VIRUS EVOLUTION

Virus Evolution, 2022, **8(2)**, 1–8
DOI: https://doi.org/10.1093/ve/veac102
Advance access publication date: 21 October 2022
Rapid Communication

Endogenous giant viruses contribute to intraspecies genomic variability in the model green alga *Chlamydomonas reinhardtii*

Mohammad Moniruzzaman,[1,2,*,†] Maria P. Erazo-Garcia,[1] and Frank O. Aylward[1,3,*,‡]

[1]Department of Biological Sciences, Virginia Tech, 926 West Campus Drive, Blacksburg, VA 24061, USA, [2]Department of Marine Biology and Ecology, Rosenstiel School of Marine, Atmospheric, and Earth Science, University of Miami, 4600 Rickenbacker Causeway, Miami, FL 33149, USA and [3]Center for Emerging, Zoonotic, and Arthropod-Borne Pathogens, Virginia Tech, 981 Kraft Dr, Room 2036, Blacksburg, VA 24060, USA
[†]https://orcid.org/0000-0001-9337-3874
[‡]https://orcid.org/0000-0002-1279-4050
*Corresponding authors: E-mail: monir@vt.edu; faylward@vt.edu

Abstract

Chlamydomonas reinhardtii is a unicellular eukaryotic alga that has been studied as a model organism for decades. Despite an extensive history as a model system, phylogenetic and genetic characteristics of viruses infecting this alga have remained elusive. We analyzed high-throughput genome sequence data of *C. reinhardtii* field isolates, and in six we discovered sequences belonging to endogenous giant viruses that reach up to several 100 kb in length. In addition, we have also discovered the entire genome of a closely related giant virus that is endogenized within the genome of *Chlamydomonas incerta*, the closest sequenced relative of *C. reinhardtii*. Endogenous giant viruses add hundreds of new gene families to the host strains, highlighting their contribution to the pangenome dynamics and interstrain genomic variability of *C. reinhardtii*. Our findings suggest that the endogenization of giant viruses may have important implications for structuring the population dynamics and ecology of protists in the environment.

Key words: endogenous virus; chlaymodonas reinhardtii; giant virus; imitervirales; genome evolution.

Introduction

Chlamydomonas reinhardtii is a widely studied unicellular green alga with a long history as a model organism that dates back to the 1950s (Sasso et al. 2018; Salomé and Merchant 2019). Despite this long history of research, no viruses that infect *C. reinhardtii* have yet been reported, and the diversity of viruses that infect this alga in nature remains unknown. In a recent study, we identified the widespread endogenization of 'giant viruses' in numerous green algae, which provides evidence of virus–host interactions that take place in nature (Moniruzzaman et al. 2020b). These Giant Endogenous Viral Elements (GEVEs) derive from giant viruses within the phylum *Nucleocytoviricota*, which possess large and complex genomes that can reach up to 2.5 Mb in length (Philippe et al. 2013). Giant viruses often encode complex functional repertoires in their genomes that include tRNA synthetases, rhodopsins, cytoskeletal components, histones, and proteins involved in glycolysis, the tricarboxylic acid cycle, and other aspects of central carbon metabolism (Aylward et al. 2021; Aylward and Moniruzzaman 2022a). Moreover, these viruses are widespread in the environment and infect a wide range of eukaryotic hosts, including green algae (Endo et al. 2020; Moniruzzaman et al. 2020a; Schulz et al. 2020; Meng et al. 2021; Ha, Moniruzzaman, and Aylward 2021). The complex genomes of giant viruses coupled with their collectively broad host range and ability to endogenize into the genomes of their hosts provides compelling evidence that they may be important vectors of gene transfer in eukaryotes.

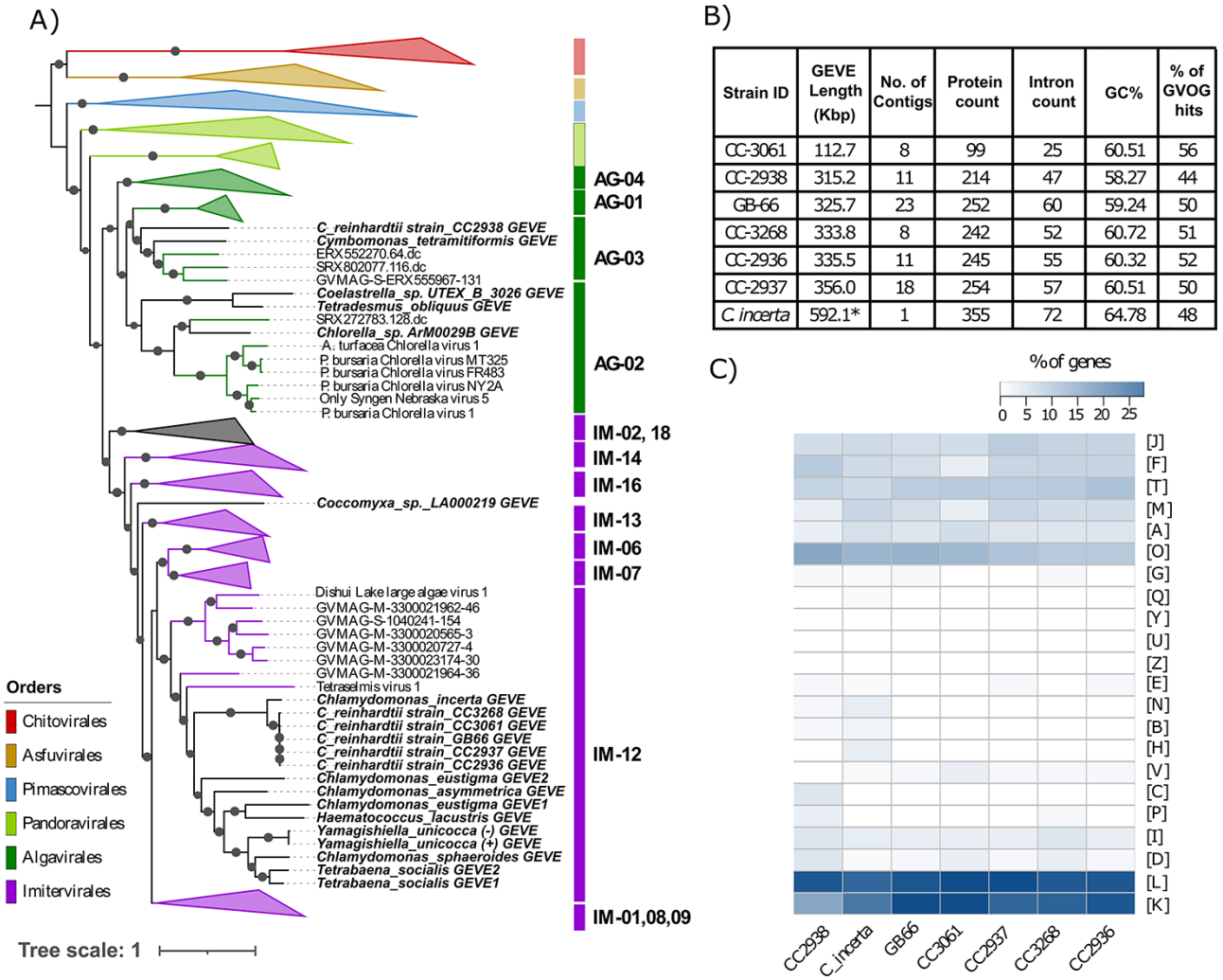
In our initial genomic survey of GEVEs we did not find evidence of endogenous giant viruses in the type strain *C. reinhardtii* (CC-503 cw92). However, several studies have recently reported high-throughput DNA sequence libraries of many *C. reinhardtii* field isolates as part of population genetics analysis. In this study we surveyed these strains for evidence of GEVEs. We report that near-complete genomes of giant viruses are present in several field isolates, and our results suggest that *C. reinhardtii* is a host to at least two distinct lineages of giant viruses. These are the first insights into the diversity and genomic complexity of viruses infecting *C. reinhardtii* in nature. We anticipate that this widely studied green alga will be a valuable model for future studies of virus–host interactions and the mechanistic aspects of giant virus endogenization.

Results

We analyzed publicly available high-throughput genome sequencing data for thirty-three wild strains of *C. reinhardtii*. These data were originally generated for population genomic studies of

diverse *C. reinhardtii* strains (Flowers et al. 2015; Craig et al. 2019; Hasan, Duggal, and Ness 2019). After *de novo* assembly and annotation (see Methods for details), we identified GEVEs in six of the wild strains (Fig. 1A, B). In five of these (CC-2936, 2937, 2938, 3268, and GB-66), the GEVEs range from 315 to 356 kb in size and harbored all but one Nucleocytoviricota hallmark genes, indicating that near-complete genomes of endogenous giant viruses have been retained in these strains (Fig. 1B, Dataset S1). In contrast, CC-3061 harbors a GEVE ~113 kb in size with five out of the ten hallmark genes, indicating that part of the GEVE was lost over the course of evolution (Supplementary Methods, Dataset S1). We also analyzed the assembled genome of *Chlamydomonas incerta*, a species phylogenetically closest to *C. reinhardtii*, for which a long-read assembled genome has been recently reported (Craig et al. 2021). This analysis revealed a GEVE ~475-kb long which is integrated within a single 592-kb contig of this alga (Fig. 1B). We developed PCR primers from the major capsid protein and DNA polymerase B genes of the GEVEs and used it on two GEVE-harboring strains (CC-2937 and CC-3268) and two strains where GEVE genomes were not detected (CC-3065 and CC-2931). A DNA fragment of the expected length was detected in the GEVE-harboring strains and were absent from others (Supplementary Fig. 1), confirming our bioinformatic predictions. Moreover, transmission electron microscopy confirmed that no visible free virions could be identified in the cultures of strains CC-2937 and CC-3268 (Supplementary Fig. 2).

Using a newly established taxonomy of Nucleocytoviricota (Aylward et al. 2021), we determined the phylogenetic position of the *C. reinhardtii* and *C. incerta* GEVEs and their relationships with other chlorophyte GEVEs that were recently reported



Strain ID	GEVE Length (Kbp)	No. of Contigs	Protein count	Intron count	GC%	% of GVOG hits
CC-3061	112.7	8	99	25	60.51	56
CC-2938	315.2	11	214	47	58.27	44
GB-66	325.7	23	252	60	59.24	50
CC-3268	333.8	8	242	52	60.72	51
CC-2936	335.5	11	245	55	60.32	52
CC-2937	356.0	18	254	57	60.51	50
C. incerta	592.1*	1	355	72	64.78	48

Figure 1. General features and phylogeny of the GEVEs. (A) Maximum likelihood phylogenetic tree of the GEVEs and representative members from diverse NCLDV families constructed from a concatenated alignment of seven NCLDV hallmark genes (see Methods). Individual families within each order are indicated with abbreviations (IM—Imitervirales and AG—Algavirales) followed by family numbers, as specified previously (Aylward et al. 2021). IDs of the GEVEs are indicated in bold-italic. (B) Basic statistics of the GEVEs present in various field strains of *C. reinhardtii* and the GEVE present in the *C. incerta* genome. (C) Functional potential of GEVEs as eggNOG categories. Categories of genes are normalized across all the NOG categories except S (function unknown) and R (general function prediction). Raw functional annotations are in Dataset S1. NOG categories: [J] Translation; [F] Nucleotide metabolism; [T] Signal Transduction; [M] Cell wall/membrane biogenesis; [A] RNA processing and modification; [O] Post-translational modification, protein turnover, and chamerone; [G] Carbohydrate metabolism; [Q] Secondary structure; [Y] Nuclear structure; [U] Intracellular trafficking and secretion; [Z] Cytoskeleton; [E] Amino acid metabolism; [N] Cell motility; [B] Chromatin structure and dynamics; [H] Coenzyme metabolism; [V] Defense mechanism; [C] Energy production and conversion; [P] Inorganic ion transport and metabolism; [I] Lipid metabolism; [D] Cell cycle control; [L] Replication and repair; [K] Transcription.

**Chlamydomonas incerta* GEVE length includes flanking eukaryotic regions.

(Moniruzzaman et al. 2020b) (Fig. 1A). Five of the strains harbored GEVEs that formed a cluster within the Imitervirales order, consistent with their high pairwise average amino acid identity (AAI). The GEVE in *C. incerta* was the closest phylogenetic relative of the Imitervirales GEVEs in *C. reinhardtii*, indicating that closely related giant viruses infect closely related *Chlamydomonas* species in nature. These GEVEs formed a sister clade with the GEVEs present in six other volvocine algae and belonged to the Imitervirales family 12 (Fig. 1A). In contrast to the GEVEs that could be classified as Imitervirales, the GEVE in CC-2938 strain belonged to the Algavirales (Fig. 1A), indicating that *C. reinhardtii* is infected by multiple phylogenetically distinct lineages of giant viruses in nature.

The coverage of the GEVE contigs was generally similar to those of the host *Chlamydomonas* contigs (see Supplementary Information), consistent with their presence as endogenous elements. The exception was the GEVE in CC-2938, in which two large contigs exhibited the same coverage as those of the host (~8 reads per kb per million), while the remaining GEVE contigs had coverage roughly twice that. This unusual pattern may be the product of large-scale duplication that recently took place in part of this GEVE. Indeed, recent work on other GEVEs in green algae found that large-scale duplications are common in GEVEs (Moniruzzaman et al. 2020b). This would explain why two large contigs with a summed length of 109 kb retain similar coverage compared to the host contigs, while the rest of the GEVE contigs have roughly double that coverage.

The %GC content of the *C. reinhardtii* GEVEs ranged from 58.27 per cent (CC-2938) to 60.72 per cent (CC-3268), which is similar to the overall genomic GC content of *C. reinhardtii* (64 per cent) (Merchant et al. 2007). Similarly, the GC content of the *C. incerta* GEVE was 64.8 per cent, resembling the overall GC content of the *C. incerta* genome (66 per cent) (Craig et al. 2021) (Fig. 1B). The GEVEs also contained several predicted spliceosomal introns, ranging from twenty-five (CC-3061) to seventy-two (*C. incerta*). Spliceosomal introns are rare in free Nucleocytoviricota but have been previously found in GEVEs present in other members of the Chlorophyta (Moniruzzaman et al. 2020b). It remains unclear if the relatively high %GC content and spliceosomal introns are features of the viruses themselves or if the evolution of these features evolved after endogenization. In addition, the GEVE in *C. incerta* was flanked by highly repetitive regions on both ends (Fig. 2A). The repetitive region at the 5′-end harbors several reverse transcriptases and transposases (Dataset S1). These regions also have higher intron density compared to the GEVE region itself and lower number of Giant Virus Orthologous Group (GVOG) hits, consistent with their eukaryotic provenance (Fig. 2A). This suggests that near-complete genomes of giant viruses may integrate within highly repetitive regions of eukaryotic genomes. It is possible that transposons may play a role in this process, potentially as loci for recombination that are found in both the algal and the viral genomes.

The GEVEs in *C. reinhardtii* encoded 99 (CC-3061) to 254 (CC-2937) predicted genes, while the *C. incerta* GEVE-encoded 355 predicted genes. Most of the genes were shared among the Imitervirales *C. reinhardtii* GEVEs, consistent with their high average AAI to each other (>98.5 per cent in all cases, Dataset S1). These GEVEs also shared a high number of orthogroups with the *C. incerta* GEVE (Dataset S1). In contrast, only a few orthogroups were shared between the Imitervirales and the Algavirales GEVEs, consistent with the large phylogenetic distance between these lineages. Between ~44 per cent and 55 per cent of the genes in the *C. reinhardtii* and *C. incerta* GEVEs have matches to GVOGs, confirming their viral provenance (Fig. 1B). In addition, different genes in these regions have best matches to giant viruses, bacteria, and eukaryotes, which is a common feature of Nucleocytoviricota members given the diverse phylogenetic origin of the genes in these viruses (Filée, Pouget, and Chandler 2008) (Fig. 2A). Based on the Cluster of Orthologous Group (COG) annotations, a high proportion of the GEVE genes are involved in transcription and DNA replication and repair; however, genes encoding translation, nucleotide metabolism and transport, signal transduction, and posttranslational modification were also present, consistent with the diverse functional potential encoded by numerous Nucleocytoviricota (Fig. 1C).

A previous study has shown that several field strains of *C. reinhardtii* harbor many genes that are absent in the reference genome (Flowers et al. 2015), which were possibly acquired from diverse sources. To quantify the amount of novel genetic material contributed by giant viruses to *C. reinhardtii*, we estimated the number of unique gene families in the analyzed *C. reinhardtii* field strains that are absent in the reference strain CC-503. On average ~1.78 per cent of the genes in the field strains were unique compared to the reference strain (Fig. 2B). Moreover, the GEVE-harboring field strains have significantly enriched in novel genes compared to those without GEVEs (two-sided Mann–Whitney U-test P-value < 0.05, Fig. 2B). These results suggest that the endogenization of giant viruses is an important contributor to interstrain genomic variability in *C. reinhardtii*. Recent studies have highlighted the importance of horizontal gene transfer in structuring the pangenome of diverse eukaryotes (Fan et al. 2020; Sibbald et al. 2020), and genes originating from endogenous Nucleocytoviricota were found to shape the genomes of many algal lineages, including members of the Chlorophyta (Moniruzzaman et al. 2020b; Nelson et al. 2021). Compared to the GEVE-free strains, GEVE-containing strains harbored a significantly higher proportion of genes from two COG categories including Transcription and Replication and Repair (two-sided Mann–Whitney U-test P-value < 0.05) (Fig. 2B). Altogether, these GEVEs contributed many genes with known functions, including glycosyltransferases, proteins involved in DNA repair, oxidative stress, and heat shock regulation (Dataset S1).

A recent comparative genomic analysis of *C. reinhardtii* analyzed the population structure of this alga and concluded that isolates from North America belong to two primary populations (NA1 and NA2) (Craig et al. 2019). Interestingly, we found Imitervirales GEVEs in both NA1 and NA2 populations, and in both cases these populations include strains for which GEVEs could not be detected. Indeed, strains CC-2931, CC-2932, and CC-3268 were all isolated from the same garden in NC, yet a GEVE could only be detected in CC-3268. This patchwork distribution of the Imitervirales GEVEs within *C. reinhardtii* populations suggests that they are the product of independent endogenization events rather than a single event in their shared evolutionary history. Moreover, the Imitervirales GEVEs we identified here fall within the same clade as most of the GEVEs we previously identified in other green algae. The prevalence of GEVEs within a particular lineage, together with their patchwork distribution across *C. reinhardtii* strains in the same population, suggests that GEVEs are the product of an active endogenization mechanism that takes place over short timescales rather than 'accidental' endogenization that may result from the illegitimate recombination that occurs during infection.

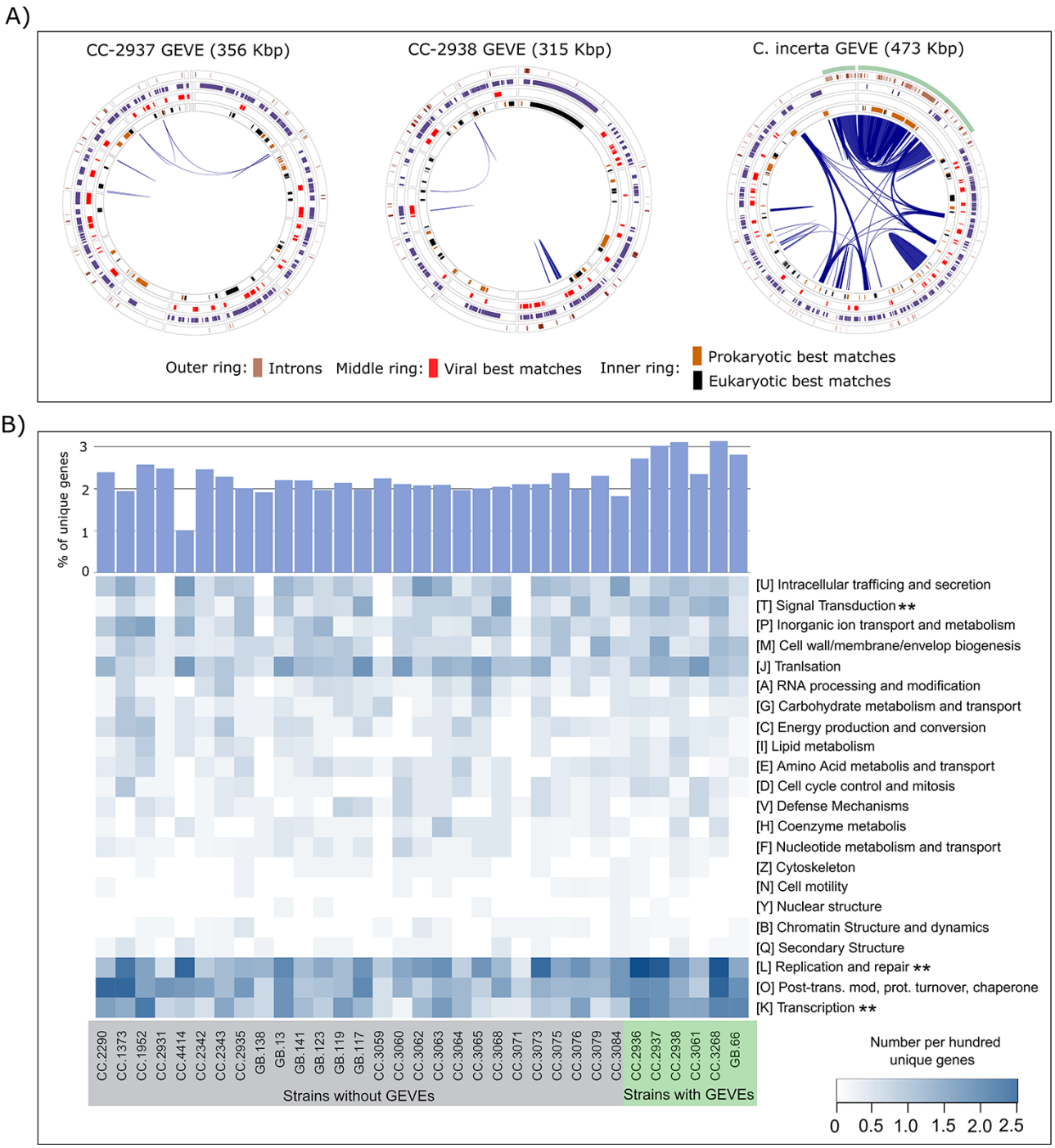


Figure 2. GEVE genomic and functional characteristics. (A) Circular plots of two representative GEVEs in *C. reinhardtii* and the GEVE present in *C. incerta*. For *C. reinhardtii* one representative Imitervirales GEVE (CC-2937) and the Algavirales GEVE (CC-2938) are shown. Circle plots show GVOG HMM hits, spliceosomal introns, and the best LAST hit matches (see Supplementary Methods). Internal blue links delineate the duplicated regions. The eukaryotic regions flanking the *C. incerta* GEVE are delineated with light blue stripes. (B) Unique genes in the field strains of *C. reinhardtii* compared to the reference strain CC-503. The heat map represents the percentage of unique genes that can be classified in different eggNOG categories (except category [R]—general function prediction and [S]—function unknown). Categories marked with '**' are significantly overrepresented in the GEVE-containing strains compared to those without GEVEs (Mann–Whitney U-test, $P < 0.05$). The bar plot on top of the heat map represents the percentage of unique genes in each strain. GEVE-containing strains have significantly higher percentages of unique genes compared to the strains without GEVEs.

Discussion

While much work remains to elucidate the role of GEVEs in shaping the ecological and evolutionary dynamics of *C. reinhardtii*, several possibilities remain open. Some genes contributed by the GEVEs could be potentially co-opted by the host, leading to changes in certain phenotypes compared to closely related strains without GEVEs. Strain-specific endogenization can also potentially lead to intraspecific variations in chromosome structure, partly

mediated by the GEVE-encoded mobile elements (Filée 2018). Finally, it is also possible that some of these GEVE loci can produce small interfering RNAs (siRNAs) that might participate in antiviral defense, and similar phenomena have been suggested for the virus-like loci in the genome of moss (*Physcomitrella patens*) (Lang et al. 2018). Recent studies on the large-scale endogenization of giant viruses into diverse green algal genomes suggest that interactions between giant viruses and their algal hosts frequently shape eukaryotic genome evolution (Moniruzzaman et al. 2020b;) and lead to the introduction of large quantities of novel genetic material. Our results indicate that these endogenization events can lead to genomic variability not only between algal species, but also between strains within the same population. Results reported in this study advance our understanding of how giant viruses shape the genome evolution of their hosts, while also expanding the scope of *C. reinhardtii* as a model organism to study the evolutionary fate and consequences of giant virus endogenization.

Methods

Raw sequence data and genome assembly

We investigated paired-end Illumina sequence data from thirty-three wild strains that were analyzed in three different studies (Flowers et al. 2015; Craig et al. 2019; Hasan, Duggal, and Ness 2019). Illumina sequence read libraries were downloaded from NCBI Sequence Read Archive (SRA) (see Dataset S1). Data from twenty-seven libraries were assembled using SPAdes v3.13.1 (Prjibelski et al. 2020) (parameters: –meta). For six of the libraries (CC-3060, CC-3062, CC-3063, CC-3064, CC-3065, and CC-3073), SPAdes assembler failed as it required more memory than was available on our computing nodes. We assembled these libraries using MEGAHIT (Li et al. 2015) with default parameters following quality trimming using TrimGalore (https://github.com/FelixKrueger/TrimGalore) (parameters: –length 36, –stringency 1, and -q 5).

Curation of GEVE contigs

We identified the preliminary candidate viral contigs from each assembled *C. reinhardtii* strains and *C. incerta* assembled genome using ViralRecall v.2.0 (using the contig screening parameter '-c') (Aylward and Moniruzzaman 2022b). We identified Nucleocytoviricota hallmark genes in these contigs using a Python script that we previously developed (https://github.com/faylward/ncldv_markersearch). After identifying the NNucleocytoviricota hallmark genes in these contig sets, we performed preliminary phylogenies using the DNA polymerase gene, which revealed that the endogenous viruses in five of the *C. reinhardtii* strains are highly similar and belong to the Imitervirales, whereas one of these strains harbored endogenous giant virus from the Algavirales group. The *C. incerta* GEVE, which was found to be endogenized in its entirety on a large contig, was found to be a close relative of the endogenous Imitervirales members from *C. reinhardtii* strains (Fig. 1). The 5′ and 3′ flanking regions of the *C. incerta* GEVEs (∼95 kb and ∼22 kb, respectively) harbored features characteristic of eukaryotic genomes, specifically, large repetitive regions that have comparatively higher intron density and low number of GVOG hits. The 5′ region also harbored a KDZ transposase (Pfam: 18,758) and two copies of zinc-binding regions associated with reverse transcriptases (Pfam: 13,966) (Dataset S1). Based on this evidence, we defined the *C. incerta* GEVE to be ∼475 kb bordered by these two flanking eukaryotic regions.

After determining the phylogenetic provenance of the endogenous viruses in each of these strains, we screened all the contigs detected by ViralRecall v2.0 to remove the contigs that originated from the *C. reinhardtii* reference chromosomal regions. We aligned the contigs to the reference genome chromosomes using Minimap2 (Li 2018) and removed contigs that were >90 per cent similar to the reference chromosomes. Contigs that shared >50–90 per cent similarity to the reference genome were manually inspected and in all cases were found to encode repetitive protein domains of diverse functions. It is possible that these regions originated from the host genome through possible duplication in different strains, and we excluded these contigs from subsequent analyses.

Following these steps, using the remaining set of contigs we delineated the GEVEs in each of these strains harboring Imitervirales GEVEs. Given the phylogenetic proximity of the *C. incerta* GEVE and its contiguous assembly in one large contig, we used this GEVE as a guide to validate the *C. reinhardtii* Imitervirales GEVEs. We aligned these contigs against the *C. incerta* GEVE at amino acid level using the promer tool implemented in MUMmer package (Delcher et al. 2002) to assess the similarity of these contigs to *C. incerta* GEVE and determined all these contigs to be originating from the same viral genome based on their alignment to this GEVE (Supplementary Fig. 3). Given the fragmented nature of assembly of individual libraries, it was possible that some of the viral regions were missed by ViralRecall in one strain, but the same region was detected in a different strain if that region was assembled into a larger contig. Since the GEVEs in the Imitervirales family are highly similar, we cross-referenced these confirmed viral contigs between libraries to detect smaller contigs that were otherwise missed by ViralRecall in one library but were detected in another. These steps ensured the maximum recovery of the viral regions from each library and allowed for a better estimation of the GEVE size and comparative analysis between GEVEs.

To define the Algavirales GEVE in CC-2938, we performed hierarchical clustering of the tetranucleotide frequency of the final ViralRecall screened contigs along with the rest of the contigs from the same host strain (>5 kb long). This analysis was performed to ensure that the viral contigs cluster together and separately from the host contigs, which will be expected based on their distinct viral origin. The results confirmed a cluster of contigs to co-cluster distinctly from the host contigs (Supplementary Fig. 4), which was determined to be the Algavirales GEVE present in CC-2938.

Hybrid gene prediction

For predicting genes on the final set of GEVE contigs, a hybrid gene prediction approach was taken, based on an approach we developed previously (Moniruzzaman et al. 2020b). Specifically, we first predicted genes using WebAUGUSTUS (Hoff and Stanke 2013) (http://bioinf.uni-greifswald.de/webaugustus/) and the *C. reinhardtii* training model on the whole assembled genomes of the *C. reinhardtii* strains. For the GEVE-harboring strains, we also predicted genes using Prodigal v.2.6.3 (Hyatt et al. 2010), which is widely used to predict genes in both prokaryotes and diverse viruses, including NCLDVs. For the GEVE contigs, we retained all the gene and intron predictions by WebAUGUSTUS and also retained the Prodigal-predicted genes only if they did not overlap with the gene boundaries predicted by WebAUGUSTUS. This hybrid approach allowed us to leverage both prediction strategies, as we previously found that some viral genes can be missed by WebAUGUSTUS but were predicted by Prodigal in these regions (Moniruzzaman et al. 2020b).

Coverage analysis of GEVE and host contigs

If GEVE contigs were truly endogenous we would expect them to have similar coverage to that of the rest of the *C. reinhardtii* genome. To test this, we compared the coverage of the GEVE and

host contigs by mapping reads from each genome onto its assembly. We performed read mapping with CoverM (https://github.com/wwood/CoverM) with the parameter '–min-covered-fraction 50'. To ensure that host contigs belonged to *C. reinhardtii* chromosomes, we compared all contigs to the reference seventeen chromosomes of *C. reinhardtii* strain CC-502 cw92 mt+ with LAST (default parameters) and retained only contigs with an *e*-value match of 1e-100. For this analysis, we only considered host contigs >20 kb in length and GEVE contigs >10 kb in length. The results are provided in Supplementary Fig. 5.

Read mapping to confirm GEVE absence

In the strains in which we did not identify GEVEs we sought to confirm that their absence was real and not simply due to complications arising from *de novo* assembly. For this we mapped raw sequencing reads from all genomes against the set of GEVE contigs from CC-2938 and CC-2937. These two were chosen because CC-2938 is the sole Algavirales GEVE that we found, while CC-2937 was the Imitervirales GEVE with the largest assembly recovered. We mapped reads using CoverM (https://github.com/wwood/CoverM) with the parameter '–min-covered-fraction 50'. Using this approach, we confirmed the absence of GEVEs from all strains except CC-3059, where reads could be mapped to four of the fourteen reference contigs of the Imitervirales GEVE. This suggests that CC-3059 contains a partial GEVE that could not be resolved in the *de novo* assemblies, although in all other cases no GEVE contigs could be recovered.

Functional annotation

We predicted the function of the protein sequences in each of the GEVEs and the unique genes present in each field strain by searching the proteins against hidden Markov model (HMM) profiles from COG (Tatusov et al. 2000), Pfam v. 32 (Sara et al. 2019), eggNog v. 5.0 (Huerta-Cepas et al. 2019), eggNOG Viral (Huerta-Cepas et al. 2019), and VOG (vogdb.org) databases using 'hmmsearch' command implemented in HMMER v.3.21 (Eddy 2011) with an *e*-value threshold of <0.00001. The best hit for a protein was evaluated based on the highest bit score to a HMM profile.

Identification of unique genes in diverse field strains

For the identification of unique genes that are present in different field strains of *C. reinhardtii* but absent in the reference genome (CC-503), we first predicted genes in all these genomes using WebAUGUSTUS (Hoff and Stanke 2013) as described in the 'Hybrid gene prediction' section. Some strain assemblies contained contigs with coverage greater than twenty times the longest contigs in the assembly, and manual inspection revealed that they likely derived from bacterial contamination. To mitigate the impact of this on our unique gene estimates, for this analysis we did not consider contigs that had coverage greater than one standard deviation above the mean for a given assembly. For the remainder of the contigs, the predicted proteins from the field strains with and without GEVEs were searched against the reference CC-503 proteins (*C. reinhardtii* assembly version 5.5) using 'Blastp' (parameters: -max_hsps 1 and -max_target_seqs 1). To obtain a conservative estimate of the unique gene families in each field strain, only genes that had no homology at an *e*-value threshold of 0.001 to the reference proteome were considered. Although we predicted additional genes using Prodigal in the GEVE contigs for GEVE functional analysis and homology searches, for estimating unique genes, we excluded the Prodigal predicted proteins. This was done as we compared results across all field strains—since Prodigal prediction is only relevant for the GEVE contigs and cannot be included for the other contigs in the genome.

GVOG analysis

For identifying GEVE genes with similarities to diverse giant viruses, we used a curated GVOG database that we recently constructed from 1,380 quality-checked genomes that include 1,253 metagenome-assembled genomes and 127 complete genomes available in culture for Nucleocytoviricota members. GVOGs are publicly available at https://faylward.github.io/GVDB/ (Aylward et al. 2021). To evaluate hits to the GVOGs, we used 'hmmsearch' implemented in HMMER v.3.2.1 with an *e*-value threshold of <0.00001.

Duplication and synteny analysis

We compared synteny between different GEVE regions using the 'progressiveMauve' tool implemented in Mauve package (Darling 2004). For determining the similarity of the *C. reinhardtii* GEVE contigs to the *C. incerta* GEVE at the amino acid level, we used the 'promer' tool implemented in MUMmer (Delcher et al. 2002) with the '–maxmatch' option. We estimated the amount of repetitive regions within each GEVE using RECON 1.0.8 (Bao and Eddy 2002), with a nucleotide identity of >90 per cent.

AAI and orthogroup analysis

AAI between GEVE proteomes was calculated using a custom Python script (https://github.com/faylward/lastp_aai), which carries out pairwise LAST (v. 959) searches (parameter: -m 500) of protein sequences and calculates the average AAIs between all possible pairs of genomes (Kiełbasa et al. 2011). Orthogroups of proteins among the GEVEs were calculated using Proteinortho v.6.0.6 (Lechner et al. 2011) with default parameters (-identity = 25 and -cov = 50).

GEVE phylogenies

For the phylogenetic reconstruction of the GEVEs along with known NCLDVs, we used a subset of high-quality genomes recently curated to develop a phylogenomic framework of Nucleocytoviricota (Aylward et al. 2021). The GEVEs from this study and a previous study (Moniruzzaman et al. 2020b) were included. We used a concatenated alignment of a set of nine core genes as described previously to be ideal for the phylogenetic reconstruction of Nucleocytoviricota (Aylward et al. 2021). Alignments were generated using Clustal Omega (Sievers et al. 2011) and trimmed with trimAl (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009). The tree was constructed using IQ-TREE v1.6.9 (Nguyen et al. 2015), and 1,000 ultrafast bootstrap replicates were performed to assess the statistical support at the nodes (parameters: -wbt, -bb 1000 and -m LG+I+G4). The tree was visualized using iTOL (Letunic and Bork 2019).

Homology search

To identify the best match of the GEVE proteins in diverse domains of life and viruses, we compared the GEVE proteins against a database of NCBI RefSeq v. 99. To this end, we employed LASTAL v. 959 with the parameter '-m 5000' for the increased sensitivity of homology detection. Before evaluating the best hits, all the hits to Chlorophyta were removed, to avoid self-hits. Taxonomic profile of each best hit was determined by cross-referencing the hits to the NCBI Taxonomy database (Federhen 2012). For this, we used the Python API implemented in the ETE3 Toolkit (Huerta-Cepas, Serra, and Bork 2016).

Assessing the partial loss of GEVE in CC-3061

The total length of the final set of screened GEVE contigs in CC-3061 was ~115 kb long, which is much smaller than the other *C. reinhardtii* strains harboring GEVEs. This suggests that the GEVE in CC-3061 went through partial loss over the course of genome evolution. However, it is also possible that due to fragmented assemblies obtained from the raw data, some of the GEVE regions failed to assemble and hence were missed by our screening approach. If that is the case, we should still be able to find the reads corresponding to such small contigs. As the Imitervirales GEVEs in the *C. reinhardtii* strains are highly similar to each other, we mapped the raw reads from CC-3061 library to one of the near-complete GEVEs from strain CC-2937. We found that although several of the CC-2937 GEVE contigs had good coverage, other contigs had zero or near-zero coverage, indicating that reads originating from these regions are absent in the CC-3061 library (Coverage values available in Dataset S1). This analysis confirmed that the regions missing in the CC-3061 GEVE are due to partial loss and not an artifact of lower quality assembly or low sequencing depth.

PCR amplification of GEVE-specific marker genes in select strains

Chlamydomonas reinhardtii strains CC-2931, CC-2937, CC-3065, and CC-3268 were purchased from the Chlamydomonas Resource Center (Minneapolis, MN, USA) and cultured on agar plates with tris-acetate phosphate (TAP) medium at 25°C, under a 16:8 h light:dark photoperiod. Total genomic DNA was isolated from colonies using the protocol described by Nouemssi et al. (2020). The ITS1-5.8S-ITS2 region was PCR-amplified as a positive control using the primers Fw_ITS1 and Rv_ITS4 described by White et al. (1990). Evidence of GEVEs was confirmed through PCR using viral polymerase B and capsid-specific primers: GEVE_PolB_Fw (5′–AACTCCCTTTACGGCCAGAT–3′), GEVE_PolB_Rv (5′–CACGCAG TGTCCGAGTAGAA–3′), GEVE_cap_Fw (5′—GACGGCTACGACCGT ATGAT—3′), and GEVE_cap_Rv (5′–CATCACCCAAATCAGCTC CT–3′), which were designed to amplify 248- and 450-bp fragments, respectively. PCR amplification was performed in a final volume of 25 μl containing 1× of ReadyMix™ Taq PCR Reaction Mix (Sigma-Aldrich, St. Louis, MO, USA), 0.3 μM of each primer, and 2 μl of DNA template. For all reactions, the PCR program consisted of an initial denaturation step at 95°C for 5 min, followed by thirty-five cycles of 95°C for 30 s, annealing at 55°C for 30 s and 72°C for 1 min, and followed by a final extension step at 72°C for 5 min. PCR products were separated by gel electrophoresis on a 1.5 per cent agarose gel in Tris-Acetate-EDTA (TAE) buffer.

Transmission electron microscopy

For electron microscopy, GEVE strains CC-2937 and CC-3268 were grown in liquid TAP media for 3 days and cells were fixed overnight at 4°C with 2 per cent glutaraldehyde prepared in 0.1 M Phosphate Buffer Saline (PBS) buffer. After fixation, cells were centrifuged and rinsed three times with 0.1 M PBS buffer to remove the fixative and the supernatant was replaced with 500 μl of low melting agarose (4 per cent). Agarose-embedded samples were postfixed for 1 h in 2 per cent OsO_4 and rinsed three times with deionized water (Saikachi, Sugasawara, and Suzuki 2021). Dehydration was achieved by submerging the samples in ethanol solutions of increasing concentration and further transferring to a 1:1 propylene oxide:ethanol solution for 30 min and then to pure propylene oxide for 30 min (Graham and Orenstein 2007). Samples were infiltrated in 1:1 propylene oxide/Poly/Bed 812 resin mixture for 120 min and 24 h, placed into pure Poly/Bed 812 for 24 h, and finally into silicone molds that were filled with fresh Poly/Bed 812. The resin was polymerized at 60°C for 72 h (Graham and Orenstein 2007). Ultra-thin sections of 100 nm were cut with a Leica EM UC7 ultramicrotome and an ultra 45° DiATOME® diamond knife, which were placed on 200-mesh copper grids (Rey, Faruqui, and Ryadnov 2021). Samples were stained with uranyl acetate 3 per cent and Reynold's lead citrate solution for 10 min and 4 min, respectively (Graham and Orenstein 2007; Rey, Faruqui, and Ryadnov 2021). Images were obtained with a JEOL JEM-1400 series 80 kV Transmission Electron Microscope.

Data and code availability

Dataset S1 contains information regarding the raw data source, GEVE functional annotations, hallmark gene distribution in each GEVE, and coverage information of the partial GEVE in CC-3061. All the GEVE fasta files, unique gene fasta in each of the strains and their annotations, and concatenated alignment file used to build the phylogenetic tree in Fig. 1 are available in Zenodo: https://zenodo.org/record/4958215. Code and instructions for ViralRecall v2.0 and NCLDV marker search scripts are available at: github.com/faylward.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Acknowledgements

We acknowledge the use of the Virginia Tech Advanced Research Computing Center for bioinformatic analyses performed in this study. We thank Nathalie del Pilar Becerra Mora and members of the Virginia Tech Electron Microscopy Laboratory for assistance visualizing the TEM samples.

Funding

National Science Foundation (IIBR-1918271 to F.O.A.); Simons Early Career Award in Marine Microbial Ecology and Evolution (to F.O.A.), National Institutes of Health (1R35GM147290-01 to F.O.A).

Conflict of interest: The authors declare no conflict of interest relevant to the content of the manuscript.

References

Aylward, F. O. et al. (2021) 'A Phylogenomic Framework for Charting the Diversity and E6volution of Giant Viruses', *PLoS Biology*, 19: e3001430.

Aylward, F. O., and Moniruzzaman, M. (2022a) 'Viral Complexity', *Biomolecules*, 12: 1061.

——— (2022b) 'ViralRecall - A flexible command-line tool for the detection of giant virus signatures in 'Omic data'', *Viruses*, 13: 150.

Bao, Z., and Eddy, S. R. (2002) 'Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes', *Genome Research*, 12: 1269–76.

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: 1972–3.

Craig, R. J. et al. (2019) 'Patterns of Population Structure and Complex Haplotype Sharing among Field Isolates of the Green Alga *Chlamydomonas reinhardtii*', *Molecular Ecology*, 28: 3977–93.

——— et al. (2021) 'Comparative Genomics of *Chlamydomonas*', *The Plant Cell*, 33: 1016–41.

Darling, A. C. E. (2004) 'Mauve: Multiple Alignment of Conserved Genomic Sequence with Rearrangements', *Genome Research*, 14: 1394–403.

Delcher, A. L. et al. (2002) 'Fast Algorithms for Large-Scale Genome Alignment and Comparison', *Nucleic Acids Research*, 30: 2478–83.

Eddy, S. R. (2011) 'Accelerated Profile HMM Searches', *PLoS Computational Biology*, 7: e1002195.

Endo, H. et al. (2020) 'Biogeography of Marine Giant Viruses Reveals Their Interplay with Eukaryotes and Ecological Functions', *Nature Ecology & Evolution*, 4: 1639–49.

Fan, X. et al. (2020) 'Phytoplankton Pangenome Reveals Extensive Prokaryotic Horizontal Gene Transfer of Diverse Functions', *Science Advances*, 6: eaba0111.

Federhen, S. (2012) 'The NCBI Taxonomy Database', *Nucleic Acids Research*, 40: D136–43.

Filée, J. (2018) 'Giant Viruses and Their Mobile Genetic Elements: The Molecular Symbiosis Hypothesis', *Current Opinion in Virology*, 33: 81–8.

Filée, J., Pouget, N., and Chandler, M. (2008) 'Phylogenetic Evidence for Extensive Lateral Acquisition of Cellular Genes by Nucleocytoplasmic Large DNA Viruses', *BMC Evolutionary Biology*, 8: 320.

Flowers, J. M. et al. (2015) 'Whole-Genome Resequencing Reveals Extensive Natural Variation in the Model Green Alga *Chlamydomonas reinhardtii*', *The Plant Cell*, 27: 2353–69.

Graham, L., and Orenstein, J. M. (2007) 'Processing Tissue and Cells for Transmission Electron Microscopy in Diagnostic Pathology and Research', *Nature Protocols*, 2: 2439–50.

Ha, A. D., Moniruzzaman, M., and Aylward, F. O. (2021) 'High Transcriptional Activity and Diverse Functional Repertoires of Hundreds of Giant Viruses in a Coastal Marine System', *mSystems*, 6: e00293–21.

Hasan, A. R., Duggal, J. K., and Ness, R. W. (2019) 'Consequences of Recombination for the Evolution of the Mating Type Locus in *Chlamydomonas reinhardtii*', *The New Phytologist*, 224: 1339–48.

Hoff, K. J., and Stanke, M. (2013) 'WebAUGUSTUS—A Web Service for Training AUGUSTUS and Predicting Genes in Eukaryotes', *Nucleic Acids Research*, 41: W123–28.

Huerta-Cepas, J. et al. (2019) 'eggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses', *Nucleic Acids Research*, 47: D309–14.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016) 'ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data', *Molecular Biology and Evolution*, 33: 1635–8.

Hyatt, D. et al. (2010) 'Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification', *BMC Bioinformatics*, 11: 119.

Kiełbasa, S. M. et al. (2011) 'Adaptive Seeds Tame Genomic Sequence Comparison', *Genome Research*, 21: 487–93.

Lang, D. et al. (2018) 'The *Physcomitrella patens* Chromosome-Scale Assembly Reveals Moss Genome Structure and Evolution', *The Plant Journal: For Cell and Molecular Biology*, 93: 515–33.

Lechner, M. et al. (2011) 'Proteinortho: Detection of (Co-)Orthologs in Large-Scale Analysis', *BMC Bioinformatics*, 12: 124.

Letunic, I., and Bork, P. (2019) 'Interactive Tree of Life (iTOL) V4: Recent Updates and New Developments', *Nucleic Acids Research*, 47: W256–59.

Li, H. (2018) 'Minimap2: Pairwise Alignment for Nucleotide Sequences', *Bioinformatics*, 34: 3094–100.

Li, D. et al. (2015) 'MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph', *Bioinformatics*, 31: 1674–6.

Meng, L. et al. (2021) 'Quantitative Assessment of Nucleocytoplasmic Large DNA Virus and Host Interactions Predicted by Co-Occurrence Analyses', *mSphere*, 6: e01298–20.

Merchant, S. S. et al. (2007) 'The *Chlamydomonas* Genome Reveals the Evolution of Key Animal and Plant Functions', *Science*, 318: 245–50.

Moniruzzaman, M. et al. (2020a) 'Dynamic Genome Evolution and Complex Virocell Metabolism of Globally-Distributed Giant Viruses', *Nature Communications*, 11: 1710.

——— et al. (2020b) 'Widespread Endogenization of Giant Viruses Shapes Genomes of Green Algae', *Nature*, 588: 141–5.

Nelson, D. R. et al. (2021) 'Large-Scale Genome Sequencing Reveals the Driving Forces of Viruses in Microalgal Evolution', *Cell Host & Microbe*, 29: 250–66.e8.

Nguyen, L.-T. et al. (2015) 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies', *Molecular Biology and Evolution*, 32: 268–74.

Nouemssi, S. B. et al. (2020) 'Rapid and Efficient Colony-PCR for High Throughput Screening of Genetically Transformed *Chlamydomonas reinhardtii*', *Life*, 10: 186.

Philippe, N. et al. (2013) 'Pandoraviruses: Amoeba Viruses with Genomes up to 2.5 Mb Reaching That of Parasitic Eukaryotes', *Science*, 341: 281–6.

Prjibelski, A. et al. (2020) 'Using SPAdes De Novo Assembler', *Current Protocols*, 70: e102.

Rey, S., Faruqui, N., and Ryadnov, M. G. (2021) 'Ultramicrotomy Analysis of Peptide-Treated Cells', *Methods in Molecular Biology*, 2208: 255–64.

Saikachi, A., Sugasawara, K., and Suzuki, T. (2021) 'Analyses of the Effect of Peptidoglycan on Photocatalytic Bactericidal Activity Using Different Growth Phases Cells of Gram-Positive Bacterium and Spheroplast Cells of Gram-Negative Bacterium', *Catalysts*, 11: 147.

Salomé, P. A., and Merchant, S. S. (2019) 'A Series of Fortunate Events: Introducing *Chlamydomonas* as a Reference Organism', *The Plant Cell*, 31: 1682–707.

Sara, E.-G. et al. (2019) 'The Pfam Protein Families Database in 2019', *Nucleic Acids Research*, 47: D427–32.

Sasso, S. et al. (2018) 'From Molecular Manipulation of Domesticated to Survival in Nature', *eLife*, 7: e39233.

Schulz, F. et al. (2020) 'Giant virus diversity and host interactions through global metagenomics', *Nature*, 578: 432–6.

Sibbald, S. J. et al. (2020) 'Lateral Gene Transfer Mechanisms and Pangenomes in Eukaryotes', *Trends in parasitology*, 36: 927–41.

Sievers, F. et al. (2011) 'Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega', *Molecular systems biology*, 7: 539.

Tatusov, R. L. et al. (2000) 'The COG database: a tool for genome-scale analysis of protein functions and evolution', *Nucleic acids research*, 28: 33–6.

White, T. J., Bruns, T., Lee, S. J., and Taylor, J. (1990) 'Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics', *PCR protocols: a guide to methods and applications*, 18: 315–22.