Virus Evolution, 2025, **11(1)**, veaf009
DOI: https://doi.org/10.1093/ve/veaf009
Advance Access Publication 27 February 2025
Research Article

Multiple, diverse endogenous giant virus elements within the genome of a brown alga

Dean Mckeown[1,2,3], Alexandre Cormier[4], Declan Schroeder[3], Arnaud Couloux[5], Nachida Tadrent[5], J. Mark Cock[1,*], Erwan Corre[2,*]

[1]CNRS, Sorbonne Université, UPMC University Paris 06, Algal Genetics Group, UMR 8227, Integrative Biology of Marine Models, Station Biologique de Roscoff, Place Georges Teissier, CS90074, Roscoff F-29688, France
[2]CNRS, Sorbonne Université, FR2424, ABiMS-IFB, Station Biologique, Place Georges Teissier, Roscoff, France
[3]Department of Veterinary Population Medicine, College of Veterinary Medicine, University of Minnesota, 1365, Gortner Ave, Falcon Heights, Minneapolis, MN 55108, United States
[4]Ifremer, Service de Bioinformatique de l'Ifremer, Centre Ifremer Bretagne - ZI de la pointe du diable, CS 10070, Plouzané 29280, France
[5]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université Evry, Université Paris-Saclay, 2 rue Gaston Crémieux CP 5706, Evry 91057, France

*Corresponding authors. Erwan Corre, CNRS, Sorbonne Université, FR2424, ABiMS-IFB, Station Biologique, Roscoff, France. E-mail: corre@roscoff.fr; J. Mark Cock, CNRS, Sorbonne Université, UPMC University Paris 06, Algal Genetics Group, UMR 8227, Integrative Biology of Marine Models, Station Biologique de Roscoff, CS90074, Roscoff F-29688, France. E-mail: cock@sb-roscoff.fr.

Abstract

Endogenous viral elements (EVEs) have been found in diverse eukaryotic genomes. These elements are particularly frequent in the genomes of brown algae (Phaeophyceae) because these seaweeds are infected by viruses (*Phaeovirus*) of the phylum *Nucleocytoviricota* (NCV) that are capable of inserting into their host's genome as part of their infective cycle. A search for inserted viral sequences in the genome of the freshwater brown alga *Porterinema fluviatile* identified seven large EVEs, including four complete or near-complete proviruses. The EVEs, which all appear to have been derived from independent insertion events, correspond to phylogenetically diverse members of the *Phaeovirus* genus and include members of both the A and B subgroups of this genus. This latter observation is surprising because the two subgroups were thought to have different evolutionary strategies and were therefore not expected to be found in the same host. The EVEs contain a number of novel genes including a H4 histone-like sequence but only one of the EVEs possesses a full set of NCV core genes, indicating that the other six probably correspond to nonfunctional, degenerated viral genomes. The majority of the genes within the EVEs were transcriptionally silent and most of the small number of genes that showed some transcriptional activity were of unknown function. However, the existence of some transcriptionally active genes and several genes containing introns in some EVEs suggests that these elements may be undergoing some degree of endogenization within the host genome over time.

Keywords: brown algae; genome; endogenization; endogenous viral elements; *Phaeovirus*; *Porterinema fluviatile*

Introduction

Endogenous viral elements (EVEs) are widespread in eukaryotic genomes. Most research has focused on retroviral EVEs (Feschotte and Gilbert 2012), but increasing attention is being paid to elements originating from large DNA viruses of the phylum *Nucleocytoviricota* (hereafter referred to as NCVs) (Maumus et al. 2014, Wang et al. 2014, Gallot-Lavallée and Blanc 2017, Moniruzzaman et al. 2020, Denoeud et al. 2024, Sarre et al. 2024). Indeed, evidence is accumulating that NCVs may have played an important general role in shaping eukaryotic genome evolution (Moniruzzaman et al. 2020, Denoeud et al. 2024, Sarre et al. 2024).

NCV EVEs have been found in diverse eukaryotic genomes (Maumus et al. 2014, Wang et al. 2014, Gallot-Lavallée and Blanc 2017, Moniruzzaman et al. 2020, Denoeud et al. 2024, Sarre et al. 2024), but there is little understanding of the origins and biological functions of these elements. Most NCV EVEs appear to be fragments of viral genomes and they are often found in species for which no extant integrating viruses have been identified, indicating that they represent remnants of ancient infections. If such EVEs still perform any functions, it would likely be as part of host processes, rather than viral ones. Currently, the only known exceptions to this situation are EVEs corresponding to NCVs of the genus *Phaeovirus* (family Phycodnaviridae). Phaeoviruses have lysogenic life cycles that involve integration of the virus into the host genome following infection (this life cycle is also referred to as a latent infection strategy) (Müller and Knippers 2011, Schroeder and McKeown 2021). Consequently, the genomes of their brown algal (phylum Ochrophyta, class Phaeophyceae) hosts may contain active integrated viruses (proviruses; Delaroque and Boland 2008, Gallot-Lavallée and Blanc 2017). Moreover, there is evidence that a host alga can be infected by multiple integrating viruses (Meints et al. 2008, Stevens et al. 2014), even including EVEs of different NCV taxa (Ruiz Martínez et al. 2023), and these multiple insertions could co-occur with other NCV EVEs within the same genome. Together these features provide considerable potential for recombination between viral genomes and horizontal gene transfer (HGT) to the host.

Phaeoviruses have been classed into two subgroups: A and B (Stevens et al. 2014). Phaeoviruses of subgroup A have previously been correlated with larger genomes, single insertions per genome, and lower DNA polymerase divergence between viral strains (Stevens et al. 2014). In contrast, phaeoviruses of subgroup B were found to have smaller genome sizes, be associated with multiple insertions per host, and exist as a greater variety of strains. These features suggest contrasting evolutionary strategies, with the K-selected subgroup A thought to rely more on transmission by inheritance of the provirus between host generations (i.e. vertical transmission), while the *r*-selected subgroup B would rely more on virions for transmission (i.e. the lytic cycle or horizontal transmission).

Analysis of active *Phaeovirus* proviruses together with other brown algal EVEs of recent or ancient origin could be a uniquely useful approach to understand viral genome endogenization and virus-derived HGT in eukaryotes. However, analysis of the phenomenon of *Phaeovirus* integration is complicated by the lack of a common integration site (Meints et al. 2008), and evidence for distinct genotypes that deploy different evolutionary strategies (subgroups A and B; Stevens et al. 2014). Until recently, only one *Phaeovirus* provirus had been studied in full genomic context, and, based on its gene content, it is unlikely that this EVE is a functional provirus (Cock et al. 2010). In an earlier study, another integrated *Phaeovirus* strain was capable of producing virus particles but, surprisingly, the provirus may have been present as multiple fragments dispersed throughout the host genome (Delaroque and Boland 2008). The recent completion of 60 new brown algal genome assemblies, corresponding to all of the major orders of the Phaeophyceae, has provided a broad overview of NCV insertions across this phylogenetic class (Denoeud et al. 2024).

In this study, we carried out a detailed characterization of multiple *Phaeovirus* EVEs in the genome of a single strain of the brown alga *Porterinema fluviatile*. Four complete or near-complete proviruses were detected, together with three shorter *Phaeovirus* EVEs. Analysis of the presence of single-copy NCV core genes in the seven EVEs indicated that they each originated from a separate integration event. The EVEs correspond to novel phaeoviruses and, interestingly, they include both multiple subgroup A viruses and a subgroup B virus with a large genome. The cooccurrence of both virus subtypes in the same genome raises questions about the proposed evolutionary strategies of the two subgroups. This study also provided an intriguing insight into the various stages of degeneration and endogenization of NCV proviruses, which warrants further study in the context of HGT.

Materials and methods

Reference genomes

Porterinema fluviatile is a filamentous brown algal that occurs in freshwater, brackish, and marine habitats, where it is either epilithic or grows epiphytically or endophytically on other algae or plants (Waern 1952, Kawai et al. 2021). Phylogenetic analysis has indicated that it is a member of the Ectocarpales (Kawai et al. 2021, Akita et al. 2022). *Porterinema fluviatile* strain SAG 2381 was obtained from the Culture Collection of Algae at Göttingen University (SAG) and its 167 Mbp genome was sequenced by the Phaeoexplorer project (Denoeud et al. 2024). The initial set of gene models, which consisted of 15 519 genes predicted by Gmove using eukaryote optimized annotation, did not identify all the viral genes in the genome. *De novo* gene prediction was therefore carried out using GeneMarkS and an additional 3241 predicted genes were added to the genome annotation (European Nucleotide Archive accession number PRJEB76691).

The genome sequences and annotations of additional reference genomes were acquired using the following accession numbers (downloaded from GenBank unless stated otherwise): *Ectocarpus siliculosus* virus 1 (EsV-1; NC_002687.1), *Feldmannia* species virus 158 (FsV-158; NC_011183.1), *Cladosiphon okamuranus* (GCA_001742925.1), *Ectocarpus* species 7 (previously *Ectocarpus siliculosus*; Orcae database, https://bioinformatics.psb.ugent.be/orcae/overview/EctsiV2), *Nemacystus decipiens* (PRJDB7493), *Saccharina japonica* (GCA_000978595.1), and *Undaria pinnatifida* (GCA_012845835.1).

Identification of viral genes

Genes in the *P. fluviatile* genome derived from viral insertions were identified by a differential BLAST approach. All searches were performed with the BLASTp algorithm, implemented through the software package DIAMOND (Buchfink et al. 2021). The first round searched for viral proteins in the *P. fluviatile* predicted proteome by screening for matches in the virus-only database, Reference Viral Database (RVDB; Goodacre et al. 2018), using BLASTp. Genes with a significant match in the database (based on an *e*-value cut-off of <1-e3) represented putative viral genes (only the best match was retained for each protein).

To compare the BLASTp scores for viral matches with scores for the best match to a cellular protein (i.e. to a protein not encoded by a viral genome), the proteins that had detected matches in RVDB were compared with the NCBI RefSeq nonredundant (NR) protein database, again using BLASTp (with an *e*-value cut-off of <1-e5), excluding all NR proteins which had been assigned to the 'Viruses' category by RefSeq from the subject database. In addition, to exclude matches to other viral proteins which had not been assigned as viral by RefSeq (such as integrated viruses categorized as cellular sequences), all the NR proteins were compared with RVDB using BLASTp. NR proteins that detected a matching protein in RVDB (*e*-value of <1-e40) were considered to be viral and were removed. Only the best NR match was retained after BLASTp comparison of each algal protein with the remaining NR proteins.

To compare the best viral with the best cellular BLAST match for each protein, relative bitscores (rbitscores) were calculated by dividing the BLASTp bitscore for each matching protein by the algal protein's self-hit bitscore (Maumus et al. 2014). Self-hit scores were acquired by comparing the complete set of predicted *P. fluviatile* proteins with itself using BLASTp. Proteins were only analysed if they had significantly matched a viral sequence in RVDB (i.e. had been identified as putative viral sequences by the process described earlier). If a protein had not detected a cellular match, the cellular rbitscore was set to zero. Proteins with a viral rbitscore (RVDB match) at least 20% greater than its cellular rbitscore (NR match) were designated as 'viral'. All other proteins were designated as 'cellular'. The rbitscores were used to generate the scatterplots shown in Supplementary Fig. S1.

A BLASTp search was also performed against the Nucleo-Cytoplasmic Virus Orthologous Group (NCVOG) database, which is comprised of sets of orthologous viral proteins (Yutin et al. 2009). A small subset of NCVOGs (∼50 genes) are known as core genes because they encode functions that are essential for NCVs to complete their life cycles. These core functions have been categorized into four groups based on their level of conservation, ranging from genes that are required by all or nearly all NCVs (group 1) to those that are only conserved within specific NCV families (group 4) (Iyer et al. 2001, Koonin and Yutin 2010).

The *P. fluviatile* genome was also searched for viral genes using Virsorter 2 (Roux et al. 2015) with the NCV parameter and ViralRecall (Aylward and Moniruzzaman 2021) to identify viral regions, for comparison purposes. Following completion of the two screens for viral genes (i.e. the custom pipeline described earlier plus Virsorter 2 and ViralRecall), manual inspection was performed on all contigs/regions with around five or more viral genes, prioritizing those containing NCV core genes and/or those that had been identified by both our custom pipeline, Virsorter 2, and ViralRecall. This manual examination aimed to distinguish *bona fide* EVEs from false positives.

Factors taken into consideration during the manual assessment included genomic organization (e.g. the presence of contiguous arrays of viral genes with no or few intervening cellular genes), exon number (viral genes are almost always monoexonic), and the predicted functions of the viral proteins (some cellular protein families, such as histones, e.g., share similarity with viral sequences due to ancient HGTs; Koonin et al. 2006).

Frameshifts in the open reading frames of NCV core genes within EVEs were manually verified by mapping Illumina DNA-seq reads (ENA accession numbers ERR12682651, ERR12682652, and ERR12682653) to the assembled genome with HISAT2 (Kim et al. 2015) version 2.1.0 and visualizing the read mapping using the Integrative Genomics Viewer (Robinson et al. 2011) version 2.5.

Phylogeny

All algal proteins with homology to NCV DNA polymerase B were aligned with publicly available NCV DNA polymerase proteins (Pfam domain PF00136) using Mafft v7.407 (FFT-NS-i method). A maximum likelihood phylogenetic tree was then constructed using RaxML v8.2.12 (200 bootstraps, and the GTR substitution model with gamma rate distribution). The phylogenetic tree was visualized with iTOL (Letunic and Bork 2019). To analyse the phylogeny of EVEs that lacked DNA polymerase B, equivalent phylogenetic analyses were carried out for late transcription factor 2 (Pfam domain PF06467) and for A18 helicase of superfamily II (Pfam domain PF04851), except that the maximum likelihood trees were generated using FastTree v2.1.10 (JTT ML model, 24 rounds of minimum-evolution nearest-neighbour interchanges, 2 rounds of minimum-evolution subtree-prune-regraft, and 12 rounds of maximum-likelihood nearest-neighbour interchanges).

Whole-genome alignments and annotation

To compare genome collinearity, *P. fluviatile* contigs were aligned using the PROmer package of MUMmer4 (Marçais et al. 2018). As PROmer uses a six-frame nucleotide translation, this analysis also served to identify any viral regions that were not detected by the above similarity searches. Alignments were also made to two reference *Phaeovirus* genomes, EsV-1 and FsV-158. In addition, a search for repeat sequences was performed using the Nucmer package of MUMmer4, by aligning the best viral contigs against themselves and visualizing the repeats using dot plots in CGView.

To create circular representations of contigs and EVEs, the MUMmer4 similarity search results were visualized using CGView (Stothard et al. 2018). The guanine cytosine (GC) content and GC skew values were generated by CGView. The MUMmer PROmer alignments between EVEs and reference viruses were also visualized using genoPlotR (Guy et al. 2011).

RNA expression

Gene expression analysis was carried out using publicly available RNA-seq data (Denoeud et al. 2024). Briefly, these RNA-seq data were obtained from cultures of *P. fluviatile* strain SAG 2381 grown in Petri dishes in either natural seawater (three biological replicates) or 5% natural seawater diluted in water (three biological replicates). RNA was extracted using the Qiagen RNeasy kit (Qiagen, Courtaboeuf, France) and sequencing libraries prepared with the TruSeq Stranded mRNA kit (Illumina). Paired-end reads were generated for each replicate and the reads were mapped onto the *P. fluviatile* genome using Stringtie (Pertea et al. 2015) with default settings and minimum junction coverage of three to obtain read counts per gene, which were then converted to transcripts per million (TPMs) for each annotated gene. Expression levels were expressed as log2(TPM + 1), where the TPM was the mean from three replicates. Quantification of gene expression using RNA-seq data was performed in mapping-based mode with '-l ISR' for paired-end libraries and '-l SR' for single-end libraries.

A set of reference transcripts was created by concatenating the reference transcriptome (Gmove mRNA genes) with the GeneMarkS-predicted viral gene transcripts and indexing with Salmon v1.3.0 (Patro et al. 2017).

Results

Detection of multiple NCV EVEs in *P. fluviatile*

A good quality assembly of the genome of *P. fluviatile* strain SAG 2381 (110 contigs, N50: 2 616 382 bp) has recently been made available (Denoeud et al. 2024). To detect EVEs in this genome, a screen was carried out for proteins that shared similarity with viral proteins in the RVDB database, which is a manually curated, low redundancy database of all viral and virus-related sequences (excluding bacterial viruses) from GenBank (Goodacre et al. 2018). For each *P. fluviatile* protein, rbitscores were then calculated both for the best RVDB match and for the best match to nonviral proteins in the NR database. This analysis identified a number of proteins with high viral and low cellular rbitscores that were comparable to that observed for the *Ectocarpus* species 7 strain Ec32 genome (previously referred to as *E. siliculosus*; Cock et al. 2010), which contains an EsV-1 provirus (Supplementary Fig. S1). Most of these proteins were found in seven regions located on six contigs of the *P. fluviatile* genome assembly. Four of the contigs contained EVEs of similar size and gene content to *Phaeovirus* genomes (EVEs 7, 12, 22, and 40-1). One of the contigs also contained a shorter *Phaeovirus*-like EVE (EVE 40-2) and two additional short *Phaeovirus*-like EVEs were found in the remaining two contigs (EVEs 21 and 94). All six contigs contained host DNA regions in addition to the viral DNA insertions supporting the conclusion that the viral regions corresponded to EVEs and not to contaminating DNA from viral particles. The EVEs were named after their contig numbers, with an additional number in the case of multiple EVEs per contig, i.e. for contig 40 (Fig. 1, Table 1). In total, ∼0.7% of the *P. fluviatile* genome was located within the seven EVEs (1.2 Mbp of 167 Mbp total), corresponding to 6.5% of the protein-coding genes (1211 of 18 760 genes).

Only one EVE possessed a full set of NCV core genes

Sixteen NCV core genes are present in the genomes of both EsV-1 and FsV-158 (shown in bold in Fig. 1), which is currently the minimum known set of core genes encoded by phaeoviruses. Only EVEs

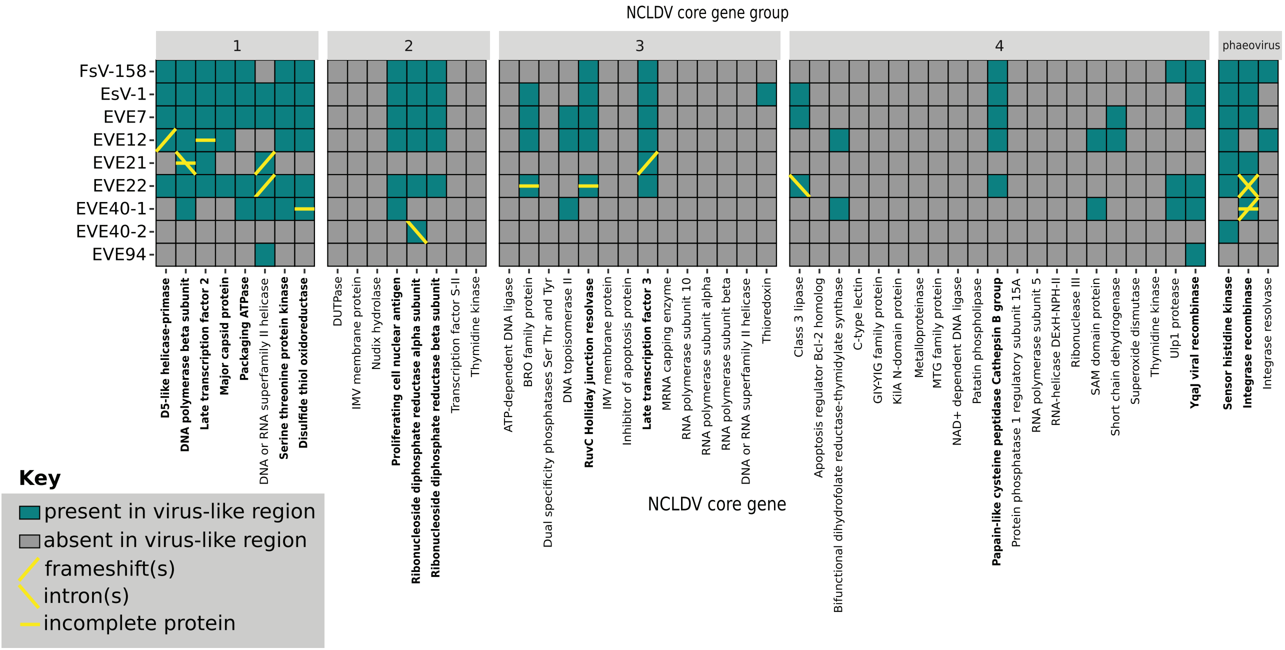

Figure 1. Heatmap illustrating the presence and absence of NCV core genes in *P. fluviatile* EVEs and in two reference phaeoviruses, EsV-1and FsV-158. NCV core genes are ordered by group (top of the figure). The core gene group 'phaeovirus' is constituted of *Phaeovirus*-only genes. Key NCV core genes shared by both EsV-1 and FsV-158 are shown in bold.

Table 1. Summary of the properties of the *P. fluviatile* EVEs.

EVE	Size of EVE (kbp)	Genes in EVE	Genes in EVE with introns	GC% of EVE	GC% of non-EVE region of contig	Contig size (kbp)	EVE position start to end (kbp)	*Phaeovirus* subgroup	Percentage of expressed genes in EVE
7	317	315	**0**	**48.9**	53.9	4036	3498-3815	B	**2.9**
12	323	318	**0**	**46.7**	54.2	3611	22-345	A	**1.3**
21	117	108	8	54.0	54.3	2694	297-414	A	27.8
22	245	237	10	53.8	54.1	2616	1209-1454	A	**8**
40-1	139	156	**0**	**42.3**	54.2	1696	1557-1696	B	**5.1**
40-2	28	30	1	55.6	54.2	1696	220-248	nd	**3.3**
94	42	47	2	56.6	53.2	92	43-85	A	**0**

The values are approximate as no precise virus–host boundaries were identified. EVE features that differ markedly from those of the host genome are highlighted in bold. Expressed genes were defined as having a TPM of ≥1. nd, not determined.

7 and 22 had all 16 of these core genes. EVEs 12, 21, 40, and 94 possessed only 13, 5, 9, and 1 of the 16 core genes, respectively (Fig. 1). EVEs that lack key NCV core genes (DNA polymerase, packaging ATPase, major capsid protein, and integrase recombinase) or exhibit structural modifications of core genes such as frameshifts, partial proteins, and introns are unlikely to be capable of completing a viral life cycle (i.e. are probably no longer functional proviruses). Such core gene set degeneration was observed in all EVEs except EVE 7, which is therefore the only candidate for a complete and functional provirus (Fig. 1).

The genomic context of EVEs

EVEs 7 (Fig. 2a and b), 21 (Supplementary Fig. S4a and b), 22 (Fig. 2e and f), and 40-2 (Supplementary Fig. S4c and d) were flanked by multiple intron-rich genes that mostly detected cellular BLASTp matches, indicating that both EVE-host boundaries were present within these contigs. The context of the other EVEs was less complete. Contig 12 may contain one virus–host boundary of EVE 12 (Fig. 2c and d) because one end of the EVE was bordered by a single intron-rich, cellular gene. Similarly, only one boundary of EVE 40-1 (Supplementary Fig. S4c and d) appears to be present in contig 40 because the EVE extends to the end of the contig. The least complete was EVE 94 (Supplementary Fig. S5a and b), which was on the smallest of the contigs (Table 1), with only a small number of mostly monoexonic genes located at either end of the contig.

The shortest EVE, 40-2, was bordered by a pair of inverted terminal repeats (ITRs; ~99% similarity, ~6 kbp long), but ITRs were not detected in the flanking regions of the other EVEs (Supplementary Fig. S2). The GC contents of EVEs 7 (Fig. 2a), 12 (Fig. 2c), and 40-1 (Supplementary Fig. S4c) differed markedly from the GC content of the flanking algal DNA (Table 1) and abrupt changes in the GC content coincided with the approximate positions of the EVE ends (Fig. 2, Supplementary Figs S4 and S5).

Phylogeny

A DNA polymerase B gene was detected in EVEs 7, 12, 21, 22, and 40-1. However, the DNA polymerase sequence from EVE 21 was excluded from further analysis due to its short length (300 amino acids compared to ~1000 for the DNA polymerases of the other EVEs and EsV-1). All four remaining sequences cluster with *Phaeovirus* DNA polymerases. The EVE 7 and 40-1 sequences fall

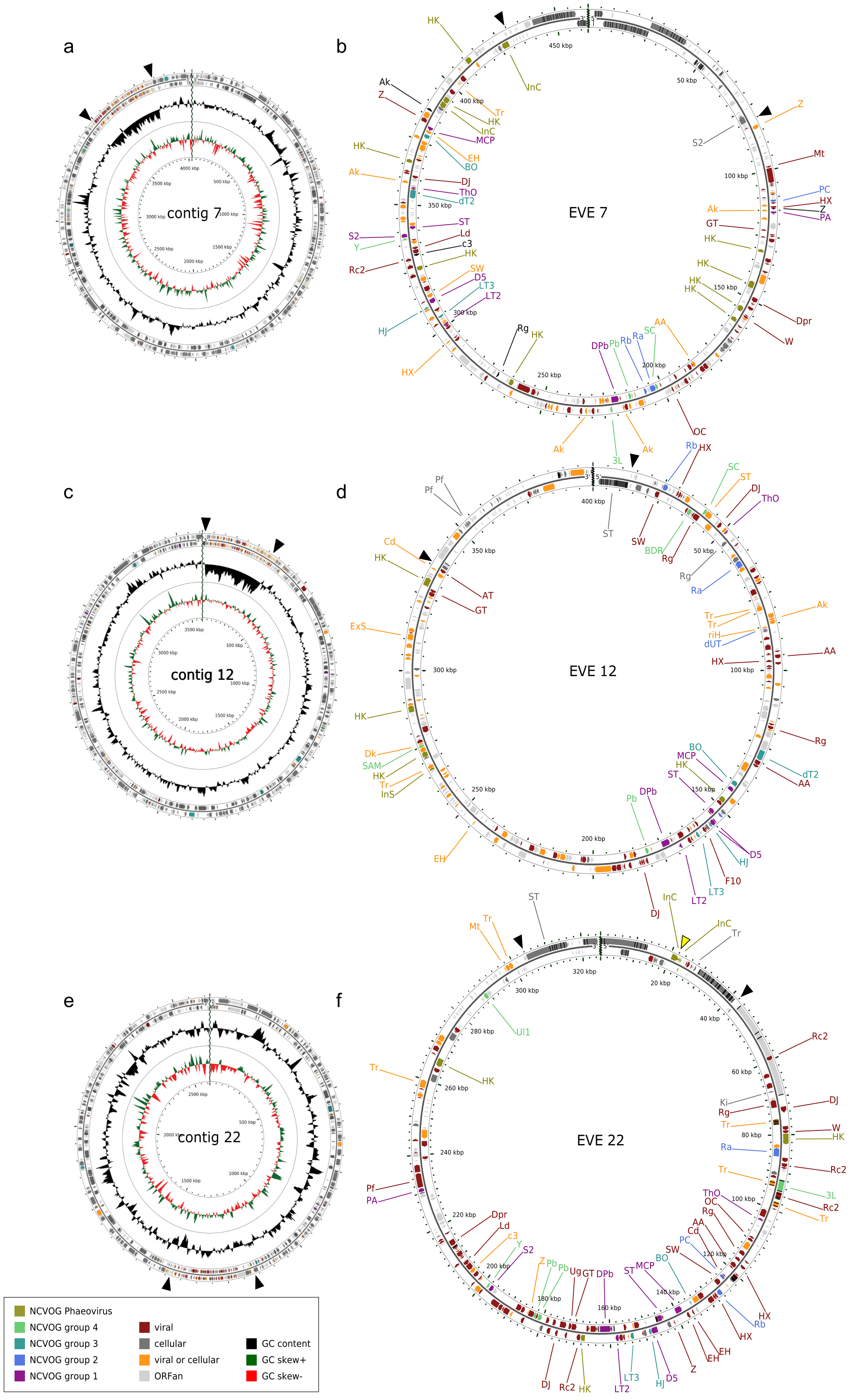

Figure 2. Circular representations of selected *P. fluviatile* EVEs. Whole contig views are shown for contigs 7 (a), 12 (c), and 22 (e), with zoomed views of their EVEs in (b), (d), and (f), respectively. Arrowheads indicate approximate virus–host boundaries. Length is shown in kilobase pairs (kbp). From the exterior inwards, circles 1–4 represent the forward strand, reverse strand, GC content, and GC skew. Genes are coloured according to the categories in the key. Exons are represented by black lines on genes (not to scale). Zigzag lines represent the start and end of each contig in (a), (c), and (e) and the end of the zoomed region in (b), (d), and (f). The yellow arrowhead indicates the location of a fragmented integrase gene in EVE 22 (InC). Gene labels for NCVOGs with characterized functions are abbreviated as follows (with NCV core gene group in brackets): 3L, Class 3 lipase (4); AA, AAA ATPase; AT, aminotransferase; Ak, ankyrin repeat protein; BDR, bifunctional dihydrofolate reductase–thymidylate synthase (4); BO, BRO family protein (3); Cd, cytidine and deoxycytidylate deaminase; D5, D5-like helicase-primase (1); DJ, DnaJ family protein; DPb, DNA polymerase beta subunit (1); Dk, deoxynucleoside kinase; Dpr, DNA primase; EH, HNH endonuclease; ExS, exostosin; F10, F10-like kinase; GT, glycosyltransferase family protein; HJ,

within subgroup B, whereas those of EVEs 12 and 22 cluster with subgroup A (Fig. 3). For the EVEs that lack DNA polymerase, but have other group 1 core genes (EVEs 21 and 94), the late transcription factor 2 of EVE 21 is highly similar to that of EVE 22 (Supplementary Fig. S6a) and the A18 helicase of superfamily II of EVE 94 is most similar to that of EVE 22 (Supplementary Fig. S6b), indicating that both EVE 21 and EVE 94 belong to subgroup A.

Transposon-related genes

EVEs 22 and 40-1 have two and three IS4-like transposases, respectively, and these elements share high similarity with those of EsV-1 (EsV-1-155 and EsV-1-170; Van Etten et al. 2002). A second type of transposase (IS200/IS605-like) is present in all EVEs except EVE 21 and 40-1. A single copy of this transposase is present in FsV-158 and it is likely part of a transposable element acquired by phaeoviruses during the provirus phase (Schroeder et al. 2009). Multiple copies of this transposase were present in EVEs 12 (three copies) and 22 (six copies). In EVE 22, one transposase interrupted a class 3 lipase gene and another was part of a short region containing the putative translocated integrase recombinase (Fig. 2). EVE 22 also had two additional divergent IS200/IS605-like transposases and an IS630-like transposase.

Retrotransposon-like elements were detected in EVEs 7 (~6kbp long) and 40-1 (~12 kbp long). Both retrotransposon-like elements were predicted to encode a retrotransposon-like reverse transcriptase, multiple polyproteins, and integrase.

Integrases

Integrases are virus-produced proteins that catalyse the integration of the viral genome into its host genome. The catalytic sites of *Phaeovirus* integrase recombinases include four invariant amino acids. Algal genomes contain EsV-1 integrase-like (EsV-1-213-like) proteins that are similar to *Phaeovirus* integrases, possibly due to HGT (Delaroque and Boland 2008). Only EVEs 7 and 21 encode intact integrases that both possess the characteristic catalytic site and are more similar to *Phaeovirus* integrases than to algal integrase-like proteins (Supplementary Fig. S9). Similar integrase sequences are present in EVEs 22 and 40-2, but the open reading frames of these integrase genes contain multiple frameshifts (Fig. 1, Supplementary Fig. S9). In addition, the EVE 22 integrase gene is separated from the rest of the EVE by an ~30 kbp region containing three algal genes (Fig. 2). Various EsV-1-213-like genes were found on contigs 7, 21, and 40 within the algal DNA part of the contig (i.e. outside the EVEs). All of the proteins encoded by these genes are predicted to lack the catalytic integrase site and the genes are not located close to the EVEs, apart from the integrase gene on contig 7, which is located near the approximate border of the EVE (Fig. 2, Supplementary Fig. S9).

Other NCV orthologous genes

The *P. fluviatile* EVEs contain eight NCVOGs that have not been found previously in *Phaeovirus* genomes (Supplementary Table S1). Four of these NCVOGs have novel predicted functions for phaeoviruses, three of which are shared with mimiviruses (aminotransferase, ribonuclease H, and deoxyribonucleoside monophosphate kinase) and one with chloroviruses (exostosin glycosyltransferase) (Supplementary Table S1). Some of these NCVOGs may be involved in nucleic acid metabolism [ribonuclease H (Yutin et al. 2013) and deoxyribonucleoside monophosphate kinase (Bäckström et al. 2019)], whereas others may have roles in the glycosylation of capsid proteins [aminotransferase (Piacente et al. 2012) and exostosin glycosyltransferase (Van Etten et al. 2017)].

Other novel genes

In addition, the *P. fluviatile* EVEs contain a number of genes that are very poorly or not at all conserved across the NCV families, including genes that are rare or have not yet been detected in *Phycodnaviridae* (Supplementary Table S2). EVEs 7 and 12 had the most novel genes and most are shared with at least one NCV family (mostly *Mimiviridae*; Supplementary Table S2). Many of these protein families have been hypothesized to have been horizontally transferred from cellular organisms to viruses and now serve to modulate cellular processes (e.g. components of a bacterial stress response pathway, PF15632, and the electron transfer chain, PF00355) or capsid structure (tetratricopeptide repeat, PF13424) (Supplementary Table S2). One unusual protein identified was an H4 histone-like protein found in EVE 40-1 (Supplementary Fig. S10). Among viruses, H4 histone-like genes have only been found in marseilleviruses (NCVs) and bracoviruses. In marseilleviruses, H4 histone-like proteins are an ancient family that likely predates eukaryotes (Erives 2015) and are required for viral replication (Liu et al. 2021). In bracoviruses (*Polydnaviridae*), H4 histone-like proteins, which are likely to correspond to a recent HGT from host to virus, are used by the bracoviruses to suppress the host immune system (Gad and Kim 2008). The EVE 40-1 H4 histone-like protein is more similar to bracovirus/eukaryotic-type H4 histone proteins than to the H4 histone-like proteins of NCVs (Supplementary Fig. S10). Unlike the situation observed in bracoviruses, the EVE 40-1 histone-like protein is not closely related to that of its host (i.e. *P. fluviatile*), the two comparisons detecting 84.2% and 68.3% identity, respectively. The EVE 4-1 protein may be a novel type of H4 histone-like protein that has diverged from eukaryotic homologs but still retains the most conserved features of H4 histones.

Histidine kinases

Histidine protein kinases (HPKs) are cell signalling molecules that generally function as part of two-component signalling systems. HPK genes are known to occur in *Phaeovirus* genomes (Delaroque et al. 2001), but their functions in this group of viruses are unknown. A total of nine HPKs were detected in EVE 7, more than has been detected in previously studied phaeoviruses; EsV-1 has six HPKs, whereas FsV-158 and FirrV-1 have three (Delaroque et al. 2003, Schroeder et al. 2009). EsV-1 and FsV-158/FirrV-1 possess distinct sets of HPKs with four being EsV-1-specific and one FsV-158/FirrV-1-specific. Interestingly, EVE 7 has HPKs that are similar to all three of the FsV-158/FirrV-1 HPKs and to five of the six EsV-1

Figure 2. (Continued) RuvC Holliday junction resolvase (3); HK, sensor histidine kinase (5); HX, helicase exonuclease; InC, integrase recombinase (5); InS, integrase resolvase (5); Ki, KilA N-domain protein; LT2, late transcription factor 2 (1); LT3, late transcription factor 3 (3); Ld, Divergent lipase; MCP, (1); Mt, methyltransferase; OC, OTU-like cysteine protease; PA, packaging ATPase (1); PC, proliferating cell nuclear antigen (2); Pb, papain-like cysteine peptidase cathepsin B group (4); Pf, Pif1 helicase; Ra, ribonucleoside diphosphate reductase alpha subunit (2); Rb, ribonucleoside diphosphate reductase beta subunit (2); Rc2, replication factor C small subunit 2; Rg, RING domain protein; S2, DNA or RNA superfamily II helicase (1); SAM, SAM domain protein (4); SC, short-chain dehydrogenase (4); ST, serine threonine protein kinase (1); SW, SWIB MDM2 domain protein; ThO, disulfide thiol oxidoreductase (1); Tr, transposase; Ug, UDP-glucose 6-dehydrogenase; Ul1, Ulp1 protease (4); W, Von Willebrand factor; Y, YqaJ viral recombinase (4); Z, zinc finger family protein; c3, collagen triple helix repeat protein; dT2, DNA topoisomerase II (3); dUT, DUTPase (2); riH, ribonuclease H.

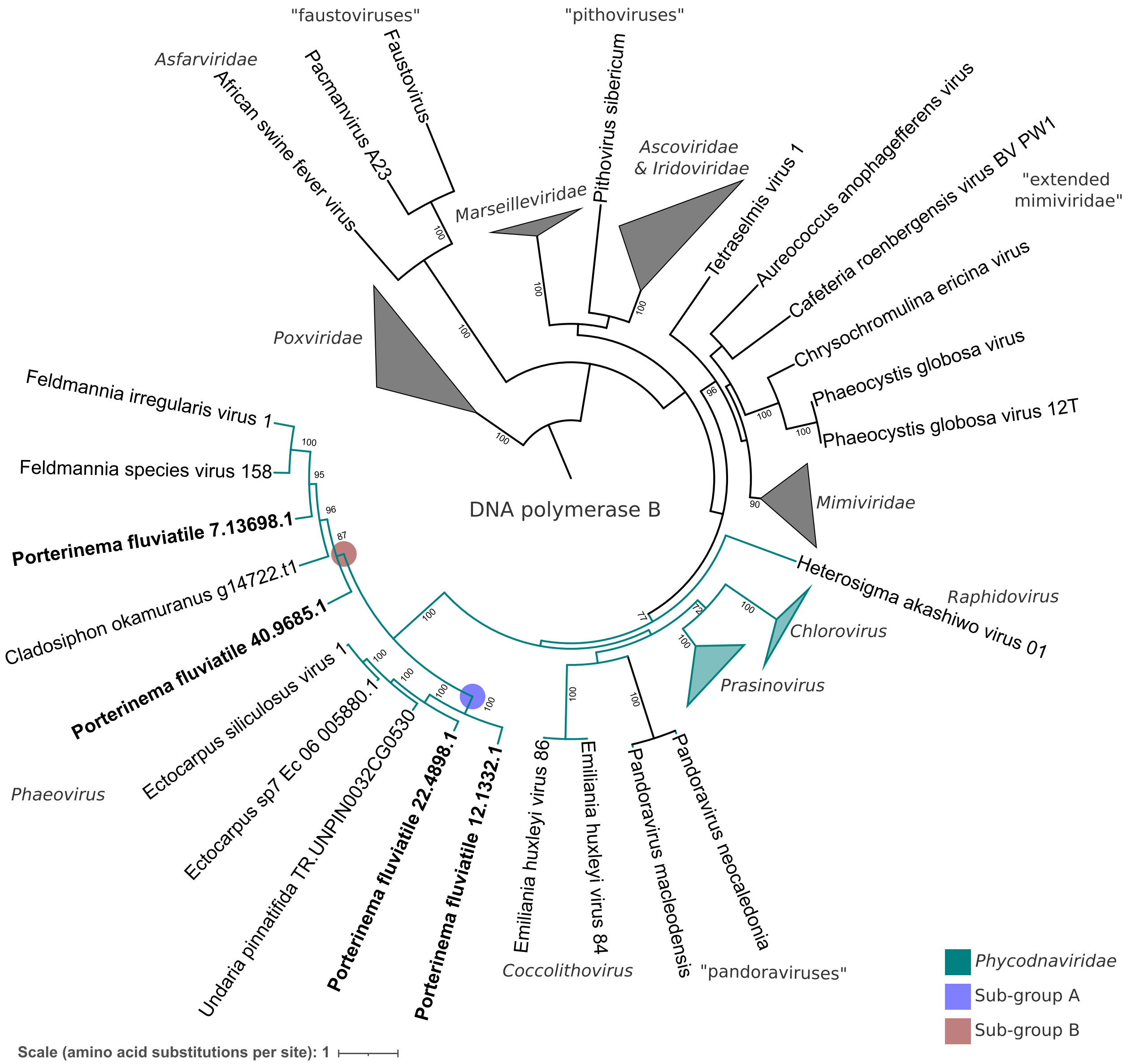
"faustoviruses"
"pithoviruses"
Asfarviridae
African swine fever virus
Pacmanvirus A23
Faustovirus
Pithovirus sibericum
Marseilleviridae
Ascoviridae & Iridoviridae
Tetraselmis virus 1
Aureococcus anophagefferens virus
Cafeteria roenbergensis virus BV PW1
"extended mimiviridae"
Chrysochromulina ericina virus
Phaeocystis globosa virus
Phaeocystis globosa virus 12T
Mimiviridae
Poxviridae
Feldmannia irregularis virus 1
Feldmannia species virus 158
Porterinema fluviatile 7.13698.1
Cladosiphon okamuranus g14722.t1
Porterinema fluviatile 40.9685.1
Ectocarpus siliculosus virus 1
Ectocarpus sp7 Ec 06 005880.1
Undaria pinnatifida TR.UNPIN0032CG0530
Porterinema fluviatile 22.4898.1
Porterinema fluviatile 12.1332.1
Phaeovirus
DNA polymerase B
Heterosigma akashiwo virus 01
Raphidovirus
Chlorovirus
Prasinovirus
Emiliania huxleyi virus 86
Emiliania huxleyi virus 84
Coccolithovirus
Pandoravirus macleodensis
Pandoravirus neocaledonia
"pandoraviruses"
100
96
90
95
87
77
72
🟩 Phycodnaviridae
🟪 Sub-group A
🟥 Sub-group B
Scale (amino acid substitutions per site): 1

Figure 3. Phylogenetic tree based on viral DNA polymerases encoded by *P. fluviatile* EVEs and by diverse NCV viruses. Maximum likelihood tree with GTR substitution model with gamma rate distribution and 200 bootstrap replicates. Bootstrap values are shown as percentages next to nodes (only >70% values are shown). Triangles represent collapsed clades. *Porterinema fluviatile* proteins are shown in bold. The EVE that encodes each protein can be deduced from the first number in the protein name, e.g. *P. fluviatile* 40.9685.1 is encoded by EVE 40-1.

HPKs (Supplementary Table S3). The three additional HPKs in EVE 7 correspond to additional copies of previously described HPKs.

Comparative analysis of *P. fluviatile* EVEs and other phaeoviruses

EVEs 12, 21, 22, 40, and 94 show greater colinearity with EsV-1 than with FsV-158, whereas EVE 7 is more colinear with FsV-158 (Supplementary Fig. S7). Most of the EVEs share only short colinear regions with these reference viruses, except for EVEs 22 and 94 (Supplementary Fig. S7), which, consistently, also exhibited higher levels of percent alignment when EsV-1 or FsV-158 were aligned with these EVEs as reference (Fig. 4, Supplementary Fig. S7). We also noted that EVE 94 is strongly colinear with (Fig. 4) and similar to (Fig. 4, Supplementary Fig. S7) part of EVE 22.

In almost all cases, pairs of EVEs share at least one orthologous NCV core gene, making it unlikely that they are derived from the same provirus. The only exception is for EVEs 12 and 94, which share no common NCV core genes (Fig. 1) and are colinear with different regions of EsV-1 (Fig. 4a). Moreover, EVE 94 encodes NCV core genes (A18 superfamily helicase II and YaqJ recombinase) and NCVOGs (DNA primase, class 3 lipase, and divergent lipase) that are missing from EVE 12 but present in EVEs 7, 22, and 40. Given the close proximity of EVEs 12 and 94 to the ends of their contigs (Fig. 2, Supplementary Fig. S5), it is possible that they could be two parts of the same larger EVE. However, association of these two EVEs would probably not result in the formation of a functional provirus, as the reconstructed EVE would still lack important NCV core genes (Fig. 1). An alternative, and perhaps more likely, sce-

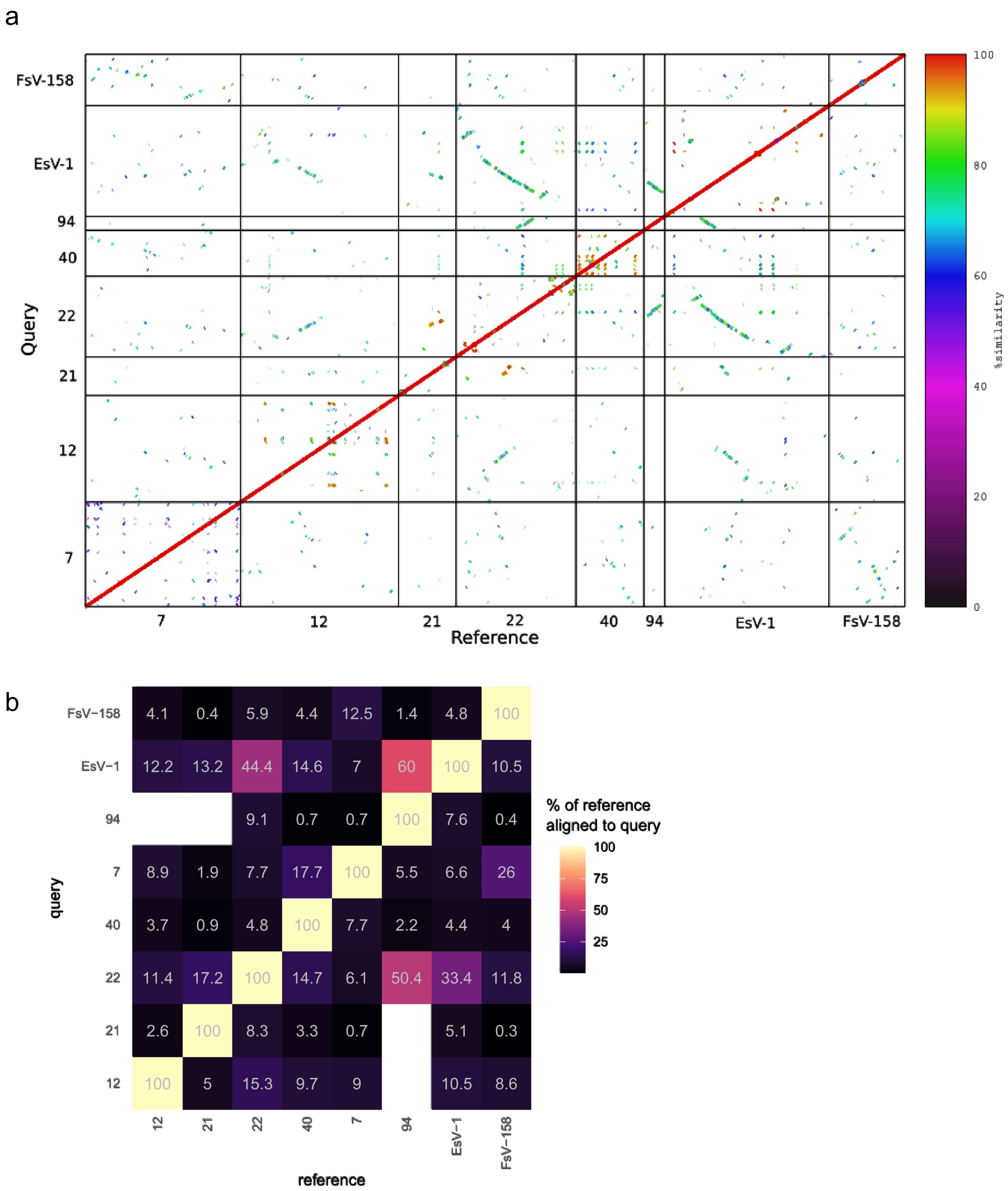

Figure 4. (a) Dot plot comparison of *P. fluviatile* EVEs with each other and with two reference *Phaeovirus* genomes. Only the EVE regions of the *P. fluviatile* contigs were aligned. The two reference viral genomes were EsV-1 and FsV-158. The red diagonal line corresponds to the self-alignments. Alignment drawn using MUMmer4 PROmer. (b) Heatmap summarizing the similarity between the *P. fluviatile* EVEs and two reference phaeoviruses, EsV-1 and FsV-158. Values indicate the percent of the reference (above the diagonal) or the percent of the query (below the diagonal) sequence that was aligned to the query or the reference sequence, respectively, by the PROmer package of MUMmer4. Numbers refer to *P. fluviatile* EVEs. White tiles, no alignment.

nario is that EVE 94 is a remnant of a provirus closely related to EVE 22, which is not similar to EVE 12.

Virus gene expression

The expression patterns of genes within the *P. fluviatile* EVEs were measured under two different conditions: culture in freshwater or in seawater. The majority of the EVE genes were transcriptionally silent under both conditions, with genes being considered expressed only if their TPM was equal to or greater than 1 (Supplementary Fig. S8). EVE 21 had the highest proportion of expressed genes (Table 1).

The expression level of multiexonic genes within the seven EVEs was analysed to determine whether there was a correlation between intron acquisition and transcriptional activation, which would be consistent with both phenomena being features of endogenization of viral genes. The mean TPM (under freshwater conditions, Supplementary Table S4) for the intron-containing genes was higher than that for the monoexonic genes (1.7 ± 0.40 and 0.7 ± 0.09 mean plus standard error for 19 and 302 genes, respectively), but the difference was not statistically significant (Wilcoxon–Mann–Whitney test, Holm–Bonferroni adjusted *p*-value: 0.062).

Most of the EVE genes identified as being expressed were classed as proteins of unknown function based on BLASTp searches against the complete NCBI NR database and a search of Pfam (Supplementary Table S4). The next most abundant class consisted of proteins that matched *Phaeovirus* homologues of unknown function (Supplementary Table S4). Only a small subset of the expressed genes (∼10 genes from across all EVEs) were assigned functions. In EVE 7, these genes were mainly located within a short (∼6 kb) Copia-type retrotransposon-like region and were categorized as reverse transcriptase, polyproteins, and integrase. These retrotransposon genes were highly expressed, with TPMs of between 5.7 and 10.6 (Supplementary Table S4). The only expressed putative *Phaeovirus* gene in EVE 7 that was orthologous to a *Phaeovirus* gene encoded a thaumatin-like protein. Expressed EVE 21 genes encoded a tetratricopeptide repeat protein, a *Phaeovirus* DNA polymerase, an ankyrin repeat protein, and a methyltransferase. A tetratricopeptide repeat gene was also expressed in EVE 22, in addition to a gene encoding a *Phaeovirus* replication factor 2.

Discussion

In this study, we have characterized seven *Phaeovirus*-like EVEs that are present in a single brown algal genome. The EVEs are of different sizes and have different NCV core gene contents, ranging from complete or near-complete *Phaeovirus* genomes (EVEs 7, 12, and 22) to large (EVEs 21 and 40-1) and small (EVEs 40-2 and 94) fragments (Fig. 1; Table 1). Only one EVE possesses a minimum set of intact NCV core genes (EVE 7) and the remaining EVEs are probably nonfunctional due to the absence or degeneration of core genes (Fig. 1). Multiple *Phaeovirus* infections have been detected previously using a PCR approach (Stevens et al. 2014), but in this study, we observed a greater evolutionary diversity of inserted viral sequences, as they included members of both subgroup A (EVEs 12, 21, 22, and 94) and subgroup B (EVEs 7 and 40-1); EVE 40-2 was unclassified because it has no core genes suitable for phylogeny. The availability of a complete genome sequence also allowed us to obtain a more complete description of the EVEs than in most previous studies. For example, we observed the presence of NCV core genes such as major capsid protein (MCP) in apparently nonfunctional proviruses, indicating that PCR detection of core viral genes does not necessarily demonstrate that an active *Phaeovirus* is present.

Overall, the *P. fluviatile* EVEs were mostly transcriptionally silent (Supplementary Table S4, Supplementary Fig. S8). This has also been shown to be the case for the provirus present in the *Ectocarpus* species 7 genome (Cock et al. 2010, Cormier et al. 2017). Interestingly, ChIP-seq experiments have shown that nucleosomes in the region of the *Ectocarpus* species 7 genome corresponding to the inserted provirus exhibit post-translational dimethylation of lysine 79 of histone H3 and trimethylation of lysine 20 of histone H4 (Bourdareau et al. 2021). Both of these histone modifications have been correlated with reduced gene expression (Bourdareau et al. 2021).

The lack of expression of EVE genes could correspond to a transcriptionally inactive phase of the virus lysogenic life cycle or could be silencing imposed by the host as a defence mechanism. To distinguish between these two possibilities, it will be necessary to analyse the transcriptional profile of latent, genome-inserted phaeoviruses that have been shown experimentally to be functional, i.e. to be capable of producing viral particles.

It is possible that at least part of the gene expression observed in this study corresponds to the latency transcription profile (i.e. expression of the genes that control the switch between the latent and lytic phases). Equivalent profiles for other viruses are diverse and complex, often involving multiple viral and host factors (Traylen et al. 2011). In this context, the main expressed genes of interest, detected in this study, were a partial NCV DNA polymerase gene in EVE 21 and a thaumatin and a putative retrotransposon-like element in EVE 7. The functions of these genes warrant further study, especially as thaumatin has been implicated in host defence (Delaroque et al. 2001).

The sets of NCV core genes detected in the *P. fluviatile* EVEs are typical of those previously observed for phycodnaviruses. Novel genes that had not been previously described in phaeoviruses include noncore NCVOGs and genes with little or no NCV conservation. For example, a large number of histidine kinase genes (nine genes) was detected in EVE 7 and these included histidine kinases from both *Phaeovirus* subgroups A and B. It has been previously hypothesized that there are *Phaeovirus* lineages that must encode a larger set of histidine kinases, including those from both subgroups (Delaroque et al. 2003). Other noncore NCVOGs include genes involved in DNA metabolism and an expanded set of carbohydrate metabolism genes in EVE 12 (Supplementary Tables S1 and S2), many of which were previously found in mimiviruses. The presence of these carbohydrate metabolism genes suggests that some phaeoviruses may have unexplored mechanisms for adding sugar groups to capsid components, similar to those reported for chloroviruses (Van Etten et al. 2017). Despite being phylogenetically related to previously characterized phaeoviruses, EVE 12 has a surprising number of such genes.

The most novel genes in the *P. fluviatile* EVEs are homologous to proteins with roles in defence against pathogens and in electron transfer (Supplementary Table S2). Two EVEs (12 and 40-1) share a homolog of the cellular carotenoid synthesis gene, which has been suggested to play a role in altering host cell processes in other phycodnaviruses (Needham et al. 2019). One surprising observation was the presence of a novel H4 histone-like gene in EVE 40-1, which was likely acquired by an HGT event independent of that reported for NCVs known as marseilleviruses (Supplementary Fig. S10). The EVE 40-1 H4 histone-like protein is not closely related to that of marseilleviruses and is more similar to the H4 histone of polydnaviruses. Therefore, this may be a case of convergent evolution in which several virus groups (phaeoviruses, marseilleviruses, and polydnaviruses) may have independently acquired histone H4 to modify host cell processes (Gad and Kim 2008). Overall, analysis of the *P. fluviatile* EVEs highlighted an under-explored level of diversity among phaeoviruses, but the gene contents of the EVEs also indicated that viruses with divergent strategies may exist within the same host.

Compared to viruses with lytic life strategies, viruses with latent strategies are expected to have more stable genomes with slower rates of evolutionary divergence due to them being inherited vertically between host generations. In some instances, viruses with latent strategies may even abandon the lytic phase altogether, as is the case for bracoviruses. It is not yet clear to what extent phaeoviruses benefit from being vertically inherited along with the host algal genome because there are processes that disrupt integrated phaeoviruses, as shown clearly in the *P. fluviatile* genome. Presumably, this leads to fragmentation of the viral genome, loss or inactivation of NCV core genes, and loss of overall viral function.

The processes responsible for viral inactivation and fragmentation remain unclear, but several hypotheses can be proposed. First, an inserted viral genome may become nonfunctional, e.g. due to aberrations occurring during the insertion process or to insertion of a transposon into an essential gene following insertion of the virus. The inactive virus might then be expected to degenerate and fragment over time. Interestingly, functional annotation of some expressed genes (approximately seven genes; Supplementary Table S4) within EVEs indicated that they may be components of transposable elements, which may be an indicator of a viral element becoming part of the host genome. Second, inactivation and fragmentation may be a consequence of systems actively deployed by the brown algal host that allow it to interfere with infections involving insertion of proviruses into the genome. Inhibitor experiments have indicated that methylation of inserted viral DNA silences expression of viral genes in the ichthyosporean protist *Amoebidium appalachense* (Sarre et al. 2024). Most brown algae, including *P. fluviatile*, have lost DNA methyltransferase 1 (DNMT1; Denoeud et al. 2024), but low levels of DNA methylation have been detected in some species, where they are thought to be mediated by DNMT2 (Fan et al. 2020). Therefore, DNA methylation could potentially be a mechanism to suppress the activity of inserted viral genomes in *P. fluviatile*. In another brown alga, *Ectocarpus* species 7, peaks of two histone post-translational modifications that have been correlated with reduced gene expression, dimethylation of lysine 79 of histone H3 and trimethylation of lysine 20 of histone H4, were detected at the inserted EsV-1 provirus (Bourdareau et al. 2021), which has been shown to be transcriptionally silent (Cock et al. 2010, Cormier et al. 2017). It is therefore possible that histone post-translational modifications may also play a role in the host alga's defence against inserting viruses, particularly in species with little or no DNA methylation.

The fate of fragments of integrated viral genomes is another intriguing question. Such viral fragments may serve as a reservoir of viral genes that could be acquired, via recombination, by an active virus during the time it is integrated into the host genome. Also, the viral genes in these fragments could be recruited by the host, and such a process may be ongoing for the shorter EVEs in the *P. fluviatile* genome such as EVE 21, which includes several expressed viral genes. This type of process may be an important route for HGT between NCVs and their hosts. EVEs represent a significant portion of the genome of the *P. fluviatile* strain sequenced here (0.7% of the genome, 6.5% of the genes) providing considerable scope for HGT events to occur. The large proportion of EVEs in the *P. fluviatile* genome also raises the question as to whether accumulation of inserted viral genomes might lead to a significant level of genetic load or disruption of host gene organization, even if the inserted sequences are silenced, and whether the algal host might need to have mechanisms to address such a problem.

It has been suggested that at least some phaeoviruses may be capable of regenerating a complete viral genome by recombining viral fragments located at different sites in the host genome (Delaroque and Boland 2008) and that this process may be part of the normal viral replication strategy. This phenomenon is very rare in viruses and, so far, has only been described for plant pararetroviruses (Aoki et al. 1999). The data presented here tend not to support this hypothesis for phaeoviruses because at least one *P. fluviatile* EVE is potentially a complete viral genome and because the EVEs shared orthologous NCV core genes, including single copy genes such as DNA polymerase, suggesting that each EVE is derived from an independent insertion event. Moreover, based on genomic structure and gene content, it does not seem possible that two or more of the EVEs could be combined to reform a functional virus (Fig. 1).

Finally, several of the EVEs show clear signs of host endogenization and, consequently, these elements are unlikely to still be capable of functioning as part of a viral infection cycle. The processes of EVE fragmentation and activation of gene expression described earlier are indications of endogenization, as is the acquisition of introns by several viral genes (EVEs 21, 22, 40-2, and 94; Figs 1 and 2; Table 1). Interestingly, EVEs that contained multiexonic genes also tended to have percent GC contents that were similar to that of the host genome (Table 1). It is not clear whether these EVEs already had a GC content similar to the host when they were inserted into its genome or whether their GC content changed after insertion. Several processes are thought to drive changes in the GC content (Meunier and Duret 2004, Glémin et al. 2014), and these mechanisms could theoretically have led to viral insertions acquiring a similar GC content to the surrounding host DNA. However, given that selection to maintain intact coding regions would be weak because inserted viral genes are usually silenced and, in principal, do not carry out any essential functions for the host, it is difficult to see how this conversion could have occurred without creating numerous mutations, resulting in rapid gene degradation. We therefore favour the hypothesis that differences in percent GC contents were features of the inserted viral genomes before they were inserted into their host genome, but further work is needed to investigate this interesting observation.

The above endogenization-related features may provide a means to very approximately date viral genome insertions. For example, EVEs 7, 12, and 40-1 lack these endogenization-related features and are therefore likely to correspond to more recent integrations than the EVEs that do exhibit signs of endogenization. Detailed genomic study of EVEs in additional host individuals of the same species might help test this hypothesis. For example, if, as we propose, EVE 7 corresponds to an active virus whereas the other EVEs are ancient, inactive remnants, we might expect the latter to be found at conserved loci in diverse *P. fluviatile* individuals, whereas EVE 7-related sequences would be expected to be present in only a subset of individuals or to be found at multiple loci (representing multiple independent insertions).

Clearly, phaeoviruses and brown algae are engaged in an intriguing system of novel host–virus evolutionary relationships, and future analyses involving both experimental virology and genomic approaches will be necessary to answer the many questions raised by this study.

Acknowledgements

We thank the Institut Français de Bioinformatique and the Roscoff Analysis and Bioinformatics for Marine Science platform ABiMS (http://abimssb-roscoff.fr) for providing computing and data storage resources.

Author contributions

D.M. carried out the analyses and wrote the initial manuscript draft. A.C. carried out gene expression analyses. D.S. provided expert input and helped analyse the data. A.C. contributed genomic data. N.T. contributed genomic data. J.M.C. and E.C.

obtained funding and supervised the project. All authors reviewed and edited the manuscript.

Supplementary data

Supplementary data is available at *VEVOLU Journal* online.

Conflict of interest: None declared.

Funding

This work was supported by the Région Bretagne's Stratégie d'Attractivité Durable program (project Algannot) and the Agence Nationale de la Recherche project Idealg (ANR-10-BTBR-04-01), which jointly financed a postdoctoral fellowship for D.M., and by the France Génomique National Infrastructure project Phaeoexplorer (ANR-10-INBS-09), the Centre National de la Recherche Scientifique, and Sorbonne University.

Data availability

The *P. fluviatile* genome has been deposited in the EBI/ENA database under the project accession number PRJEB76691 and is available from the Phaeoexplorer website (https://phaeoexplorer.sb-roscoff.fr/). All codes used in the bioinformatic analyses are available at https://github.com/dmckeow/bioinf/tree/547d015d66ae3d55e12a7358e58d245e387eaf6f/bin/SBR.

References

Akita S, Vieira C, Hanyuda T *et al.* Providing a phylogenetic framework for trait-based analyses in brown algae: phylogenomic tree inferred from 32 nuclear protein-coding sequences. *Mol Phylogen Evol* 2022;**168**:107408.

Aoki T, Mohs G, Gonokami MK *et al.* Influence of exciton-exciton interaction on quantum beats. *Phys Rev Lett* 1999;**82**:3108–11.

Aylward FO, Moniruzzaman M. ViralRecall—a flexible command-line tool for the detection of giant virus signatures in 'omic data. *Viruses* 2021;**13**:Article2.

Bäckström D, Yutin N, Jørgensen SL *et al.* Virus genomes from deep sea sediments expand the ocean megavirome and support independent origins of viral gigantism. *mBio* 2019;**10**:10–1128.

Bourdareau S, Tirichine L, Lombard B *et al.* Histone modifications during the life cycle of the brown alga *Ectocarpus*. *Genome Biol* 2021;**22**:12.

Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 2021;**18**:366–68.

Cock JMM, Sterck L, Rouzé P *et al.* The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* 2010;**465**:617–21.

Cormier A, Avia K, Sterck L *et al.* Re-annotation, improved large-scale assembly and establishment of a catalogue of noncoding loci for the genome of the model brown alga *Ectocarpus*. *New Phytol* 2017;**214**:219–32.

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* 2008;**8**:110.

Delaroque N, Boland W, Müller DG *et al.* Comparisons of two large phaeoviral genomes and evolutionary implications. *J Mol Evol* 2003;**57**:613–22.

Delaroque N, Müller DG, Bothe G *et al.* The complete DNA sequence of the *Ectocarpus siliculosus* virus EsV-1 genome. *Virology* 2001;**287**:112–32.

Denoeud F, Godfroy O, Cruaud C *et al.* Evolutionary genomics of the emergence of brown algae as key components of coastal ecosystems. *Cell* 2024;**187**:6943–65.

Erives A. Eukaryotic core histone diversification in light of the histone doublet and DNA topo II genes of Marseilleviridae. bioRxiv, 2015:022236,

Fan X, Han W, Teng L *et al.* Single-base methylome profiling of the giant kelp *Saccharina japonica* reveals significant differences in DNA methylation to microalgae and plants. *New Phytol* 2020;**225**:234–49.

Feschotte C, Gilbert C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat Rev Genet* 2012;**13**: 283–96.

Gad W, Kim Y. A viral histone H4 encoded by *Cotesia plutellae* bracovirus inhibits haemocyte-spreading behaviour of the diamondback moth, *Plutella xylostella*. *J Gen Virol* 2008;**89**:931–36.

Gallot-Lavallée L, Blanc G. A glimpse of nucleo-cytoplasmic large DNA virus biodiversity through the eukaryotic genomics window. *Viruses* 2017;**9**:17.

Glémin S, Clément Y, David J *et al.* GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. *Trends Genet* 2014;**30**:263–70.

Goodacre N, Aljanahi A, Nandakumar S *et al.* A Reference Viral Database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. *mSphere* 2018;**3**:1–18.

Guy L, Kultima JR, Andersson SGE *et al.* GenoPlotR: comparative gene and genome visualization in R. *Bioinformatics* 2011;**27**: 2334–35.

Iyer LM, Aravind L, Koonin EV. Common origin of four diverse families of large eukaryotic DNA viruses. *J Virol* 2001;**75**: 11720–34.

Kawai H, Hanyuda T, Henry EC. Transfer of *Pilinia* from Ectocarpales to Ishigeales (Phaeophyceae) with proposal of Piliniaceae fam. Nov., and taxonomy of *Porterinema* in Ectocarpales. *Eur J Phycol* 2021;**57**:1–10.

Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015;**12**:357–60.

Koonin EV, Senkevich TG, Dolja VV. The ancient virus world and evolution of cells. *Biol Direct* 2006;**1**:29.

Koonin EV, Yutin N. Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. *Intervirology* 2010;**53**:284–92.

Letunic I, Bork P. Interactive Tree of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019;**47**:256–59.

Liu Y, Bisio H, Toner CM *et al.* Virus-encoded histone doublets are essential and form nucleosome-like structures. *Cell* 2021;**184**:4237–4250.e19.

Marçais G, Delcher AL, Phillippy AM *et al.* MUMmer4: a fast and versatile genome alignment system. *PLoS Comput Biol* 2018;**14**:1–14.

Maumus F, Epert A, Nogué F *et al.* Plant genomes enclose footprints of past infections by giant virus relatives. *Nat Commun* 2014;**5**: 4268.

Meints RH, Ivey RG, Lee AM *et al.* Identification of two virus integration sites in the brown alga *Feldmannia* chromosome. *J Virol* 2008;**82**:1407–13.

Meunier J, Duret L. Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol* 2004;**21**:984–90.

Moniruzzaman M, Weinheimer AR, Martinez-Gutierrez CA *et al.* Widespread endogenization of giant viruses shapes genomes of green algae. *Nature* 2020;**588**:141–45.

Müller DG, Knippers R. Phaeovirus. In: Tidona C, Darai G (eds), *The Springer Index of Viruses*, 2nd edn. Springer, Berlin, Heidelberg, 2011, 1259–64.

Needham DM, Yoshizawa S, Hosaka T *et al.* A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proc Natl Acad Sci USA* 2019;**116**:20574–83.

Patro R, Duggal G, Love MI *et al.* Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nat Methods* 2017;**14**:417–19.

Pertea M, Pertea GM, Antonescu CM *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnol* 2015;**33**:290–95.

Piacente F, Marin M, Molinaro A *et al.* Giant DNA virus mimivirus encodes pathway for biosynthesis of unusual sugar 4-amino-4,6-dideoxy-D-glucose (Viosamine). *J Biol Chem* 2012;**287**:3009–18.

Robinson JT, Thorvaldsdóttir H, Winckler W *et al.* Integrative genomics viewer. *Nat Biotechnol* 2011;**29**:24–26.

Roux S, Enault F, Hurwitz BL *et al.* VirSorter: mining viral signal from microbial genomic data. *PeerJ* 2015;**2015**:1–20.

Ruiz Martínez E, Mckeown DA, Schroeder DC *et al.* Phaeoviruses present in cultured and natural kelp species, *Saccharina latissima* and *Laminaria hyperborea* (Phaeophyceae, Laminariales), in Norway. *Viruses* 2023;**15**:2331.

Sarre LA, Kim IV, Ovchinnikov V *et al.* DNA methylation enables recurrent endogenization of giant viruses in an animal relative. *Sci Adv* 2024;**10**:eado6406.

Schroeder DC, McKeown DA. Viruses of seaweeds. In: Hurst CJ (ed.), *Studies in Viral Ecology*, 2nd edn. Wiley, 2021, 121–38.

Schroeder DC, Park Y, Yoon HM *et al.* Genomic analysis of the smallest giant virus—*Feldmannia* sp. Virus 158. *Virology* 2009;**384**:223–32.

Stevens K, Weynberg K, Bellas C *et al.* A novel evolutionary strategy revealed in the phaeoviruses. *PLoS One* 2014;**9**:e86040.

Stothard P, Grant JR, Van Domselaar G. Visualizing and comparing circular genomes using the CGView family of tools. *Brief Bioinf* 2018;**20**:1576–82.

Traylen CM, Patel HR, Fondaw W *et al.* Virus reactivation: a panoramic view in human infections. *Future Virol* 2011;**6**:451–63.

Van Etten JL, Agarkova I, Dunigan DD *et al.* Chloroviruses have a sweet tooth. *Viruses* 2017;**9**:88.

Van Etten JL, Graves MV, Müller DG *et al.* Phycodnaviridae—large DNA algal viruses. *Arch Virol* 2002;**147**:1479–516.

Waern M. Rocky-shore algae in the Öregrund archipelago. *Svenska Växtgeografiska Sällskapet* 1952;**30**:1–352.

Wang L, Wu S, Liu T *et al.* Endogenous viral elements in algal genomes. *Acta Oceanol Sin* 2014;**33**:102–07.

Yutin N, Colson P, Raoult D *et al.* Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. *Virol J* 2013;**10**:1–13.

Yutin N, Wolf YI, Raoult D *et al.* Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. *Virol J* 2009;**6**:1–13.

Virus Evolution, 2025, 11(1), veaf009 , DOI: https://doi.org/10.1093/ve/veaf009, Advance Access Publication 27 February 2025, Research Article