* assembly. Other downloaded long-read assemblies, such as *the Chlamydomonas reinhardtii* (Merchant et al., 2007) and *Chromochloris zofingiensis* (Roth et al., 2017) genomes, also showed VFAM integration in contigs generated from long, continuous reads.

were uniquely enriched for the GO terms "negative regulation of cation transmembrane transport," "flavonoid metabolic process," "osmosensory signaling pathway," "acylglycerol metabolic process," "neutral lipid metabolism," and "glutathione metabolism" (Data S10), (FDR-corrected p values < 0.001). The GO terms that were uniquely enriched in Ochrophytes included "cellular metal ion homeostasis," "positive regulation of lyase activity," "organelle disassembly," "cellular response to reactive oxygen species," "lignin metabolic process," "sorbitol catabolic process," "anion transmembrane transport," "organic cation transport," "cyclic-nucleotide-mediated signaling," "divalent inorganic cation transport," "divalent metal ion transport," "response to salt," and "sulfur compound transport" (FDR-corrected p values < 0.001).

Because Trebouxiophytes, like *Chloroidium*, can use a wide variety of carbon sources for heterotrophic growth (Nelson et al., 2017), and viruses have been known to transfer carbohydrate-acting genes to their hosts (Van Etten et al., 2017), we searched for viral-associated carbohydrate-acting genes in the *Chloroidium sp*. UTEX3077 genome (Nelson et al., 2019). Chloroidium species showed an exception to a VFAM-salt correlation. They are salt-tolerant (up to 60 g/L) but had less genomic VFAMs when compared to other marine Chlorophytes and Ochrophytes. The *Chloroidium sp*. *CF* genome was assembled with Illumina long-range mate-pair libraries, and generated contigs also show VFAMs interspersed with conserved microalgal genes. For example, a chitinase was found with three strong VFAM matches. This gene has homologs in distantly related organisms, including insects, fungi, and viruses. A MUSCLE alignment of these homologs and maximum likelihood phylogeny reconstruction revealed that this gene was of viral origin and prone to HGT in organisms from separate kingdoms (see Figure S4).

Viral family domains in *Chlamydomonas*

A variety of transcription factors, helicases, and other chromatin-acting enzymes in *Chlamydomonas reinhardtii* had strong evidence (E value < 1e-9 from serial HMM matches, see also Figure 6) of viral origin (Data S6 and S8). Other proteins with VFAM domains found in *C. reinhardtii* included membrane occupation and recognition nexus (MORN) repeats from Pandoraviruses (VFAM 86), a histone H2A from Lausannevirus (VFAM 6324), a histone H2b from Pandoravirus, an adenosohomocysteinase

Please cite this article in press as: Nelson et al., Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host & Microbe (2020), https://doi.org/10.1016/j.chom.2020.12.005

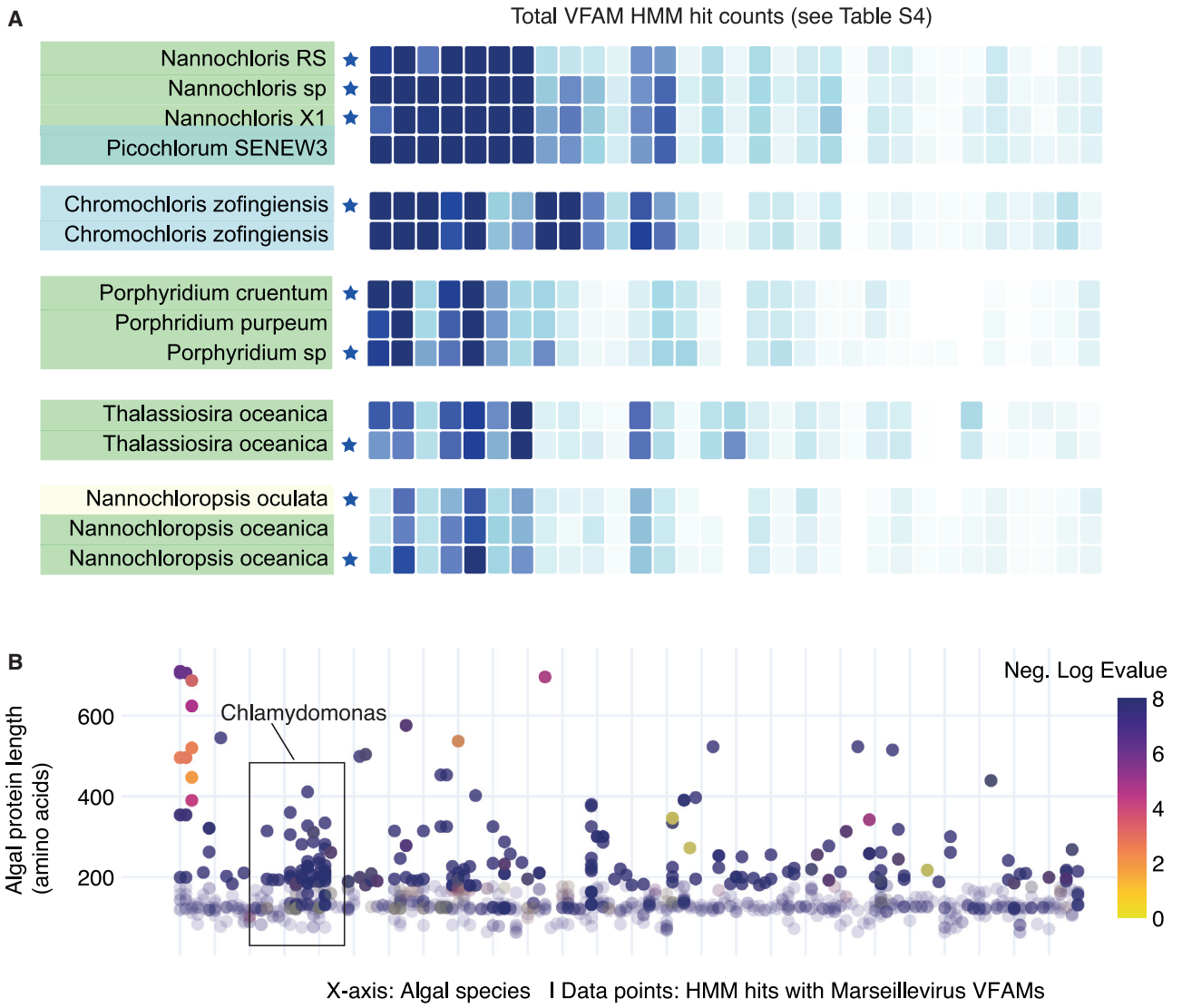


Figure 6. Viral family types, including Chlorovirus, Mimivirus, Hordeovirus, distinguish microalgal lineages

Viral-origin sequences are species- and lineage-specific;

(A) VFAM patterns were reproducible across independent sequencing and genome assemblies for the same or closely related species. These viral fingerprints are core, distinguishing features of the microalgal clades we surveyed and were reproducible in genomic assemblies from separate cultures, sequencing technologies, and computational workflows. Species shading: blue, saltwater; green, freshwater; yellow, euryhaline. A blue star indicates the sequencing was done by our group; other genomes were downloads.

(B) HMM matches of microalgal proteins to Marseilleviruses. The y axis indicates the length of the microalgal protein. *Chlamydomonas* species had the greatest range of VFAM dCNs, and *Chlamydomonas asymmetrica* had the most matches overall.

from Pandoravirus (VFAM 4000), a cathepsin from a *Trichoplusia nucleopolyhedrovirus* (VFAM 5985), as well as functional domains for heat shock proteins, ankyrin repeats, chitinases, proteases, helicases, topoisomerases, ATPases, reverse transcriptases, and protein kinases from a wide variety of viruses.

C. reinhardtii was recently found to be able to assimilate more nitrogen sources than previously believed (Calatrava et al., 2019; Chaiboonchoe et al., 2014); therefore, we investigated if any nitrogen-related metabolic reactions could be traced to viral-origin genes. We saw VFAM-containing genes with asparagine synthase (PF00733.18, E value = 7.1e-53) and aspartyl protease (PF13650.3, E value = 7.9e-13) PFAM domains, indicating that EVOPs might be involved in *C. reinhardtii* nitrogen metabolism. Our evidence shows that VFAM 1034 (also seen in *Phaeocystis globosa* virus Santini et al., 2013, *Ostreococcus tauri* viruses Derelle et al., 2008; Steele et al., 2018; Weynberg et al., 2009, 2011, *Acanthamoeba polyphaga* moumouvirus, *Ostreococcus lucimarinus* virus OlV1, *Micromonas sp*. RCC1109 virus MpV1, *Cafeteria roenbergensis* virus BV-PW1, and the *Megavirus* Iba virus Wagstaff et al., 2017) has integrated into *C. reinhardtii*'s genome (XP_001700322.1; NCBI, https://www.ncbi.nlm.nih.gov, VFAM HMM alignment score: E value = 1.5e-57). This

Please cite this article in press as: Nelson et al., Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host & Microbe (2020), https://doi.org/10.1016/j.chom.2020.12.005

protein is currently annotated as a protein of unknown function; however, the top 20 BLASTP hits from other organisms in NCBI are asparagine synthases. The VFAM contents and multifunctionality, as inferred from its PFAM domains (Asn synthase and two peptidase domains), of this gene indicate a viral origin. The peptidase domains are located in the C terminus of the predicted protein and include S8 (subtilisin and kexin) and S53 (sedolisin) endopeptidases and exopeptidases. Chlamydomonas species inhabit a wide range of terrestrial environments, and the integration of multifunctional genes, such as nitrogen-metabolism genes from viruses, could significantly impact the local fitness of these green microalgae.

Saltwater microalgae show significant increases in genomic viral family content

Compared with freshwater species, saltwater species had significantly higher numbers of genomic VFAMs ($p < 0.008$, Figures 3 and 4; Data S6). This correlation of VFAMs with salinity habitat status had notable exceptions, including *Chromochloris zofingiensis* (Figure 4). *C. zofingiensis* had many VFAMs (2,286 VFAMs in the assembly from Roth et al., 2017, 2,365 VFAMs in our assembly [E value < 1e-9]) but is a freshwater alga. Saltwater high-VFAM species, including Haptophytes like *Phaeocystis Antarctica* (AAC, 1.8% genomic repeats, ~49,000 VFAMs) and *Emiliana huxleyi* (PRE, ~24,000 VFAMs), had low levels of genomic repeats (1.8% and 6.6%, respectively). These observations led us to hypothesize that viral content might inversely correlate with repeat content and microalgae. We tested this hypothesis in Chlorophytes ($n = 75$). A ranked Spearman's test of correlation gave a Rho (rs) of −0.207 and a combined covariance of −98.17 (=0.0752). Repeat content in Ochrophytes ($n = 46$) were also negatively correlated with VFAMs with a Rho value of −0.264 ($p = 0.099$).

Exceptions included *Chromochloris zoofingensis*, a freshwater microalga with high levels of VFAMs in both the reference assembly ((Roth et al., 2017); 2,286 VFAMs in 33,513 exons) and our assembly (2,365 VFAMs in 33,910 exons). Species of Cercozoa, Cryptophyta, Haptophyta, and Ochrophyta had significantly higher viral contents than species from Chlorophyta and Myzozoa phyla. In contrast, Chlorophytes and Myzozoa had, on average, 3.6% more repetitive, non-coding elements in their genomes ($p < 0.0001$). Diatoms, members of the Ochrophyte phylum, had the most VFAMs, while all *Chlamydomonas* species surveyed ($n = 8$) had low VFAM content (Data S6). Algae with high repeat content, including *Amorphochlora amoeboformis* (AAC, 43.6%, 65 Mbps), *Lingulodinium polyedra* (AAC, 17.81%, 31 Mbps), and *Heterocapsa arctica* (AAC, 15.32%, 33 Mbps), had few VFAMs. Conversely, species with low repeat content had many VFAMs.

Saltwater versus freshwater genomes

Freshwater species were compared with saltwater species to find differences in amino acid coding and functional potential (Figure 7). Threonine (T), arginine (R), valine (V), methionine (M), etc., showed no significant differences between the two groups. Each other residue was significantly different in the two groups (e.g., alanine (A): fresh = 12.17%, salt = 9.35%, $p < 0.0001$; aspartic acid (D): fresh = 3.47%, salt = 4.87%; $p < 0.0001$; lysine (K): fresh = 2.86%, salt = 4.42%, $p < 0.0001$; glutamic acid (E), fresh = 3.83%, salt = 5.72%, $p < 0.001$; tyrosine (Y): fresh = 1.61%, salt = 1.99%, $p = 0.0011$. Lysine in saltwater species proteins was, on average, nearly double that in freshwater species ($p < 0.0001$). We also found that predicted enzymes in lysine biosynthesis pathways from the Kyoto encyclopedia of genes and genomes (KEGG) had higher counts, on average, in most saltwater species (Table S4; Data S10). Lysine was recently found to be important for oxidative stress resistance (Chang et al., 2020), and higher abundances of lysine in saltwater microalgal proteins might be necessary for their survival in the face of salt-induced oxidative stress.

We applied principal component analysis (PCA) to determine if the variance in PFAM domains (Figure 7; Table S4) generated biologically or ecologically relevant clustering (see Table S3 for PCA scores and variance). We found that 20% of the variance was attributed to PC1; this distribution correlated with saline habitat status. Thus, PFAMs contributing to loading on PC1 were implicated in salt tolerance. The species with negative scores for PC1 were mostly marine. A PC1vPC2 graph revealed that oceanic Chlorophytes clustered close to Ochrophytes. The results for separation across PC2 were not informative. Separation across PC3 yielded distinct clusters for marine Chlorophytes and marine Diatoms (Ochrophytes). Freshwater species formed a tight cluster away from the other, more dispersed, saltwater species. However, there were exceptions to an exclusive salt-freshwater species clustering for PC1vPC3, including *Axilococcus clingmanii* and *Chlorella desiccata*, in the seven-member picoeukaryote-majority cluster (Figure 7A). The clustering of marine Chlorophytes and Ochrophytes separate from most freshwater species in a directional vector was consistent with the environment (salt)-dependent PFAM loading variance results.

The top PFAM scoring toward PC1 for saltwater species was PF00004.26, a nucleotide-hydrolyzing AAA ATPase domain. This PFAM was also prominently enriched, or duplicated, in our other analyses comparing salt- and freshwater species and positive selection from dN/dS ratios within CDSs (see Data S13). The second and third top-negative scoring PFAMs in PC1 were DEAH/DEAD box helicases (PF00270.26 and PF00271.28, respectively); CDSs with DEAD box helicase domains had homology to a variety of viral proteins (see Data S11 for BLAST results). The fourth, PF00005.24, is an ATP-binding cassette transporter domain. The top negatively scoring PFAMs along PC1 or those correlating with the direction of the cluster of saltwater species should contribute heavily to active transport during salt stress. A cluster of marine picoeukaryotic green microalgae was separated from diatoms and other, freshwater, Chlorophytes. Other PFAMs scoring toward PC1 were also identified as duplicated or under positive selection in saltwater species and correlated with enriched domain-centric ontology (Fang and Gough, 2013a), including small molecule binding (dCN increase = 10.33×), nucleoside phosphate-binding (dCN increase = 10.79×), and ion binding (dCN increase = 8.37×).

In order to determine PFAMs that were unique to saltwater species but shared across multiple clades, we isolated four groups of nine genomes each that were freshwater or saltwater and Chlorophyta or Ochrophyta. In this analysis, PFAM count arrays were graphed in clade- or environment- two-thirds distributions (Figure S6). Saltwater Chlorophytes (SC) and saltwater

Please cite this article in press as: Nelson et al., Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host & Microbe (2020), https://doi.org/10.1016/j.chom.2020.12.005

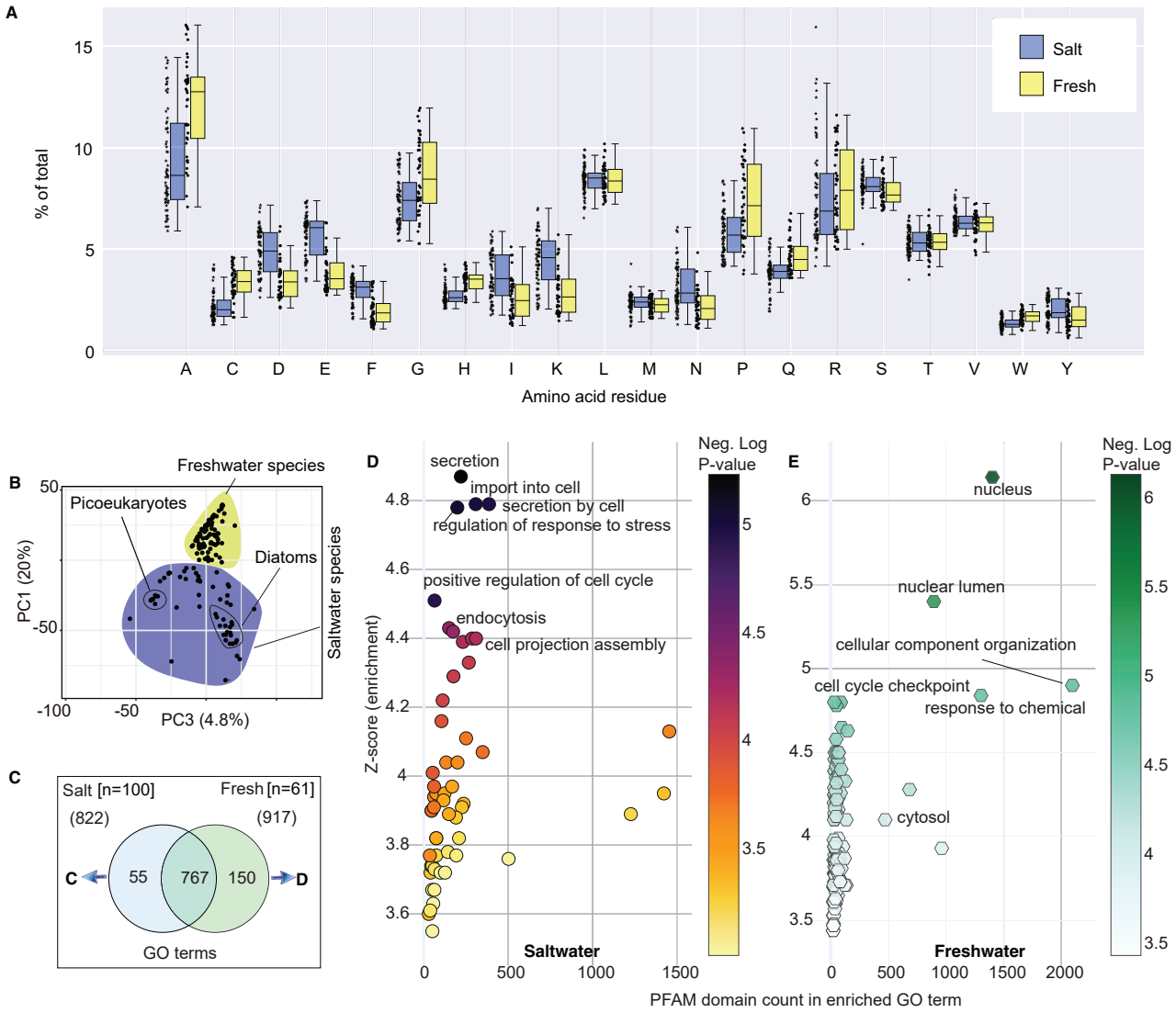


Figure 7. Principal component analysis (PCA) and uniquely enriched GO terms reveal the main differences between saltwater and freshwater species

(A) Amino acid counts per species. Box plots show amino acid residues per species (data points), lower and upper quartiles for each group (box ends), median (line inside box), and range (whiskers). Roughly one-third of the 20 residues were not significantly different between the two groups, one-third had significantly higher counts in saltwater species ($p < 0.05$), and one-third had significantly higher counts in saltwater species.

(B) From PFAM dCNs, we used non-property linked PCA; the data variances determined the principal components (PCs). When PCA was performed, there were no assumptions about the data nor were PCs chosen in advance. PC1 accounted for 20% of the variance, and species could mostly be grouped along PC1 in a gradient of salinity tolerance (loadings and scores in Table S3). PC2 accounted for 5.6% of the variance, but the variance was not informative because only one species accounted for the variance. The distance along PC3, which accounted for 4.8% of variance, separated diatoms and halotolerant, picoeukaryotic chlorophytes.

(C) Domain-centric GO term enrichment (Fang and Gough, 2013a) from all saltwater species compared with freshwater species (enrichment at a false discovery rate [FDR]-corrected $p < 0.001$ for all GO terms; Table S4).

(D) Uniquely enriched GO terms in saltwater species. "Secretion," "import into cell," and "regulation of response to stress" had high enrichment scores (*Z* scores) and low FDR-corrected p values. The unique inclusion of these GO terms suggests that the intracellular trafficking of molecules is essential for microalgal survival in saline environments.

(E) The GO terms "nucleus" and "nuclear lumen" had high enrichment scores (*Z* scores) and low FDR-corrected p values in freshwater species (Table S4). These unique enrichment results suggest these functions are essential for the habitation of terrestrial environments.

Ochrophytes (SO) shared more PFAMs when compared to freshwater Ochrophytes (FO) or FC as seen by the PFAM distributions shifted toward the SC/SO axis. We found 84 PFAMs shared across saltwater-dwelling species of Chlorophyta and Ochrophyta that were absent in freshwater species from either phylum (Data S12). Included in this salt-specific gene subset

Please cite this article in press as: Nelson et al., Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host & Microbe (2020), https://doi.org/10.1016/j.chom.2020.12.005

was a salt-dependent phosphate transporter (PF02690.12), an oxoglutarate-Fe dioxygenase (PF10014.6), an N-acetylmuramoyl-L-alanine amidase (PF01510.22), a vacuole effluxer (PF11700.5), a capsular polysaccharide synthesis protein (PF05704.9), a galactose-3-O-sulfotransferase (PF06990.8), a hepatitis A-like viral protein (PF06777.8), a lipid-binding putative hydrolase (PF08557.7), and the transmembrane protein domains PF01823.16 and PF15383.3.

Freshwater microalgae are enriched in nuclear processes and carbohydrate metabolism

The freshwater species sequenced for this project included Chlorophytes (n = 22), Euglenophytes (n = 2), Streptophyta (n = 1), Haptophyta (n = 7), and Rhodophyta (n = 9) (Table S1). Freshwater species had lower numbers of PFAMs per genome. We found that 1,324 and 778 PFAMS were unique to salt- (n = 31) and freshwater (n = 57) species, respectively (Figure 7B). Freshwater species were enriched for the nucleus, and nuclear membrane GO terms (Figures 7D and S5; see Data S9 for GO terms and S10 for enzyme commission [EC] terms). For example, "small-RNA loading onto a RISC complex" (GO: 0070922, FDR-corrected p = 8.9e-6), "regulation of nuclear-transcribed mRNA poly(A) tail shortening" (GO: 0060211 and GO:0060213, FDR-corrected p = 8e–7) were uniquely enriched in freshwater species. The increase in PFAM domains involved in nuclear processes in freshwater species correlated with the differences in the ionic environment. Specifically, RNA-binding protein activity is modulated by intracellular ionic conditions (Chong et al., 2018; Shorter, 2019; Yan et al., 2018).

The sample size for freshwater species was smaller, but they had more enriched GO terms; this difference indicated the presence of functional redundancies in saltwater-unique GO terms. While saltwater species were enriched for the Biological Process (BP) GO term "positive regulation of cell cycle," freshwater species were enriched in the BP "cell-cycle checkpoint." This result indicated that gene families regulating cell-cycle progression are expanding and contracting in salt- and freshwater species, respectively. For example, DNA Damage Checkpoint (GO000077, dCN increase = 4.49×, FDR-corrected p = 0.0000019) and non-coding mediated negative regulation of translation (GO:0040033, dCN increase = 4.25×, FDR-corrected p = 3.10E-04) were enriched in freshwater species. The GO term "carbohydrate derivative binding" was enriched in freshwater species (GO:0097367, FDR-corrected P=5.1e-10), which likely represents the abundance of heterotrophism-supporting genes in inland species, including Trebouxiophytes. The recent expansion of many saltwater lineages into diverse, freshwater terrestrial environments (Alverson et al., 2011; Alverson and Theriot, 2005; Onyshchenko et al., 2019) is likely a major driver for the functional divergence seen in their genomes (Figure 1).

Unique enrichment of membrane functions in saltwater microalgae

Sequences coding for predicted membrane functions (e.g., topology-determination, transport, signaling, and adhesion) were highly diversified in saltwater species compared with freshwater species. Saltwater species showed the unique enrichment of GO terms in a domain-centric gene ontology analysis. We found PFAM domains with membrane functions, including transporters with predicted affinities for a wide range of solutes, such as sodium, potassium, peptide, sugar, nitrogen, metal, and phosphate (Data S5). Protein family duplication suggests widespread gene duplication and diversification in marine species; this hypothesis is supported by previous study findings (Nakov et al., 2018a; Parks et al., 2018). Overrepresentation of membrane proteins in the GO enrichment analysis for saltwater species (Table S4) suggests that maintaining membrane integrity and intracellular ion homeostasis in the face of increased extracellular salts is required. The enrichment of membrane proteins likely contributes to the fitness of single-celled microalgae in dynamic ocean environments. The fine-tuning of membrane topology, in part via the transporter proteins and their adaptor ankyrins, appears to be an underlying mechanism responsible for this adaptive capacity.

Saltwater species had 99 domains with an average copy number more than 4-fold higher than their freshwater counterparts (Table S3; Data S5). Saltwater species had, on average, more copies of phospholipase D. Osmotic stress, cold, and ABA activate different phospholipases to generate lipid messengers (e.g., Phosphatidic acid (PA), DAG, and IP3) that have roles in regulating stress tolerance (Zhu, 2002). Recent studies indicate that PLD and PA help plants cope with drought and salt stress (Bargmann et al., 2009; Wang, 2005). PA, which was suggested to be generated via activation of a phospholipase D (PLD) pathway and the combined action of a phospholipase C/diacylglycerol kinase (PLC/DGK) pathway, has also been shown to increase as a response to hypersalinity (Arisz and Munnik, 2013) Other mechanisms related to signaling that were expanded in saltwater species include hydroxysteroid metabolism. Loss of dihydroxysteroid dehydrogenase is known to cause salt sensitivity in humans, and its higher prevalence in the genomes of saltwater species could aid survival in marine environments (Kerstens et al., 2004; Lovati et al., 1999; Luft, 2016).

A more resilient cell membrane would be necessary when microalgal outer barriers disintegrate after exposure to conditions, such as increased atmospheric CO_2-associated ocean acidification (Dell'Acqua et al., 2019). Acidic conditions degrade carbon-based external cellular structures (e.g., shells) (Dell'Acqua et al., 2019). Thus, a robust defense against detrimental chemical forces is required by marine species, and this can be seen in the unique collection of membrane genes found in saltwater species (Figure 7). For example, ankyrins had markedly increased dCN in the saltwater group. Ankyrins control membrane topology by acting as pins between 7–10 nm rafts of phospholipids and other proteins (Bennett, 1992) and thus, have a major role in specialized membrane organization and viral replication (Bennett, 1992; Wang et al., 2014).

The membrane architecture of marine microalgae likely protects cells from the salt stress and mechanical stress resulting from the constant limnological mixing that occurs in the upper photic zones they inhabit. Cadherin copy numbers were also increased in marine species (Data S5), which indicates the presence of complex systems for mechanical stress response. We noted an increase in lysine in saltwater species proteins. Lysine plays a crucial role in crosslinking collagen molecules, and collagen plays a key role in building an extracellular matrix (ECM) that promotes resistance to mechanical stress (Jansen et al., 2018). Overall, our results indicated the presence of a

Please cite this article in press as: Nelson et al., Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host & Microbe (2020), https://doi.org/10.1016/j.chom.2020.12.005

much more comprehensive mechanochemical response system in marine species compared with their freshwater counterparts.

Freshwater species lacked many of the membrane structural (ankyrins) and functional (transporters) components seen in saltwater species. This result suggests they are not necessary for survival in non-marine conditions. Freshwater-obligates, or non-euryhaline, species have rigid cell walls (Cronmiller et al., 2019). Thus, the membrane protein-dense genetic repertoire of marine microalgae forms the foundation of this flexible, ion-resistant phenotype. Trebouxiophytes, or mainly aeroterrestrial green microalgae, which are tolerant of desiccation and salt (Nelson et al., 2017), have also been suggested to have flexible, ion-tolerant cellular membranes (Holzinger and Karsten, 2013).

Picophytes have increased relative and absolute levels of genes under positive selection

We searched for genes under positive selection in the salt- versus freshwater species, specifically within the Chlorophyceae (Data S13). Sixty genomes from small microalgae, 30 freshwater, and 30 saltwater species, and their CDSs and amino acid sequences were used in the FUSTr workflow (Cole and Brewer, 2018) to calculate dS/dN ratios for genes and their phylogenies (see Data S13). A total of 4,403,010 CDSs were used as input; CDSs and predicted proteins from these genomes were organized into 2,991,159 families; 792 had >15 sequences per family. We detected 335 families under strong positive selection; 401,500 PFAMs (HMMsearch, E value < 1e-9, from 60 species) were identified from genes in these families as being under strong positive selection. Picoeukaryotes, including *Micromonas* and *Nannochloris* species, had the highest overall numbers of genes under strong positive selection (Data S13).

Ankyrin domains were the most numerous in the subsets under positive selection; Ank_4 had 10,131 copies in the 60-species subset (Ank_5[dCN] = 10,100, Ank_3[dCN] = 8,474, Ank_2 [dCN] = 8,323, Ank[dCN] = 8,088). Because membrane integrity is sensitive to temperature shifts, evolving ankyrin domains could increase host survival probabilities during global warming conditions. The coalescence of ankyrin domain abundance in saltwater species in the enrichment and PCA analyses further supports the hypothesis that membrane topology alteration is the primary adaptation to dynamic marine environments in microalgae. For example, because *Chlorovirus* ankyrins control host ubiquitination and cell-cycle progression machinery (Noel et al., 2014), they are implicated in massive algal blooms and the resultant biogeochemical shifts (Rosenwasser et al., 2016).

A cluster of mostly viral-associated and transporter PFAMs had a higher dCN in the subset under positive selection in SC (n = 30). PCA of the subset of genes under positive selection revealed viral-type PFAMs that scored highly toward PCs 1–3 (e.g., reverse transcriptases, proteases, topoisomerases, and resolvases of viral origin). Compared with freshwater species, marine microalgal genomes had more dCNs of membrane-related functions that were under strong positive selection (Data S13). These membrane functions ranged from active and passive ion, or solute, transport through the cell membrane to secretion domains. Our results suggested that virus-origin genes that mostly encode for membrane functions are under strong positive selection and likely have major roles in the continued survival of marine microalgae populations.

Conclusions

A subset of genes containing mainly membrane and viral proteins was shared among marine microalgae from different lineages, indicating their importance for the maintenance of membrane integrity in a saline environment. The identification of these genes predicted to confer halotolerance opens a path for their use in the development of agriculture in marginalized, saline, regions, as well as for the development of new microalgal strains for biotechnological purposes. Furthermore, this dataset helps to clarify fundamental distinctions in microalgal genomics and will inform future ecogenomics studies in addition to serving as a database for genetic engineering projects.

New genetic resources are essential for progress in biotechnology development. Agronomical engineering depends on the use of foreign genes, such as the pest resistance cassettes used in a variety of productive GMO crops. Transgenic approaches using plant and non-plant photosynthetic organisms have proved to be useful in conferring tolerance to drought and high soil salinity (Cabello et al., 2017; Dabi et al., 2019; Fan et al., 2019; He et al., 2019; Huang et al., 2017; Li et al., 2017, 2018; Mushke et al., 2019; Passricha et al., 2019; Sun et al., 2017; Tang et al., 2019; Udawat et al., 2017; Wu et al., 2017; Xu et al., 2018; Yao et al., 2016). Generally, these studies have focused on overexpression or ablation of plant genes to augment halotolerance (Mushke et al., 2019; Passricha et al., 2019; Sun et al., 2017; Tang et al., 2019; Udawat et al., 2017). Still, the use of genes from more distantly related organisms, such as *Physcomitrella* (Li et al., 2017, 2018), has been remarkably effective in increasing plant halotolerance. Genes from saltwater photosynthetic microbes, especially microalgae, are excellent candidates for agronomical biotechnology applications. Thus, more genetic resources are needed for the development of halotolerant plants for use as food crops, ornamental cultivars, and agents of bioremediation and ecological restoration.

Recent studies have shown that giant viruses frequently acquire host genes (Moniruzzaman et al., 2020; Schulz et al., 2020); here, we show that the reverse has occurred repeatedly throughout algal evolution. Viral sequence accumulation was drastically different between different microalgal lineages with regard to frequency and function. Functional EVOPs likely confer some niche-specific fitness benefit to the host. One example supported by this work is the acquisition of membrane proteins, including transporters and topology-altering proteins, from viruses in marine species. High levels of VFAMs and membrane proteins are hallmarks that distinguish saltwater from freshwater microalgal genomes. Our results are not indicative of either purifying selection in freshwater or horizontal transfer of virus genes to saltwater species; instead, both processes likely contributed to the heterogeneity of VFAM content from microalgae in various environments.

In this report, we provided evidence for unexpected host ranges of several microalgal virus families. Genomic integrations of sequences from viruses from unrelated clades highlight possible horizontal gene transfer events, with broad-host range viruses acting as shuttle vectors. Our data are consistent with microbial eukaryotes acting as hosts for a wider variety of viruses than was previously appreciated. Environmental sources of viruses, especially microbial eukaryotes, should be more seriously considered and evaluated to anticipate future potential public health crises.

Availability of supporting data and materials

This resource is made available with the aim of unrestricted accessibility, and the datasets are hosted by multiple data repositories, including Dryad:https://doi.org/10.5061/dryad.7wm37pvnv and NCBI,Bioproject:PRJNA517804). All microalgae strains used in this study are available to the general public through their respective culture collection centers (see Table S1).

STAR★methods

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHODS DETAILS
  - Microalgal strain selection and cultivation
  - DNA extraction
  - Sequencing
  - *De novo* genome assembly
  - Artificial degradation experiments
  - Contamination screening
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Genome annotation
  - Domain-centric GO enrichment analysis
  - Validation using reference genomes
  - Phylogeny reconstructions
  - Amino acid residue comparisons
  - Dimension reduction analyses
  - Ternary graphs
  - Positive selection analysis
  - Viral family sequence identification

Supplemental Information

Supplemental Information can be found online at https://doi.org/10.1016/j.chom.2020.12.005.

Acknowledgments

We thank the NYUAD High-Performance Computing Center for providing computational resources. This work was supported by Tamkeen under the NYU Abu Dhabi Research Institute grant to the NYUAD Center for Genomics and Systems Biology (73 71210 CGSB9), and by NYUAD Faculty Research Funds (AD060).

Author contributions

K.S.-A. and D.R.N. conceived the study. D.N., M.H., J.S., M.P., J.B., M.L., S.D., and D.R.N. cultured the microalgae. D.R.N. oversaw sequencing and bioinformatic analyses. D.R.N., K.M.H., A.K.J., and A.C. performed the bioinformatic analyses. K.S.-A., D.R.N., K.M.H, K.M.A.A., and K.L. interpreted the results. K.S.-A., D.R.N., K.M.H, and K.L. wrote the manuscript with input from the other co-authors.

Declaration of interests

The authors declare no competing interests.

Received: June 22, 2020
Revised: October 8, 2020
Accepted: November 18, 2020
Published: January 11, 2021

References

Abrahão, J., Silva, L., Silva, L.S., Khalil, J.Y.B., Rodrigues, R., Arantes, T., Assis, F., Boratto, P., Andrade, M., Kroon, E.G., et al. (2018). Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. Nat. Commun. *9*, 749.

Alborzi, S.Z., Devignes, M.D., and Ritchie, D.W. (2017). ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains. BMC Bioinformatics *18*, 107.

Alverson, A.J., Beszteri, B., Julius, M.L., and Theriot, E.C. (2011). The model marine diatom Thalassiosira pseudonana likely descended from a freshwater ancestor in the genus Cyclotella. BMC Evol. Biol. *11*, 125.

Alverson, A.J., and Theriot, E.C. (2005). Comments on recent progress toward reconstructing the diatom phylogeny. J. Nanosci. Nanotechnol. *5*, 57–62.

Arisz, S.A., and Munnik, T. (2013). Distinguishing phosphatidic acid pools from de novo synthesis, PLD, and DGK. Methods Mol Biol. *1009*, 55–62, https://doi.org/10.1007/978-1-62703-401-2_6.

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., et al. (2004). The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science *306*, 79–86.

Bargmann, B.O., Laxalt, A.M., ter Riet, B., van Schooten, B., Merquiol, E., Testerink, C., Haring, M.A., Bartels, D., and Munnik, T. (2009). Multiple PLDs required for high salinity and water deficit tolerance in plants. Plant Cell Physiol. *50*, 78–89.

Bennett, V. (1992). Ankyrins. Adaptors between diverse plasma membrane proteins and the cytoplasm. J. Biol. Chem. *267*, 8703–8706.

Blanc, G., Agarkova, I., Grimwood, J., Kuo, A., Brueggeman, A., Dunigan, D.D., Gurnon, J., Ladunga, I., Lindquist, E., Lucas, S., et al. (2012). The genome of the polar eukaryotic microalga Coccomyxa subellipsoidea reveals traits of cold adaptation. Genome Biol. *13*, R39.

Blanc-Mathieu, R., Krasovec, M., Hebrard, M., Yau, S., Desgranges, E., Martin, J., Schackwitz, W., Kuo, A., Salin, G., Donnadieu, C., et al. (2017). Population genomics of picophytoplankton unveils novel chromosome hypervariability. Sci. Adv. *3*, e1700239.

Bonetti, G., Trevathan-Tackett, S.M., Carnell, P.E., and Macreadie, P.I. (2019). Implication of viral infections for greenhouse gas dynamics in freshwater wetlands: challenges and perspectives. Front. Microbiol. *10*, 1962.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.

Bzhalava, Z., Hultin, E., and Dillner, J. (2018). Extension of the viral ecology in humans using viral profile hidden Markov models. PLoS One *13*, e0190938.

Cabello, J.V., Giacomelli, J.I., Gómez, M.C., and Chan, R.L. (2017). The sunflower transcription factor HaHB11 confers tolerance to water deficit and salinity to transgenic Arabidopsis and alfalfa plants. J. Biotechnol. *257*, 35–46.

Calatrava, V., Hom, F., Llamas, A., Fernandez, E., and Galvan, A. (2019). Nitrogen scavenging from amino acids and peptides in the model alga Chlamydomonas reinhardtii. The role of extracellular L-amino oxidase. Algal Res. *38*, 101395.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.

Chaiboonchoe, A., Dohai, B.S., Cai, H., Nelson, D.R., Jijakli, K., and Salehi-Ashtiani, K. (2014). Microalgal metabolic network model refinement through High-throughput functional metabolic profiling. Front. Bioeng. Biotechnol. *2*, 68.

Chang, R., Stanley, J., Robinson, M., Sher, J., Li, Z., Chan, Y., Omdahl, A., Wattiez, R., Godzik, A., and Matallana-Surget, S. (2020). Unraveling oxidative

stress resistance: molecular properties govern proteome vulnerability. bioRxiv. https://doi.org/10.1101/2020.03.09.983213v1.

Chong, P.A., Vernon, R.M., and Forman-Kay, J.D. (2018). RGG/RG motif regions in RNA binding and phase separation. J. Mol. Biol. *430*, 4650–4665.

Cohen, J. (2020). New coronavirus threat galvanizes scientists. Science *367*, 492–493.

Cohen, J., and Kupferschmidt, K. (2020a). Labs scramble to produce new coronavirus diagnostics. Science *367*, 727.

Cohen, J., and Kupferschmidt, K. (2020b). Strategies shift as coronavirus pandemic looms. Science *367*, 962–963.

Cole, T.J., and Brewer, M.S. (2018). FUSTr: a tool to find gene families under selection in transcriptomes. PeerJ *6*, e4234.

Correa, A.M., Ainsworth, T.D., Rosales, S.M., Thurber, A.R., Butler, C.R., and Vega Thurber, R.L. (2016). Viral outbreak in corals associated with an in situ bleaching event: atypical herpes-like viruses and a new megavirus infecting Symbiodinium. Front. Microbiol. *7*, 127.

Cronmiller, E., Toor, D., Shao, N.C., Kariyawasam, T., Wang, M.H., and Lee, J.H. (2019). Cell wall integrity signaling regulates cell wall-related gene expression in Chlamydomonas reinhardtii. Sci. Rep. *9*, 12204.

Curtis, B.A., Tanifuji, G., Burki, F., Gruber, A., Irimia, M., Maruyama, S., Arias, M.C., Ball, S.G., Gile, G.H., Hirakawa, Y., et al. (2012). Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature *492*, 59–65.

Dabi, M., Agarwal, P., and Agarwal, P.K. (2019). Functional validation of JcWRKY2, a Group III transcription factor toward mitigating salinity stress in transgenic tobacco. DNA Cell Biol. *38*, 1278–1291.

Darzi, Y., Letunic, I., Bork, P., and Yamada, T. (2018). iPath3.0: interactive pathways explorer v3. Nucleic Acids Res. *46*, W510–W513.

Dell'Acqua, O., Ferrando, S., Chiantore, M., and Asnaghi, V. (2019). The impact of ocean acidification on the gonads of three key Antarctic benthic macroinvertebrates. Aquat. Toxicol. *210*, 19–29.

Derelle, E., Ferraz, C., Escande, M.L., Eychenié, S., Cooke, R., Piganeau, G., Desdevises, Y., Bellec, L., Moreau, H., and Grimsley, N. (2008). Life-cycle and genome of OtV5, a large DNA virus of the pelagic marine unicellular green alga Ostreococcus tauri. PLoS One *3*, e2250.

Do, C.B., Mahabhashyam, M.S., Brudno, M., and Batzoglou, S. (2005). ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res. *15*, 330–340.

Fan, Y., Yin, X., Xie, Q., Xia, Y., Wang, Z., Song, J., Zhou, Y., and Jiang, X. (2019). Co-expression of SpSOS1and SpAHA1 in transgenic Arabidopsis plants improves salinity tolerance. BMC Plant Biol. *19*, 74.

Fang, H., and Gough, J. (2013a). DcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. Nucleic Acids Res. *41*, D536–D544.

Fang, H., and Gough, J. (2013b). A domain-centric solution to functional genomics via dcGO Predictor. BMC Bioinformatics *14*, S9.

Filée, J. (2015). Genomic comparison of closely related giant viruses supports an accordion-like model of evolution. Front. Microbiol. *6*, 593.

Filée, J., and Chandler, M. (2010). Gene exchange and the origin of giant viruses. Intervirology *53*, 354–361.

Flowers, J.M., Hazzouri, K.M., Pham, G.M., Rosas, U., Bahmani, T., Khraiwesh, B., Nelson, D.R., Jijakli, K., Abdrabu, R., Harris, E.H., et al. (2015). Whole-genome resequencing reveals extensive natural variation in the model green alga Chlamydomonas reinhardtii. Plant Cell *27*, 2353–2369.

Fujisaki, K., Tateda, C., Shirakawa, A., Iwai, M., and Abe, Y. (2018). Identification and characterization of a Tombusvirus isolated from Japanese gentian. Arch. Virol. *163*, 2477–2483.

Galindo-González, L., Mhiri, C., Deyholos, M.K., and Grandbastien, M.A. (2017). LTR-retrotransposons in plants: engines of evolution. Gene *626*, 14–25.

Gao, D., Chu, Y., Xia, H., Xu, C., Heyduk, K., Abernathy, B., Ozias-Akins, P., Leebens-Mack, J.H., and Jackson, S.A. (2018). Horizontal transfer of non-LTR retrotransposons from arthropods to flowering plants. Mol. Biol. Evol. *35*, 354–364.

Graves, M.V., Burbank, D.E., Roth, R., Heuser, J., DeAngelis, P.L., and Van Etten, J.L. (1999). Hyaluronan synthesis in virus PBCV-1-infected chlorella-like green algae. Virology *257*, 15–23.

Gregory, A.C., Zayed, A.A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., et al. (2019). Marine DNA viral macro- and microdiversity from pole to pole. Cell *177*, 1109–1123.e14.

Gunawardene, C.D., Donaldson, L.W., and White, K.A. (2017). Tombusvirus polymerase: structure and function. Virus Res. *234*, 74–86.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072–1075.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. *8*, 1494–1512.

Hayes, W.B., and Mamano, N. (2018). SANA NetGO: a combinatorial approach to using Gene Ontology (GO) terms to score network alignments. Bioinformatics *34*, 1345–1352.

He, K., Zhao, X., Chi, X., Wang, Y., Jia, C., Zhang, H., Zhou, G., and Hu, R. (2019). A novel Miscanthus NAC transcription factor MINAC10 enhances drought and salinity tolerance in transgenic Arabidopsis. J. Plant Physiol. *233*, 84–93.

Heberle, H., Meirelles, G.V., da Silva, F.R., Telles, G.P., and Minghim, R. (2015). InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinformatics *16*, 169.

Holzinger, A., and Karsten, U. (2013). Desiccation stress and tolerance in green algae: consequences for ultrastructure, physiological and molecular mechanisms. Front. Plant Sci. *4*, 327.

Hron, T., Farkašová, H., Gifford, R.J., Benda, P., Hulva, P., Görföl, T., Pačes, J., and Elleder, D. (2018). Remnants of an ancient Deltaretrovirus in the genomes of horseshoe bats (Rhinolophidae). Viruses *10*, 185.

Huang, Y., Guan, C., Liu, Y., Chen, B., Yuan, S., Cui, X., Zhang, Y., and Yang, F. (2017). Enhanced Growth Performance and Salinity Tolerance in Transgenic Switchgrass via Overexpressing Vacuolar Na+ (K+)/H+ Antiporter Gene (PvNHX1). Front. Plant Sci. *8*, 458.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol. Biol. Evol. *33*, 1635–1638.

Jackman, S.D., Vandervalk, B.P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S.A., Jahesh, G., Khan, H., Coombe, L., Warren, R.L., and Birol, I. (2017). ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res. *27*, 768–777.

Jackson, A.O., Lim, H.S., Bragg, J., Ganesan, U., and Lee, M.Y. (2009). Hordeivirus replication, movement, and pathogenesis. Annu. Rev. Phytopathol. *47*, 385–422.

Jansen, K.A., Licup, A.J., Sharma, A., Rens, R., MacKintosh, F.C., and Koenderink, G.H. (2018). The role of network architecture in collagen mechanics. Biophys. J. *114*, 2665–2678.

Jover, L.F., Effler, T.C., Buchan, A., Wilhelm, S.W., and Weitz, J.S. (2014). The elemental composition of virus particles: implications for marine biogeochemical cycles. Nat. Rev. Microbiol. *12*, 519–528.

Kajitani, R., Yoshimura, D., Okuno, M., Minakuchi, Y., Kagoshima, H., Fujiyama, A., Kubokawa, K., Kohara, Y., Toyoda, A., and Itoh, T. (2019). Platanus-allee is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions. Nat. Commun. *10*, 1702.

Katoh, K., Rozewicki, J., and Yamada, K.D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. *20*, 1160–1166.

Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., et al. (2014). The marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. *12*, e1001889.

Kerstens, M.N., Navis, G., and Dullaart, R.P. (2004). Salt sensitivity and 11beta-hydroxysteroid dehydrogenase type 2 activity. Am. J. Hypertens. *17*, 283–284.

Krupovic, M., and Koonin, E.V. (2015). Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. Nat. Rev. Microbiol. *13*, 105–115.

Kupferschmidt, K. (2020). Genome analyses help track coronavirus' moves. Science *367*, 1176–1177.

Lananan, F., Jusoh, A., Ali, N., Lam, S.S., and Endut, A. (2013). Effect of Conway Medium and f/2 Medium on the growth of six genera of South China Sea marine microalgae. Bioresour. Technol. *141*, 75–82.

Lavigne, R., Seto, D., Mahadevan, P., Ackermann, H.W., and Kropinski, A.M. (2008). Unifying classical and molecular taxonomic classification: analysis of the Podoviridae using BLASTp-based tools. Res. Microbiol. *159*, 406–414.

Lefort, V., Longueville, J.E., and Gascuel, O. (2017). SMS: smart model selection in PhyML. Mol. Biol. Evol. *34*, 2422–2424.

Legendre, M., Fabre, E., Poirot, O., Jeudy, S., Lartigue, A., Alempic, J.M., Beucher, L., Philippe, N., Bertaux, L., Christo-Foroux, E., et al. (2018). Diversity and evolution of the emerging Pandoraviridae family. Nat. Commun. *9*, 2285.

Li, Q., Zhang, X., Lv, Q., Zhu, D., Qiu, T., Xu, Y., Bao, F., He, Y., and Hu, Y. (2017). Physcomitrella patens Dehydrins (PpDHNA and PpDHNC) confer salinity and drought tolerance to transgenic Arabidopsis Plants. Front. Plant Sci. *8*, 1316.

Li, Q., Zhang, X., Lv, Q., Zhu, D., Qiu, T., Xu, Y., Bao, F., He, Y., and Hu, Y. (2018). Corrigendum: Physcomitrella patens Dehydrins (PpDHNA and PpDHNC) confer salinity and drought tolerance to transgenic Arabidopsis plants. Front. Plant Sci. *9*, 919.

Lin, S., Cheng, S., Song, B., Zhong, X., Lin, X., Li, W., Li, L., Zhang, Y., Zhang, H., Ji, Z., et al. (2015). The Symbiodinium kawagutii genome illuminates dinoflagellate gene expression and coral symbiosis. Science *350*, 691–694.

Lovati, E., Ferrari, P., Dick, B., Jostarndt, K., Frey, B.M., Frey, F.J., Schorr, U., and Sharma, A.M. (1999). Molecular basis of human salt sensitivity: the role of the 11beta-hydroxysteroid dehydrogenase type 2. J. Clin. Endocrinol. Metab. *84*, 3745–3749.

Luft, F.C. (2016). 11beta-hydroxysteroid dehydrogenase-2 and salt-sensitive hypertension. Circulation *133*, 1335–1337.

Magis, C., Taly, J.F., Bussotti, G., Chang, J.M., Di Tommaso, P., Erb, I., Espinosa-Carrasco, J., and Notredame, C. (2014). T-Coffee: tree-based consistency objective function for alignment evaluation. Methods Mol. Biol. *1079*, 117–129.

Martínez Martínez, J., Poulton, N.J., Stepanauskas, R., Sieracki, M.E., and Wilson, W.H. (2011). Targeted sorting of single virus-infected cells of the coccolithophore Emiliania huxleyi. PLoS One *6*, e22520.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. arXiv https://arxiv.org/abs/1802.03426.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Maréchal-Drouard, L., et al. (2007). The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science *318*, 245–250.

Miele, V., Penel, S., and Duret, L. (2011). Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics *12*, 116.

Mock, T., Otillar, R.P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J., Salamov, A., Sanges, R., Toseland, A., Ward, B.J., et al. (2017). Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. Nature *541*, 536–540.

Moniruzzaman, M., Martinez-Gutierrez, C.A., Weinheimer, A.R., and Aylward, F.O. (2020). Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. Nat. Commun. *11*, 1710.

Mushke, R., Yarra, R., and Kirti, P.B. (2019). Improved salinity tolerance and growth performance in transgenic sunflower plants via ectopic expression of a wheat antiporter gene (TaNHX2). Mol. Biol. Rep. *46*, 5941–5953.

Nakov, T., Beaulieu, J.M., and Alverson, A.J. (2018a). Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). New Phytol. *219*, 462–473.

Nakov, T., Beaulieu, J.M., and Alverson, A.J. (2018b). Insights into global planktonic diatom diversity: the importance of comparisons between phylogenetically equivalent units that account for time. ISME J. *12*, 2807–2810.

Nelson, D.R., Chaiboonchoe, A., Fu, W., Hazzouri, K.M., Huang, Z., Jaiswal, A., Daakour, S., Mystikou, A., Arnoux, M., Sultana, M., and Salehi-Ashtiani, K. (2019). Potential for Heightened sulfur-Metabolic Capacity in Coastal Subtropical Microalgae. iScience *11*, 450–465.

Nelson, D.R., Khraiwesh, B., Fu, W., Alseekh, S., Jaiswal, A., Chaiboonchoe, A., Hazzouri, K.M., O'Connor, M.J., Butterfoss, G.L., Drou, N., et al. (2017). The genome and phenome of the green alga Chloroidium sp. UTEX 3007 reveal adaptive traits for desert acclimatization. eLife *6*, e25783.

Nilsson, J. (2019). Protein phosphatases in the regulation of mitosis. J. Cell Biol. *218*, 395–409.

Noel, E.A., Kang, M., Adamec, J., Van Etten, J.L., and Oyler, G.A. (2014). Chlorovirus Skp1-binding ankyrin repeat protein interplay and mimicry of cellular ubiquitin ligase machinery. J. Virol. *88*, 13798–13810.

Ochoa Cruz, E.A., Cruz, G.M., Vieira, A.P., and Van Sluys, M.A. (2016). Virus-like attachment sites as structural landmarks of plants retrotransposons. Mob. DNA *7*, 14.

Onyshchenko, A., Ruck, E.C., Nakov, T., and Alverson, A.J. (2019). A single loss of photosynthesis in the diatom order Bacillariales (Bacillariophyta). Am. J. Bot. *106*, 560–572.

Parks, M.B., Nakov, T., Ruck, E.C., Wickett, N.J., and Alverson, A.J. (2018). Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). Am. J. Bot. *105*, 330–347.

Passricha, N., Saifi, S.K., Kharb, P., and Tuteja, N. (2019). Marker-free transgenic rice plant overexpressing pea LecRLK imparts salinity tolerance by inhibiting sodium accumulation. Plant Mol. Biol. *99*, 265–281.

Piacente, F., De Castro, C., Jeudy, S., Molinaro, A., Salis, A., Damonte, G., Bernardi, C., Abergel, C., and Tonetti, M.G. (2014). Giant virus Megavirus chilensis encodes the biosynthetic pathway for uncommon acetamido sugars. J. Biol. Chem. *289*, 24428–24439.

Porollo, A. (2014). EC2KEGG: a command line tool for comparison of metabolic pathways. Source Code Biol. Med. *9*, 19.

Potter, S.C., Luciani, A., Eddy, S.R., Park, Y., Lopez, R., and Finn, R.D. (2018). HMMER web server: 2018 update. Nucleic Acids Res. *46*, W200–W204.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One *5*, e9490.

Romani, G., Piotrowski, A., Hillmer, S., Gurnon, J., Van Etten, J.L., Moroni, A., Thiel, G., and Hertel, B. (2013). A virus-encoded potassium ion channel is a structural protein in the Chlorovirus Paramecium bursaria chlorella virus 1 virion. J. Gen. Virol. *94*, 2549–2556.

Rosenwasser, S., Ziv, C., Creveld, S.G.V., and Vardi, A. (2016). Virocell metabolism: metabolic innovations during host-virus interactions in the ocean. Trends Microbiol. *24*, 821–832.

Roshan, U. (2014). Multiple sequence alignment using Probcons and Probalign. Methods Mol. Biol. *1079*, 147–153.

Roth, M.S., Cokus, S.J., Gallaher, S.D., Walter, A., Lopez, D., Erickson, E., Endelman, B., Westcott, D., Larabell, C.A., Merchant, S.S., et al. (2017). Chromosome-level genome assembly and transcriptome of the green alga Chromochloris zofingiensis illuminates astaxanthin production. Proc. Natl. Acad. Sci. USA *114*, E4296–E4305.

Rowe, J., Drou, N., Youssef, A., and Gunsalus, K.C. (2019). BioSAILs: versatile workflow management for high-throughput data analysis. bioRxiv https://www.biorxiv.org/content/10.1101/509455v1.

Ruiz, E., Baudoux, A.C., Simon, N., Sandaa, R.A., Thingstad, T.F., and Pagarete, A. (2017). Micromonas versus virus: new experimental insights challenge viral impact. Environ. Microbiol. *19*, 2068–2076.

Ryu, J.Y., Kim, H.U., and Lee, S.Y. (2019). Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc. Natl. Acad. Sci. USA *116*, 13996–14001.

Santini, S., Jeudy, S., Bartoli, J., Poirot, O., Lescot, M., Abergel, C., Barbe, V., Wommack, K.E., Noordeloos, A.A., Brussaard, C.P., and Claverie, J.-M. (2013). Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. Proc. Natl. Acad. Sci. USA *110*, 10800–10805.

Schulz, F., Roux, S., Paez-Espino, D., Jungbluth, S., Walsh, D.A., Denef, V.J., McMahon, K.D., Konstantinidis, K.T., Eloe-Fadrosh, E.A., Kyrpides, N.C., and Woyke, T. (2020). Giant virus diversity and host interactions through global metagenomics. Nature *578*, 432–436.

Schvarcz, C.R., and Steward, G.F. (2018). A giant virus infecting green algae encodes key fermentation genes. Virology *518*, 423–433.

Seitzer, P., Jeanniard, A., Ma, F., Van Etten, J.L., Facciotti, M.T., and Dunigan, D.D. (2018). Gene gangs of the Chloroviruses: conserved clusters of collinear monocistronic genes. Viruses *10*.

Seligmann, H. (2018). Giant viruses as protein-coated amoeban mitochondria? Virus Res. *253*, 77–86.

Service, R.F. (2020). Coronavirus epidemic snarls science worldwide. Science *367*, 836–837.

Sheyn, U., Rosenwasser, S., Lehahn, Y., Barak-Gavish, N., Rotkopf, R., Bidle, K.D., Koren, I., Schatz, D., and Vardi, A. (2018). Expression profiling of host and virus during a coccolithophore bloom provides insights into the role of viral infection in promoting carbon export. ISME J. *12*, 704–713.

Shoguchi, E., Shinzato, C., Kawashima, T., Gyoja, F., Mungpakdee, S., Koyanagi, R., Takeuchi, T., Hisata, K., Tanaka, M., Fujiwara, M., et al. (2013). Draft assembly of the Symbiodinium minutum nuclear genome reveals dinoflagellate gene structure. Curr. Biol. *23*, 1399–1408.

Shorter, J. (2019). Phase separation of RNA-binding proteins in physiology and disease: an introduction to the JBC Reviews thematic series. J. Biol. Chem. *294*, 7113–7114.

Sievers, F., and Higgins, D.G. (2018). Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. *27*, 135–145.

Skewes-Cox, P., Sharpton, T.J., Pollard, K.S., and DeRisi, J.L. (2014). Profile hidden Markov models for the detection of viruses within metagenomic sequence data. PLoS One *9*, e105067.

Slimani, M., Pagnier, I., Raoult, D., and La Scola, B. (2013). Amoebae as battlefields for bacteria, giant viruses, and virophages. J. Virol. *87*, 4783–4785.

Sobhy, H. (2018). Virophages and their interactions with giant viruses and host cells. Proteomes *6*, 23.

Sorensen, G., Baker, A.C., Hall, M.J., Munn, C.B., and Schroeder, D.C. (2009). Novel virus dynamics in an Emiliania huxleyi bloom. J. Plankton Res. *31*, 787–791.

Steele, D.J., Kimmance, S.A., Franklin, D.J., and Airs, R.L. (2018). Occurrence of chlorophyll allomers during virus-induced mortality and population decline in the ubiquitous picoeukaryote Ostreococcus tauri. Environ. Microbiol. *20*, 588–601.

Sun, C., Feschotte, C., Wu, Z., and Mueller, R.L. (2015). DNA transposons have colonized the genome of the giant virus Pandoravirus salinus. BMC Biol. *13*, 38.

Sun, H., Li, L., Lou, Y., Zhao, H., Yang, Y., Wang, S., and Gao, Z. (2017). The bamboo aquaporin gene PeTIP4;1-1 confers drought and salinity tolerance in transgenic Arabidopsis. Plant Cell Rep. *36*, 597–609.

Taly, J.F., Magis, C., Bussotti, G., Chang, J.M., Di Tommaso, P., Erb, I., Espinosa-Carrasco, J., Kemena, C., and Notredame, C. (2011). Using the T-Coffee package to build multiple sequence alignments of protein, RNA, DNA sequences and 3D structures. Nat. Protoc. *6*, 1669–1682.

Tang, Y., Bao, X., Zhi, Y., Wu, Q., Guo, Y., Yin, X., Zeng, L., Li, J., Zhang, J., He, W., et al. (2019). Overexpression of a MYB family gene, OsMYB6, increases drought and salinity stress tolerance in transgenic rice. Front. Plant Sci. *10*, 168.

Tempel, S. (2012). Using and understanding RepeatMasker. Methods Mol. Biol. *859*, 29–51.

Teng, Z., Guo, M., Liu, X., Tian, Z., and Che, K. (2017). Revealing protein functions based on relationships of interacting proteins and GO terms. J. Biomed. Semant. *8*, 27.

Trainic, M., Koren, I., Sharoni, S., Frada, M., Segev, L., Rudich, Y., and Vardi, A. (2018). Infection Dynamics of a Bloom-Forming Alga and Its Virus Determine Airborne coccolith Emission from Seawater. iScience *6*, 327–335.

Udawat, P., Jha, R.K., Mishra, A., and Jha, B. (2017). Overexpression of a plasma membrane-localized SbSRP-like protein enhances salinity and osmotic stress tolerance in transgenic tobacco. Front. Plant Sci. *8*, 582.

Ummat, A., and Bashir, A. (2014). Resolving complex tandem repeats with long reads. Bioinformatics *30*, 3491–3498.

Ustyantsev, K., Blinov, A., and Smyshlyaev, G. (2017). Convergence of retrotransposons in oomycetes and plants. Mob. DNA *8*, 4.

Van Etten, J.L. (2003). Unusual life style of giant chlorella viruses. Annu. Rev. Genet. *37*, 153–195.

Van Etten, J.L., Agarkova, I., Dunigan, D.D., Tonetti, M., De Castro, C., and Duncan, G.A. (2017). Chloroviruses have a sweet tooth. Viruses *9*, 88.

Van Etten, J.L., and Meints, R.H. (1999). Giant viruses infecting algae. Annu. Rev. Microbiol. *53*, 447–494.

Vardi, A., Haramaty, L., Van Mooy, B.A., Fredricks, H.F., Kimmance, S.A., Larsen, A., and Bidle, K.D. (2012). Host-virus dynamics and subcellular controls of cell fate in a natural coccolithophore population. Proc. Natl. Acad. Sci. USA *109*, 19327–19332.

Villain, A., Gallot-Lavallée, L., Blanc, G., and Maumus, F. (2016). Giant viruses at the core of microscopic wars with global impacts. Curr. Opin. Virol. *17*, 130–137.

Vondrak, T., Ávila Robledillo, L., Novák, P., Koblížková, A., Neumann, P., and Macas, J. (2020). Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats. Plant J. *101*, 484–500.

Wagstaff, B.A., Vladu, I.C., Barclay, J.E., Schroeder, D.C., Malin, G., and Field, R.A. (2017). Isolation and characterization of a double stranded DNA megavirus infecting the toxin-producing haptophyte Prymnesium parvum. Viruses *9*, 40.

Wang, C., Wei, Z., Chen, K., Ye, F., Yu, C., Bennett, V., and Zhang, M. (2014). Structural basis of diverse membrane target recognitions by ankyrins. eLife *3*, e04353.

Wang, X. (2005). Regulatory functions of phospholipase D and phosphatidic acid in plant growth, development, and stress responses. Plant Physiol. *139*, 566–573.

Watanabe, T., Yamazaki, S., Maita, C., Matushita, M., Matsuo, J., Okubo, T., and Yamaguchi, H. (2018). Lateral gene transfer Between Protozoa-related giant viruses of family Mimiviridae and Chlamydiae. Evol. Bioinform. Online *14*, 1176934318788337.

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., and Zdobnov, E.M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. *35*, 543–548.

Weynberg, K.D., Allen, M.J., Ashelford, K., Scanlan, D.J., and Wilson, W.H. (2009). From small hosts come big viruses: the complete genome of a second Ostreococcus tauri virus, OtV-1. Environ. Microbiol. *11*, 2821–2839.

Weynberg, K.D., Allen, M.J., Gilg, I.C., Scanlan, D.J., and Wilson, W.H. (2011). Genome sequence of Ostreococcus tauri virus OtV-2 throws light on the role of picoeukaryote niche separation in the ocean. J. Virol. *85*, 4520–4529.

Whelan, S., and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. *18*, 691–699.

Woodrow, P., Ciarmiello, L.F., Fantaccione, S., Annunziata, M.G., Pontecorvo, G., and Carillo, P. (2012). Ty1-copia group retrotransposons and the evolution of retroelements in several angiosperm plants: evidence of horizontal transmission. Bioinformation *8*, 267–271.

Wu, M., Liu, H., Han, G., Cai, R., Pan, F., and Xiang, Y. (2017). A moso bamboo WRKY gene PeWRKY83 confers salinity tolerance in transgenic Arabidopsis plants. Sci. Rep. *7*, 11721.

Xiao, C., and Rossmann, M.G. (2011). Structures of giant icosahedral eukaryotic dsDNA viruses. Curr. Opin. Virol. *1*, 101–109.

Xu, Z., Raza, Q., Xu, L., He, X., Huang, Y., Yi, J., Zhang, D., Shao, H.B., Ma, H., and Ali, Z. (2018). GmWRKY49, a salt-responsive nuclear protein, improved root length and governed better salinity tolerance in transgenic Arabidopsis. Front. Plant Sci. *9*, 809.

Yamada, T. (2011). Giant viruses in the environment: their origins and evolution. Curr. Opin. Virol. *1*, 58–62.

Yan, L., Kuang, G., and Lin, N. (2018). Phase separation and selective guest-host binding in multi-component supramolecular self-assembly on Au(111). Chem. Commun. (Camb) *54*, 10570–10573.

Yao, W., Wang, S., Zhou, B., and Jiang, T. (2016). Transgenic poplar overexpressing the endogenous transcription factor ERF76 gene improves salinity tolerance. Tree Physiol. *36*, 896–908.

Yergaliyev, T.M., Nurbekova, Z., Mukiyanova, G., Akbassova, A., Sutula, M., Zhangazin, S., Bari, A., Tleukulova, Z., Shamekova, M., Masalimov, Z.K., and Omarov, R.T. (2016). The involvement of ROS producing aldehyde oxidase in plant response to Tombusvirus infection. Plant Physiol. Biochem. *109*, 36–44.

Yolken, R.H., Jones-Brando, L., Dunigan, D.D., Kannan, G., Dickerson, F., Severance, E., Sabunciyan, S., Talbot, C.C., Jr., Prandovszky, E., Gurnon, J.R., et al. (2014). Chlorovirus ATCV-1 is part of the human oropharyngeal virome and is associated with changes in cognitive functions in humans and mice. Proc. Natl. Acad. Sci. USA *111*, 16106–16111.

Yoosuf, N., Yutin, N., Colson, P., Shabalina, S.A., Pagnier, I., Robert, C., Azza, S., Klose, T., Wong, J., Rossmann, M.G., et al. (2012). Related giant viruses in distant locations and different habitats: Acanthamoeba polyphaga moumouvirus represents a third lineage of the Mimiviridae that is close to the megavirus lineage. Genome Biol. Evol. *4*, 1324–1330.

Zhang, H., Gao, S., Lercher, M.J., Hu, S., and Chen, W.H. (2012). EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. Nucleic Acids Res. *40*, W569–W572.

Zhu, J.K. (2002). Salt and drought stress signal transduction in plants. Annu. Rev. Plant Biol. *53*, 247–273.

Please cite this article in press as: Nelson et al., Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host & Microbe (2020), https://doi.org/10.1016/j.chom.2020.12.005

STAR★methods

Key Resources Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical Commercial Assays		
Illumina Nextera libraries	Illumina	Cat # 20018705
10x Genomics Chromium Linked Read kit	10x Genomics	Cat # CG00044
QIAGEN MagAttract HMW DNA Kit (48)	Qiagen	Cat# 67563
Deposited Data		
All sequencing reads and *de novo*-assembled genomes generated in this study	This study	Uploaded to NCBI -[https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA517804]
All supplementary datasets (Data S1–S10)	This study	Uploaded to Dryad: [https://doi.org/10.5061/dryad.7wm37pvnv]
Genomic assemblies for 67 previously sequenced microalgae	45 assemblies from other studies, hosted at NCBI and Phytozome (JGI); references in Table S1	NCBI [https://www.ncbi.nlm.nih.gov] Phytozome [https://phytozome.jgi.doe.gov/]
Transcriptomic assemblies from the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MME-TSP)	(Keeling et al., 2014)	https://figshare.com/articles/Marine_Microbial_Eukaryotic_Transcriptome_Sequencing_Project_re-assemblies/3840153/3
Viral Family (VFAM) HMMs	(Skewes-Cox et al., 2014)	http://derisilab.ucsf.edu/software/vFam/
Experimental Models: Organisms/Strains		
Table 1 contains all species used in this study for cultivation, DNA extraction, and sequencing and for species with genomes downloaded from public databases.		
Software and Algorithms		
ABySS v2.0	(Jackman et al., 2017)	https://github.com/bcgsc/abyss
HMMer v3.1b2	(Potter et al., 2018)	hmmer.org
Platanus v1.2.4	(Kajitani et al., 2019)	platanus.bio.titech.ac.jp
RepeatMasker v1.332	(Tempel, 2012)	www.repeatmasker.org
Quast v3.2	(Gurevich et al., 2013)	bioinf.spbau.ru/quast
Supernova v2.01	10x Genomics	https://support.10xgenomics.com/de-novo-assembly/software
ec2kegg v1	(Porollo, 2014)	https://sourceforge.net/projects/ec2kegg/
Diamond v2.0.2.140	(Buchfink et al., 2015)	https://github.com/bbuchfink/diamond
MUSCLE v3.8.31	Robert C. Edgar	http://etetoolkit.org
pmodeltest.py v1.4	Francos Serra, Jaime Huerte-Cepas	https://github.com/etetoolkit/pmodeltest/blob/master/pmodeltest.py
MAFFT v6.861b	(Katoh et al., 2019)	https://mafft.cbrc.jp/alignment/software/
PROBCONS v1.12	(Do et al., 2005; Roshan, 2014)	Probcons.stanford.edu/
RAxML v8.1.20	Alexandros Stamatakis	http://etetoolkit.org
readAl v1.4.rev6	(Capella-Gutiérrez et al., 2009)	http://trimal.cgenomics.org
statAl v1.4.rev6	(Capella-Gutiérrez et al., 2009)	https://github.com/etetoolkit/ext_apps/blob/master/src/trimal-1.4/source/statAl.cpp
Morpheus	Broad Institute	https://software.broadinstitute.org/morpheus/
T-COFFEE v11.00.8cbe486	(Magis et al., 2014; Taly et al., 2011)	tcoffee.crg.cat/
trimAl v1.4.rev6	(Capella-Gutiérrez et al., 2009)	http://trimal.cgenomics.org/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
PhyML v20160115	(Lefort et al., 2017)	http://phylogeny.lirmm.fr/phylo_cgi/one_task.cgi?task_type=phyml
pyfasta v0.5.2	Brent Pedersen	https://github.com/brentp/pyfasta
ETE3 toolkit	(Huerta-Cepas et al., 2016)	http://etetoolkit.org
Hierarchical Genome Assembly Process (HGAP)	Pacific Biosciences	https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/HGAP
SMRTlink	Pacific Biosciences	https://www.pacb.com/support/software-downloads/
BioSails	(Rowe et al., 2019)	https://github.com/biosails
Tensor Flow Embedding Projector	Google	https://www.tensorflow.org/resources/tools
InteractiVenn	(Heberle et al., 2015)	http://www.interactivenn.net/
Plotly	Plotly, inc.	https://plot.ly
EvolView phylogenetic tree viewer	(Zhang et al., 2012)	https://www.evolgenius.info/evolview/
iPath3	(Darzi et al., 2018)	pathways.embl.de

Resource Availability

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Kourosh Salehi-Ashtiani (ksa3@nyu.edu)

Materials Availability

There are no restrictions to the availability of the reagents. The sequencing reagents used in this study are available through their respective vendors (see Key Resources Table).

Additional information on microalgae strains used in this study can be found in Table S1.

Data and Code Availability

The datasets generated in this study and the code used for their production and analyses are available at Dryad: https://doi.org/10.5061/dryad.7wm37pvnv. The genomes are also hosted at NCBI,Bioproject:PRJNA517804.

Experimental Model and Subject Details

Details for the species used in this study, including culture collection center of origin, geographic location of origin, environmental preference, date of isolation, and others are in Table S1.

Methods Details

Microalgal strain selection and cultivation

Cultivation, DNA extraction, and sequencing of isogenic microalgae were done in several international culture collections and sequencing centers; UTEX (Austin, TX, USA), Bigelow laboratories (NMCA culture collection center, East Boothbay, ME, USA), New York University Abu Dhabi Center for Genomics and Systems Biology (Abu Dhabi, UAE), Admera Health LCC (South Plainfield, NJ, USA), and Novogene (HK).

The UTEX strains were grown on slants using one of the following media as appropriate: BG11 Medium, Bristol Medium, Cyandidium Medium, f/2 Medium, Modified Artificial Seawater Medium, Modified Bold's 3N Medium, Porphyridium Medium, Proteose Medium, Soil Extract Medium, Trebouxia Medium, or Volvox Medium with 1.5% agar as described on the UTEX website (https://utex.org/pages/algal-culture-media-recipes); grown under cool white fluorescent lights at 20ºC on a 12-hour light cycle. For species isolated and cultured at NYUAD, f/2 medium (Lananan et al., 2013), or Tris-minimal medium (https://chlamycollection.org/), was used (https://utex.org/pages/algal-culture-media-recipes).

The species chosen for sequencing were intended to represent as many microalgae phyla and as many different environments as possible. We sequenced representatives from 11 phyla (see Table S1). Most of the species were from the Chlorophyta or the Ochrophyta phyla. The project designations were algalICODE phase II (n=107, this manuscript), algalICODE phase I (n=22), NCBI-hosted (n=43), and Phytozome-hosted (n=2). Individual strain cultures were selected as representative species for their lineages or as standards to confirm workflow reproducibility (see Table S1; Data S1).

Please cite this article in press as: Nelson et al., Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host & Microbe (2020), https://doi.org/10.1016/j.chom.2020.12.005

We emphasized maximizing the sample size of each saltwater and freshwater species (Figure 1; Table S1). Of our initial effort to culture >150 species, 24 failed to produce sufficient biomass, six were contaminated, and 3 yielded reads inadequate for an assembly matching the expected size (Data S1). The *de novo* assemblies from the final batch of 107 sequenced species were combined with publicly available algal genomes for downstream analyses, including CDS predictions (Data S2), hidden Markov model (HMM)-based functional predictions, including viral and protein family domain identification, hierarchical bi-clustering (Figure 1), enrichment analyses (Figure 7), principal component analyses (Figure 7), and ternary graph-based analyses (Figure S6; Data S12).

The natural habitats of these microalgae include a range of diverse geographic locations and all climatic zones, with various temperatures, wind speeds, precipitation, and solar radiation. To allow the study of their evolution, we included species from different types of environments (from the arctic to the tropics) and both salt- and freshwater habitats (Figure 1A; Table S1). The freshwater species sequenced in this project included members of the Chlorophyta and Ochrophyta; most of the Haptophytes, Rhodophytes, and Myzozoa we sequenced were saltwater species. Most of the UTEX accessions were freshwater species (28/40); most of the NCMA accessions were saltwater species (50/57). *Alexandrium andersonii*, a mixotrophic dinoflagellate (1.7Gb), *Heterocapsa arctica*, an arctic dinoflagellate (1.3 Gb), *Lingulodinium polyedra*, a red-tide dinoflagellate (1.2 Gb), *Amphidinium gibbosum* (1.1Gb), and *Karena brevis* (1.0 Gb) were the largest *de novo*-assembled genomes in this work (Table S1).

Long-read assemblies, including those from other studies (i.e., *Chromochloris zofingiensis* (Roth et al., 2017), Thalassiosira pseudonana (Armbrust et al., 2004), and *Chlamydomonas reinhardtii* (Merchant et al., 2007)), were used to validate our high-throughput short-read assembly process. Four subtropical axenic isolates (from the United Arab Emirates) were sequenced for this study using long-read technologies, including 10x Genomics (Pleasanton, CA, USA) linked-reads, and Pacific Biosciences (Menlo Park, CA, USA) Sequel long reads. Long reads were used to validate assemblies, viral element insertions, and to resolve repeat-containing regions (Ummat and Bashir, 2014; Vondrak et al., 2020). Our results indicated that the CDSs that provided the foundational information for the comparative analyses in this manuscript were reliably determined using short reads (Illumina HiSeq X or Novoseq6000). For example, the *Chromochloris zofingiensis* genome is the highest quality algal genome published to date (Roth et al., 2017) and has 33,513 exons; our short-read assembly for *Chromochloris zofingiensis* had 33,910 exons.

Other reference microalgae used in this study as resequencing standards included *Thalassiosira psuedonana* ((Armbrust et al., 2004), *Chlamydomonas reinhardtii* (Merchant et al., 2007), *Scenedesmus* sp., *Guillardia theta* (Curtis et al., 2012), *Fragilariopsis cylindricus* (Mock et al., 2017), *Coccomyxa subellipsoidea* (Blanc et al., 2012), and *Bigelowiella natans* (Curtis et al., 2012). A comparison of the assemblies generated from the monoculture and sequencing performed in this study and previous whole-genome sequencing projects is presented in Figure 2, and QUAST and BUSCO assembly metrics are in Table S1. Hidden Markov models were used to predict structure and function from the whole-genome sequences (Figures 1B–1D; Data S6 and S7). The results for functional characterization using Enzyme Commission (EC) codes (Alborzi et al., 2017; Ryu et al., 2019), Kyoto Encyclopedia of Genes and Genomes (KEGG) designations (Porollo, 2014), and Gene Ontology (GO) terms (Hayes and Mamano, 2018; Teng et al., 2017) are in Tables S6 and S8.

DNA extraction

DNA was extracted from mature cultures with QIAGEN DNeasy Plant Maxi kits for HiSeqX 150x2 paired-end (short read) sequencing or QIAGEN MagAttract High Molecular Weight DNA Kits (48) for long-read sequencing. DNA was quantified and assayed for integrity as per the kit manufacturer's protocol. For HMW DNA extraction, briefly, DNA concentration was measured using a Qubit Fluorometer and checked for size by pulsed-field electrophoresis. A length-weighted mean of 50-70 kb was obtained, or the sample was rejected for sequencing. See Data S2 for FastQC reports and Figure S1 for the gel images showing DNA integrity. Extracted DNA with low integrity was not included in library preparations. More than 30 cultures were grown whose DNA did not meet the quality threshold; in these instances, substitute strains were chosen. The final, sequenced strains are listed in Table S1.

Sequencing

Genomic DNA sequencing was performed with Illumina paired-end (Illumina, San Diego, CA, USA), PacBio Sequel (Pacific Biosciences, Menlo Park, CA, USA), and 10x Genomics linked-reads, where indicated, (10x Genomics, Pleasanton, CA, USA) to enable reliable coverage, contig assembly, and *de novo* genomic sequence assembly. For Illumina paired-end sequencing, Nextera 2x150 bp libraries (Illumina, San Diego, CA, USA) with approximately 72 million reads per sample passing quality filters (Data S2) were used for sequencing with a HiSeqX (https://emea.illumina.com/systems/sequencing-platforms/hiseq-x.html). All reads are uploaded to the National Center for Biotechnology Information (NCBI) under the Bioproject accession PRJNA517804). The target coverage was 100x on a 100 Mbp genome. Quality control for library preparation for Illumina sequencing was done with Qubit, Tapestation (Figure S1), and qPCR. Combining these technologies assisted the validation of VFAM placement within selected genomes and ensured reliable assembly.

De novo genome assembly

De novo assembly can produce variable output depending on the source species and software used; we used both ABySS 2.0 (Jackman et al., 2017) and the Platanus (Kajitani et al., 2019) pipelines for each species sequenced with short-reads (Illumina HiSeqX (Illumina, San Diego, CA, USA)). The ABySS 2.0 command was: 'unset SLURM_NTASKS && mkdir -p $READFILE && TMPDIR=/tmp ABySS-pe -j 18 lib=pe1 k=64 name=$READFILE pe1='$READFILE R1_001.fastq.gz $READFILE R2_001.fastq.gz' –directory=/data/analysis/ABySS_pe/$READFILE'. The Platanus commands were: */platanus assemble -o $READ.OUT -f $READ-1.trimmed*

Please cite this article in press as: Nelson et al., Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host & Microbe (2020), https://doi.org/10.1016/j.chom.2020.12.005

$READ-2.trimmed -t 4 -m 72 2>assemble.$READ.log'. Details of all YML workflows used are in Data S3 and the Key Resources Table lists all essential software used in the creation and analysis of these genomes.

The output with the most single-copy, universally conserved orthologs, according to ''Based on evolutionary-informed expectations of the gene content of near-universal single-copy orthologs,'' BUSCO, was chosen for subsequent analyses (Table S1; Data S2). This step produces some bias, as ABySS produced assemblies that were much closer in size to the estimated genome sizes from close relatives. For example, *Alexandrium andersonii*, a dinoflagellate expected to have a large (>1Gb) genome, yielded an ABySS assembly of 1.6Gb, whereas its Platanus assembly was only 302 Kb. The final selected assemblies are in Data S1; they are designated with '.AAC' extensions (algalICODE), and previously published genomes are included with a '.PRE' extension. BUSCO (Waterhouse et al., 2018) and QUAST metrics (Table S1) were used to select the higher quality assembly for downstream analyses (Data S1). All versions of assemblies are in Data S1. The genome assemblies presented in this work are draft versions; they may be updated or edited in subsequent versions or as per data repository regulations.

For ~ 50% of the assemblies, Platanus yielded higher BUSCO and QUAST scores. However, Platanus was unable to assemble several genomes (see: Platanus assemblies, Data S1), producing a final assembly significantly smaller than expected for the target species. For these unassembled genomes, and others assembled with Platanus having low BUSCO scores, the ABYSS 2.0 assembly was selected for downstream analysis. For long-read assemblies, PacBio Sequel reads, SMRT sequencing data or 10X Genomics Chromium linked reads were used. PacBio long reads were assembled using the Hierarchical Genome Assembly Process (HGAP) and long-read polishing (https://www.pacb.com/) to produce genome assemblies with >99% accuracy. 10X Genomics linked reads were assembled using Supernova (v2.0.1, https://support.10xgenomics.com/de-novo-assembly/software/downloads/latest) using the megabubbles *de novo* assembly parameter.

Artificial degradation experiments

The genomes' BUSCO scores varied by genome; thus, we assessed the robustness of our assemblies and downstream comparative genomics analyses through the downsampling of the assembled genomic reads. We downsampled FASTA records from the CDSs to 25 thousand randomly selected sequences; these were used as input for the same downstream analyses we applied to the original genome set. This experiment addresses the effects of genome size and possible quality degradation on our primary conclusions. We found that comparative analyses using these downsampled genomes could reproduce our principal findings regarding the differences of microalgae from different clades and environments.

The downsampling normalized for set numbers of FASTA records and addresses potential biases introduced by genome size or overall quality differences. Our approach is similar to the rarifying approach to genome normalization. From a computational perspective, the rarifying approach is essentially the opposite process but also aims to standardize input sequences in forcing sequences to a selected number by stripping away overflow sequences. Overall, this resulted in a 74% decrease in total sequences as input for our comparative pipelines. This reduction in data is equivalent to 26% of predicted sequences, well below our average BUSCO completeness scores (see Figure 2). This technique has the added benefit of normalizing for genome size by feature scaling. For example, total gene counts are ignored, and 'whale' species (e.g., Dinoflagellates) have less impact on the conclusions. Gene content is summarized as a percentage of the genome, but fewer overall genes are used, and potentially essential genes are lost. All of the computational analyses involved in this downsampled genome set are included in Data S1.

Downsampled genomes could yield the conclusion that saltwater species had more VFAMs in their genomes on average (~2x, average= 946 VFAMs in saltwater species, 510 VFAMs in freshwater species (P=0.00096 in a one-tailed t-test)). The same unique GO term enrichment trends were seen after downsampling. For example, four GO terms involved in slowing cell cycle progression were uniquely enriched in the freshwater species' downsampled genomes (FDR P<0.003 for all terms). The GO term 'negative regulation of cell cycle' was uniquely enriched in freshwater species in the original (FDR P=5.3e-3) and the downsampled genome sets (FDR P= 1.56e-3).

Saltwater species were uniquely enriched for membrane-related functions, including vesicle formation, in the original (FDR P=5.5e-3 (Table S4)), and downsampled (FDR P=2.46e-3) genome sets. The GO term GO:0015291 for active membrane transport activity was also uniquely enriched in the downsampled saltwater genome set (FDR P=2.38e-3), which is consistent with our conclusions of uniquely enriched membrane processes in saltwater species. Finally, the 25% sequence retention downsampled genome set also showed less uniquely enriched GO terms in saltwater species than freshwater species (28 salt / 116 fresh unique GO terms in the downsampled set vs. 56 salt / 151 fresh unique GO terms in the original set). Our primary conclusions did not lose significance in these artificial data degradation experiments.

Contamination screening

All cultures in this study were routinely defined by their transfer cycle, examined by microscopy to ensure that they were unialgal and were the intended algal species and that there is no overt evidence of bacterial contamination. Axenic strains were regularly tested using an internal bacterial test media (f/2 + bactopeptone and methylamine HCl; Marine (MTM) and freshwater (FTM), both of which contain bactopeptone and malt extract protocols to ensure that they remain axenic. None of the assemblies included in the manuscript were determined to be non-axenic at the culturing phase.

In the computational phase, genomic assemblies were screened for bacterial contamination by examining G+C% distributions and BLAST hits to common contaminants. Assemblies without visible contamination, apparent from abnormal G+C% distributions or high numbers of BLASTP hits with contaminant species (Table S1, 'contamination report' sheet), were considered to be axenic.

Please cite this article in press as: Nelson et al., Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host & Microbe (2020), https://doi.org/10.1016/j.chom.2020.12.005

The BLASTP search also included the 62 previously published genomes included in the study, and all of these genomes were found to have multiple hits to bacteria; thus, genomes having more bacterial BLASTP hits than the well-curated *Emiliania huxleyi* genome (Martínez Martínez et al., 2011; Sheyn et al., 2018; Trainic et al., 2018; Vardi et al., 2012) were considered to be contaminated. This included *Entomoneis* spp, Prorocentrum minimum, Cladosiphon okamuranus, Coelasterella sp. M60, Rhodella maculata, Dunaliella sp. *RO*, *Halamphora MG11*, *Aureococcus anophagefferens*, *Prymnesium polylepis*, *Chloromonas AAM2*, and *Rhodomonas salina*.

The Diamond BlastP command was: 'diamond blastp –db uniref100 –query $GENOME.aa.fa –out $GENOME.diamond.blastp.txt –more-sensitive –threads 4 -k 1 –outfmt 6 qseqid qlen sseqid sallseqid slen qstart qend sstart send qseq sseq evalue bitscore score length pident nident mismatch positive gapopen gaps ppos qframe btop stitle salltitles qcovhsp qtitle'. The remaining assemblies were presumably axenic and used for downstream comparative analyses.

Quantification and Statistical Analysis

Genome annotation

Hidden Markov Models were used to identify repeats (Repeatmasker (Tempel, 2012), coding sequences (CDSs, https://github.com/KorfLab/SNAP), protein families (PFAMs, http://hmmer.org/), and viral families (VFAMs, (Skewes-Cox et al., 2014)). This workflow uses Bayesian classification architecture based on amino acid sequence alignment from a subset of model organisms; thus, the predictive capacity of this dataset is limited by the sample size used to create HMM models. An advantage to this type of predictive approach is that all samples are treated equally, and biases from different annotation methods are eliminated.

Overall, 102 marine strains, 62 freshwater strains, and ten strains that grow in both environments (i.e., euryhaline) were included in comparative analyses. Genomes assembled in this project (algallCODE, n=107), and the downloaded genomes (n=67), were assessed according to genome completeness using N50, total contig number, and integrity of genes from near-universally conserved ortholog groups (BUSCO, Bench-Marking Single Copy Orthologs (Waterhouse et al., 2018)).

Annotations for the 174 genomes were based on the unsupervised machine learning algorithms in HMMer (hmmer.org), with the exception of Diamond BLASTP search results. For example, HMMsearch results for VFAM HMM alignments are in Table S2. The 174 combined genomes were used as input for RepeatMasker (Tempel, 2012) and SNAP (https://github.com/KorfLab/SNAP). RepeatMasker was run with a RepeatMasker Combined Database: Dfam Consensus-20170127 run with rmblastn version 2.6.0+. RepeatMasker results are in Data S4 with the VFAM results. SNAP was run with an *Arabidopsis thaliana* HMM for all species and provided the base CDSs. The SNAP command was: 'cd /snap/ &&./snap A.thaliana.hmm $GENOME.fa -aa $GENOME.aa.fa -tx $GENOME.tr.fa -gff -quiet -name $GENOME > $GENOME.gff'. The SNAP-produced CDSs were used in downstream comparative analyses (e.g., PFAM prediction). Multiple newly sequenced genomes were predicted to be diploid or polyploid, and PFAM domains from identical sequences were not predicted to be duplication; however, heterozygous genes yielded additional PFAM counts.

HMMsearch was done with these CDSs against the PFAM-A-v31.1 database. The command was: *'hmmsearch –noali -E 0.000000001 –cpu 28 –domtblout $OUT $IN.aa.fa'*. The HMMer user guide (http://eddylab.org/software/hmmer/Userguide.pdf) provides in-depth descriptions of the reported values, including sequence coordinates for alignment and matches with HMM models, E-values, percent identity, and bias. Protein families from HMMsearch (E value < 1e-9, see Data S5) were grouped according to whether they belonged to freshwater or saltwater species (see Table S1), divided by the number of species in that group (freshwater n =102; saltwater n= 62), then compared across groups to determine the relative abundance of each PFAM domain in saltwater v. freshwater species. Euryhaline species were included as saltwater species where indicated. Exclusion analyses (e.g., ternary graphs) did not include euryhaline species.

Each subgroup of CDSs was individually used for the diamond blast command /software/diamond/diamond blastp –db /db/nrD.dmnd –query./$seqs.aa.fa –out FUSTr-"$".diamond.blastp.txt –more-sensitive –threads 24 -k 11 –outfmt 6 qseqid qlen sseqid sallseqid slen qstart qend sstart send qseq sseq evalue bitscore score length pident nident mismatch positive gapopen gaps ppos qframe btop stitle salltitles qcovhsp qtitle. These results informed investigations into inheritance patterns, including horizontal gene transfers (HGTs) for selected genes.

Homology analysis with the Basic Local Alignment Search Tool (BLAST, https://blast.ncbi.nlm.nih.gov/Blast.cgi) was performed with the accelerated Diamond Blast algorithm algorithm (Buchfink et al., 2015). The Diamond BlastP command was: 'diamond blastp –db uniref100 –query $GENOME.aa.fa –out $GENOME.diamond.blastp.txt –more-sensitive –threads 4 -k 1 –outfmt 6 qseqid qlen sseqid sallseqid slen qstart qend sstart send qseq sseq evalue bitscore score length pident nident mismatch positive gapopen gaps ppos qframe btop stitle salltitles qcovhsp qtitle'. In a BLASTP search using a cutoff of 1e-10, species queried had from 10%-90% of their proteins with significant hits (See Data S11).

EC designations were obtained using PFAM associations described in the EC domain-miner manuscript (Alborzi et al., 2017). The association list is a downloadable file, and EC designations were obtained with the command: *'while read l*; *do grep -w $l $PFAM.txt > $EC.out'* and the resultant EC designations are in Data S10. KEGG designations were obtained from EC IDs using EC2KEGG (Porollo, 2014). The EC codes were used to retrieve KEGG terms using EC2KEGG (Porollo, 2014). Briefly, the scripts in this package use 'libwww-perl' (for internet communication with KEGG, http://search.cpan.org/~mschilli/libwww-perl-6.06/lib/LWP.pm), Text-NSP (for computing the Fisher exact test, http://search.cpan.org/~tpederse/Text-NSP-1.27/), and 'Statistics-Multtest' (for correcting p

Please cite this article in press as: Nelson et al., Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host & Microbe (2020), https://doi.org/10.1016/j.chom.2020.12.005

values on multiple hypotheses testing, http://search.cpan.org/~jokergoo/Statistics-Multtest-0.13/). Shared and unique predicted KEGG pathways for saltwater and freshwater species, and an expanded description of the genes involved in these pathways is in Figure S5.

Data S10 informed the pathway comparison in Figure S5 using iPath3.0 (Darzi et al., 2018), pathways.embl.de); EC numbers for PFAMs with higher multiplicity in either saltwater or freshwater species are in Table S3. Nodes in the iPath diagrams represent enzymatic reactions with corresponding EC designations. For the iPATH analysis, EC designations were uploaded to the iPath website (https://pathways.embl.de/), and primary metabolism, secondary metabolism, and metabolism associated with microbes from diverse environments maps were used to generate the images in Figure S5.

Domain-centric GO enrichment analysis

Gene Ontology (GO) names and enrichment scores were obtained using dcGO (http://suPFAM.org/SUPERFAMILY/dcGO/, (Fang and Gough, 2013a, 2013b)) from CDSs, EVOPs, or genes under positive selection with significant (e=1e-9) hits against the PFAM-A database. Algorithms used in dcGO convert Z-scores and FDR-corrected P-values were calculated based on hypergeometric distributions.

Shared and unique enriched GO terms were delineated with InteractiVenn (www.interactivenn.net). All supplementary files with the '.ivenn' extension can be uploaded to www.interactivenn.net, and gene lists corresponding to the displayed numbers are provided as copiable text (see Supplemental Datasets).

Validation using reference genomes

Cultures of microalgal species that have previously undergone whole-genome sequencing have provided much insight into microalgal evolution and ecology (see references in Table S1). Cultures were grown of reference species to validate the sequencing and assembly workflows. Estimates of size (Data S1), coding sequences (CDSs, Data S2 and S3), aligned homologs in known species, repeat composition (Data S4) were similar for our reconstructed genomic assemblies and the NCBI-hosted reference genomes (Figure S2; Table 1). The similarity across sequencing and bioinformatics pipelines suggests reproducibility for whole-genome sequencing and their functional inferences. All genomes, including those sequenced in this project (algalICODE,.AAC), and those downloaded (previously sequenced,.PRE) were pooled for parallel CDS predictions (Data S1) and used for downstream analyses.

The genome of *Chromochloris zofingiensis*, sequenced because of its capacity for oil and pigment production, is regarded to be the highest quality genome currently available (Roth et al., 2017). We compared the functional predictions inferred from this genome with our re-sequenced *de novo* assembly for validation and found that their HMM profiles were identical in their regards to matches with the PFAM-A and the VFAM (viral family, available at http://derisilab.ucsf.edu/software/VFAM/, (Skewes-Cox et al., 2014)) databases. Other assemblies in re-sequencing validation controls included *Micromonas pusilla* and *Galdieria sulfaria* (Figure S3). Our assemblies were nearly identical in their HMM matches to public databases (Figure 1; Data S3 and S5). The similarity of these HMM matches confirms the robustness of the technical aspects of our sequencing and genomic assembly and the use of HMM profiles for gene function prediction. All of the assemblies resulting from our resequencing yielded highly similar counts regarding gene content, G+C% and size estimates, and functional domain prediction, including BUSCO results and BLAST hits.

Phylogeny reconstructions

Rubisco large subunit genes (RbcL) were identified using BlastP (Lavigne et al., 2008), extracted, and trimmed using trimAl (Capella-Gutiérrez et al., 2009) in ETE3 (http://etetoolkit.org/). Roughly half of the species used in this project, including both newly sequenced and downloaded genomes, had RbcL sequences that could be used to create a robust phylogeny. For lists of species, and to view the tree, see the 'ETE3' folder in Data S1. This folder contains all input, intermediate, and output files from the ETE3 workflow (e.g., RAxML, MAFFT (Katoh et al., 2019), pmodeltest, PROBCONS, readAl and statAl (Capella-Gutiérrez et al., 2009), T-COFFEE (Magis et al., 2014; Taly et al., 2011), PhyML (Lefort et al., 2017), and pyfasta. Clean sequences (~1400 bp) were aligned with ClustalOmega (Sievers and Higgins, 2018). The command used was: 'ete3 build -w full_ultrafast_modeltest_bootstrap -a RBCL_AAC_corrected.aa.fa -o RBCL_AAC_corrected.ETE3.out –clearall -C 28'. The Newick format tree output from ETE3 was visualized in Evolview (Zhang et al., 2012). (www.evolgenius.info). Other tracks loaded in Evolview included HMMer VFAM results and other tracks summarized by Table S1. For the inference of horizontal gene transfer events, phylogenies reconstructed from aligned sequences of HGT candidates (e.g., Figure S4) were compared with their rbcL-based phylogenies and published species trees (Alverson and Theriot, 2005).

Amino acid residue comparisons

Translated coding sequences produced as described in the 'Genome Annotation' section were used. The SNAP program adds asterisks to indicate a peptide terminal; these and headers were removed, and the twenty amino acids were quantified and classified into salt or fresh based on their designations in Table S1. Counts for each amino acid were divided by the total number of residues within each species, then pooled counts were divided by the number of species in each group (salt n=100; fresh n=61). The compositions of multiple amino acids were significantly different between these two groups, as indicated in the legend of Figure 7. A two-tailed Student's t-test with Welch correction was used for each amino acid comparison.

Please cite this article in press as: Nelson et al., Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host & Microbe (2020), https://doi.org/10.1016/j.chom.2020.12.005

Dimension reduction analyses

PFAM dCNs from total CDSs, genes under positive selection, and endogenous, viral-origin PFAMs (EVOPs)) in the microalgal genomes were used as inputs for principal component analysis (PCA) in ClustVis (https://biit.cs.ut.ee). Loadings and scores, relative to Figure 7, are in Table S3. Marine and terrestrial species were observed to separate from PC1; thus, scores contributing to this separation were inferred to contribute to the phenotypes of the respective, clustered species.

T-distributed neighbor embedding (t-SNE) and uniform manifold approximation projection (UMAP (McInnes et al., 2018) used PFAM dCNs as in the PCAs using the TensorFlow projector online tool available at https://www.tensorflow.org/resources/tools. Briefly, PFAM counts were used as vectors for n=174 species in 4,217 different Pfams. PFAM descriptions were used to trace accessions and functions. Embeddings in Figure S3 were obtained using eight nearest neighbors in high dimensional space, according to (McInnes et al., 2018). For Figure S3 panels B-D, the point with the largest text was selected to reveal the nearest neighbors, annotated with smaller text, in high-dimensional space.

Ternary graphs

Ternary plots were graphed in Plotly (https://plot.ly). Briefly, PFAM counts were used as variable criteria in the groups. Values for the combined groups were weighted as proximity to a corner such that higher counts minimized distance. Genes with an expression value of zero in one out of three groups were found along edges and considered to be excluded from the other two groups.

Positive selection analysis

Sixty genomes from small microalgae, 30 freshwater, and 30 saltwater species, and their CDSs and amino acid sequences were used in the FUSTr workflow (Cole and Brewer, 2018) to calculate dS/dN ratios for genes and their phylogenies (see Data S13). This multi-software computational pipeline is similar to PAML in its scope of analysis. Briefly, FUSTr finds genes in transcriptomic datasets under strong positive selection and automatically detects isoform designation patterns in transcriptome assemblies to maximize phylogenetic independence in downstream analysis. Non-nucleotide characters were detected and removed from sequences. Header patterns are examined to identify isoforms and eliminate redundancies. Coding sequences are extracted from transcripts with Transdecoder v3.0.1 (Haas et al., 2013). For the purposes of this manuscript, the longest open reading frame (ORF) for the transcript under consideration was used. Then, homologies in peptide sequences are assessed with BLASTP in DIAMOND (v.0.9.10, (Buchfink et al., 2015)) with an *e*-value cutoff of 10^{-5}. This homology matrix is parsed into putative gene families with SiLiX (v.1.2.11) transitive clustering. SiLiX has been shown to be faster, have more efficient memory allocation, and reduce domain chaining compared to other clustering algorithms such as MCL (Miele et al., 2011). Predicted peptides included in families had 35% minimum identity, 90% minimum overlap, at least 100 amino acids.

Multiple amino acid sequence alignments are generated by the software selected by MAFFT v7.221 (Katoh et al., 2019). Hanging sequence ends were removed with Trimal v1.4.1's *gappyout* algorithm (Capella-Gutiérrez et al., 2009). Phylogenies of each family's untrimmed amino acid multiple sequence alignment are reconstructed using FastTree v2.1.9 (Price et al., 2010). Trimmed multiple sequence codon alignments are then generated by reverse translation of the amino acid alignment using the CDS sequences. The final output shows detected gene families that are under strong selection and the average d_N/d_S per family. Details per codon position within family alignments include: alpha mean posterior synonymous substitution rate at a site; beta mean posterior non-synonymous substitution rate at a site; mean posterior beta-alpha; posterior probability of negative selection at a site; posterior probability of positive selection at a site; Empirical Bayes Factor for positive selection at a site; potential scale reduction factor; and estimated effective sample site for the probability that beta exceeds alpha.

For our FUSTr analysis, a total of 4,403,010 CDSs were used as input; CDSs and predicted proteins from these genomes were organized into 2,991,159 families; 792 had >15 sequences per family. The FUSTr analysis detected 335 families under strong positive selection; 401,500 PFAMs (HMMsearch, e<0.0000000001, n = 60 species) were identified from genes in these families as being under strong positive selection. The FUSTr dataset generated for this project includes 792 phylogenetic trees and includes 13,841 VFAM-CDSs and other proteins with strong positive selection signatures.

Viral family sequence identification

A caveat in using expression-only (RNA) data to describe VFAMs is the possibility that the assembled RNA sequence was contamination and not expressed from the microalgal genome. Here, long-read sequencing technologies, including 10x Genomics (Pleasanton, CA, USA) linked-reads, and Pacific Biosciences (Menlo Park, CA, USA) Sequel long reads were combined to ensure scaffold integrity. In contigs with high coverage of long reads, VFAMS were interspersed with known microalgal genes in nuclear genome contigs (e.g., Figure 2). Synteny was maintained in a manner that suggested integrations of VFAM sequences in core microalgal assembled contigs. In addition to our long-read validations, we used high-quality, previously published, microalgal genomes that were also created from long reads such as *Chromochloris zofingiensis* (Roth et al., 2017)and confirmed that VFAMs were integrated into these published chromosomes.

Viral family sequences within genomes were discovered using HMMs built from Markov clustering and multiple sequence alignments (Skewes-Cox et al., 2014) from all of the virally annotated proteins in RefSeq as of 2014. These VFAMs can detect viral sequences with high accuracy and did not cluster non-viral sequences into the Markov clusters in the published test sets. We used a strict 1e-9 E-value to call VFAM domains in the translated CDS (tCDS) from the microalgal genomes. Altogether, 91,757 distinct VFAM tCDSs were discovered and extracted using in-house scripts. The VFAM HMMsearch results, extracted tCDSs, and the scripts

Please cite this article in press as: Nelson et al., Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host & Microbe (2020), https://doi.org/10.1016/j.chom.2020.12.005

used to process them are in Data S3 and S6. Extracted tCDS were used as input for dimension reduction analyses, BLAST, dcGO enrichment based on clade or environment (see Data S13), and DFAM searches to find transposable elements (Data S5). VFAM-CDSs were used as input for HMMsearch against the Dfam repeat database (DFAM.HMM) to search for transposable elements and 720 out of 91,757 of these sequences (0.78%) were predicted to code for transposon elements (E-value=1e-9).

After establishing the presence of VFAMs in microalgal chromosomes, we validated their presence in expressed genes. Transcriptomic data from the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP, (Keeling et al., 2014)) was downloaded from https://figshare.com/articles/Marine_Microbial_Eukaryotic_Transcriptome_Sequencing_Project_re-assemblies/3840153/3to. VFAM predictions we performed on the MMETSP dataset are in Data S9. From HMM-based VFAM detection performed in the same manner as with the predicted CDSs from *de novo* and downloaded assemblies, MMETSP *de novo* transcriptomes were confirmed to contain substantial portions of VFAMs. Briefly, HMMsearch was done with transcriptome fasta files against the PFAM-A-v31.1 database. The command was: *'hmmsearch –noali -E 0.000000001 –cpu 28 –domtblout $OUT $IN.aa.fa'*. The HMMer user guide (http://eddylab.org/software/hmmer/Userguide.pdf) describes the reported values, including sequence coordinates for alignment and matches with HMM models, E-values, percent identity, and bias.

Additional confirmation for the viral origins of VFAM HMM-identified sequences was carried out by examining BLAST results, extracting high-confidence hit sequences from public repositories, and performing alignments and maximum likelihood phylogenies (see Figures S7 and S8). The BLAST command for the 91,757 VFAM-CDSs was 'diamond blastp –db nrD.dmnd –query all_vfam_extracted.aa.fa –out all_vfam_extracted.diamond.blastp.txt –more-sensitive –threads 4 -k 10 –outfmt 6 qseqid qlen sseqid sallseqid slen qstart qend sstart send qseq sseq evalue bitscore score length pident nident mismatch positive gapopen gaps ppos qframe btop stitle salltitles qcovhsp qtitle''. Altogether, BLAST results were inspected to discern phylogenies for the viral sequences. The top 25 hits were analyzed for phylogeny and Biosample location. Further confirmation of the likelihood of HGT events was instilled if multiple high-confidence hits came from unrelated lineages but shared a similar habitat.

Alignments were performed using MUSCLE with 100 iterations and with using diagonal sequence alignments to detect islands of homologous sequence with greater distance. Maximum likelihood phylogenies using these alignments and a neighbor-joining tree construction method. The substitution model for branch reconstruction was the Whelan and Goldman (WAG) matrix (Whelan and Goldman, 2001) with a 2.0 transition/transversion ratio. Rate variation was included in likelihood calculations under four categories and a gamma distribution parameter of 1.0.

A caveat to this study is that we cannot conclusively attribute every VFAM as having been directly transferred from a virus to the algae containing that VFAM in its genome, although recent studies have shown a high likelihood of ultimate viral origin (i.e., *de novo* gene generation) (Legendre et al., 2018). Additional evidence supporting this scenario has been presented in (Graves et al., 1999; Hron et al., 2018; Romani et al., 2013; Schvarcz and Steward, 2018; Seligmann, 2018; Slimani et al., 2013; Sobhy, 2018; Sun et al., 2015; Watanabe et al., 2018).

Cell Host & Microbe, Volume *29*

Supplemental Information

Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution

David R. Nelson, Khaled M. Hazzouri, Kyle J. Lauersen, Ashish Jaiswal, Amphun Chaiboonchoe, Alexandra Mystikou, Weiqi Fu, Sarah Daakour, Bushra Dohai, Amnah Alzahmi, David Nobles, Mark Hurd, Julie Sexton, Michael J. Preston, Joan Blanchette, Michael W. Lomas, Khaled M.A. Amiri, and Kourosh Salehi-Ashtiani

SUPPLEMENTAL INFORMATION

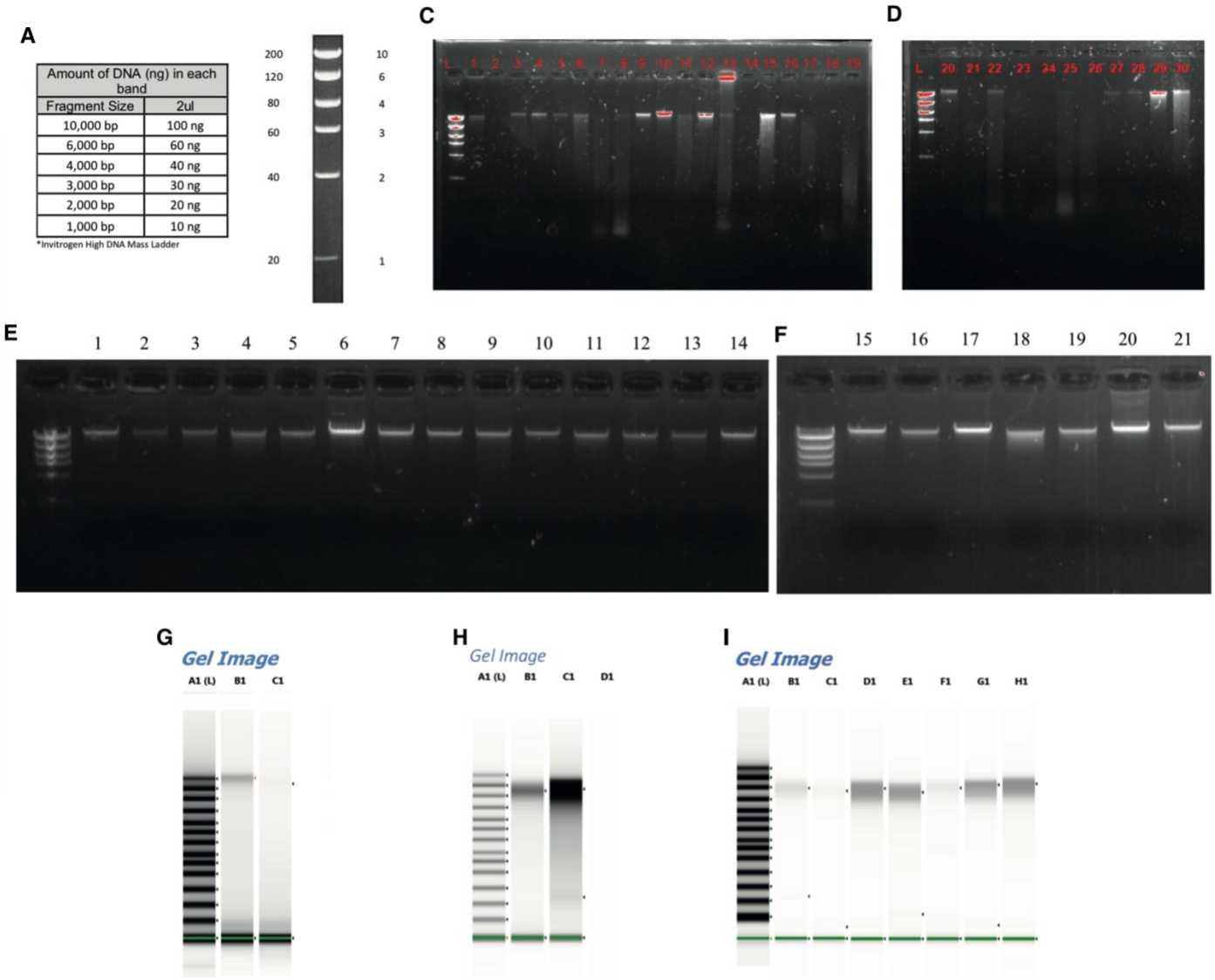


Figure S1. Initial DNA integrity evaluations. Related to Fig. 1. (A): gDNA integrity from gel images for genomic DNA (gDNA) extractions from representative strains (51/100, or approximately half of the extractions). gDNA extraction, library preparation, and sequencing were done in batches as culturing was completed. Ladder composition (Lambda 10kb ladder). (B-E): Extracted gDNA was run on a 1% agarose gel to ensure integrity; gDNA extractions with highly fragmented DNA, as evidenced by band smearing, were not included in library preps. (G-I): Final multiplexed library (Nextera) Tapestation images.

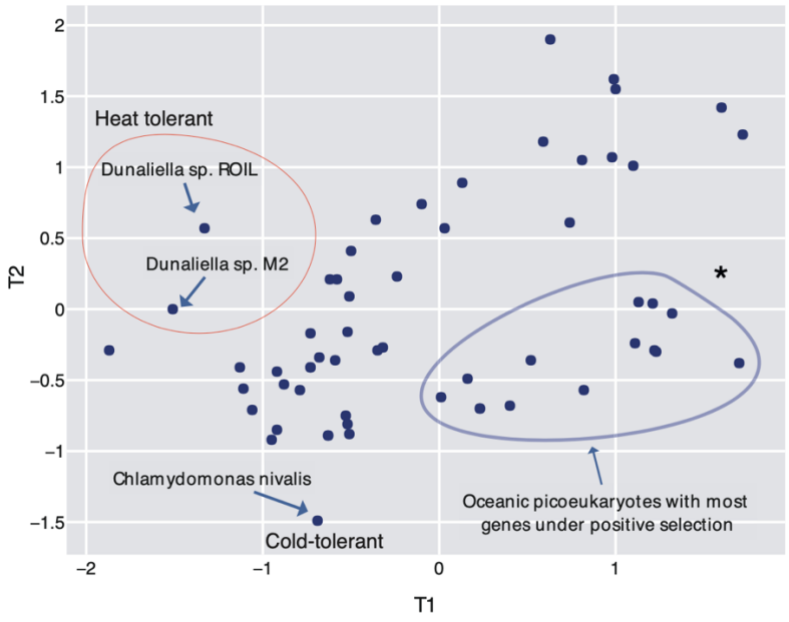


Figure S2. tSNE of newly sequenced species based on their PFAMs. Related to Fig. 1. (A) PFAM SNE scatterplot shows separate clustering of saltwater and freshwater species based on their functional domains (PFAM count, n=8416, E-value < 1e-9). Notable exceptions included large genome species, including multicellular Xanthophytes (yellow-green algae) and unicellular dinoflagellates, including *Symbiodinium* and *Alexandria* species. These results agree with our Pearson correlation and PCA analyses in showing PFAM clustering based on the environment.

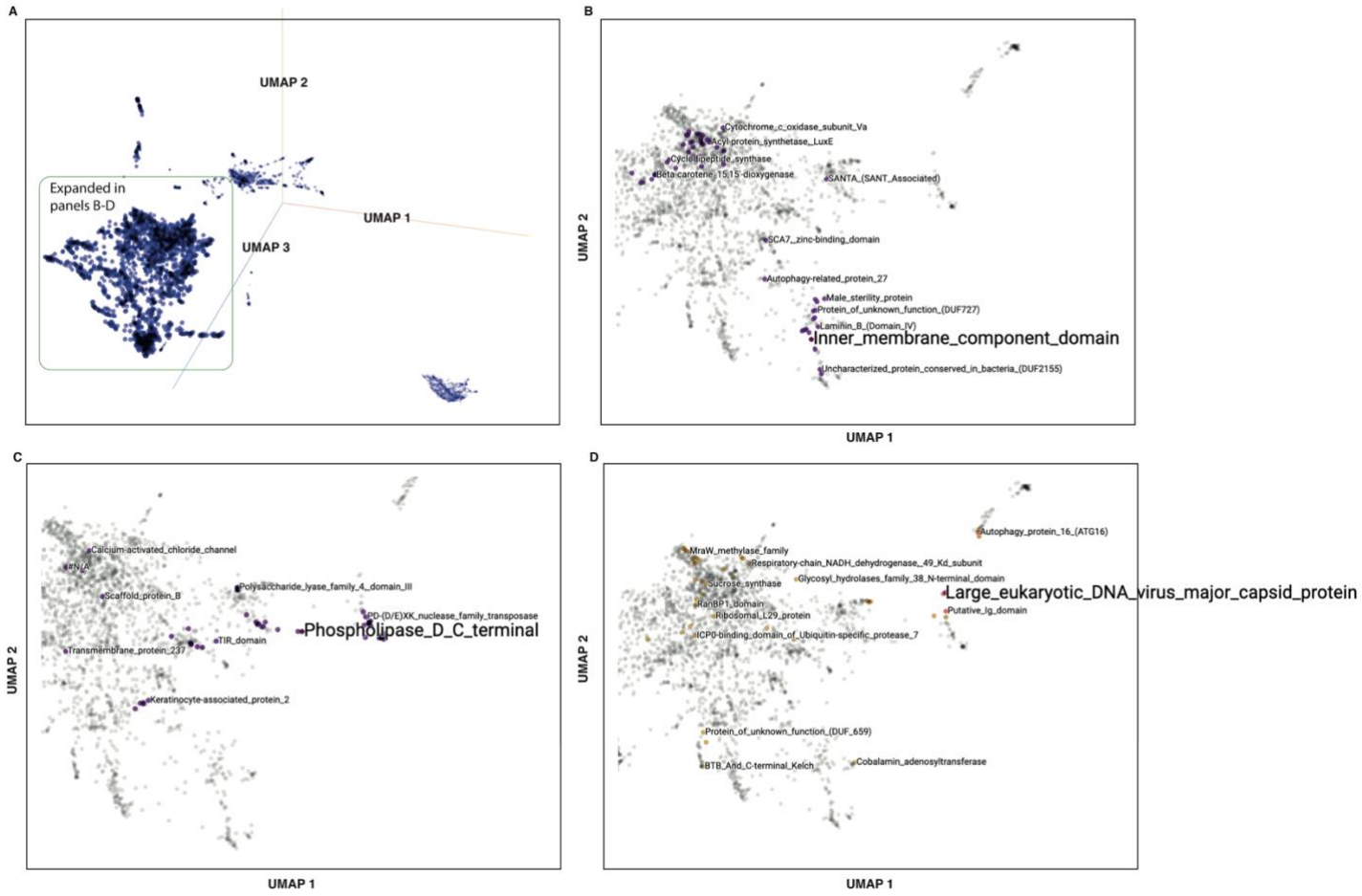


Figure S3. UMAP analyses revealed clusters of genes with unintuitive conservation patterns. Related to Fig. 1. UMAP uses local neighbor distances for embedding, and this information can be used to find the nearest neighbors for a given data point representing either a species' PFAM count array (n=8042) or a PFAM's distribution across species (n=180). From manual inspection of several PFAMS identified as enriched in saltwater species, and their nearest neighbors, we observed collections of nearest neighbors that were membrane-centric and contained one or two viral-related PFAMs. For example, querying proteins from the highlighted cluster in Fig. 1 for saltwater species, viral-related PFAM nearest neighbors could be identified. Conservation of shared functions within clusters is implied; however, we further propose that the nearest-neighbor clusters represent sets of co-evolving, or co-conserved, genes.

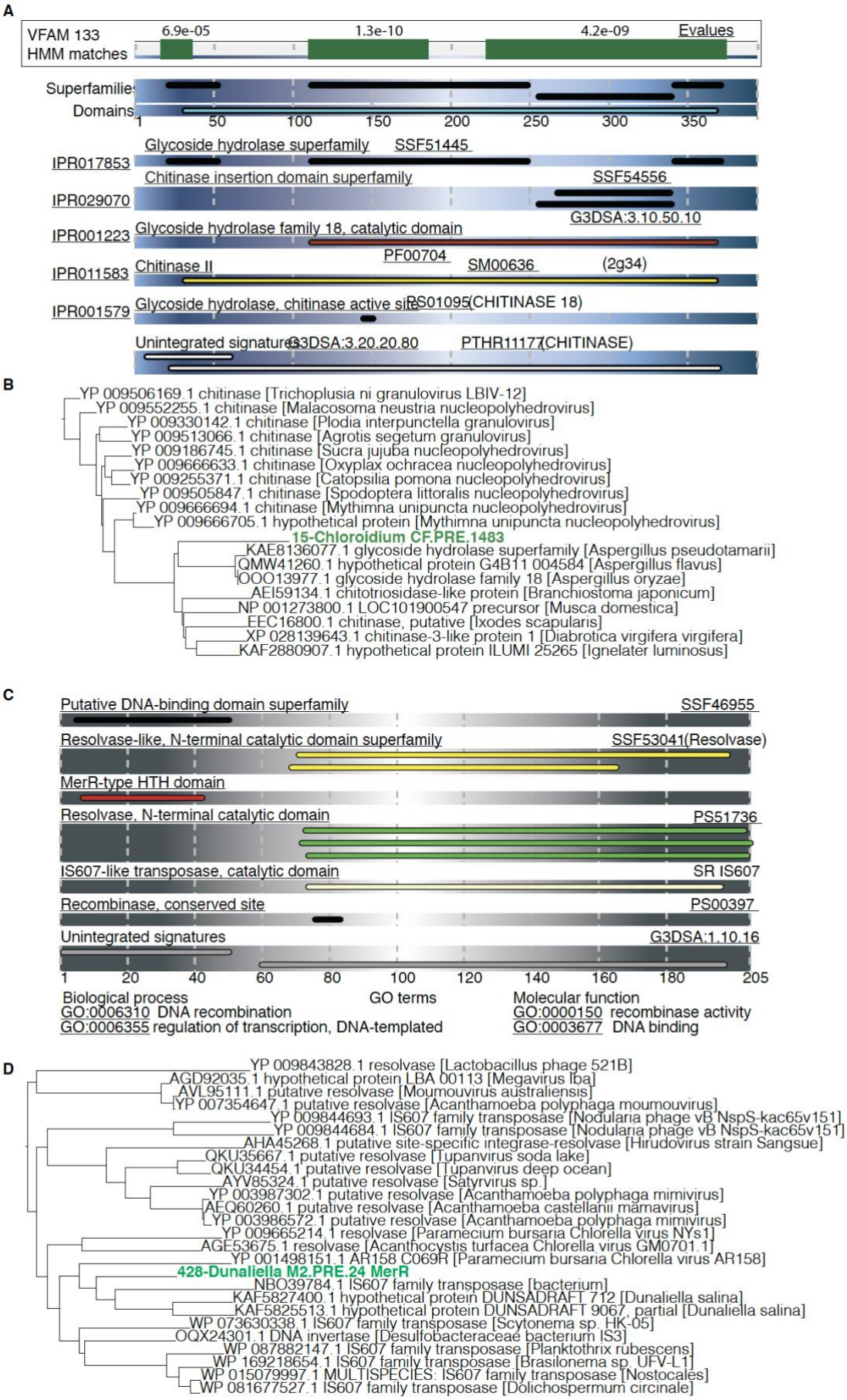

A
VFAM 133
HMM matches
6.9e-05
1.3e-10
4.2e-09
Evalues
Superfamilies
Domains
1
50
100
150
200
250
300
350
Glycoside hydrolase superfamily
SSF51445
IPR017853
Chitinase insertion domain superfamily
SSF54556
IPR029070
G3DSA:3.10.50.10
IPR001223
Glycoside hydrolase family 18, catalytic domain
PF00704
IPR011583
Chitinase II
SM00636
(2g34)
IPR001579
Glycoside hydrolase, chitinase active site
PS01095(CHITINASE 18)
Unintegrated signatures
G3DSA:3.20.20.80
PTHR11177(CHITINASE)
B
YP 009506169.1 chitinase [Trichoplusia ni granulovirus LBIV-12]
YP 009552255.1 chitinase [Malacosoma neustria nucleopolyhedrovirus]
YP 009330142.1 chitinase [Plodia interpunctella granulovirus]
YP 009513066.1 chitinase [Agrotis segetum granulovirus]
YP 009186745.1 chitinase [Sucra jujuba nucleopolyhedrovirus]
YP 009666633.1 chitinase [Oxyplax ochracea nucleopolyhedrovirus]
YP 009255371.1 chitinase [Catopsilia pomona nucleopolyhedrovirus]
YP 009505847.1 chitinase [Spodoptera littoralis nucleopolyhedrovirus]
YP 009666694.1 chitinase [Mythimna unipuncta nucleopolyhedrovirus]
YP 009666705.1 hypothetical protein [Mythimna unipuncta nucleopolyhedrovirus]
15-Chloroidium CF.PRE.1483
KAE8136077.1 glycoside hydrolase superfamily [Aspergillus pseudotamarii]
QMW41260.1 hypothetical protein G4B11 004584 [Aspergillus flavus]
OOO13977.1 glycoside hydrolase family 18 [Aspergillus oryzae]
AEI59134.1 chitotriosidase-like protein [Branchiostoma japonicum]
NP 001273800.1 LOC101900547 precursor [Musca domestica]
EEC16800.1 chitinase, putative [Ixodes scapularis]
XP 028139643.1 chitinase-3-like protein 1 [Diabrotica virgifera virgifera]
KAF2880907.1 hypothetical protein ILUMI 25265 [Ignelater luminosus]
C
Putative DNA-binding domain superfamily
SSF46955
Resolvase-like, N-terminal catalytic domain superfamily
SSF53041(Resolvase)
MerR-type HTH domain
Resolvase, N-terminal catalytic domain
PS51736
IS607-like transposase, catalytic domain
SR IS607
Recombinase, conserved site
PS00397
Unintegrated signatures
G3DSA:1.10.16
1
20
40
60
80
100
120
140
160
180
205
Biological process
GO:0006310 DNA recombination
GO:0006355 regulation of transcription, DNA-templated
GO terms
Molecular function
GO:0000150 recombinase activity
GO:0003677 DNA binding
D
YP 009843828.1 resolvase [Lactobacillus phage 521B]
AGD92035.1 hypothetical protein LBA 00113 [Megavirus lba]
AVL95111.1 putative resolvase [Moumouvirus australiensis]
YP 007354647.1 putative resolvase [Acanthamoeba polyphaga moumouvirus]
YP 009844693.1 IS607 family transposase [Nodularia phage vB NspS-kac65v151]
YP 009844684.1 IS607 family transposase [Nodularia phage vB NspS-kac65v151]
AHA45268.1 putative site-specific integrase-resolvase [Hirudovirus strain Sangsue]
QKU35667.1 putative resolvase [Tupanvirus soda lake]
QKU34454.1 putative resolvase [Tupanvirus deep ocean]
AYV85324.1 putative resolvase [Satyrvirus sp.]
YP 003987302.1 putative resolvase [Acanthamoeba polyphaga mimivirus]
AEQ60260.1 putative resolvase [Acanthamoeba castellanii mamavirus]
YP 003986572.1 putative resolvase [Acanthamoeba polyphaga mimivirus]
YP 009665214.1 resolvase [Paramecium bursaria Chlorella virus NYs1]
AGE53675.1 resolvase [Acanthocystis turfacea Chlorella virus GM0701.1]
YP 001498151.1 AR158 C069R [Paramecium bursaria Chlorella virus AR158]
428-Dunaliella M2.PRE.24 MerR
NBO39784.1 IS607 family transposase [bacterium]
KAF5827400.1 hypothetical protein DUNSADRAFT 712 [Dunaliella salina]
KAF5825513.1 hypothetical protein DUNSADRAFT 9067, partial [Dunaliella salina]
WP 073630338.1 IS607 family transposase [Scytonema sp. HK-05]
OQX24301.1 DNA invertase [Desulfobacteraceae bacterium IS3]
WP 087882147.1 IS607 family transposase [Planktothrix rubescens]
WP 169218654.1 IS607 family transposase [Brasilonema sp. UFV-L1]
WP 015079997.1 MULTISPECIES: IS607 family transposase [Nostocales]
WP 081677527.1 IS607 family transposase [Dolichospermum circinale]

Figure S4. Algal genes with putative viral origins and inferred adaptive functions. Related to Figs. 3-6. (A): Putative Chitinase with viral origin in the Chloroidium sp. UTEX 3077 genome. Locations of viral and functional sequences in a Chloroidium sp. UTEX 3077 [CF] gene [15-Chloroidium_CF.PRE.1483]. Three high confidence matches to VFAM_133 were found in a Chloroidium gene with predicted chitinase activity. VFAM_133 includes NPHVs with diverse known host ranges including insects, fungi, plants, and algae. Interproscan reported associated GO terms of GO:0005975 carbohydrate metabolic process, Molecular Function: GO:0004553 hydrolase activity, hydrolyzing O-glycosyl compounds, GO:0008061 chitin-binding. Homologs found through BLASTP searches with the highest identity to the Chloroidium amino acid sequence were either known or predicted chitinases (see Dataset S11). Maximum likelihood phylogenies support the hypothesis that this putative carbon assimilation gene was spread through a viral vector. This sequence has multiple domains for glycoside hydrolases and chitinases characteristic of other multifunctional, viral-origin genes.
(B): Maximum likelihood phylogeny of 15-Chloroidium_CF.PRE.1483 and 66 homologs with the greatest percent identity as revealed by a BLASTP against the non-redundant (NR) and NCBI Virus databases. Portions of this figure were generated with InterProScan 5.46-81.0.
(C) Alignment of the *Dunaliella sp. M2*'s Tupanvirus homolog with its closest BlastP hits from the NR database. (A) A recombinase/DNA invertase/transposase domain was found directly adjacent to a MerR mercury-responsive transcription factor PFAM domain (E-value<2.6e-11, InterProScan 5.46-81.0), and both domains were determined to be of viral origin in *Dunaliella sp. M2*, an extremophilic desert alga (428-Dunaliella_M2.PRE.24).
(D) This sequence was used as a query for BLASTP searches against the NR database, which revealed homologs in a variety of giant dsDNA viruses (see Dataset S11). Top hits included proteins from *Tupanviruses* (E-value < 1e-51) and *Acanthamoeba polyphaga moumouvirus* (E-value < 5e-54, see also Dataset S11). The homologs of the *Dunaliella* MerR/Transposase/Resolvase (MerRtr) CDSs with at least 97% coverage across the >200 aa alignment region were aligned with MUSCLE, and 3/18 of the sequences shared a unique ~ ten aa track with the *Dunaliella* sequence. The only two proteins that retained synteny with the *Dunaliella* protein were from 'Deep Ocean' and 'Soda Lake' *Tupanviruses*. The unique retention of these ten amino acids in *Dunaliella* and the *Tupanviruses* suggests that *Dunaliella* was also a host for this giant virus. A maximum likelihood phylogeny was reconstructed showing inheritance from viruses to the Dunaliella genus, with representatives of *Dunaliella sp. M2* and *Dunaliella salina* (Polle et al., 2017).

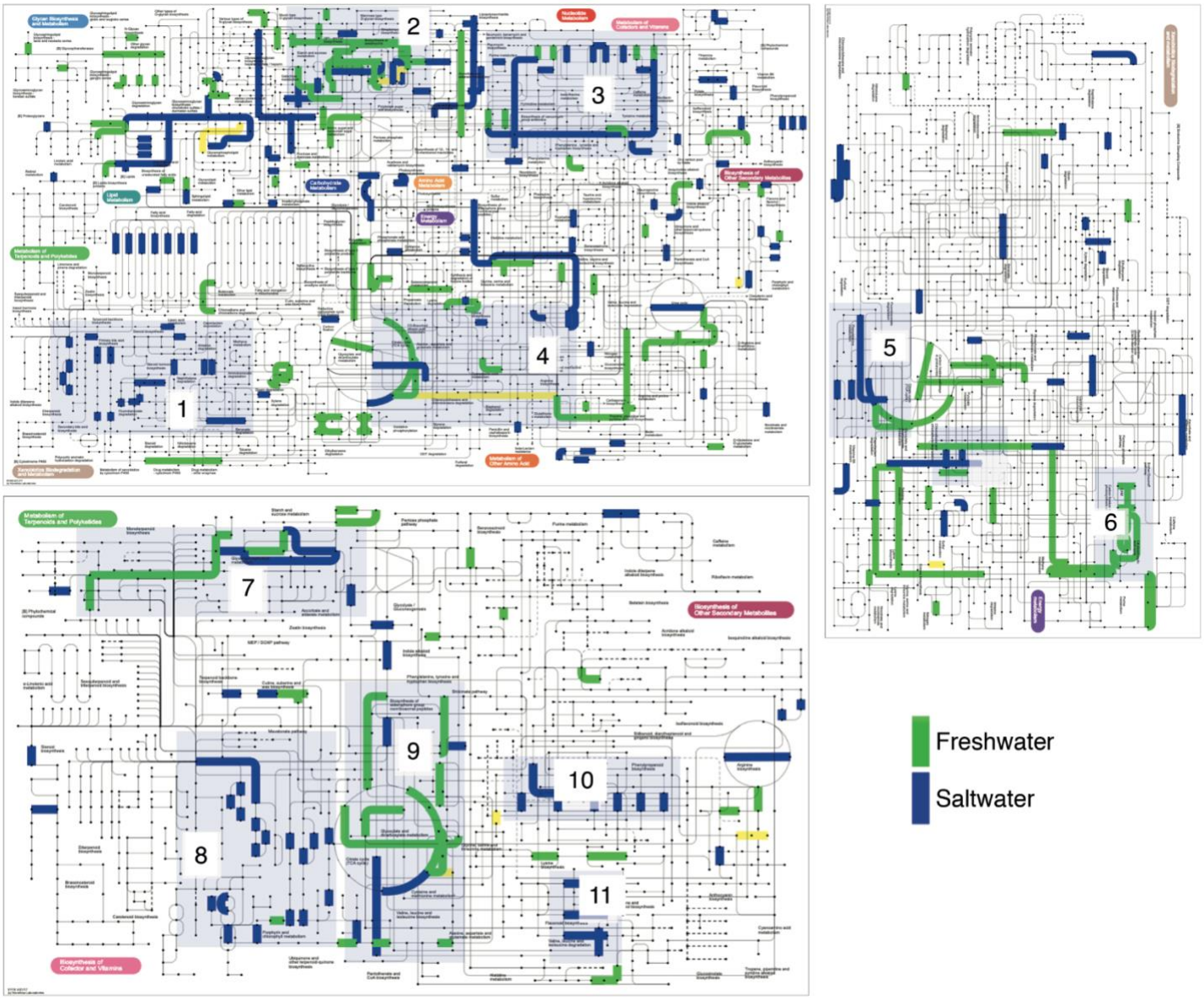


Figure S5. Pathways (iPath3, EMBL (Darzi et al., 2018)) with more protein families in saltwater species (blue), freshwater species (green), or in both groups (yellow). Related to Fig. 7. See Dataset S10 for source data. The ECs for either fresh or saltwater species originate from PFAM domains that had higher multiplicity in the respective group. The KEGG orthology (KO) designations in Table S3 can be uploaded to the iPath 3.0 website (https://pathways.embl.de/ipath3.cgi) to reproduce the maps and highlighted reactions shown in this figure. At yellow edges, different PFAMs link to the same reaction. Maps consist of primary metabolism (KEGG map01100, highlighted regions numbered 1-4), biosynthesis of secondary metabolites (KEGG map01110, highlighted areas numbered 5,6), and microbial analysis in diverse environments (KEGG map01120, highlighted regions numbered 8-10). [1]: Steroid hormone biosynthesis. Saltwater species had more reactions in these pathways. [2] Freshwater species had more sugar metabolism reactions. [3] Different reactions for purine metabolism were more abundant in each group; reactions had more representative PFAMs in saltwater species while other reactions were overrepresented in freshwater species. [4] Glyoxylate cycle reactions were also differentially abundant in fresh- and saltwater species. [5] Reactions in phenylalanine metabolism were more abundant in saltwater species. [6] Freshwater species had more reactions for carbon fixation, including C4 and CAM-related metabolic reactions. [7]: Glycerophospholipid biosynthesis. [8]: Terpenoid biosynthesis. [9] Bicarboxylate biosynthesis. [10] Phenylpropanoid biosynthesis. For further details on specific reactions, upload the KOs in Table S3 to the iPath 3.0 website (https://pathways.embl.de/ipath3.cgi); the highlighted edges can be selected and are linked to biochemical reaction information and references in KEGG.

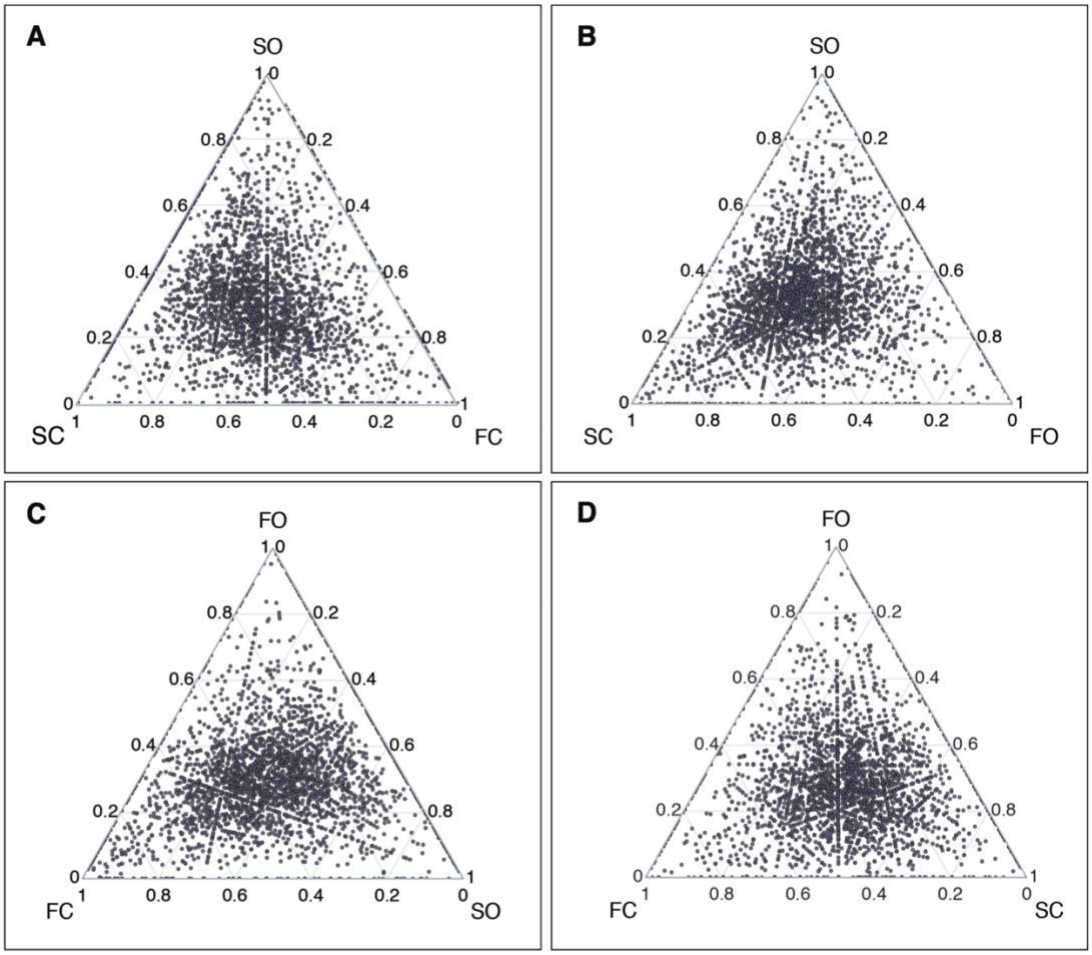


Figure S6: Ternary graphs show salt-specific, lineage-independent PFAMs. Related to Fig. 7. (A) Saltwater species' functional domains within genes, independent of lineage, were obtained using four groups with nine genomes each (see Dataset S12). Ternary analyses using normalized PFAM dCNs averaged across nine representatives from 4 groups: Saltwater Chlorophytes (SC), Freshwater Chlorophytes (FC), Saltwater Ochrophytes (SO), and Freshwater Ochrophytes (FO)). (A-D) Each panel shows PFAM-weighted scores as distributions with environment or lineage as variables. For example, these distributions show that SCs and SOs have more PFAM domains in common than FCs and FOs. More functional divergence is seen in freshwater species from the same lineages, while saltwater species maintain the conservation of functional domains (PFAMs). We identified 86 PFAMs from SOs (n=9) and SCs (n=9) that were not found in either FCs (n=9) or FOs (n=9).