HHS Public Access

Author manuscript

Science. Author manuscript; available in PMC 2025 June 09.

Published in final edited form as:

Science. 2025 May 15; 388(6748): eads6303. doi:10.1126/science.ads6303.

Cryptic infection of a giant virus in a unicellular green alga

Maria P. Erazo-Garcia[1,†], **Uri Sheyn**[1,†], **Zachary K. Barth**[1], **Rory J. Craig**[2], **Petronella Wessman**[3], **Abdeali M. Jivaji**[1], **W. Keith Ray**[4], **Maria Svensson-Coelho**[3,‡], **Charlie K. Cornwallis**[3], **Karin Rengefors**[3], **Corina P. D. Brussaard**[3,5,6,*], **Mohammad Moniruzzaman**[7,*], **Frank O. Aylward**[1,8,*]

[1]Department of Biological Sciences, Virginia Tech, Blacksburg, VA, USA.

[2]Department of Algal Development and Evolution, Max Planck Institute for Biology Tübingen, Tübingen, Germany.

[3]Department of Biology, Lund University, Lund, Sweden.

[4]Mass Spectrometry Incubator, Fralin Life Sciences Institute, Virginia Tech, Blacksburg, VA, USA.

[5]Department of Marine Microbiology and Biogeochemistry, Royal Netherlands Institute for Sea Research (NIOZ), Texel, Netherlands.

[6]Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, Amsterdam, Netherlands.

[7]Department of Marine Biology and Ecology, University of Miami, Coral Gables, FL, USA.

[8]Center for Emerging, Zoonotic, and Arthropod-Borne Infectious Disease, Virginia Tech, Blacksburg, VA, USA.

Abstract

Latency is a common strategy in a wide range of viral lineages, but its prevalence in giant viruses remains unknown. In this work, we describe a 617–kilo–base pairs integrated giant viral element in the model green alga *Chlamydomonas reinhardtii*. We resolved the integrated viral genome using long-read sequencing, identified a putative polintovirus-like integrase, and show that viral particles accumulate primarily during the stationary growth phase. A diverse array of viral-encoded selfish genetic elements is expressed during viral activity, including several Fanzor nuclease–encoding transposable elements. In addition, we show that field isolates of *Chlamydomonas* spp. harbor signatures of endogenous giant viruses related to the *C. reinhardtii* virus that exhibit similar infection dynamics, suggesting that giant virus latency is prevalent in

*Corresponding author. corina.brussaard@nioz.nl (C.P.D.B.); m.monir@miami.edu (M.M.); faylward@vt.edu (F.O.A.).
†These authors contributed equally to this work.
‡Present address: Division of Molecular Biology, Department of Laboratory Medicine, Ryhov County Hospital, Jönköping, Sweden.

Author contributions: Conceptualization: K.R., C.P.D.B., M.M., F.O.A.; Methodology: M.P.E.-G., U.S., Z.K.B., R.J.C., P.W., A.M.J., W.K.R., M.S.-C., C.K.C., K.R., C.P.D.B.; Investigation: M.P.E.-G., U.S., Z.K.B., R.J.C., P.W., A.M.J., W.K.R., M.S.-C.; Visualization: M.P.E.-G., U.S., Z.K.B., R.J.C., P.W., F.O.A.; Supervision: K.R., C.P.D.B., F.O.A.; Provision of reagents: K.R., C.P.D.B., F.O.A.; Funding acquisition: K.R., C.P.D.B., F.O.A.; Writing – original draft: M.P.E.-G., U.S., Z.K.B., F.O.A.; Writing – review & editing: M.P.E.-G., U.S., Z.K.B., R.J.C., P.W., A.M.J., W.K.R., M.S.-C., C.K.C., K.R., C.P.D.B., M.M., F.O.A.

Competing interests: The authors declare that they have no competing interests.

natural host communities. Our work describes an unusually large temperate virus of a unicellular eukaryote, substantially expanding the scope of cryptic viral infections in the virosphere.

Abstract

INTRODUCTION: The recent discovery of giant endogenous viral elements (GEVEs) across a wide range of protist genomes presents an opportunity to investigate a possible latent viral infection strategy within giant viruses. Although these elements can be prominent features of eukaryotic genomes, GEVEs frequently exhibit clear signs of genomic erosion, including duplications, methylation, and intron invasion, raising questions about their viability. *Chlamydomonas reinhardtii* is a unicellular green alga long recognized as a model organism, but its potential interaction with viruses in the environment has remained elusive. The recent observation that some field isolates of *C. reinhardtii* harbor signatures of GEVEs suggests that this alga could also serve as a valuable model for investigating the dynamics of endogenous giant viruses in nature.

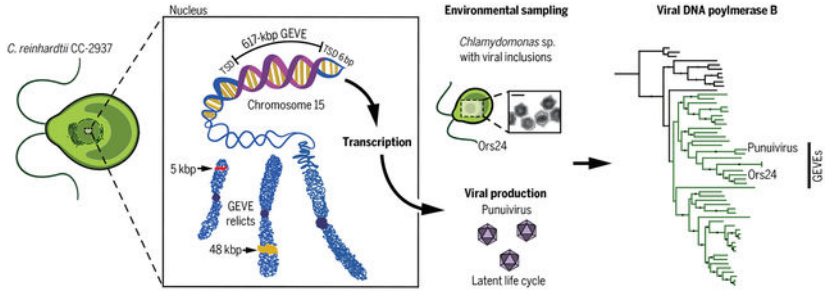
RATIONALE: Although latency is a common strategy in a wide range of viral lineages, it remains unknown whether GEVEs originate from a viral life cycle involving a latent phase or whether they are the result of another route of gene transfer. Latency was proposed as a potential strategy used by viruses of eukaryotic algae by numerous studies dating to the 1970s that described the formation of large icosahedral particles in otherwise healthy cultures. Methodological challenges, particularly in isolating and tracing the origin of these large particles, have hindered efforts to determine whether these particles result from active latent giant viruses or other factors such as persistent infections or contaminated cultures.

RESULTS: Using long-read sequencing, we resolved a 617–kilo–base pair (kbp) GEVE located on chromosome 15 of *C. reinhardtii* strain CC-2937. The GEVE is flanked by 6-bp target site duplications (TSDs), which are signatures of distinct families of DD(E/D) integrase enzymes. We found a candidate integrase encoded by the GEVE, which is related to polintovirus (phylum *Preplasmiviricota*) integrases, suggesting that interactions with hyperparasites could drive the evolution of cryptic infection strategies in giant viruses. We demonstrate that the GEVE is transcriptionally active and produces viral particles that accumulate primarily during the stationary growth phase of liquid cultures that exhibit no evident signatures of infection. The GEVE encodes several selfish genetic elements, including several transposases that encode Fanzor nucleases, which are active during infection and show signatures of recent mobility. In addition, we provide evidence that *Chlamydomonas* spp. isolates from freshwater environments harbor giant viruses closely related to the *C. reinhardtii* CC-2937 GEVE, suggesting that cryptic infections involving genome integration are prevalent among large DNA viruses of green algae.

CONCLUSION: Our study describes an unusually large temperate virus that infects the model green alga *C. reinhardtii*. Our evidence indicates that the GEVE can reactivate and produce viral particles, although many aspects of the infection program, including the potential molecular signals that control reactivation, remain unclear. Additionally, the presence of several viral-encoded selfish genetic elements suggests that giant viruses may serve as vectors of selfish DNA in eukaryotes. Last, our discovery of signatures of giant viruses related to the GEVE in field isolates of *Chlamydomonas* spp. points to cryptic infections as potentially widespread among natural algal populations. Our findings broaden the scope of cryptic infections in the virosphere

and emphasize genome integration as a potentially important component of the infection cycle of many giant viruses.

Graphical Abstract



An active GEVE in *C. reinhardtii* produces virions and establishes a cryptic infection. A 617-kbp GEVE and two relicts were resolved in *C. reinhardtii* CC-2937 using long-read sequencing. The GEVE is transcriptionally active in stationary-phase cultures, producing particles of the virus “punuivirus,” named after the Inca deity of untroubled sleep, Puñuy. Related giant viruses are associated with natural *Chlamydomonas* isolates and exhibit comparable infection dynamics, which suggests that cryptic infection strategies are common among large protist viruses. Ors24 is a *Chlamydomonas* isolate from Örsjön, a lake located in southern Sweden. Scale bar is 200 nm.

Endogenous viral elements (EVEs) are prevalent features in eukaryotic genomes that play key roles in regulation, antiviral defense, and other cellular processes (1–3). Once linked primarily to integrated retroviruses, it is now recognized that EVEs are derived from a wide range of viral lineages, including single-stranded DNA and double-stranded DNA (ds-DNA) viruses (4–7). To date, the largest EVEs discovered are derived from large DNA viruses in the phylum *Nucleocytoviricota*, often called “giant viruses” because of their large genomes and virions. Large EVEs derived from nucleocytoviruses, called giant endogenous viral elements (GEVEs), are ubiquitous in green algae, brown algae, various fungi, a wide range of other protists, and even some plants (8–12). GEVEs are prominent features that can contribute large quantities of viral genes to the genomes of their hosts; for example, the genome of the green alga *Tetrabaena socialis* includes two GEVEs totaling >3 mega–base pairs (Mbp), whereas the fungus *Rhizophagus irregularis* has the longest contiguously resolved GEVE at 1.5 Mbp (8, 9).

Despite the large contribution of GEVEs to many eukaryotic genomes, it remains unknown whether these elements are derived from the active integration of nucleocytoviruses as part of their infection cycle or merely accidental integrations that occur during stalled infections. Interestingly, studies dating to the 1970s have observed the formation of large icosahedral particles from otherwise healthy cultures of protists, but it has remained unclear whether this can be attributed to the activity of latent viruses or other factors such as persistent infection of a subpopulation (13, 14). The best-studied example of a putatively active GEVE to date is in the multicellular brown alga *Ectocarpus siliculosus*, where a 330-kbp endogenous nucleocytovirus has been linked to virus-like particle (VLP) formation in reproductive tissues (15–17), but even here, the specific activity of the GEVE remains unclear. Indeed, the viability of many GEVEs is questionable, and many appear to be silenced through

methylation and chromatin remodeling, whereas others have undergone large-scale erosion and genomic rear-rangements that likely led to their inactivation (9, 12, 18).

Given the recent widespread discovery of GEVEs in protist genomes (8, 12, 19), it is important to determine whether these elements arise from a viral infection strategy involving latency and genome integration. To elucidate the activity of GEVEs and their potential for virion production, we studied the model green alga *Chlamydomonas reinhardtii*, which has been used for decades in detailed analyses of cilia, photosynthesis, and other aspects of eukaryotic biology (20, 21). The recent observation that some field isolates of *C. reinhardtii* harbor signatures of GEVEs suggests that this alga may also be a useful system for in-depth analysis of endogenous giant viruses (22). We used a combination of long-read sequencing, transcriptomics, proteomics, and additional surveys of field isolates to examine the activity of GEVEs in *C. reinhardtii* and their potential role as part of the latent infection cycle of giant viruses. Our work describes a large temperate virus that infects the model green alga *C. reinhartdii*, thereby highlighting the importance of cryptic infections to protists in the biosphere.

Results and discussion

Long-read sequencing resolves a contiguous GEVE

We used long-read Oxford Nanopore sequencing to obtain a high-quality draft assembly of *C. reinhardtii* strain CC-2937. This strain was selected because a previous study using short-read sequencing found that it contained the most GEVE signatures among all *C. reinhardtii* strains surveyed (22). We recovered a high-quality assembly with an estimated genome size consistent with the latest *C. reinhardtii* reference genome (see Materials and methods) (23). We screened the polished contigs of the assembly for nucleocytovirus signatures using ViralRecall (24) and recovered a 617-kbp GEVE flanked by eukaryotic sequences within a 2.8-Mbp contig (Fig. 1A). Other than the GEVE region, the contig corresponds to *C. reinhardtii* chromosome 15 in the reference genome. The GEVE was contiguous and delimited by terminal inverted repeats (TIRs) 10.8 and 14.8 kbp in length, with the difference attributable to a variable-length satellite array present in the TIRs. This GEVE is almost twice as long as previously estimated using short-read sequencing (22), underscoring the importance of long-read sequencing to accurately delineate large EVEs.

We predicted 579 open reading frames (ORFs) (data S1) from the GEVE that include a complete set of *Nucleocytoviricota* hallmark genes, such as family B DNA polymerase (PolB), two double–jelly roll major capsid proteins (MCPs), multisubunit RNA polymerase homologs, an A32 packaging adenosine triphosphatase (ATPase), and a VLTF3 transcription factor (25). We did not observe recently duplicated marker genes, in contrast to GEVEs in other green algal genomes that contained large-scale genome duplications (8). The percentage GC content of the GEVE was only slightly less than that of the flanking regions (60.6 versus 62.8%) and also slightly below the genome-wide GC content reported for *C. reinhardtii* (64%) (23, 26). Many GEVEs show a clear deviation in nucleotide composition compared with the genomes of their hosts (8), but here we found no detectable discrepancy (fig. S1).

To determine the integration site of the GEVE, we compared our CC-2937 assembly with the genomes of the reference strain CC-4532 and two field isolates, CC-1952 and CC-2931 (27). We mapped the TIRs to an intergenic region downstream of the Cre15.g635700 gene. This region exhibits substantial structural variation, and the sequence flanking the TIRs corresponds to an ~9-kbp interspersed repetitive element that is absent from the other strains at this locus. Independent copies of this repeat are found in CC-2937 at two other regions on chromosomes 4 (contig_813) and 3 (contig_174) and at single loci in CC-1952 (chromosome 9) and CC-2931 (chromosome 10). By aligning these five repeat copies, we determined the exact insertion site and TIR boundaries of the GEVE (Fig. 1B). The TIRs feature the terminal motif “ACC-GGT” and are flanked by a 6-bp target site duplication (TSD).

We identified sequences homologous to the GEVE TIRs at two other regions in the CC-2937 genome, on contig_437 (chromosome 16) and contig_337 (chromosome 7). Comparison to the other strains revealed that these sequences also correspond to insertions specific to CC-2937, although, unlike the GEVE, the flanking sequences are nonrepetitive, and the signatures of integration can be directly resolved (Fig. 1C). The insertions in chromosomes 16 and 7 were 48.2 and 5.28 kbp long, respectively. The termini of these insertions perfectly match the left and right ends of the GEVE TIRs, and both are flanked by distinct 6-bp TSDs. Most of the integrated sequences can be mapped to regions of the GEVE, suggesting that they represent relicts of closely related viruses that have undergone deletion following endogenization. We also found relicts of the TIR sequence, several of which were flanked by 6-bp TSDs, among the other available *C. reinhardtii* genomes (table S1). Altogether, the widespread presence of GEVE relicts in other strains, as well as in other locations in the CC-2937 genome, indicates that viral integration is a common occurrence and that there is likely strong selection for large mutations to deactivate GEVEs.

TSDs of fixed lengths are associated with distinct families of DD(E/D) integrase enzymes, which introduce staggered nicks in the target DNA that, after repair, result in duplications that correspond to the length of the stagger between the two DNA strands (28). Integrases that form 6-bp TSDs are typically associated with specific members of a broad assemblage of enzymes that includes retroviral-like integrases [from long terminal repeat (LTR) retrotransposons and polintoviruses] and integrases of specific eukaryotic and prokaryotic DNA transposons (29). To identify candidate integrases encoded by the GEVE that could form these characteristic 6-bp TSDs, we performed homology searches using HHblits on all GEVE-encoded proteins with unknown functions. Most of the integrase candidates that we found belong to the *IS630-Tc1-Mariner* superfamily that introduces “TA” dinucleotide TSDs and are encoded by virus-specific selfish elements. However, we identified one additional retroviral-like integrase (GEVE_506) featuring two chromodomains at the C terminus (Fig. 1D). Integrases of some LTR retrotransposons and polintoviruses (phylum *Preplasmiviricota)* feature a single C-terminal chromodomain (30, 31), and in the case of Chromovirus LTRs, this chromodomain is associated to targeted integration into heterochromatin, potentially limiting the deleterious effects of insertion on the host (32). We found homologs of this protein in diverse preplasmiviruses, a broad assemblage of mid-size DNA viruses that includes several parasites of giant viruses (i.e., virophages) (fig. S2). Virophages often integrate into eukaryotic genomes and produce TSDs of 5 or 6 bp and,

in some cases, act as a kind of inducible antiviral defense against giant virus infection (33). It is plausible that this integrase is responsible for integration into the host *C. reinhardtii* genome, and if this is true, it would point to a potential case in which a giant virus evolved a latent infection strategy by co-opting a virophage enzyme.

Viral particles are produced in *C. reinhardtii* CC-2937 cultures

Next, we sought to assess whether viral particles could be detected in cultures of *C. reinhardtii* CC-2937. We monitored virion production in cultures from inoculation to stationary phase by performing a polymerase chain reaction (PCR) assay targeting the viral *mcp* gene on 0.45-μm-filtered supernatants treated with deoxyribonuclease (DNase) to eliminate nonencapsidated host DNA (see Materials and methods). Our results indicate that free virions begin to accumulate as the cultures reach stationary phase at 6 days after inoculation (average 1.0×10^7 cells ml^{-1}) (Fig. 2A and fig. S3).

To examine this trend in more detail, we quantified viral DNA in these samples through quantitative PCR (qPCR) by targeting the GEVE *mcp* gene sequence and comparing the amplification results with a calibration curve generated from amplifying the *mcp* sequence in a DNA construct of known concentrations (fig. S4; see details in Materials and methods). Increased virion abundance in culture occurred in two waves, peaking at days two and seven after inoculation with an average of 5.0×10^4 and 44×10^4 *mcp* copies ml^{-1}, respectively (Fig. 2B). The cultures appeared healthy and did not crash, demonstrating that the production of virions did not result in widespread cell death. These results indicate that low levels of viral production were maintained at high host cell densities, leading to a ratio of virions to host cells of ~0.05:1. We also verified the presence of free viral particles using flow cytometry on the viral-fraction material concentrated by tangential flow filtration. We identified a distinct population of particles that had a staining signature comparable to that of large dsDNA nucleocytoviruses (Fig. 2C and fig. S5, A to C) (positive controls taken along in our analysis) (34). Negative-stain electron microscopy of concentrated CC-2937 supernatants consistently showed spherical particles ~200 nm in diameter with electron-dense cores (Fig. 2C and fig. S6). Last, we performed short-read DNA sequencing of purified virions that confirmed packaging of the full GEVE region into the viral particles (fig. S7).

The presence of viral particles suggests that the GEVE is active, and it is therefore appropriate to coin a name to refer to this previously uncharacterized viral isolate. For the species taxon, we propose the binomial name “punuivirus latens.” The genus name draws inspiration from the Inca deity Puñuy (“who grants untroubled sleep”), and the species name refers to the cryptic infection strategy of this virus. For the viral isolate, we use the trivial name punuivirus cr2937.

Prevalent transcriptional activity of GEVE genes during activation

To examine the viral activity during culture growth in more detail, we grew duplicate cultures over 7 days and harvested cells at different time points within the growth cycle to assess transcript abundance using RNA sequencing (RNA-seq) (fig. S8, A and B). Consistent with our qPCR results, we found that the expression of viral genes peaked at

the late-exponential and early stationary phases of host growth (~6 days after inoculation, 4.5×10^6 to 5.4×10^6 cells ml^{-1}) (Fig. 3A). Almost all GEVE genes were expressed during at least one point along the time course ($n = 499$, 86%), including the complete set of nucleocytovirus markers (Fig. 3B), which shows that full activation of viral gene expression occurred (data S2). We used self-organizing maps to demarcate the genes into two distinct clusters based on whether they were primarily expressed before peak viral production (BVP cluster; days 3 and 4, $n = 24$ genes) or during peak viral production (DVP cluster; days 5 to 7, $n = 167$ genes; Fig. 3C). Genes expressed before peak viral production tended to be colocalized near the ends of the GEVE and within the TIRs (in seven clusters of at least two genes, with only two not being colocalized with another gene), whereas the central region was populated mostly by genes expressed during peak viral production (Fig. 3A). The early expression of the BVP cluster before peak viral production, together with the colocalization of many of these genes on the GEVE, indicates that they may have a potential role in the suppression of viral activation. DESeq2 analyses revealed that approximately one-third of the GEVE genes (191 out of 579) were differentially expressed (Fig. 3D and fig. S9). Transcripts enriched during peak viral production include those of *mcp* and other structural genes needed for virion formation, consistent with the activation of these genes during virion biogenesis.

Prevalence of GEVE-encoded selfish genetic elements

The GEVE is predicted to encode several selfish genetic elements, and our RNA-seq time course showed these elements to be transcriptionally active. Included among the GEVE's transcriptionally active selfish elements were a single *Metaviridae* LTR retrotransposon, at least seven putative homing endonuclease genes (HEGs), and several Fanzor-encoding elements. The GEVE-encoded HEGs include five inteinic LAGLIDADG nucleases and two freestanding HNH-3 endonucleases. The inteinic LAGLIDADG HEGs are located within the RNA polymerase alpha subunit (RNAPL) ($n = 2$), RNA polymerase beta subunit (RNAPS) ($n = 2$), and DNA polymerase family B (PolB) ($n = 1$) genes (fig. S10A). The freestanding HNH-3 endonucleases are located proximal to the GEVE's *mcp* gene (fig. S10B). The LTR retrotransposon belongs to the family *Gypsy-4_cRei*, which introduces 5-bp TSDs, and is present at several locations in the *C. reinhardtii* genome (data S3).

The GEVE-encoded Fanzor elements can be split into three distinct families that we refer to as A, B, and C, and copies within a family exhibit high nucleotide identity (>99%). For each family, the full-length element also includes a gene encoding a *IS360-Tc1–Mariner* transposase. Within the GEVE, we observed three full-length copies of family A, five of family B, and four of family C. In addition to full-length elements, we also observed other arrangements, including elements that consisted of only the Fanzor gene and right-end guide, nonautonomous transposons that had both ends maintained but gene content was absent or highly degraded, and other element fragments (fig. S11A and data S4). We were able to find homologs to our Fanzor proteins encoded in the GEVEs of other green algae, and we constructed a phylogeny of all these elements together with other references (fig. S11B). Fanzor families A and B are related, whereas family C belongs to a distinct lineage. All of the *C. reinhardtii* GEVE Fanzors belonged to the previously defined Fanzor 1 lineage that is associated with diverse mobile elements in eukaryotic and giant virus genomes

(35). Moreover, we found a Fanzor fragment within chromosome 17 of the *C. reinhardtii* reference genome (strain CC-4532) that bore 80% nucleotide identity to the sequence from Fanzor C. Together with the apparent mobility of the viral-encoded LTR retrotransposon, these findings show widespread sharing of selfish genetic elements between virus and host, indicating that endogenous giant viruses could be important vectors of selfish DNA in eukaryotes.

Diverse proteins are packaged into punuivirus virions

To confirm the presence of free virions and identify the suite of proteins that are likely packaged, we performed liquid chromatography–tandem mass spectrometry (LC-MS/MS) on the supernatants of aging *C. reinhardtii* CC-2937 cultures. A total of 43 proteins were identified with high confidence (at least two peptide-spectrum matches in distinct samples; see Materials and methods and data S5). Among these, the MCP was by far the most abundant protein detected, as expected for free virions (Fig. 4). Other abundant proteins included both multisubunit RNA polymerase subunits, several putative viral helicases, DNA topoisomerase II, a putative glycosyltransferase, a putative capsid fiber, and a viral scaffold protein, most of which have been found to be packaged in other nucleocytoviruses (36–38). Most of the other packaged proteins had no predicted function. Given the particle diameter of virions (~200 nm), this number of encoded proteins is within the expected trends observed for other nucleocytoviruses with a similar virion size, including coccolithoviruses, schizomimiviruses, and marseilleviruses (36). Our proteomic analysis also detected group B and C Fanzors, as well as the Gag-Pol polyprotein from the LTR retrotransposon (Fig. 4). This may suggest that the effectors are active immediately upon cellular entry during viral infection. Work in bacteriophages has shown that HEGs can mediate interviral competition during coinfection (39, 40), and it is possible that Fanzors and other selfish genetic elements encoded in GEVEs may also play a similar role.

Prevalence of punuivirus relatives associated with *Chlamydomonas* populations in Swedish lakes

To assess the prevalence of cryptic infections in a distinct natural population of *Chlamydomonas* spp., we analyzed monoclonal culture strains isolated in 2016 from Örsjön and Krageholmssjön, two lakes located in southern Sweden. These isolates fall within the *Chlamydomonas* genus and are closely related to *C. reinhardtii*, as indicated by molecular analysis of 18*S* ribosomal RNA (rRNA) gene amplicons (fig. S12). Thirteen of the 18 isolates (72%) from Örsjön and 12 of the 20 from Krageholmssjön (60%) tested positive for amplification of nucleocytovirus *mcp* genes (fig. S13 and table S2). Similar to *C. reinhardtii* CC-2937, these monocultures grow well in the laboratory and have not undergone any crashes, verifying that the viruses associated with these strains do not cause observable levels of cell death. Sequencing of the *mcp* PCR products from two isolates of Örsjön and Krageholmssjön (Ors24 and Kgh18, respectively) confirmed that these viruses are related to GEVEs within the order *Imitervirales* that were previously found in green algal genomes (fig. S14). In addition, we performed low-coverage PacBio sequencing on strain Ors24 that yielded a viral DNA polymerase B sequence. Phylogenetic analysis of this gene confirmed the placement of this isolate within a GEVE clade (fig. S15), suggesting that giant viruses related to punuivirus are prevalent in this population.

We selected strain Ors24 for thin-section transmission electron microscopy (TEM) at different growth phases to investigate whether viral particles could be observed. Consistent with our findings for *C. reinhardtii* strain CC-2937, we observed viral particles of ~225 nm in diameter that appeared primarily in the mid-exponential phase (Fig. 5A). We also observed virions in up to 3% of the cells (fig. S16); because virions would only be expected to be visible in the later stages of a lytic infection program, this suggests that a larger fraction of cells were undergoing active viral infection at that time. Virions with clear icosahedral symmetry were formed from apparent virus factories (Fig. 5, B and C), indicating that these structures are formed in the cytoplasm during viral activation. The similar infection dynamics we observed in both strains Ors24 and *C. reinhardtii* CC-2937 suggest that latent virus activation is taking place in both cultures during active growth. The phylogenetic proximity of the viruses involved, together with the previous discovery of a large clade of endogenous giant viruses in diverse green algae (8), suggests that this viral lineage is associated with a range of different green algae in nature.

Conclusions

We have identified a giant virus integrated into the *C. reinhardtii* genome and present evidence that it can actively produce viral particles from seemingly healthy algal cultures. A notable feature of this virus, referred to as punuivirus, is the presence of a putative integrase that shares homology with enzymes encoded by virophages, which suggests that interactions with hyperparasites may have driven the evolution of its cryptic infection program. Despite our current insights, the details of how a ~600-kbp genomic payload can be integrated into a host genome remain unclear, however, and it will be important for future work to clarify the mechanistic details of this process. Moreover, in the case of GEVE reactivation, it is also unknown what signals induce viral activation and virion production. In natural populations, only a small fraction of cells appears to produce virions, even during peak viral activity, suggesting that population heterogeneity during cellular growth plays a role. As virion production appears to peak during mid-exponential– or stationary phase, it is possible that a buildup of metabolic byproducts may signal viral activation. Last, although our results are consistent with a model of GEVE reactivation and virion production, other scenarios involving the persistent infection of a subpopulation of host cells remain a possibility and should be investigated further.

The GEVE genome encodes a variety of selfish genetic elements that are expressed, and their presence in several locations indicates that they can mobilize to other areas of the host and viral genomes. Among these, Fanzor elements are programmable RNA-guided nucleases that are of interest for genetic engineering applications; in this context, one may consider that punuivirus is a vector for these enzymes as part of its normal infection program. In future work, it will be revealing to understand the molecular details of how transposon mobility occurs during infection, as well as the long-term consequences of the multipartite coevolution between virus, host, and selfish genetic elements.

A central aspect of punuivirus latency is its ability to integrate into the *C. reinhardtii* genome, but viral integration is not necessarily a requirement for long-term persistent infections. Indeed, some virulent nucleocytoviruses can stably coexist with their hosts by

infecting only a subset of the population, thereby leading to long-term viral persistence without host population collapse (41, 42). Moreover, other giant viruses have low virulence in a particular host but appear to compensate with a broader host range (43). We surmise that the integration of punuivirus into the *C. reinhardtii* genome provides an added benefit to the virus by ensuring that it can be maintained even during long periods of host dormancy, for example, in the durable zygospores that form during sexual reproduction. Zygospores are highly resistant to environmental perturbations, remaining viable in soil for several years (44), and integration into these cells may promote long-term viability of the virus.

The cryptic infection program of punuivirus is potentially a common strategy among large protist viruses in nature that has traditionally been overlooked because of methodological challenges. For example, most cultivated giant viruses have been discovered because of their pronounced impact on cultures of their host (i.e., “culture crashes”), and the lack of any clear phenotypic effect of a cryptic virus, even during peak viral production, has likely impeded the earlier discovery of this phenomenon. Studies dating to the 1970s observed viral production in otherwise healthy cultures of green algae and speculated that it may be due to the activity of latent viruses, but this was difficult to prove owing to technological limitations and the possibility of environmental contamination (13, 14). Our observations of endogenous viral activity in *C. reinhardtii*, together with our discovery of widespread related viruses in freshwater *Chlamydomonas* isolates, clarifies these earlier observations and revives the view that latency may be commonplace in large DNA viruses of protists. Indeed, eukaryotic genomes have also been shown to harbor sequences derived from the recently discovered mirusvirus lineage of large DNA viruses (18, 45, 46). Altogether, these findings highlight genome integration as a potentially common strategy used by diverse lineages of large eukaryotic DNA viruses as part of a latent infection cycle.

Materials and methods

Maintenance and culture conditions for *C. reinhardtii* CC-2937

C. reinhardtii strain CC-2937 was acquired from the Chlamydomonas Resource Center (Minneapolis, MN, USA) and maintained on 2% agar TAP media (no. T8224, Plant Phytotech Labs, Lenexa, KS, USA) slants supplemented with 4 g liter^{-1} of yeast extract and 1 ml liter^{-1} of glacial acetic acid, adjusted to pH 7 with glacial acetic acid. Liquid TAP media was prepared identically, omitting the agar and yeast extract. All liquid and agar cultures were maintained at 24°C under a 12 hour:12 hour light:dark cycle at an intensity of 100 μmol quanta m^{-2} s^{-1}. Liquid cultures were agitated using an orbital shaker (Fisherbrand Multi-Platform shaker, Thermo Fisher Scientific, Waltham, MA, USA) at 150 rpm in conical flasks with a total capacity twice that of the media volume used. Cell densities were measured using the CytoFLEX-S Flow Cytometer (Beckman Coulter, Brea, CA, USA) equipped with violet (405 nm) and blue (488 nm) lasers. Chlorophyll autofluorescence was excited by the 488-nm blue laser and collected using a 780/60-nm band-pass filter. A total of 10,000 events were recorded per measurement, with a medium flow rate (30 μl min^{-1}). Cultures were diluted accordingly to maintain reads between 100 and 1500 events μl^{-1}. The Forward scatter and chlorophyll autofluorescence channels were set with automatic thresholds and gain values of 42 and 124, respectively.

Genomic DNA extraction

High–molecular weight genomic DNA (gDNA) was isolated from three *C. reinhardtii* CC-2937 late-exponential cultures (~10^7 cells ml^{-1}). First, 50 ml of the culture was centrifuged at 4500*g* for 5 min in a Sorvall ST1R Plus-MD centrifuge with the TX-400 rotor (Thermo Fisher Scientific, Waltham, MA, USA). The resulting pellet was washed once with phosphate-buffered saline (PBS) 1X and gently resuspended in 5 ml of SDS buffer (50 mM Tris-HCl pH 8, 200 mM NaCl, 20 mM EDTA, 2% SDS) and 5 ml of CTAB buffer (100 mM Tris-HCl pH 8, 20 mM EDTA pH 8, 1.4 M NaCl, 2% CTAB, 1% PVP M.W. 40,000) preheated at 65°C. Five microliters of RNase A 100 mg ml^{-1} (#19101, Qiagen, Germantown, MD, USA) and Proteinase K 20 mg ml^{-1} (#P50220, RPI, Mount Prospect, IL, USA) were added and incubated at 65°C for 1 hour, mixing every 15 min. The lysate was centrifuged at 4500*g* for 5 min and decanted into a new falcon tube. One volume of phenol-chloroform (1:1) was added and mixed by inversion for 10 min, followed by centrifugation at 3000*g* for 10 min. The supernatant was transferred to a new tube using wide-bore pipette tips and extracted again using one volume of chloroform. Five microliters of proteinase K 20 mg ml^{-1} and 10 µl of RNase A 100 mg ml^{-1} were added and incubated at 50°C for 1 hour, followed by the addition of one volume of chloroform and centrifugation at 3000*g* for 10 min. The DNA in the supernatant was precipitated with 2.5 volumes of cold 100% ethanol and recovered by centrifugation at 4500*g* for 5 min. The pellet was transferred to a DNA LoBind tube containing 70% ethanol and dried at 39°C for approximately 10 min. Finally, the DNA was resuspended in 400 µl of molecular grade water and stored at 4°C.

DNA shearing, cleanup, and library preparation for long-read sequencing

The extracted gDNA was sheared 30 times with a 27-gauge syringe needle and purified using 0.7x AMPure XP beads (Beckman Coulter Inc., Indianapolis, IN, United States) for 15 min. The DNA was eluted in 60 µl of preheated elution buffer (10mM Tris-HCl pH8) for 20 min at 40°C. The DNA purity was measured with a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA), and integrity was assessed by performing agarose gel electrophoresis and running 4200 TapeStation Genomic DNA ScreenTape assays (Agilent Technologies, Santa Clara, CA, USA). The DNA concentration was measured using a Qubit fluorometer, and approximately 3.5 µg of DNA was used for library preparation using the SQK-LSK114 Ligation Sequencing Kit V14 from Oxford Nanopore (Oxford Science Park, UK) with modifications. The formalin-fixed paraffin-embedded (FFPE) repair step was omitted, and the end-prep step was performed using the NEBNext Ultra II End Repair Module (New England BioLabs, Ipswich, MA, USA) according to the manufacturer's instructions. The end-prepped DNA was diluted with two volumes of elution buffer and cleaned up with 1x AMPure XP beads as described previously, using 60 µl of elution buffer. For the adapter ligation, the reaction was incubated for 1 hour, then diluted with one volume of elution buffer and cleaned with 0.8x AMPure XP beads, which were washed twice with 250 µl of a mix of SFB:LFB buffer (1:2). The DNA was eluted with 30 µl of the provided elution buffer as described before. About 2.7 µg of library was recovered, which was used to load a PromethION Flow Cell (R10.4.1) three times (370 ng per load) after washing it for 2 hours every 24 hours using the Nanopore EXP-WSH004 Flow Cell Wash Kit. Before this run, two sequencing attempts without flow cell reloads were

performed using 1 μg of unsheared and sheared gDNA (from two different culture flasks) as the input for library preparation following the kit instructions.

Genome assembly and polishing

The raw reads were base-called in real time using the MinKNOW software (v23.07.5) and Guppy (v7.0.9) with the high-accuracy model (400 bps, 5 khz). Reads generated with the three sequencing runs were pooled, and those shorter than 1 kb were discarded. A draft genome assembly was generated with Flye v2.8.3 (47) with the parameters "-nano-raw" and a genome size of 120 Mb. The assembly was polished using long reads and four rounds of Racon v1.4.20 (https://github.com/isovic/racon) with default settings, followed by one round of Medaka v1.11.1 (https://github.com/nanoporetech/medaka) specifying the model r1041_e82_400bps_hac_v4.2.0. The Medaka consensus assembly was further polished with two rounds of Racon using Illumina short reads that had previously been generated for the CC-2937 strain [National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) accession no. SRR1734616] (48). The estimated genome size recovered was ~110 Mb, which included 77 contigs and two scaffolds with an N50 value of 3.9 Mb. Contigs below 10 kbp were filtered for downstream analyses. The assembly was mapped against the telomere-to-telomere assembly of *C. reinhardtii* strain CC-5816 (49) using Minimap2 v2.12.0 (50) (option -cx asm20) to assign contigs to specific chromosomes.

GEVE contig identification

ViralRecall v2.1 (24) was run on the final polished assembly (using the contig screening parameter "-c") to identify the contig(s) containing nucleocytovirus marker genes. One contig (contig_536), which mapped to chromosome 15 in *C. reinhardtii* CC-5816, was found to contain clear signatures of an endogenous nucleocytovirus. We used Minimap2 (parameters: -X -N 50 -p 0.1 -cx asm20) to find the synteny blocks, which were then visualized using the R package gggenomes (https://thackl.github.io/gggenomes/).

The TIRs were annotated using Minimap2 by mapping the contig against itself as described previously (7). The precise GEVE region was determined by the TIR boundaries, Minimap2 alignment, and manual comparison of the flanking sequence among CC-2937, the reference genome (CC-4532 v6), and the genomes of two other field isolates (CC-1952, CC-2937) (27). GEVE relicts were identified by BLASTn (51) searches using the GEVE TIRs as query sequences. The repeat content was calculated with Tandem Repeats Finder v4.09 (52) with parameters described elsewhere (49). The Repeat and GC fractions were calculated from 10-kb nonoverlapping sliding windows. Tetranucleotide frequency deviation of the GEVE compared with the rest of chromosome 15 was calculated using methods previously described (8).

Functional annotation

ORFs from the GEVE region were predicted using Prodigal v. 2.6.3 (default parameters) (53), and a preliminary set of functional predictions was generated by running eggNOG-mapper v2 (54), as well as ViralRecall 2.0. Multiple sequence alignments for proteins of unknown function were queried using the HHblits search module of HH-suite v3.3.0 (55, 56) against the UniRef30_2023_02 database (two iterations, e-value cutoff 1×10^{-5}).

Repeat elements, including Fanzor elements in complete, partial, or variant configurations, were identified through a BLASTn search of the GEVE against itself in QIAGEN CLC Main Workbench v7.9.1 using default parameters. The predicted functions of the proteins encoded in repeated regions were generated using HHpred (57). To identify the putative GEVE integrase, we searched the HHblits output for hits to DDE-type integrases (probability ≥95%, amino acid length ≥500) and generated protein structure predictions for the candidates using AlphaFold2 (58). The viral factory scaffold protein was predicted using VFCpredict (59).

Integrase and Fanzor elements phylogenetic tree

Homologs of the GEVE integrase candidate (GEVE_506) were identified with a BLASTp search against a curated database of proteins predicted from preplasmiviruses (polintoviruses, virophages, polinton-like viruses, and relatives) compiled from previous studies (7, 60–62). Sequences with conserved DD(E/D) residues were retained and aligned with a collection of DD(E/D) integrase sequences from transposable elements (29) and selected HHblits hits (probability ≥97%) using Muscle v5.2 (63). We trimmed the alignments to remove all sites with ≥90% gaps using trimAl v1.5 (64) and constructed the phylogeny using IQ-TREE v2.3.6 (65) with the Q.pfam+F+I+G4 model, as chosen by ModelFinder (66), and 1000 ultrafast bootstraps (67). Fanzor nuclease sequences were aligned and trimmed as described before. The phylogeny was constructed using IQ-TREE with VT+F+R10 as the best-fit model.

Exponential-to-stationary growth experiment and RNA sequencing of *C. reinhardtii* CC-2937

C. reinhardtii strain CC-2937 was taken from a slant and cultured in liquid TAP media. A starter culture was maintained at a cell density of approximately 150×10^3 to 200×10^3 cells ml^{-1} through daily dilution. From this starter culture, two biological replicate cultures were inoculated at an initial cell density of ~100×10^3 cells ml^{-1} for five consecutive days starting 31 March 2023 (fig. S8A). The cell density of each culture was monitored daily using flow cytometry as previously described (fig. S8B). On the seventh day (6 April 2023), 1.5 ml from each culture was harvested by centrifugation (4000*g*, 4 min, 4°C). The supernatant was discarded, and the cell pellets, containing between 3 million and 8 million cells each, were rapidly frozen in liquid nitrogen and stored at −80°C for future analysis.

RNA was extracted from cell pellets using TRIzol Plus RNA Purification Kit (Invitrogen, 12183555) following the manufacturer's instructions. Total RNA was quantified using a Qubit RNA HS kit (Invitrogen, Q32852), and the quality was assessed with the HS RNA ScreenTape on an Agilent TapeStation system. RNA was converted into a strand-specific library using Illumina's Stranded Total RNA Prep, Ligation with Ribo Zero Plus Sample Prep Kit (Illumina, 20040529) for subsequent cluster generation and sequencing on Illumina's NovaSeq 6000. Supplemental probes (table S3) were used in the hybridize probes step in the RiboZero Plus reactions. The libraries were enriched by 13 cycles of PCR, validated using Agilent TapeStation, and quantitated by qPCR (P5 Primer: AATGATACGGCG-ACCACCGA, P7 Primer: CAAGCAGAAGACGGCATACGAGAT). Individually indexed cDNA libraries were pooled and sequenced on NovaSeq 6000 SP 150

cycle PE using Illumina NovaSeq Control Software v1.8.0. The BCL files were converted to FASTQ files, and adapters were trimmed and demultiplexed using bcl2fastq Conversion Software.

Gene prediction of the RNA-seq assembled transcripts

A transcriptome coassembly was generated from raw RNA-seq illumina reads using rnaSPAdes v3.13.0 (68). The polished *C. reinhardtii* CC-2937 assembly was soft-masked using tantan v22 (69), and then assembled transcripts were mapped onto this reference using BLAT v35 (70) (parameters "-minIdentity=92") to obtain a psl file. The psl file was converted to a hints file using the blat2hints.pl PERL script provided with AUGUSTUS (https://github.com/nextgenusfs/augustus), and genes were then predicted using AUGUSTUS v. 3.5.0 (71) with the hints file and polished assembly used as input (parameters–species=chlamy2011–soft-masking=1). For the GEVE region, the coding density was lower than expected for a viral genome, likely because AUGUSTUS was not designed for viral gene prediction. To resolve this issue, AUGUSTUS gene predictions in the GEVE region were replaced with genes predicted using Prodigal (default parameters). To estimate the expression level of genes in the different RNA-seq experiments, raw RNA-seq reads were trimmed with Trim Galore v0.6.4 (https://github.com/FelixKrueger/TrimGalore; parameters "–length 36 -q 5–stringency 1") and then mapped onto the predicted transcripts using CoverM v0.4.0 with a minimum covered fraction of 20% (https://github.com/wwood/CoverM).

Differential expression analysis

A differential expression analysis of the RNA-seq count data was performed with the DESeq2 package (72). The reference level for all contrasts was set to the youngest culture (3 days after inoculation). Significant differentially expressed genes (DEGs) were identified based on an absolute log2-fold change (LFC) of 1.5 (Wald test: LFC threshold = 0.585, alpha = 0.05) and a false discovery rate (FDR) adjusted p value <0.05. The shrunken LFCs were estimated using the "normal" shrinkage estimator and visualized with volcano plots generated by the EnhancedVolcano (https://github.com/kevinblighe/EnhancedVolcano) R package. DEGs were clustered using self-organizing maps (SOMs) via the kohonen package in R (73). We used the SOM codebook vectors in combination with K-means clustering to determine the optimal number of gene clusters. Then, we performed an unsupervised hierarchical clustering using the Ward method and Euclidean distance on the SOM codes to assign each gene to a cluster. The normalized DESeq2 counts belonging to the GEVE genes were scaled across samples with the min-max method prior to visualization on a heatmap generated using the pheatmap R package (https://github.com/raivokolde/pheatmap).

Detection of viral particles by PCR and qPCR

A loopful of freshly growing CC-2937 cells on TAP agar plates (7 days old) was used to inoculate a 25-ml TAP media starter culture, which was grown for six days to a density of 9.5×10^6 cells ml^{-1}. The starter culture was then used to inoculate four flasks with 125 ml of TAP media to a final cell density of ~3×10^5 cells ml^{-1}. For a total of 11 days, cell density was monitored daily with flow cytometry, and cells were observed with a Nikon Ti2-E inverted microscope with transmitted illumination. Five hundred–microliter samples

were collected and centrifuged at 900*g* for 1 min, and the supernatants were then filtered using 0.45-μm PES syringe filters and stored at 4°C until further processing. Unpackaged host DNA contamination in the filtrates was removed by treating the samples with 1 U of DNase I (no. EN0521, Thermo Fisher Scientific, Waltham, MA, USA) per 8 μl of sample, following the manufacturer's instructions and stored at −20°C until use. Viral DNA was detected by PCR using 40 cycles and 2 μl of the DNase-treated filtrates as template in 10-μl final volume reactions, using the GEVE major capsid protein (*mcp*) primer pair and conditions described in a previous study (22). Host DNA contamination was detected by amplifying the ITS1–5.8S-ITS2 region using the primer pair Fw_ITS1/Rv_ITS4 (74, 75) and the same conditions used previously.

To quantify free virions in terms of MCP copies μl^{-1} using qPCR, we initially generated a standard curve constructed from a dilution series ranging from 2.82×10^0 to 1.41×10^5 molecules per μl of a linear construct (Twist Bioscience, South San Francisco, CA, USA) containing a fragment of the GEVE *mcp* gene (GEVE_395) (5′ CAATCCGCCCTCACTACAACCGGAGTGCTACTATTTCAGCCAGCTCACAAACCAG AAGAACGGGTCCATTACCATAGGCAACTTGGATGCTTCGATGTACCTGGATTACGT GTATCTGGACACAGATGAGCGCAAGAAGTTTGCCCAAGCCGCTCACGAATACCTG GTGGAGCAGCTGCAGTATACCGGCGAGGAGTCGCTGCAGGGAAGCCAGGGCAAG GTGAAGCTGAGCCTGAACCACCCCGTTAAGGAGCTGATTTGGGTGATGCAGAAG GATGACTGGCTGACCAACACCGGCGCCAGGGTGATTGTGCCTACCTCTGCTACTC TG-GCGTCGATGAGGGA 3′). The target sequence was amplified with the primers GEVE_MCP_qPCR forward (5′ GCAAGAAGTTTGCCCAAGCCGC 3′) and reverse (5′ CTCAGCTTCACCTTGCCCTGGC 3′), which amplified a product of 100 bp. We conducted triplicate qPCR reactions of 24 μl containing 1x Platinum SYBR Green qPCR SuperMix-UDG w/ROX (no. 11744500, Thermo Fisher Scientific, Waltham, MA, USA), a final concentration of 300 nM of each primer, and 4 μl of the DNase-treated filtrates. The thermal cycling was performed in the CFX96 Real-Time PCR system (BioRad, Hercules, CA, USA) using the following settings: 50°C for 2 min, 95°C for 2 min, 45 cycles of 95°C for 15 s followed by 60°C for 30 s, and a final melting curve from 65° to 95°C. Between all qPCR runs, the equation of the standard curve was Cq = 46.09 − 3.29 log10 (*mcp* copies ml^{-1}) and the R^2 value was 0.94.

Viral population detection through flow cytometry

After 15 days of incubation, and considering the low viral loads, the contents of the four replicate flasks used for qPCR were pooled to ensure sufficient material for concentration by tangential flow filtration (TFF) before detecting viral particles by flow cytometry. Cultures were centrifuged at 4695*g* for 20 min at 4°C to pellet the cells in a Sorvall ST1R Plus-MD centrifuge with the TX-400 rotor (Thermo Fisher Scientific, Waltham, MA, USA). The collected supernatant was filtered sequentially through decreasing pore size filters using a peristaltic pump. A total of six 5-μm filters, two 3-μm filters, and one 0.8-μm filter were used to filter out cellular debris without clogging. About 350 ml of filtrate was recovered and stored overnight at 4°C. Viral particle concentration was performed using a 100-kDa (MWCO) PES Vivaflow 200 TFF unit (Sartorius, Göttingen, Germany), keeping an inlet pressure <10 psi, to a final volume of approximately 25 ml (~14-fold concentration).

Fifteen milliliters of this sample was further concentrated with an Amicon 100-kDa Ultra Centrifugal Filter (Millipore Sigma, St. Louis, MO) to about 1.1 ml (total ~190-fold concentration). This concentrate was diluted 1:2 with molecular grade water, and 50 µl was fixed with 25% glutaraldehyde (stored at 4°C, Electron Microscopy Sciences, Hatfield, PA, USA) to a final concentration of 0.25% for 20 min in the dark at room temperature. The rest of the sample was filtered using a 0.8-µm PES filter, and a 50-µl aliquot was fixed as described before. This step was repeated with a 0.45-µm PES filter to visualize the particles using different pore size cutoffs (fig. S5, D to F).

Based on the virus detection and enumeration protocol by Brussaard *et al.* (76, 77), the glutaraldehyde fixed samples (2x diluted) were further diluted (50x) with TE-buffer (10 mM Tris-HCl pH 7, 1 mM EDTA pH8) and stained with nucleic acid–specific SYBR Safe (diluted to 1X in the TE-buffer, Thermo Fisher Scientific, Waltham, MA, USA) for 20 min at 80°C. Particles with virus-like green fluorescence were excited using the 405-nm violet and 488-nm blue lasers using the CytoFLEX-S flow cytometer. The Violet Side Scatter (V-SSC, 200 gain) and the Green Fluorescence (B525, 2,000 gain) channels were fitted with a 405/10 and a 525/40 band-pass filter, respectively (78, 79). The thresholds were set to 1200 for V-SSC and 20,000 for B525 to keep the abort rate below 1%.

As a positive control for the detection of large DNA viruses, we used Paramecium bursaria chlorella virus 1 (PBCV-1) and Acanthamoeba polyphaga mimivirus lysates. For the PBCV-1 positive control, the B525 channel threshold was decreased to 15,000 because it had a slightly lower green fluorescence signal under the set parameters. Both controls were easily detected (fig. S5, A and B). The filtrate (permeate; <100 kDa) collected from the Vivaflow concentration unit was treated identically and used as a negative control (fig. S5I).

DNA extraction from viral particles

The previously obtained Vivaflow concentrate was used to extract DNA from free virions using a phenol-chloroform protocol. Briefly, nonencapsidated DNA was removed by treating 12 tubes, each containing 450 µl of the sample, with 1 µl of RNase A 100 mg ml^{-1}, 3 U of DNase, and 50 µl of 10 X DNase Reaction Buffer. Samples were incubated for 1 hour at 37°C, and the nuclease reaction was subsequently stopped by adding 50 µl of EDTA (50 mM) and incubating for 10 min at 65°C. Viral capsids were digested by adding 27.5 µl of SDS (10%) and 5 µl of Proteinase K (20 mg ml^{-1}) and incubating for 3 hours at 65°C. Two rounds of extraction were performed using one volume of phenol-chloroform-isoamyl alcohol (25:24:1), followed by centrifugation at 13,000*g* for 5 min. Two additional rounds of extraction were performed using one volume of chloroform only. DNA was precipitated by adding 1/10 volume of sodium acetate (3 M, pH 5.2) and 2.5 volumes of cold 100% ethanol. Samples were incubated overnight at −20°C, and DNA was recovered by centrifugation at 20,000*g* for 30 min at 4°C. Pellets were washed with 70% ethanol and allowed to dry for 5 min at 39°C. The dried pellets were resuspended in molecular-grade water and pooled into a single sample with a final volume of ~100 µl. To confirm the presence of viral DNA devoid of host DNA, we performed a PCR targeting the *mcp* gene and the ITS1–5.8S-ITS2 region using 1 µl of DNA, 35 cycles of amplification, and conditions as described previously (22). The PCR showed positive amplification of only viral DNA (fig. S17).

Viral particle whole-genome sequencing

Illumina libraries were prepared by SeqCoast Genomics (Portsmouth, NH, USA) using the Illumina DNA Prep tagmentation kit (no. 20060059) with Illumina Unique Dual Indexes. Sequencing was performed twice using the Illumina NextSeq2000 platform (2 × 150-bp paired-end reads), yielding a total of 152,114 read pairs after filtering out contaminating human reads using Kraken (80). Eighty-seven percent of the reads successfully mapped to the *C. reinhardtii* CC-2937 assembly using Minimap2 v2.17. Of these, 95.84% mapped to the viral contig_536, containing the full-length 617 kb GEVE, followed by the viral contig_437 (harboring the 48 kb GEVE relict), with 2.71% of the mapped reads. Coverage was calculated with SAMtools v1.16.1 (81) and smoothed using a rolling average (window size = 1000 bp). Comparison of the raw sequencing reads to the GEVE revealed a mean sequence identity of 99.7%. Single-nucleotide variants (SNVs) identified between the sequencing reads and reference GEVE sequence were identified with inStrain v1.9.0 (default parameters) (82), which found 124 possible SNVs and two single-nucleotide substitutions (data S6). Of these, 103 of the SNVs (83%) were found with a frequency of <10%, suggesting that they represent minor variants in the population.

Identification of virions using proteomics

Three biological replicates were conducted by filling two 1-liter Erlenmeyer flasks with 500 ml of TAP media and inoculating each with a loopful of freshly growing CC-2937 cells from 7-day-old TAP agar plates. The flasks were incubated for 9 days, and the cultures were concentrated using a 100-kDa PES Vivaflow 200 TFF unit, following previously described methods, to a final volume of approximately 22 ml (~40-fold concentration). The concentrated supernatant (~21 ml) was distributed into 1.5-ml tubes and centrifuged at 15,000*g* for 1 hour at 4°C. From each tube, 1400 µl of the supernatant was carefully removed, and the remaining volumes were pooled into a single tube, which was then centrifuged under the same conditions for an additional hour. The supernatant was discarded, and the resulting pellets were frozen at −80°C until proteome analysis.

Protein was solubilized in S-trap lysis buffer (10% w/v SDS in 100 mM triethylammonium bicarbonate pH 8.5). Proteins were reduced using DTT (4.5 mM) and alkylated with IAA (10 mM). Unreacted IAA was quenched with DTT (10 mM), and samples were acidified using o-phosphoric acid. Protein was precipitated using methanol and incubation at −80°C overnight. An aliquot of each sample corresponding to 100 µg of protein was loaded onto a mini S-trap (Protifi, Fairport, NY, USA) and washed with methanol. Proteins were then digested overnight with 2 µg of trypsin in 25 µl of 50 mM triethyammonium bicarbonate (pH 8.5).

LC-MS/MS was performed in duplicate using a Thermo Fisher Scientific Vanquish Neo HPLC and autosampler (Waltham, MA, USA) system controlled by Chromeleon 7.2.10 coupled online to a Bruker timsTOF fleX mass spectrometer via a Bruker Captive Spray ion source (Billerica, MA, USA). Three micrograms (3 µl) of peptide solution was separated on a PharmaFluidics 50 cm µPAC capLC C18 column (Thermo Fisher Scientific, Waltham, MA, USA) at a flow rate of 350 nl min^{-1} in an oven compartment heated to 40°C. The LC gradient that was used started with a linear increase (solvent A: 2% acetonitrile, 98% water,

and 0.1% formic acid; solvent B: 80% acetonitrile, 20% water, and 0.1% formic acid) from 2% to 10% B over 3 min, followed by a linear increase from 10% to 50% B over 88 min, and followed by a wash of 4 min at 98% B.

For the DDA-PASEF acquisition mode, one survey TIMS-MS and 10 PASEF MS/MS scans were performed per acquisition cycle. We analyzed an ion mobility (IM) range from 1/K0 = 0.6 to 1.6 volt-seconds (Vs) cm^{-2} using equal ion accumulation and ramp time in the dual-TIMS analyzer of 100 ms each. Suitable precursor ions for MS/MS analysis were isolated in a window of 2 Thomson (Th) for $m/z < 700$ and 3 Th for $m/z > 700$ by rapidly switching the quadrupole position in sync with the elution of precursors from the TIMS device. The collision energy was lowered stepwise as a function of increasing IM, starting from 20 eV for 1/K0 = 0.6 Vs cm^{-2} and 59 eV for 1/K0 = 1.6 Vs cm^{-2}, making use of the m/z and IM information to exclude singly charged precursor ions with a polygon filter mask and further used "dynamic exclusion" to avoid resequencing of precursors that reached a "target value" of 20,000 arbitrary units (au). The IM dimension was calibrated linearly using three ions from the Agilent ESI LC-MS tuning mix (m/z, 1/K0: 622.0289, 0.9848 Vs cm^{-2}; 922.0097, 1.1895 Vs cm^{-2}; and 1221.9906, 1.3820 Vs cm^{-2}).

Data files were processed with Mascot Distiller 2.8.5 (Matrix Science, Boston, MA, USA) using the default settings for data generated using Bruker timsTOF instruments. Processed data were then searched using Mascot 2.8.3 (Matrix Science, Boston, MA, USA). The search used the UniProt reference *C. reinhardtii* proteome database, a common protein contaminant database, and the FASTA formatted GEVE proteome. The search assumed trypsin-specific peptides with the possibility of two missed cleavages, a precursor mass tolerance of 100 parts per million (ppm) and a fragment mass tolerance of 0.1 Da, a fixed modification of carbamidomethyl at Cys, the variable modifications of oxidation of Met, and cyclization of a peptide N-terminal Gln to pyro-Glu. To identify a final list of proteins packaged in the virions, only proteins that were detected at least twice across all technical runs and with a Mascot database search score ≥35 were retained. Additionally, proteins with a lower score were also retained if at least two peptide-spectrum matches were recorded in at least two runs.

Electron microscopy of *C. reinhardtii* CC-2937 concentrated supernatants

The remaining concentrated supernatant (~5.5 ml) was used for viral particle screening with TEM. Viral particles were pelleted by centrifugation at 15,000*g* for 1 hour at 4°C. Approximately 1300 µl of the supernatant was removed, and the remaining volume was pooled into a single tube and topped up to 1.5 ml with molecular-grade water. After homogenization, the sample was filtered through a 0.8-µm PES filter (13 mm diameter) and centrifuged for an additional hour under the same conditions. The supernatant was discarded, and the pellet was resuspended in 100 µl of molecular-grade water and then filtered through a 0.45-µm PVDF filter (4 mm diameter). A non-GEVE strain supernatant (*C. reinhardtii* CC-2935) was also concentrated and pelleted in the same manner, serving as a negative control to differentiate VLPs from non-VLPs. Flow cytometry was performed on these samples, further confirming that concentrating the non-GEVE strain CC-2935 results in the absence of a small particle population compared with CC-2937 (fig. S5, G and H).

Formvar/Carbon 200-mesh copper grids (Electron Microscopy Sciences, Hatfield, PA) were hydrophilized with UV-C light (254 nm) radiation for 2 hours in a PCR Station Enclosure (no. 3970305, Labconco, Kansas City, MO, USA). Ten microliters of the concentrated sample were placed onto the grid and incubated for 10 min. The excess sample was blotted with filter paper and stained with 3 μl of uranyl acetate 2% for 30 s. This step was repeated once, and the grid was allowed to dry at room temperature. The VLPs were visualized with a JEM 2100 Transmission Electron Microscope (JEOL, Tokyo, Japan) operated at 200 kV. The average diameters of 23 viral particles (four measurements per virion) were measured with the software ImageJ (83).

Assessing the prevalence of viral activity in freshwater *Chlamydomonas* spp. field isolates

Water samples were collected in 2016 from Örsjön (56.2828485, 14.6838229) and Krageholmssjön (55.50159, 13.74462), two lakes located in southern Sweden, using a 10-μm plankton net. Single *Chlamydomonas*-like green algal cells were isolated by hand using an inverted microscope. Individual cells were washed with five drops of filtered lake water (0.2-μm pore size) and placed into 96-well plates containing 0.5X MWC+Se media diluted with filtered lake water (84). Once wells turned visibly green, the cultures were sequentially transferred to larger wells, until finally transferring them to 30 ml of 1X MWC+Se media in Nunc T25 tissue culture flasks (no. 169900, Thermo Fisher Scientific, Sweden). A final total of 18 and 20 monoalgal cultures were established from Örsjön and Krageholmssjön, respectively. These monocultures were maintained at 15°C using a 12 hour:12 hour light:dark cycle and reduced light intensity (10 μmol quanta m^{-2} s^{-1}) with monthly transfers to fresh media. For the experiments and DNA extraction, growth was increased by elevating light intensity to 90 μmol quanta m^{-2} s^{-1}.

To confirm that isolates were *Chlamydomonas* spp., we sequenced the 18*S* rRNA gene following methods described elsewhere (85). For phylogenetic analysis, we aligned the 18*S* sequences of two representative isolates from both lakes (Ors24 and Kgh138), the *C. reinhardtii* CC-2937 strain, and a collection of reference 18*S* sequences from the PR2 database (86). To select references, we compared the 18*S* sequences of the isolates to the PR2 sequences using BLASTn and retained the top 100 hits for each. We then dereplicated the reference 18*S* sequences at 99% identity using CD-HIT (87) and aligned all sequences together using Muscle v5.1. We trimmed the alignments to remove all sites with ≥20% gaps using trimAl v1.4, and constructed the phylogeny using IQ-TREE v2.03 with the TIM2+F+I+G4 model, as chosen by ModelFinder, and 1000 ultrafast bootstraps.

To screen the monocultures for giant viruses, we grew them to mid-exponential phase (approximately 75×10^3 cells ml^{-1}), followed by DNA extraction and PCR amplification of the viral *mcp* gene using the degenerate primers, “mcp Fwd” and “mcp Rev,” which target the *mcp* sequence of large algal viruses (88). For DNA extraction, cultures were centrifuged at 3000*g* for 10 min, and after most of the supernatant was decanted, the pellet was gently resuspended in the remaining liquid. The cell fraction was further transferred to 1.5-ml tubes and centrifuged, and the resulting pellets were stored at −80°C until further processing. DNA was extracted from the cell pellets using the DNeasy Plant Mini kit (Qiagen, Valencia, CA, USA) with modifications. First, the pellets were transferred to 2-ml

screw-cap tubes with a small amount of glass beads (212 to 300 μm) and shock frozen at −150°C for 5 min. Then, 100 μl of buffer AP1 was added, and lysis was performed with the TissueLyser II (Qiagen, Valencia, CA, USA) at 30 Hz for 30 to 60 s. An additional 300 μl of buffer AP1 was added, and the procedure was continued according to the manufacturer's instructions. The DNA was quantified using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). The PCR reactions were performed in a final volume of 25 μl with final concentrations of 1.5 mM MgCl2, 0.25 mM dNTPs, 0.2 mg ml^{-1} BSA, 0.8 μM of primers, 0.12 U of AmpliTaq (Thermo Fisher Scientific, Waltham, MA, USA), and approximately 40 ng of DNA. Initial denaturation was performed at 94°C for 4 min, followed by 39 cycles of denaturation at 94°C for 30 s, annealing at 50°C for 45 s, elongation at 72°C for 1 min, and a final elongation at 72°C for 10 min.

The two representative isolates from both lakes (Ors24 and Kgh138) were selected for purification and sequencing of the *mcp* PCR products using the MinElute PCR Purification Kit (Qiagen, Valencia, CA, USA) following the manufacturer's instructions. Sequencing was performed in an Applied Biosystems 8-capillary 3500 Genetic Analyzer using the BigDye Terminator Cycle Sequencing Kit (Thermo Fisher Scientific, Waltham, MA, USA). High-quality chromatograms were manually trimmed and aligned in Geneious 11.0.2, using the Geneious alignment algorithm. An online BLASTn search of the consensus sequence revealed matches with other viral MCP sequences of large dsDNA viruses. For phylogenetic analysis, we aligned the Ors24, Kgh138, and punuivirus MCP sequences together with a set of homologous protein sequences from known green algal GEVEs and reference nucleocytoviruses available in the Giant Virus Database (https://faylward.github.io/GVDB/) using Muscle5 v5.1 (default parameters). The tree was constructed using IQ-TREE v2.03 using the -alrt support option and the LG+G+R10 substitution model. For the Ors24 strain, we also obtained low-coverage long-read sequencing using the PacBio Revio Technology Platform with the multiplex HiFi protocol at the Uppsala Genome Center, Science for Life Laboratory. This sequencing recovered a near full-length viral family B polymerase, which we placed into a tree together with other representative viral sequences using the same method as for the MCP tree.

To screen for VLPs inclusions using TEM, Ors24 was harvested at different stages of growth, including early exponential, mid-exponential, late-exponential, and stationary phases (~2×10^5, 4×10^5, 6×10^5, and 8×10^5 cells ml^{-1}, respectively). Sample preparation was performed according to Hoops and Witman (89) with modifications. Cells were double fixed with 4% glutaraldehyde (GA) in MWC+Se media for 15 min at room temperature and then transferred to 4% GA in 100 mM sodium cacodylate (NaCac) for 4 hours at room temperature. The GA+NaCac was removed, and 100 mM of NaCac buffer was added in the dark at 4°C. Samples were postfixed in 1% osmium tetroxide water for 2 hours at 4°C, and the pellets were dehydrated in graded ethanol series and embedded in epoxy resin (Agar 100) using acetone. Semi-thin sections (1.5 μm) were made using a Leica EM UC7 ultramicrotome with a glass knife and stained with Richardson's solution (90) to examine the orientation of the tissue in the trimmed block. Ultra-thin sections (50 nm) were made using a Leica EM UC7 ultratome with a diamond knife. The sections were mounted on pioloform-coated, single-slot, copper grids and stained with uranyl acetate (2%, 30 min) and lead citrate (4 min). The grids were visualized using a JEOL JEM 1400 Plus transmission

electron microscope (Jeol, Tokyo, Japan). A total of 40 icosahedral VLPs were measured for size, with the diameter of each VLP calculated as the average of three measurements across vertices.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The genomic sequencing and proteomics work were performed by the Genomics Sequencing Center and the Mass Spectroscopy Incubator facilities, respectively, of the Fralin Life Science Institute at Virginia Tech. We thank R. Helm (Department of Biochemistry) for assistance with the proteomics and S. McCartney and H. Wang for assistance with TEM at the Nanoscale Characterization and Fabrication Lab, Virginia Tech. We are grateful to C. Abergel and J. Van Etten for providing cultures of mimivirus and PBCV-1, respectively, for their use as positive controls in our flow cytometry assays. The authors also acknowledge support from the National Genomics Infrastructure (NGI)–Uppsala Genome Center and UPPMAX for providing assistance in massive parallel sequencing and computational infrastructure. We acknowledge the Microscopy and DNA Sequencing Facility at the Department of Biology, Lund University, for assistance with TEM and Sanger sequencing of the freshwater *Chlamydomonas* strains. We thank A. Vardi and members of his lab for comments on an earlier version of this manuscript.

Funding:

This work was funded by National Institutes of Health R35 grant no. 1R35GM147290-01 (F.O.A.). The Lund University Hedda Andersson Guest Professor Program provided a visiting guest professorship (C.P.D.B.). This work was also funded by Swedish Research Council (VR) grants 2017-03860 (K.R.) and 2022-03503 (C.K.C.), Templeton Foundation grant 60501 (C.K.C.), Knut and Alice Wallenberg Foundation grant 2018.0138 (C.K.C.), Council for Research Infrastructures (RFI), and the Science for Life Laboratory (SciLifeLab), Sweden.

Data and materials availability:

The raw data for the three Oxford Nanopore sequencing runs, RNA-seq reads, and viral particle Illumina short-reads have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under BioProject number PRJNA1131777. The genome sequence of punuivirus is available in GenBank under accession number PV354230.1. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (91) partner repository with the dataset identifier PXD060118. Supplementary data associated with this study, including the assembled *C. reinhardtii* CC-2937 contigs, gene predictions, protein predictions, alignments for the phylogenetic trees, and the integrase structure prediction, are available through Zenodo (92).

REFERENCES AND NOTES

1. Feschotte C, Gilbert C, Endogenous viruses: Insights into viral evolution and impact on host biology. Nat. Rev. Genet 13, 283–296 (2012). doi: 10.1038/nrg3199 [PubMed: 22421730]
2. Katzourakis A, Gifford RJ, Endogenous viral elements in animal genomes. PLOS Genet. 6, e1001191 (2010). doi: 10.1371/journal.pgen.1001191 [PubMed: 21124940]
3. Takahashi H, Fukuhara T, Kitazawa H, Kormelink R, Virus latency and the impact on plants. Front. Microbiol 10, 2764 (2019). doi: 10.3389/fmicb.2019.02764 [PubMed: 31866963]
4. Horie M et al. , Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature 463, 84–87 (2010). doi: 10.1038/nature08695 [PubMed: 20054395]

5. Holmes EC, The evolution of endogenous viral elements. Cell Host Microbe 10, 368–377 (2011). doi: 10.1016/j.chom.2011.09.002 [PubMed: 22018237]
6. Chiba S et al. , Widespread endogenization of genome sequences of non-retroviral RNA viruses into plant genomes. PLOS Pathog. 7, e1002146 (2011). doi: 10.1371/journal.ppat.1002146 [PubMed: 21779172]
7. Bellas C et al. , Large-scale invasion of unicellular eukaryotic genomes by integrating DNA viruses. Proc. Natl. Acad. Sci. U.S.A 120, e2300465120 (2023). doi: 10.1073/pnas.2300465120 [PubMed: 37036967]
8. Moniruzzaman M, Weinheimer AR, Martinez-Gutierrez CA, Aylward FO, Widespread endogenization of giant viruses shapes genomes of green algae. Nature 588, 141–145 (2020). doi: 10.1038/s41586-020-2924-2 [PubMed: 33208937]
9. Zhao H et al. , A 1.5-Mb continuous endogenous viral region in the arbuscular mycorrhizal fungus *Rhizophagus irregularis*. Virus Evol. 9, vead064 (2023). doi: 10.1093/ve/vead064 [PubMed: 37953976]
10. Gong Z, Zhang Y, Han G-Z, Molecular fossils reveal ancient associations of dsDNA viruses with several phyla of fungi. Virus Evol. 6, veaa008 (2020). doi: 10.1093/ve/veaa008 [PubMed: 32071765]
11. Maumus F, Epert A, Nogué F, Blanc G, Plant genomes enclose footprints of past infections by giant virus relatives. Nat. Commun 5, 4268 (2014). doi: 10.1038/ncomms5268 [PubMed: 24969138]
12. Sarre LA et al. , DNA methylation enables recurrent endogenization of giant viruses in an animal relative. Sci. Adv 10, eado6406 (2024). doi: 10.1126/sciadv.ado6406 [PubMed: 38996012]
13. Dodds JA, Viruses of marine algae. Experientia 35, 440–442 (1979). doi: 10.1007/BF01922694
14. Reisser W, Viruses and virus-like particles of freshwater and marine eukaryotic algae — a review. Arch. Protistenkd 143, 257–265 (1993). doi: 10.1016/S0003-9365(11)80293-9
15. Delaroque N, Maier I, Knippers R, Müller DG, Persistent virus integration into the genome of its algal host, *Ectocarpus siliculosus* (Phaeophyceae). J. Gen. Virol 80, 1367–1370 (1999). doi: 10.1099/0022-1317-80-6-1367 [PubMed: 10374952]
16. Müller DG, Kawai H, Stache B, Lanka S, a virus infection in the marine brown alga *Ectocarpus siliculosus* (Phaeophyceae). Bot. Acta 103, 72–82 (1990). doi: 10.1111/j.1438-8677.1990.tb00129.x
17. Cock JM et al. , The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. Nature 465, 617–621 (2010). doi: 10.1038/nature09016 [PubMed: 20520714]
18. Gyaltshen Y et al. , Long-read-based genome assembly reveals numerous endogenous viral elements in the green algal bacterivore *Cymbomonas tetramitiformis*. Genome Biol. Evol 15, evad194 (2023). doi: 10.1093/gbe/evad194 [PubMed: 37883709]
19. Filée J, Multiple occurrences of giant virus core genes acquired by eukaryotic genomes: The visible part of the iceberg? Virology 466–467, 53–59 (2014). doi: 10.1016/j.virol.2014.06.004
20. Salomé PA, Merchant SS, A series of fortunate events: Introducing Chlamydomonas as a reference organism. Plant Cell 31, 1682–1707 (2019). doi: 10.1105/tpc.18.00952 [PubMed: 31189738]
21. Sasso S, Stibor H, Mittag M, Grossman AR, The natural history of model organisms: From molecular manipulation of domesticated *Chlamydomonas reinhardtii* to survival in nature. eLife 7, e39233 (2018). doi: 10.7554/eLife.39233 [PubMed: 30382941]
22. Moniruzzaman M, Erazo-Garcia MP, Aylward FO, Endogenous giant viruses contribute to intraspecies genomic variability in the model green alga *Chlamydomonas reinhardtii*. Virus Evol. 8, veac102 (2022). doi: 10.1093/ve/veac102 [PubMed: 36447475]
23. Craig RJ et al. , The Chlamydomonas Genome Project, version 6: Reference assemblies for mating-type plus and minus strains reveal extensive structural mutation in the laboratory. Plant Cell 35, 644–672 (2023). doi: 10.1093/plcell/koac347 [PubMed: 36562730]
24. Aylward FO, Moniruzzaman M, ViralRecall-A flexible command-line tool for the detection of giant virus signatures in ‘omic data. Viruses 13, 150 (2021). doi: 10.3390/v13020150 [PubMed: 33498458]

25. Aylward FO, Moniruzzaman M, Ha AD, Koonin EV, A phylogenomic framework for charting the diversity and evolution of giant viruses. PLOS Biol. 19, e3001430 (2021). doi: 10.1371/journal.pbio.3001430 [PubMed: 34705818]
26. Merchant SS et al. , The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science 318, 245–250 (2007). doi: 10.1126/science.1143609 [PubMed: 17932292]
27. López-Cortegano E, Craig RJ, Chebib J, Balogun EJ, Keightley PD, Rates and spectra of de novo structural mutations in *Chlamydomonas reinhardtii*. Genome Res. 33, 45–60 (2023). doi: 10.1101/gr.276957.122 [PubMed: 36617667]
28. Craig NL, "A moveable feast: An introduction to mobile DNA" in Mobile DNA III, Craig NL et al., Eds. (Wiley, 2015), pp. 1–39.doi: 10.1128/9781555819217.ch1
29. Kojima KK, Bao W, *IS481EU* shows a new connection between eukaryotic and prokaryotic DNA transposons. Biology 12, 365 (2023). doi: 10.3390/biology12030365 [PubMed: 36979057]
30. Kordiš D, A genomic perspective on the chromodomain-containing retrotransposons: Chromoviruses. Gene 347, 161–173 (2005). doi: 10.1016/j.gene.2004.12.017 [PubMed: 15777633]
31. Pritham EJ, Putliwala T, Feschotte C, Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene 390, 3–17 (2007). doi: 10.1016/j.gene.2006.08.008 [PubMed: 17034960]
32. Gao X, Hou Y, Ebina H, Levin HL, Voytas DF, Chromodomains direct integration of retrotransposons to heterochromatin. Genome Res. 18, 359–369 (2008). doi: 10.1101/gr.7146408 [PubMed: 18256242]
33. Fischer MG, Hackl T, Host genome integration and giant virus-induced reactivation of the virophage mavirus. Nature 540, 288–291 (2016). doi: 10.1038/nature20593 [PubMed: 27929021]
34. Brussaard CPD, Optimization of procedures for counting viruses by flow cytometry. Appl. Environ. Microbiol 70, 1506–1513 (2004). doi: 10.1128/AEM.70.3.1506-1513.2004 [PubMed: 15006772]
35. Yoon PH et al. , Eukaryotic RNA-guided endonucleases evolved from a unique clade of bacterial enzymes. Nucleic Acids Res. 51, 12414–12427 (2023). doi: 10.1093/nar/gkad1053 [PubMed: 37971304]
36. Gann ER et al. , Structural and proteomic studies of the Aureococcus anophagefferens virus demonstrate a global distribution of virus-encoded carbohydrate processing. Front. Microbiol 11, 2047 (2020). doi: 10.3389/fmicb.2020.02047 [PubMed: 33013751]
37. Fischer MG, Kelly I, Foster LJ, Suttle CA, The virion of *Cafeteria roenbergensis* virus (CroV) contains a complex suite of proteins for transcription and DNA repair. Virology 466–467, 82–94 (2014). doi: 10.1016/j.virol.2014.05.029
38. Dunigan DD et al. , *Paramecium bursaria* chlorella virus 1 proteome reveals novel architectural and regulatory features of a giant virus. J. Virol 86, 8821–8834 (2012). doi: 10.1128/JVI.00907-12 [PubMed: 22696644]
39. Birkholz EA et al. , An intron endonuclease facilitates interference competition between coinfecting viruses. Science 385, 105–112 (2024). doi: 10.1126/science.adl1356 [PubMed: 38963841]
40. Goodrich-Blair H, Shub DA, Beyond homing: Competition between intron endonucleases confers a selective advantage on flanking genetic markers. Cell 84, 211–221 (1996). doi: 10.1016/S0092-8674(00)80976-9 [PubMed: 8565067]
41. Yau S et al. , Virus-host coexistence in phytoplankton through the genomic lens. Sci. Adv 6, eaay2587 (2020). doi: 10.1126/sciadv.aay2587 [PubMed: 32270031]
42. Joffe N et al. , Cell-to-cell heterogeneity drives host-virus coexistence in a bloom-forming alga. ISME J. 18, wrae038 (2024). doi: 10.1093/ismejo/wrae038 [PubMed: 38452203]
43. Blanc-Mathieu R et al. , A persistent giant algal virus, with a unique morphology, encodes an unprecedented number of genes involved in energy metabolism. J. Virol 95, e02446–e20 (2021). doi: 10.1128/JVI.02446-20 [PubMed: 33536167]
44. Harris EH, *Chlamydomonas* as a model organism. Annu. Rev. Plant Physiol. Plant Mol. Biol 52, 363–406 (2001). doi: 10.1146/annurev.arplant.52.1.363 [PubMed: 11337403]

45. Collier JL et al. , The protist *Aurantiochytrium* has universal subtelomeric rDNAs and is a host for mirusviruses. Curr. Biol 33, 5199–5207.e4 (2023). doi: 10.1016/j.cub.2023.10.009 [PubMed: 37913769]
46. Zhao H, Meng L, Hikida H, Ogata H, Eukaryotic genomic data uncover an extensive host range of mirusviruses. Curr. Biol 34, 2633–2643.e3 (2024). doi: 10.1016/j.cub.2024.04.085 [PubMed: 38806056]
47. Kolmogorov M, Yuan J, Lin Y, Pevzner PA, Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol 37, 540–546 (2019). doi: 10.1038/s41587-019-0072-8 [PubMed: 30936562]
48. Flowers JM et al. , Whole-genome resequencing reveals extensive natural variation in the model green alga *Chlamydomonas reinhardtii*. Plant Cell 27, 2353–2369 (2015). doi: 10.1105/tpc.15.00492 [PubMed: 26392080]
49. Payne ZL, Penny GM, Turner TN, Dutcher SK, A gap-free genome assembly of *Chlamydomonas reinhardtii* and detection of translocations induced by CRISPR-mediated mutagenesis. Plant Commun. 4, 100493 (2023). doi: 10.1016/j.xplc.2022.100493 [PubMed: 36397679]
50. Li H, Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100 (2018). doi: 10.1093/bioinformatics/bty191 [PubMed: 29750242]
51. Camacho C et al. , BLAST+: Architecture and applications. BMC Bioinformatics 10, 421 (2009). doi: 10.1186/1471-2105-10-421 [PubMed: 20003500]
52. Benson G, Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580 (1999). doi: 10.1093/nar/27.2.573 [PubMed: 9862982]
53. Hyatt D et al. , Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119 (2010). doi: 10.1186/1471-2105-11-119 [PubMed: 20211023]
54. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J, eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol. Biol. Evol 38, 5825–5829 (2021). doi: 10.1093/molbev/msab293 [PubMed: 34597405]
55. Steinegger M et al. , HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 20, 473 (2019). doi: 10.1186/s12859-019-3019-7 [PubMed: 31521110]
56. Remmert M, Biegert A, Hauser A, Söding J, HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods 9, 173–175 (2011). doi: 10.1038/nmeth.1818 [PubMed: 22198341]
57. Zimmermann L et al. , A completely reimplemented MPI Bioinformatics Toolkit with a new HHpred server at its core. J. Mol. Biol 430, 2237–2243 (2018). doi: 10.1016/j.jmb.2017.12.007 [PubMed: 29258817]
58. Jumper J et al. , Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). doi: 10.1038/s41586-021-03819-2 [PubMed: 34265844]
59. Rigou S et al., Nucleocytoviricota viral factories are transient organelles made by phase separationbioRxiv 2024.09.01.610734 [Preprint] (2024); https://doi.org/10.1101/2024.09.01.610734.doi: 10.1101/2024.09.01.610734
60. Starrett GJ et al. , Adintoviruses: A proposed animal-tropic family of midsize eukaryotic linear dsDNA (MELD) viruses. Virus Evol. 7, veaa055 (2021). doi: 10.1093/ve/veaa055 [PubMed: 34646575]
61. Paez-Espino D et al. , Diversity, evolution, and classification of virophages uncovered through global metagenomics. Microbiome 7, 157 (2019). doi: 10.1186/s40168-019-0768-5 [PubMed: 31823797]
62. Stephens D, Faghihi Z, Moniruzzaman M, Widespread occurrence and diverse origins of polintoviruses influence lineage-specific genome dynamics in stony corals. Virus Evol. 10, veae039 (2024). doi: 10.1093/ve/veae039 [PubMed: 38808038]
63. Edgar RC, Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. Nat. Commun 13, 6968 (2022). doi: 10.1038/s41467-022-34630-w [PubMed: 36379955]
64. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973 (2009). doi: 10.1093/bioinformatics/btp348 [PubMed: 19505945]

65. Minh BQ et al. , Corrigendum to: IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol 37, 2461 (2020). doi: 10.1093/molbev/msaa015 [PubMed: 32556291]
66. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS, ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587–589 (2017). doi: 10.1038/nmeth.4285 [PubMed: 28481363]
67. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS, UFBoot2: Improving the ultrafast bootstrap approximation. Mol. Biol. Evol 35, 518–522 (2018). doi: 10.1093/molbev/msx281 [PubMed: 29077904]
68. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD, rnaSPAdes: A de novo transcriptome assembler and its application to RNA-seq data. Gigascience 8, giz100 (2019). doi: 10.1093/gigascience/giz100 [PubMed: 31494669]
69. Frith MC, A new repeat-masking method enables specific detection of homologous sequences. Nucleic Acids Res. 39, e23 (2011). doi: 10.1093/nar/gkq1212 [PubMed: 21109538]
70. Kent WJ, BLAT—The BLAST-like alignment tool. Genome Res. 12, 656–664 (2002). [PubMed: 11932250]
71. Keller O, Kollmar M, Stanke M, Waack S, A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics 27, 757–763 (2011). doi: 10.1093/bioinformatics/btr010 [PubMed: 21216780]
72. Love MI, Huber W, Anders S, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014). doi: 10.1186/s13059-014-0550-8 [PubMed: 25516281]
73. Wehrens R, Buydens LMC, Self- and super-organizing maps in R: The kohonen package. J. Stat. Softw 21, 1–19 (2007). doi: 10.18637/jss.v021.i05
74. Hadi SIIA et al. , DNA barcoding green microalgae isolated from neotropical inland waters. PLOS ONE 11, e0149284 (2016). doi: 10.1371/journal.pone.0149284 [PubMed: 26900844]
75. White TJ, Bruns T, Lee S, Taylor J, “Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics” in PCR Protocols: A Guide to Methods and Applications, Innis MA, Gelfand DH, Sninsky JJ, White TJ, Eds. (Academic Press, 1990), pp. 315–322.
76. Brussaard CPD, Payet JP, Winter C, Weinbauer MG, “Quantification of aquatic viruses by flow cytometry” in Manual of Aquatic Viral Ecology, Wilhelm S, Weinbauer M, Suttle C, Eds. (ASLO Waco, 2010), pp. 102–109. doi: 10.4319/mave.2010.978-0-9845591-0-7.102
77. Brussaard CPD, Marie D, Bratbak G, Flow cytometric detection of viruses. J. Virol. Methods 85, 175–182 (2000). doi: 10.1016/S0166-0934(99)00167-6 [PubMed: 10716350]
78. Zucker RM, Ortenzio JNR, Boyes WK, Characterization, detection, and counting of metal nanoparticles using flow cytometry. Cytometry A 89, 169–183 (2016). doi: 10.1002/cyto.a.22793 [PubMed: 26619039]
79. Zhao Y et al. , Enhanced resolution of marine viruses with violet side scatter. Cytometry A 103, 260–268 (2023). doi: 10.1002/cyto.a.24674 [PubMed: 35929601]
80. Wood DE, Salzberg SL, Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15, R46 (2014). doi: 10.1186/gb-2014-15-3-r46 [PubMed: 24580807]
81. Li H et al. , The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009). doi: 10.1093/bioinformatics/btp352 [PubMed: 19505943]
82. Olm MR et al. , inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat. Biotechnol 39, 727–736 (2021). doi: 10.1038/s41587-020-00797-0 [PubMed: 33462508]
83. Schneider CA, Rasband WS, Eliceiri KW, NIH Image to ImageJ: 25 years of image analysis. Nat. Methods 9, 671–675 (2012). doi: 10.1038/nmeth.2089 [PubMed: 22930834]
84. Guillard RRL, Lorenzen CJ, Yellow-green algae with chlorophyllide C. J. Phycol 8, 10–14 (1972).
85. Cornwallis CK et al. , Single-cell adaptations shape evolutionary transitions to multicellularity in green algae. Nat. Ecol. Evol 7, 889–902 (2023). doi: 10.1038/s41559-023-02044-6 [PubMed: 37081145]

86. Guillou L et al. , The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic Acids Res. 41, D597–D604 (2013). doi: 10.1093/nar/gks1160 [PubMed: 23193267]

87. Fu L, Niu B, Zhu Z, Wu S, Li W, CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152 (2012). doi: 10.1093/bioinformatics/bts565 [PubMed: 23060610]

88. Larsen JB, Larsen A, Bratbak G, Sandaa R-A, Phylogenetic analysis of members of the Phycodnaviridae virus family, using amplified fragments of the major capsid protein gene. Appl. Environ. Microbiol 74, 3048–3057 (2008). doi: 10.1128/AEM.02548-07 [PubMed: 18359826]

89. Hoops HJ, Witman GB, Outer doublet heterogeneity reveals structural polarity related to beat direction in Chlamydomonas flagella. J. Cell Biol 97, 902–908 (1983). doi: 10.1083/jcb.97.3.902 [PubMed: 6224802]

90. Richardson KC, Jarett L, Finke EH, Embedding in epoxy resins for ultrathin sectioning in electron microscopy. Stain Technol. 35, 313–323 (1960). doi: 10.3109/10520296009114754 [PubMed: 13741297]

91. Perez-Riverol Y et al. , The PRIDE database at 20 years: 2025 update. Nucleic Acids Res. 53, D543–D553 (2025). doi: 10.1093/nar/gkae1011 [PubMed: 39494541]

92. Erazo-Garcia MP, Aylward F, Latent infection of an active giant endogenous virus in a unicellular green alga, Version 2, Zenodo (2025); 10.5281/zenodo.13645514.

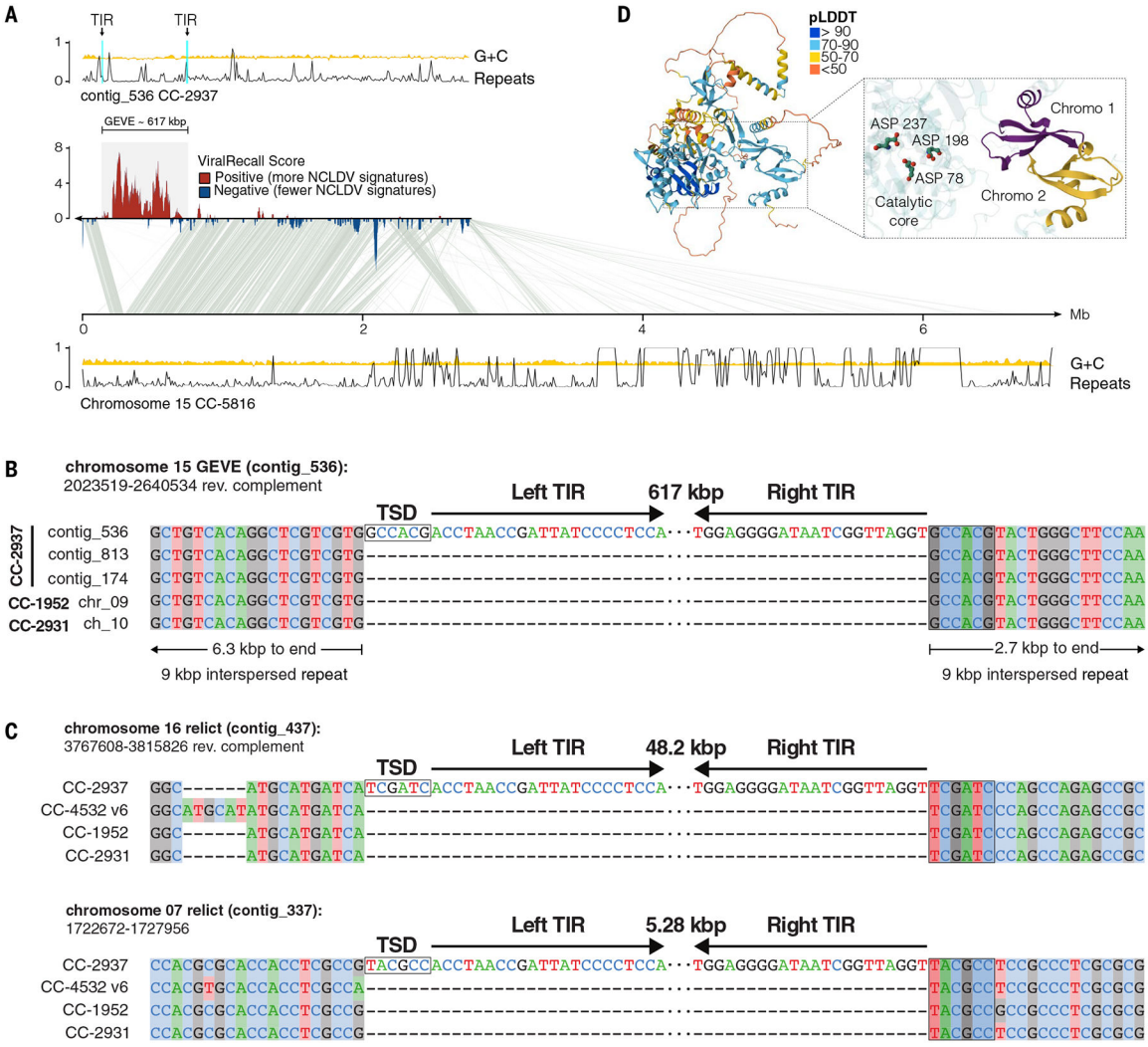


Fig. 1. Features of the *C. reinhardtii* CC-2937 GEVE and its insertion site.

(**A**) Synteny plot of *C. reinhardtii* CC-2937 contig_536 mapped against chromosome 15 of *C. reinhardtii* CC-5816. Line plots are depicted for both sequences and show the tandem repeats (black) and GC fractions (yellow) along the sequence. TIRs flanking the GEVE are marked with arrows. Regions of contig_536 enriched in nucleocytovirus (NCLDV) signatures are shown as ViralRecall scores >0 (red), whereas regions with fewer NCLDV signatures have scores <0 (blue). (**B**) Alignment of five independent copies of the interspersed repetitive element into which the GEVE is integrated. Only the ends of the TIRs are represented, and the 6-bp TSD is highlighted with a box. (**C**) Integration sites of two putative GEVE relicts on chromosomes 16 and 7. The TIRs and TSDs of the insertions in CC-2937 are shown relative to three divergent strains that do not carry the insertions. Coordinates for each of the presented viral insertions in *C. reinhardtii* CC-2937 are provided. (**D**) Predicted protein structure of the putative GEVE integrase (GEVE_506) colored by the per-residue model confidence score (pLDDT). The enlarged panel highlights two chromodomains (Chromo) and the DD(E/D) catalytic core, consisting of three aspartate residues (ASP).

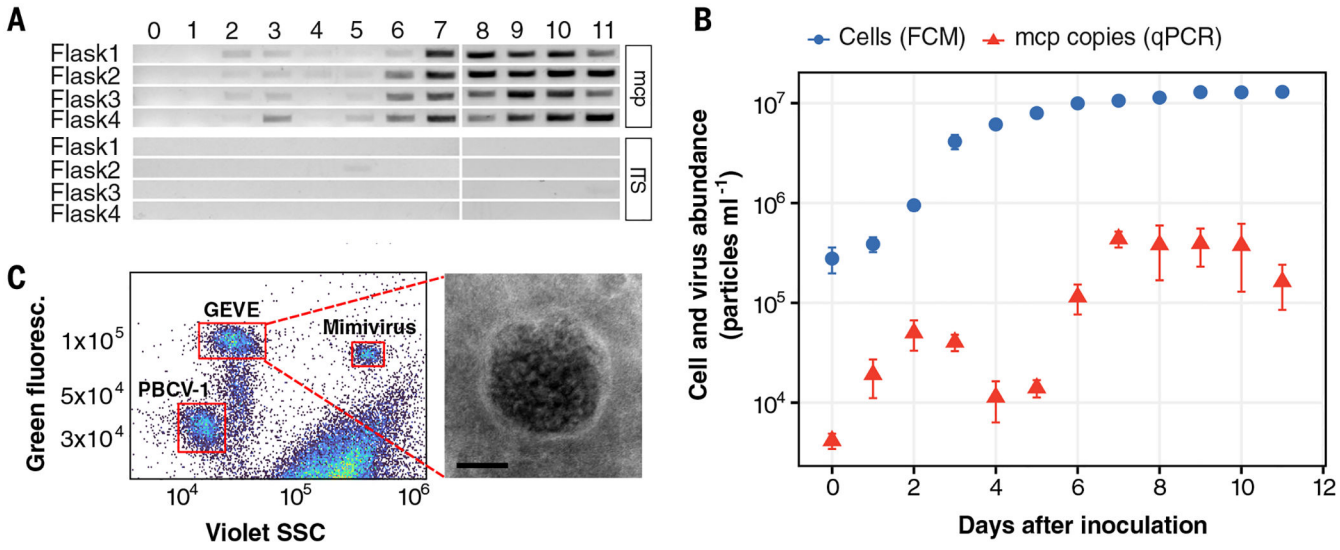


Fig. 2. Late-exponential *C. reinhardtii* CC-2937 cultures show evidence of virion production.

(**A**) Gel image of the PCR assay targeting the viral major capsid protein (*mcp*) gene and host internal transcribed spacer (ITS) region. The assay was performed on DNase-treated supernatants from four biological replicates ($n = 4$) sampled daily for 11 days. (**B**) Quantification of viral and host abundances over 11 days ($n = 4$). Data represent mean ± SD. Viral abundance in supernatants was measured by qPCR in triplicate (red), whereas host cell density was assessed by flow cytometry (FCM) (blue). (**C**) Flow cytometry analysis (left) of concentrated supernatants alongside two positive controls of known large DNA viruses mixed with the sample. The right panel displays an electron micrograph of the concentrated viral fraction from a 9-day-old culture, showing a VLP identified by negative staining. Scale bar is 100 nm.

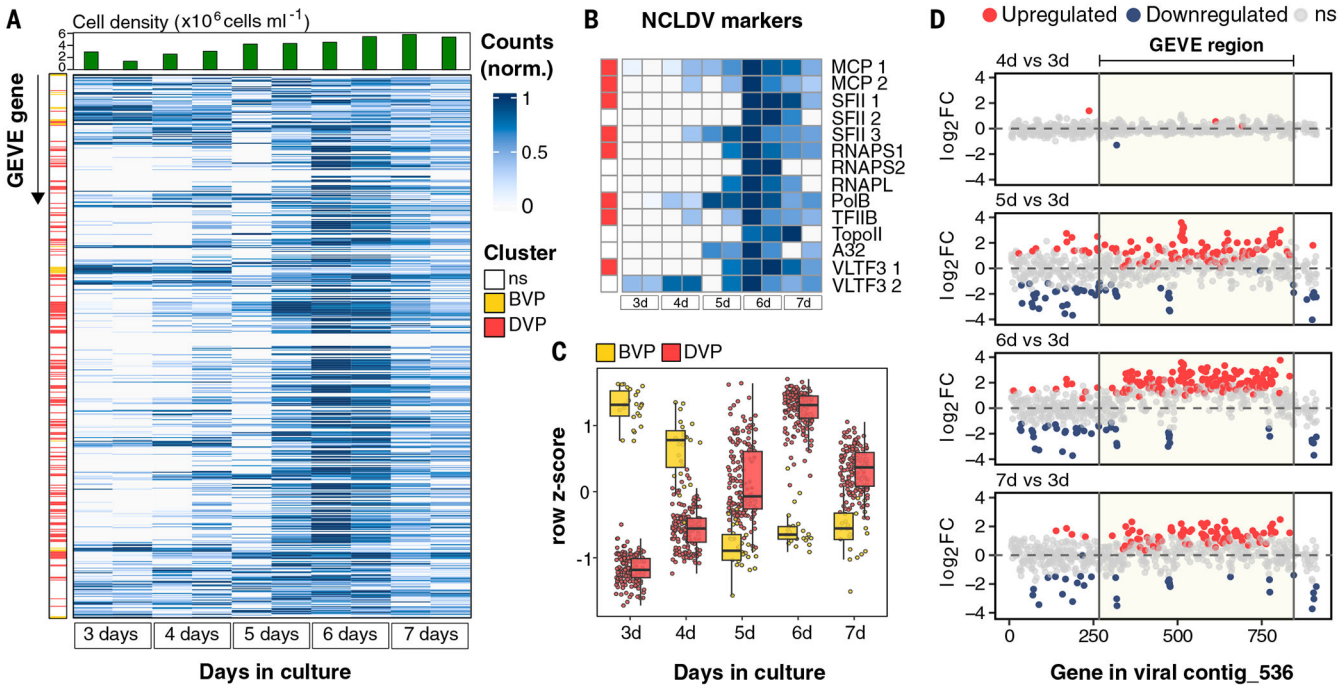


Fig. 3. RNA-seq analysis of *C. reinhardtii* CC-2937 cultures at different growth phases.

(**A**) Heatmap showing minimum-maximum normalized transcript counts for each GEVE gene across cultures harvested 3 to 7 days after inoculation. Rows represent individual GEVE genes, ordered according to their genomic position along viral contig_536. Each pair of columns represents biological replicates for a specific time point ($n = 2$ per time point). Row annotations indicate gene clusters based on expression patterns (see Materials and methods): before peak viral production (BVP, yellow) and during peak viral production (DVP, red). Nonsignificantly differentially expressed genes (ns) were excluded from the clustering analysis. Column annotations display the average cell count on the day of sampling (top panel). (**B**) Heatmap showing minimum-maximum normalized transcript counts for hallmark nucleocytovirus markers (25), with row annotations indicating assigned clusters. (**C**) Box plot showing the expression patterns of differentially expressed genes identified with DESeq2 (Wald test, $\alpha = 0.05$; log2 fold-change threshold ≥0.585; adjusted *p* value <0.05). Colors correspond to assigned clusters from self-organizing maps analysis. The center line represents the median, box limits are upper and lower quartiles, and whiskers are minimum and maximum values. (**D**) Shrunken $\log_2$ fold-change ($\log_2$FC) of genes along the viral contig; day three serves as the reference for all comparisons. Each dot represents a gene in its genomic positions along contig_536. The dashed horizontal line marks no change in gene expression ($\log_2$FC = 0).

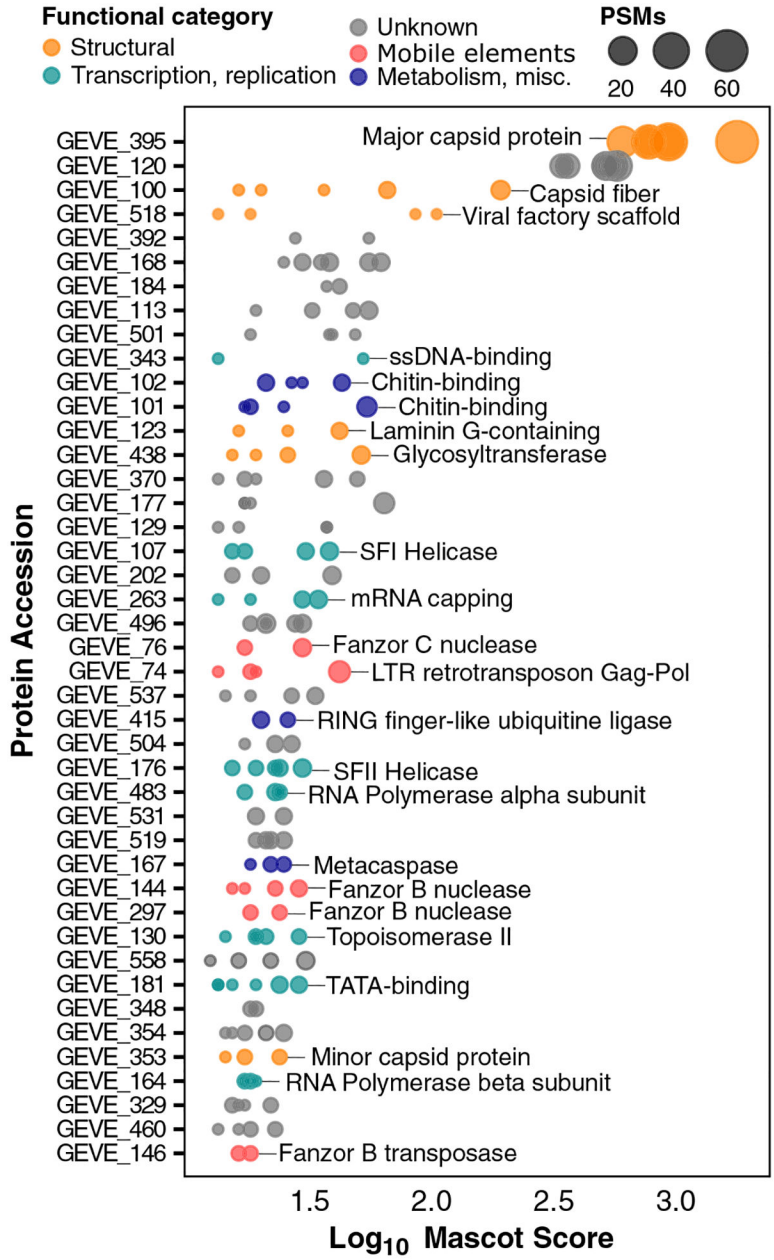


Fig. 4. Proteomic analysis of GEVE virions purified from concentrated *C. reinhardtii* CC-2937 supernatants.

Peptides were identified using LC-MS/MS across three biological replicates ($n = 3$), each analyzed in duplicate. Each dot represents a protein identified in a technical replicate, with dot size indicating the number of peptide-spectrum matches (PSMs) as a proxy for relative protein abundance and dot color denoting the assigned functional category. Proteins are arranged based on their Mascot database search scores.

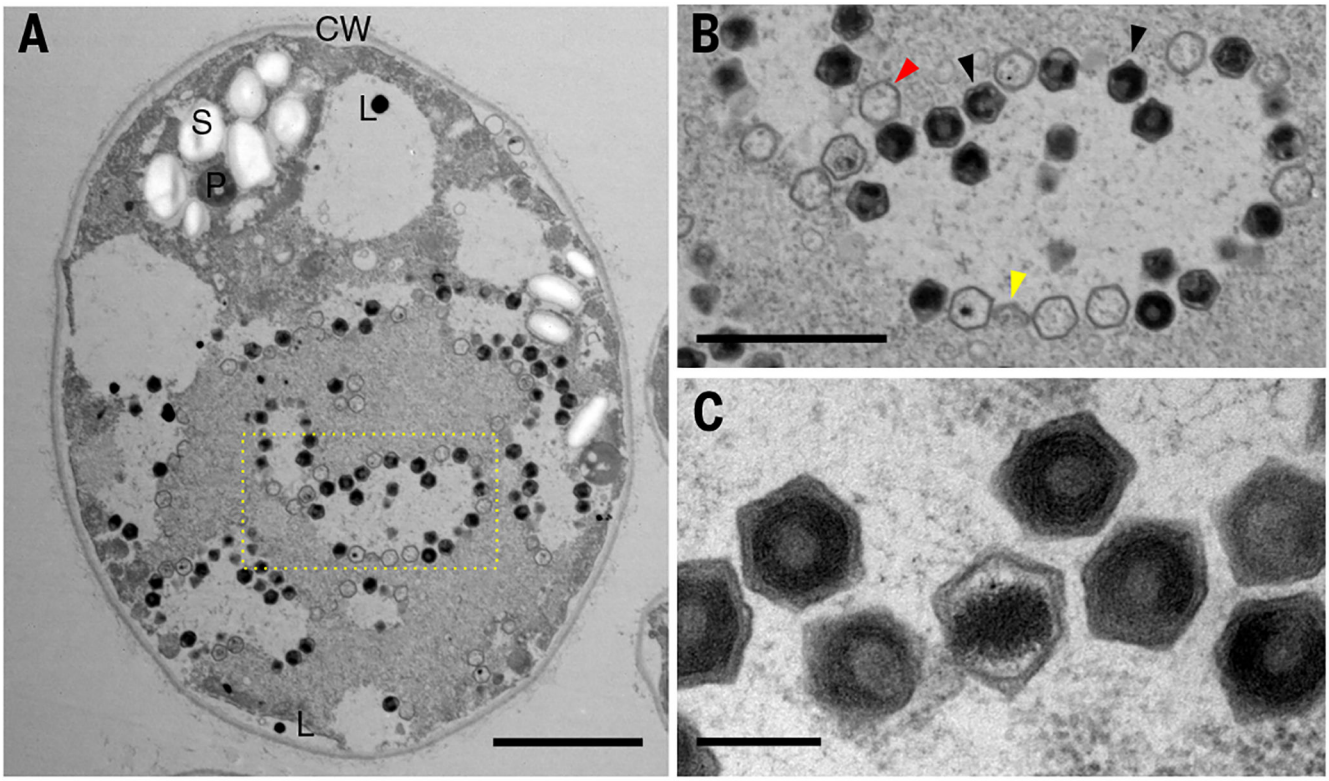


Fig. 5. TEM of ultrathin-sectioned Ors 24 *Chlamydomonas* sp. cells.

(**A**) Infected *Chlamydomonas* sp. cell in early exponential phase. The nucleus and chloroplast are not clearly distinguishable. CW, cell wall; L, lipid vesicle or plastoglobule; P, pyrenoid; S, starch sheath. Scale bar is 2 μm. (**B**) Enlarged view of the boxed region in (A), showing hexagonal viral particles. Virion production is observed in a clearly delineated, lighter colored area with virions in later stages of completion accumulating at the edges of the production area, that is, the virus factory or viroplasm. Red arrows mark empty capsids, black arrows mark full capsids, and yellow arrows mark partially assembled capsids. Scale bar is 1 μm. (**C**) Magnified view of assembled virions. Scale bar is 200 nm.