Eukaryotic genomic data uncover an extensive host range of mirusviruses

Hongda Zhao, Lingjie Meng, Hiroyuki Hikida, Hiroyuki Ogata*

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

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

Highlights

- Mirusvirus signals detected in genomic data from eight eukaryotic supergroups.
- Habits of putative mirusvirus hosts not limited to marine environments.
- Major capsid sequences from these assemblies show new mirusviral lineages.
- Three circular mirusvirus genomes were identified.

Summary

A recent metagenomic study has revealed a novel group of viruses designated mirusviruses, which are proposed to form an evolutionary link between two double-stranded DNA virus realms, *Varidnaviria* and *Duplodnaviria*. Metagenomic data suggest that these viruses infect marine microeukaryotes, but their host range remains largely unknown. In this study, we investigated the presence of mirusvirus marker genes in 1,901 eukaryotic assemblies, mainly derived from unicellular eukaryotes, to identify potential hosts of mirusviruses. Mirusvirus marker sequences were identified in 1,348 assemblies spanning 284 eukaryotic genera across eight supergroups. The habitats of the putative mirusvirus hosts included not only marine but also other diverse environments. Among the major capsid protein (MCP) sequence signals in the genome assemblies, we identified 85 sequences that showed high sequence and structural similarities to reference mirusvirus MCPs. A phylogenetic analysis of these sequences revealed their distant evolutionary relationships with the seven previously reported mirusvirus clades. Most of the scaffolds with these MCP sequences encoded multiple mirusvirus homologs, underscoring the impact of mirusviral infection on the evolution of the host genome. We also identified three circular mirusviral genomes within the genomic data of *Schizochytrium* sp. and *Ostreobium quekettii*. Overall, mirusviruses probably infect a wide spectrum of eukaryotes and are more diverse than previously reported.

Keywords: virus–host relationship, mirusvirus, major capsid protein, protist, endogenous virus

Introduction

Viruses pervade diverse environments across the Earth and play crucial ecological and evolutionary roles.[1–3] A recent metagenomic analysis identified a previously unrecognized but diverse group of double-stranded (ds) DNA viruses, designated mirusviruses, that are abundant, widespread, and active in the global marine ecosystem.[4] Based on their virion morphogenesis genes (e.g., HK97-fold

MCP), mirusviruses are proposed to form the phylum '*Mirusviricota*' in the realm *Duplodnaviria*, which includes herpesviruses and caudoviruses. However, a large number of genes encoded by mirusviruses, including informational genes, are more closely related to homologs in nucleocytoviruses. Nucleocytoviruses belong to another dsDNA viral realm, *Varidnaviria*, and are known to play important roles in marine ecosystems. The genomic similarities between mirusviruses and nucleocytoviruses suggest an evolutionary interplay between the two distinct groups of viruses and their similar habitats.

Viral groups have different host ranges. Nucleocytoviruses infect diverse eukaryotes, including protists, green algae, and animals.[5] Within *Duplodnavira*, caudoviruses infect a broad spectrum of prokaryotes,[6,7] whereas herpesviruses have a relatively narrow host range, limited to animals.[8,9] Until the discovery of mirusviruses, no duplodnaviruses were known to infect early-branching eukaryotes, such as protists. Mirusviruses were the first group of duplodnaviruses suggested to infect unicellular eukaryotes. This suggestion was based on the horizontal gene transfer of heliorhodopsin genes between mirusviruses and green algae, and the fact that mirusvirus signals were detected in the metagenomes and metatranscriptomes derived from size fractions corresponding to unicellular planktonic eukaryotes.[4] Therefore, the discovery of mirusviruses was considered to fill the host gap in the duplodnaviruses (between prokaryotes and animals). However, the evidence of mirusvirus hosts was limited and the taxonomic breadth of their hosts remained poorly investigated.

Two recent studies investigating the viral signals in eukaryotic genomes revealed some potential hosts of mirusviruses. Specifically, endogenized mirusviral genomes have been detected in the thraustochytrids species *Aurantiochytrium limacinum* and *Hondaea fermentalgiania*,[10] and the green algal species *Cymbomonas tetramitiformis*.[11] Notably, an additional circular mirusvirus-like genome (probably in the form of an episome) was identified in the *A. limacinum* genome assembly, suggesting an episomal form of mirusvirus in the host cells.[10] In general, the viral sequences within eukaryotic assemblies can be categorized into three types: transferred genes, integrated viral genome fragments, and free viral genomes.[12–14] The previously reported endogenized mirusviral sequences and the circular genome correspond to the latter two types, respectively. These three types of evidence strongly suggest that eukaryotes containing viral signals have acted as the hosts of viruses.

In the present study, we systematically screened 1,901 genome assemblies from a diverse group of predominantly unicellular eukaryotes for mirusvirus signals. Of the 318 eukaryotic genera analyzed, 284 contained mirusvirus signals. The mirusviral MCP sequences identified in this study formed distinct phylogenetic clades, separate from previously reported mirusviral MCP sequences derived from marine environments. Moreover, three circular mirusviral genomes encoding a nearly complete set of marker genes were identified. This study suggests that the host range of the mirusviruses is broad, and that the mirusviruses display previously unrecognized diversity.

Results

Mirusviral marker sequences detected in a wide range of eukaryotes

The dataset for this study was collected from GenBank and comprised 1,901 eukaryotic genomic assemblies. These assemblies spanned over 318 minor lineages (mostly at the genus level, and are hereafter referred to as ‘genera’ for simplicity) and 16 major lineages of eukaryotic organisms (mostly at the phylum level or higher; Supplementary Table S1, Data S1), and cover eight of the nine eukaryotic supergroups.[15] In these assemblies, we identified nearly 118 million open reading frames (ORFs). Of these protein sequences, 6,659 showed significant sequence similarities (E-value < 10^{-5}) to the hidden Markov models (HMMs) of five selected mirusviral marker sequences: MCP, capsid triplex subunit 1 (Triplex1), capsid triplex subunit 2 (Triplex2), capsid portal protein (Portal), and capsid maturation protease (Maturation). To ensure their specificity to the mirusviruses, we aligned these sequences with other duplodnavirus (caudovirus and herpesvirus) sequences and HMMs in the PFAM database, and excluded possible false positives. In this way, we identified 6,042 marker sequences (541 MCP, 1,202 Portal, 509 Triplex1, 625 Triplex2, and 3,165 Maturation) that are specifically related to mirusvirus counterparts in 1,348 eukaryotic genome assemblies (71% of the analyzed assemblies).

The 1,348 eukaryotic genome assemblies containing viral marker sequences included 284 genera spanning 15 major lineages and eight eukaryotic supergroups (Data S2). Specifically, MCP was detected in 98 genera, Portal in 171 genera, Triplex1 in 114 genera, Triplex2 in 135 genera, and Maturation in 249 genera. Of these 284 genera, we selected 90 genera that showed strong signals for mirusviral marker sequences based on the criterion that the assemblies in the genus contained four or five distinct marker sequences (Fig. 1). Of the 6,042 marker sequences identified, 4,365 belonged to these 90 genera. These genera were spread across 11 major lineages and seven supergroups of eukaryotes.

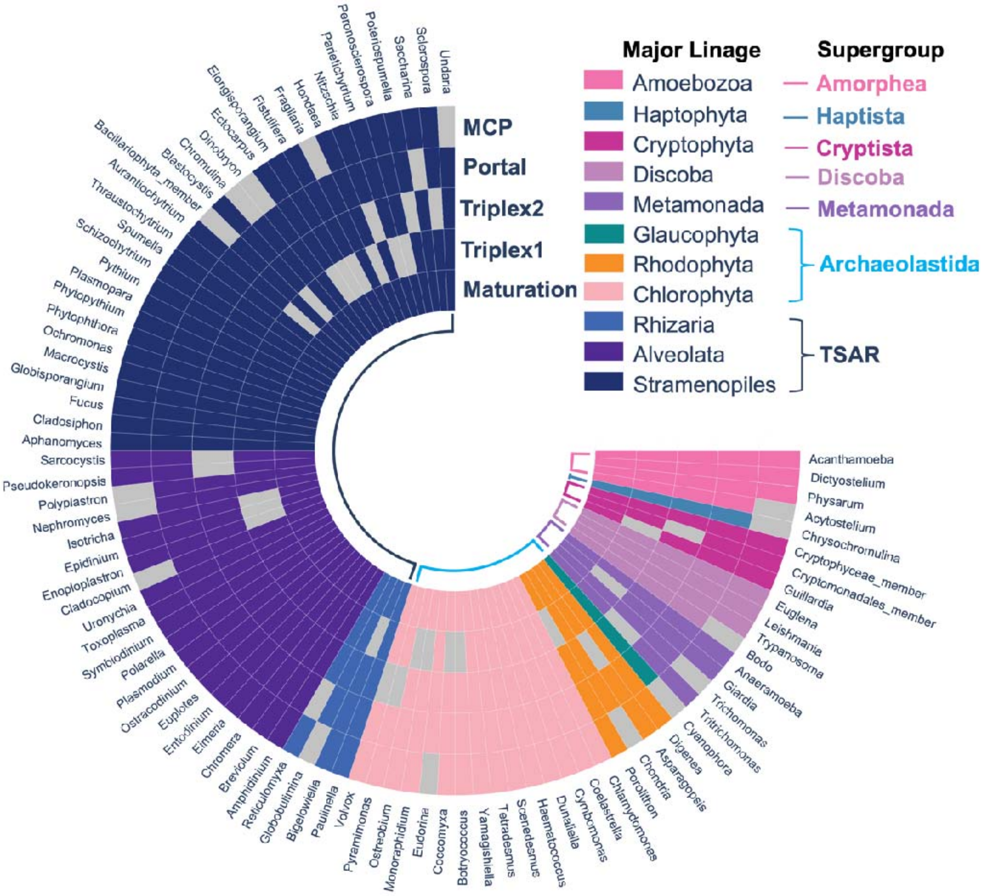


Fig. 1 Eukaryotic genera showing four or five different mirusvirus markers in

their assemblies. Colors of the cells represent the major lineages of the genus. Layers of cells represent the presence (colored) or absence (gray) of MCP, Portal, Triplex2, Triplex1, or Maturation sequences within each genus from the outermost to the innermost, respectively. Supergroups of eukaryotes are indicated in colored bold letters.

We investigated the sequence similarities between all 6,042 mirusviral marker sequences detected in the eukaryotic assemblies and reference marker sequences encoded in previously described mirusvirus genomes derived from marine metagenomes.[4] Most of the marker sequences shared low sequence similarity scores with the reference markers, and the median bit-score calculated with hmmscan ranged from 10 to 15 for all five marker genes (Fig. 2A). The median lengths of the marker sequences in the eukaryotic assemblies were smaller than those of the reference markers and ranged from 100 to 200 amino acids (Fig. 2B). These features suggest that many of the marker sequences detected were decaying nonfunctional genes (Fig. 2C). Nevertheless, there were also marker sequences of normal length but that shared low bit scores with reference sequences, implying the existence of diversified viral lineages.

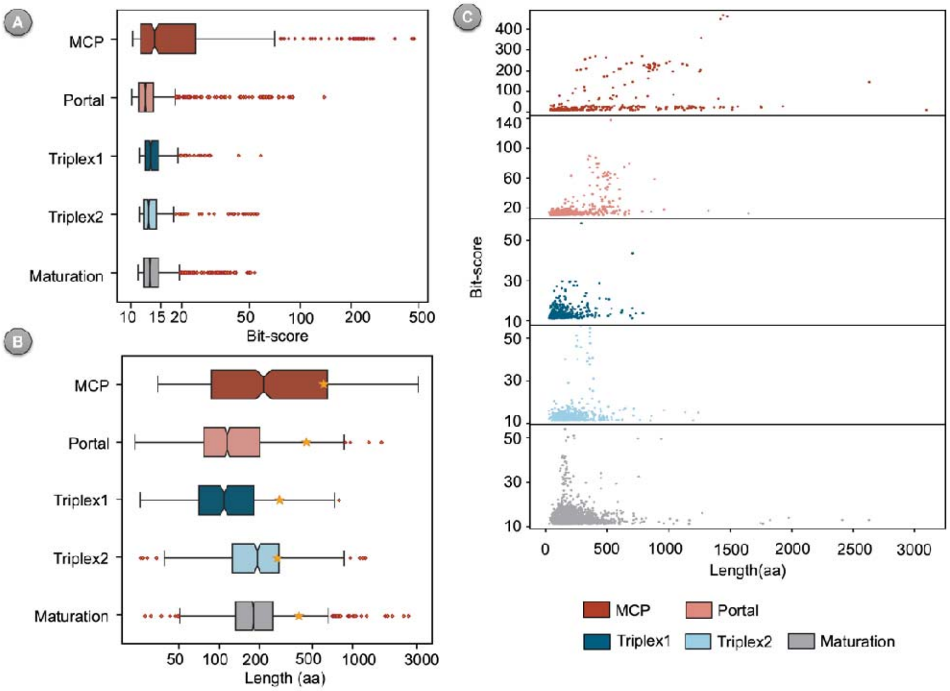


Fig. 2 Lengths of detected marker sequences and their similarities to reference sequences. (A) Boxplots of bit-scores against five reference marker genes. **(B)** Boxplots of the length distributions of the detected marker sequences. Yellow stars represent the median lengths of the reference mirusvirus orthogroups. **(A–B)** Red diamonds represent the outlier points. **(C)** Relationships between length and bit-score for each marker.

MCPs from eukaryotic assemblies form clades distinct from those of known mirusviruses

We next investigated the evolutionary relationships between the mirusvirus signals in the eukaryotic genome assemblies and previously described mirusviruses in marine metagenomes, focusing on MCP. This protein is the most important viral structural component and is critical for viral taxonomic classification.[1] Furthermore, MCP sequences are reported to represent an evolutionary path for the mirusviruses, similar to informational genes.[4] To minimize the interference degraded genes impose on phylogenetic analyses, we specifically filtered MCP marker sequences suitable for evolutionary analyses as follows. First, from the initial 541 MCP sequences identified in eukaryotic assemblies, we selected 262 sequences suitable for structural prediction. The selection was based on the absence of ambiguous amino acids (due to sequencing quality) and a length of 200–1,500 amino acids. We predicted the three-dimensional (3D) structures of these 262 MCPs and then compared their structural similarities with the marine mirusviral MCPs. Of the 262 structural models, 170 showed a template modeling score (TM-score) of > 0.5 (a criterion for the potential same fold[16,17]) with at least one reference mirusviral MCP, supporting their structural similarities.

We further refined our selection of MCPs from eukaryotic assemblies for phylogenetic analysis based on the conservation of functional domains. The HK97-fold is an important shared feature of duplodnaviral MCP structures. There are multiple conserved elements in the HK97-fold, including the axial domain (A-domain), peripheral domain (P-domain), extended loop (E-loop), and N-terminal arm (N-arm).[18] The floor domain, which includes the last three elements, exists in all HK97 MCPs. We used the E-loop, a long two-stranded β-sheet hairpin, as an indicator of the presence of the floor domain. Of the 170 sequences with high structural similarities to reference mirusviral MCPs, 150 were found to contain the two-stranded β-sheet hairpin with a β-strand longer than 10 amino acids on each side.

Finally, we compared the structures of these 150 MCP sequences with those of other representative duplodnaviral MCPs (from caudoviruses and herpesviruses). Of these 150 MCP sequences, 85 showed higher structural similarities to mirusviruses than to other viruses (Supplementary Fig. S1). Therefore, we considered these 85 MCP sequences as the most appropriate set of sequences for the subsequent phylogenetic analysis. These 85 MCP sequences were derived from 15 assemblies (14 species), including organisms from Alveolata, Amoebozoa, Chlorophyta, Cryptophyta, Rhizaria, and Stramenopiles (Table 1). Most of the homologs showed a bit-score of > 100 and a TM-score to the reference mirusvirus MCPs of > 0.7 (Supplementary Fig. S2). However, homologs from Rhizaria generally displayed low bit-scores, but a wide range of TM-scores to the reference mirusviral MCPs.

Table 1: Species and assemblies containing the 85 MCP homologs.

Species	Major lineage	Assembly	Location of isolation	Environment	Clade
Cymbomonas tetramitiformis	Chlorophyta	GCA_001247695.1	English Channel	Sea	E01
Aurantiochytrium sp.	Stramenopiles	GCA_001462505.1	Madeira	Sea	E02
Aurantiochytrium sp.	Stramenopiles	GCA_003116975.1	Hiroshima	Sea	E02
Schizochytrium sp.	Stramenopiles	GCA_004764695.1	missing	Sea	E02
Parietichytrium sp.	Stramenopiles	GCA_012862575.1	Okinawa	Sea	E02
Hondaea fermentalgiana	Stramenopiles	GCA_014084085.1	Mayotte	Mangrove	E02

Aurantiochytrium acetophilum	Stramenopiles	GCA_004332575.1	Biscayne Bay	Mangrove	E02
Ostreobium quekettii	Chlorophyta	GCA_905146915.1	Catanduanes Island	Sea	E02
Euplotes weissei	Alveolata	GCA_021440005.1	Qingdao	Sea	E02
Cryptophyta sp.	Cryptophyta	GCA_026770585.1	Northern Baffin Bay	Sea	E03
Uncultured Cryptomonadales	Cryptophyta	GCA_947538865.1	Římov Reservoir	Fresh water	E03
Guillardia theta	Cryptophyta	GCA_000315625.1	Connecticut	Sea	E03
Acanthamoeba pearcei	Amoebozoa	GCA_000826505.1	missing	missing	E04
Paulinella micropora	Rhizaria	GCA_009731375.1	Ibaraki	Fresh water	E05
Paulinella micropora	Rhizaria	GCA_019918135.1	Chungnam	Fresh water	E05

After all filters were applied, we reconstructed a phylogenetic tree comprising these 85 MCP homologs, together with 79 reference mirusviral MCPs derived from marine metagenomes (Fig. 3). The topology of the subtree of the reference mirusviral MCPs (clades M01 to M07) was generally consistent with a previous report.[4] The MCP homologs detected in eukaryotic assemblies were not grouped within the reference mirusvirus clades. Instead, they formed five distinct clades, E01 to E05, which were distantly related to the reference MCPs. The representative sequences for the individual clades (i.e., the longest homolog from each clade) displayed predicted 3D structures similar to those of the reference MCP, including the floor domain (Fig. 3). Notably, clades E01–E03 had an additional conserved antiparallel β-strand adjacent to the E-Loop.

The newly identified clades corresponded to distinct eukaryotic lineages, except

Clade E02. Clade E01 consisted of nine homologs from a single genomic assembly of the green algal species *C. tetramitiformis*.[11] Clade E02 was the only clade consisting of homologs (n = 13) from different organisms, including *Euplotes weissei* (Alveolata), *Ostreobium quekettii* (Chlorophyta), and six thraustochytrids (Stramenopiles) members.[10] Clade E03 consisted of 18 homologs from three Cryptophyta assemblies. Clade E04 contained a single homolog from an assembly of *Acanthamoeba pearcei*. Clade E05 consisted of 44 homologs from two different strains of *Paulinella micropora* (Rhizaria).

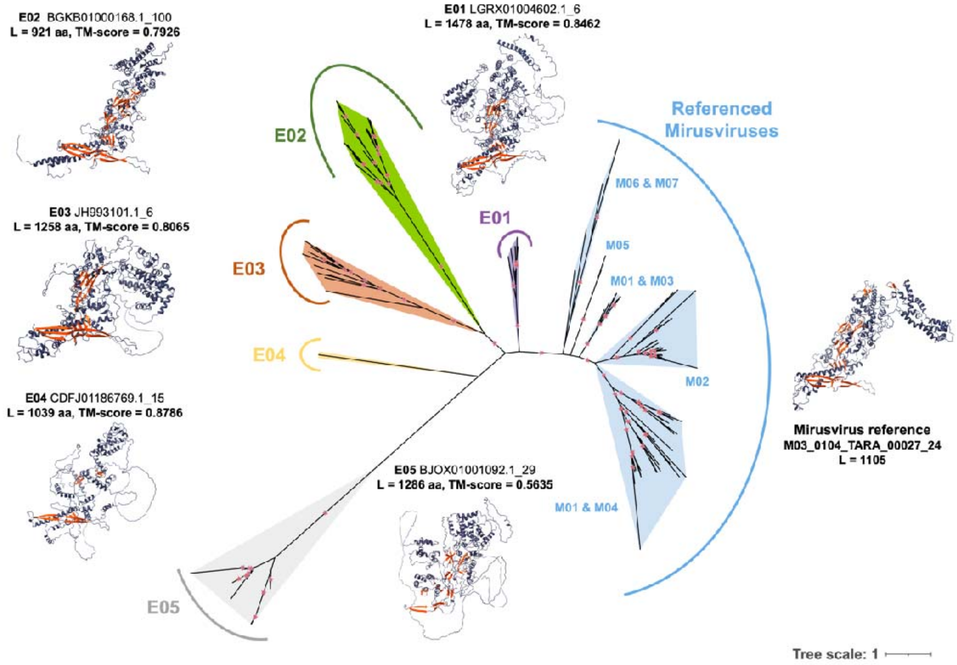


Fig. 3 Maximum-likelihood phylogenetic tree of MCPs. Different clades are indicated with different colors: blue represents the reference mirusviral MCPs. Predicted protein structures of the longest homolog within each clade are displayed around the tree. β-sheets are colored orange. Above the structure, the clade number is

followed by the scaffold from which the homolog originated and the serial ORF number predicted with prodigal. L, length of this homolog. TM-score is the highest TM-score to mirusvirus reference structures. Small red stars indicate branches with bootstrap support of > 95%. Best-fit model of this tree was Q.pfam+F+R6.

Existence of viral regions and three circular viral genomes

We also investigated the genomic context around the 85 most highly conserved MCP homologs. These homologs were found in 83 contigs or scaffolds (hereafter referred to as ‘scaffolds’ for simplicity). The lengths of these scaffolds ranged from 1.2 kilobases to 19.7 megabases. We analyzed all the predicted ORFs on these scaffolds and found that most of the scaffolds encoded many homologs of mirusviral genes, except the very short scaffolds (Fig. 4). Most of the scaffolds displayed a higher predicted ORF density than the average level within the same assembly (Supplementary Fig. S3).

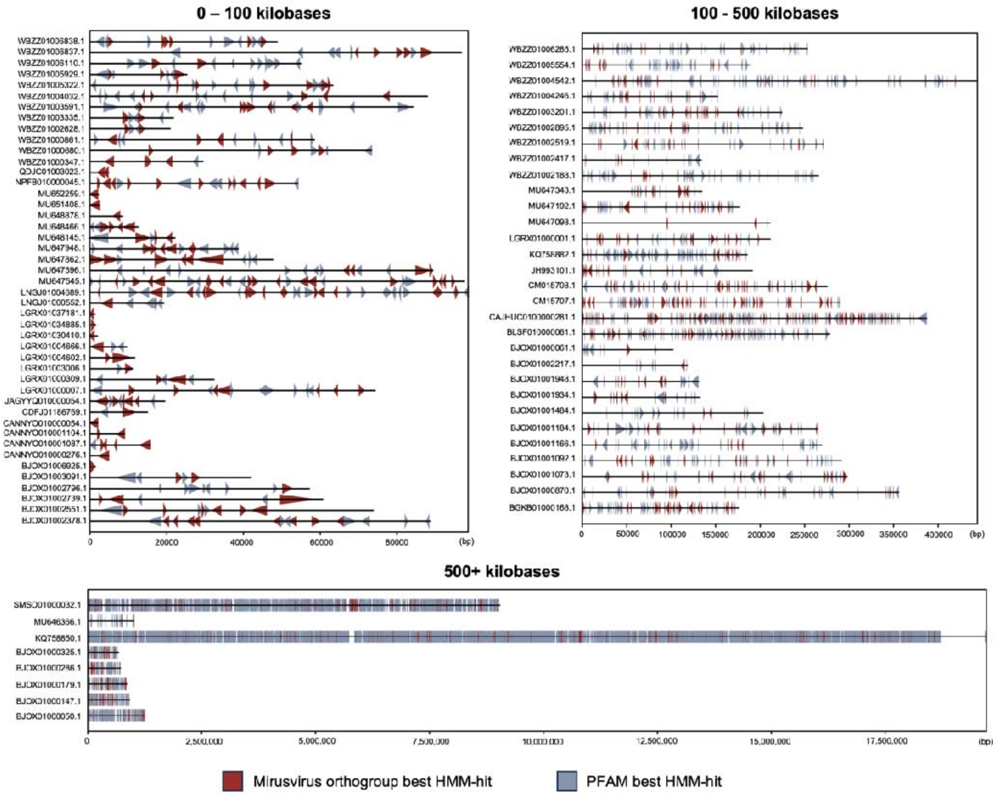


Fig. 4 Genomic maps of the 83 scaffolds. Scaffolds were divided into three groups according to their lengths. The predicted ORFs with mirusvirus orthogroup best hits are colored red. The predicted ORFs with PFAM best hits are colored blue. The tip of the triangle represents the direction of the ORF.

From database records at the National Center for Biotechnology Information (NCBI), we found that CM015707.1 and CM015708.1 are circular contigs from *Schizochytrium* sp. (Stramenopiles), with lengths of 289 kilobases and 275 kilobases, respectively. When we examined the remaining 81 scaffolds, we found that CAJHUC010000281.1 (from *Ostreobium quekettii*, Chlorophyta) is also likely to be a circular contig (325 kilobases) (Supplementary Fig. S4). These circular contigs encode 3–5 mirusvirus markers (of the five selected for this study) and additional functionally important proteins, such as terminases, DNA/RNA polymerases, and heliorhodopsins (Fig. 5). Notably, the three circular genomes lack RNA polymerase subunit B, even though RNA polymerase subunit B is the most commonly detected gene and is conserved in 89% of the reported mirusviral genomes. In contrast, the circular genomes encoded two copies of RNA polymerase subunit A. These circular contigs belonged to Clade E02. A family B DNA polymerase (PolB)-based phylogenetic analysis confirmed the close relationships between these three circular genomes (Supplementary Fig. S5). No split genes were detected in these circular contigs, indicating that most genes were intact and not pseudogenized.

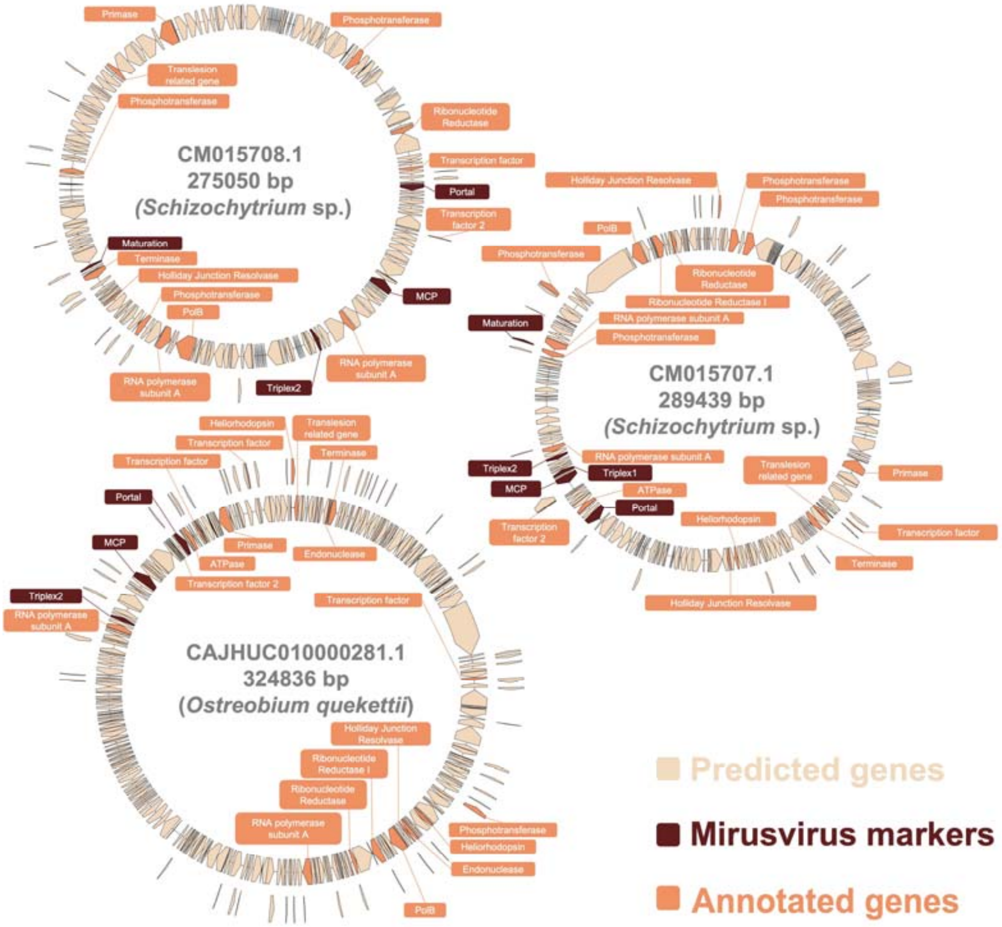


Fig. 5 Three circular mirusviral genomes recovered from eukaryotic assemblies. The five markers used in this study are colored brown. Functionally annotated genes are colored orange. Other predicted genes are colored yellowish.

Discussion

Mirusviruses are reported to be a deep-branching diverse group of dsDNA viruses, but previous studies have provided only limited information on their potential hosts in marine environemnts.[4,10,11] In the present study, we detected mirusviral maker sequence signals in a large number of eukaryotic (mostly unicellular) genome assemblies (1,348 assemblies; 71% of those analyzed). These eukaryotes included nearly 90% of the eukaryotic genera analyzed and 15 of the 16 major lineages analyzed (Data S2). These potential mirusvirus hosts included not only organisms living in marine environments (e.g., *Aurantiochytrium*, *Euplotes*) but also those living in other environments (e.g., freshwater: *Paulinella*, *Yamagishiella*; soil: *Dictyostelium*, *Physarum*; parasites of animals or plants: *Trypanosoma*, *Phytophthora*). Therefore, mirusviruses probably infect a broad spectrum of eukaryotes in various types of habitat.

Many of the mirusviral marker sequences identified in eukaryotic assemblies showed low sequence similarity scores to the reference mirusviral sequences (Fig. 2A). This sequence divergence from the reference sequences is probably attributable to the accumulation of mutations (e.g., pseudogenization[19]) after their ancient insertion into the host genomes. This is supported by the existence of abnormally short ORFs in the mirusviral marker sequences (Fig. 2B). Despite the presence of decaying genes, 90 genera displayed a comprehensive set of mirusviral marker genes (i.e., four or five markers). Such a high level of marker gene conservation may reflect the relatively recent insertion of the mirusviral genomes into the eukaryotic genomes (Fig. 1).

Our analysis also revealed a wide range of sequence similarity scores, even for ORFs in the normal size range (Fig. 2C). This suggests that factors beyond pseudogenization have contributed to the low sequence similarities between the newly detected sequences and the reference sequences. We analyzed the phylogeny of 85 selected MCP sequences that showed strong similarities (both in sequence and structure) to reference mirusviral MCPs and were thus considered to represent relatively recent insertions (Supplementary Fig. S2). These homologs were only distantly related to the reference mirusviral MCPs and formed five distinct clades (Fig. 3). Therefore, these MCP marker sequences probably originated from viral lineages that are distinct from the previously described mirusviral lineages.

Current research into mirusviruses is based on genomic data from environmental samples or eukaryotes, because no mirusvirus has yet been isolated. Despite their widespread presence in marine environments, the uncertainty regarding their original hosts poses challenges for isolation studies. In this study, we identified organisms showing strong and fresh mirusvirus signals. Notably, a three-stranded antiparallel β-sheet was observed in clade E01, E02, and E03 MCP homologs (Fig. 3). This structure consists of two β-strands in the E-loop at the N-terminus and an additional β-strand at the C-terminus,[1] and its presence indicates the high structural integrity of the MCP. A previous study reported a circular mirusviral genome in a thraustochytrids species.[10] We detected three circular mirusviral genomes, without evidence of pseudogenization, in another thraustochytrids species and a green algal species (Fig. 5). The circular status of the genomes suggests that these mirusviruses either latently[10,20] or persistently[21] infect their hosts. Eukaryotes harboring circular mirusviral genomes or high-integrity MCP sequences (E01–E03 in Table 1) represent promising candidates for the isolation of mirusviruses because their infections are probably recent or on-going events.

Viral genes are known to contribute to the critical evolution of their hosts.[22,23] The invasion of eukaryotic host genomes by DNA viral genomes is widespread among unicellular eukaryotes.[13] In particular, large dsDNA viruses have been reported to form giant viral endogenous elements within their host genomes.[24,25] The detection of mirusvirus signals in various eukaryotes indicates that mirusviruses are not an exception to this phenomenon (Data S1, S2). We identified multiple mirusvirus homologs in the scaffolds that harbored mirusviral MCPs (Fig. 4), suggesting that the genomic regions are giant endogenous viral elements originating from mirusvirus genomes. It has been suggested that giant viral endogenous elements confer unique capabilities upon their hosts (usually unicellular eukaryotes) by providing various genes of viral origin,[24] in a similar way to endogenous viral elements in higher eukaryotes.[26,27] Given the extensive presence of mirusvirus-like genomic regions in eukaryotic genomes, mirusviruses may have contributed to the genomic innovation of their hosts.[28]

Methods

Unicellular eukaryotic genome assembly data

We compiled the unicellular eukaryotic genome assembly data by retrieving all the relevant eukaryotic genome assemblies from the GenBank database (as of June 2023). Assemblies from Fungi (NCBI Taxonomy ID 4751), Metazoa (NCBI Taxonomy ID 33208), and Streptophyta (NCBI Taxonomy ID 35493) were excluded. A total of 1,901 assemblies were collected for subsequent analysis. Organisms that were not classified at the genus level were treated as distinct genera. To identify mirusvirus-originating sequences within these assemblies, gene calling was performed on all compiled data with the Prodigal v2.6.3 software, with the parameter '-p meta'.[29] Predicted sequences shorter than 20 amino acids or longer than 4,000 amino acids were discarded.

Creation of mirusvirus marker protein models

The reference metagenome-assembled mirusvirus genomes were obtained from a previous study.[4] Gene calling was performed with Prodigal, and orthologous groups were then generated with Orthofinder v2.5.2.[30] The core gene orthologous groups were annotated with a BLASTp search against the RefSeq database with Diamond v2.1.8,[31] and subsequent manual curation. For the five virion module protein markers, redundant sequences with > 90% sequence identity were removed from the five orthologous groups with cd-hit v4.8.1 and the parameter '-c 0.9'.[32] The HMMs of the five markers were built with HMMER v3.3.2. [32]

Detection mirusvirus marker sequences in eukaryotic assemblies

To detect mirusviral marker sequences within the eukaryotic assemblies, we used HMMs corresponding to five marker genes. We screened our predicted protein database, retaining hits with an E-value < 10^{-5} as determined with HMMER. To ensure the specificity of our results for mirusviruses, we established additional controls to exclude similar sequences from other viruses. First, we downloaded caudovirus and herpesvirus protein sequences from NCBI Virus (as of October 26, 2023). We then curated a dataset of homologs for the selected marker genes (for caudoviruses: MCP, Portal, Maturation; for herpesviruses: MCP, Triplex1, Triplex2, Portal, Maturation) using keyword filtering based on the NCBI annotations. Sequence redundancy was removed with cd-hit, with a cut-off of 90% sequence identity. We then constructed HMMs for each of the caudovirus and herpesvirus proteins and used these models to examine our mirusvirus hits. Any protein sequences from the eukaryotic assemblies that demonstrated a lower E-value when matched with the caudovirus or herpesvirus protein models than when compared with the mirusvirus models were excluded from further analysis. We also used hmmscan to scan the remaining mirusvirus hits against HMMs in the PFAM database (as of November 7, 2023),[33] and discarded hits with lower E-values against any PFAM HMM.

3D structural analyses of MCP homologs

Protein 3D structural predictions were made with AlphaFold v2.3.2 (-t 2023-11-14).[34] For structural comparisons with known HK97-fold MCPs, we also predicted the 3D structures of reference MCP sequences from mirusviruses, herpesviruses, and caudoviruses. Sequences shorter than 200 amino acids or with > 70% sequence identity to other sequences in the same viral group were removed from the reference MCP sequences. In total, 37 herpesvirus MCP models and 252 caudovirus MCP models were referenced. Structural similarities between proteins were calculated with Foldseek v6-29e2557 with the program ‘easy-search’.[35] Protein structures were visualized with ChimeraX v1.7.[36]

Phylogenetic analysis

Multiple-sequence alignments of MCP and PolB sequences longer than 200 amino acids were generated with Clustal-Omega v1.2.4.[37] We trimmed the alignments with trimAl v1.4.1, with the parameter ‘-gt 0.1’.[38] Maximum-likelihood phylogenetic trees were constructed with IQ-TREE v2.2.2.6,[39] with 1,000 ultrafast bootstrap replicates. Models were selected with ModelFinder.[40] The phylogenetic trees were visualized with iTOL v6.8.1.[41] For the phylogenetic tree of PolB, we collected all mirusviral PolB homologous sequences on those 83 scaffolds and also used eukaryotic reference sequences from a previous study.[42]

Annotation of contigs and scaffolds

We identified mirusvirus orthologous groups that were shared by more than 10 reference mirusviral metagenome-assembled genomes. Using the HMMs of these orthologous groups as the queries, we scanned all 83 scaffolds (E-value < 10^{-3}). We also used PFAM models as queries to scan these scaffolds (E-value < 10^{-3}). For the predicted genes that showed similarities to both mirusvirus orthologous groups and PFAM models, the annotation result was based on the hit with the lower E-value. To determine whether these scaffolds were likely to be circular, we performed a similarity comparison of different parts of the same sequence with NCBI web-based versions of BLASTn and DigAlign to determine whether the extremities of the sequence showed the same sequence (https://www.genome.jp/digalign/).[43] For the three circular mirusvirus contigs, we re-predicted their ORFs with the default parameters of Prodigal. The genomic maps of the circular contigs were generated with the Python Package 'DNA features viewer',[44] and the annotations were based on the best hits of the predicted ORFs to the mirusvirus orthogroups (E-value $< 10^{-3}$). To investigate pseudogenization, we used Diamond BLASTp to query all ORFs derived from the circular genomes against the RefSeq database, using the parameter '-k 3 -e 10'. We looked for adjacent ORFs showing similarity to different parts of the same reference sequence as potential signals of pseudogenization.

Acknowledgments

We thank Junyi Wu and Ruixuan Zhang for valuable discussions. This work was supported, in part, by JSPS/KAKENHI (nos 22H00384 and 19H05667). Computational time was provided by the Supercomputer System, Institute for Chemical Research, Kyoto University (Kyoto, Japan). Molecular graphics and analyses were performed with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (CA, USA), with support from National Institutes of Health (R01-GM129325) and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases (MD, USA). We thank Edanz (https://jp.edanz.com/ac) for editing a draft of this manuscript.

Author Contributions

HZ designed the work, performed all the analyses, and wrote the initial draft of the manuscript. LM contributed to the data preparation and assisted with the analyses. HH and HO supervised HZ and contributed to the finalization of the manuscript. All authors have read and approved the final version of the manuscript.

Declaration of Interests

The authors declare no competing interests.

Data and Code Availability

- The annotations, predicted ORFs, HMM files, alignments, predicted structure files, and tree files are available through GenomeNet (https://www.genome.jp/ftp/db/community/Mirusvirus_host_range/).
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the authors upon request.

References

1. Krupovic, M., and Koonin, E.V. (2017). Multiple origins of viral capsid proteins from cellular ancestors. Proceedings of the National Academy of Sciences *114*, E2401–E2410. 10.1073/pnas.1621061114.

2. Forterre, P. (2006). The origin of viruses and their possible roles in major evolutionary transitions. Virus Research *117*, 5–16. 10.1016/j.virusres.2006.01.010.

3. Koonin, E.V., Dolja, V.V., and Krupovic, M. (2015). Origins and evolution of viruses of eukaryotes: The ultimate modularity. Virology *479–480*, 2–25. 10.1016/j.virol.2015.02.039.

4. Gaïa, M., Meng, L., Pelletier, E., Forterre, P., Vanni, C., Fernandez-Guerra, A., Jaillon, O., Wincker, P., Ogata, H., Krupovic, M., et al. (2023). Mirusviruses link herpesviruses to giant viruses. Nature *616*, 783–789. 10.1038/s41586-023-05962-4.

5. Schulz, F., Abergel, C., and Woyke, T. (2022). Giant virus biology and diversity in the era of genome-resolved metagenomics. Nat Rev Microbiol, 1–16. 10.1038/s41579-022-00754-5.

6. Pagaling, E., Haigh, R.D., Grant, W.D., Cowan, D.A., Jones, B.E., Ma, Y., Ventosa, A., and Heaphy, S. (2007). Sequence analysis of an Archaeal virus isolated from a hypersaline lake in Inner Mongolia, China. BMC Genomics *8*, 410. 10.1186/1471-2164-8-410.

7. Piel, D., Bruto, M., Labreuche, Y., Blanquart, F., Goudenège, D., Barcia-Cruz, R., Chenivesse, S., Le Panse, S., James, A., Dubert, J., et al. (2022). Phage–host coevolution in natural populations. Nat Microbiol *7*, 1075–1086. 10.1038/s41564-022-01157-1.

8. Bandín, I., and Dopazo, C.P. (2011). Host range, host specificity and hypothesized host shift events among viruses of lower vertebrates. Veterinary Research *42*, 67. 10.1186/1297-9716-42-67.

9. Rosani, U., Gaia, M., Delmont, T.O., and Krupovic, M. (2023). Tracing the invertebrate herpesviruses in the global sequence datasets. Frontiers in Marine Science *10*.

10. Collier, J.L., Rest, J.S., Gallot-Lavallée, L., Lavington, E., Kuo, A., Jenkins, J., Plott, C., Pangilinan, J., Daum, C., Grigoriev, I.V., et al. (2023). The protist Aurantiochytrium has universal subtelomeric rDNAs and is a host for mirusviruses. Current Biology. 10.1016/j.cub.2023.10.009.

11. Gyaltshen, Y., Rozenberg, A., Paasch, A., Burns, J.A., Warring, S., Larson, R.T., Maurer-Alcalá, X.X., Dacks, J., Narechania, A., and Kim, E. (2023). Long-Read-Based Genome Assembly Reveals Numerous Endogenous Viral Elements in the Green Algal Bacterivore Cymbomonas tetramitiformis. Genome Biol Evol *15*, evad194. 10.1093/gbe/evad194.

12. Irwin, N.A.T., Pittis, A.A., Richards, T.A., and Keeling, P.J. (2022). Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. Nat Microbiol *7*, 327–336. 10.1038/s41564-021-01026-3.

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

14. Zhao, H., Zhang, R., Wu, J., Meng, L., Okazaki, Y., Hikida, H., and Ogata, H. (2023). A 1.5-Mb continuous endogenous viral region in the arbuscular mycorrhizal fungus Rhizophagus irregularis. Virus Evolution *9*, vead064. 10.1093/ve/vead064.

15. Burki, F., Roger, A.J., Brown, M.W., and Simpson, A.G.B. (2020). The New Tree of Eukaryotes. Trends in Ecology & Evolution *35*, 43–55. 10.1016/j.tree.2019.08.008.

16. Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res *33*, 2302–2309. 10.1093/nar/gki524.

17. Xu, J., and Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics *26*, 889–895. 10.1093/bioinformatics/btq066.

18. Duda, R.L., and Teschke, C.M. (2019). The amazing HK97 fold: versatile results of modest differences. Curr Opin Virol *36*, 9–16. 10.1016/j.coviro.2019.02.001.

19. Katzourakis, A., and Gifford, R.J. (2010). Endogenous Viral Elements in Animal Genomes. PLoS Genet *6*, e1001191. 10.1371/journal.pgen.1001191.

20. Cohen, J.I. (2020). Herpesvirus latency. J Clin Invest *130*, 3361–3369. 10.1172/JCI136225.

21. Blanc-Mathieu, R., Dahle, H., Hofgaard, A., Brandt, D., Ban, H., Kalinowski, J., Ogata, H., and 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. 10.1128/JVI.02446-20.

22. Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature *405*, 299–304. 10.1038/35012500.

23. Koonin, E.V., and Krupovic, M. (2018). The depths of virus exaptation. Current Opinion in Virology *31*, 1–8. 10.1016/j.coviro.2018.07.011.

24. Moniruzzaman, M., Weinheimer, A.R., Martinez-Gutierrez, C.A., and Aylward, F.O. (2020). Widespread endogenization of giant viruses shapes genomes of green algae. Nature *588*, 141–145. 10.1038/s41586-020-2924-2.

25. Moniruzzaman, M., Erazo-Garcia, M.P., and Aylward, F.O. (2022). Endogenous giant viruses contribute to intraspecies genomic variability in the model green alga Chlamydomonas reinhardtii. Virus Evolution *8*, veac102. 10.1093/ve/veac102.

26. Fujino, K., Horie, M., Honda, T., Merriman, D.K., and Tomonaga, K. (2014). Inhibition of Borna disease virus replication by an endogenous bornavirus-like element in the ground squirrel genome. Proceedings of the National Academy of Sciences *111*, 13175–13180. 10.1073/pnas.1407046111.

27. Suzuki, Y., Frangeul, L., Dickson, L.B., Blanc, H., Verdier, Y., Vinh, J., Lambrechts, L., and Saleh, M.-C. (2017). Uncovering the Repertoire of Endogenous Flaviviral Elements in Aedes Mosquito Genomes. Journal of Virology *91*, 10.1128/jvi.00571-17. 10.1128/jvi.00571-17.

28. Aylward, F.O. (2023). Microbiology: The curious case of the mysterious mirusvirus. Current Biology *33*, R1234–R1235. 10.1016/j.cub.2023.10.037.

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

30. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology *20*, 238. 10.1186/s13059-019-1832-y.

31. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat Methods *12*, 59–60. 10.1038/nmeth.3176.

32. Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLOS Computational Biology *7*, e1002195. 10.1371/journal.pcbi.1002195.

33. Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nucleic Acids Research *42*, D222. 10.1093/nar/gkt1223.

34. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. 10.1038/s41586-021-03819-2.

35. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., and Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. Nat Biotechnol, 1–4. 10.1038/s41587-023-01773-0.

36. Meng, E.C., Goddard, T.D., Pettersen, E.F., Couch, G.S., Pearson, Z.J., Morris, J.H., and Ferrin, T.E. (2023). UCSF ChimeraX: Tools for structure building and analysis. Protein Science *32*, e4792. 10.1002/pro.4792.

37. Sievers, F., and Higgins, D.G. (2018). Clustal Omega for making accurate alignments of many protein sequences. Protein Sci *27*, 135–145. 10.1002/pro.3290.

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

39. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and 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. 10.1093/molbev/msaa015.

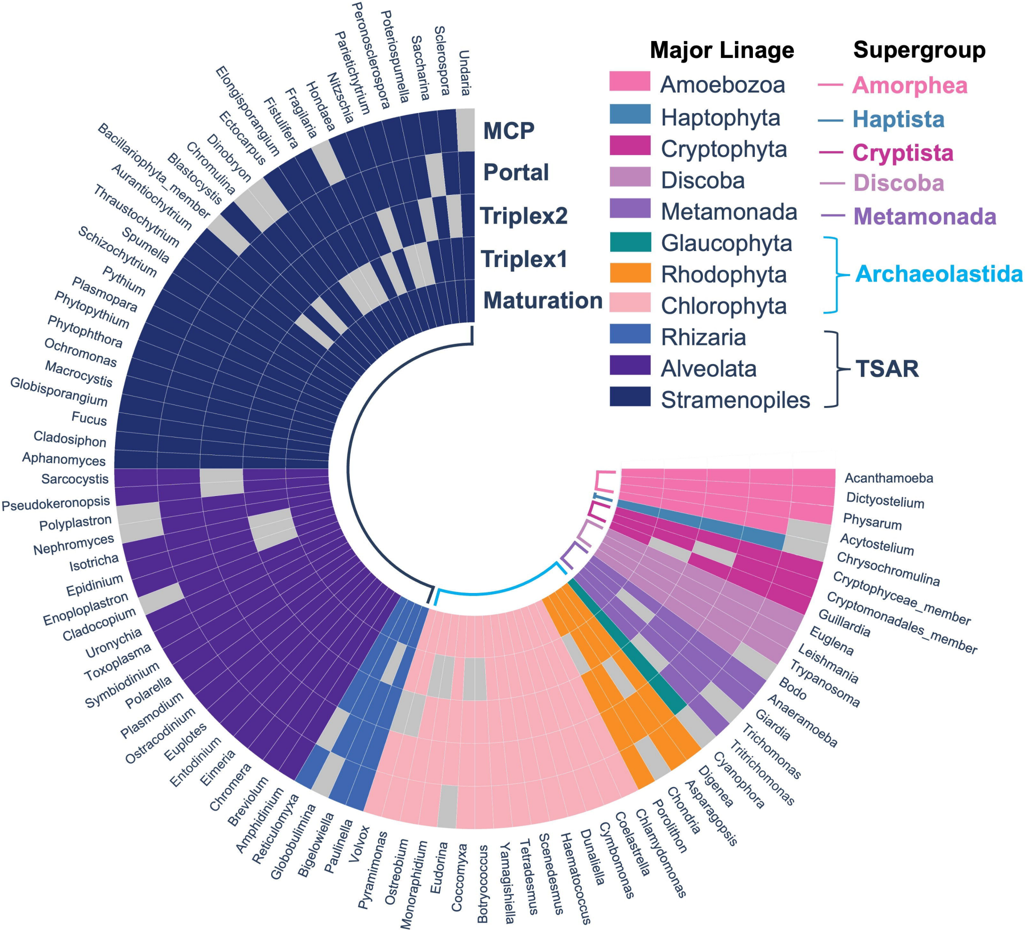
40. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods *14*, 587–589. 10.1038/nmeth.4285.

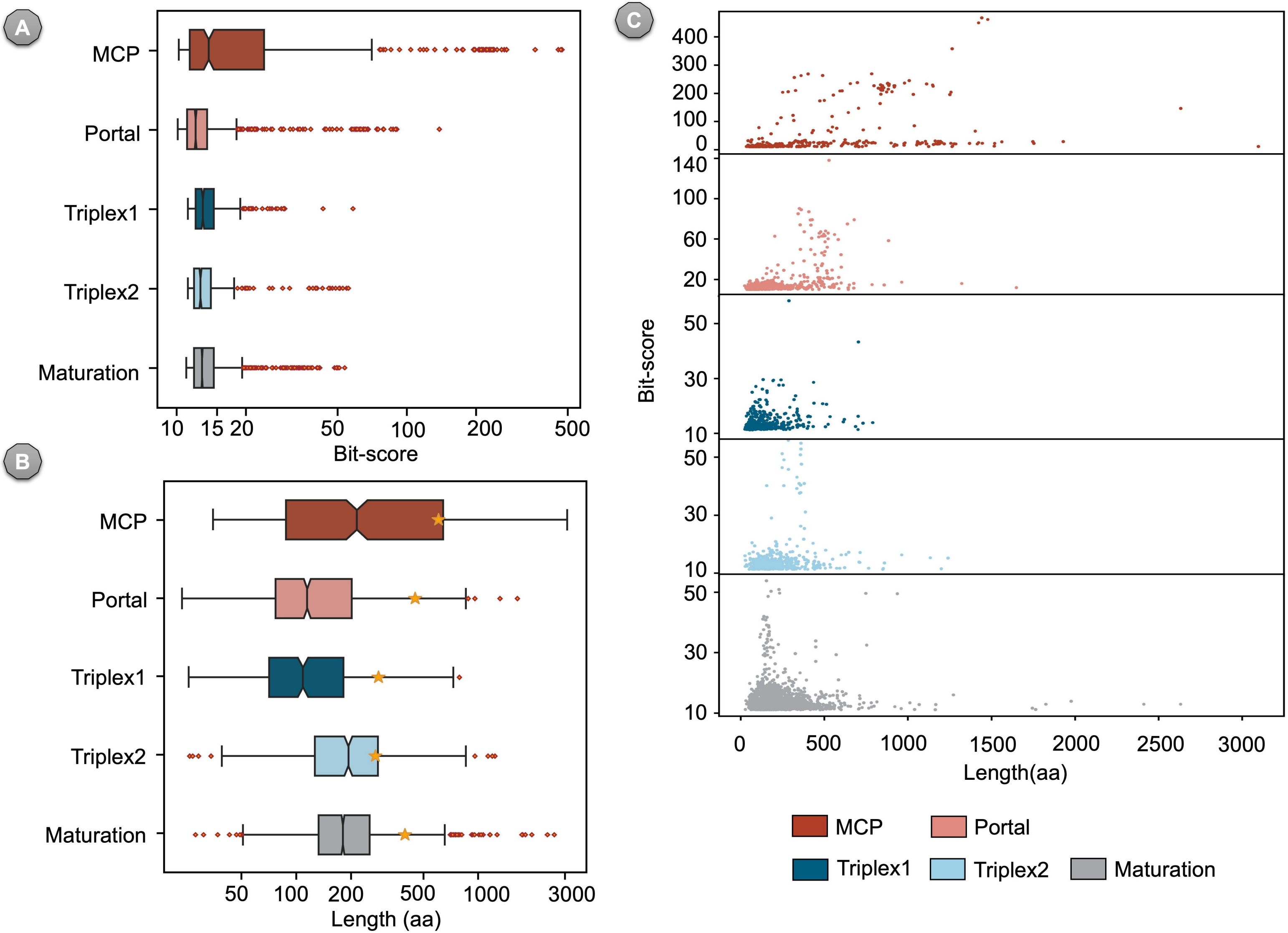
41. Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics *23*, 127–128. 10.1093/bioinformatics/btl529.

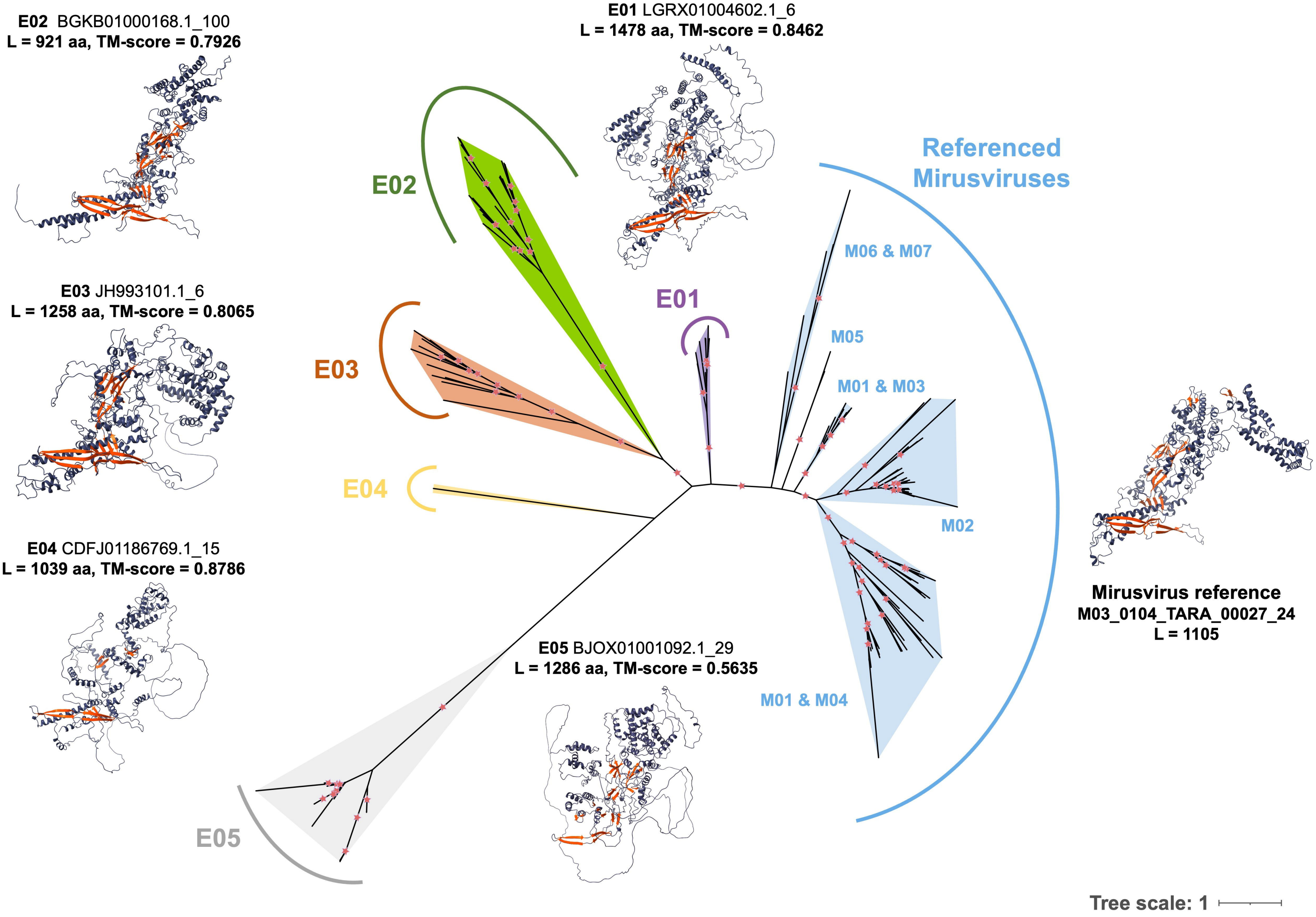
42. Kazlauskas, D., Krupovic, M., Guglielmini, J., Forterre, P., and Venclovas, Č. (2020). Diversity and evolution of B-family DNA polymerases. Nucleic Acids Res *48*, 10142–10156. 10.1093/nar/gkaa760.

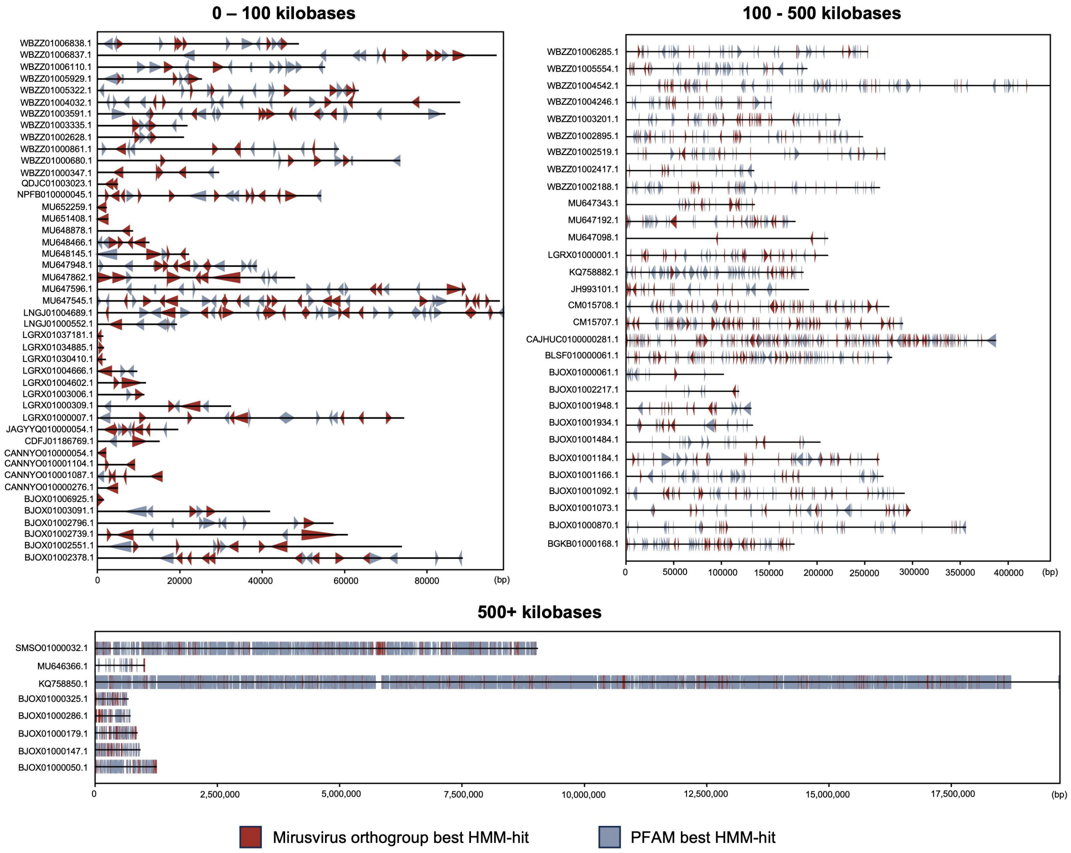
43. Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T.L. (2008). NCBI BLAST: a better web interface. Nucleic Acids Research *36*, W5–W9. 10.1093/nar/gkn201.

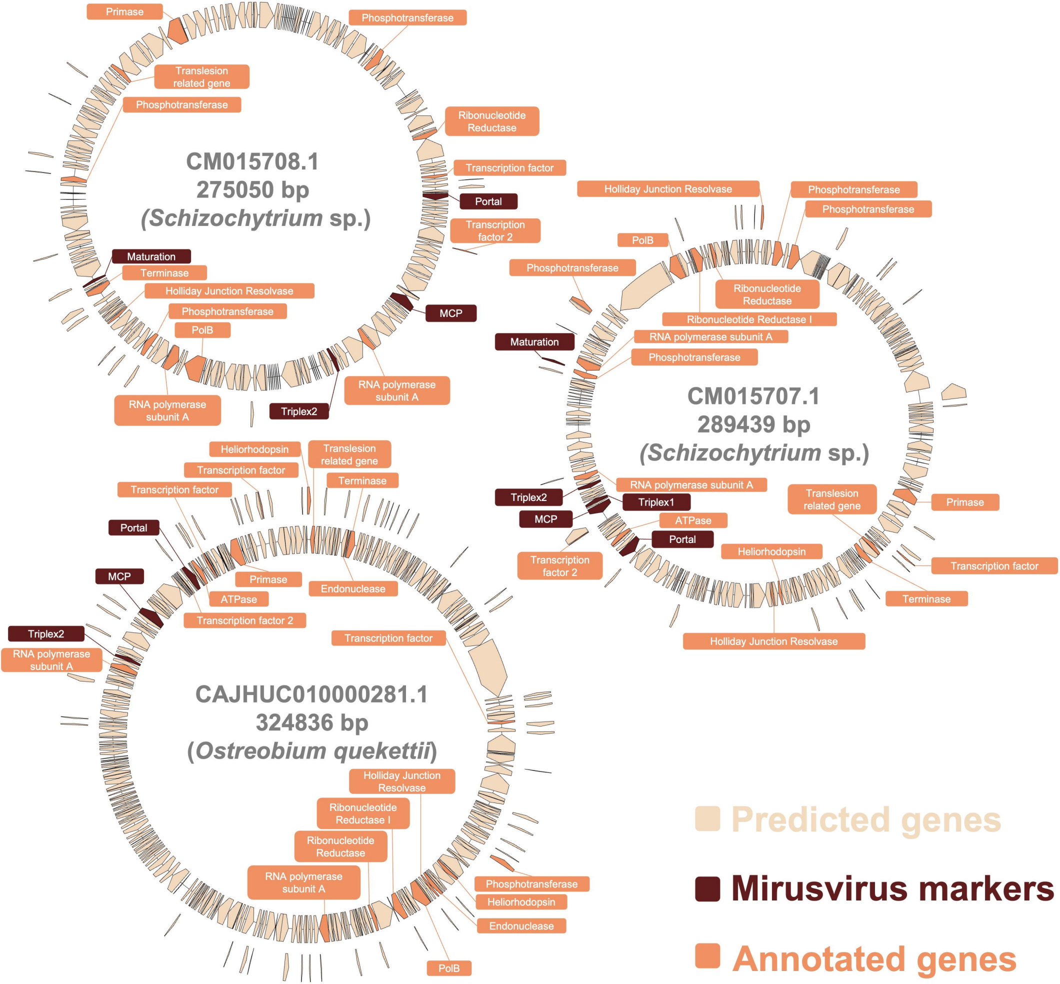
44. Zulkower, V., and Rosser, S. (2020). DNA Features Viewer: a sequence annotation formatting and plotting library for Python. Bioinformatics *36*, 4350–4352. 10.1093/bioinformatics/btaa213.


Major Linage
Amoebozoa
Haptophyta
Cryptophyta
Discoba
Metamonada
Glaucophyta
Rhodophyta
Chlorophyta
Rhizaria
Alveolata
Stramenopiles
Supergroup
Amorphea
Haptista
Cryptista
Discoba
Metamonada
Archaeolastida
TSAR
MCP
Portal
Triplex2
Triplex1
Maturation
Undaria
Sclerospora
Saccharina
Poteriospumella
Peronosclerospora
Parietichytrium
Nitzschia
Hondaea
Fragilaria
Fistulifera
Elongisporangium
Ectocarpus
Dinobryon
Chromulina
Blastocystis
Bacillariophyta_member
Aurantiochytrium
Thraustochytrium
Spumella
Schizochytrium
Pythium
Plasmopara
Phytopythium
Phytophthora
Ochromonas
Macrocystis
Globisporangium
Fucus
Cladosiphon
Aphanomyces
Sarcocystis
Pseudokeronopsis
Polyplastron
Nephromyces
Isotricha
Epidinium
Enoploplastron
Cladocopium
Uronychia
Toxoplasma
Symbiodinium
Polarella
Plasmodium
Ostracodinium
Euplotes
Entodinium
Eimeria
Chromera
Breviolum
Amphidinium
Reticulomyxa
Globobulimina
Bigelowiella
Paulinella
Volvox
Pyramimonas
Ostreobium
Monoraphidium
Eudorina
Coccomyxa
Botryococcus
Yamagishiella
Tetradesmus
Scenedesmus
Haematococcus
Dunaliella
Cymbomonas
Coelastrella
Chlamydomonas
Porolithon
Chondria
Asparagopsis
Digenea
Cyanophora
Tritrichomonas
Trichomonas
Giardia
Anaeramoeba
Bodo
Trypanosoma
Leishmania
Euglena
Guillardia
Cryptomonadales_member
Cryptophyceae_member
Chrysochromulina
Acytostelium
Physarum
Dictyostelium
Acanthamoeba


A
MCP
Portal
Triplex1
Triplex2
Maturation
10
15
20
50
100
200
500
Bit-score
B
MCP
Portal
Triplex1
Triplex2
Maturation
50
100
200
500
1000
3000
Length (aa)
C
400
300
200
100
0
140
100
60
20
50
30
10
50
30
10
50
30
10
Bit-score
0
500
1000
1500
2000
2500
3000
Length(aa)
MCP
Portal
Triplex1
Triplex2
Maturation


E02 BGKB01000168.1_100
L = 921 aa, TM-score = 0.7926
E01 LGRX01004602.1_6
L = 1478 aa, TM-score = 0.8462
E03 JH993101.1_6
L = 1258 aa, TM-score = 0.8065
E04 CDFJ01186769.1_15
L = 1039 aa, TM-score = 0.8786
E05 BJOX01001092.1_29
L = 1286 aa, TM-score = 0.5635
E02
E03
E04
E05
E01
Referenced Mirusviruses
M06 & M07
M05
M01 & M03
M02
M01 & M04
Mirusvirus reference
M03_0104_TARA_00027_24
L = 1105
Tree scale: 1


0 – 100 kilobases
WBZZ01006838.1
WBZZ01006837.1
WBZZ01006110.1
WBZZ01005929.1
WBZZ01005322.1
WBZZ01004032.1
WBZZ01003591.1
WBZZ01003335.1
WBZZ01002628.1
WBZZ01000861.1
WBZZ01000680.1
WBZZ01000347.1
QDJC01003023.1
NPFB010000045.1
MU652259.1
MU651408.1
MU648878.1
MU648466.1
MU648145.1
MU647948.1
MU647862.1
MU647596.1
MU647545.1
LNGJ01004689.1
LNGJ01000552.1
LGRX01037181.1
LGRX01034885.1
LGRX01030410.1
LGRX01004666.1
LGRX01004602.1
LGRX01003006.1
LGRX01000309.1
LGRX01000007.1
JAGYYQ010000054.1
CDFJ01186769.1
CANNYO010000054.1
CANNYO010001104.1
CANNYO010001087.1
CANNYO010000276.1
BJOX01006925.1
BJOX01003091.1
BJOX01002796.1
BJOX01002739.1
BJOX01002551.1
BJOX01002378.1
0
20000
40000
60000
80000
(bp)
100 - 500 kilobases
WBZZ01006285.1
WBZZ01005554.1
WBZZ01004542.1
WBZZ01004246.1
WBZZ01003201.1
WBZZ01002895.1
WBZZ01002519.1
WBZZ01002417.1
WBZZ01002188.1
MU647343.1
MU647192.1
MU647098.1
LGRX01000001.1
KQ758882.1
JH993101.1
CM015708.1
CM15707.1
CAJHUC0100000281.1
BLSF010000061.1
BJOX01000061.1
BJOX01002217.1
BJOX01001948.1
BJOX01001934.1
BJOX01001484.1
BJOX01001184.1
BJOX01001166.1
BJOX01001092.1
BJOX01001073.1
BJOX01000870.1
BGKB01000168.1
0
50000
100000
150000
200000
250000
300000
350000
400000
(bp)
500+ kilobases
SMSO01000032.1
MU646366.1
KQ758850.1
BJOX01000325.1
BJOX01000286.1
BJOX01000179.1
BJOX01000147.1
BJOX01000050.1
0
2,500,000
5,000,000
7,500,000
10,000,000
12,500,000
15,000,000
17,500,000
(bp)
🟥 Mirusvirus orthogroup best HMM-hit
🟦 PFAM best HMM-hit


CM015708.1
275050 bp
(*Schizochytrium* sp.)
Primase
Phosphotransferase
Translesion related gene
Phosphotransferase
Ribonucleotide Reductase
Transcription factor
Portal
Transcription factor 2
Maturation
Terminase
Holliday Junction Resolvase
Phosphotransferase
PolB
MCP
RNA polymerase subunit A
Triplex2
RNA polymerase subunit A
CM015707.1
289439 bp
(*Schizochytrium* sp.)
Holliday Junction Resolvase
Phosphotransferase
Phosphotransferase
PolB
Phosphotransferase
Ribonucleotide Reductase
Ribonucleotide Reductase I
RNA polymerase subunit A
Maturation
Phosphotransferase
RNA polymerase subunit A
Triplex2
Triplex1
Translesion related gene
Primase
MCP
ATPase
Portal
Transcription factor 2
Heliorhodopsin
Transcription factor
Terminase
Holliday Junction Resolvase
CAJHUC010000281.1
324836 bp
(*Ostreobium quekettii*)
Heliorhodopsin
Translesion related gene
Transcription factor
Terminase
Transcription factor
Portal
MCP
Primase
Endonuclease
ATPase
Transcription factor 2
Triplex2
RNA polymerase subunit A
Transcription factor
Holliday Junction Resolvase
Ribonucleotide Reductase I
Ribonucleotide Reductase
RNA polymerase subunit A
Phosphotransferase
Heliorhodopsin
Endonuclease
PolB
Predicted genes
Mirusvirus markers
Annotated genes