Endogenization of ssDNA fragment from an infectious virus in the genome of the *Chaetoceros tenuissimus*

Yuki Hongo (hongoy@affrc.go.jp)
Japan Fisheries Research and Education Agency https://orcid.org/0000-0003-1434-1220
Kei Kimura
Saga University
Yoshihiro Takaki
Yukari Yoshida
Shuichiro Baba
Genta Kobayashi
Saga University
Keizo Nagasaki
Takeshi Hano
Yuji Tomaru

Article

Keywords:

Posted Date: April 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-418097/v1

Abstract

One of the smallest diatoms, *Chaetoceros tenuissimus*, maintains their population despite coexisting with infectious viruses during blooms. To further understand this relationship, here, we sequenced the *C. tenuissimus* NIES-3715 genome. A gene fragment of a replication-associated gene from its own infectious ssDNA virus (designated endogenous virus-like fragment, EVLF) was found to be integrated in a total of 41 Mbp of both haploid assemblies. In addition, the EVLF was transcriptionally active and conserved in nine other *C. tenuissimus* strains from different geographical areas, although the primary structures of their proteins varied. The phylogenetic tree further suggested that the EVLF was acquired by the ancestor of *C. tenuissimus*. A target site duplication, a hallmark for long interspersed nuclear element retrotransposons, flanked the EVLF. Therefore, the EVLF was likely integrated by a retrotransposon during viral infection. The present study used genome information provides further insights into the diatom-virus evolutionary relationship.

Introduction

Diatoms (Bacillariophyta) are an important group of oceanic eukaryotic phytoplankton accounting for approximately 40% of primary marine production[1, 2]. The TARA Oceans project, a global plankton sampling campaign, has highlighted the significance of diatoms in global biogeochemical cycles[3, 4]. Research on the dynamics of diatoms is, therefore, important to understand global ecosystems. Generally, the growth and photosynthetic activity of diatoms are higher than those of other phytoplanktons and are often responsible for the blooms found in coastal and upwelling regions, thus playing a role as a major food resource for zooplankton, larvae, and filter feeders[5, 6] among others. Diatom dynamics, in addition to predation, are primarily controlled by environmental factors, such as water temperature, light, and nutrients[5, 7]. Other than these abiotic factors, diatom populations are exposed to diverse biological stressors. Over the past three decades, numerous studies have suggested that viral infections are a major determinant of phytoplankton fate in aquatic environments[8]. Indeed, knowledge regarding diatom viruses has accumulated rapidly since their first report in 2004. Roughly two diatom-virus groups have been identified, namely single-stranded (ss) DNA and ssRNA diatom viruses[9]. Diatom cell death due to viral infections can be readily observed in the laboratory, with infected cultures dying off in a few days[9]. In nature, however, the diatom population does not exhibit a rapid decrease in numbers, even in the presence of these infectious viruses; therefore, they seem to have the ability to coexist in the same region[10, 11]. Hence, diatoms are thought to have evolved a mechanism to resist viral infections, as has been observed in other phytoplankton-virus systems.

The virus resistance mechanisms of eukaryotic microalgae have been reported by numerous researchers, e.g., variation in susceptibility in the host cell[12–14], blockade of intracellular virus genome replication[15], bacterially mediated virus resistance[16], and host cell physiological barriers[17, 18]. In addition, recent genomic and transcriptomic analyses have revealed virus resistance systems at higher resolutions. For example, variabilities in genomic islands containing the viral-attachment genes of the cyanobacteria

Prochlorococcus facilitate host-virus coexistence[19]. Moreover, for the smallest eukaryotic phytoplankton, *Ostreococcus tauri* (Mamiellophyceae), the size of the hypervariable region in chromosome 19 differs greatly among strains, which is assumed to be closely related to its susceptibility to infection by its dsDNA virus, OtV[20, 21]. Furthermore, in many host organism-virus relationships, whole- or partial- viral genome sequences present in the host DNA can act as a viral resistance factor, e.g., superinfection exclusion and RNA interference related mechanisms[22–25]. To gain a more in-depth understanding of the host-virus systems, recent studies have highlighted the importance of using host genome analyses.

The marine planktonic diatom, *Chaetoceros tenuissimus* Meunier (Bacillariophyta, Centrales), is rectangular in the girdle view and is one of the smallest (~ 5 µm) diatoms. This species is widely distributed and is observed in Japanese coastal waters[10], the Narraganset Bay[26], the Mediterranean Sea[27, 28], and the San Matıas Gulf[29]. A previous study showed that *C. tenuissimus* has a high growth rate of at least three divisions per day and blooms during spring and autumn to levels of ~ 10^7 cells/l[10]. To date, four different viruses capable of infecting *C. tenuissimus* have been isolated and characterised, two different ssDNA viruses (CtenDNAV type-I and type-II), and two different ssRNA viruses (CtenRNAV type-I and type-II)[9, 30]. However, the *C. tenuissimus* population in natural environments sustains its bloom size even in the presence of these viruses[10, 11]. The tolerance to viral infection, along with their fast growth rate, suggests that *C. tenuissimus* might play an important role in maintaining primary productions in coastal environments. Knowledge on the relationship between *C. tenuissimus* and its viruses has gradually accumulated from the viewpoint of growth-physiology studies based on traditional culture experiments[17, 31], however, studies focusing on the aspect of cell biology are lacking. Here, we have explored the utility of genomic sequencing for this diatom species to further the current understanding regarding host-virus interactions at the molecular level. We believe that novel genomic discoveries can provide critical insights into the evolution of the diatom-virus relationship.

Results

Genome assembly and gene prediction

A total of 16.6 Gigabases (Gb) of sequence, providing a 150-fold coverage of the genome sequence, was obtained using the Illumina Miseq and Nanopore MinION platforms and assembled into 41 megabases (Mb) in a total of 85 scaffolds, ranging in size from approximately 1 Kb to 4.46 Mb (Supplementary Table 1). Using the kmer-based statistic, the haploid genome size and its heterozygosity were estimated as 39.7 Mb and 1.56%, respectively (Supplementary Fig. 1), and the size corresponded to the assembly. The accuracy of the genome assembly was confirmed by comparing with the *T. pseudonana and T. oceanica* genomes using BUSCO software[32]. The 82 complete genes assigned by BUSCO were identified in the *C. tenuissimus* genome assembly, and the number was larger than that for the *Thalassiosira* species (Supplementary Table 2). Thus, a successful genome assembly of *C. tenuissimus* NIES-3715 was achieved. A summary of the genome structure is shown in Table 1. A total of 18,705 protein coding genes were predicted in the haploid nuclear genome. Of the predicted proteins, 14,860 had significant

similarities (e-value < 1e-5) to protein sequences in the non-redundant proteins database (nr), and 9,544 had recognisable Interpro domains. Complete chloroplast and mitochondrial genomes were identified from the assembly, with sizes of 116 Kb and 36 Kb containing 131 and 33 predicted genes, respectively (Table 1, Supplementary Figs. 2a and b). The chloroplast genome had synteny to a related species, *Chaetoceros simplex*[33], with an identify of 99.4 % (Supplementary Fig. 3a). In contrast, the mitochondrial genome in this study is the first reported in a *Chaetoceros* species (Supplementary Fig. 2b) and showed similarity to the mitochondrial genome of *T. pseudonana* (83.3% identity) which is classified with *Chaetoceros* as being Centrales (Supplementary Fig. 3b). All sequence reads obtained in this study were deposited in the DDBJ Sequence Read Archive under accession number DRA009158, and the assembly scaffolds for the nuclear genome, as well as the chloroplast and mitochondrial genomes of *C. tenuissimus* NIES-3715, were also deposited under accession numbers BLLK01000001-BLLK01000085, LC537471, and LC537470, respectively.

EVLF in the nuclear genome of C. tenuissimus strains

An EVLF, which was similar to a replication-associated protein in *C. tenuissimus* DNA virus, was found to be integrated between the predicted proteins in the nuclear genome (Fig. 1a). The 181 amino acids were encoded by 546 bases of the sequence and partially aligned with the sequence of the replication-associated protein in *C. tenuissimus* DNA virus SS12-43V (67% identity; Fig. 1b). Although EVLF translation was stopped in the middle by a termination codon (TGA), the amino acid sequence could be followed in the next translation frame, and was found to be similar in sequence up to amino acid residue 381 in the *C. tenuissimus* DNA virus (Fig. 1b), although a termination codon was also located in the middle of the 3' end (Fig. 1b). A poly-adenine (A) like sequence was observed downstream of the fragment and common sequence “CATAAAA” flanked the fragment (Fig. 1a).

PCR analysis using specific primers designed to amplify the EVLF over its 1.5 Kbp length revealed that nine out of 10 other strains of *C. tenuissimus* also possessed the EVLF in their genomes (Fig. 2a). However, no target region in *C. tenuissimus* strain B was amplified in this study. In this case, it is possible that the primer set used to amplify the 1.5 kb region did not bind to the target site in strain B because other primer sets used for RT-PCR were able to amplify the EVLF region (Fig. 2a). As a result, strain B was also found to possess the EVLF in its genome. From the sequences of the cloned amplified gene fragments, a total of 15 clones with distinct sequences were obtained. The ML tree of these protein sequences formed two clades (Fig. 2b). Moreover, in clade I, three clades were divided with over 93% of bootstrap probability (BP, Fig. 2b). While the 10 sequences in clade I had stop codons and/or frame shifts (Fig. 2b), the translated amino acid sequences were aligned with the other sequences as well as the virus replication-associated protein from *C. tenuissimus* (Supplementary Fig. 4). All the EVLFs from the *C. tenuissimus* strains were clustered with bacilladnavirus replication-associated proteins (Fig. 2c). In particular, the EVLFs were formed by the replication-associated proteins in the *C. tenuissimus* DNA virus type-V clade as a sister group, with moderate statistical support (BP = 85%; Fig. 2c).

Transcription of EVLF in C. tenuissimus NIES-3715

In the growth phase, the EVLF was found to be transcribed in three replicate cultures of *C. tenuissimus* NIES-3715; however, in two of the cultures, the EVLFs at 6 days were found to be more weakly transcribed than those at other days, although the actin gene was transcribed constantly during sampling (Fig. 3). Moreover, as no amplification of the negative control was observed, all potentially contaminating DNA was completely digested in the samples (Fig. 3).

Comparative genes among organisms

The orthologous genes among *C. tenuissimus*, *T. pseudonana*, *P. tricornutum*, and *Cyanidioschyzon merolae* were identified from a total of 18,705, 11,673, 10,408, and 4,803 protein sequences, respectively. The genes included in the 2,102 ortholog groups were common among these organisms (Fig. 4a). In the *C. tenuissimus* genes, 3,265 ortholog groups were common in both *T. pseudonana* and *P. tricornutum*, while in the other groups a total of 1,011 and 822 were common in *T. pseudonana* and *P. tricornutum*, respectively (Fig. 4a). The specific paralog genes in *C. tenuissimus* also formed 29 groups (Fig. 4a).

In the ortholog group, OG0000, 385 paralog genes in *C. tenuissimus* were most abundant among these organisms and its 376 amino acid sequences possessed a leucine rich repeat 5 domain (IPR026906; Supplementary Table 3). Subsequently, seven ortholog groups, OG0004, OG0006, OG0021, OG0024, OG0065, OG0094, and OG0121, noticeably formed groups with genes possessing reverse transcriptase domains (IPR013103 and IPR000477), although the number of genes that contained these domains were not equal to that of the paralog genes (Fig. 4b). In the above ortholog groups, except for OG0006, the genes in *C. tenuissimus* were more abundant than those of other organisms (Fig. 4b).

The number of genes that possessed a functional domain is shown in Fig. 4c. The genes that possessed a protein kinase domain (IPR000719) were found to be abundant in all organisms, while genes that possessed the following four distinct domains, leucine rich repeat 5 (IPR026906), zinc finger, MYND-type (IPR002893), heat shock factor (HSF)-type, DNA-binding (IPR000232), and ankyrin repeat-containing domain (IPR020683; Fig. 4c), were more abundant in *C. tenuissimus* than in those of the other three diatom species. Moreover, the number of genes that possessed the following eight domains in *C. tenuissimus* was notably higher than that in other organisms; leucine rich repeat 5 (IPR026906), zinc finger, RING-type (IPR001841), reverse transcriptase, RNA-dependent DNA polymerase (IPR013103), integrase, catalytic core (IPR001584), reverse transcriptase domain (IPR000477), peptidase M11, gametolysin (IPR008752), sulfatase (IPR000917), and notch domain (IPR000800; Fig. 4c).

Transposable elements

Although the total number of retroelements in *C. tenuissimus* was less than that in *P. tricornutum*, three retroelements of *C. tenuissimus* were detected at higher levels than those in other organisms: short interspersed nuclear elements (SINEs, 0.02%), long interspersed nuclear elements (LINEs, 0.64%), and Gypsy/DIRS1 of long terminal repeat elements (1.11%; Fig. 4d).

Discussion

Diatoms and their infectious viruses seem to be able to coexist in natural waters and are closely related to each other[10, 11]. To understand the ecological relationships between host diatoms and their viruses, we sequenced the *C. tenuissimus* genome. The haploid genome size in *C. tenuissimus* NIES-3715 was estimated to be 39.7 Mb, and the sequences were assembled into 41 Mb with 85 scaffolds. Although the assembly is not yet at the chromosome level, a total of 18,705 genes were predicted from the haploid assembly. Surprisingly, a predicted gene having a similarity to a putative replication-associated protein of its infectious ssDNA virus was found to be integrated between two coding sites (Fig. 1a). The predicted EVLF protein sequence was highly similar to the mid-region of the replication-associated protein in *C. tenuissimus* ssDNA virus SS12-43V, which infects and lyses *C. tenuissimus* NIES-3715. Moreover, a frameshift caused by a two-base deletion, as well as nonsense mutations, was observed in the sequence (Fig. 1b). Therefore, the integration of the virus replication-associated gene appears to have not fully occurred. Like *C. tenuissimus*, an integrated virus fragment has been shown in the diatom *P. tricornutum* genome and its transcription has been detected by an EST analysis[34]. However, this fragment is not similar to that in *C. tenuissimus* but instead is similar to a viral replication gene identified only from viral metagenomic data (data not shown). Therefore, this is the first case of the integration of an extant closely related infectious virus fragment into its host genome.

When a virus first infects the host, its replication relies on the host cellular machinery[35]. In particular, the genes of DNA viruses must be transcribed as mRNA[35]. The ssDNA diatom viruses encode three putative proteins in their genomes[30]. To integrate the virus gene into the host genome, the nucleic acids of both the host and virus should be in the same cell during the infection. There are several key molecular processes required for the successful integration of a viral gene into the host genome. A Bornavirus-like nucleoprotein gene has been found to be integrated into mammalian genomes, and this integration is thought to be mediated through mRNAs by retrotransposons, which are mobile genetic elements with their own reverse transcriptase (RT), such as LINEs[36]. The banana genome also contains an integrated infectious virus gene, and it is thought to have been integrated by the Ty3/gypsy retrotransposon, which is a retroelement with long terminal repeats[37]. In the *C. tenuissimus* genome, the number of proteins that possessed a RT domain was higher than that in *T. pseudonana*, *P. tricornutum*, and *F. solaris* (Fig. 4c). Similarly, the percentage of predicted retroelements in *C. tenuissimus* was higher than that in *T. pseudonana* and *F. solaris* (Fig. 4d). Notably, the percentage of LINEs in the *C. tenuissimus* genome was the highest among the different species (Fig. 4d). In addition, considering that the number of paralog genes possessing or lacking an RT domain in *C. tenuissimus* was higher than that in other species, it is assumed that the *C. tenuissimus* genome contains more retrotransposons (Fig. 4b). These results indicated that more retrotransposons have been copied and integrated into the genome, which might be a factor in explaining the large genome size of *C. tenuissimus*. It is, therefore, possible that the retrotransposons possessing the RT domain converted the mRNA of the virus replication-associated gene into cDNA during the infection process. However, this then leads to the question of how the cDNA of the virus replication-associated gene integrates into *C. tenuissimus* genome. In general, once retroelements enter the nucleus, an endonuclease encoded by a LINE makes a single-stranded endonucleolytic nick in genomic DNA at a degenerate consensus sequence (5'-TTTT/A-3', with “/” indicating the scissile

phosphate), exposing a poly-(T) tail and a 3' hydroxyl group that serves as a primer for the reverse transcription of retroelement mRNA, and subsequent integration[38, 39]. The integrated retroelements have a structural hallmark in that they are flanked by variable size target site duplications (TSDs) [38, 39]. This structural hallmark was observed in the region of the integration of EVLF in the *C. tenuissimus* strains (Supplementary Fig. 5). A poly-(A) like sequence was observed downstream of the stop codons, which corresponds to the terminus of the virus replication-associated gene (Supplementary Fig. 5). In addition, the endonuclease cleavage sites “5'-TTTTATG-3'” flanked the EVLF and were characterised as TSDs (Fig. 1a). In addition, a 5' truncation and point mutations within the integrated retrotransposon are characteristic of integration by LINEs[39]. These structural hallmarks were also observed in the EVLFs of the *C. tenuissimus* strains (Fig. 2b) but not in the *P. tricornutum* genome[34]. In the case of *C. tenuissimus*, this evidence suggested that the virus replication-associated gene was possibly integrated into the *C. tenuissimus* genome by host LINEs.

In the culture experiments, CtenDNAV type-II was inoculated into the host cells at the log phase, and the cell density decreased after the cells reached the stationary phase[30]. In short, the inoculated cells maintained a high growth rate, suggesting that *C. tenuissimus* can resist virus infection[40]. Contrary to *C. tenuissimus*, in many marine microbial host-virus relationships, the viral mortality of the host increases when there is a high growth rate of host cells[41–44]. Considering this phenomenon together with the fact that EVLF is transcribed (Fig. 3), we speculate that the EVLF might act as an antiviral immunity-like RNA interference (RNAi). RNA silencing has been shown to act as a defence response against viral infections in plants[45–47] and mosquitos[48], Moreover, this gene silencing mechanism functions in the diatom, *P. tricornutum*[49]. The *C. tenuissimus* genome also encoded genes important in RNAi function, such as genes encoding Dicer and Argonaute proteins (accession nos. GFH46084.1 and GFH61989.1, respectively). From the RT-PCR results, the EVLF expression levels for 6 days culture were lower than those for other culture periods. Therefore, there is a possibility that *C. tenuissimus* represses viral proliferation through the replication-specific RNAi machinery during periods of high growth rate.

ssDNA viruses infecting *C. tenuissimus* are classified into two types (type-T and type-V) based on the sequences of replication-associated proteins[50]. The EVLFs were certainly derived from the type-V clade of *C. tenuissimus* DNA virus (Fig. 2c). Although the geographic distribution of these strains is different in Japan (Supplementary Fig. 6), the EVLFs were highly conserved among strains and in both alleles (Fig. 2b). Moreover, nine out of 10 strains were found to have EVLF at the same locus (Fig. 2a). These results indicate that the ancestor of *C. tenuissimus* had acquired the replication-associated gene from one type of virus, and then might have acquired an antiviral immunity-like RNAi machinery to resist against infectious viruses. Consequently, the populations have survived to date, and the EVLFs remain encoded in their genome as a fossil. Although the enigma of the survival strategy against the infectious viruses used by *C. tenuissimus* still remains to be solved, it is hoped that the genome information from *C. tenuissimus* will shed further light on it.

In this study, we sequenced the *C. tenuissimus* genome and discovered that a replication-associated gene associated with its own infectious virus has been integrated into the genome. This discovery represents the first case of an extant, closely related infectious virus fragment being integrated into a host genome; meanwhile, the EVLF may repress viral proliferation by RNA interference during periods of high growth rate. Finally, our analysis of their relationship suggests a close evolutionary relatedness.

Methods

DNA extraction and DNA library construction

The *C. tenuissimus* strain NIES-3715 was isolated from Seto Inland Sea, Japan (Supplementary Fig. 6) in Aug 2002. To check for bacterial contamination, the cultures were observed using epifluorescence microscopy after staining with SYBR-Gold. Briefly, the lysate was fixed with glutaraldehyde at a final concentration of 1%, and SYBR-Gold (Thermo Fisher Scientific, Waltham, MA, USA) was added to each fixed sample at a final dilution of 1.0×10^{-4} of the commercial stock. The stained samples were filtered onto 0.2-μm polycarbonate membrane filters (Nuclepore membrane; Cytiva, Sheffield, UK), after which the filters were mounted on a glass slide with a drop of low-fluorescence immersion oil, and covered with another drop of immersion oil and a cover slip. The slides were viewed at 1000× magnification with an Olympus BX50 epifluorescence microscope. The axenic algal cultures were grown in a modified SWM3 medium enriched with 2 nM Na_2SeO_3 [51] under a 12/12-h light-dark cycle at 20°C. Light irradiance was 850 μmol m^{-2} s^{-1} using white LED illumination. The algal strain was cultured for 7 days. Approximately 3×10^6–5×10^6 cells/ml in the stationary phase were used for DNA extraction. The cells in the cultures were harvested by centrifugation at 860 × *g* and 4°C for 15 min, after which the cell pellets were stored at − 80°C until analysis. DNA was extracted from the samples using a DNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA), according to the manufacturer's instructions. DNA libraries for paired-end and mate-paired sequencing were constructed in accordance with KAPA Hyper Prep Kit (F. Hoffmann-La Roche Ltd., Basel, Switzerland) and Nextera Mate Pair Sample Prep Kit (Illumina, Inc. San Diego, CA, USA), respectively. These libraries were sequenced into 300 bp paired-end reads using MiSeq (Illumina, Inc.) at the Japan Agency for Marine-Earth Science and Technology, Yokosuka, Japan.

For long-read sequencing of genomic DNA using MinION (Oxford Nanopore Technologies, Oxford, UK), total nucleic acid was extracted from the pellet using the DNAs-ici!-F (RIZO Inc., Tsukuba, Japan), according to the manufacturer's protocol. To extract genomic DNA from the total nucleic acid sample, the sample was treated with RNase A (Nippon Gene, Tokyo, Japan) and subsequently purified with phenol/chloroform prior to construction of a DNA sequencing library. The sequencing library was constructed using the ligation sequencing kit (SQK-LSK109, Oxford Nanopore Technologies) and sequenced by MinION, according to the respective manufacturer's instructions. After sequencing, base-calling was performed with Albacore (v2.3.1, Oxford Nanopore Technologies).

De novo genome assembly

To estimate the genome size and heterozygosity, k-mer counting was performed using the short paired-end reads and the Jellyfish programme[52]. The histogram of 21mer counts was visualised using GenomeScope[53].

Hybrid assembly of all Illumina short reads and MinION reads was performed using MaSuRCA[54] (v3.3.0) with default parameters. The haploid genome sequence was constructed from the assembled genome using HaploMerger2[55] (v20180603) with default parameters. The assembly quality was evaluated by QUAST[56]. To evaluate the assembly accuracy, single-copy ortholog genes were searched using BUSCO[32] with alveolata_stramenophiles_ensembl datasets.

Gene prediction and annotation

To predict the gene regions in the genome sequence, we first obtained the complete open reading frames (ORFs) from RNAseq (DNA Data Bank of Japan (DDBJ) Sequence Read Archive under accession number DRA011082). In RNAseq, the *de novo* assembly procedure was performed following that described by Hongo et al[57]. The complete ORFs were extracted from the assembled sequences and translated using TransDecoder[58] with default parameters. Next, the quality controlled paired-end reads of RNAseq were mapped to the assembled genome using TopHat2[59] with default parameters. Using the mapping data and the protein sequences of complete ORFs, the gene model was predicted using BRAKER2[60] (v2.1.0) with default parameters. Proteins predicted from the gene model were annotated based on their homology to sequences in the nr database from NCBI using the BLASTP programme with a threshold e-value of < 1e-5, and protein domains were found using Interproscan with a threshold e-value of < 1e-5.

Confirmation of an EVLF in the nuclear genome

To analyse the EVLF in the *C. tenuissimus* genome, we used nine other strains of this diatom species other than strain NIES-3715 (Supplementary Fig. 6). One millilitre of a stationary growth phase *C. tenuissimus* culture was centrifuged at 17,400 × *g* for 3 min at 4°C. The resulting diatom cell pellets were then preserved at − 80°C until analysis. DNA was extracted from stored cell pellets using the DNeasy Plant Mini Kit (Qiagen), according to the manufacturer's instructions. The EVLF was amplified using a primer pair, ctEVLFout_v1_F: 5'-GCAAACACGTKTGTTGATATATCGG-3' and ctEVLFout_v1_R: 5'-CGATCCTCTTGAAGACCCAGT-3', (Fig. 1). PCR amplification was conducted in a reaction mixture with a 20 µl final volume, containing 0.5 µl DNA, 1 × BlendTaq buffer (Toyobo, Japan), 200 nM dNTPs, 0.2 µM of each primer, and 1 U BlendTaq DNA polymerase. PCR was conducted using GeneAmp PCR system 9700 with the following cycle parameters: 30 cycles of denaturation at 94°C for 30 s, annealing at 55°C for 30 s, and extension at 72°C for 30 s. The PCR products were then electrophoresed on 1% (w/v) agarose ME gels (Wako Pure Chemical Industries, Osaka, Japan), and the nucleic acids were visualised using Midori green nucleic acid stain (Nippon Genetics, Tokyo, Japan). PCR amplicons of approximately 1.5 kb were excised, and their nucleic acids were extracted (NucleoSpin® Gel and PCR Clean-up; Macherey-Nagel GmbH and Co., KG, Düren, Germany). The PCR products were ligated into the pGEM-T Easy vector (Promega, Madison, WI, USA) and transformed into *Escherichia coli* DH5α-competent cells (Toyobo,

Japan). Sequencing was conducted using the dideoxy method with ABI PRISM 3130 Genetic Analyzer (Thermo Fisher Scientific).

RT-PCR analysis

The axenic algal cultures of *C. tenuissimus* strain NIES-3715 were grown in a modified SWM3 medium under a 12/12 h light-dark cycle of ca. 500 to 600 μM of photons m^{-2} s^{-1}, using cool white, fluorescent illumination at 25°C for 3 days. For the RT-PCR analysis, preconditioned cultures were inoculated into 1-l of fresh SWM3 medium at a final density of 2.5×10^3 cells/ml using a 2-l polycarbonate Erlenmeyer flask (431255; Corning Inc, Glendale, AZ, USA). This experiment was performed in triplicate. The cultures were subsampled at the early logarithmic growth phase (day 1 and day 2) and at late logarithmic growth phase (day 4 and day 7). One each sampling day, diatom cells in the cultures were retained on 0.4-μm polycarbonate membrane filters (Nuclepore membrane; Cytiva). The number of diatom cells on the filters ranged from 10^7 to 10^8 cells per filter, which were frozen in liquid nitrogen and stored at − 80 ºC until analysis.

The retained filters containing the diatom cell samples were cut into small pieces in the TRIzol reagent (Thermo Fisher Scientific), and total RNA was extracted using a TRIzol Plus RNA Purification Kit (Thermo Fisher Scientific), with PureLink DNase (Thermo Fisher Scientific) digesting any contaminating DNA, in accordance with the manufacturer's instructions. Moreover, to completely digest any contaminating DNA, DNase treatment was performed using TURBO DNase free kit (Thermo Fisher Scientific). The quantity of the total RNA was measured using a Qubit RNA HS assay kit (Thermo Fisher Scientific). cDNA was constructed from 1 μg of total RNA using an oligo$(dT)_{15}$ primer and SuperScript IV Reverse transcriptase (Thermo Fisher Scientific), in accordance with the manufacturer's instructions.

The EVLF sequence was amplified from the constructed cDNA using *Ex Taq* hot start version (Takara, Shiga, Japan) using the following conditions: initial denaturation phase of 98 ºC for 1 min, followed by 30 cycles of 98 ºC for 10 s, 60 ºC for 30 s, and 72 ºC for 40 s. Actin was also amplified using the cDNA and the total RNA to be used as a positive and a negative control, respectively. The primers used for this analysis were as follows: ctEVLFin_v1_F, 5'-AAGAAGAAGAGTCGACTGGATCAAC-3'; ctEVLFin_v1_R, 5'-ACAATAACGGTCTCATGATTGAGC-3'; ctActin_F, 5'-CTGGATGTGTTCTTGATTCTGGAG-3'; ctActin_R, 5'-CTTAGACATACGCTCACTGATTCCTG-3'. The amplicon lengths using these primers pairs were 456 bp for the EVLF and 500 bp for actin.

Phylogenetic analysis of EVLFs and virus replication-associated genes

The genome sequence of EVLF was obtained for nine strains of *C. tenuissimus* (Supplementary Fig. 6) using the above genome sequence confirmation. To clarify the evolutionary relationship of the EVLFs, a maximum-likelihood (ML) tree analysis was conducted. First, comparing within the strains, the EVLF protein sequences, including all alleles, were aligned with a replication-associated protein from *C. tenuissimus* DNA virus SS12-43V (accession no. BBE21064.1) using MAFFT[61] (v7.212) with default parameters. All stop codons and the frame-shifted amino acids in the alignment were removed by manual editing. The best-fit evolutionary model for the optimum alignment was calculated using

ModelFinder[62] and the Akaike information criterion. The ML tree was inferred from an evolutionary model using RAxML[63] (v8.2.4) with 100 bootstrap replicates. Second, to compare with virus replication-associated proteins, the virus protein sequences in the NCBI database were retrieved based on similarity to the EVLF using the BLASTP programme. These accession numbers are shown in Supplementary Table 4. All retrieved protein sequences and the EVLF were aligned using MAFFT[61] with default parameters, and gaps were automatically trimmed using trimAl[64] using the '-automated1' command option and default settings for all the other options. The subsequent procedure was the same as that described above.

Identification of orthologous genes

Protein sequences from *T. pseudonana*, *P. tricornutum*, and *Cyanidioschyzon merolae* were retrieved from public databases (accession no. GCF_000149405.2, GCF_000150955.2, and GCF_000091205.1, respectively.) Orthologous gene groups in all the protein sequences, including those in *C. tenuissimus*, were found using OrthoFinder[65] with default parameters. Protein domains in the sequences of reference organisms were found using Interproscan using a threshold e-value of < 1e-5.

Prediction of transposable elements

Transposable elements (TEs) in the *C. tenuissimus* genome were predicted using RepeatModeler2[66] and RepeatMasker[67] programmes with default parameters. To compare the TEs statistically among diatoms, the genomes of *T. pseudonana*, *P. tricornutum*, and *F. solaris* were analysed using the same programmes and parameters and compared to TEs that have already been reported for these three genomes[68, 69].

Declarations

Data availability

Sequence data generated during the current study are available in DDBJ repository, under accession number DRA009158, and the assembly data analysed during the current study are also available in DDBJ repository, under accession numbers BLLK01000001-BLLK01000085 (nuclear genome), LC537471 (chloroplast genome), and LC537470 (mitochondrial genome).

Code availability

All software packages are described in the Methods with the corresponding versions and references. No custom code is used in this study.

Acknowledgments

This study was supported by Research Fellowships for Young Scientists and Grants-in-Aid for Young Scientists (A) (22688016) and KAKENHI (19H00956) from the Japan Society for the Promotion of Science.

Author contributions

Y.H. analysed the data and wrote the manuscript. K.K. and Y.To. designed and performed the research, analysed the data and wrote the manuscript. Y.Ta. and Y.Y. performed the sequencing and analysed the data. S.B., G.K., K.N. and T.H. contributed to culture preparation and sequencing.

Competing Interests

The authors declare no conflict of interest.

References

1. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998).
2. Nelson, D. M., Tréguer, P., Brzezinski, M. A., Leynaert, A. & Quéguiner, B. Production and dissolution of biogenic silica in the ocean: revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Global Biogeochem. Cycles* **9**, 359–372 (1995).
3. Benoiston, A. S. et al. The evolution of diatoms and their biogeochemical functions. *Phil. Trans. R. Soc. B* **372**, 20160397 (2017).
4. Tréguer, P. et al. Influence of diatom diversity on the ocean biological carbon pump. *Nat. Geosci.* **11**, 27–37 (2018).
5. Sarthou, G., Timmermans, K. R., Blain, S. & Tréguer, P. Growth physiology and fate of diatoms in the ocean: a review. *J. Sea Res.* **53**, 25–42 (2005).
6. Werner, D. Introduction with a note on taxonomy. *Botanical monographs* **13**, 1–17 (1977).
7. Raven, J. A. & Waite, A. M. The evolution of silicification in diatoms: inescapable sinking and sinking as escape? *New Phytol.* **162**, 45–61 (2004).
8. Mojica, K. D. A. & Brussaard, C. P. D. Factors affecting virus dynamics and microbial host-virus interactions in marine environments. *FEMS Microbiol. Ecol.* **89**, 495–515 (2014).
9. Tomaru, Y., Toyoda, K. & Kimura, K. Marine diatom viruses and their hosts: Resistance mechanisms and population dynamics. *Perspect Phycol.* **2**, 69–81 (2015).
10. Tomaru, Y., Toyoda, K. & Kimura, K. Occurrence of the planktonic bloom-forming marine diatom *Chaetoceros tenuissimus* Meunier and its infectious viruses in western Japan. *Hydrobiologia* **805**, 221–230 (2018).
11. Tomaru, Y., Fujii, N., Oda, S., Toyoda, K. & Nagasaki, K. Dynamics of diatom viruses on the western coast of Japan. *Aquat. Microb. Ecol.* **63**, 223–230 (2011).
12. Zingone, A. et al. Diversity in morphology, infectivity, molecular characteristics and induced host resistance between two viruses infecting *Micromonas pusilla*. *Aquat. Microb. Ecol.* **45**, 1–14 (2006).
13. Thomas, R. et al. Acquisition and maintenance of resistance to viruses in eukaryotic phytoplankton populations. *Environ. Microbiol.* **13**, 1412–1420 (2011).

14. Waters, R. E. & Chan, A. T. *Micromonas pusilla* virus: the virus growth cycle and associated physiological events within the host cells; host range mutation. *J. Gen. Virol.* **63**, 199–206 (1982).
15. Tomaru, Y., Mizumoto, H. & Nagasaki, K. Virus resistance in the toxic bloom-forming dinoflagellate *Heterocapsa circularisquama* to single-stranded RNA virus infection. *Environ. Microbiol.* **11**, 2915–2923 (2009).
16. Kimura, K. & Tomaru, Y. Coculture with marine bacteria confers resistance to complete viral lysis of diatom cultures. *Aquat. Microb. Ecol.* **73**, 69–80 (2014).
17. Tomaru, Y., Kimura, K. & Yamaguchi, H. Temperature alters algicidal activity of DNA and RNA viruses infecting *Chaetoceros tenuissimus*. *Aquat. Microb. Ecol.* **73**, 171–183 (2014).
18. Arsenieff, L. et al. First viruses infecting the marine diatom *Guinardia delicatula*. *Front Microbiol.* **9**, 3235 (2019).
19. Avrani, S., Wurtzel, O., Sharon, I., Sorek, R. & Lindell, D. Genomic island variability facilitates *Prochlorococcus*–virus coexistence. *Nature* **474**, 604–608 (2011).
20. Blanc-Mathieu, R. et al. Population genomics of picophytoplankton unveils novel chromosome hypervariability. *Sci. Adv.* **3**, e1700239 (2017).
21. Yau, S. et al. A viral immunity chromosome in the marine picoeukaryote, *Ostreococcus tauri*. *PLoS Pathog.* **12**, e1005965 (2016).
22. Bertsch, C. et al. Retention of the virus-derived sequences in the nuclear genome of grapevine as a potential pathway to virus resistance. *Biol. Direct* **4**, 1–11 (2009).
23. Bondy-Denomy, J. et al. Prophages mediate defense against phage infection through diverse mechanisms. *ISME J.* **10**, 2854–2866 (2016).
24. Koonin, E. V. Taming of the shrewd: novel eukaryotic genes from RNA viruses. *BMC Biol.* **8**, 1–4 (2010).
25. Middelboe, M. & Brussaard, C. P. D. Marine viruses: key players in marine ecosystems. *Viruses* **9**, 302 (2017).
26. Rines, J. E. B. The Chaetoceros Ehrenberg (Baciullariophyceae) flora of Narragansett Bay, Rhode Island, USA. *Bibl. Phycol.* **79**, 1–196 (1988).
27. Kooistra, W. H. C. F. et al. Comparative molecular and morphological phylogenetic analyses of taxa in the Chaetocerotaceae (Bacillariophyta). *Phycologia* **49**, 471–500 (2010).
28. Montresor, M., Di Prisco, C., Sarno, D., Margiotta, F. & Zingone, A. Diversity and germination patterns of diatom resting stages at a coastal Mediterranean site. *Mar. Ecol. Prog. Ser.* **484**, 79–95 (2013).
29. Sar, E. A., Hernández-Becerril, D. U. & Sunesen, I. A morphological study of *Chaetoceros tenuissimus* Meunier, a little-known planktonic diatom, with a discussion of the section Simplicia, subgenus Hyalochaete. *Diatom Res.* **17**, 327–335 (2002).
30. Kimura, K. & Tomaru, Y. Discovery of two novel viruses expands the diversity of single-stranded DNA and single-stranded RNA viruses infecting a cosmopolitan marine diatom. *Appl. Environ. Microbiol.* **81**, 1120–1131 (2015).

31. Kimura, K. & Tomaru, Y. Effects of temperature and salinity on diatom cell lysis by DNA and RNA viruses. *Aquat. Microb. Ecol.* **79**, 79–83 (2017).
32. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness. In: Kollmar M (eds). Gene Prediction. Methods in Molecular Biology, vol 1962. Humana:New York, NY, 2019, pp 227–245.
33. Sabir, J. S. M. et al. Conserved gene order and expanded inverted repeats characterize plastid genomes of Thalassiosirales. *PLoS One* **9**, e107854 (2014).
34. Liu, H. et al. Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. *BMC Evol. Biol.* **11**, 1–15 (2011).
35. O'Carroll, I. P. & Rein, A. Viral nucleic acids. *Encyclopedia of Cell Biology.* 10.1016/B978-0-12-394447-4.10061-6 (2016).
36. Horie, M. et al. Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* **463**, 84–87 (2010).
37. Gayral, P. et al. A single Banana streak virus integration event in the banana genome as the origin of infectious endogenous Pararetrovirus. *J. Virol.* **82**, 6697–6710 (2008).
38. Richardson, S. R. et al. The influence of LINE-1 and SINE retrotransposons on mammalian genomes. In: Craig NL, Chandler M, Gellert M, Lambowitz AM, Rice PA, Sandmeyer SB (eds). Mobile DNA III. ASM Press:Washington, DC, 2015, pp 1165–1208.
39. Beck, C. R., Garcia-Perez, J. L., Badge, R. M. & Moran, J. V. LINE-1 elements in structural variation and disease. *Annu. Rev. Genomics Hum. Genet.* **12**, 187–215 (2011).
40. Tomaru, Y., Yamaguchi, H. & Miki, T. Growth rate-dependent cell death of diatoms due to viral infection and their subsequent coexistence in a semi-continuous culture system. *Microbes Environ.* **36**, ME20116 (2020).
41. Nagasaki, K. & Yamaguchi, M. Effect of temperature on the algicidal activity and stability of HaV (*Heterosigma akashiwo* virus). *Aquat. Microb. Ecol.* **15**, 211–216 (1998).
42. Van Etten, J. L., Burbank, D. E., Xia, Y. & Meints, R. H. Growth cycle of a virus, PBCV-1, that infects *Chlorella*-like algae. *Virology* **126**, 117–125 (1983).
43. Mojica, K. D. A. & Brussaard, C. P. D. Factors affecting virus dynamics and microbial host–virus interactions in marine environments. *FEMS Microbiol. Ecol.* **89**, 495–515 (2014).
44. Middelboe, M. Bacterial growth rate and marine virus–host dynamics. *Microb. Ecol.* **40**, 114–124 (2000).
45. Moissiard, G. & Voinnet, O. Viral suppression of RNA silencing in plants. *Mol. Plant Pathol.* **5**, 71–82 (2004).
46. Waterhouse, P. M., Wang, M. B. & Lough, T. Gene silencing as an adaptive defence against viruses. *Nature* **411**, 834–842 (2001).
47. Voinnet, O. RNA silencing as a plant immune system against viruses. *TRENDS Genet.* **17**, 449–459 (2001).

48. Suzuki, Y. et al. Non-retroviral endogenous viral element limits cognate virus replication in *Aedes aegypti* ovaries. *Curr. Biol.* **30**, 3495–3506 (2020).
49. De Riso, V. et al. Gene silencing in the marine diatom *Phaeodactylum tricornutum*. *Nucleic Acids Res.* **37**, e96-e96 (2009).
50. Tomaru, Y. & Kimura, K. Novel protocol for estimating viruses specifically infecting the marine planktonic diatoms. *Diversity* **12**, 225 (2020).
51. Imai, I., Itakura, S., Matsuyama, Y. & Yamaguchi, M. Selenium requirement for growth of a novel red tide flagellate *Chattonella verruculosa* (Raphidophyceae) in culture. *Fish. Sci.* **62**, 834–835 (1996).
52. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
53. Vurture, G. W. et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
54. Zimin, A. V. et al. The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
55. Huang, S., Kang, M. & Xu, A. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics* **33**, 2577–2579 (2017).
56. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
57. Hongo, Y., Yabuki, A., Fujikura, K. & Nagai, S. Genes functioned in kleptoplastids of *Dinophysis* are derived from haptophytes rather than from cryptophytes. *Sci. Rep.* **9**, 1–11 (2019).
58. Haas, B. J. et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
59. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, 1–13 (2013).
60. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. Bioinform.* **3**, lqaa108 (2021).
61. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **3**, 3059–3066 (2002).
62. 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).
63. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
64. 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).
65. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 1–14 (2015).

66. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).

67. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2015; Retrieved 1st March 2021, from http://www.repeatmasker.org.

68. Tanaka, T. et al. Oil accumulation by the oleaginous diatom *Fistulifera solaris* as revealed by the genome and transcriptome. *Plant Cell* **27**, 162–176 (2015).

69. Maumus, F. et al. Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC Genomics* **10**, 1–19 (2009).

Table 1

Table 1. Summary of genome assembly

	C. tenuissimus NIES-3715	T. pseudonana	P. tricornutum
Nuclear genome			
Size (Mbp)	41.0	34.5	27.4
G+C content (overall %)	38.9	47	49
G+C content (coding %)	40.3	48	51
Protein-coding genes	18,705	11,242	10,402
Average gene size (bp)	1526	992	1527
Chloroplast genome			
Size (bp)	116,464	128,814	117,369
G+C content (overall %)	32.1	30.7	32.6
Protein-coding genes	131	144	130
tRNAs	75	33	30
Mitochondrial genome			
Size (bp)	36,047	43,827	77,356
G+C content (overall %)	30.8	30.1	35.0
Protein-coding genes	33	40	32
tRNAs	48	22	24

Figures

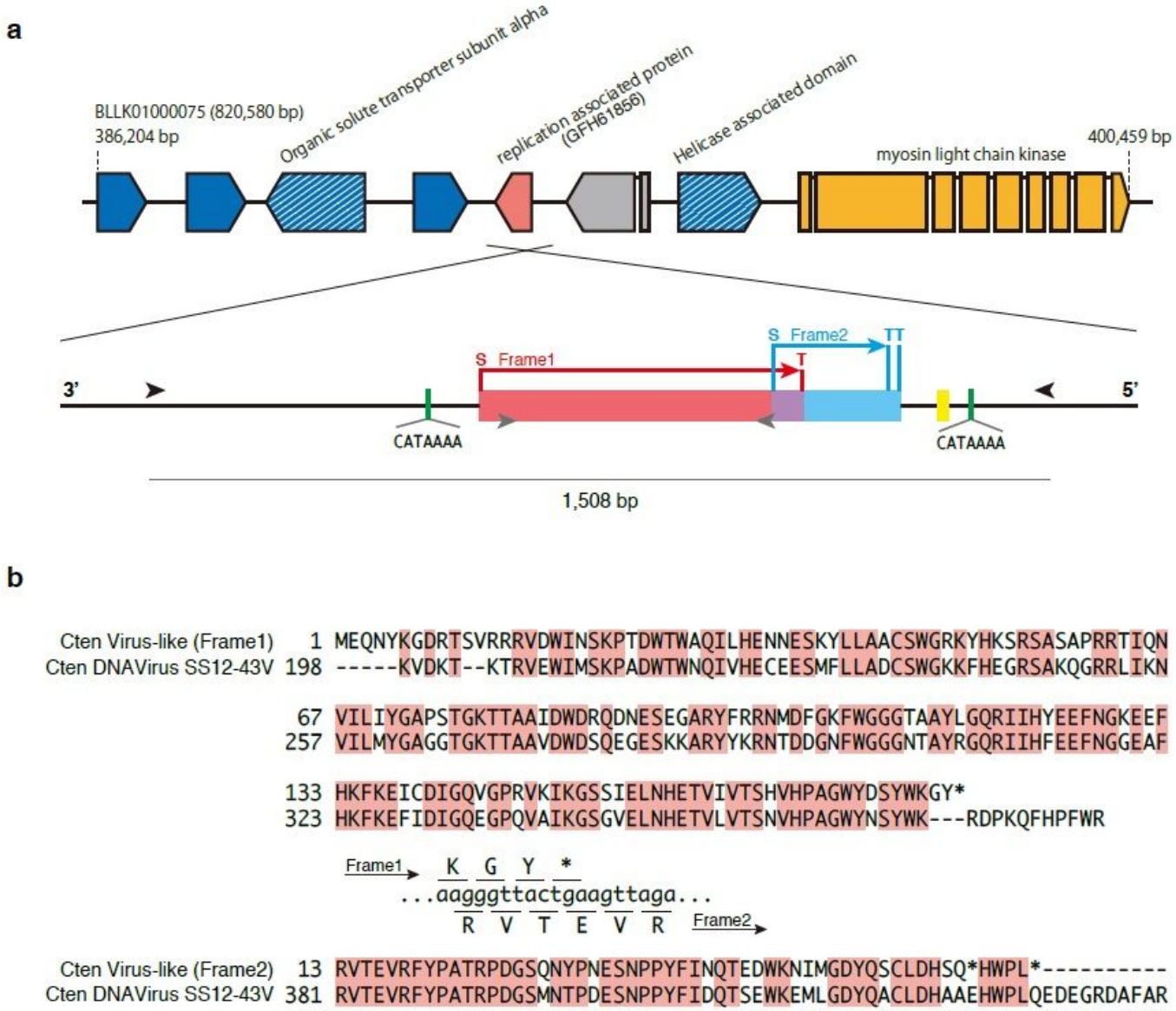

Figure 1

(a) Location and structure of the EVLF in the host genome. Gene directions are indicated by their arrowheads. Blue and blue with a slash indicate genes coding hypothetical proteins and hypothetical proteins possessing known domains, respectively. Orange indicates a myosin light chain kinase, and a pink indicates a virus-like fragment. A structure of the EVLF is shown in a close-up with reverse chain direction. The pink and blue boxes indicate each open reading frame which merged partially (purple), and the “S” and “T” on the boxes refer to the start and terminal codons, respectively. A yellow box indicates a poly-(A) like sequence, and green boxes indicate a “CATAAAA” sequence. Black and gray arrowheads

indicate the amplification primer sets, ctEVLFout_v1 and ctEVLFin¬_v1¬, which amplified the 1508 bp and 456 bp PCR products, respectively. (b) Alignment of amino acid sequences between the EVLF of C. tenuissimus NIES-3715 and a replication-associated protein in its infectious DNA virus. The alignment was constructed by MAFFT61. Pink boxes indicate consensus amino acid residues. Hyphens and asterisks indicate gaps and termination codons, respectively. A two-nucleotide deletion led to a separation into two open reading frames.

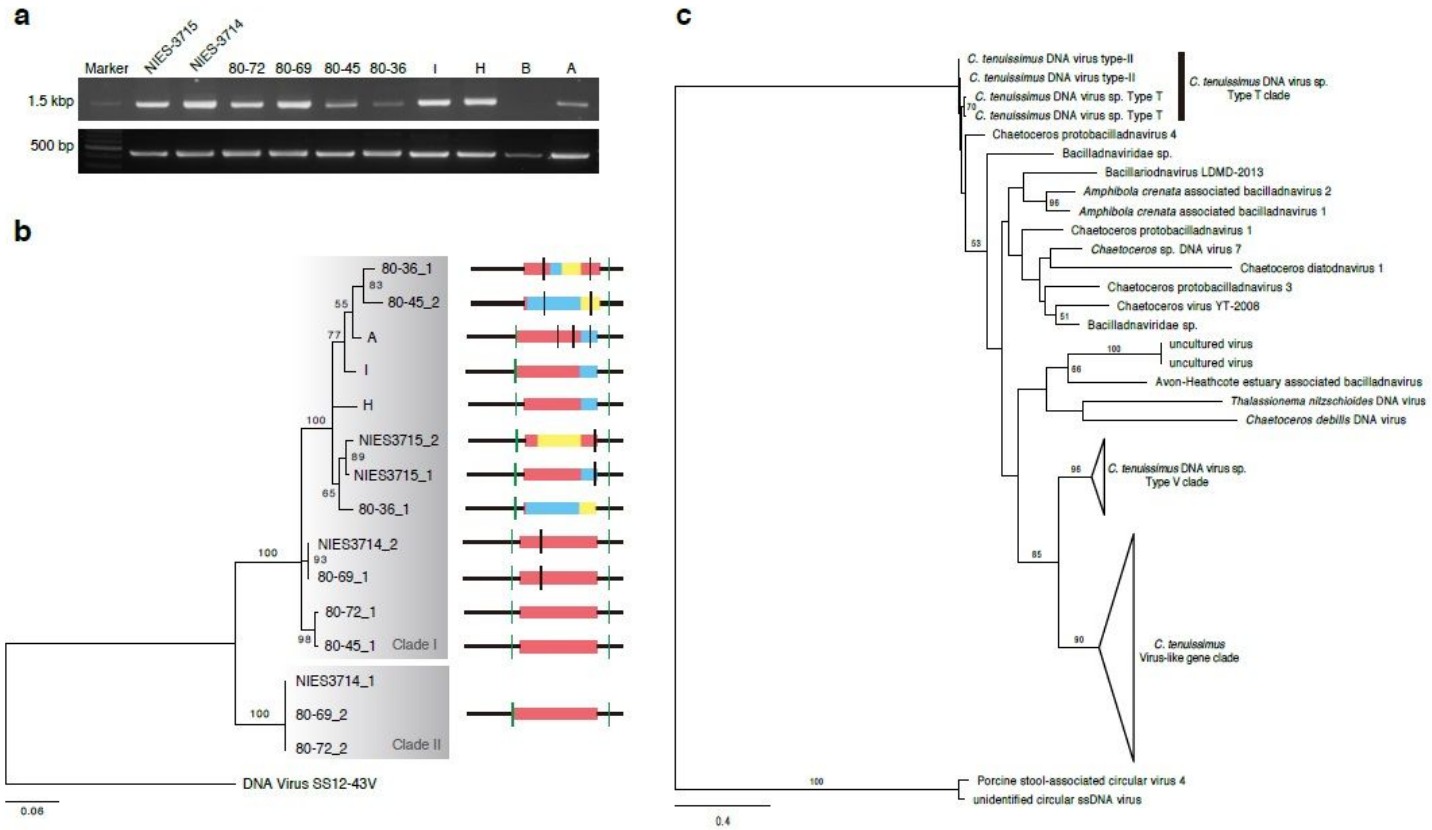

Figure 2

(a) Amplification of the EVLF from the genomes of C. tenuissimus strains, with the regions as indicated in Supplementary Figure 6. (b) Maximum-likelihood trees of EVLFs with the fragment structures. The red, blue, and yellow boxes indicate the 1st, 2nd, and 3rd frames, respectively. The green and black vertical lines indicate TSD sequences and the terminal codons, respectively. (c) Maximum-likelihood trees of EVLFs with the replication-associated protein from bacilladnavirus. Bootstrap values greater than 50% are shown.

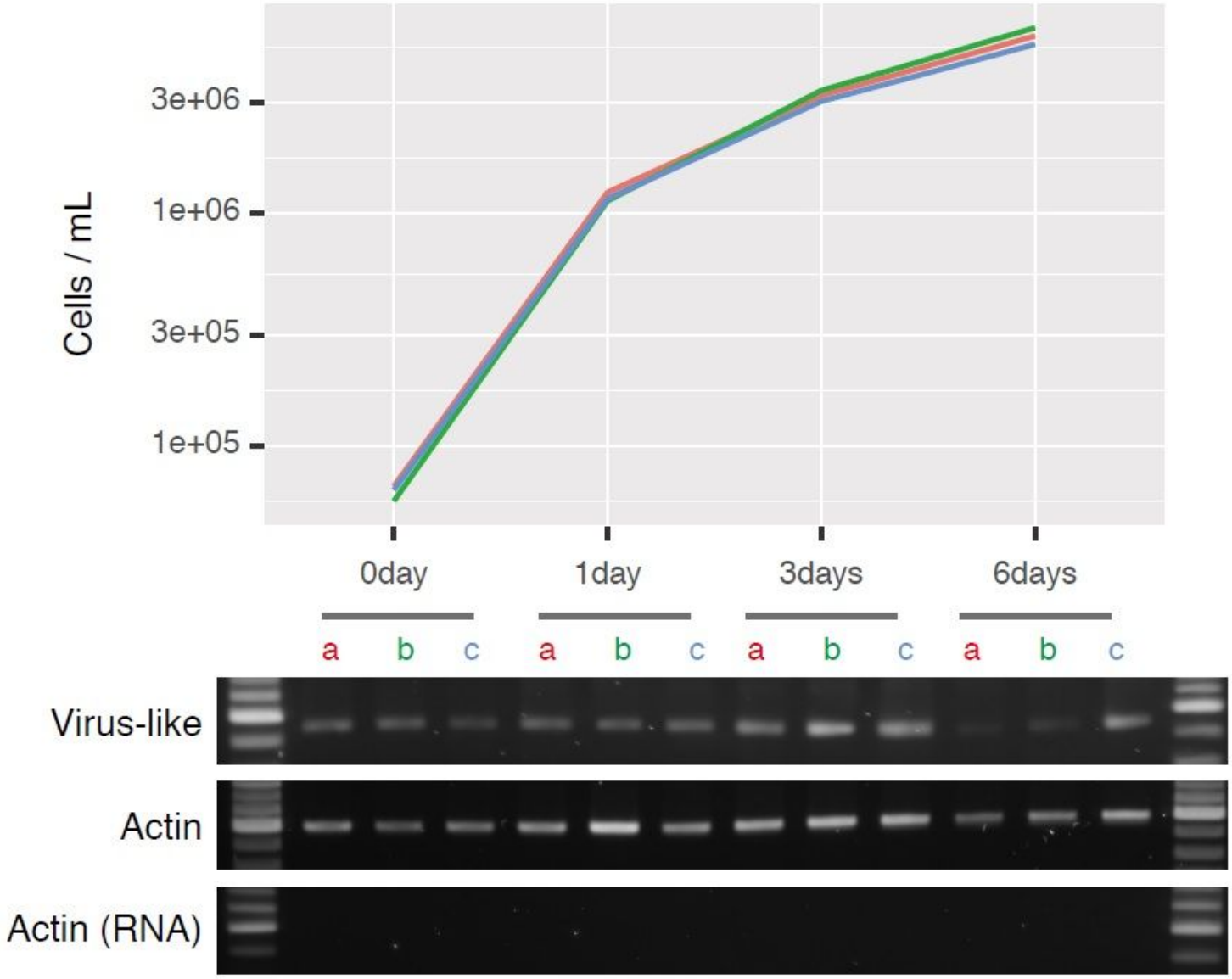

Figure 3

Cell densities after 6 days of culture and the expression pattern of the EVLF. Cultures were performed in triplicate flasks (a, b, c), and the line colours in the plot correspond to the respective flasks. The actin gene was amplified from the cDNA as a positive control but not from RNA as a negative control.

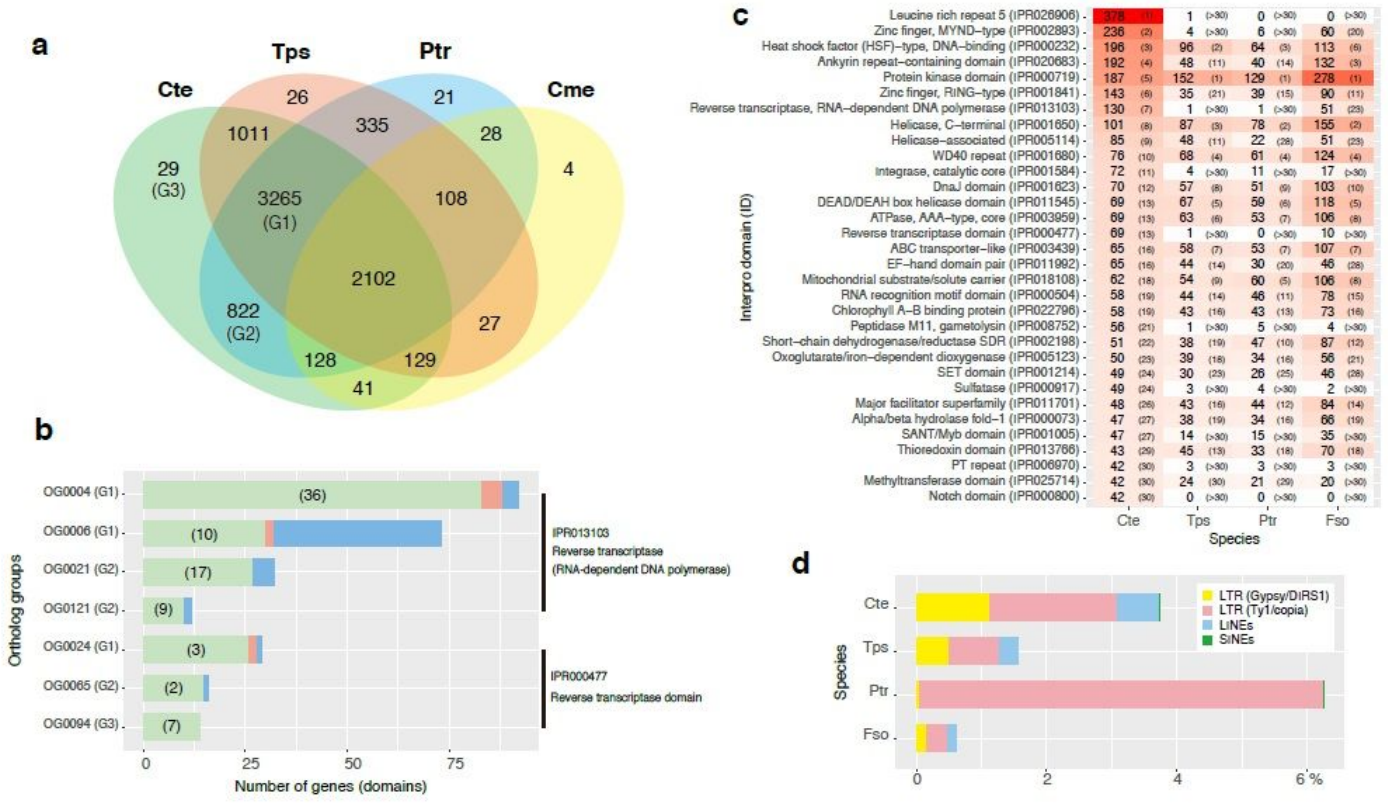

Figure 4

(a) Venn diagram of orthologous gene groups. Orthologous gene groups were identified from C. tenuissimus (Cte), T. pseudonana (Tps), P. tricornutum (Ptr), and C. merolae (Cme). The numbers in the Venn diagram indicate the number of orthologous gene groups. Reverse transcriptase domains (Interpro ID, IPR013103 and IPR000477) were included in the ortholog gene groups (G1, OG0004, OG0006, and OG0024; G2, OG0021, OG0121, and OG0065; G3, OG0094). (b) The number of genes in the above ortholog gene groups. The parentheses indicate the number of genes that possessed reverse transcriptase domains (IPR013103 and IPR000477). (c) Top 30 of the Interpro domains detected in the C. tenuissimus genome. The gene numbers are in descending order based on Cte and the numbers in parentheses indicate their ranking. The shade of red becomes lighter as the gene number decreases. (d) The percentage of retroelements in the four species genomes predicted by RepeatMasker.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.pdf
- SupplementaryFigures.pdf