Article

Widespread endogenization of giant viruses shapes genomes of green algae

https://doi.org/10.1038/s41586-020-2924-2

Received: 7 June 2019

Accepted: 1 September 2020

Published online: 18 November 2020

Check for updates

Mohammad Moniruzzaman[1], Alaina R. Weinheimer[1], Carolina A. Martinez-Gutierrez[1] & Frank O. Aylward[1]✉

Endogenous viral elements (EVEs)—viruses that have integrated their genomes into those of their hosts—are prevalent in eukaryotes and have an important role in genome evolution[1,2]. The vast majority of EVEs that have been identified to date are small genomic regions comprising a few genes[2], but recent evidence suggests that some large double-stranded DNA viruses may also endogenize into the genome of the host[1]. Nucleocytoplasmic large DNA viruses (NCLDVs) have recently become of great interest owing to their large genomes and complex evolutionary origins[3–6], but it is not yet known whether they are a prominent component of eukaryotic EVEs. Here we report the widespread endogenization of NCLDVs in diverse green algae; these giant EVEs reached sizes greater than 1 million base pairs and contained as many as around 10% of the total open reading frames in some genomes, substantially increasing the scale of known viral genes in eukaryotic genomes. These endogenized elements often shared genes with host genomic loci and contained numerous spliceosomal introns and large duplications, suggesting tight assimilation into host genomes. NCLDVs contain large and mosaic genomes with genes derived from multiple sources, and their endogenization represents an underappreciated conduit of new genetic material into eukaryotic lineages that can substantially impact genome composition.

Much research has been devoted to the study of small EVEs in eukaryotic genomes that derive from retroviruses[1], but our knowledge of the prevalence of larger EVEs that originate from double-stranded DNA (dsDNA) viruses remains relatively unexplored. In this study, we assess the incidence of large EVEs that derive from NCLDVs, a diverse group of eukaryotic viruses that include the largest viruses characterized to date. We developed a bioinformatic approach to identify large NCLDV-derived EVEs in eukaryotic genomes (Extended Data Fig. 1; Methods) and used it to assess the incidence of NCLDV integration in the phylum Chlorophyta, a diverse group of green algae closely related to land plants[7]. Interactions with NCLDVs are known to occur in several ecologically important chlorophytes such as *Chlorella*, *Micromonas* and *Ostreococcus*[6,8], although the breadth of such interactions across different chlorophyte lineages is not well understood.

We surveyed 65 publicly available genomes spanning six classes within the Chlorophyta, and in 24 of these, we identified genomic signatures of NCLDVs (Fig. 1b, Extended Data Fig. 2, Extended Data Table 1, Supplementary Data 1). Widely known as 'giant viruses', NCLDVs are notable for their large genomes that often exceed several hundred kilobases (kb) and encode diverse functional repertoires involved in virion production and modulation of host metabolism[3–5,9]. We used multiple metrics to identify viral genomic loci, including the occurrence of viral hallmark genes, enrichment of viral proteins, homology to known NCLDVs and nucleotide composition, and we ultimately identified 18 giant EVEs (GEVEs) that ranged in size from 78 to 1,925 kb and can be traced to individual viruses (Fig. 1a, Extended Data Table 2, Supplementary Data 2; see details in the Methods). GEVEs were present in 12 of the algal genomes that we surveyed, with several genomes containing more than one GEVE (Figs. 1, 2a). The presence of complete or nearly complete sets of NCLDV hallmark genes with congruent phylogenetic signals in 14 of the 18 GEVEs allowed for their classification into the *Phycodnaviridae* (4) and *Mimiviridae* (10) families (Fig. 1a). Although four GEVEs lacked these hallmark genes and remained unclassified at the family level, their origin from within the NCLDV can still be ascertained by their clustering together with known NCLDVs based on overall gene content analysis (Extended Data Fig. 3), the homology of encoded proteins to other *Mimiviridae* and *Phycodnaviridae*, the presence of the NCLDV-specific D5 helicase/primase (Supplementary Fig. 1e), and the nucleotide composition that is consistent within GEVEs but distinct from the host genomic regions. In addition, we identified viral hallmark genes in another 12 chlorophyte genomes for which GEVEs could not be recovered (Extended Data Table 1), indicating that these genomes contained more fragmented signatures of past NCLDV integration. Overall, we identified signatures of NCLDV endogenization ranging from complete NCLDV genomes to sets of hallmark genes in 37% of the genomes surveyed.

The GEVEs contain between 76 and 1,782 predicted genes, consistent with the large size and diverse genomic repertoires of NCLDVs[10–12], and therefore represent a large and underexplored reservoir of viral genes in algal genomes (Extended Data Table 2). Remarkably, the *Tetrabaena socialis* and *Chlamydomonas eustigma* genomes both contain two large GEVEs (428–1,925 kb; Fig. 1a, b) that derive from distinct NCLDVs based on their complete set of NCLDV hallmark genes, unique tetranucleotide signatures and low

[1]Department of Biological Sciences, Virginia Tech, Blacksburg, VA, USA. ✉e-mail: faylward@vt.edu

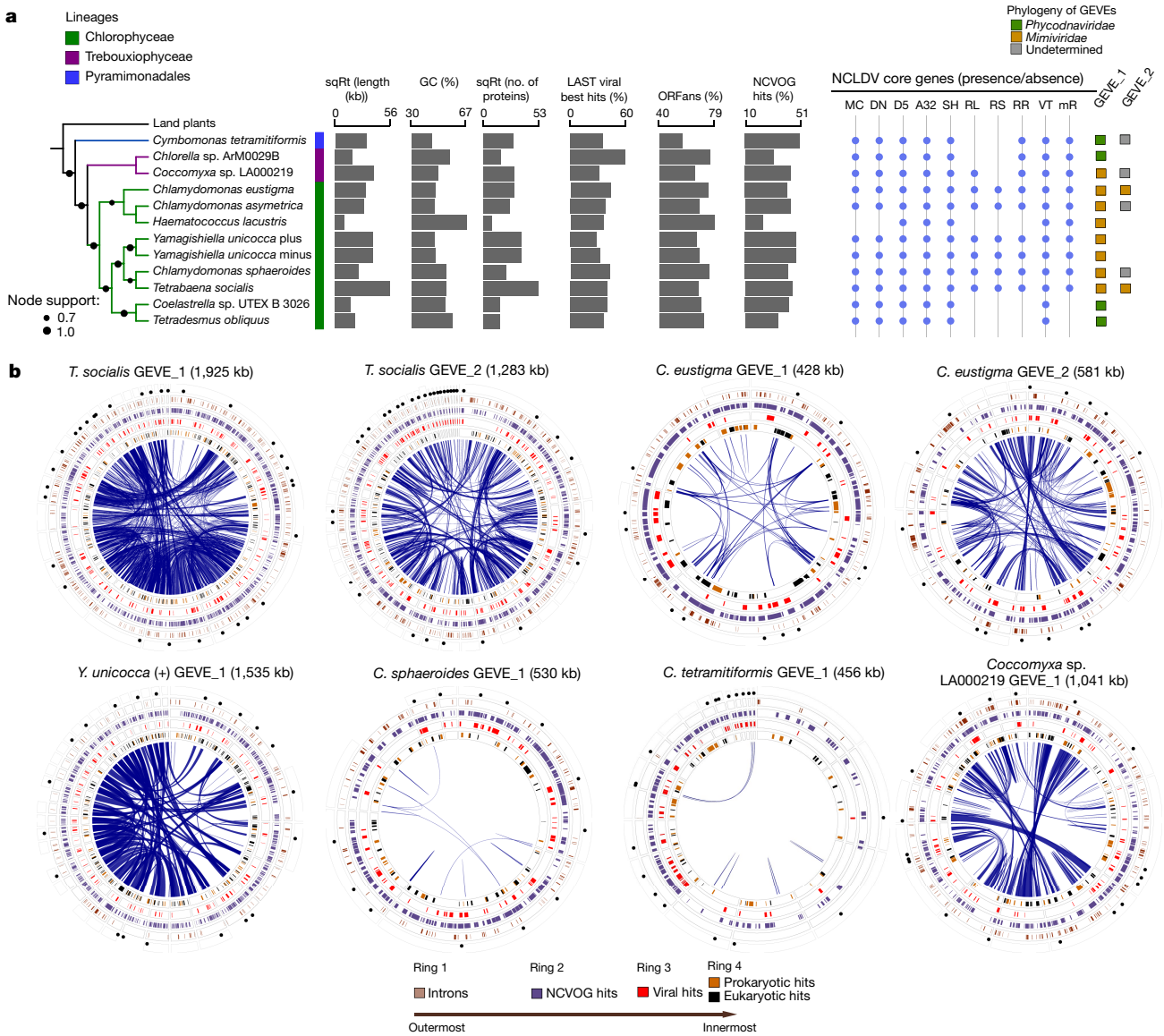


Fig. 1 | Distribution and general features of the GEVEs. a, Phylogenetic tree of GEVE-containing chlorophytes and the main features of select GEVEs. The length and total number of proteins are square-root (sqRt) converted. Several chlorophyte genomes contain more than one GEVE; in these cases, the bar plots for these represent the combined values for both of the GEVEs. A32, A32-like virion packaging ATPase; D5, D5 helicase/primase; DN, DNA polymerase; GC, guanine + cytosine content; MC, major capsid protein; mR, mRNA capping enzyme; RL, RNA polymerase large subunit; RR, ribonucleotide reductase; RS, RNA polymerase small subunit; SH, superfamily II helicase; VT, VLTF3-like transcription factor. **b**, Circular genome plots of eight representative GEVEs showing nucleocytoplasmic virus orthologous group (NCVOG) hidden Markov model (HMM) hits (see Methods), the location of spliceosomal introns and the best LAST hit matches. The black dots above the outermost track mark the locations of the core genes, while the blue links inside the circles delineate the duplicated regions. Genome plots for the rest of the GEVEs are presented in Extended Data Fig. 2.

amino acid identity to each other (46–62%) (Supplementary Fig. 2c, h, Supplementary Data 3, Supplementary Discussion). In general, GEVEs contain substantially higher coding density than the rest of the eukaryotic genomes in which they are integrated, and thereby contribute a disproportionately large number of genes to overall genomic inventories. *T. socialis* is a particularly extreme example in which two GEVEs contain a total of 2,846 genes and represent around 10% of the total coding potential of the genome (Extended Data Fig. 4, Supplementary Table 1). Similarly, in four other chlorophyte hosts, 6–7.5% of the coding sequences are contributed by the GEVEs (Supplementary Table 1). Many of the viral genes that we identified in the GEVEs were not identified with standard eukaryotic gene prediction methods alone (Extended Data Fig. 4), indicating that GEVEs represent a previously unrecognized reservoir of genomic novelty in chlorophyte genomes.

Many of the GEVEs showed signs of segmental duplications and gene loss, indicating that genomic rearrangements had taken place since endogenization (Fig. 1b, Extended Data Fig. 2). Several of the GEVEs, including those in *T. socialis*, *Yamagishiella*, *C. eustigma*, *Coccomyxa* and *Tetradesmus obliquus*, contained large proportions of duplicated regions (Fig. 1b, Extended Data Fig. 5) exemplified by multiple nearly identical copies of NCLDV hallmark genes that were typically present in only one copy in free viruses (Fig. 3, Extended Data Fig. 5, Supplementary Fig. 3). Moreover, a comparison of duplicated regions in GEVEs to reference NCLDV genomes revealed that most had a significantly higher level of duplications (Extended Data Fig. 5). Four of the GEVEs also notably lacked nine of the ten NCLDV hallmark genes that we analysed (retaining only the D5 helicase/primase), and the GEVE in *Haematococcus* was also missing several (Fig. 1a, Extended Data Fig. 5), indicating that some GEVEs have experienced gene loss. Together, this suggests

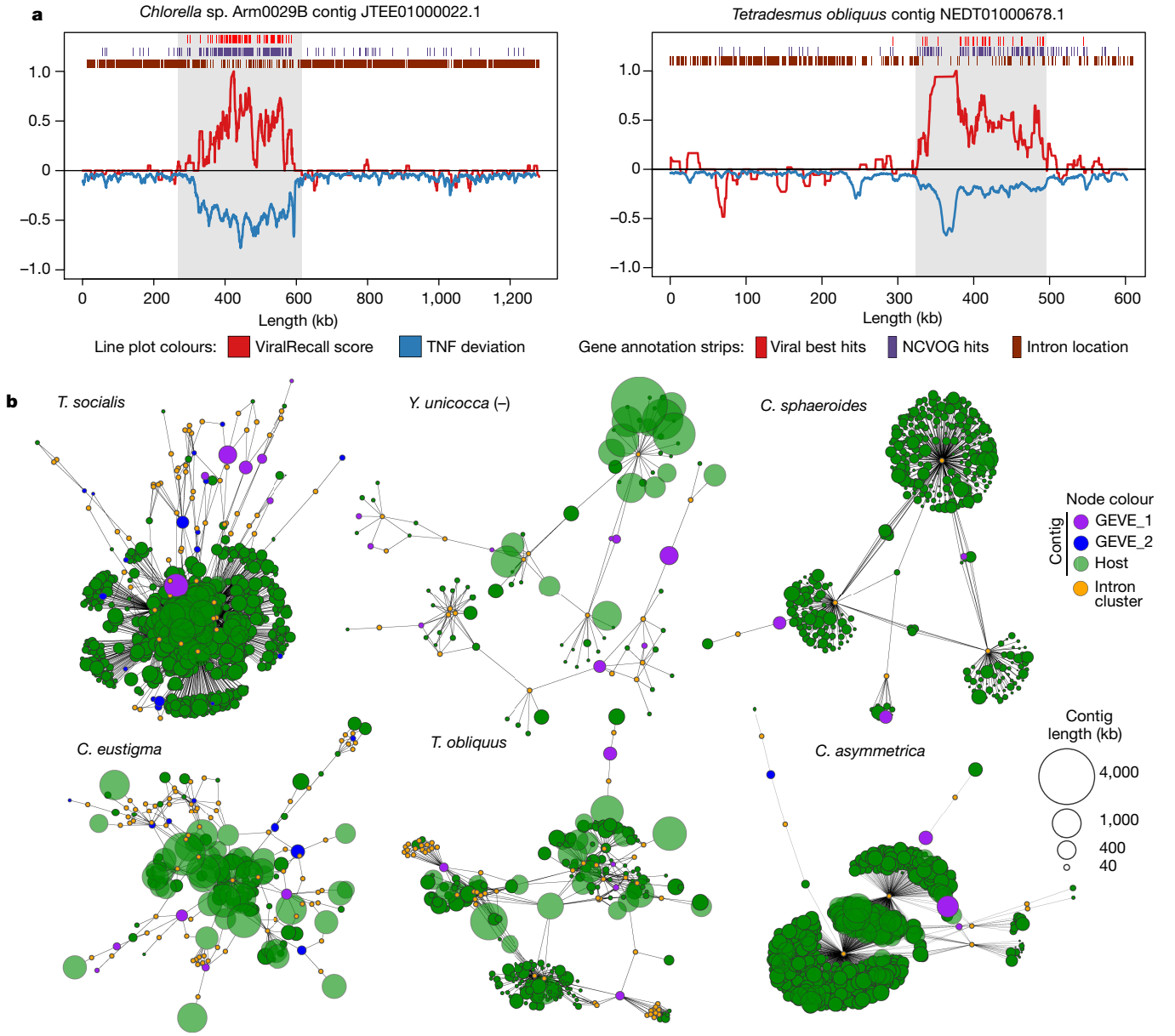
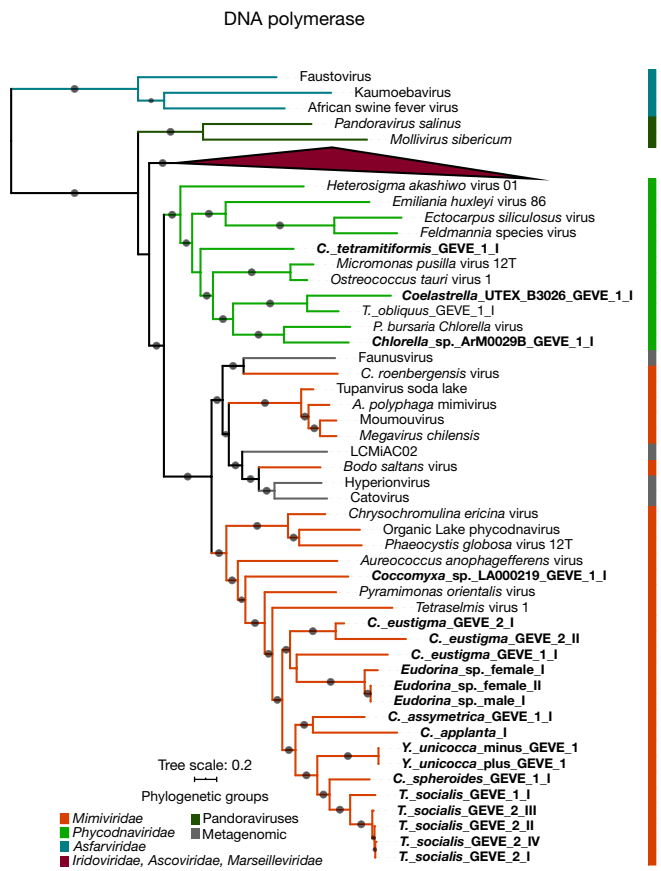


Fig. 2 | Signatures of endogenization. a, High similarity to known viruses and strong deviation from genome-averaged tetranucleotide frequency (TNF) delineate the viral regions integrated in eukaryotic contigs (see Methods for details). Viral regions are shaded in grey. Viral regions show distinctive signatures, including the presence of genes with best hits to viruses, a high proportion of NCVOGs and biased intron density compared to surrounding regions. Example contigs from *Chlorella* sp. ArM0029B-specific (left) and *T. obliquus*-specific (right) GEVEs are shown. **b**, Bipartite networks of intron sharing between host-specific and GEVE-specific contigs in selected chlorophytes. For each of these chlorophytes, only the largest networks are shown. Each intron cluster (orange nodes) contains homologous introns from both host (>10 kb in length) and GEVE contigs. The rest of the intron sharing networks from all of the chlorophytes are available in Supplementary Fig. 4.

that GEVEs are highly dynamic and have experienced varying degrees of duplications, rearrangements and deletions. In addition, we identified numerous transposons in these GEVEs (Supplementary Data 4), consistent with previous findings that mobile elements are abundant in NCLDV genomes[10,13]. These mobile elements could provide sites for homologous recombination within GEVEs or between GEVEs and host genomic loci, thereby potentially contributing to the plasticity observed in these viral regions.

Several lines of evidence confirm that the GEVEs are endogenized components of the host genome. First, although the majority of the chlorophyte genomes that we analysed are in draft status, some contigs exhibited clear signs of integration; one of the contigs in the *T. obliquus* and the sole region representing the *Chlorella* sp. ArM0029B GEVE contained clear boundaries between viral and eukaryotic regions (Fig. 2a), demonstrating their integration into the host genome. Second, in all 18 of the GEVEs, we identified numerous spliceosomal introns (Extended Data Table 2, Supplementary Table 2), which are absent in reference *Mimiviridae* genomes and have only been reported in two genes of Chloroviruses (members of the *Phycodnaviridae*) to date (see Supplementary Discussion). By contrast, the introns that we detected were found in various GEVE genes, including NCLDV hallmark genes (Supplementary Table 3). This is consistent with previous studies that have shown that genes horizontally transferred into eukaryotes frequently acquire introns[14,15]. In 17 GEVEs, we also found homologous introns that were shared between GEVE and host genomic loci (Supplementary Table 2), and intron occurrence networks revealed extensive intron



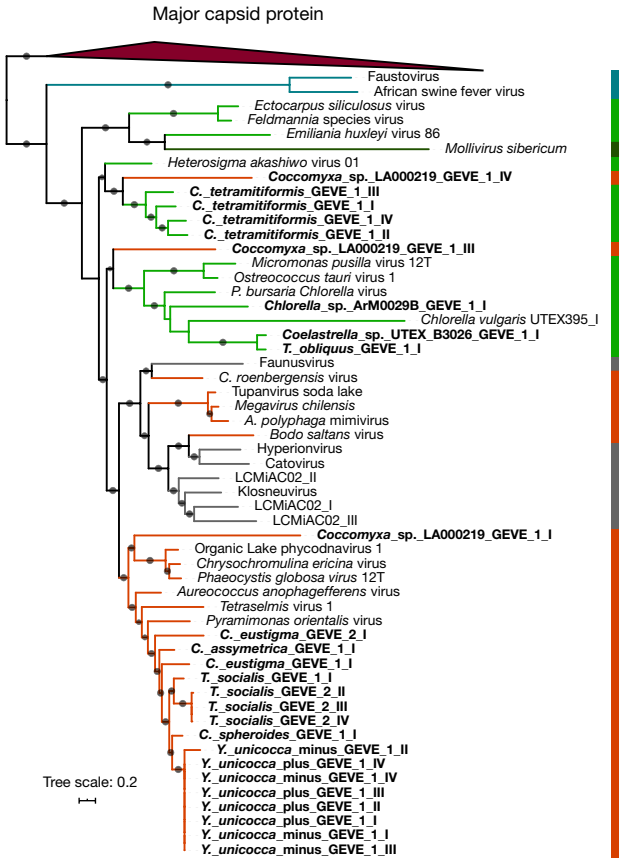


Fig. 3 | Evolutionary history of the GEVEs. a, **b**, Maximum-likelihood phylogenetic trees of the NCLDV core genes DNA polymerase (**a**) and major capsid protein (**b**). Sequences recovered from the chlorophytes are presented in bold, with the suffix 'GEVE_' denoting the sequences that are specific to GEVEs. Branches representing *Ascoviridae*, *Iridoviridae* and *Marseilleviridae* members are collapsed. The dark dots at the nodes represent an approximate likelihood-ratio test (aLRT) Shimodaira–Hasegawa (SH)-like support value of >0.7.

sharing between these genomic regions, confirming that GEVEs have been assimilated into host genomes (Fig. 2b, Supplementary Fig. 4). Third, in 13 GEVEs, we also found shared protein-coding genes (>95% amino acid identity) between GEVE and host loci, indicating the presence of gene exchange between these genomic regions (Supplementary Table 2, Supplementary Fig. 5). These shared genes included several mobile genetic elements (Supplementary Fig. 6). Fourth, NCLDV hallmark genes in GEVEs contained signatures of relaxed selection (a higher dN/dS ratio; one-sided Mann–Whitney *U*-test, $P < 0.001$), which is consistent with lower levels of purifying selection that would be expected in endogenized viruses[2] (Extended Data Fig. 6; see Supplementary Discussion). Last, analysis of transcriptomes for 6 of the 12 genomes in which GEVEs were identified revealed the absence of expression of viral structural and information processing genes that would be expected to be highly expressed under a scenario of viral infection (Extended Data Fig. 7, Supplementary Data 5, Supplementary Discussion). Together with unique features of GEVEs that are incompatible with free viruses (that is, high repeat content, occasional loss or duplication of hallmark genes), these results demonstrate that GEVEs represent bona fide components of algal genomes.

To evaluate the evolutionary history of NCLDV endogenization events, we sought to assess whether the GEVEs are the result of independent integration of distinct viruses or ancient endogenization events followed by inheritance throughout subsequent algal speciation. Because GEVEs derive from both the *Mimiviridae* and *Phycodnaviridae*, it is clear that multiple endogenization events must be responsible for their observed phylogenetic distribution. Moreover, within NCLDV families, the GEVE phylogenies did not generally mirror that of the host algae, and several GEVEs clustered with isolate viruses (Figs. 1a, 3), indicating that most of the GEVEs have not co-diversified within host genomes and are the product of individual endogenization events. In some cases, GEVEs have a phylogeny that mirrors their host, as with the *Mimiviridae* GEVEs found in *Yamagishiella unicocca*, *T. socialis* and *Chlamydomonas sphaeroides*, and the *Phycodnaviridae* GEVEs in *T. obliquus* and *Coelastrella*, and so it may seem plausible that they originate from the same ancient endogenization event. However, the amino acid identity between these GEVEs was markedly low (32–64%) (Supplementary Data 3, Supplementary Discussion) and there was no detectable synteny between these GEVEs (Supplementary Fig. 7), providing evidence against shared descent from the same ancestral NCLDV. Moreover, we found no evidence of shared introns in these GEVEs based on detectable nucleic acid similarity (Supplementary Fig. 8), and intron distributions in NCLDV hallmark genes were also not consistent across these GEVEs (Supplementary Table 3), both of which would be expected if they derived from a shared endogenization event (Supplementary Discussion).

The prevalence of GEVEs in chlorophyte genomes underscores the important role of NCLDV on the genome evolution of eukaryotes. Other studies have identified NCLDV hallmark genes in diverse eukaryotic lineages, including amoeba, metazoa and several protist groups[16–18], providing intriguing evidence of potential horizontal gene transfer between host–virus pairs. Genomic regions with signatures of NCLDV have also been found in two moss genomes and a multicellular alga (Charophyta), where it has been postulated that they are the result of ancient horizontal gene transfer from NCLDV[18,19]. Moreover, one recent study focusing on the evolutionary origins of NCLDV found

evidence that early eukaryotes may have acquired a DNA-dependent RNA polymerase from these viruses[20], supporting the hypothesis that NCLDV–eukaryote gene exchange occurred early in the evolution of these groups[21]. Here we demonstrate that large-scale transfer of giant virus genes into host genomes frequently occurs via endogenization. In addition to the GEVEs that we identified in 12 genomes, we found encoded proteins with homology and best matches to reference NCLDVs in all 65 chlorophyte genomes analysed (Supplementary Fig. 9), indicating that host–virus interactions have shaped a broad range of eukaryotic genomes to varying degrees. Chlorophyte genomes thereby contain a spectrum of NCLDV-derived genetic material; while some have hundreds to thousands of genes from near-complete endogenized viruses, others contain viral regions that have gone through extensive rearrangements and gene loss, but nevertheless retain signatures of their viral origin.

NCLDVs are well known for their mosaic genetic repertoires that derive from horizontal gene transfer from multiple viral and cellular sources[10,11,22], and many of these viruses contain genes that are specific to central metabolic processes previously found only in cellular lineages[9,22–25]. NCLDVs often contain genes that can alter host metabolic state during infection[9,23], akin to 'auxiliary metabolic genes' in bacteriophages[26]. Consistent with this, the GEVEs also contained genes that were predicted to be involved in carbohydrate metabolism, chromatin remodelling, signal transduction, energy production and translation (Extended Data Fig. 8, Supplementary Data 4). Moreover, in several GEVEs, we identified ion channels, ammonia transporters, cell-death-mediating caspases, a light-harvesting complex protein and photolyases (Supplementary Data 4), which were recently reported to be present in a wide range of NCLDVs[9,23]. The introduction of large quantities of diverse viral genes into eukaryotic genomes leads to numerous opportunities for co-option by the host; work on traditional EVEs has identified many such exaptations, ranging from defence against other viruses to membrane transport functions[27].

Bacteriophage integration into host genomes has long been recognized as a major driver of genomic innovation[28]; indeed, many key physiological adaptations of bacteria can be traced to prophage-encoded genes that confer unique capabilities to their hosts[29]. It has been traditionally thought that this mode of genome evolution is less common in eukaryotes[30,31], but our identification of widespread GEVEs in chlorophyte genomes potentially challenges this view. Examples in which large dsDNA viruses endogenize are largely restricted to a narrow range of viruses with specific infection strategies, such as the phaeovirus *Ectocarpus siliculosus* that can integrate into the genome of host gametes as part of replication cycles[32–34]. The widespread endogenization of NCLDV into chlorophytes therefore represents an underappreciated aspect of eukaryotic genome evolution and suggests that many eukaryotic lineages have access to a much larger array of genomic material than previously thought.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2924-2.

1. Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet*. **13**, 283–296 (2012).
2. Holmes, E. C. The evolution of endogenous viral elements. *Cell Host Microbe* **10**, 368–377 (2011).
3. Fischer, M. G. Giant viruses come of age. *Curr. Opin. Microbiol*. **31**, 50–57 (2016).
4. Wilhelm, S. W. et al. A student's guide to giant viruses infecting small eukaryotes: from *Acanthamoeba* to zooxanthellae. *Viruses* **9**, 46 (2017).
5. Abergel, C., Legendre, M. & Claverie, J.-M. The rapidly expanding universe of giant viruses: *Mimivirus*, *Pandoravirus*, *Pithovirus* and *Mollivirus*. *FEMS Microbiol. Rev*. **39**, 779–796 (2015).
6. Weynberg, K. D., Allen, M. J. & Wilson, W. H. Marine prasinoviruses and their tiny plankton hosts: a review. *Viruses* **9**, 43 (2017).
7. Bhattacharya, D. & Medlin, A. L. Algal phylogeny and the origin of land plants. *Plant Physiol*. **116**, 9–15 (1998).
8. Jeanniard, A. et al. Towards defining the chloroviruses: a genomic journey through a genus of large DNA viruses. *BMC Genomics* **14**, 158 (2013).
9. Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R. & Aylward, F. O. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nat. Commun*. **11**, 1710 (2020).
10. Filée, J. Genomic comparison of closely related giant viruses supports an accordion-like model of evolution. *Front. Microbiol*. **6**, 593 (2015).
11. Van Etten, J. L. et al. Chloroviruses have a sweet tooth. *Viruses* **9**, 88 (2017).
12. Schvarcz, C. R. & Steward, G. F. A giant virus infecting green algae encodes key fermentation genes. *Virology* **518**, 423–433 (2018).
13. Sun, C., Feschotte, C., Wu, Z. & Mueller, R. L. DNA transposons have colonized the genome of the giant virus *Pandoravirus salinus*. *BMC Biol*. **13**, 38 (2015).
14. Marcet-Houben, M. & Gabaldón, T. Acquisition of prokaryotic genes by fungal genomes. *Trends Genet*. **26**, 5–8 (2010).
15. Rossoni, A. W. et al. The genomes of polyextremophilic cyanidiales contain 1% horizontally transferred genes with diverse adaptive functions. *eLife* **8**, e45017 (2019).
16. 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).
17. Maumus, F. & Blanc, G. Study of gene trafficking between *Acanthamoeba* and giant viruses suggests an undiscovered family of amoeba-infecting viruses. *Genome Biol. Evol*. **8**, 3351–3363 (2016).
18. Gallot-Lavallée, L. & Blanc, G. A glimpse of nucleo-cytoplasmic large DNA virus biodiversity through the eukaryotic genomics window. *Viruses* **9**, 17 (2017).
19. Maumus, F., Epert, A., Nogué, F. & Blanc, G. Plant genomes enclose footprints of past infections by giant virus relatives. *Nat. Commun*. **5**, 4268 (2014).
20. Guglielmini, J., Woo, A. C., Krupovic, M., Forterre, P. & Gaia, M. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc. Natl Acad. Sci. USA* **116**, 19585–19592 (2019).
21. Forterre, P. & Gaïa, M. Giant viruses and the origin of modern eukaryotes. *Curr. Opin. Microbiol*. **31**, 44–49 (2016).
22. Piacente, F., Gaglianone, M., Laugieri, M. E. & Tonetti, M. G. The autonomous glycosylation of large DNA viruses. *Int. J. Mol. Sci*. **16**, 29315–29328 (2015).
23. Schulz, F. et al. Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–436 (2020).
24. Abrahão, J. et al. Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. *Nat. Commun*. **9**, 749 (2018).
25. Wilson, W. H. et al. Complete genome sequence and lytic phase transcription profile of a *Coccolithovirus*. *Science* **309**, 1090–1092 (2005).
26. Roux, S. et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).
27. Koonin, E. V. & Krupovic, M. The depths of virus exaptation. *Curr. Opin. Virol*. **31**, 1–8 (2018).
28. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
29. Groisman, E. A. & Ochman, H. Pathogenicity islands: bacterial evolution in quantum leaps. *Cell* **87**, 791–794 (1996).
30. Martin, W. F. Too much eukaryote LGT. *BioEssays* **39**, 1700115 (2017).
31. Keeling, P. J. & Palmer, J. D. Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet*. **9**, 605–618 (2008).
32. Cock, J. M. et al. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* **465**, 617–621 (2010).
33. Delaroque, N., Maier, I., Knippers, R. & Müller, D. G. Persistent virus integration into the genome of its algal host, *Ectocarpus siliculosus* (Phaeophyceae). *J. Gen. Virol*. **80**, 1367–1370 (1999).
34. Delaroque, N. & Boland, W. The genome of the brown alga *Ectocarpus siliculosus* contains a series of viral DNA pieces, suggesting an ancient association with large dsDNA viruses. *BMC Evol. Biol*. **8**, 110 (2008).

Methods

Genomes analysed

We downloaded all available genomes from the phylum Chlorophyta from the NCBI on 1 December 2018. A full list of these genomes and associated genome statistics can be found in Supplementary Data 6.

Initial identification of virus-like regions in chlorophyte genomes

For the first step of our analysis, we used a tool called ViralRecall to identify virus-like regions in eukaryotic genomes. An overview of this workflow is presented in Extended Data Fig. 1. ViralRecall is implemented in Python3 and it has been tested on an Ubuntu 16.04 OS. ViralRecall is open-source, and complete code and instructions for use are available on GitHub (https://github.com/faylward/viralrecall). We then performed multiple additional analyses to confirm the virus-like regions derived from endogenization events involving NCLDVs. Below, we first describe the tool ViralRecall, and later sections describe the additional analyses that we performed for confirmation and curation of these results.

Initially genes were predicted from eukaryotic contigs using Prodigal v. 2.6.3[35] (-p meta option), which efficiently predicts genes from bacterial, archaeal and viral genomes owing to their similar genomic architecture. Prodigal would be expected to perform poorly on bona fide eukaryotic genes due to their more complex coding structure (that is, introns), but for viral regions, it provides robust gene predictions. The open reading frames (ORFs) predicted by Prodigal were then compared to two custom databases of HMMs using the hmmsearch tool in HMMER3 v. 3.2.1[36]. The first database used was based on the Pfam database[37] and contains HMMs that are found in cellular organisms and absent from viruses, while the second, the Viral Orthologous Groups (VOG) database (vogdb.org), contains HMMs that are present in viruses and absent from cellular genomes. Both the Pfam and VOG databases were manually curated to remove sequences that were not reliable signatures of cellular organisms or viruses, respectively (see section below on ViralRecall databases for details). After the scores of all HMM hits have been recorded, a final ViralRecall score was calculated for each ORF by subtracting the VOG score from the Pfam score, with no hits given a score of 0. A rolling mean of the ViralRecall score was calculated across all contigs/chromosomes using a window size of 15 ORFs. Genomic regions that had a net positive ViralRecall score were marked as putative viral regions. Cut-offs were employed to remove low-confidence viral regions, including those that were very short (<10 kb), contained few viral genes (number of VOG hits of <4) or had low-confidence hits (a mean ViralRecall score below 10).

Rationale and databases for ViralRecall

The general purpose of ViralRecall is to identify genomic regions that are enriched in encoded proteins that bear homology to known viral protein families. This is challenging, in part because the protein families encoded in viruses and cellular lineages are overlapping; for example, DNA and RNA polymerase subunits are found in cellular lineages and NCLDVs[38]. To overcome this challenge, we used two HMM databases: one for viral protein families (the VOG database; vogdb.org), and one for protein families of cellular organisms (Pfam v. 31[37]). Because the Pfam database includes various protein families, some of which are found in viruses, we removed all Pfam HMMs that were present in viruses available in Viral RefSeq (accessed November 2018; Pfam HMMs were considered represented if a viral protein had a hit above the noise threshold for that model using hmmsearch in the HMMER3 tool[36]).

Curation and final demarcation of GEVEs

Because the majority of the chlorophyte genomes that we analysed were not complete and were composed of multiple contigs, it was necessary to bin viral contigs together to arrive at final GEVEs. In addition to the viral regions identified with ViralRecall, we also identified contigs that contained NCLDV hallmark genes but were not predicted using the ViralRecall tool (see Supplementary Data 2 for details). We manually inspected these contigs to remove those with potentially spurious hits to NCLDV hallmark genes or cases of isolated hallmark genes in otherwise eukaryotic contigs, and if these features were not found, we included these in our preliminary set of putative viral contigs. For the annotation of hallmark genes, we compared the amino acid sequences of all protein annotations against a set of ten custom HMMs that we created for NCLDV hallmark genes (see the method on the construction of NCLDV core gene HMM below). We used the hmmsearch tool in HMMER3[36] with an *e*-value cut-off of 1×10^{-5}.

Final GEVEs were constructed through manual inspection of the viral contigs using several lines of evidence. After arriving at a final set of putative viral contigs for each genome, we visualized similarities in their TNF through hierarchical clustering (Supplementary Fig. 2a–k). We identified cases in which contigs clustered together, and cross-referenced this information with several additional lines of evidence to generate the final GEVEs.

For one line of evidence, we generated a custom set of NCLDV-specific protein families that we refer to as NCVOGs (see detailed methods below). We confirmed that NCVOGs were present in all contigs used to construct GEVEs. We also confirmed that the NCVOG content of GEVEs was consistent with known NCLDVs using an ordination analysis (Extended Data Fig. 3a). Moreover, we conducted a bipartite network analysis of GEVEs and known NCLDVs using the NCVOGs, which confirmed that GEVEs cluster together with known NCLDVs (Extended Data Fig. 3b).

For another line of evidence, the presence of ten NCLDV hallmark genes was assessed in all GEVEs, and phylogenies were created for these genes to assess their phylogenetic provenance. We confirmed consistent phylogenetic signals in NCLDV hallmark genes that were present in the same GEVE (Fig. 3, Supplementary Fig. 1a–e). In four cases, GEVEs lacked nine of ten hallmark genes and encoded only the D5 helicase/primase, but we still included them here due to the high number of NCVOGs present in these GEVEs and their clustering with known NCLDVs in our ordination and bipartite network analysis.

In addition, we assessed the overall number of best hits that encoded proteins had to viruses and prokaryotes (see 'Homology searches' below). As a general guideline, each candidate GEVE had to have at least 50% of the total best hits of encoded ORFs to NCLDVs and prokaryotes combined. This pattern of mixed evolutionary provenance of genes is consistent with reference NCLDVs, which usually contain genes with best matches to genes in all three domains of life[5,39,40].

Finally, we assessed patterns of intron density in the multi-exon genes (genes intervened by at least one intron) on GEVE contigs compared to the host (see details on intron prediction in subsequent sections). Bona fide eukaryotic genes are frequently intron-rich, and we would expect the intron density to be lower in GEVE genes since these features are extremely rare in free NCLDVs and would only propagate after endogenization (similar to what has been shown for horizontally transferred genes in eukaryotes[14,15]). Indeed, we found the median intron density to be low in all GEVEs compared to their host counterparts, and it was statistically significant in 15 out of the 18 cases (Supplementary Fig. 10, see Supplementary Discussion for details), and the existence of a sharp change in intron density can be clearly observed in the two cases that we found where GEVE regions were integrated into host contigs (Fig. 2a).

Hybrid gene prediction

To predict final genes in GEVEs after they had been demarcated, we used a hybrid gene prediction strategy that leveraged both eukaryotic-focused and virus-focused gene prediction algorithms. For this approach, we first predicted genes using AUGUSTUS v. 2.5.5 and the *Chlamydomonas reinhardtii* training model[41,42]. We extracted gene and intron predictions from this analysis, and then subsequently ran

Prodigal v. 2.6.2[35] on the GEVEs and retained prodigal-annotated genes if they did not overlap with any features annotated by AUGUSTUS. This hybrid gene prediction strategy allowed us to take advantage of the benefits of both prediction strategies; for example, while AUGUSTUS is effective at predicting exon–intron boundaries, we found that it can sometimes under-predict viral genes in GEVEs (Extended Data Fig. 4). Conversely, while Prodigal does not predict exon–intron boundaries, it effectively predicts a broader range of NCLDV genes and has been used for this purpose when dealing with free NCLDV genomes[9]. The strength of this approach was validated by our finding that Prodigal successfully predicted some NCLDV hallmark genes that were missed by AUGUSTUS, confirming that using both together in a hybrid prediction strategy is synergistic. The visualizations that compared the predicted ORFs of both approaches (Extended Data Fig. 4) were created using the R package Circlize[43].

Homology searches

To identify the phylogenetic provenance of genes encoded within GEVEs, we compared the amino acid sequences of all predicted GEVE ORFs to NCBI RefSeq v. 92[44]. We used LASTAL v. 959 with the parameter '-m 5000', which increases the number of initial matches per query sequence position and thereby increases the sensitivity of homology detection[45]. Only matches with *e* values of <0.001 were considered for downstream analysis. We removed all hits to Chlorophyta, since the genomes under analysis are present in RefSeq and would thereby have best matches to themselves. Consequently, this also ensured that when a viral gene is present in multiple chlorophytes, it would be recorded as a viral hit, rather than a hit to different chlorophyte genomes. To retrieve taxonomic profiles for the best hits of each GEVE ORF, we cross-referenced each best hit with the NCBI Taxonomy database[46] using the Python API available in the ETE3 Toolkit[47]. An identical LASTAL search was also conducted on all of the proteins found in the 65 chlorophyte genomes analysed in this study (not just those found in GEVEs). This was done to identify viral signatures that were present in the chlorophyte genomes but was not necessarily strong enough to identify full GEVEs, possibly because some GEVEs had degraded over time or if only a small number of giant virus genes were integrated. For this analysis, we predicted proteins from all chlorophyte genomes using Prodigal. The results for this analysis are provided in Supplementary Fig. 9.

Analysis of GEVEs integrated into eukaryotic contigs

In *Chlorella* sp. ArM0029B and *T. obliquus*, we found clear evidence of viral/eukaryotic chimaerism in individual contigs consistent with the integration of viral genomes into eukaryotic chromosomes. To identify possible changes in nucleotide composition that corresponded to the viral verus eukaryotic regions, we calculated TNFs across a 300-bp sliding window for these contigs using the R package Biostrings[48]. We then calculated the Pearson correlation of these frequencies compared to the TNF of a set of the largest contigs in these genome assemblies, which we used as a representation of the core eukaryotic sequences. For this analysis, a large negative Pearson correlation value denotes TNFs dissimilar to the core eukaryotic genome, thereby indicating exogenous DNA. The regions that exhibited the largest negative Pearson correlation values were gene-dense regions in which many ORFs bore sequence homology to known viral sequences, confirming that these regions belong to GEVEs.

Duplication level and synteny analysis

For estimating the level of duplications within each GEVE and reference genomes, we used RECON 1.0.8[49], with a nucleotide alignment identity of >90%. Given a BLASTn output, RECON identifies repetitive regions and their lengths within a genome. To assess potential gene order conservation in GEVEs (synteny), we used the PROMER tool implemented in the MUMMER package[50] with the parameter '--maxmatch'.

GEVE genome plots

Duplications within the GEVEs, best LAST hits of GEVE-specific coding sequences and other gene annotation features were plotted on circular tracks using the R package Circlize[43]. We plotted 14 out of the 18 GEVEs that are classified into *Mimiviridae* and *Phycodnaviridae* based on the phylogenetic provenance of the NCLDV hallmark genes (Fig. 1, Extended Data Fig. 2).

Functional annotation

Predicted protein sequences in each of the viral bins were searched against HMM profiles from the following protein family databases: COG[51], Pfam[37], TIGRfam[52], eggNog[53], eggNOG Viral[53] and VOG (vogdb.org) using hmmsearch of HMMER v.3.2.1. (hmmer.org) with an *e*-value threshold of <0.00001. Best hits for each protein were assessed based on maximum bit score. Functional categories for each hit to the eggNOG database were manually examined.

Construction of NCLDV hallmark protein HMM profiles

For assessment of the occurrence of NCLDV hallmark genes in the GEVEs and chlorophyte genomes, we built custom HMM profiles from ten giant virus 'hallmark genes' used in a previous study[54]. In addition to the proteins used to make these initial HMMs, we used these models to identify additional proteins using 126 reference NCLDVs that represent all established families using hmmsearch (e-value: $e = 1 \times 10^{-10}$). These new proteins were included with those used to make the original HMMs, and new HMMs were created using the 'hmmbuild' command in the HMMER3 suite.

Phylogenetic analysis of NCLDV hallmark genes

The ten hallmark gene HMM profiles (described in the previous paragraph) were used to query the GEVE proteins using the 'hmmsearch' command in the HMMER3 package. A hallmark gene hit was recorded if the *e* value was $<1 \times 10^{-5}$ and the bit score was above a threshold established in a previous study[9]. We found cases in which some of the core genes were split into two individual coding sequences that were located close to each other; we implemented a previously developed Python script[9] to identify these cases, to ensure that they hit the same HMM profile in a non-overlapping manner and to concatenate them (code available at https://github.com/faylward/ncldv_markersearch). As the final quality check, these core gene candidates were queried by LAST against the NCBI RefSeq database to confirm that they had best hits to known NCLDVs.

Finally, selected core gene candidates and reference sequences were used to construct maximum-likelihood phylogenetic trees using PhyML[55] implemented in the ETE3 Toolkit[47] with the workflow 'standard_trimmed_phyml'. As part of this workflow, Clustal Omega[56] was used for sequence alignment and trimAl[57] for alignment trimming. The trees were visualized and annotated in iTOL[58].

NCVOG HMM database

We generated protein families from reference NCLDV genomes to aid in the identification of NCLDV proteins in eukaryotic genomes. For this, we generated orthologous groups from 127 reference NCLDV genomes using Proteinortho v. 6.0.6[59], with proteins that we predicted using Prodigal v. 2.6.3[35] as input. For each orthologous group, we aligned the corresponding proteins using Clustal Omega v. 1.2.3[56], and we then generated HMMs using the hmmbuild command in HMMER3[36]. When searching for NCVOGs in eukaryotic genomes or GEVEs, we used 'hmmsearch' with an *e*-value cut-off of 1×10^{-10}.

Bipartite network of GEVEs and reference NCLDV genomes

To examine patterns of shared gene content between GEVEs and reference NCLDVs, we created a bipartite network based on NCVOG content. For this, all predicted proteins of GEVEs and reference NCLDVs

were searched against the NCVOG database using hmmsearch (*e*-value cut-off of 1×10^{-10}). A bipartite network was then generated in igraph[60] in which two types of nodes were present: small nodes representing NCVOGs, and large nodes representing reference genomes or GEVEs, with edges in between the nodes present if an NCVOG was detected in a given genome. The network was represented using a spring-directed layout, also called a Fruchterman–Reingold layout (layout.fruchterman.reingold in igraph). For purposes of network visualization only, NCVOGs present in more than three genomes were shown.

Transcriptomic data analysis

For transcriptomic analysis, we used RNA sequencing data sets that are publicly available for six of the chlorophytes containing GEVEs, with the following NCBI Sequence Read Archive IDs: *T. socialis* (SRR6260814), *C. eustigma* (DRP003789), *Chlorella* sp. ArM0029B (SRR5416917), *T. obliquus* (ERR2699865), *H. lacustris* (SRR2148810) and *Cymbomonas tetramitiformis*[61] (personal communication). RNA sequencing reads were mapped using Bowtie v. 2.2.6[62] with default parameters. The average expression coverage of the genes and introns were calculated using the 'coverage' command of the bedtools[63] package.

Independent validation of introns using RNA sequencing data

AUGUSTUS predicts intron–exon boundaries defined by the canonical 5′-GT-3′-AG splice site that is conserved in spliceosomal introns[41], and in this study, we were able to leverage the gene prediction model designed for *C. reinhardtii*, which is closely related to the chlorophytes in our analysis and would therefore be expected to provide high-confidence results. To provide further confirmation of these intron predictions, we investigated the expression levels of each of the introns compared to that of their cognate exons. Even after considering alternative splicing and intron retention, the vast majority of introns would be expected to have lower expression than the exons in their corresponding genes. After mapping the RNA sequencing data, we extracted the coverage value of the introns and their cognate exons (according to the method described in the 'Transcriptomic data analysis' section). The results of this analysis are presented in the Supplementary Discussion and Supplementary Fig. 11.

Intron sharing and gene sharing analysis

For the detection of cases of homologous introns that are shared between GEVE and host genomic loci, we clustered all introns of >40 bp from the GEVEs and their corresponding hosts using the CD-HIT-est program[64] with a nucleotide similarity value of >80%. We conducted a second round of analysis using BLASTn (>80% similarity, $e<1\times10^{-10}$) to detect cases of shared introns that were probably missed by CD-HIT. This analysis revealed additional shared introns in *Coelastrella* GEVEs and host loci. For the detection of shared genes across host and GEVE genomic loci, we constructed orthologous groups of proteins with a similarity threshold of >95% amino acids using Proteinortho v. 6.0.6[59], to only allow the detection of cases in which highly similar proteins were present in host and GEVE loci. Functional annotation of the shared gene clusters were done as discussed in the 'Functional annotation' section. For annotation, protein representatives from each cluster were chosen randomly. For both the intron and the gene sharing analyses, only host contigs of >10 kb were used to avoid the inclusion of possible cryptic viral regions with host contigs. Many large host contigs of >100 kb in length frequently contained signatures of shared genes and introns (Fig. 2b, Supplementary Fig. 5). Bipartite networks of intron and gene sharing were constructed in igraph[60].

Calculation of dN/dS ratios

dN/dS values were estimated using four hallmark NCLDV genes (PolB, VLTF3, A32 and SFII). MCP was not used because multiple copies of this gene can sometimes be found in NCLDVs. To compare dN/dS values between GEVEs and free NCLDVs, we identified close relatives of the GEVEs in a set of recently generated metagenome-assembled genomes of NCLDVs in the environment[9]. For this, we constructed diagnostic phylogenetic trees to identify groups of metagenome viruses that were similar to the GEVEs; we ultimately identified three separate clades of metagenome-assembled viruses used in this analysis, and a full list can be found in Supplementary Data 6. Hallmark genes from GEVEs and reference NCLDV metagenome-assembled genomes were aligned separately, and dN/dS values were estimated using codeml in the PAML v.4 package[65] using methods previously described[66]. To ensure that sequences from GEVEs were not too similar or divergent to allow for appropriate estimation of dN/dS ratios, we only considered values where dN > 0, dS < 2 and dN/dS < 10. To evaluate whether GEVEs experience relaxed selection compared to free viruses (as expected if they were endogenized), we compared all dN/dS values using a one-sided Mann–Whitney *U*-test, with a significance threshold of 0.001.

Phylogeny of host algae

To construct a phylogeny of host green algae, we used the large subunit of RuBisCO (rbcL). We identified homologues in host genomes by searching predicted proteins against the COG1850 HMM downloaded from the eggNOG 4.5 database[67], and by retaining best hits. For an outgroup, we used the RbcL protein of *Arabidopsis thaliana* downloaded from the NCBI (BAA84393.1). Proteins were aligned using Clustal Omega (default parameters) and a phylogeny was constructed using IQ-TREE v. 1.6.6[68]. The best substitution model was selected using the ModelFinder tool, which selected the LG+G4 model[69].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Nucleotide and protein sequences specific to each of the GEVEs, hallmark gene set used for phylogenetic analyses, alignments for all phylogenies presented, HMM profiles of the core genes and NCVOG families, and other data products are available at: https://zenodo.org/record/3975964#.XzFj0hl7mfZ.

Code availability

A custom bioinformatic pipeline (ViralRecall) was developed in Python 3.5 for purposes of this study. This code is already publicly available on GitHub for the Aylward lab: https://github.com/faylward/viralrecall. For NCLDV marker gene detection, we also used a custom Python script available on GitHub: https://github.com/faylward/ncldv_markersearch. Other bioinformatic analyses performed in this study were done using publicly available bioinformatic tools and are described in the Methods.

35. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
36. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
37. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
38. Yutin, N., Wolf, Y. I., Raoult, D. & Koonin, E. V. Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. *Virol. J.* **6**, 223 (2009).
39. Filée, J., Siguier, P. & Chandler, M. I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. *Trends Genet.* **23**, 10–15 (2007).
40. Filée, J., Pouget, N. & Chandler, M. Phylogenetic evidence for extensive lateral acquisition of cellular genes by nucleocytoplasmic large DNA viruses. *BMC Evol. Biol.* **8**, 320 (2008).
41. Hoff, K. J. & Stanke, M. Predicting genes in single genomes with AUGUSTUS. *Curr. Protoc. Bioinformatics* **65**, e57 (2019).
42. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).
43. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
44. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).

45. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res*. **21**, 487–493 (2011).
46. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res*. **40**, D136–D143 (2012).
47. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol*. **33**, 1635–1638 (2016).
48. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. Biostrings: efficient manipulation of biological strings. R package version 2.56.0 https://bioconductor.org/packages/Biostrings (2020).
49. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res*. **12**, 1269–1276 (2002).
50. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res*. **30**, 2478–2483 (2002).
51. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*. **28**, 33–36 (2000).
52. Haft, D. H. et al. TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res*. **29**, 41–43 (2001).
53. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res*. **47**, D309–D314 (2019).
54. Moniruzzaman, M. et al. Virus–host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. *Nat. Commun*. **8**, 16054 (2017).
55. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol*. **59**, 307–321 (2010).
56. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol*. **7**, 539 (2011).
57. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
58. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*. **47**, W256–W259 (2019).
59. Lechner, M. et al. Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* **12**, 124 (2011).
60. Csardi G, N. T. The igraph software package for complex network research. *InterJournal Complex Systems* **1695**, 1–9 (2006).
61. Burns, J. A., Paasch, A., Narechania, A. & Kim, E. Comparative genomics of a bacterivorous green algae reveals evolutionary causalities and consequences of phago-mixotrophic mode of nutrition. *Genome Biol. Ecol*. **7**, 3047–3061 (2015).
62. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
63. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
64. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
65. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol*. **24**, 1586–1591 (2007).
66. Martinez-Gutierrez, C. A. & Aylward, F. O. Strong purifying selection is associated with genome streamlining in epipelagic Marinimicrobia. *Genome Biol. Evol*. **11**, 2887–2894 (2019).
67. Huerta-Cepas, J. et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res*. **44**, D286–D293 (2016).
68. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol*. **32**, 268–274 (2015).
69. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

Acknowledgements We thank J. Burns from the Bigelow Laboratory of Ocean Sciences and E. Kim from the American Museum of Natural History for providing access to the RNA sequencing data of *C. tetramitiformis*. We acknowledge use of the Virginia Tech Advanced Research Computing Center for bioinformatic analyses performed in this study. This work was supported by a Simons Early Career Investigator Award in Marine Microbial Ecology and Evolution (grant no. 620443) and NSF grant IIBR-1918271 to F.O.A.

Author contributions F.O.A. and M.M. designed the project and wrote the paper. M.M. curated GEVEs, performed gene annotations and phylogenetic analysis. A.R.W. performed the GEVE protein annotations. C.A.M.-G. performed the dN/dS analysis.

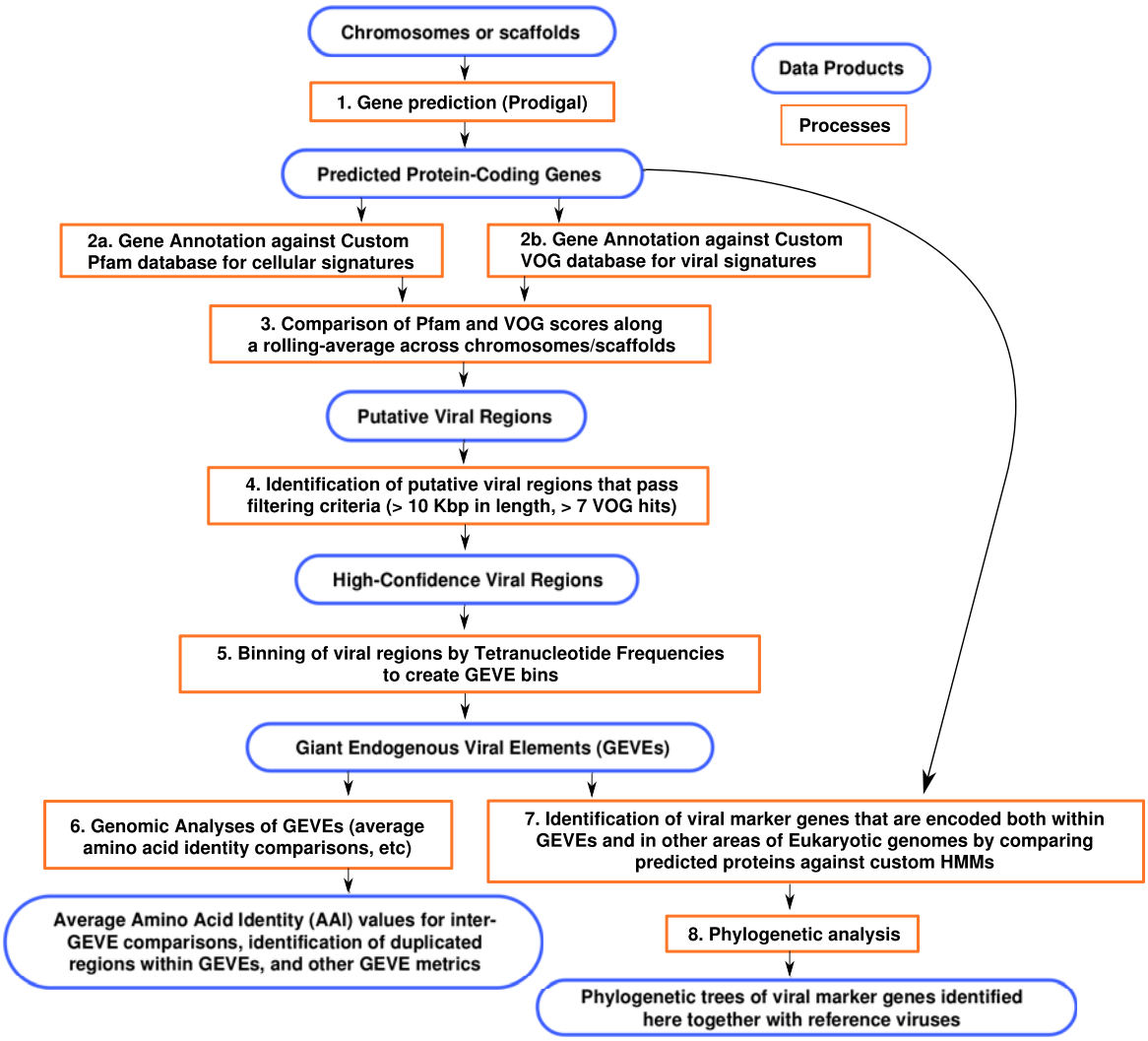
Competing interests The authors declare no competing interests.

Additional information

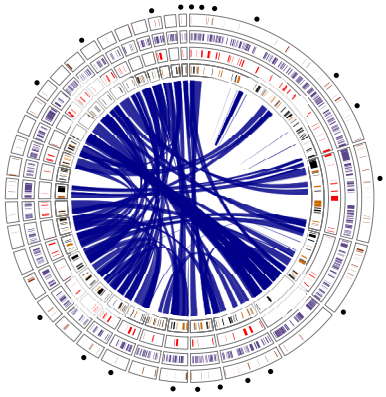
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2924-2.

Correspondence and requests for materials should be addressed to F.O.A.

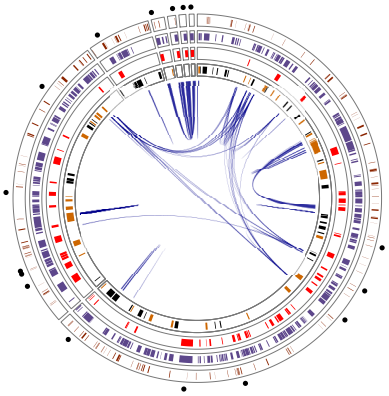
Peer review information *Nature* thanks Chantal Abergel, Matthew Sullivan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



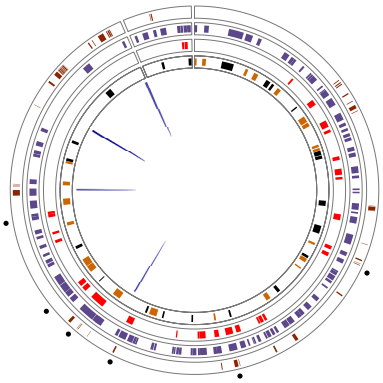
Extended Data Fig. 1 | Workflow for GEVE detection. Overview of the initial steps to identify virus-like regions in chlorophyte genomes and subsequent steps to curate Giant Endogenous Viral Elements (GEVEs). Steps in the grey box are implemented in the ViralRecall tool; steps outside this box represent additional analyses we performed to validate our findings and further analyse the GEVEs.



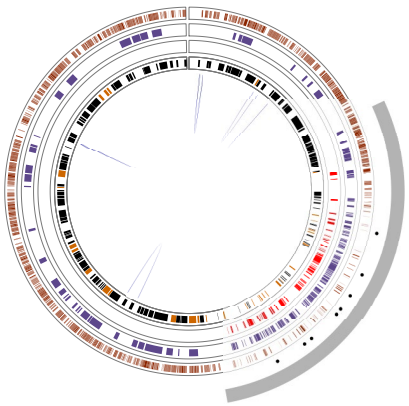
I) Y. unicocca (-) GEVE 1



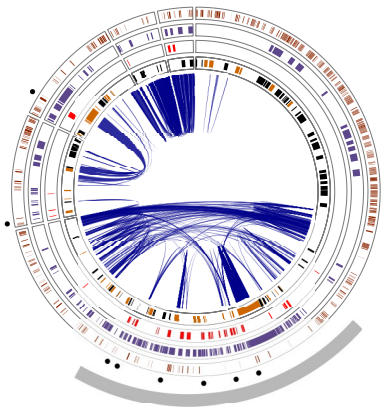
II) C. asymmetrica GEVE 1



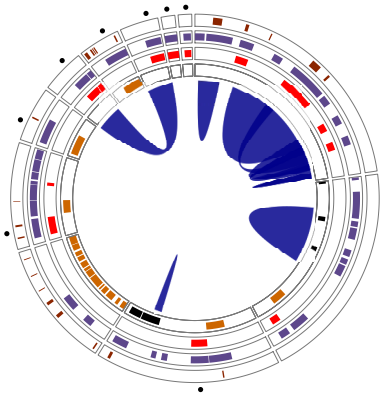
III) Coelastrella sp. B3026 GEVE 1



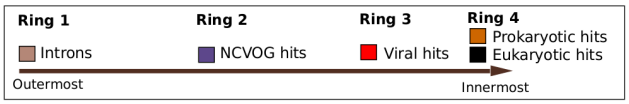
IV) Chlorella sp. ArM 0029B GEVE 1



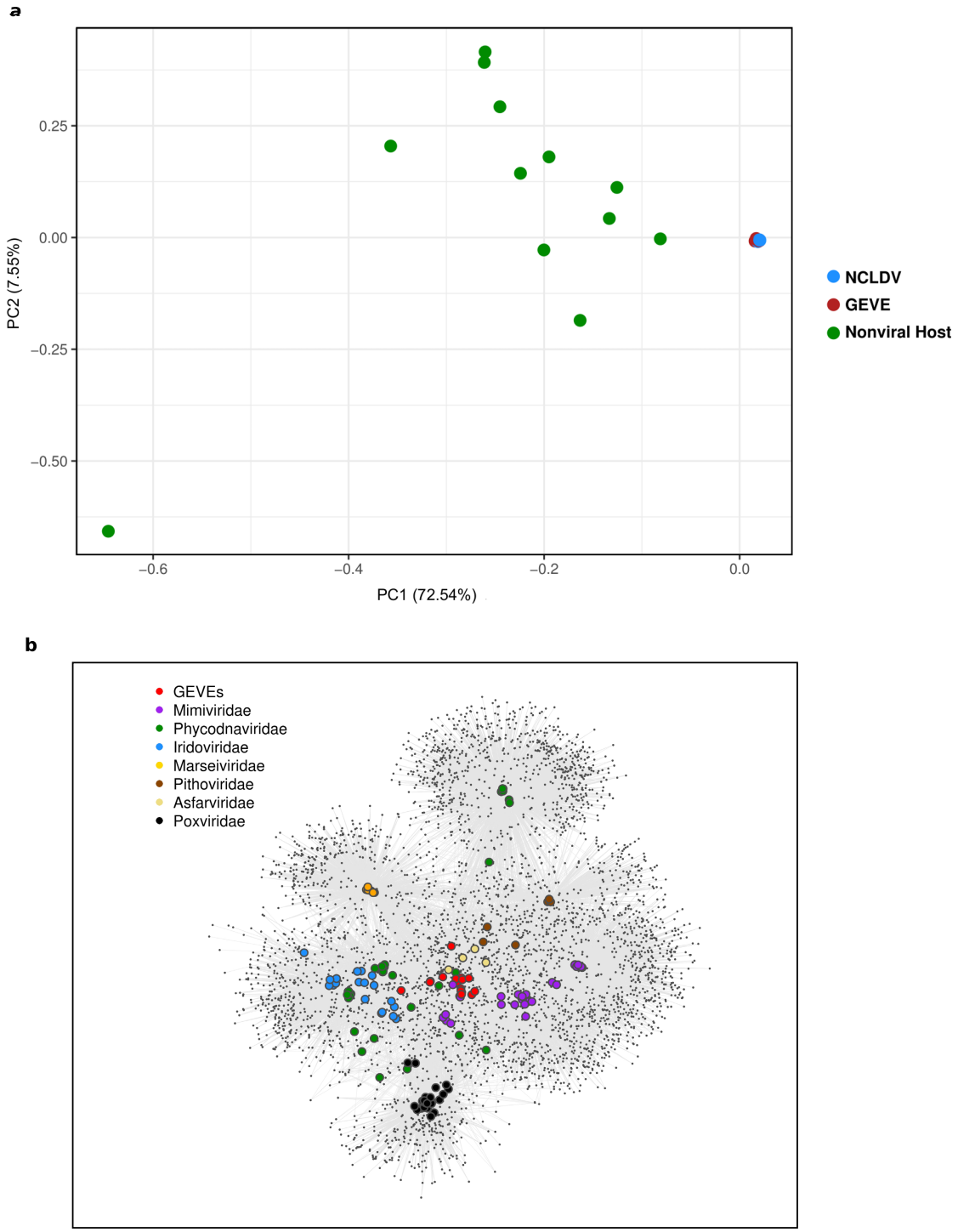
V) T. obliquus GEVE 1



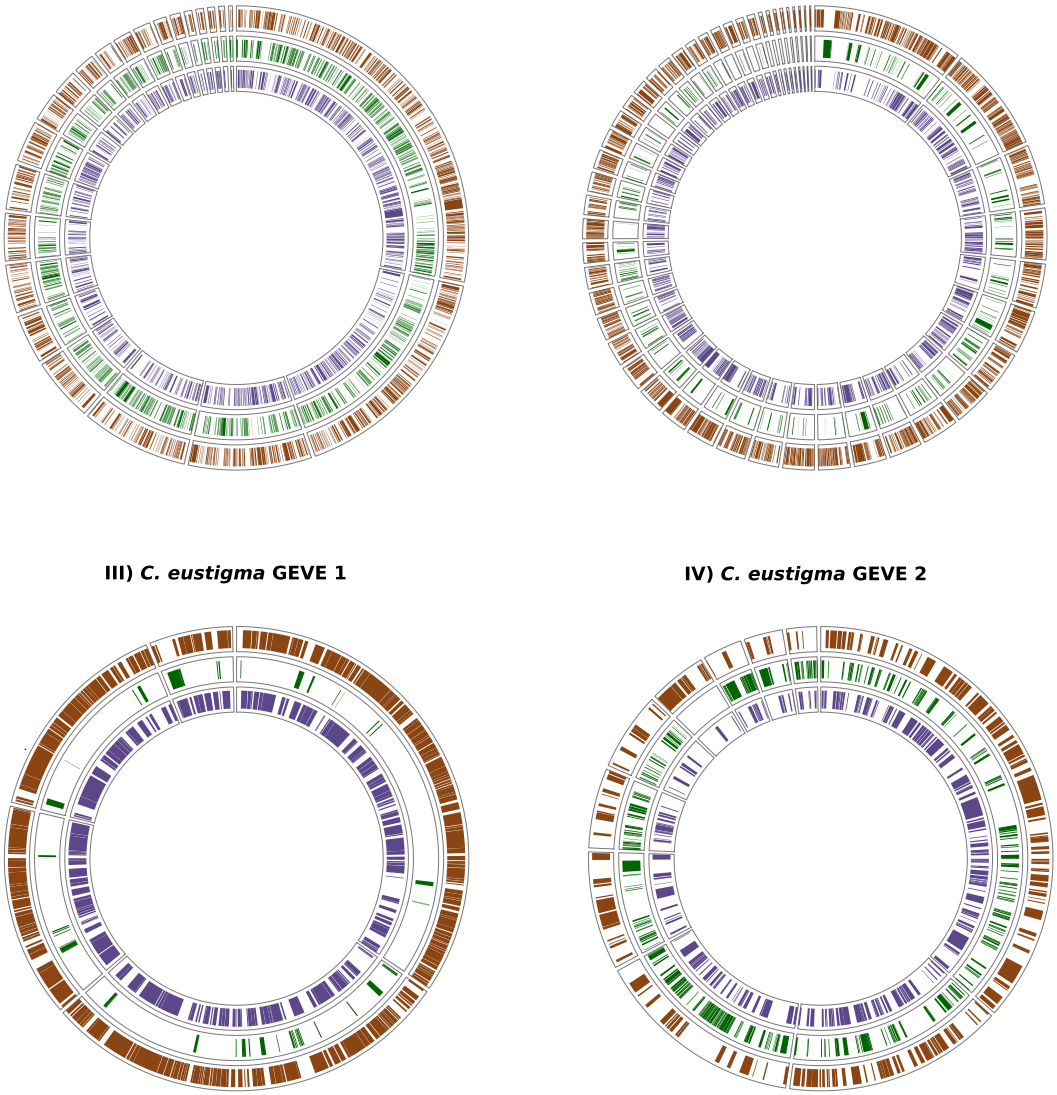
VI) H. lacustris GEVE 1



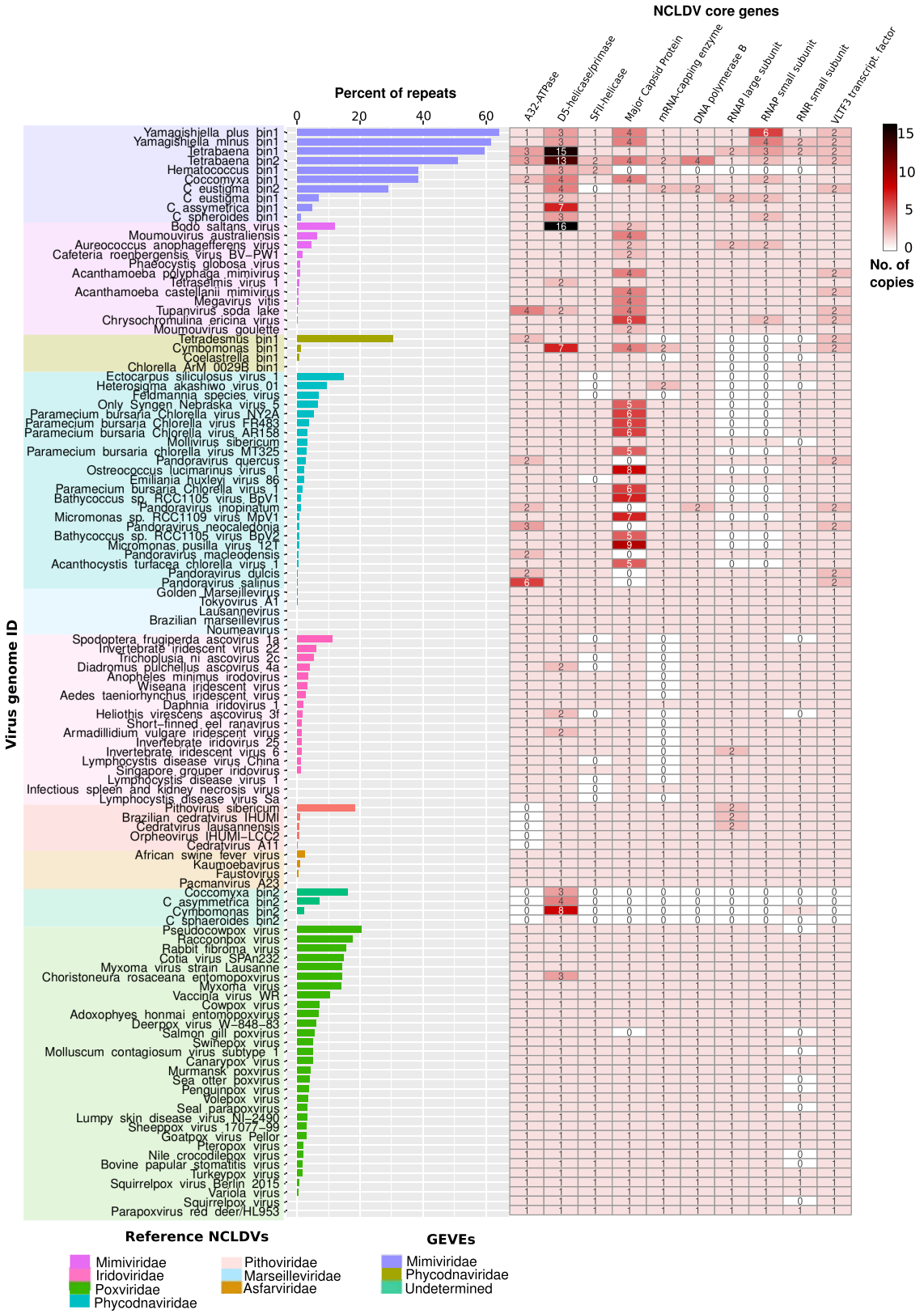
Extended Data Fig. 2 | General features of additional GEVEs. Circular genome plots of 6 additional GEVEs (apart from those shown in Fig. 1b) showing NCVOG HMM hits, spliceosomal intron locations, and best LAST hit matches. Black dots atop the outermost track mark the locations of the core genes, while the blue links inside the circles represent duplicated regions. The grey shading demarcates the location of integrated GEVE as determined by ViralRecall in case of *Chlorella* and *Tetradesmus obliquus*.



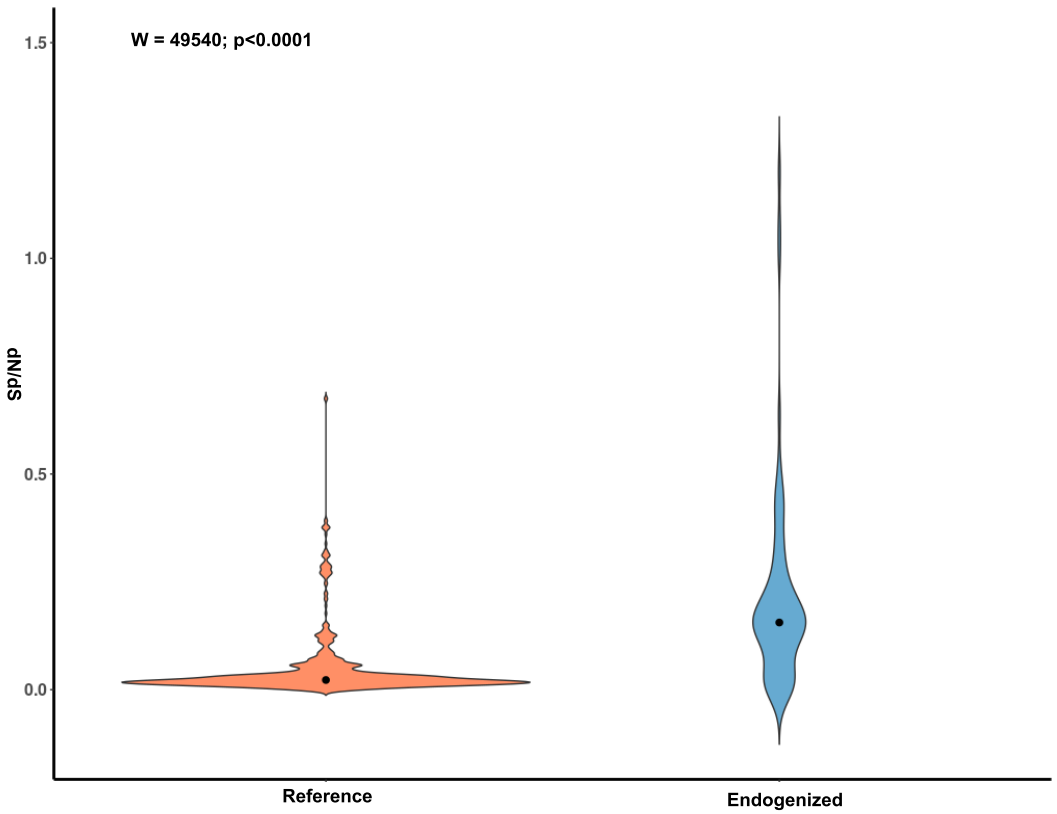
Extended Data Fig. 3 | GEVEs have coding potential similar to known giant viruses. a, Principal component analysis (PCA) of the coding potential of the GEVE genomes, corresponding host genomes and reference giant viruses based on the presence/absence of Nucleocytoplasmic virus orthologous group (NCVOG) specific proteins in these genomes. The plot demonstrates the similarity in coding content of GEVEs and reference giant viruses, whereas the eukaryotic hosts are distinct in terms of coding potential. Nonviral chlorophyte host chromosomes have a much more scattered distribution due to the sporadic occurrence and low abundance of some NCVOGs in these genomes (ankyrin repeat proteins and transposons are represented in NCVOGs and are present in the nonviral portion of host chromosomes, for example). Eukaryotic-specific proteins are not included in NCVOGs, and so the host chlorophyte genomes don't show tight clustering, since this aspect of their genomic repertoires is not captured by NCVOGs. The prcomp() function in R was used to calculate the values. **b**, Bipartite network of 18 GEVEs and 126 reference giant viruses based on shared gene content. The network is constructed by profiling the presence of NCVOGs across all the virus and GEVE genomes represented. Large nodes represent NCLDV or GEVE genomes, smaller nodes represent NCVOG protein families and edges denote gene families represented in different genomes.



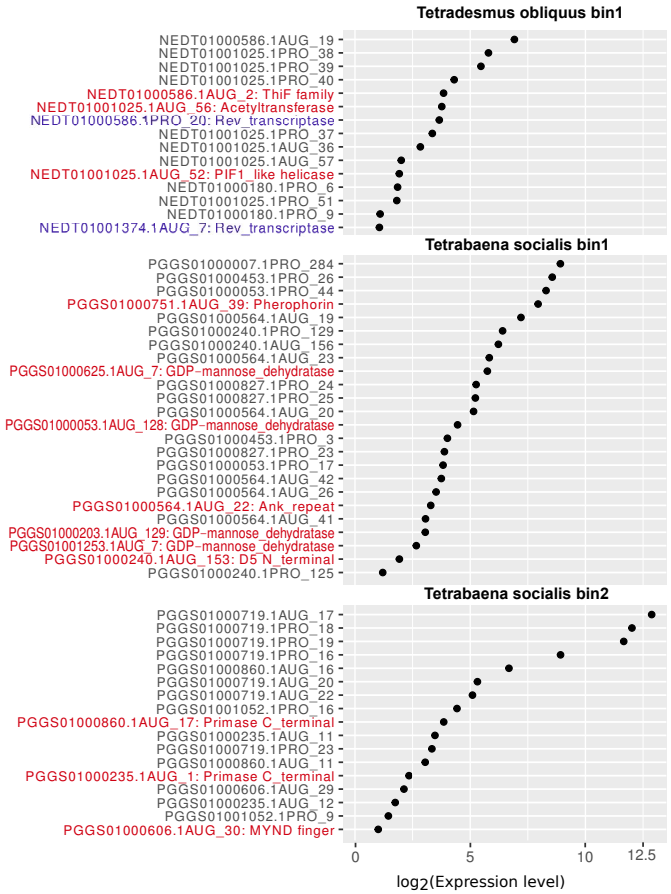
Extended Data Fig. 4 | Example of gene prediction approach within the GEVEs. Genes predicted by AUGUSTUS (outer ring, brown) and non-overlapping Prodigal predicted genes (middle ring, green) in the GEVEs within *Chlamydomoans eustigma* and *Tetrabaena socialis* are shown as examples. In most cases, Prodigal predicted many genes that were not detected by eukaryotic gene prediction algorithms. Many of the Prodigal predicted genes originally missed by AUGUSTUS have hits to NCVOGs (innermost right, purple) - including NCLDV core genes.

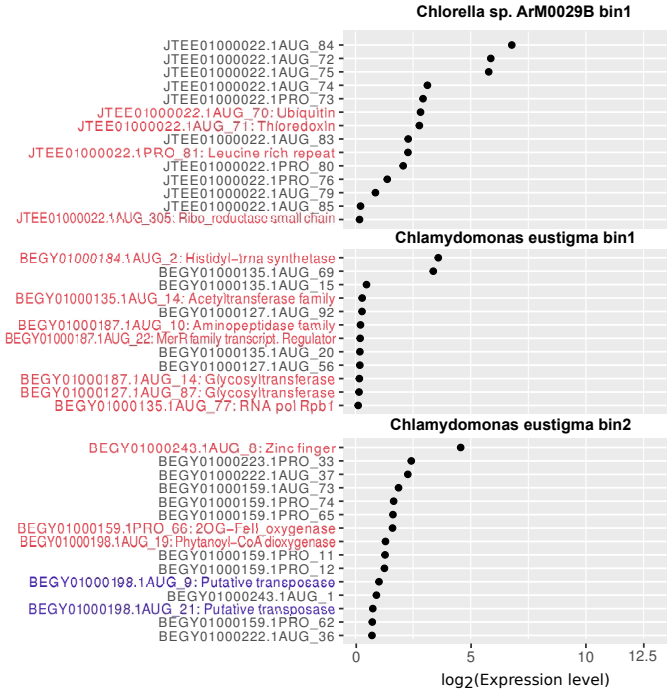


Extended Data Fig. 5 | Level of duplications and core gene copy numbers in GEVE genomes versus reference giant virus genomes. The left panel shows duplication level (repeated genomic regions at >90% nucleotide similarity) as estimated using RECON 1.08. The right panel shows copy numbers of NCLDV core genes in each of the GEVEs and reference genomes (see Methods for details).

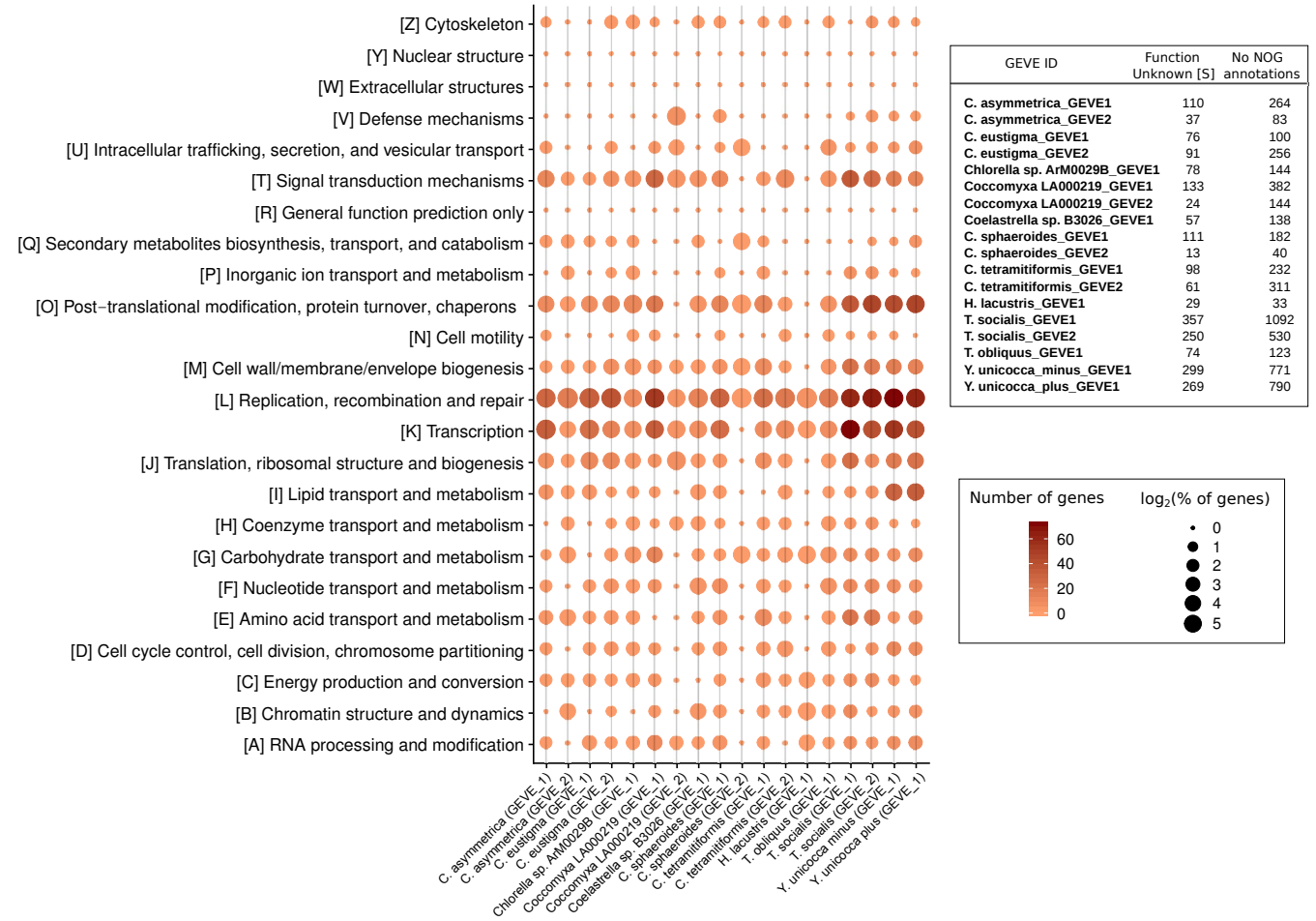


Extended Data Fig. 6 | Signature of relaxed selection in the GEVEs compared to free viruses. Violin plot representing median dN/dS values of endogenized and free reference giant viruses. Statistical significance of differences between dN/dS values of the compared groups according to a non-paired, one-sided Mann–Whitney Wilcoxon test is denoted by: ***$P < 0.0001$. 'W' denotes the Wilcoxon test statistic. For this test 79 values were for GEVE-GEVE dN/dS values and 775 were for comparisons between free viruses. The IDs of the reference genomes used for calculating the dN/dS values are provided in Supplementary Data 6.





Extended Data Fig. 7 | Expression profiles of GEVE genes. Selected set of expressed genes in 6 of the GEVEs. For each GEVE, up to 15 genes with highest expressions are shown, with exception of *Tetrabaena socialis* GEVE_1, for which all genes having >1 expression coverage are presented. For a particular gene, expression is measured as the average read mapping coverage of the CDS(s) in that gene. Genes having putative functions (based on PFAM or COG annotations) are shown in red, while mobile elements are shown in blue.



GEVE ID	Function Unknown [S]	No NOG annotations
C. asymmetrica_GEVE1	110	264
C. asymmetrica_GEVE2	37	83
C. eustigma_GEVE1	76	100
C. eustigma_GEVE2	91	256
Chlorella sp. ArM0029B_GEVE1	78	144
Coccomyxa LA000219_GEVE1	133	382
Coccomyxa LA000219_GEVE2	24	144
Coelastrella sp. B3026_GEVE1	57	138
C. sphaeroides_GEVE1	111	182
C. sphaeroides_GEVE2	13	40
C. tetramitiformis_GEVE1	98	232
C. tetramitiformis_GEVE2	61	311
H. lacustris_GEVE1	29	33
T. socialis_GEVE1	357	1092
T. socialis_GEVE2	250	530
T. obliquus_GEVE1	74	123
Y. unicocca_minus_GEVE1	299	771
Y. unicocca_plus_GEVE1	269	790

Extended Data Fig. 8 | Functional potential coded by the GEVEs. Functional profiles (EggNOG) of the GEVEs normalized across all the NOG functional categories except category S (Function unknown). No gene was found to be in category R (General function prediction only). Number of genes having no hits or in category S (Function unknown) are shown in the table on the right.

Extended Data Table 1 | NCLDV hallmark genes in diverse chlorophyte genomes without GEVEs

Chlorophyte genome ID	D5 helicase/primase	A32-ATPase	DNA polymerase	RNAP large subunit	Major capsid protein	mRNA capping enzyme
Botryococcus braunii	4					
Chlamydomonas applanta	5		1			
Chlorella vulgaris					1	
Dunaliella salina	5					1
Eudorina sp. female			2			
Eudorina sp. male			1			
Gonium pectorale	2					
Monoraphidium sp. 549	4			1		
Picochlorum sp. SENEW3	1					
Picochlorum sp. 'soloecismus'	1					
Trebouxia sp. TZW2008	2	1**				
Chlamydomonas debryana					1**	

NCLDV hallmark genes recovered from 12 additional chlorophyte genomes that do not harbour any GEVEs. Two of these sequences have slightly lower bit scores in HMM search than the specified score cut-off (marked with **), but have best hits to NCLDV hallmark genes in NCBI RefSeq; indicating divergent or degraded hallmark genes.

Extended Data Table 2 | GEVE feature summaries

GEVE_ID	Length (bp)	No. of contigs	Total gene	GC(%) ±STDEV	No. of NCLDV core genes										NCVOG hits	Best LAST hits				ORFANs (%)	No. of introns
					VLTF3	DNAP	RNR	MCP	D5	A32	mRc	RPL	RPS	SFII		Total	Viral	Euk	Prok		
T_socialis GEVE1*	1924766	23	1782	51.28±1.8	2	1	2	1	15	3	1	2	3	1	629	419	151	158	110	76.49	416
T_socialis GEVE2*	1283177	50	1064	55.1±4.67	2	4	1	4	13	3	2	1	2	2	485	333	152	107	74	68.7	397
Y.unicocca (+) GEVE1*	1535269	37	1354	45.7±1.56	2	1	1	4	3	1	1	1	6	1	516	318	92	152	74	76.51	157
Y.unicocca (-) GEVE1*	1524417	31	1371	46.0±1.11	2	1	2	4	3	1	1	1	4	1	547	322	104	145	73	76.51	147
Coccomyxa GEVE1*	1041436	14	705	46.4±5.0	1	1	1	4	4	2	1	1	2	1	303	224	73	84	67	68.23	263
C_asymmetrica GEVE1*	730928	7	491	54.9±4.90	1	1	1	1	7	1	1	1	1	1	207	142	62	45	35	71.08	166
C_eustigma GEVE2*	580957	10	466	38.2±5.97	2	2	1	1	4	1	2	1	1	0	188	123	60	32	31	73.61	165
C_sphaeroides GEVE1*	529923	7	421	55.45±0.74	1	1	1	1	3	1	1	1	2	1	216	138	59	43	36	67.22	125
C_eustigma GEVE1*	427616	5	272	54.30±5.17	1	1	1	1	2	1	1	2	2	1	159	99	38	37	24	63.6	206
H.lacustris GEVE1*	88022	11	76	67.19±5.16	1	0	0	0	3	1	1	0	0	2	39	36	13	5	18	52.63	22
C.tetramitiformis GEVE1†	455621	19	426	35.48±6.42	2	1	1	4	7	1	2	0	0	1	165	124	50	21	53	70.89	12
T.obliquus GEVE1†	424147	5	271	57.8±2.33	2	1	0	1	1	2	0	0	0	1	118	99	36	33	30	63.47	157
Chlorella 0029B GEVE1†	341407	1	289	55.78	1	1	1	1	1	1	1	0	0	1	140	115	69	27	19	60.21	170
Coelastrella GEVE1†	250669	2	250	53.07±1.20	1	1	0	1	1	1	0	0	0	1	104	72	29	21	22	71.2	47
Coccomyxa GEVE2‡	235940	3	202	49.98±2.79	0	0	0	0	3	0	0	0	0	0	30	24	5	10	9	88.12	27
C_asymmetrica GEVE2‡	171705	6	159	37.99±5.06	0	0	0	0	4	0	0	0	0	0	52	38	8	8	22	76.1	14
C_sphaeroides GEVE2‡	77956	3	81	51.18±0.83	0	0	0	0	2	0	0	0	0	0	19	6	4	1	1	92.59	18
C.tetramitiformis GEVE2‡	596198	11	432	51.77±3.48	0	0	0	0	8	0	0	0	0	0	70	57	15	24	18	86.81	64

Summary statistics of the giant endogenous virus elements (GEVEs) described in this study. Abbreviations of the NCLDV core genes: MCP, major capsid protein; DNAP, DNA polymerase; D5, D5 helicase-primase; A32, A32-like virion packaging ATPase; SF_II, superfamily II helicase, RNA_L, RNA polymerase large subunit; RNA_S, RNA polymerase small subunit; RNR_S, ribonucleotide reductase; VLTF3, VLTF3-like transcription factor; mRc, mRNA capping enzyme. Viral_best, Prok_best and Euk_best indicate the number of best hits to different domains (Viruses, Prokaryotes and Eukaryotes) out of the total LAST hits.

**Mimiviridae*.

†*Phycodnaviridae*.

‡Phylogeny undetermined.

nature research

Corresponding author(s): Frank Aylward

Last updated by author(s): Aug 7, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No new data was collected as part of this study. Genomic data that was analyzed is already publicly available in NCBI, and accession numbers are provided in Supplementary Dataset S1. Accession numbers for transcriptomes analyzed are available in the Methods.
Data analysis	A bioinformatic pipeline (ViralRecall) was developed in Python 3.5 for purposes of this study. This code is already publicly available on the GitHub site for the Aylward lab: https://github.com/faylward/viralrecall. Other bioinformatic analyses performed in this study were done using publicly available bioinformatic tools and are described in the Methods section. These tools include: Prodigal v. 2.6.3, HMMER3 v. 3.2.1, AUGUSTUS v. 2.5.5, LASTAL v. 959, ETE3 Toolkit v. 3.1.1, RECON 1.0.8, BLAST 2.9.0+, trimAl v1.4.rev22, Clustal Omega v. 1.2.1, CD-HIT v. 4.6, PAML v. 4

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

No new data was collected as part of this study; all data analyzed is already available in public data repositories. Accession numbers for all genomes used in this study are available in Supplementary Dataset S1. Accession numbers for the transcriptome datasets used in this study are available in the Methods section. Several public databases were used in this study for gene annotation purposes; these databases include the Pfam, TIGRfam, EggNOG, EggNOG Viral, and VOG

databases. The database versions and appropriate citations of literature describing these databases and their location are available in the Methods section. We have made available several processed data products that were generated in this study, including GEVE nucleotide sequences, gene/protein predictions, alignments used for phylogenies, intron annotations, and gene annotations. These files have been uploaded to the Zenodo archive and is available under the following DOI: 10.5281/zenodo.3975964.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We analyzed genomic signatures of Large Nucleo-Cytoplasmic DNA Viruses in publicly-available genomes of green algae. We developed a novel bioinformatic method to identify these signatures, and subsequently performed many comparative genomic analyses to assess their evolutionary and ecological significance.
Research sample	We analyzed 66 publicly available genomes of green algae.
Sampling strategy	We analyzed all relevant data.
Data collection	We did not collect any new data for this study.
Timing and spatial scale	There is no relevant spatial or temporal scale in this study. We analyzed all available genomes in NCBI.
Data exclusions	We did not exclude any data.
Reproducibility	We provide detailed methods and bioinformatic workflows that ensure the results are reproducible.
Randomization	We did not have any treatment groups in our study, so no randomization was needed.
Blinding	We did not have any treatment groups in our study, so no blinding was needed.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging