A 1.5 Mb continuous endogenous viral element region in the arbuscular mycorrhizal fungus *Rhizophagus irregularis*

Hongda Zhao, Hiroyuki Hikida, Yusuke Okazaki, Hiroyuki Ogata*

Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan

* Corresponding author: Hiroyuki Ogata (ogata@kuicr.kyoto-u.ac.jp)

Abstract

Most fungal viruses are RNA viruses and no double-stranded DNA virus that infects fungi is known to date. A recent study identified DNA polymerase genes that originated from large dsDNA viruses in genomes of basal fungi, suggesting the existence of dsDNA viruses infecting fungi. In this study, we searched for viral infection signatures in chromosome-level assemblies of an arbuscular mycorrhizal fungus, *Rhizophagus irregularis*. We identified a continuous 1.5 Mb putative viral region on a chromosome of *R. irregularis* strain 4401. Phylogenetic analysis revealed that the viral region is related to members of family *Asfarviridae* in phylum *Nucelocytoviricota*. Single-copy marker genes of the *Nucelocytoviricota* were detected as single copies in the viral region. Furthermore, this viral region was not found in other four strains of *R. irregularis* and showed less invasive signals of fungal transposable elements than other genomic regions. These results suggest a recent and single insertion of a large dsDNA viral genome in this fungal strain, providing strong evidence for the recent infection of a dsDNA virus to the fungus.

Introduction

The fungal virosphere is dominated by RNA viruses, and few ssDNA viruses are known in phytopathogen fungi (Li et al., 2020; Yu et al., 2010). Fungal viruses (also called mycoviruses) were classified into 23 viral families. Among these, 22 RNA virus families contain 204 mycovirus species, and a single-stranded (ss) DNA virus family contains two mycovirus species (Kondo et al., 2022). No double-stranded (ds) DNA virus has been identified in fungi, but recent studies suggest their existence. A single virion sequencing study recovered genomes of dsDNA viruses belonging to *Nucleocytoviricota* (nucleocytovirus) on subsurface oceanic crustal fluids, where *Ascomycota* fungi are the main eukaryotes (Bhattacharjee et al., 2023). These viral genomes had genes of fungal origin, suggesting that the viruses infect the fungi. DNA polymerase genes likely originating from *Nucleocytoviricota* were also identified in genomes of basal fungi, including *Rhizophagus irregularis*, an arbuscular mycorrhizal fungus (Gong et al., 2020). *Nucleocytoviricota* is a phylum of viruses with large dsDNA genomes ranging from 70 kb to 2.5 Mb (Aylward et al., 2021). Their hosts cover a diverse array of eukaryotes, from protists to animals (Schulz et al., 2022).

Endogenous viral elements (EVEs) are a form or trace of viral genomes integrated into the host genome (Feschotte and Gilbert, 2012). Some previous studies identified genes from *Nucleocytoviricota* in the genomes of eukaryotes (Maumus et al., 2014; Maumus and Blanc, 2016; Yoshikawa et al., 2019). A recent study discovered giant EVEs (GEVEs) of nucleocytoviruses in genomes of green algae (Moniruzzaman et al., 2022, 2020). The total size of a GEVE can reach several hundred kilobases or even over a thousand kilobases, although it is often scattered across multiple contigs. Genomic integration of nucleocytoviruses into host genomes has significant implications for the evolution of eukaryotes, as up to 10% of open read frames (ORFs) can be originated from GEVEs (Moniruzzaman et al., 2020). Such evolutionary events may also act as a bridge for horizontal gene transfer among eukaryotic organisms (Cheng et al., 2021). Furthermore, multiple signals of virophages (parasites of large DNA

viruses) were reported in a variety of eukaryote genomes (Bellas et al., 2023), implying the vast diversity of unknown dsDNA virosphere among the eukaryotic radiation. Investigating EVEs helps explore the virosphere of eukaryotes and currently unidentified virus-host relationships, as in the previous reports of *Nucleocytoviricota* genes in the genomes of plants (moss and fern) (Maumus et al., 2014) and oomycetes (Hannat et al., 2021), from which no nucleocytovirus has been isolated.

One of the difficulties in identifying GEVE is the highly fragmented assemblies of eukaryotic genomes, leading to the fragmentation of GEVEs. Recently, chromosome-level assemblies of *R. irregularis* were reported (Yildirir et al., 2022), allowing us to search for more complete viral fragments in genomes of this species. The present study searched genomes of five strains of *R. irregularis* for *Nucleocytoviricota* signals and identified a 1.5 Mb GEVE region of a *Nucleocytoviricota*-like virus in a chromosome of strain 4401, which is the largest continuous GEVE region known to date. This GEVE exhibits homology to the genomes of *Asfarviridae*, a family of *Nucleocytoviricota*. The results provided strong evidence for the recent infection of this fungal strain by a putative nucleocytovirus, demonstrating that dsDNA viruses exist in the fungal virosphere.

Results

Datasets of *R. irregularis*

We collected five chromosome-level assemblies of *R. irregularis*, corresponding to five different strains (Yildirir et al., 2022) (Table 1). All these genomes were based on Illumina sequencing, Nanopore sequencing, and high-throughput chromatin conformation capture (Hi-C) sequencing. Long reads generated from Nanopore could be longer than a repeat or hypervariable region, which overcomes genomes' previous highly fragmented nature. Hi-C sequencing detects spatial proximity and was used for scaffolding.

Table 1 Genomic data of *R. irregularis*

Strain	Genome size (bp)	Number of scaffolds	Isolated place	Accession
A1	147,088,061	33	Switzerland	GCA_020716765.1
B3	146,830,121	33	Switzerland	GCA_020716685.1
DAOM-197198	147,209,168	33	Canada	GCA_020716725.1
C2	161,924,360	33	Switzerland	GCA_020716745.1
4401	146,854,905	33	Canada	GCA_020716705.1

A continuous 1.5 Mb GEVE region on a fungal chromosome

ViralRecall is an HMM-based tool designed for detecting viral regions based on *Nucleocytoviricota* orthologous groups, and ten *Nucleocytoviricota* marker genes were identified by this tool. By employing ViralRecall for the five genomes of *R. irregularis*, we identified a giant continuous viral signal as a 1,550 kb region (from 1,370,742 bp to 2,921,549 bp) on chromosome 8 of strain 4401 (Fig. 1a).

This 1.5 Mb viral region had distinctive sequence and structural features. Out of ten *Nucleocytoviricota* marker genes analyzed, the 1.5 Mb viral region contained five marker genes: B-family DNA polymerase (PolB), RNA polymerase large subunit (RNAPL), RNA polymerase small subunit (RNAPS), mRNA capping enzyme (mRNAc), and viral late transcription factor 3 (VLTF3) genes (Supplementary Table S1, Supplementary Figs S1–S6). The average GC content of the whole genomes of *R. irregularis* was 27.89%, whereas the viral region showed a GC content of 36.58%. Throughout the region, GC content was also higher than those of the remaining parts of the chromosome (Fig. 1b). Furthermore, Hi-C sequencing data indicate a higher level of condensation of DNA in the region compared with other genomic regions (Fig. 1d).

To confirm whether this viral region corresponds to a bona fide insertion in the fungal chromosome, we mapped

raw long reads to the fungal genome. Most regions of the chromosome showed comparable coverage including the viral region (Fig. 1c). We found many short regions with a high read coverage, but they mostly corresponded to repetitive elements. We identified a total of 259 long reads directly connecting the viral region and cellular regions (Fig. 1e). These results confirmed that the viral region represents a GEVE integrated into the fungal chromosome.

In the GEVE region, we identified 705 ORFs, corresponding to 27.39% coding density, representing a low level compared to other typical nucleocytoviruses. Of the 705 ORFs, 172 were functionally annotated. Among them, 53, 39, and 26 ORFs were related to the functions in "Replication, recombination, repair", "Signal transduction mechanisms", and "Transcription", respectively (Supplementary Fig. S7). Using PseudoFinder, we also identified 18 putative viral pseudogenes including RNAPL.

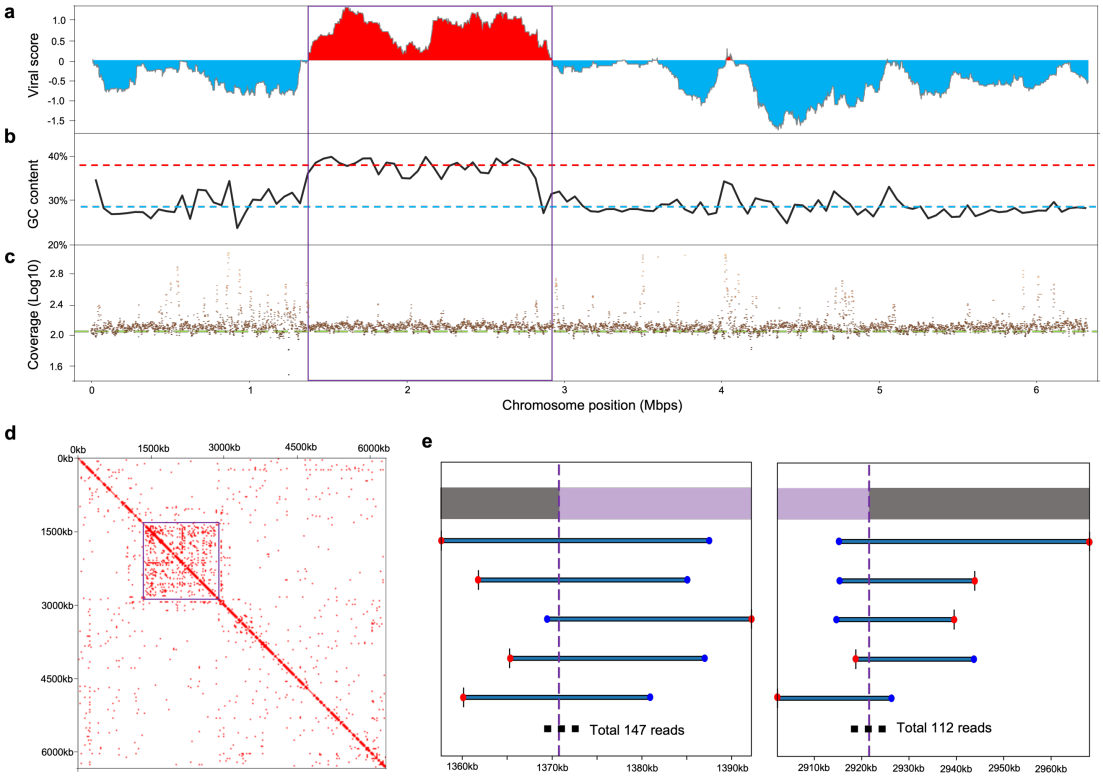


Fig. 1 The 1.5 Mb GEVE region in the fungal chromosome. (a) ViralRecall score of chromosome 8 of strain 4401. Viral scores were evaluated with a rolling window of 150 ORFs on the chromosome. The score of each ORF was based on HMM scores of the viral reference database and cellular reference database. A higher score means viral, and a lower score means cellular. **(b)** GC content along the chromosome. The window size is 50,000 bp. The red and blue dashed line represents the average GC content of the GEVE region (36.58%) and that of the remaining parts of the chromosome 8 (28.00%), respectively. **(c)** Average read coverages of the chromosome. Each dot represents the average coverage of 1,000 continuous base pairs. The green line represents the average sequencing depth (bases/genome size = 113×). **(d)** Heat map showing Hi-C signals of the chromosome 8. The resolution is 25 kb. Squares with signal values above 1 are marked in red, representing high contact probabilities. **(a–d)** The squares highlight the viral region. **(e)** Examples of long reads connecting the GEVE regions with the remaining parts of the chromosome. We present the five longest reads representatively. Blue and red points represent the start and end of the read, respectively. The purple vertical dashed line is the extremities of the GEVE region (1,370,742 bp in the left figure and 2,921,549 bp in the right figure). Purple and grey bards on the top indicate the GEVE and other cellular chromosomal regions, respectively.

The GEVE is specific to strain 4401

In all five fungal strains, a total of 59 putative viral regions were detected, including the 1.5 Mb GEVE region (Supplementary Table S2). However, all these regions except the 1.5 Mb GEVE region were shorter than 300 kb and contained only PolB genes (15 to 33 copies per genome) and/or mRNAc genes (one copy in strains A1 and B3)

(Supplementary Table S3). Therefore, we compared translated sequence level similarity of the chromosome 8 between the five strains of *R. irregularis* (Fig. 2). Most parts of the chromosome 8 of strain 4401 showed more than 90% sequence identity to the corresponding regions of the chromosome 8 of the other strains. However, other strains lacked the region corresponding to the 1.5 Mb GEVE region found in strain 4401. Consistently, the chromosome 8 of strain 4401 is about 1 Mb longer than the chromosome 8 of other strains (Table 2). These results indicate that the integration of the GEVE on the chromosome 8 is unique to strain 4401.

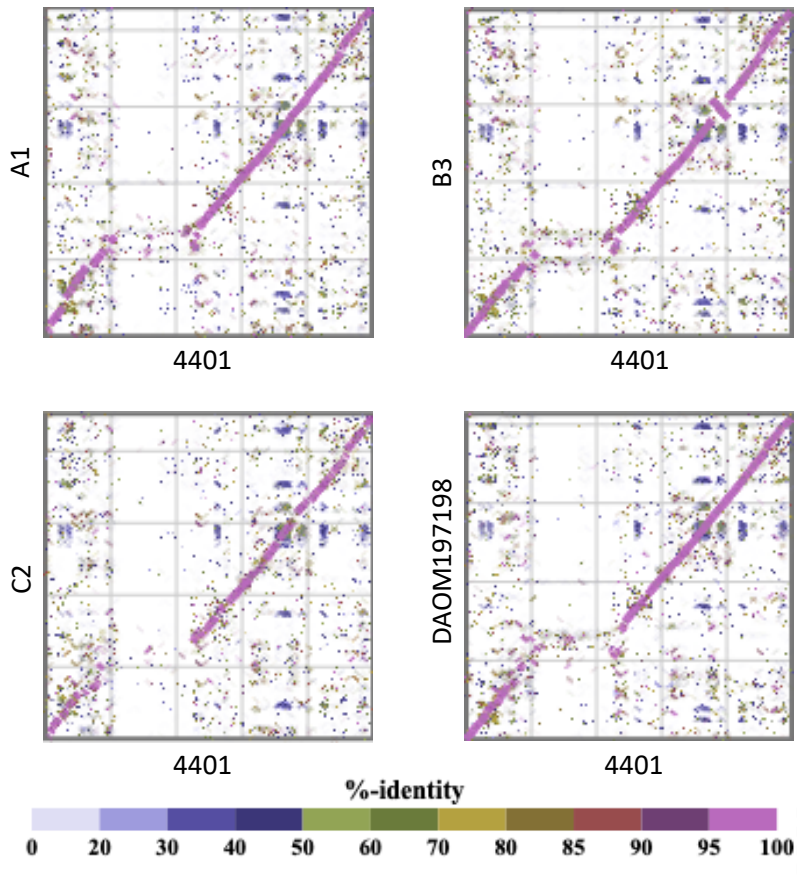


Fig. 2 Comparison of chromosome 8 between different strains. Each dot represents an amino acid sequence similarity based on tBLASTx.

Table 2 The lengths of chromosome 8 of each strain.

Strain	Length (bp)
A1	5,441,258
B3	5,350,159
C2	5,721,097
DAOM-197198	5,236,836
4401	6,327,528

The genomes of *R. irregularis* strains have many repeats, and the average proportion of repeats is nearly 50% (Yildirir et al., 2022). For each chromosome of the five strains of *R. irregularis*, they contained 40% to 65% of repeats. In contrast, the 1.5 Mb GEVE region showed 18.28% of repeat content (Fig. 3a). Similarly, the GEVE region showed a smaller proportion of known transposable elements (TEs; 2.31%) identified by RepeatModeler compared with the entire genomes (average, 13.45%) (Fig. 3b).

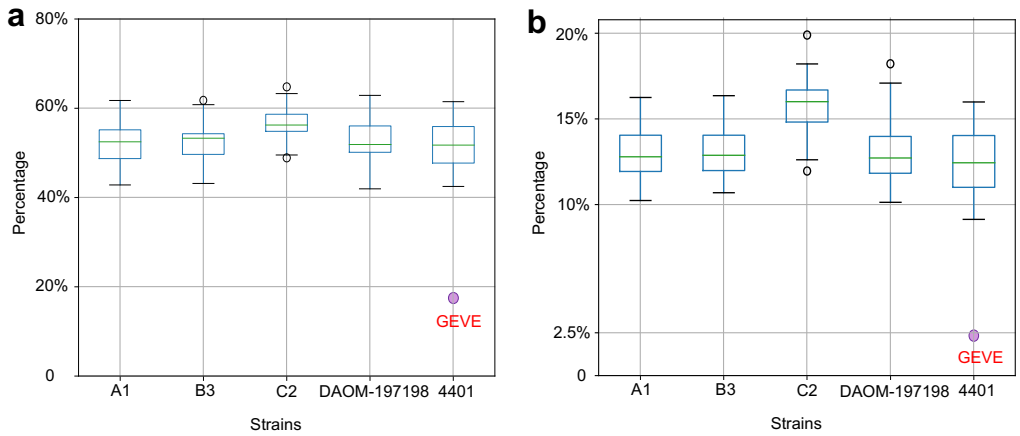


Fig. 3 Length percentages of repeat and TEs across each chromosome of *R. irregularis*. (a) Percentages of repeated sequences across different chromosomes of five strains and the 1.5 Mb GEVE region. The purple dot represents the repeat content within the GEVE region (18.28%). **(b)** Percentages of known TE sequences across different chromosomes of five strains and the 1.5 Mb GEVE region. The purple dot represents the TEs content within the GEVE region (2.31%).

The GEVE is highly homologous to *Asfarviridae*

The five *Nucleocytoviricota* marker genes in the 1.5 Mb GEVE region were found as single copies and dispersed in the region (Fig. 4a). Phylogenetic analysis using the five marker genes indicated that they are closely related to homologs from *Asfarviridae* (Supplementary Figs S1–S5). We further performed phylogenetic analysis using a concatenated sequence of the three longest and the most universal NCLDV marker genes (PolB, RNAPL, RNAPS) (Fig. 4b). The constructed tree indicated that the GEVE represents a sister group of the clade including *African swine fever virus* and *Abalone asfa-like virus* (100% ultrafast bootstrap supports).

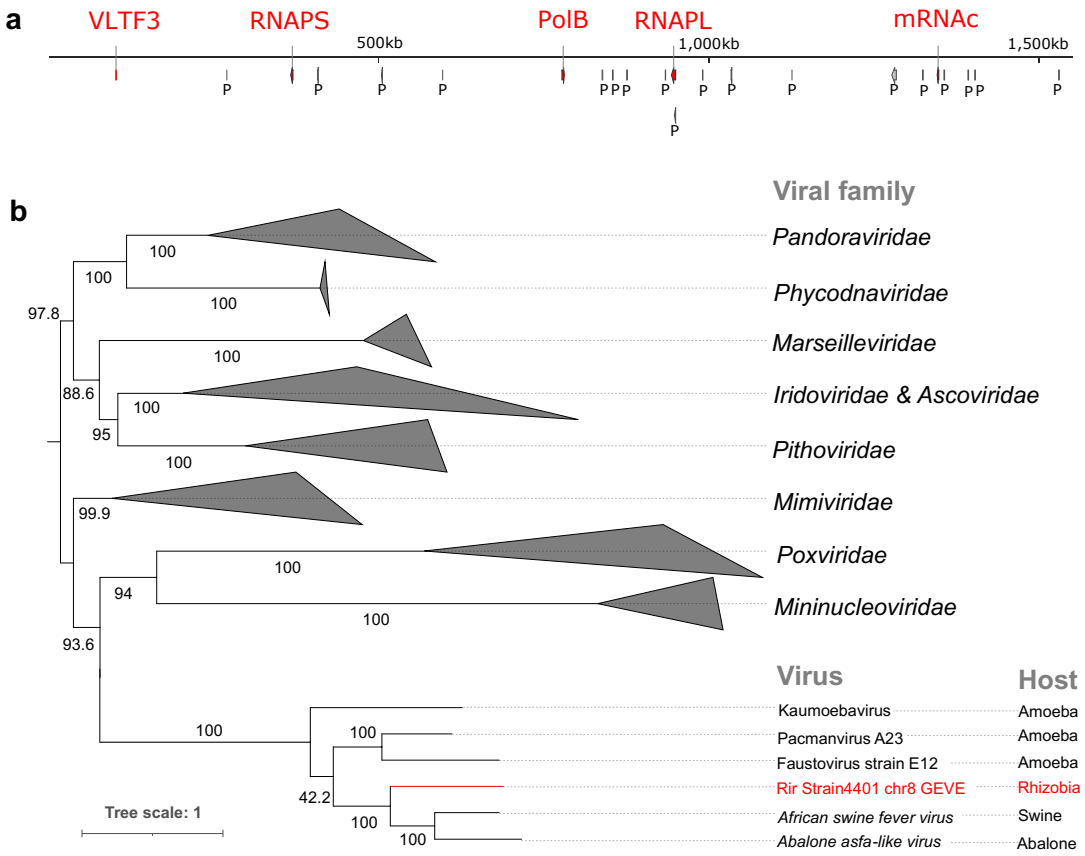


Fig. 4 Marker genes on the 1.5Mb GEVE region. (a) Distribution of *Nucleocytoviricota* marker genes and putative viral pseudogenes on this 1.5 Mb GEVE region. Marker genes were colored in red and viral pseudogene candidates were marked as "P". RNAPL gene is a marker gene and a putative pseudogene. **(b)** A concatenated maximum-likelihood (ML) phylogenetic tree using three marker genes (PolB, RNAPL, RNAPS). Rir is the abbreviation of *R. irregularis*. Ultrafast bootstrap supports are shown along the branches. The best-fit model was Q.pfam+F+I+I+R8.

Lack of MCP homologs in the fungal genomes

Among ten analyzed marker genes, the 1.5 Mb GEVE region did not encode genes for major capsid protein (MCP), A32-ATPase (A32), D5 primase/helicase (D5), ribonucleotide reductase (RNR), or superfamily II helicase (SFII). None of these five genes were detected in the genomes of the five strains. We found homologs of *Asfarviridae* minor capsid protein in this GEVE region (from 171,954 bp to 176,120 bp). As for MCP, we performed tBLASTn and additional sensitive searches using HMM models of MCP sequences from multiple viral groups including typical MCPs of nucleocytovirus (see methods for detail). However, no MCP gene was detected in the genome of strain 4401 or the other four strains.

The fungal GEVE is highly related to virus-like sequences from a sea slug

Based on the BLASTp best hits against the NR database, we could assign the best hit taxonomy to 184 ORFs (E-value < 10^{-5}). Among these ORFs, 102 ORFs showed the best hit to eukaryotes, 56 to prokaryotes, and 26 to viruses (of which 21 to *Asfarviridae*). We manually checked those ORFs with the best hit of prokaryotes. There were 21 ORFs best-hit to a bacterial gene but had multiple *Asfarviridae* genes in the following hits. These prokaryotic hits were environmental sequences, so we consider these ORFs as viral ORFs (Fig.5a). Similar pattern of best hits was also found in previously reported nucleocytovirus because of the novelty of the virus (Blanc-Mathieu et al., 2021).

Notably, 23 out of 102 eukaryotic best hits came from a single genome assembly of a sea slug species, *Elysia marginata* (Fig. 5b). Although the best hits were *E. marginata*, most of the following hits for these ORFs were genes

from viruses belonging to *Asfarviridae*, suggesting the presence of viral-like sequences in the genomic data of *E. marginata*. By scanning the genomic data of *E. marginata*, we detected 13 putative viral regions and 10 *Nucleocytoviricota* marker genes (PolB, RNAPS, RNAPL, MCP, RNR, VLTF3, A32, SFII, D5, mRNAc) (Supplementary Table S1 and S4). Six regions showed GC content of around 35%, similar to *E. marginata* (36.5%). However, seven regions showed a higher GC content (56.7%). The total lengths of these seven regions reached 124 kb, where all the identified marker genes were encoded. The length of each region almost coincided with that of the contig.

Phylogenetic analysis of the marker genes confirmed that the viral sequences of *E. marginata* are related to *Asfarviridae* and the GEVE found in *R. irregularis* (Fig. 5c) The viral sequences from *E. marginata* erewere positioned between the fungal GEVE and the clade including *African swine fever virus* and *Abalone asfa-like virus*. We also found the consistency of the phylogenies among different marker genes (Supplementary Figs S1-S6). In conclusion, the genomic data from *E. marginata* contains sequences related to *Asfarviridae*.

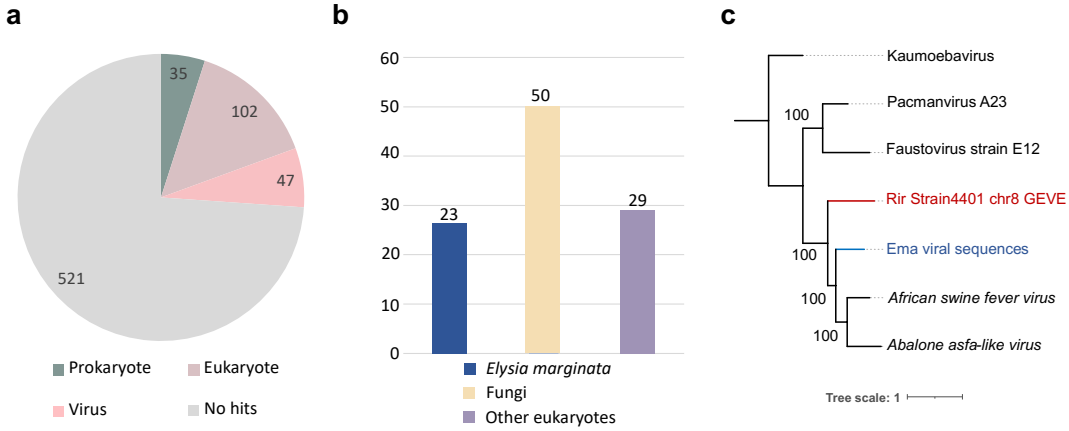


Fig. 5 Relationship between the GEVE on chromosome 8 of strain 4401 and viral sequences in *E. marginata*. (a) Taxonomic attribution of the BLASTp best hit based annotation of all 705 ORFs in the 1.5 Mb GEVE region against the edited NR database. **(b)** Details of 102 eukaryotic annotations. 23 best hits for *E. marginata* were from one assembly, GCA_019649035.1 **(c)** A ML tree of the viral sequences from the genome of *E. marginata* using three concatenated marker genes (PolB, RNAPL, RNAPS). Rir and Ema are abbreviations of *R. irregularis* and *E. marginata*, respectively. The best-fit model was LG+I+G4.

Discussion

In the fungal virosphere where most members are RNA viruses, no dsDNA virus is known (Kondo et al., 2022). In the present study, we discovered a continuous 1.5 Mb *Nucleocytoviricota* GEVE region in the arbuscular mycorrhizal fungus *R. irregularis*. Phylogenetic analysis indicated this GEVE is close to the clade of *African swine fever virus* and *Abalone asfa-like virus* (Fig. 4b, Supplementary Fig. S1-S5). Also, the GEVE region contains single copies of five *Nucleocytoviricota* marker genes (Fig. 4a) and showed distinct GC content and Hi-C profile throughout the region compared with the other parts of the chromosome (Figs. 1b and d). These results suggest that the GEVE region originated from a single insertion of an *Asfarviridae*-related virus. The compaction of the GEVE genomic region implies the possible silencing of genes in this region as known in the mammal genome (Shukron et al., 2019), which may function as a defense of fungi against the invasion of extracellular DNA. Alternatively, the Hi-C signals of this GEVE region might result from unidentified region-specific repeats as described below. The GEVE was specific to a strain among the five analyzed strains (Fig. 2, Supplementary Table S2). This GEVE region showed low repeat and TEs density compared with other regions of the fungal, implying that the region experienced a shorter time for invasions of TEs than the other chromosomal regions. (Fig. 3). Overall, this viral integration is probably a recent event, suggesting there are dsDNA viruses that are still actively infecting the fungal species.

The genome size of currently isolated *Asfarviridae* members is from 155 kb to 466 kb (Matsuyama et al., 2020; Reteno et al., 2015), presenting a significant discrepancy compared to the 1.5 Mb GEVE region identified in fungal chromosomes. The integration of fungal sequences was suggested as one of the factors for the expansion of the GEVE region. However, despite the identified 27.39% of coding sequences and 18.28% of repeats in this GEVE region, a substantial portion of the genomic regions remains unidentified. One speculative explanation for this observation is the presence of potential pseudogenized genes that have undergone loss of coding function. These pseudogenes may have triggered the expansion of region-specific repeats that evade detection by RepeatModeler. Further investigation is needed to validate this hypothesis.

Although such a long and contiguous GEVE region was found only in chromosome 8 of strain 4401, PolB genes possibly originating from *Asfarviridae* were found on multiple chromosomes of all strains (Supplementary Table S1). These PolBs do not have fixed positions or copy numbers across the genomes of *R. irregularis* (Supplementary Table S3). This suggests that infection of *R. irregularis* by *Asfarviridae*-like viruses may be a widespread and on-going event.

Members of *Asfarviridae* infect a variety of eukaryotes, such as swine, abalone, and amoebae, although only five members have been isolated and completely sequenced (Karki et al., 2021). In addition to them, previous studies suggested that marine dinoflagellate (Ogata et al., 2009) and oomycetes (Hannat et al., 2021) as potential hosts of *Asfarviridae*. In the present study, we identified *Asfarviridae*-like genome sequences within the genomic data of a sea slug species, *E. marginata*, as a close relative of the fungal GEVE (Fig. 4c). The viral sequences of *E. marginata* encode all of the core genes (Supplementary Table S1), suggesting that the virus of these viral sequences can form virion particles and be infectious. Unlike GEVEs in *R. irregularis*, the seven viral regions with high GC contents covered almost the entire length of their contigs, suggesting that these viral regions are not insertions in the sea slug genome but derived from viral genomes concomitantly sequenced with the *E. marginata* genome (Supplementary Table S2). Taken together, members of *Asfarviridae* are likely more widespread and diverse than previously understood.

MCP is one of the major components of virions (Krupovic et al., 2022), but no MCP gene was detected in the genomes of *R. irregularis* (Supplementary Table S1). However, as we discovered *Asfarviridae* minor capsid protein homologs in the 1.5 Mb GEVE region, this virus may have a capsid structure. There are two possible explanations for the lack of MCP genes. First, gene deletion or genome rearrangement might have occurred after the virus genome integration (Moniruzzaman et al., 2020) and caused gene loss for MCP from the GEVE region. However, this possibility may be low given that the GEVE appears nearly complete (by conserving many single-copy marker genes) and that no MCP homologs were detected in any of the genomes of the analyzed five strains. Second, there is a possibility that the virus giving rise to the GEVE encodes an MCP whose sequence is distinct from the known MCPs. The MCPs of nucleocytoviruses are highly divergent, and some nucleocytoviruses (such as pandoraviruses) utilize different types of major virion proteins (Krupovic et al., 2020). Previously unknown types of MCPs may be undetected as all of our detection methods are based on known references.

Arbuscular mycorrhizal fungi are obligate plant mutualistic organisms, which provide significant benefits to plants, such as improvement of nutrition and enhanced resistance to disease (Gosling et al., 2006). As a model organism of arbuscular mycorrhizal fungi, *R. irregularis* forms a robust tripartite association with its endobacteria and plants in its life cycle, and horizontal gene transfers among these organisms contribute to their evolution and symbiotic adaptation (Li et al., 2018). For example, foreign genes participate in the life cycle of *R. irregularis* including gene expression, mitosis, and signal transduction. This study found that dsDNA viruses could also be an important contributor to horizontally transferred genes in *R. irregularis*. Future functional analysis of these viral-originating genes in *R. irregularis* may provide novel insight into the ecology and evolution of these beneficial microorganisms.

Methods

Detection of viral regions

To identify virus-like regions in eukaryotic genomes, we used ViralRecall v2.1 (Aylward and Moniruzzaman, 2021) to search genomic data with a window size and viral score set at 150 ORFs and at 0, respectively. ViralRecall is a tool designed for identifying virus-like regions, especially this tool uses *Nucleocytoviricota* orthologous groups and the Pfam database to detect signals of this viral phylum. This tool will evaluate viral and cellular scores for each ORF in the genome and calculate how likely the regions are virus-like. Chromosome-level genomic data for five strains of *R. irregularis* were retrieved from GenBank, and genomic data for *E. marginata* was retrieved from GCA_019649035.1 (Maeda et al., 2021). GC content was calculated by in-house Python scripts and the window size was set at 50,000 base pairs to minimize the effect of repeats. Insertion of the viral region into the fungal chromosome was validated as follows. First, long reads of strain 4401 (accession no. SRR15461860) were mapped onto the whole genome by Minimap2 v2.24 (Li, 2018). We applied the "view" function of Samtools v1.16.1 (Li et al., 2009) to select reads mapped regions connecting the virus and host region. For visualization, the "bamtobed" function of Bedtools v2.29.2(Quinlan and Hall, 2010) was used. Comparisons of chromosome 8 at the amino-acid sequence level across different strains were performed using DiGAlign v1.3 (http://www.genome.jp/digalign/).

Generation of Hi-C contact map and detection of repeats

To determine the chromosomal structure of the chromosome on which the 1.5 Mb GEVE region is located, we transform raw Hi-C sequencing data of strain 4401 from NCBI Sequence Read Archive (accession no. SRR15461854) into normalized contact maps using Juicer v2.0 (Durand et al., 2016) and visualized the results by Juicebox (http://aidenlab.org/juicebox/). Repeated elements in the genome were identified and masked by using RepeatModeler v2.0.2 (Flynn et al., 2020) and RepeatMasker v4.1.2 (http://www.repeatmasker.org). RepeatModeler could identify repeats and annotate TEs including retroelements and DNA transposons with distinct discovery algorithms. We also performed the "-LTRStruct" parameter while running RepeatModeler to discover long terminal repeat retroelements.

Phylogenetic analyses

Nucleocytoviricota marker genes were predicted by using the built-in HMMER of ViralRecall. To verify whether these genes are indeed viral homologs, we collected reference sequences of these marker genes from a previous study (Kazlauskas et al., 2020) and GenBank, then built phylogenetic trees as follows. Multiple sequence alignments were performed by Clustal-Omega v1.2.4 (Sievers and Higgins, 2018) and trimmed by using trimAl v1.4.1 (Capella-Gutiérrez et al., 2009) with the parameter "-gt 0.1". Maximum likelihood trees were generated by using IQ-TREE v2.2.0 (Minh et al., 2020) with 1,000 ultrafast bootstraps (Hoang et al., 2018). The best-fit model was selected using ModelFinder (Kalyaanamoorthy et al., 2017).

MCP genes were further searched in the genome of strain 4401 with hmm files constructed from the MCP sequences of phages, *Nucleocytoviricota*, *Mirusviricota*, and *Herpesvirales* by using HMMER v3.3.2 (Eddy, 2011) ($e < 0.05$). We also searched the major virion capsid 1 and 2 of pandoraviruses using the same method. Reference sequences of phages, *Herpesvirales*, pandoraviruses, and other members of *Nucleocytoviricota* are manually collected from GenBank and NCBI protein database. MCP sequences of *Mirusviricota* were obtained from a previous study (Gaïa et al., 2022). Considering the possibility of missing results if the sequence in fungal genome is not identified as an ORF, we performed tBLASTn using NCBI BLAST (Johnson et al., 2008) to search for those MCP sequences against all nucleotide sequences of *R. irregularis* (NCBI: txid588596) in the nr/nt database (e <

0.05).

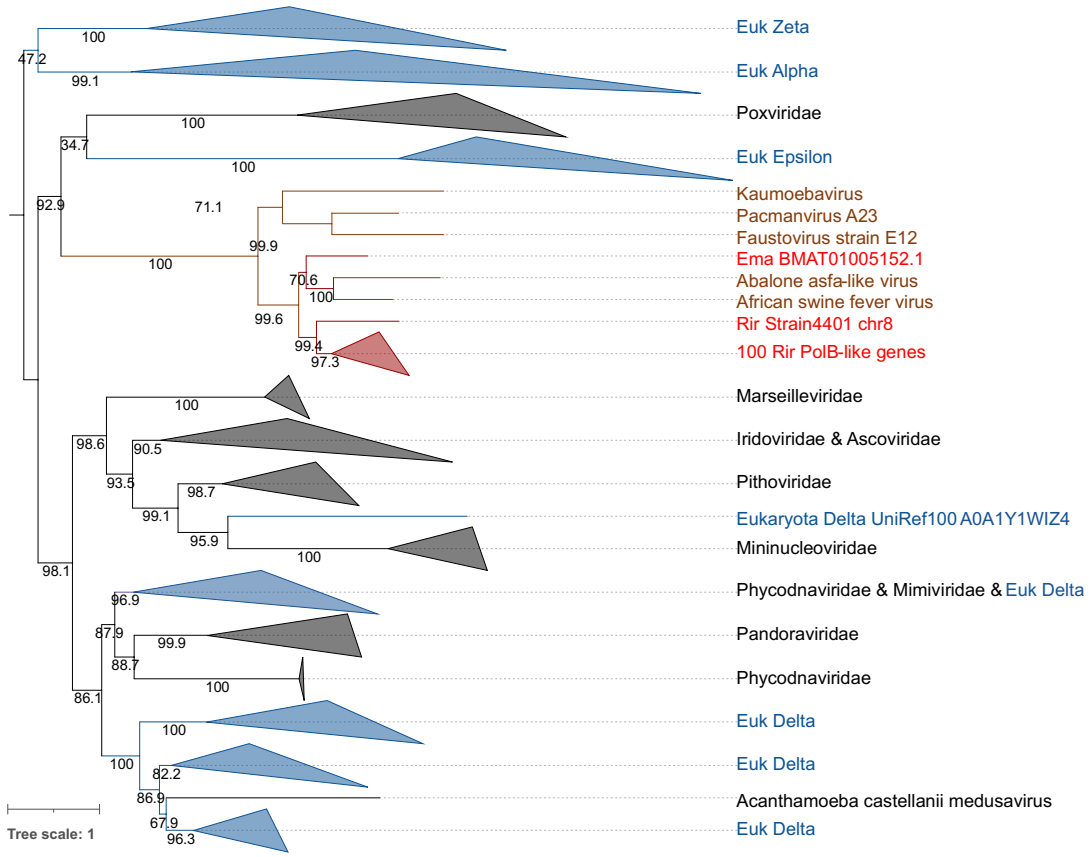
Annotation of viral regions

ORFs in the GEVE were predicted by ViralRecall built-in Prodigal v2.6.3 (Hyatt et al., 2010). As a tool designed for the identification of prokaryotic genes, it efficiently predicted viral genes due to the similarity of genomic architecture with prokaryotes, but some eukaryotic genes might be performed poorly (Moniruzzaman et al., 2020). ORFs were annotated by using BLASTp in BLAST+ v2.13.0 (Camacho et al., 2009) ($e < 10^{-5}$). Because previous studies may have annotated viral insertions in fungi as fungal genes, we used the NR database from which the sequences from a fungal class *Glomeromycetes* (NCBI: txid214506, to which *R. irregularis* belongs) were removed. The best hit of each ORF was used to determine the related species of the ORF, and we classified them into three parts, eukaryote, prokaryote, and virus. Functional annotation was retrieved through eggNOG-mapper v2.1.9 (Cantalapiedra et al., 2021). Pseudogene candidates were predicted by Prokka v1.14.6 (Seemann, 2014) and Pesudofinder v1.1.0 (Syberg-Olsen et al., 2022) using the Viral-Host database 11/2022 (Mihara et al., 2016) as the reference database.

Acknowledgments

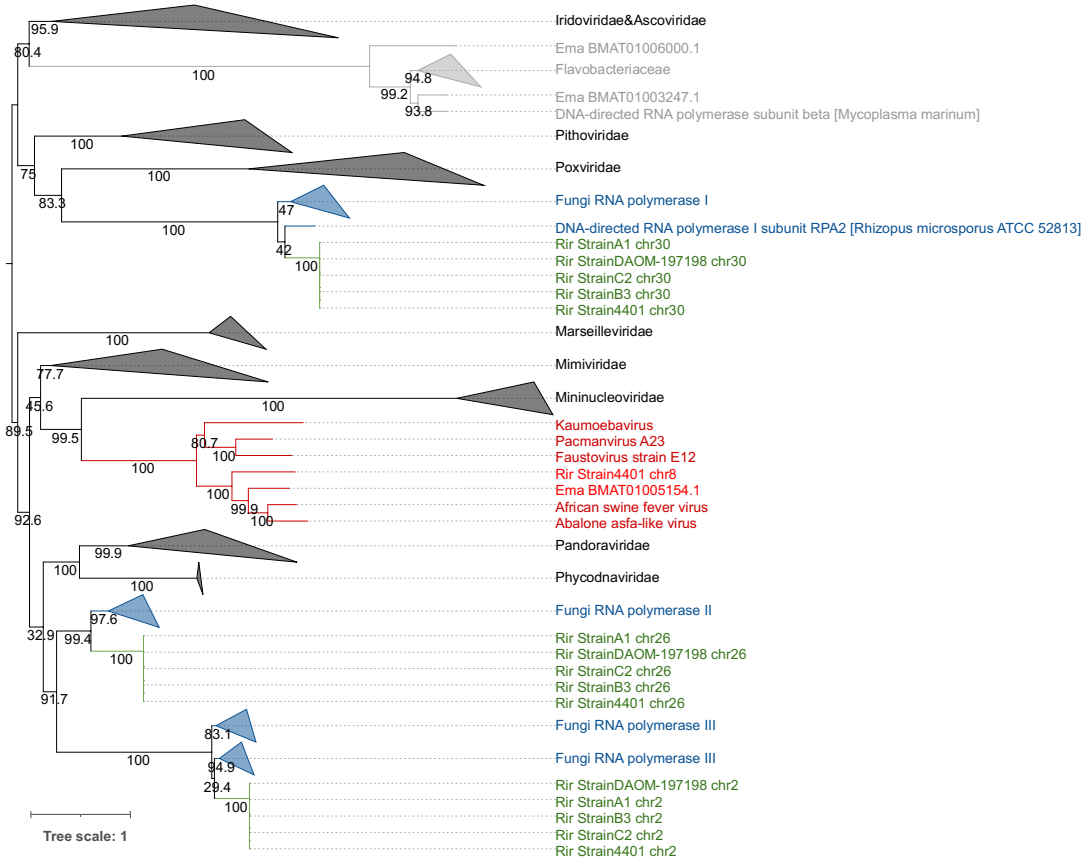
We thank Lingjie Meng, Ruixuan Zhang, and Junyi Wu for the helpful discussion on the present study. This work was supported by JSPS/KAKENHI (No. 22H00384 to H.O.), Computational time was provided by the Supercomputer System, Institute for Chemical Research, Kyoto University.

Supplementary figures

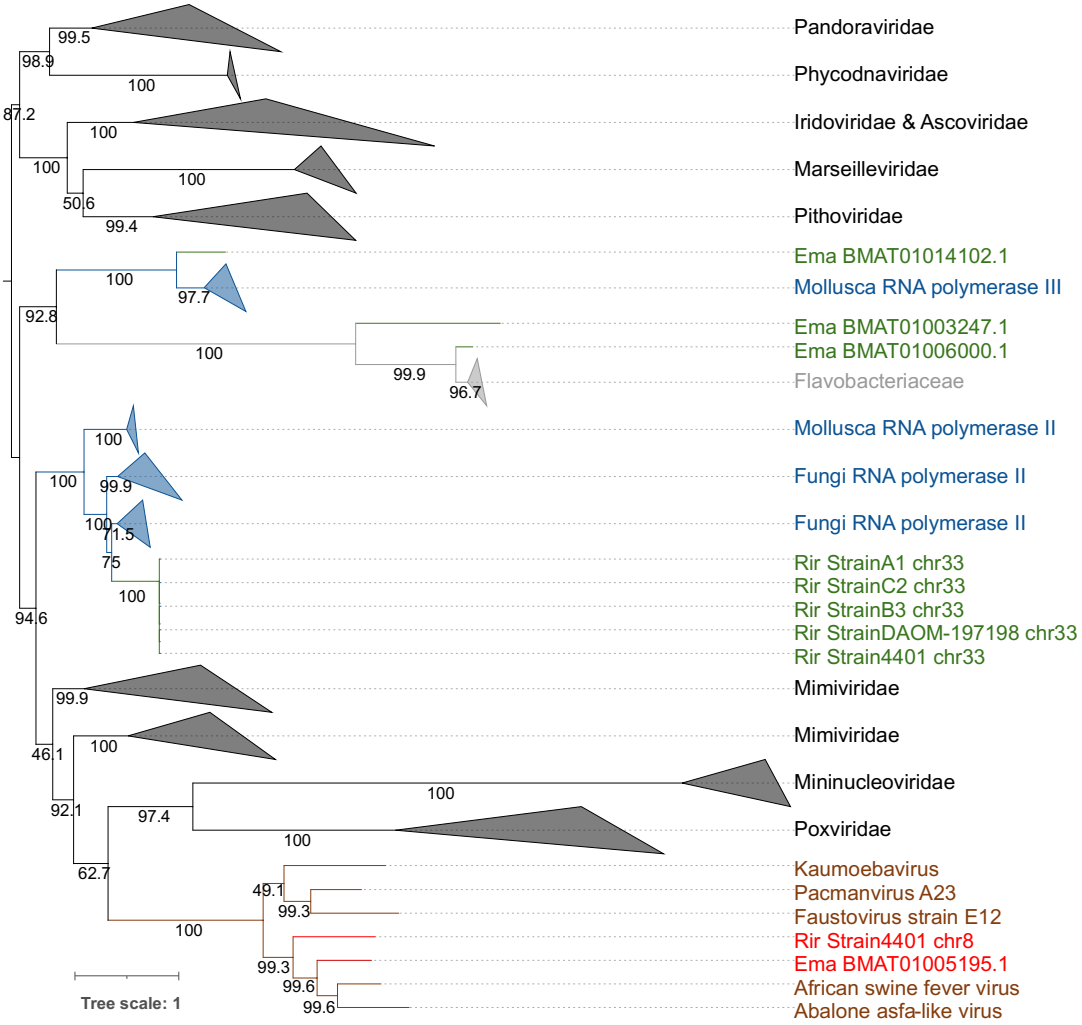


Supplementary Fig S1: Phylogenetic tree of 101 PolB candidates predicted by ViralRecall. Clades containing eukaryotic sequences are colored blue. *Asfarviridae* sequences are colored brown. Viral sequences from *R. irregularis* and *E. marginata* are shown as red labels. Ultrafast

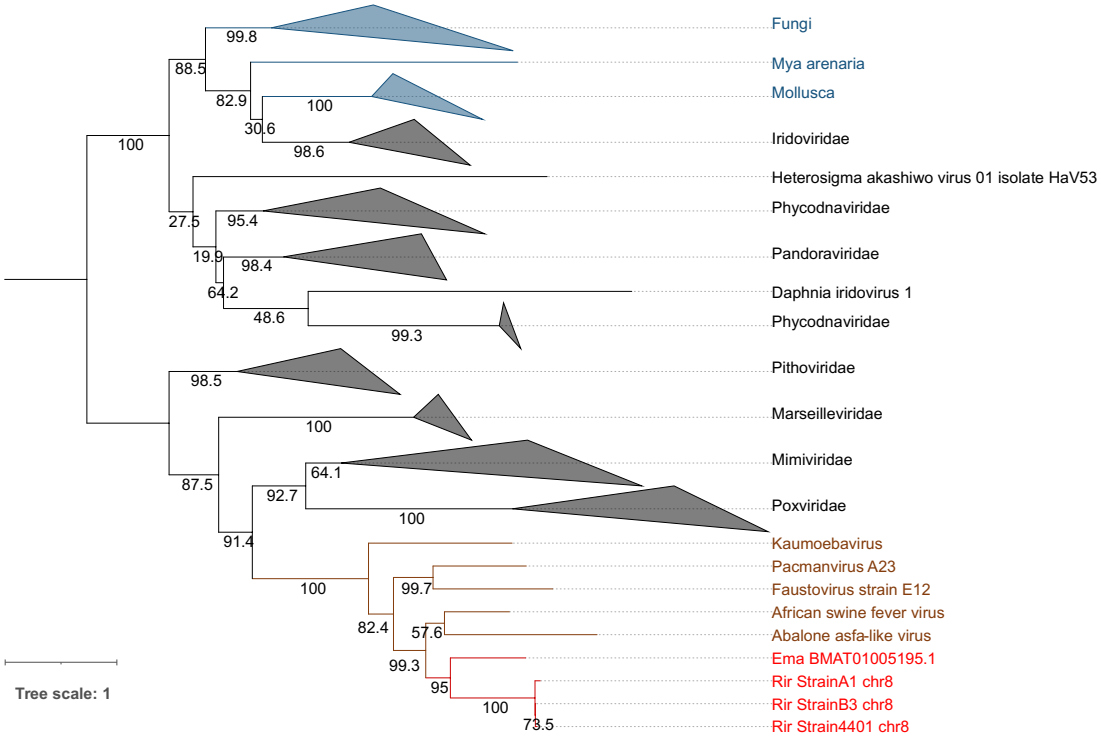
bootstrap supports are shown along the branches. The best-fit model was Q.pfam+F+R10.



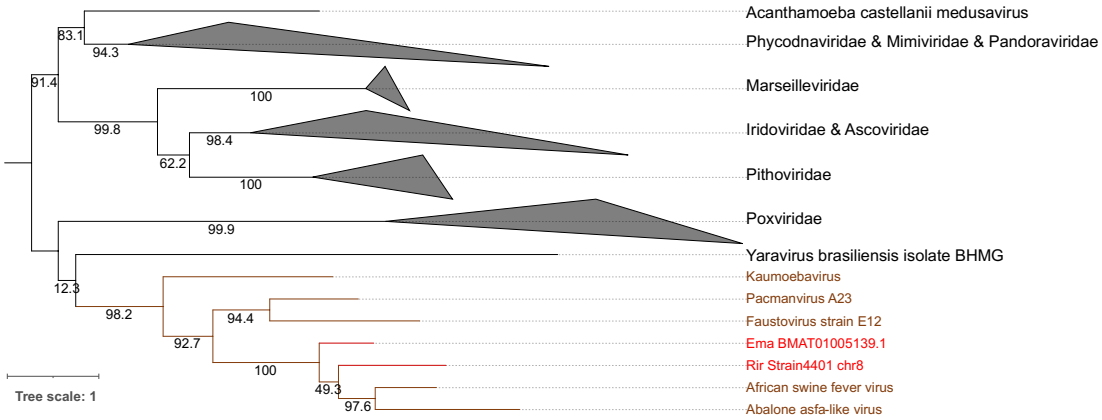
Supplementary Fig S2: Phylogenetic tree of RNAPL candidates predicted by ViralRecall. Coloring is the same as Fig. S1, except that eukaryotic and bacterial sequences from *R. irregularis* and *E. marginata* are colored in green and bacterial sequences are in light grey. Ultrafast bootstrap supports are shown along the branches. The best-fit model was LG+F+I+I+R8.



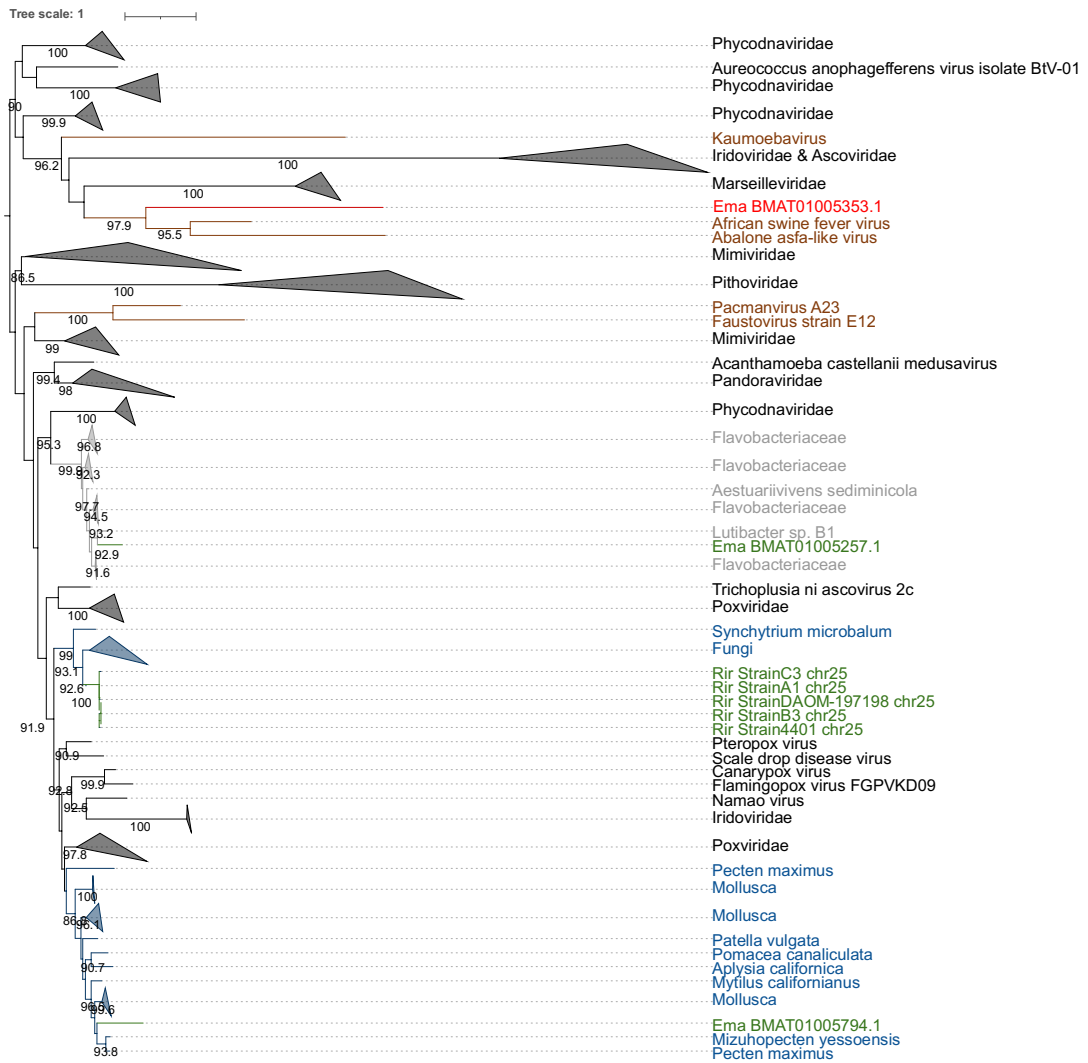
Supplementary Fig S3: Phylogenetic tree of RNAPS candidates predicted by ViralRecall. The coloring is the same as in Fig. S2. Ultrafast bootstrap supports are shown along the branches. The best-fit model was LG+F+R10.



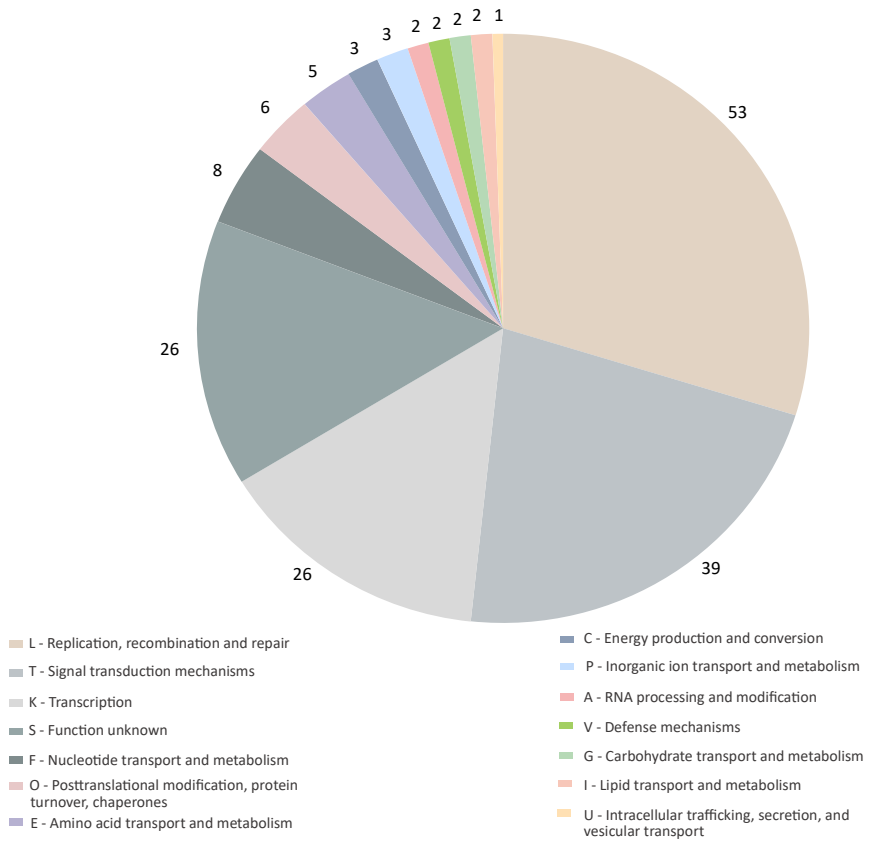
Supplementary Fig S4: Phylogenetic tree of mRNAc candidates predicted by ViralRecall. The coloring is the same as in Fig. S2. Ultrafast bootstrap supports are shown along the branches. The best-fit model was Q.pfam+F+I+I+R6.



Supplementary Fig S5: Phylogenetic tree of VLTF3 candidates predicted by ViralRecall. The coloring is the same as in Fig. S1. Ultrafast bootstrap supports are shown along the branches. The best-fit model was Q.insect+F+R6.



Supplementary Fig S6: Phylogenetic tree of RNR candidates predicted by ViralRecall. The coloring is the same as in Fig. S2. Ultrafast bootstrap supports are shown along the branches. The best-fit model was Q.yeast+I+I+R7.



Supplementary Fig S7: Functional annotation of the GEVE. The number of ORFs with the COG functional categories was shown outside each sector.

Supplementary tables

Supplementary Table S1: Distribution of ten *Nucleocytoviricota* marker genes predicted by ViralRecall. All the above marker gene candidates have scores of HMM above the threshold set by ViralRecall (RNAPL, RNAPS, PolB: 200; SFII: 100, Others: 80). While the MCP gene was also not detected in some viruses, all five *Asfarviridae* viruses are known to have this gene. These false negatives are probably due to the limitation of built-in Prodigal and HMM algorithms.

Supplementary Table S2: Viral regions predicted by ViralRecall in the five strains of *R. irregularis*.

Supplementary Table S3: Distribution of *Nucleocytoviricota* marker genes in the five strains of *R. irregularis*.

Supplementary Table S4: Predicted viral regions and marker genes in the genome of *E. marginata*.

References

Aylward, F.O., Moniruzzaman, M., 2021. ViralRecall—A Flexible Command-Line Tool for the Detection of Giant Virus Signatures in 'Omic Data. Viruses 13, 150. https://doi.org/10.3390/v13020150

Aylward, F.O., Moniruzzaman, M., Ha, A.D., Koonin, E.V., 2021. A phylogenomic framework for charting the diversity and evolution of giant viruses. PLOS Biology 19, e3001430. https://doi.org/10.1371/journal.pbio.3001430

Bellas, C., Hackl, T., Plakolb, M.-S., Koslová, A., Fischer, M.G., Sommaruga, R., 2023. Large-scale invasion of unicellular eukaryotic genomes by integrating DNA viruses. Proceedings of the National Academy of Sciences 120, e2300465120. https://doi.org/10.1073/pnas.2300465120

Bhattacharjee, A.S., Schulz, F., Woyke, T., Orcutt, B.N., Martínez Martínez, J., 2023. Genomics discovery of giant fungal viruses from subsurface oceanic crustal fluids. ISME Communications 3, 1–9.

https://doi.org/10.1038/s43705-022-00210-8

Blanc-Mathieu, R., Dahle, H., Hofgaard, A., Brandt, D., Ban, H., Kalinowski, J., Ogata, H., Sandaa, R.-A., 2021. A Persistent Giant Algal Virus, with a Unique Morphology, Encodes an Unprecedented Number of Genes Involved in Energy Metabolism. Journal of Virology 95, e02446-20. https://doi.org/10.1128/JVI.02446-20

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421. https://doi.org/10.1186/1471-2105-10-421

Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., Huerta-Cepas, J., 2021. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Molecular Biology and Evolution 38, 5825–5829. https://doi.org/10.1093/molbev/msab293

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Cheng, S., Wong, G.K.-S., Melkonian, M., 2021. Giant DNA viruses make big strides in eukaryote evolution. Cell Host & Microbe 29, 152–154. https://doi.org/10.1016/j.chom.2021.01.008

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., Aiden, E.L., 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Systems 3, 95–98. https://doi.org/10.1016/j.cels.2016.07.002

Eddy, S.R., 2011. Accelerated Profile HMM Searches. PLOS Computational Biology 7, e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Feschotte, C., Gilbert, C., 2012. Endogenous viruses: insights into viral evolution and impact on host biology. Nature Reviews Genetics 13, 283–296. https://doi.org/10.1038/nrg3199

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., Smit, A.F., 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences 117, 9451–9457. https://doi.org/10.1073/pnas.1921046117

Gaïa, M., Meng, L., Pelletier, E., Forterre, P., Vanni, C., Fernandez-Guerra, A., Jaillon, O., Wincker, P., Ogata, H., Krupovic, M., Delmont, T.O., 2022. Plankton-infecting relatives of herpesviruses clarify the evolutionary trajectory of giant viruses. BioRxiv, https://doi.org/10.1101/2021.12.27.474232

Gong, Z., Zhang, Y., Han, G.-Z., 2020. Molecular fossils reveal ancient associations of dsDNA viruses with several phyla of fungi. Virus Evolution 6, veaa008. https://doi.org/10.1093/ve/veaa008

Gosling, P., Hodge, A., Goodlass, G., Bending, G.D., 2006. Arbuscular mycorrhizal fungi and organic farming. Agriculture, Ecosystems & Environment 113, 17–35. https://doi.org/10.1016/j.agee.2005.09.009

Hannat, S., Pontarotti, P., Colson, P., Kuhn, M.-L., Galiana, E., La Scola, B., Aherfi, S., Panabières, F., 2021. Diverse Trajectories Drive the Expression of a Giant Virus in the Oomycete Plant Pathogen *Phytophthora parasitica*. Frontiers in Microbiology 12, 662762. https://doi.org/10.3389/fmicb.2021.662762

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., Vinh, L.S., 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution 35, 518–522. https://doi.org/10.1093/molbev/msx281

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119. https://doi.org/10.1186/1471-2105-11-119

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., Madden, T.L., 2008. NCBI BLAST: a better web interface. Nucleic Acids Research 36, W5–W9. https://doi.org/10.1093/nar/gkn201

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., Jermiin, L.S., 2017. ModelFinder: fast model

selection for accurate phylogenetic estimates. Nature Methods 14, 587–589. https://doi.org/10.1038/nmeth.4285

Karki, S., Moniruzzaman, M., Aylward, F.O., 2021. Comparative Genomics and Environmental Distribution of Large dsDNA Viruses in the Family *Asfarviridae*. Frontiers in Microbiology 12, 657471. https://doi.org/10.3389/fmicb.2021.657471

Kazlauskas, D., Krupovic, M., Guglielmini, J., Forterre, P., Venclovas, Č., 2020. Diversity and evolution of B-family DNA polymerases. Nucleic Acids Research 48, 10142–10156. https://doi.org/10.1093/nar/gkaa760

Kondo, H., Botella, L., Suzuki, N., 2022. Mycovirus Diversity and Evolution Revealed/Inferred from Recent Studies. Annual Review of Phytopathology 60, 307–336. https://doi.org/10.1146/annurev-phyto-021621-122122

Krupovic, M., Makarova, K.S., Koonin, E.V., 2022. Cellular homologs of the double jelly-roll major capsid proteins clarify the origins of an ancient virus kingdom. Proceedings of the National Academy of Sciences 119, e2120620119. https://doi.org/10.1073/pnas.2120620119

Krupovic, M., Yutin, N., Koonin, E., 2020. Evolution of a major virion protein of the giant pandoraviruses from an inactivated bacterial glycoside hydrolase. Virus Evolution 6, veaa059. https://doi.org/10.1093/ve/veaa059

Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, M., Zhao, J., Tang, N., Sun, H., Huang, J., 2018. Horizontal Gene Transfer From Bacteria and Plants to the Arbuscular Mycorrhizal Fungus *Rhizophagus irregularis*. Frontiers in Plant Science 9, 701. https://doi.org/10.3389/fpls.2018.00701

Li, P., Wang, S., Zhang, L., Qiu, D., Zhou, X., Guo, L., 2020. A tripartite ssDNA mycovirus from a plant pathogenic fungus is infectious as cloned DNA and purified virions. Science Advances 6, eaay9634. https://doi.org/10.1126/sciadv.aay9634

Maeda, T., Takahashi, S., Yoshida, T., Shimamura, S., Takaki, Y., Nagai, Y., Toyoda, A., Suzuki, Y., Arimoto, A., Ishii, H., Satoh, N., Nishiyama, T., Hasebe, M., Maruyama, T., Minagawa, J., Obokata, J., Shigenobu, S., 2021. Chloroplast acquisition without the gene transfer in kleptoplastic sea slugs, *Plakobranchus ocellatus*. Elife 10, e60176. https://doi.org/10.7554/eLife.60176

Matsuyama, T., Takano, T., Nishiki, I., Fujiwara, A., Kiryu, I., Inada, M., Sakai, T., Terashima, S., Matsuura, Y., Isowa, K., Nakayasu, C., 2020. A novel Asfarvirus-like virus identified as a potential cause of mass mortality of abalone. Scientific Reports 10, 4620. https://doi.org/10.1038/s41598-020-61492-3

Maumus, F., Blanc, G., 2016. Study of Gene Trafficking between Acanthamoeba and Giant Viruses Suggests an Undiscovered Family of Amoeba-Infecting Viruses. Genome Biology and Evolution 8, 3351–3363. https://doi.org/10.1093/gbe/evw260

Maumus, F., Epert, A., Nogué, F., Blanc, G., 2014. Plant genomes enclose footprints of past infections by giant virus relatives. Nature Communications 5, 4268. https://doi.org/10.1038/ncomms5268

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S., Ogata, H., 2016. Linking Virus Genomes with Host Taxonomy. Viruses 8, 66. https://doi.org/10.3390/v8030066

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution 37, 1530–1534. https://doi.org/10.1093/molbev/msaa015

Moniruzzaman, M., Erazo-Garcia, M.P., Aylward, F.O., 2022. Endogenous giant viruses contribute to intraspecies genomic variability in the model green alga *Chlamydomonas reinhardtii*. Virus Evolution 8, veac102.

https://doi.org/10.1093/ve/veac102

Moniruzzaman, M., Weinheimer, A.R., Martinez-Gutierrez, C.A., Aylward, F.O., 2020. Widespread endogenization of giant viruses shapes genomes of green algae. Nature 588, 141–145. https://doi.org/10.1038/s41586-020-2924-2

Ogata, H., Toyoda, K., Tomaru, Y., Nakayama, N., Shirai, Y., Claverie, J.-M., Nagasaki, K., 2009. Remarkable sequence similarity between the dinoflagellate-infecting marine girus and the terrestrial pathogen African swine fever virus. Virology Journal 6, 178. https://doi.org/10.1186/1743-422X-6-178

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033

Reteno, D.G., Benamar, S., Khalil, J.B., Andreani, J., Armstrong, N., Klose, T., Rossmann, M., Colson, P., Raoult, D., La Scola, B., 2015. Faustovirus, an asfarvirus-related new lineage of giant viruses infecting amoebae. Journal of Virology 89, 6585–6594. https://doi.org/10.1128/JVI.00115-15

Schulz, F., Abergel, C., Woyke, T., 2022. Giant virus biology and diversity in the era of genome-resolved metagenomics. Nature Reviews Microbiology 20, 721-736. https://doi.org/10.1038/s41579-022-00754-5

Shukron, O., Piras, V., Noordermeer, D., Holcman, D., 2019. Statistics of chromatin organization during cell differentiation revealed by heterogeneous cross-linked polymers. Nature Communications 10, 2626. https://doi.org/10.1038/s41467-019-10402-x

Sievers, F., Higgins, D.G., 2018. Clustal Omega for making accurate alignments of many protein sequences. Protein Science 27, 135–145. https://doi.org/10.1002/pro.3290

Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Syberg-Olsen, M.J., Garber, A.I., Keeling, P.J., McCutcheon, J.P., Husnik, F., 2022. Pseudofinder: Detection of Pseudogenes in Prokaryotic Genomes. Molecular Biology and Evolution 39, msac153. https://doi.org/10.1093/molbev/msac153

Yildirir, G., Sperschneider, J., Malar C, M., Chen, E.C.H., Iwasaki, W., Cornell, C., Corradi, N., 2022. Long reads and Hi-C sequencing illuminate the two-compartment genome of the model arbuscular mycorrhizal symbiont *Rhizophagus irregularis*. New Phytologist 233, 1097–1107. https://doi.org/10.1111/nph.17842

Yoshikawa, G., Blanc-Mathieu, R., Song, C., Kayama, Y., Mochizuki, T., Murata, K., Ogata, H., Takemura, M., 2019. Medusavirus, a Novel Large DNA Virus Discovered from Hot Spring Water. Journal of Virology 93, e02130-18. https://doi.org/10.1128/JVI.02130-18

Yu, X., Li, B., Fu, Y., Jiang, D., Ghabrial, S.A., Li, G., Peng, Y., Xie, J., Cheng, J., Huang, J., Yi, X., 2010. A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. Proceedings of the National Academy of Sciences 107, 8387–8392. https://doi.org/10.1073/pnas.0913535107